WILEY

Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics

1

Genetics LYNN B. JORDE

Genomics PETER F. R. LITTLE

Proteomics MICHAEL J. DUNN

Bioinformatics SHANKAR SUBRAMANIAM

VISIT…

LANZAROTE Caliente.com

Introductory Review

Population genomics: patterns of genetic variation within populations

Greg Gibson
North Carolina State University, Raleigh, NC, USA

1. Polymorphism

Polymorphism at the nucleotide level ranges over at least an order of magnitude within species, and average polymorphism ranges over two orders of magnitude between species. *Homo sapiens* is among the least polymorphic of all species, with a heterozygous single nucleotide polymorphism (SNP) generally occurring once every 500 to 1000 bp (International SNP Map Working Group, 2001). By contrast, marine invertebrates such as the sea squirt and echinoderms have an astonishing level of sequence diversity with a SNP every 5 to 10 bp (Dehal *et al*., 2002). Diversity is a function of organism-level factors such as population size, generation time, and breeding structure (Aquadro *et al*., 2001), but variation within and among chromosomes signifies that recombination and mutation rates are also critical (Begun and Aquadro, 1992; Charlesworth *et al*., 1995). In most species, centromeric and telomeric regions are less recombinogenic, hence have smaller effective population sizes, and tend to be less polymorphic (Nachman, 2002). Even within a locus, polymorphism can vary over an order of magnitude, according primarily to functional constraint: synonymous substitution rates tend to be uniform, whereas replacements can be excluded from highly conserved domains. Noncoding gene sequences are typically more polymorphic than exons and less polymorphic than intergenic DNA, but core regulatory sequences up to several hundred basepairs in length may often be the most conserved of all sequences (Wray *et al*., 2003).

Significant disparity between two measures of polymorphism, namely, the number of segregating sites and the average heterozygosity, provides evidence for departure from "neutrality" (Hudson *et al*., 1987; Kreitman, 2000). However, neutrality comes in many flavors, and demographic processes are just as likely to affect the difference between these two measures as is selection (Nielsen, 2001). Heterozygosity is a function of allele frequency as well as density, so unexpectedly high or low numbers of heterozygotes relative to the number of SNPs in a population can arise as a result of several processes that may be superimposed on random drift. Thus, rapid population expansion or strong purifying selection both reduce

heterozygosity, whereas admixture or balancing selection will increase heterozygosity. Tests such as Tajima's D (Tajima, 1989) have remained useful descriptors of diversity, but have been joined by a new series of tests that are more firmly rooted in coalescent theory (Wall and Hudson, 2001). Rather than strictly interpreting test scores relative to theoretical expectations, comparison of the distribution of test scores across tens or hundreds of loci among species emphasizes that diversity is affected by a complex interplay of factors and that it is the location of a gene at either extreme of the continuum that marks it as a candidate target of selection, rather than a p-value per se (Hey, 1999; Bustamante *et al*., 2002).

A trend toward empirical evaluation of significance by permutation in light of genomic data is also seen in relation to population structure. Standard F-statistics introduced by Sewall Wright based on differences in genotype frequencies among populations (Weir and Hill, 2002) have been extended into an analysis of molecular variance (AMOVA) framework, one popular implementation of which is the Arlequin software (Schneider *et al*., 2000). Estimates of SNP, indel, haplotype, or microsatellite allele frequency differences are sensitive to sample size, so samples of at least 100 individuals per population are recommended. Using genomic data, the multiple comparison issue also arises: in a set of 500 sites, a single site with a testwise p-value of 0.0001 is not unexpected, but in a large sample this may correspond to an allele frequency difference of just 10%. Consequently, population structure is best estimated from multilocus data. For example, Pritchard *et al*. (2000) have introduced Bayesian statistics to assign individuals to likely subpopulations with numerous applications in evolutionary, conservation, quantitative, and human genetics. It is well known that over 90% of all human polymorphism is common to all populations, but the ability to genotype hundreds of loci has led to the recognition that given sufficient data there is a detectable signature of demographic history even in our species (Rosenberg *et al*., 2002). Similarly, long-held assumptions of panmixia in *Drosophila melanogaster* are being challenged by deeper sampling (Glinka *et al*., 2003), as are commonly held notions about the genetic uniformity of crops such as maize (Matsuoka *et al*., 2002), and in fact the power to discriminate population structure in most species will have a profound impact on quantitative biology. An important implication of the ability to detect population structure is inference of departure from neutrality, by comparison of the observed F-statistics with those obtained from a collection of assumed neutral markers (Lewontin and Krakauer, 1973; Rockman *et al*., 2003).

The advent of new sequencing and genotyping technologies will only accelerate the data-driven nature of evolutionary genetic research (*see* Article 7, **Single molecule array-based sequencing**, Volume 3). ABI 3730 automated DNA sequencing machines routinely generate traces with over 1 kb of high-quality sequence and have a throughput capacity exceeding 1 Mb per day. Single-molecule sequencing methods are expected to make the sequencing of complete eukaryotic genomes for $1000 each a reality, possibly in the next decade (Meldrum, 2000), while massively parallel resequencing by hybridization to wafers of tiled oligonucleotides has already been used to characterize polymorphism between primate species (Frazer *et al*., 2003). Such studies have identified hundreds of loci that are candidates for the adaptive evolution in the recent human lineage, some of which are likely to contribute to the etiology of common disease (Tishkoff and Verrilli, 2003;

Clark *et al*., 2003). Molecular evolutionary studies of single genes in samples of 30 individuals have been typical but will soon be dwarfed by genome-scale sampling, and increasingly, attention will be placed on the efficient sampling design and formulation of hypotheses that utilize patterns of variation across the genome to interpret unusual patterns of variation at focal loci. Describing the variance of standard population-genetic parameters at a genome-wide scale is unprecedented territory, and developing approaches to quantify this variation across these expansive contiguous regions is the challenge for the near future. This type of data will also allow reexamination of some of the most basic assumptions underlying many population-genetic approaches, such as the infinite sites and island migration models.

2. Recombination and linkage disequilibrium

Recombination and mutation are the two biochemical processes that influence the distribution of molecular variation. Recombination can be directly measured by monitoring the coinheritance of markers transmitted from parent to offspring, but with the exception of technically demanding single sperm typing (Jeffreys *et al*., 2000); the resolution of this method is of the order of just centimorgans or hundreds of kilobases. Since an important consequence of recombination is its effect on linkage disequilibrium over scales from tens of bases to tens of kilobases, indirect methods for measuring recombination have been introduced based on population-genetic measurement of the cosegregation of markers (Hudson and Kaplan, 1985; Stumpf and McVean, 2003). Linkage disequilibrium (LD) is the nonrandom assortment of genetic markers: given two alleles each at a frequency of 20%, just 4% of individual chromosomes should have both alleles if assortment is random, but physically adjacent markers will often cosegregate more often. In this case, the maximum possible LD would have 20% of the chromosomes with both less common alleles, and 80% with both common alleles. Two commonly used statistics measure this departure from randomness, D' and r^2, only the latter of which explicitly takes allele frequencies into account (Hill and Robertson, 1966; Weir, 1996). A further technical challenge in the measurement of LD is establishing the linkage phase of double heterozygotes, which can be addressed directly by studying trios of parents and their offspring (which is however impractical for many species) or computationally with EM likelihood algorithms (Fallin and Schork, 2000; Stephens *et al*., 2001).

Quantitative geneticists have long been interested in LD because detection of association between markers and phenotypes is dependent on LD between anonymous markers and the causative disease or quantitative trait nucleotide(s) (Zondervan and Cardon, 2004). This idea has given rise to the human HapMap project, which is an effort to describe the complete pattern of haplotypes in the human genome (International HapMap Consortium, 2003). Haplotypes are sets of multilocus alleles, and because of LD they tend to be less common than chance would predict: there are 32 possible ways that five biallelic alleles can combine, but typically just a handful of these will be at any appreciable frequency in a population. Standard population-genetic theory predicts that LD should decay monotonically with distance, but at least in the human genome it now appears that there are often

fairly discrete boundaries that define haplotype blocks that range in length from 10 to 100 kb or more (Gabriel *et al*., 2002; *see also* Article 12, **Haplotype mapping**, Volume 3 and Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4). Consequently, while there are in excess of 5 million SNPs in the human genome, there may be as few as 50 000 common haplotype blocks, and consequently it is argued that a similar number of markers will be sufficient to perform genome scans for association with disease (Risch and Merikangas, 1996). According to the common disease–common variant hypothesis, the polymorphisms that contribute to many complex human diseases are likely to have arisen early in human history, but sufficiently recently that they remain embedded in observable common haplotypes. Similarly, selected phenotypes or polymorphic traits of interest to evolutionary biologists and ecologists may be due to nucleotide variants that can be identified by LD mapping.

There is considerable debate over the reasons for the detection of haplotype blocks, with explanations ranging from sampling variance to unequal recombination rates and/or gene conversion hotspots within loci (Wall and Pritchard, 2003; Stumpf and Goldstein, 2003), and study of the population structure of haplotypes are in their infancy. With respect to evolutionary and agricultural genetics, measurement of haplotype structure is increasingly important. Domesticated crops and livestock are likely to have strong haplotype structure as a result of their breeding history (Flint-Garcia *et al*., 2003), whereas outbred and highly polymorphic species such as *Drosophila melanogaster* are almost devoid of haplotypes (*see* Article 10, **Linking DNA to production: the mapping of quantitative trait loci in livestock**, Volume 3). More recent is the advent of population genetics in nonmodel systems that are important with respect to epidemiology, particularly in humans, such as HIV and *Plasmodium* (malaria). The frequency of outcrossing or mixing among these species may contribute to these organisms' ability to evade host immunity (Awadalla, 2003). The ability to dissect quantitative traits to the nucleotide level in any species is ultimately dependent on the thorough characterization of haplotype diversity.

3. Mutation, gene content, and the transcriptome

Population genomics also encompasses several novel aspects of variation that were beyond the technical reach of classical population genetics. For example, direct measurement of mutation rates is now possible, and will complement a large body of literature on the genetic consequences of mutation accumulation (Keightley and Lynch, 2003). For many species, it has been estimated that new genetic variance for fitness or morphological traits is generated at a rate within an order of magnitude of 0.1% of the environmental variance per generation (Clayton and Robertson, 1955; Houle *et al*., 1996). Similarly, genetic evidence suggests that a typical per locus spontaneous mutation rate is approximately 10^{-6} per generation, from which nucleotides are inferred to substitute in each meiosis at a rate close to 10^{-9}. Microsatellites evolve at a much accelerated rate, but with a high variance, as directly measured by comparison of parent and offspring genotypes in several studies (Ellegren, 2000). Insertion–deletion (indel) polymorphism is prevalent,

particularly in studies of regulatory regions of genes, but has been relatively neglected by theoreticians because of the absence of good molecular data on the tempo and mode of indel generation (Li, 1997). Genomic sequence data from invertebrates such as the nematode *Caenorhabditis elegans* that can be propagated essentially clonally (with a population size of 1) will provide measurements of mutation rates independent of the filter of natural selection (Vassilieva *et al*., 2000), offering a crucial comparison with standing variation in natural populations.

Gene order and content is unlikely to be highly polymorphic within populations of multicellular eukaryotes, but has emerged as a challenging feature of microbial genetics. A mixture of processes including conjugation, horizontal transfer from other species, plasmid shuffling, and spontaneous deletion or duplication, result in differences among congeneric bacteria affecting 10% or more of the genome (Ochman and Jones, 2000; Daubin *et al*., 2003; *see also* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4). Whole-genome sequence comparisons have revealed the existence of pathogenicity and virulence islands of genes that distinguish isolates of *Bacillus*, *Escherichia*, and several other bacterial species (Whittam and Bumbaugh, 2002; Hacker and Kaper, 2000), but more generally it has been suggested that each species is defined by a core set of definitive genes that are accompanied by hundreds of variable genes whose presence defines the metabolic capacity of each isolate (Lan and Reeves, 2001). Our conception of microbial diversity is under equally profound challenge through the advent of whole-flora shotgun sequencing, an approach designed to characterize new species that cannot be cultured in vitro (De Long, 2002; Venter *et al*., 2004). As many as 90% of the microbial species in water, soil, and body cavities remain to be described, and genomic arrays will also be developed for use in monitoring diversity in microbial ecosystems.

Finally, the structure of transcriptional variation is emerging as a new field of enquiry (*see* Article 90, **Microarrays: an overview**, Volume 4). Almost no attention has been given to the prevalence of variation for alternative splicing, despite the fact that mutational studies indicate that a considerable fraction of sites affect splicing efficiency in a quantitative manner (Cartegni *et al*., 2002). Transcript abundance itself is also variable among individuals, as a result of both environmental and genetic factors (Yan and Zhou, 2004). Estimates from half a dozen species indicate that at least 10% of the transcriptome differs in abundance between any two individuals, but almost nothing is known of the tissue and temporal specificity of differential transcription (Cheung and Spielman, 2002; Gibson, 2002). Descriptors of the frequencies of qualitatively distinct levels of transcript abundance as well as the cosegregation of these "transcriptional alleles" within and among populations, as well as their heritability, will be a fundamental component of future efforts to describe the genetic architecture of complex traits.

References

Aquadro CF, Bauer-DuMont V and Reed FA (2001) Genome-wide variation in the human and fruitfly: a comparison. *Current Opinion in Genetics and Development*, **11**, 627–634.

Awadalla P (2003) The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, **4**, 50–60.

Bamshad M and Wooding SP (2003) Signatures of natural selection in the human genome. *Nature Reviews Genetics*, **4**, 99–111.
Begun DJ and Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.
Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD and Hartl DL (2002) The cost of inbreeding in arabidopsis. *Nature*, **416**, 531–534.
Cartegni L, Chew SL and Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, **3**, 285–298.
Chakravati A (1999) Population genetics – making sense out of sequence. *Nature Genetics*, **21**, 56–60.
Charlesworth B, Morgan MT and Charlesworth D (1995) The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**, 1619–1632.
Cheung VS and Spielman RS (2002) The genetics of variation in gene expression. *Nature Genetics*, **32**(Suppl), 522–525.
Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, *et al*. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
Clayton G and Robertson A (1955) Mutation and quantitative variation. *American Naturalist*, **89**, 151–158.
Daubin V, Moran NA and Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**, 829–832.
Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M and Goodstein DM (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
De Long EF (2002) Microbial population genomics and ecology. *Current Opinion in Microbiology*, **5**, 520–524.
Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, **16**, 551–558.
Fallin D and Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, **67**, 947–959.
Flint-Garcia SA, Thornsberry JM and Buckler ES IV (2003) Structure of linkage disequilibrium in plants. *Annual Reviews of Plant Biology*, **54**, 357–374.
Frazer KA, Chen X, Hinds DA, Pant PV, Patil N and Cox DR (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Research*, **13**, 341–346.
Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
Gibson G (2002) Microarrays in ecology and evolution: a preview. *Molecular Ecology*, **11**, 17–24.
Glinka S, Ometto L, Mousset S, Stephan W and De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics*, **165**, 1269–1278.
Hacker J and Kaper JB (2000) Pathogenicity islands and the evolution of microbes. *Annual Reviews of Microbiology*, **54**, 641–679.
Hartl DL and Clark AG (1997) *Principles of Population Genetics*, Third Edition, Sinauer Associates: Sunderland, MA.
Hey J (1999) The neutralist, the fly and the selectionist. *Trends in Ecology and Evolution*, **14**, 35–38.
Hey J and Machado C (2003) The study of structured populations - new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535–543.
Hill WG and Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetical Research*, **8**, 269–294.
Houle D, Morikawa B and Lynch M (1996) Comparing mutational heritabilities. *Genetics*, **143**, 1467–1483.

Hudson RR and Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Hudson RR, Kreitman M and Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Jeffreys AJ, Ritchie A and Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human Molecular Genetics*, **9**, 725–733.

Keightley PD and Lynch M (2003) Toward a realistic model of mutations affecting fitness. *Evolution; International Journal of Organic Evolution*, **57**, 683–685.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press: Cambridge.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Reviews of Genomics and Human Genetics*, **1**, 539–559.

Lan R and Reeves PR (2001) When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends in Microbiology*, **9**, 419–424.

Lewontin RC and Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li WH (1997) *Molecular Evolution*, Sinauer Associates: Sunderland, MA.

Luikart G, England PR, Tallmon D, Jordan S and Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler ES and Doebley J IV (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences (USA)*, **99**, 6080–6084.

Meldrum D (2000) Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Research*, **10**, 1288–1303.

Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics and Development*, **12**, 657–663.

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**, 641–647.

Ochman H and Jones B (2000) Evolutionary dynamics of full genome content in Escherichia coli. *The EMBO Journal*, **19**, 6637–6643.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Rockman MV, Hahn MW, Soranzo M, Goldstein DB and Wray GA (2003) Positive selection on a human-specific transcription factor binding site regulating IL4 expression. *Current Biology*, **13**, 2118–2123.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MA (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Schneider S, Roessli D and Excoffier L (2000) *Arlequin: A Software for Population Genetics Data Analysis. Version 2.000*, Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva.

Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stumpf MP and Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Current Biology*, **13**, 1–8.

Stumpf MP and McVean GA (2003) Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, **4**, 959–968.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tishkoff SA and Verrilli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Reviews of Genomics and Human Genetics*, **4**, 293–340.

Vassilieva LL, Hook AM and Lynch M (2000) The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. *Evolution; international journal of organic evolution*, **54**, 1234–1246.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74. published online March 4.

Wall JD and Hudson RR (2001) Coalescent simulations and statistical tests of neutrality. *Molecular Biology and Evolution*, **18**, 1134–1135.

Wall JD and Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**, 587–597.

Weir BS (1996) *Genetic Data Analysis II*, Sinauer Associates: Sunderland.

Weir BS and Hill WG (2002) Estimating F-statistics. *Annual Reviews of Genetics*, **36**, 721–750.

Whittam TS and Bumbaugh AC (2002) Inferences from whole-genome sequences of bacterial pathogens. *Current Opinion in Genetics and Development*, **12**, 719–725.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV and Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377–1419.

Yan H and Zhou W (2004) Allelic variations in gene expression. *Current Opinion in Oncology*, **16**, 39–43.

Zondervan KT and Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nature Reviews Genetics*, **5**, 89–100.

Specialist Review

Modeling human genetic history

Lounès Chikhi

Université Paul Sabatier, Toulouse, France

Mark A. Beaumont

University of Reading, Reading, UK

1. Introduction

The genetic patterns observed today in the human species are the result of a complex history including population expansions and collapses, migration, colonization, extinction, and admixture events. These events have taken place in different locations and times, sometimes involving populations that have been separated for centuries or millennia. Given that anatomically modern humans probably appeared some 100–200 thousand years ago (KYA) in Africa, it would certainly be ingenuous to think that the details of that history will be fully uncovered using genetic data. However, it would probably be equally misguided to refuse the use of simple models on those grounds.

Indeed, genetic data have revolutionized the way we look at our species in many ways. They have shown that the split between human, chimpanzees, and gorillas is rather recent, between 5 and 8 million years old (Nichols, 2001), instead of twice as much as was originally believed on the basis of morphological and fossil data. They have shown that the amount of genetic variation within single populations represents a major proportion (~85–90%) of the diversity present in the human species as a whole, indicating that despite the sometimes large phenotypic differences between some human groups, genetic diversity is not very strongly partitioned. They have contributed to the debate over recent migration patterns, the emergence of our species, or its relationships with Neanderthals. In particular, genetic data have been the major driving force in favor of a recent origin of our species, notably thanks to work done using mitochondrial DNA data in the late 1980s and early 1990s (Cann *et al*., 1987; *see also* Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1).

The modeling work performed in this period showed a significant genetic signal for a demographic bottleneck followed by a demographic expansion, the dating of which was recent (i.e., less than 200 KYA, Slatkin and Hudson, 1991; Rogers and Harpending, 1992), a result found across genetic markers, including independent microsatellite loci (Goldstein *et al*., 1995; Reich and Goldstein, 1998). These

inferences appeared to reject the so-called multiregional (MR) model and to favor a model of recent expansion out of Africa (RAO for recent African origin or OOA for out of Africa), which had the advantage that it correlated with the first appearance of anatomically modern humans. The MR model, which should probably be seen as a family of models (Goldstein and Chikhi, 2002), assumes an ancient origin (more than 1 million years ago, MYA) of human populations. It also posits some regional continuity between present-day populations in the different continents and the early migrations out of Africa of archaic hominids. The RAO model also had the advantage of explaining the limited diversity observed in humans and the lack of strong differentiation between present-day populations, both being higher in chimpanzees despite their much more limited geographical range and census sizes. Thus, in the mid-1990s, the picture could not have been clearer and it seemed to many that only the details needed to be worked out. Fifteen years later, one has to admit that it could not be further from the truth. The increasing use of independent genetic markers, including coding genes, and the rather different modeling assumptions used by various authors have led to a confusing literature from which opposite conclusions have been drawn (Excoffier, 2002). Even assuming that there is a general genetic signature for a recent demographic expansion, it is not necessarily straightforward to conclude in favor of the RAO, as noted by Wall and Przeworsky (2000) and others. The link between genetic results and archaeological, linguistic, or anthropological debates and controversies is not as simple as some would like to believe.

The aim of this article is to describe some of the simple models that have been used to decipher broad trends in that complex history. Principles underlying demographic inference based on genetic data are presented. In particular, we show how summary statistics are differentially influenced by demographic events such as expansions and contractions. We also present some results from the coalescent theory, which focuses on the properties of gene trees and plays a major role in population genetics modeling. We then discuss some recent methodological developments including Bayesian and the so-called approximate Bayesian computational methods (Beaumont *et al*., 2002; Marjoram *et al*., 2003; Beaumont and Rannala, 2004) and try to address a number of the issues that are the focus of ongoing research, including the use of ancient DNA data or the use of patterns of linkage disequilibrium (LD) in the genome (Stumpf and Goldstein, 2003; *see also* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1).

Given the ever-increasing importance attached to genetic data, by nongeneticists and geneticists alike, it should be stated at the outset that the amount of information that can be extracted from genetic data to infer past events is often limited and consequently requires a priori assumptions. There are two fundamental problems: (1) Genetic data only contain information on the relative rates of different processes in the history of the population. This means that it is impossible to make statements about the times of events or population sizes without some external reference, such as mutation rates, or the dating of splits of populations from archaeological evidence. (2) As discussed below, population genetics theory suggests that the vast majority of genes in humans that lived in the past have left no descendent copies

today owing to genetic drift, and consequently there will be no simple relation between the history told by archaeological remains and that told by the genes.

2. Principles of demographic inference from genetic data: information extracted from summary statistics and gene tree shapes

In this section, we describe how the demographic history influences genetic patterns by first describing its effects on some summary statistics and then on the shape of gene trees. This allows us to review some results of the coalescent theory and then address some issues regarding parameter inference.

2.1. Summary statistics are differentially affected by demographic events

The detection and quantification of past demographic events relies on the fact that specific events leave a genetic *signature* in present-day populations. Meaningful inference will therefore be possible only if these signatures are sensitive to specific demographic scenarios, and unlikely to be affected by others. A richly explored area of analysis has been the study of the effect of changes in population size on different summary statistics. For instance, a stationary (or stable) population can be characterized by the scaled mutation rate $\theta = 4N_e\mu$ (for diploids) where N_e is the effective population size and μ the locus mutation rate. There are different estimators of θ, on the assumption of a stationary population, which are based on different summary statistics calculated from the data (see below). If the population is growing or contracting, the expected values of these estimators diverge from each other, and this can be used to detect population growth or decline.

A commonly used statistic that summarizes genetic diversity is the heterozygosity expected under Hardy–Weinberg conditions $H_e = 1 - \sum(n_i/n)^2$ where n_i is the observed allelic count of the ith allele. Under an infinite allele model, Kimura and Crow (1964) showed that in a stable population $E[H_e] = \theta/(\theta + 1)$, which can be used to obtain an estimate of θ. Another way to estimate θ comes from the work of Ewens (1972) who studied the sampling properties of neutral alleles in a demographically stable population for the infinite allele model. Ewens showed that at mutation-drift equilibrium, the expected number of distinct alleles, n_A, in a sample of size n is a function of n and θ only:

$$\begin{aligned} n_A &= \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+2n-1} \\ &= \sum \frac{\theta}{\theta+i} \quad (\text{for } i = 0 \ldots 2n-1) \end{aligned} \tag{1}$$

Under these null-model conditions, he also characterized the expected allelic frequency distribution, showing that it is a function of θ, n, and n_A. This distribution obtained from what is now known as Ewens' sampling formula, provided the

means for testing departures from the null-model (Watterson, 1978). Indeed, H_e values can be computed using n_i values or using Ewens' sampling formula (i.e., ignoring allelic counts). In fact, Watterson originally used the homozygosity that is equal to $1 - H_e$, but H_e is more commonly used now. Significant differences between the two H_e values can then be interpreted in terms of departures from any of the neutrality, size constancy, or mutation model assumptions. When large populations go through a bottleneck, rare alleles are lost first, thereby reducing n_A significantly. Since H_e computed as $1 - \sum(n_i/n)^2$ is little influenced by rare alleles (their frequency is squared), high H_e values will be maintained for relatively longer periods. Thus, observed H_e values will be significantly higher than those expected conditional on n_A and equilibrium conditions. Conversely, expanding populations will tend to accumulate new (and hence rare) alleles, and significantly lower values of H_e will be observed, so long as a new equilibrium is not reached. Simple summary statistics such as n_A and H_e thus contain information on ancient demographic events. So long as simple demographic histories and a simple mutation model can be assumed, a straightforward interpretation is possible.

Since the early work of Ewens and Watterson, the approach has been extended to apply to different mutation models and hence to different genetic markers (Tajima, 1989a,b; Fu and Li, 1993) and different summaries of the data. For instance, Marth *et al*. (2004) studied the properties of the full allele frequency spectrum under different demographic scenarios, and found computationally quick ways to estimate these frequency spectrums. Similar approaches based on other aspects of the allelic distribution have been used. For microsatellites, whose alleles are defined by the size of an amplified DNA fragment consisting of small repeated units, Reich and Goldstein (1998) suggested to use an index of peakedness mathematically related to the distribution's kurtosis. The rationale was that stable populations tended to have ragged allelic distributions, whereas expanding populations had smoother, often unimodal distributions. With a similar rationale, Garza and Williamson (2001) showed that bottlenecked populations tended to have gappy allelic frequency distributions, whereas the allelic range was little affected.

For DNA sequence data, H_e and n_A are not the most appropriate measures of diversity because they do not account for the amount of mutations between alleles. The mean number of nucleotide differences between sequences, π, and S, the number of segregating sites (i.e., single-nucleotide polymorphisms or SNPs) are more suitable. Both statistics have been shown to provide two estimators, θ_S and θ_π, of the scaled mutation parameter, θ, at mutation-drift equilibrium and to be differentially affected by demographic events and selection (Tajima, 1989a,b). Tajima (1989a) hence suggested the use of $D = (\theta_\pi - \theta_S)/[\mathrm{Var}(\theta_\pi - \theta_S)]^{1/2}$, as a measure of departure from equilibrium conditions. Demographic bottlenecks and balancing selection tend to reduce S without affecting π very much and hence to produce positive D values, whereas population expansions and positive selection (selective sweeps) tend to produce negative values. As we shall see below, the situation is actually more complicated (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1 and Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1).

2.2. The shape of gene trees

Where an ordinal genetic distance can be defined between alleles, such as DNA sequences, and microsatellite alleles, it is possible to summarize or represent genetic diversity, not by a number, like π or H_e, but by a distribution of allelic pairwise distances. These pairwise difference or mismatch distributions have been extensively used and are also sensitive to demographic events (Slatkin and Hudson, 1991; Rogers and Harpending, 1992). Since, the interpretation of such distributions is easier from a gene tree perspective, we first review some results of the coalescent theory, which is the backbone of population genetic modeling. The coalescent theory provides important results on the probability distribution of genealogies (and hence on their shape) arising as a limiting case under a class of population genetics models such as the Wright–Fisher and Moran models. One major result of the standard coalescent is that in a sample of size n, the time T_n during which there are n lineages (i.e., until the first two lineages coalesce) follows an exponential distribution of parameter $\lambda_n = n(n-1)/4N_e$, where N_e is the effective population size. Its expectation is $E[T_n] = 1/\lambda_n = 4N_e/n(n-1)$.

Importantly, coalescent times are functions of N_e and n, and are independent of the mutational process, so long as neutrality can be assumed. After the first coalescent event, the number of lineages decreases to $n-1$ and the following coalescent time is sampled from the *new* exponential distribution with parameter λ_{n-1} and expectation $E[T_{n-1}] = 4N_e/(n-1)(n-2)$. This process continues until the most recent common ancestor (MRCA) of the sample is reached. The time to the MRCA represents the height of the gene tree and has an expectation of $E[T_{\text{MRCA}}] = E[T_n] + E[T_{n-1}] + \cdots + E[T_2] = 4N_e(1-1/n)$. Thus, for reasonably large sample sizes, $E[T_{\text{MRCA}}] \sim 4N_e$ at equilibrium, and the T_{MRCA} of the *sample* is nearly equal to the T_{MRCA} of the *population*. The structure of this null-model tree is interesting as most of the coalescent events take place near the branches tips. For $n = 50$, coalescence times will vary, in expectation, by more than 3 orders of magnitude, with the first 25 and 35 coalescent events taking place in the first 2 and 5% of the tree's height, respectively. In contrast, the last coalescent event has an expectation of $2N_e$, that is, half the tree's height. The tree is thus typically dominated by the last two branches in an equilibrium population.

For populations with monotonic size changes, it is straightforward to infer expected tree shapes (Figure 1). An expanding population can be seen as having its N_e decreasing backward in time. Thus, expected coalescence times will also decrease with N_e. Hence, compared to a stable population of similar size today, coalescent events will concentrate around the MRCA, producing a "starlike" tree. By contrast, in a declining population, coalescence times are moved toward the present and the tree has a "comblike" shape. Assuming a high enough mutation rate, a starlike tree is expected to produce alleles that have evolved more or less independently since the MRCA and mutations will accumulate along branches of similar lengths. Hence, going back to mismatch distributions, we expect them to be unimodal and roughly Poisson-like (Slatkin and Hudson, 1991). By contrast, stable populations are expected to produce ragged mismatch distributions with at least two modes (since the tree is dominated by the last branches).

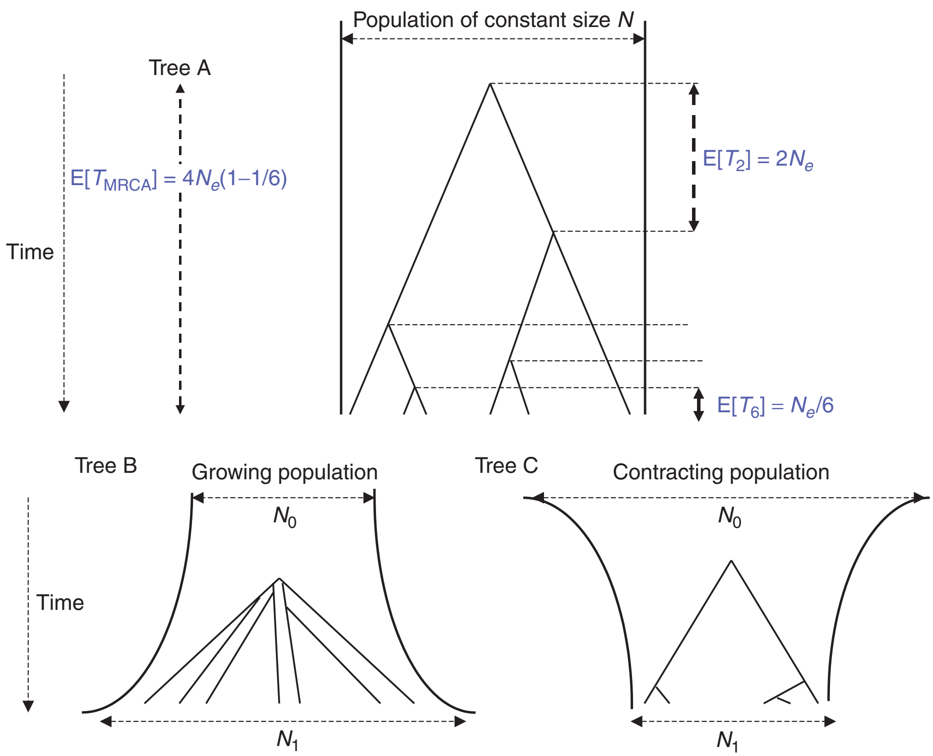


Figure 1 Coalescent trees. Tree A is the tree expected under a simple Wright–Fisher model (population size constant, no selection). Tree B is expected under a population growth. Tree C is expected under a model of population contraction

It may be worth noting that we presented the properties of the true gene trees, which are usually unknown. In practice, trees have to be estimated using polymorphism data, which are dependent on the mutation process and the time since the MRCA. When the total number of mutations is low, we may not be able to separate a starlike from a comblike tree. Similarly, when the number of mutations is high in a starlike phylogeny and if homoplasy is important, as in microsatellite loci, similar alleles may be generated by long branches that will reduce the expansion signal by creating false short branches.

Despite being a limiting case, the coalescent is extremely popular among population geneticists because it is fairly robust to details of life history (Donnelly and Tavaré, 1995), and coalescent simulations are also extremely fast (samples, rather than populations are simulated). The standard coalescent has also been extended to account for population structure, recombination within the locus of interest, and different forms of selection, a thorough review of which is given by Nordborg (2001).

In the next section, we discuss some of the ongoing research on population genetic modeling and some of problems that still need to be addressed in coalescent-based modeling.

3. Improving the use of genetic information: increasingly complex models for increasingly complex data

3.1. Bayesian and approximate Bayesian methods: the use of rejection, importance sampling, and Markov chain Monte Carlo algorithms

Following Felsenstein's (1992) remark that summary statistics discard most of the information present in genetic data, recent genetic modeling has seen the development of a number of statistical approaches that try to extract as much information as possible from the full allelic distributions.

Likelihood-based approaches aim at computing the probability $P_{\mathrm{M}}(D|\theta)$ of generating the observed data D under some demographical model M, defined by a set of parameters $\theta = (\theta_1, \ldots, \theta_k)$. This probability, which can be seen as a function of θ (since M is assumed and D is given) is the likelihood $L_{\mathrm{M}}(\theta|D)$, or simply $L(\theta)$. Some methods try to find the θ_i values that maximize $L(\theta)$, and use these maximum likelihood estimates (MLE) for the parameters of interest. Other likelihood-based approaches take a Bayesian perspective and try to estimate probability density functions for these parameters (rather than point estimates). Using Bayes formula, this can be written as:

$$P_{\mathrm{M}}(\theta|D) = P_{\mathrm{M}}(\theta) * \frac{P_{\mathrm{M}}(D|\theta)}{P_{\mathrm{M}}(D)} = P_{\mathrm{M}}(\theta) * \frac{L_{\mathrm{M}}(\theta|D)}{P_{\mathrm{M}}(D)} \tag{2}$$

Since the denominator $P_M(D)$ is constant, *given* the data, $P_M(\theta|D)$ is proportional to $P_M(\theta) * L_M(\theta|D)$. In the Bayesian framework, $P_M(\theta)$ summarizes knowledge (or lack thereof) regarding the θ_i *before* the data are observed and is referred to as the prior. $P_{\mathrm{M}}(\theta|D)$ is the posterior and represents new knowledge about the θ_i *after* the data have been observed. The posterior is thus obtained by weighting the prior with the likelihood function. While the use of priors involves some subjectivity, it has the advantage of making clear and explicit assumptions about parameters, rather than assuming point values for, say, mutation rates or generation times, as has often been the case in the past, for instance, to date population split events (Goldstein *et al*., 1995; Chikhi *et al*., 1998; Zhivotovsky, 2001).

The two previous approaches require the likelihood to be computed. It is theoretically possible to use coalescent simulations to generate samples under complex demographic models, and hence estimate $P_{\mathrm{M}}(D|\theta)$ for many parameter values using classical Monte Carlo (MC) integration. In practice, however, these probabilities are so small that they cannot be estimated in reasonable amounts of time for most sample sizes (Beaumont, 1999). Efficiency can be improved by using importance sampling (IS), which aims to sample gene trees as closely as possible from their conditional distribution, given the data, thereby avoiding most of the genealogies that would be sampled in classical MC simulations (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000). Another computer-intensive approach that has had some success in making inferences about demographic history is Markov chain Monte Carlo (MCMC), in which a Markov chain is simulated whose stationary distribution is the required Bayesian posterior distribution. Example

applications to human population data include estimation of the amount and direction of gene flow in equilibrium models of migration-drift balance (Beerli and Felsenstein, 2001), the relative contributions of source populations in an admixture model (Chikhi *et al*., 2001), and the inference of ancestral size and splitting of human and chimpanzee populations (Rannala and Yang, 2003). A particularly exciting application of MCMC is in a model of migration with population splitting (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004), which has recently been used to infer aspects of the recent demographic history of chimpanzees (Won and Hey, 2005), and undoubtedly will have useful applications in humans.

Most likelihood-based methods have not yet had the impact one would have expected. This is due to the fact that for most interesting demographic models the likelihood cannot be computed or approximated. Also, the computations are extremely time-consuming and hence cannot be applied to the increasing size of modern data sets. An exception here is computer-intensive methods for detecting admixture and cryptic population structure from multilocus genotypic data (Dawson and Belkhir, 2001; Pritchard *et al*., 2000; Falush *et al*., 2003), but these are based on relatively simple approximations that do not include genealogical structure. However, they do point to one possible avenue in population genetics, the approximation of complex genealogical models, the parameters of which can then be inferred considerably more easily. Examples in this vein are the development of PAC likelihood (for product of approximate conditionals) to simplify inference with recombining sequence data (Li and Stephens, 2003), in which the coalescent is approximated by a simpler genealogy with more tractable analytical properties; the use of composite likelihood methods (Wang, 2003; McVean *et al*., 2004), where the full likelihood is reduced to simpler components that are treated as independent and multiplied together.

An alternative to the likelihood-based methods that use all the data is the development of approximate likelihood methods that measure summary statistics from the data, and then use simulations to find parameter values that generate data sets with summary statistics that match those in the data most closely. Data are simulated according to a model or set of models for a wide range of parameter values, either using a regular grid (e.g., Weiss and von Haeseler, 1998) or sampling from prior distributions (e.g., Pritchard *et al*., 1999). Parameter values that produce simulated data sufficiently close to the observed data are used to estimate the likelihood (Weiss and von Haeseler, 1998 on mtDNA data) or construct posterior distributions (Pritchard *et al*., 1999, on linked Y chromosome microsatellite data). For instance, Weiss and von Haeseler (1998) used the number of segregating sites S and π to measure similarity between data simulated under a model of population size change (exponential increase or decrease versus stability) and observed data. They accepted the ith simulated data (and corresponding parameter values) using the indicator function:

$$I_\delta(i) = 1 \text{ if } (|\pi_i - \pi| < \delta \text{ and } S_i = S) \text{ and } I_\delta(i) = 0 \text{ otherwise} \tag{3}$$

where δ is some arbitrary tolerance level or threshold (see below).

By varying the exponential rate, the time since the population size started to change and , they constructed a grid of likelihood values. They then computed the

ratio to the highest value to define confidence intervals. Applying this approach to a Basque sample, they found virtually no support for a model of constant or decreasing population size. The best support was for a recent exponential increase from an originally small population, which is in agreement with the relative isolation of Basques from other European populations.

Pritchard *et al*. (1999) also analyzed models of exponential size change, but used three summary statistics, the number of different haplotypes, the mean (across loci) of the variance in repeat numbers, and the mean heterozygosity across loci. They used an indicator function I_δ such that for the ith simulation $I_\delta(i) = 1$ if the three relative differences $|P_{SIM} - P_{OBS}|/P_{OBS}$ are less than the tolerance δ, where P stands for any of the three parameters, and $I_\delta(i) = 0$ otherwise. They applied this method to data from different continents and found strong support for a model of population growth for the whole world population (despite genetic differentiation between continents), and for East and South Africa, America, Europe, East, and West Asia but not for West Africa or Oceania. They also found that there was a huge uncertainty on t_0, the time at which populations started to increase, and on T_{MRCA} values, confirming as well that t_0 and T_{MRCA} are generally very different. Mean T_{MRCA} values around 40 KYA had confidence intervals ranging between ~10 and more than 100 KYA. Another important result for the interpretation of results obtained under different genetic models was that constraints on allelic sizes could lead to significant changes in T_{MRCA} values (but not so much on t_0). For instance, the higher end of the T_{MRCA} confidence interval could move from ~120 to ~320 KYA when constraints were accounted for.

Bayesian versions of these summary-statistic methods (such as Pritchard *et al*., 1999) have recently been dubbed "approximate Bayesian computation (ABC)" (Beaumont *et al*., 2002; Marjoram *et al*., 2003), and made more efficient. For example, Beaumont *et al*. (2002) have suggested improvements by applying a rejection algorithm as above, weighting the accepted parameter values according to their distance to the observed data, and correcting for the relationship between the parameter value and summary statistics in the vicinity of those calculated from the observed data. Marjoram *et al*. (2003) have proposed an MCMC algorithm where the acceptance or rejection of an update depends on whether the data or summary statistics thereof that arise from the update are within some distance of those observed in the data.

ABC methods are extremely flexible since they can be applied to any demographic model under which data can be simulated. It is also easier to add nongenetic information into the priors, which could be particularly relevant for integrating archaeological data into genetic modeling. Limitations stem from the difficulty to decide which summary statistics should be used and how the value of the tolerance δ influences inference. Little has been done on these issues, but Beaumont *et al*. (2002) found that the improved ABC algorithm was little affected by δ values whereas the simple rejection approach was strongly influenced.

Summary statistics do not need to be used in a likelihood framework. An example here is the study by Pluzhnikov *et al*. (2002) in which the distribution of summary statistics for different parameter values is estimated by simulations, and then compared with summary statistics in the data. Using this approach on multilocus sequence data, they found that European and Chinese populations did not

fit with a stable population or exponentially growing population model. Recent work by Akey *et al*. (2004) studied the behavior of four summary statistics, Tajima's D and Fu and Li's D^* and F^*, and Fay and Wu's H, under the standard neutral model and four additional demographical scenarios (an exponential expansion, a bottleneck, a two-island model and an ancient population split). They computed a score function on the basis of the absolute difference in average summary statistics from simulations under specific demographic scenarios with those in the data, and then chose parameter values that minimized this score. This was carried out for 132 genes sequenced in African-American and European-American samples and they were able to (1) estimate parameter values under different scenarios, (2) select reasonable scenarios for the different data sets, and (3) detect loci potentially under selection (outliers). For instance, they found that the European-American data set was most consistent with a bottleneck occurring 40 KYA, whereas the African-American sample was most consistent with either an expansion or an old and strong bottleneck. An interesting result is that out of the 22 loci potentially under selection under a standard model, only 8 were found to be outliers in the other four scenarios, and were termed "demographically robust selection genes". In other words, the demographical history of populations can potentially account for many outlier genes, which could be interpreted as being under some form of selection based on departures from the standard neutral model. This study also showed that most of the deviations from neutrality and all the "demographically robust selection genes" were not shared by the European- and African-Americans and was interpreted as reflecting potential local adaptations. A similar approach based on mismatch distributions and a set of demographic models was followed by Marth *et al*. (2003).

As indicated by the study of Akey *et al*. (2004), summary-statistic approaches can be useful for detecting selection on individual loci or genomic regions when a large number of markers are analyzed. A recently productive area of analysis in this regard has been based on estimators of the parameter F_{ST}. This parameter, which is used to measure genetic differentiation between populations, can be interpreted as the probability that two genes chosen at random within a subpopulation share a common ancestor within that subpopulation. It is expected that neutral loci should follow some unknown distribution, and that outlier loci, potentially under selection, can then be detected, as originally suggested by Cavalli-Sforza (1966). It is possible to construct a null distribution of F_{ST} under simple models, typically island models, and see where the real data lie (Beaumont and Nichols, 1996). Another solution is to use large numbers of loci sampled randomly in the genome to account for the unknown demographic history (but common to all loci) as suggested by Goldstein and Chikhi (2002). The latter approach has been used for 26 000 SNPs in African, European, and Asians samples by Akey *et al*. (2002) who detected a number of outliers. It has also been used with microsatellite data on European and African samples (Storz *et al*., 2004), which has, remarkably, yielded similar patterns to that found by Akey *et al*. (2004). Indeed, the outlier loci all have substantially reduced variability in non-African populations, suggesting adaptive selective sweeps following the putative OOA expansion. Of course, these inferences are dependent on the demographic model assumed, and it may be possible, for

example, that during a wave of advance some loci may be carried all the way with total replacement while others may be lost and replaced by indigenous markers, in which case a complex demography could mimic selection (Eswaran *et al*. 2005).

3.2. Increasing the complexity of models and the size of the data analyzed

In recent years, there has been increasing recognition and awareness that most of the models used to infer ancient demography were too simple. In this regard, an original and interesting modeling framework was recently developed by the group of L. Excoffier (Currat *et al*., 2004; Ray *et al*., 2003), which allows explicit spatial modeling of genetic data in which ecological information can be incorporated. First, complex demographical scenarios can be simulated using digitalized geographical maps made of cells arranged in a two-dimensional grid with defined carrying capacities and friction values (expressing the difficulty in moving through them). Samples are chosen on the digitalized map and coalescent simulations can then be carried out to simulate the genetic data of the populations of interest. Theoretical expectations of a number of summary statistics can be obtained under widely different scenarios and compared to observed data. This framework was implemented in the SPLATCHE software (Currat *et al*., 2004) and it was used in its simplest form (i.e., assuming no environmental heterogeneity) to study the effect of geographical expansions and varying levels of gene flow on the form of mismatch distributions (Ray *et al*., 2003). Currat and Excoffier (2004) also simulated the most realistic model to date, of the demographic expansion of humans in Europe and their interaction with Neanderthals. In this model, a range expansion of modern humans starts in the Near East with local logistic growth and a higher carrying capacity for humans over Neanderthals (to model a better exploitation of the environment by humans). Conditional on the fact that no Neanderthal mitochondrial DNA data are observed in humans today, coalescent simulations are carried out for different levels of admixture. Thus, using a summary-statistic approach akin to ABC methods, Currat and Excoffier (2004) demonstrate that the maximum level of admixture between Neanderthals and humans was most probably much lower than 0.1% (for the female line). This value in orders of magnitude lower than the lowest bound found using coalescent simulations that did not incorporate population structure. Thus, this study shows just how incorporating spatial structure into modeling can provide extremely different conclusions on admixture estimates.

Similarly, one of the most disputed issues in human genetics has been the detection of ancient bottlenecks and patterns of population growth. Wall and Przeworsky (2000) tried to put some order in the rather confusing literature, and concluded that the signal for a population bottleneck followed by an expansion detected in the early 1990s was mostly limited to mtDNA and Y chromosome data, and hence probably the result of selection on these loci. A more recent review by Excoffier (2002) concluded that most of the data disagreeing with a population

expansion were from coding regions and could, therefore, be explained by balancing selection, as was suggested for the first time by Harpending and Rogers (2000). Thus, the signal for population expansion would actually be real. However, the balancing selection hypothesis predicts that regions of low recombination should exhibit high levels of genetic diversity, whereas Payseur and Nachman (2002) found exactly the opposite, namely, a positive relationship between local recombination rate and variability. Recently, this issue had a new development when Hellman *et al*. (2003) found that this positive correlation could actually be caused by a correlation between recombination and mutation rates. Thus, it appears that the complexity of patterns of genetic variation in the genome can hinder our ability to uncover patterns of ancient population growth (*see* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1).

Even assuming that one concentrates on neutral markers, the literature on population growth has been rather confusing. Table 1 is an attempt at summarizing some of the modeling assumptions made regarding past population size changes. The point we would like to make with this table is that it is very difficult to compare quantitatively and even qualitatively the results obtained by studies when the demographical model assumed (exponential vs. sudden growth, with or without previous bottleneck) the markers analyzed (maternally inherited mtDNA vs. nuclear coding genes vs. paternally inherited Y chromosome haplotypes) and the statistical approaches differ that much.

Another level of complexity arises when one studies the temporal behavior of the summary statistics used to detect population size changes. For instance, Tajima's D, which is probably the most popular of all statistics used to detect population expansions and bottlenecks, may be transiently positive and negative depending on the severity of the population size change and the number of generations since the original demographic event. After a bottleneck, D values will first be positive, then become negative when the population grows again, before tending towards equilibrium (Tajima, 1989b; Fay and Wu, 1999). Moreover, D is affected in a complex manner by increasing complexity of the mutation model. Simulations have shown that mutation rate heterogeneity within a locus can lower S, which could either lead to apparent signals of population bottlenecks, in an otherwise stationary population, or hide real population expansion signals by pushing D values in the opposite direction (Bertorelle and Slatkin, 1995; Aris-Brosou and Excoffier, 1996). More recently, hidden population structure was also shown to generate spurious results. Indeed, Ptak and Przeworski (2002) have shown that the number of "ethnicities" in the samples analyzed was negatively correlated to D. This can be understood by noticing that more diverse samples tend to contain a higher proportion of rare alleles, when populations are differentiated. Given that different loci can have different $N_e s$, the temporal behavior of D could lead to apparently contradictory results depending on whether mtDNA or nuclear microsatellites are analyzed, as was indeed found (Harpending and Rogers, 2000). A similarly complex temporal behavior was found by Reich and Goldstein (1998) for their peakedness statistic or by Kimmel *et al*. (1998) for their "imbalance index", making interpretations of these summary statistics more difficult than one would like.

Studies such as that of Akey *et al*. (2004) or Currat and Excoffier (2004) point to the importance of modeling population structure. However, an issue that has

Table 1 Examples of mutation and demographic models used to infer ancient human growth patterns

	Demographical model				Initial conditions		Mutation model	Approach	Marker
	Stepwise	Exponential	Logistic	Linear	Equil	Other	ISM		
Tajima (1989a)		X			X		ISM	Tajima's *D*	DNA
Slatkin and Hudson (1991)		X			X		ISM	Mismatch distribution	DNA
Rogers and Harpending (1992)	X				X		ISM	Mismatch distribution	DNA
Kimmel *et al*. (1998)	X	X	X		X	Mono[a] + Bottle[b]	aSSM	Imbalance index	Microsatellites
Reich and Goldstein (1998)	X				X		SSM	Summary statistics	Microsatellites
Wilson and Balding (1998)		X			X		SMM	Full likelihood	Linked microsatellites
Pritchard *et al*. (1999)		X			X		SSM + cSSM	Rejection algorithm	Linked microsatellites
Beaumont (2004)		X		X	X		SSM	Full likelihood	Microsatellites
Storz and Beaumont (2002)		X		X	X		SSM	Full likelihood	Microsatellites
Marth *et al*. (2003)	X				X	Bottle[b]		Mismatch distributions	SNPs
Currat and Excoffier (2004)			X		X	Spatial[c]	ISM	Summary statistics	DNA

ISM: infinite site model, aSSM: asymmetric single step model, cSSM: constrained single step model, SSM: single step model.
[a]Monomorphic initial conditions.
[b]Bottlenecked population.
[c]Spatial means that a complex spatial model is used (see corresponding references for details).

received very little attention is the influence of population extinctions despite the fact that early human groups were probably small enough to be subject to local extinctions. As noted by Eller (2002) and others (Sherry *et al*., 1998; Goldstein and Chikhi, 2002), the conservation genetics literature has developed approaches and models that could benefit human population genetics modeling. Indeed, an argument often used against the MR and in favor of the RAO model is that human diversity is low and corresponds to an effective size of approximately 10 000. Assuming a subdivided population as in the MR model would lead to census numbers too low to be compatible with any version of the MR model, which would require humans to be present across the Old World. However, as Eller (2002) showed, the argument would not necessarily hold if one accounted for extinction rates. With an extinction rate of 10%, an N_e of 10 000 could be compatible with census sizes greater than 300 000 under a wide range of conditions. This census size could even be larger if one assumed some level of intragroup inbreeding, as is the case in current human groups. Thus, accounting for extinctions would make the MR model compatible with small N_e values. Even though Eller used a simple demographic model (the infinite island model of Wright which allows every population to be in genetic continuity with every other one), this clearly shows that realistic models including extinctions, inbreeding, and differential migration rates could well add unexpected arguments to an already long debate.

The patterns and temporal behavior of genomic diversity is also the focus of ongoing and important research. Indeed, it has long been recognized that demographic events can influence patterns of linkage disequilibrium (LD, the statistical association between alleles from different loci), and hence that strong LD between physically unlinked markers could be the signature of admixture events or bottlenecks (Ardlie *et al*., 2002; Stumpf and Goldstein, 2003; Chikhi and Bruford, 2005). In the last 20 years, it has become increasingly clear that recombination rates vary enormously in the human genome with recombination hotspots separated by regions of low recombination rates (Nachman, 2003; McVean *et al*., 2004). It has recently been suggested that this structure is not simply the stochastic result in a model of more or less uniform recombination rate, but rather, corresponds to what has been called the blocklike structure of the human genome (e.g., Jeffreys *et al*., 2001; Ardlie *et al*., 2002; Stumpf and Goldstein, 2003). There is currently ongoing debate as to whether blocks exist or not (e.g., Anderson and Slatkin, 2004). But it is clear that the patterns of LD observed in many human data sets could not be explained in a model of exponential expansion, which predicted that LD should not extend much over 3–5 kb around a particular marker (Kruglyak, 1999). Moreover, recent simulations have shown that the expected pattern of LD blocks is stochastic and that even when simulations explicitly assume an extreme variation in recombination rates, some regions may not exhibit a block-like structure depending on the demographic history of the population of interest (Stumpf and Goldstein, 2003). This led these authors to suggest that populations (and regions of the genome) could go through different stages from preblock through block to postblock structure, in a complex manner in which the variance in recombination rates and the demographic history interact. Since different genomic regions have been analyzed in different populations with varied demographic histories, this issue is likely to be discussed for some years.

4. Conclusion and perspectives

We have tried to present above the general principles underlying the use of genetic data to detect past demographical events and estimate parameters of interest. We have seen that there has been a clear tendency toward incorporating more complexity in demographical and mutational models. Larger data sets including many independent loci are also becoming available to the community. For example, Marth *et al*. (2003) used data from as many as 500 000 SNPs and concluded for a significant bottleneck signal in humans. These data represent new challenges for population geneticists. Indeed, the potential risk in using SNPs is the ascertainment bias associated to their nonrandom discovery and typing. SNPs are usually detected in small samples and tend, therefore, to present balanced allele frequencies, which is not representative of neutral variation (Wakeley *et al*., 2001; Marth *et al*., 2003). Recent studies have tried to account for this sampling bias effect (e.g., Nielsen *et al*., 2004; Wakeley *et al*., 2001; Marth *et al*., 2003) but more research needs to be done. Similarly, the sampling of "populations" has to be improved. For instance, in the Akey *et al*. (2004) study, it is not clear whether differences between African-Americans and European-Americans actually reflect adaptations. Indeed, simulations show that admixture is expected to increase significance in the tests realized. Akey *et al*. concluded that their result was contrary to the expectation since African-Americans are admixed. As it happens, present-day Europeans are the result of a significant admixture event during the Neolithic transition (e.g., Chikhi *et al*., 2002), whereas the admixture level observed in African-Americans is very recent and much more limited. Thus, it could be that differences between the two groups are actually the result of their differing admixture histories.

Another avenue of research currently underexplored is the development of statistical approaches to quantitatively compared models (Burnham and Anderson, 2002). Likelihood theory has been used to compare nested models (e.g., Marth *et al*., 2003) but still much work needs to be done. Other avenues include the use of genomic patterns of diversity (such as variable recombination rates and block-structure patterns). Data from ancient DNA (assuming that contamination can be avoided) will also need to be incorporated in models that allow for genetic data to be sampled at different times. Ongoing research indicates that it is already possible to do so (e.g., Drummond *et al*., 2002; Shapiro *et al*., 2004), but future models will have to incorporate population structure since it would be extremely misleading to assume (rather than infer) a genealogical continuity between present-day people and archaeological samples from the same geographical area. A number of graphical methods such as those introduced by Nee *et al*. (1995) and further developed by Strimmer and Pybus (2001) and others may represent interesting avenues as a first exploratory step. These authors noted that it is possible to convert rates of coalescent events inferred from reconstructed gene phylogenies into N_e values and display them through time together with the corresponding gene tree (Nee *et al*., 1995). These "skyline plots", representing variations of N_e through time since the sample's MRCA, appear to generate different plots under different types of population size changes (linear versus exponential versus logistic). Thus, they could be used to construct a limited set of alternative models before some of the methods described above are applied to the data. A major limitation is that they require the genealogy

to be inferred with little error, which is difficult as we have seen. Another similar approach was used by Polanski *et al*. (1998) who tried to infer the demographic trajectory of humans (assuming monotonic growth) using mtDNA.

At this stage, we would like to make some cautionary remarks on a number of tree- or network-based methods, which have had great success among archaeologists and molecular biologists, but not so much among population geneticists. Indeed, there appears to be increasing use of such methods to devise complex scenarios, with migrations in multiple directions, back migrations, expansions, and long-distance migrations. Many of these complex scenarios are often constructed using sets of linked markers such as mtDNA or Y chromosome haplotypes. From an evolutionary point of view, such markers behave as single loci (with a number of "independent" alleles). The evidence from approximate and full likelihood analyses, which should theoretically extract close to as much information as possible from the data, suggests that they are providing limited amounts of information, even when few demographic parameters are estimated (e.g., Weiss and von Haeseler, 1998; Pritchard *et al*., 1999; Chikhi *et al*., 2001). The problem is that many demographic histories can give rise to the same genealogy, and the same demographic history is compatible with many different genealogies. Simulations have shown that population events and tree nodes are very loosely related (Nichols, 2001). For instance, we noted above with the Pritchard *et al*. (1999) study that the time at which a population expanded had little to do with the T_{MRCA}. In human studies, the already limited information is further reduced by authors who arbitrarily concentrate on subsets (haplogroups). The geographic distribution of haplogroups is then used to describe a possible pattern of past migrations of human populations (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1 and Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1). It is not clear what assumptions such methods make and under which scenarios they might "work". It is possible that they could work if one posited that each movement of populations was involving an initial bottleneck followed by an expansion of population size, so that the genealogy would be constrained to follow the demography, and the pattern not blurred by subsequent drifts or admixture. Although it is feasible that this is an accurate description of human history in some regions, it has thus far not been demonstrated. Genetic data are potentially very useful but they may be much more limited than one would like. When single locus data are used, one should always account for the possibility of a huge variance of the coalescent process, particularly when the underlying demography follows the assumptions of the standard coalescent. In this case, for instance, the variance of the T_{MRCA} is approximately $2.3 * N_e$ for reasonable sample sizes. This means that independent loci, representing independent draws from the same demographic history, are not expected to have similar gene trees or T_{MRCA} values. However, the stable population case represents an extreme, and other demographic histories, for example, expansion, will yield more similar genealogies among loci. Network-based methods are appealing because they are visual and easy to use, but it should be clear that it has not yet been demonstrated that they could provide statistically sound inferences (e.g., Knowles and Maddison, 2002). Their popularity is largely the result of a lack of credible alternative methods for complex modeling of genetic data. However, as

demonstrated in this article, the picture is now changing, and we can look forward to an exciting and challenging era based on parametric modeling of genetic data.

We hope that we have convinced the reader that it is a very exciting period to live in. A period in which the complexity of patterns of genomic diversity and of demographic trajectories of human groups is still to be discovered and explored. The field of human population genetic modeling provides, in many ways, richer research avenues than we could have imagined 20 or only 10 years ago.

Further reading

Barbujani G, Magagni A, Minch E and Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 4516–4519.

Beaumont MA (2004) Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity*, **92**, 365–379.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High resolution of human evolutionary history trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Fu YX and Li WH (1999) Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theoretical Population Biology*, **56**, 1–10.

Hey J and Machado CA (2004) The study of structure populations – new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535–543.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT and Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *American Journal of Human Genetics*, **66**, 979–988.

Kingman JFC (1982a) On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43.

Kingman JFC (1982b) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Kuhner M, Yamoto J and Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, **140**, 1421–1430.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Przeworsky M, Hudson RR and Di Rienzo A (2000) Adjusting the focus on human variation. *Trends in Genetics*, **16**, 296–302.

Tavaré S, Balding DJ, Griffiths RC and Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Templeton AR (2002) Out of Africa again and again. *Nature*, **416**, 45–51.

Wall JD and Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. *American Journal of Human Genetics*, **73**, 502–515.

Wilson IJ, Weale ME and Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society A*, **166**, 155–188.

References

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA and Krugkyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, **2**(10), 1591–1597.

Akey JM, Zhang G, Zhang K, Jin L and Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Anderson E and Slatkin M (2004) Population genetic basis of haplotype blocks in the 5q31 region. *American Journal of Human Genetics*, **74**, 40–49.

Ardlie KG, Kruglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews*, **3**, 299–309.

Aris-Brosou S and Excoffier L (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, **13**, 494–504.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont MA and Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251–261.

Beaumont MA, Zhang W and Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beerli P and Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4563–4568.

Bertorelle G and Slatkin M (1995) The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Molecular Biology and Evolution*, **12**, 887–892.

Burnham KP and Anderson DR (2002) *Model Selection and Multimodel Inference, A Practical Information-theoretic Approach*, Second Edition, Springer-Verlag: New York.

Cann R, Stoneking M and Wilson AJ (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society of London Series B*, **164**, 362–379.

Chikhi L and Bruford MB (2005) Mammalian population genetics and genomics. In *Mammalian Genomics*, Chapter 21, Ruvinsky A and Marshall Graves J (Eds.), CAB International publishing, pp. 539–584.

Chikhi L, Bruford MW and Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, **158**, 1347–1362.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V and Barbujani G (1998) Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 9053–9058.

Chikhi L, Nichols RA, Barbujani G and Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10008–10013.

Currat M and Excoffier L (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biology*, **2**, 2264–2274.

Currat M, Ray N and Excoffier L (2004) SPLATCHE: A program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.

Dawson KJ and Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.

Donnelly P and Tavaré S (1995) Coalescents and genealogical structure under neutrality. *Annual Reviews in Genetics*, **29**, 401–421.

Drummond AJ, Nicholls GK, Rodrigo AG and Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**(3), 1307–1320.

Eller E (2002) Population extinction and recolonisation in human demographic history. *Mathematical Biosciences*, **177&178**, 1–10.

Eswaran V, Harpending H and Rogers AR (2005). Genomics refutes an exclusively African origin of humans. *Journal of Human Evolution*, (EPub ahead of print).

Excoffier L (2002) Human demographic history: refining the recent African origin model. *Current Opinion in Genetics and Development*, **12**, 675–682.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fay JC and Wu CI (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution*, **16**, 1003–1005.

Felsenstein J (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, **59**, 139–147.

Fu YX and Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Garza JC and Williamson E (2001) Detection of reduction in population size using data from microsatellite DNA. *Molecular Ecology*, **10**, 305–318.

Goldstein DB and Chikhi L (2002) Human migrations and population structure: what we know and why it matters. *Annual Review of Genomics and Human Genetics*, **3**, 129–152.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL and Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6723–6727.

Griffiths RC and Tavaré S (1994) Simulating probability distributions in the coalescent. *Theoretical Population Biology*, **46**, 131–159.

Harpending H and Rogers A (2000) Genetic perspectives on human origins and differentiation. *Annual Review of Genomics and Human Genetics*, **1**, 361–385.

Hellmann I, Ebersberger I, Ptak SE, Pääbo S and Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics*, **72**, 1527–1535.

Hey J and Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**, 217–222.

Kimmel M, Chakraborty R, King JP, Bamshad M, Watkins WS and Jorde LB (1998) Signatures of population expansion in microsatellite repeat data. *Genetics*, **148**, 1921–1930.

Kimura M and Crow J (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

Knowles LL and Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2323–2635.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139–144.

Li N and Stephens M (2003) Modelling linkage disequilibrium and identifying recombination hotspots using single nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Marjoram P, Molitor J, Plagnol V and Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15324–15328.

Marth GT, Czabarka E, Murvai J and Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**, 351–372.

Marth GT, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, *et al*. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 376–381.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR and Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Nachman MW (2003) Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics and Development*, **12**, 657–663.

Nee S, Holmes EC, Rambaut A and Harvey PH (1995) Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London B*, **349**, 25–31.

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**, 358–364.

Nielsen R, Hubisz MJ and Clarck AG (2004) Reconstituting the frequency spectrum of ascertained single nucleotide polymorphism data. *Genetics*, **168**, 2373–2382.

Nielsen R and Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Nordborg M (2001) Coalescent theory. In *Handbook of Statistical Genetics*, Balding DJ, Bishop M and Cannings C (Eds.), John Wiley & Sons: New York, pp. 179–212.

Payseur BA and Nachman MW (2002) Gene density and human nucleotide polymorphism. *Molecular Biology and Evolution*, **19**, 336–340.

Pluzhnikov A, Di Rienzo A and Hudson RR (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, **161**, 1209–1218.

Polanski A, Kimmel M and Chakraborty R (1998) Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5456–5461.

Pritchard JK, Seielstad MT, Perez-Lezaun A and Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Ptak SE and Przeworski M (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, **18**, 559–563.

Rannala B and Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Ray N, Currat M and Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, **20**, 76–86.

Reich DE and Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8119–8123.

Rogers AR and Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.

Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, Gilbert MTP, Barnes I, Binladen J, *et al*. (2004) Rise and fall of the Beringian steppe bison. *Science*, **306**, 1561–1565.

Sherry ST, Batzer MA and Harpending H (1998) Modeling the genetic architecture of modern populations. *Annual Review of Anthropology*, **27**, 153–163.

Slatkin M and Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555–562.

Stephens M and Donnelly P (2000) Inference in molecular population genetics. *Journal of the Royal Statistical Society B*, **62**, 605–635.

Storz JF and Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, **56**, 154–166.

Storz JF, Payseur BA and Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution*, **21**, 1800–1811.

Strimmer K and Pybus OG (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, **18**, 2298–2305.

Stumpf MP and Goldstein DG (2003) Demography, recombination hotspot intensity, and the block-structure of linkage disequilibrium. *Current Biology*, **13**, 1–8.

Tajima F (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F (1989b) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 596–601.

Wakeley J, Nielsen R, Liu-Cordero SN and Ardlie K (2001) The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332–1347.

Wall JD and Przeworsky M (2000) When did the human population size start increasing? *Genetics*, **155**, 1865–1874.

Wang JL (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.

Watterson GA (1978) The homozygosity test of neutrality. *Genetics*, **88**, 405–417.

Weiss G and von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.

Wilson IJ and Balding DJ (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.

Won Y-J and Hey J (2005) Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, **22**, 297–307.

Zhivotovsky LA (2001) Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Molecular Biology and Evolution*, **18**, 700–709.

Specialist Review

Homeobox gene repertoires: implications for the evolution of diversity

Claudia Kappen
University of Nebraska Medical Center, Omaha, NE, USA

1. Introduction

With the completion of sequencing the DNA of numerous entire genomes, a comprehensive analysis of gene repertoires in evolutionary perspective becomes possible. Of particular interest are those types of genes that are thought to be responsible for, or contributing to, the evolutionary changes that manifest themselves in the various distinct species. One of the prominent classes of such genes are the homeobox genes, the developmental control genes that play a role in the formation of and cell differentiation in multiple tissues in organisms as diverse as plants, yeasts, and animals (Akam *et al*., 1994; Banerjee-Basu and Baxevanis, 2001; Kappen, 1995; Kenyon, 1994).

Homeobox genes encode transcription factors with a DNA binding domain (the so-called homeodomain) (Gehring *et al*., 1990), and have been shown to control the patterning and development of multiple tissue systems in animals (Tautz, 1996), including axial patterning in embryos, and of derivatives of all germ layers: nervous system in worms, flies, and vertebrates; mesodermal derivatives, such as heart and muscle in flies and vertebrates; and skeleton and skin in mammals, as well as endodermal derivatives, such as lung, pancreas, and gut in vertebrates. Furthermore, tissue differentiation in the hematopoietic system, kidney, and skeleton of vertebrates is controlled by homeobox genes, and they have been shown to be involved in cancer (Abate-Shen, 2002, 2003; Cillo, 1994; Cillo *et al*., 1999). Experimental as well as genetic changes in the function of homeodomain transcription factors alter embryonic development, tissue function, and cellular differentiation, thus making them excellent candidates for substrates of changes in the evolution of species with differences in these processes.

In plants, homeobox genes are also involved in patterning during development, but the divergence of gene sequences and function make it difficult to construct a conceptual equivalence of plant and animal homeobox genes. In fact, it is now well established that – while transcription factors encompass from 2 to 5% of the genome – the repertoires of homeobox genes in plants and animals are essentially distinct (Meyerowitz, 2002; Riechmann *et al*., 2000; Kappen, unpublished data).

For the purpose of this article, I will therefore focus on animal homeobox genes, and most specifically on those genomes for which complete sequence with appropriate annotation quality is available (*Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*). The goal of this article is to determine how the complement and composition (collectively accounting for the repertoire) of homeobox genes in diverse species can inform about the evolutionary basis of diversity between species.

To this end, I will extend my previous analysis from one completed genome (Kappen, 2000a) to multiple genomes. The earlier results strongly favored a model of "intercalation" of genes within the limits of a given repertoire of homeobox sequences. However, with only one completed (worm) and one semicomplete (at the time) genome (fly) analyzed, several important questions remained unanswered: (1) Does the complement of homeobox genes in a species relate to its complexity in body patterning? (2) Is there an identifiable pattern of gene multiplication as species with increased homeobox gene number arise? (3) Does the diversification of the homeobox gene repertoire follow a common trend along the evolutionary trajectory? or, in other words, are homeobox genes generally under common evolutionary constraints? (4) Can the patterns of history in homeobox gene evolution inform us about possible future trends? I will attempt to answer these questions on the basis of qualitative and quantitative analyses of homeodomain repertoires across relatively large evolutionary distances.

This focus will implicitly miss out on the exciting recent evidence for rapid homeobox gene evolution within shorter evolutionary time frames (Chow *et al*., 2001; Maiti *et al*., 1996; Schmid and Tautz, 1997; Sutton and Wilkinson, 1997; Ting *et al*., 1998, 2001). However, these reports are largely restricted to individual genes or individual evolutionary characters. While such investigations into individual homeobox gene function are indispensable to precisely decipher the relationship of gene regulation, gene function, and phenotypic character evolution (Averof and Patel, 1997), they cannot inform about patterns of evolution at the whole-genome level. In the same spirit, I will refer to excellent recent reviews on the evolution of individual subgroups of homeobox genes such as the Hox, Pax, En, Cut, Meis, and Knox genes (Bürglin, 1998; Bürglin and Cassata, 2002; Galliot *et al*., 1999; Gibert, 2002; Holland and Garcia-Fernandez, 1996; Kourakis and Martindale, 2000; Reiser *et al*., 2000; Steelman *et al*., 1997; Zhang and Nei, 1996) in favor of considering entire repertoires of homeodomains in selected model organisms. The expectation is that genome analyses will uncover trends of evolution at the gene as well as systems level, and in this way contribute to knowledge that may allow us to derive predictions about future trajectories of evolutionary change.

2. The evolution of homeobox gene repertoires in animals

2.1. *Content and classification of homeodomains in major animal species*

From the earliest invention of the homeodomain in a single-celled ancestral organism, expansion of the repertoire to the order of 100 homeodomains in invertebrates

and even more in vertebrates was achieved by duplication and diversification (Banerjee-Basu and Baxevanis, 2001; Kappen, 2000a). Traces of ancestry may still be evident from the relationships of homeodomains within a given genome. The information that can be derived from analyzing genomic repertoires has implications for speciation as well as evolution of diversity. There is growing recognition that the gene repertoires within the animal kingdom are largely conserved (Rubin *et al*., 2000). Indeed, for many vertebrate homeobox genes, orthologs have been identified in worm and fly, and vice versa. From whole-genome repertoires, we can now establish the extent of overlap, determine whether species-specific (or, in a broader perspective, clade-specific) homeobox genes exist, and assess degrees of conservation and divergence.

To this end, I have supplemented prior collections (Kappen *et al*., 1989, 1993; Kappen, 2000a) of homeodomain sequences with recently completed genomes by using the search functions "homeobox", "homeodomain", "homeo box", "homeo domain", and "homeo" in the NCBI Genome Browser. Nomenclature discrepancies were resolved according to sequence identity or on the basis of experimental literature. Zebrafish and Xenopus were omitted from any analyses performed here to avoid complications from tetraploidy. Classification was done as described previously (Kappen *et al*., 1993) and is in concordance with existing accepted classification schemes (Bürglin, 1995). Content of homeodomain sequences from respective species within each subgroup was determined from the sequence compilation by hand. The criteria for clade-specific (unique) genes were: (1) lack of an ortholog in other invertebrate or vertebrate species, (2) selective presence only in vertebrates, and (3) minimum distance of the "unique" sequence from any other sequence class in the same genome by at least seven residues (Kappen *et al*., 1993).

The schematic summary of these data is shown in Figure 1. Of the total of 154 homeodomain sequence classes, 80 are unique to one clade, with 53 of those found only in vertebrates. The latter include a number of proteins that contain both zinc-finger domains and homeodomains, often with multiple homeodomains in the same protein that each belong to a separate class. Under the restrictive definition of class membership used here, 53 classes are shared between two clades and 48 classes are shared between all three. The number of shared sequence classes clearly exceeds a random distribution: under Poisson distribution, one would expect 154 units to distribute into 57 single clades, 28 into two clades, and only 15 into the triple-clade category. The much higher number of sequence classes shared between all clades demonstrates the strong evolutionary conservation. This makes it possible to analyze in detail the evolution for each of the classes that are present in all clades.

2.2. Sequence subclasses and their contribution to the repertoire

Within each genome (Figure 2), subclasses of closely related sequences exist, such as the Dlx and Pbx subclasses. The subclasses are often of different size in different clades, reflecting either clade- or species-specific duplication events (as in case of the Dlx class) or strong conservation of genes known to have been duplicated in an ancestral organism, such as in the case of the Hox genes. Conserved sequence

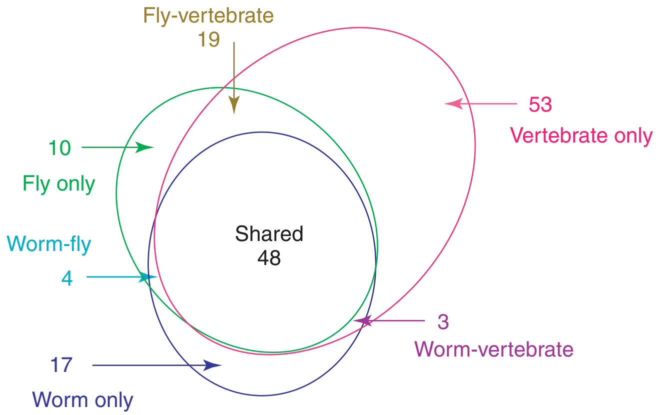


Figure 1 Overlap of homeodomain repertoires in major animal clades. The repertoire of homeodomain sequence classes for each animal clade is depicted as an ellipse whose area is approximately proportional to the number of sequence classes in each genome (72 for worm, 81 for fly, 123 for vertebrates). Areas of overlap depict sequence classes that are shared between clades

classes appear in all genomes analyzed here, as highlighted by same-color shading in Figure 2. In addition, clade-specific subclasses exist, such as a subclass of Zinc-finger homeobox genes that are restricted to the vertebrate lineage (Bayarsaihan *et al*., 2003). Genomes in separate clades (as represented by the model organisms) vary not only in total number of homeodomain sequences but also in relative contribution of each subclass to the overall repertoire (compare sizes of segment of a given color in Figure 2), motivating a more detailed analysis of genomic repertoires.

2.3. Distinct patterns of evolution for different subclasses of homeodomains

In particular, it is now possible to determine whether the increase in total number of homeodomains in vertebrates followed a specific pattern. For this analysis, I determined the number of sequences in each clade that belong to a particular sequence class. If the increase in total sequence number involved the entire genome uniformly, one would expect corresponding gene numbers in most classes. As Figure 3 shows, multiple scenarios are observed. There are sequence classes that contain the same number of genes in more than one clade, but more common are preferential expansions within one clade. When the subclass of the Hox genes – which has undergone quadruplication in vertebrates (Wagner *et al*., 2003) – is taken out of consideration, there is a general trend of doubling toward the vertebrate clade, providing the basis for the larger number of homeodomain sequences in vertebrate genomes. With this increase in gene number, the representation of most sequence classes within the genomic repertoire remains conserved (compare size of segments of same color in Figure 2), unless a subclass underwent disproportionate expansion

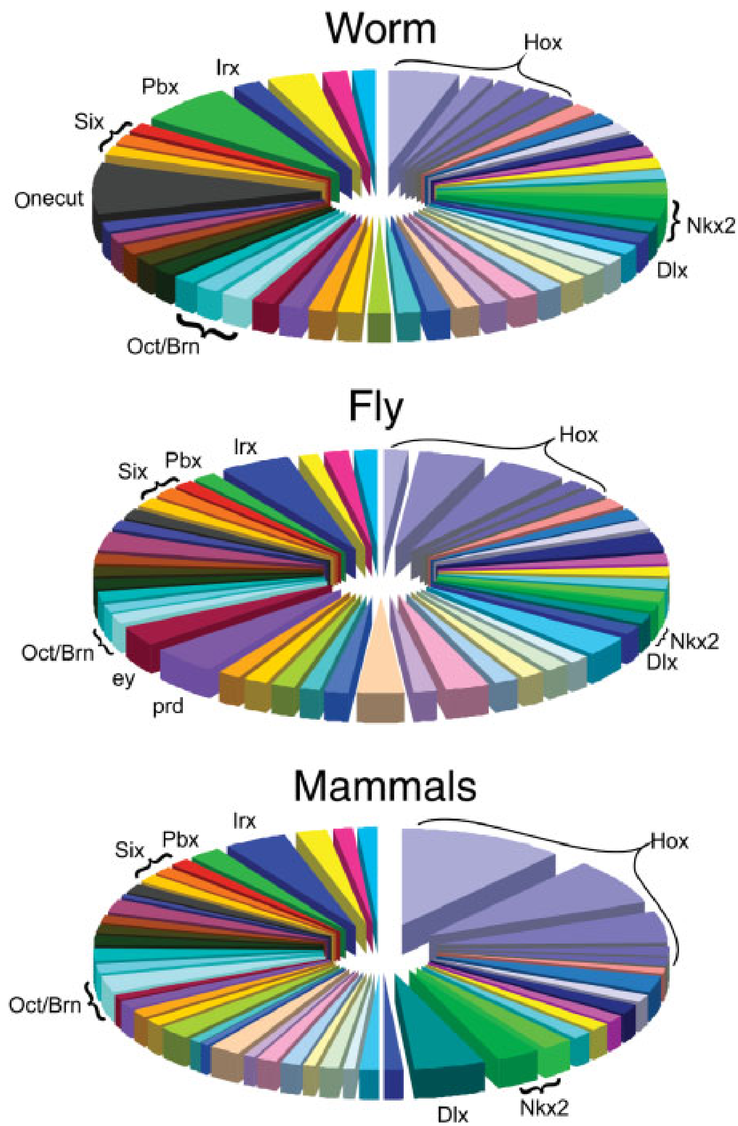


Figure 2 The repertoire of homeodomains in major animal clades. Each circle represents one genome with relative proportion of distinct homeodomain sequence classes shown as segments of a pie. Corresponding genes in the same subclass are labeled with the same color. It is evident that many subclasses are represented in all clades, but may be of different relative size in individual repertoires. For simplicity, subclasses that appear in only one clade were omitted. Several homeodomain subclasses are highlighted and specifically labeled: lilac shades: Hox-class homeodomains; green: Nkx 2 subclasses; teal: Dix class; turquoise: Oct/Brn homeodomains; red-orange-yellow: Six subclasses; green: Pbx class; blue: Irx class. The onecut (dark grey), prd (purple), and eyeless/Pax6 (wine red) subclasses show obvious clade-specific expansion in worm or fly, respectively. The Hox class contribution to the repertoire is enlarged in vertebrates

Worm	Fly	Human (*n*)
1	1	1 (9) 2 (17+2Hox) 3 (3) 4 (2) 7 (1)
1	2	1 (1) 2 (2) 3 (2) 6 (1)
1	3	2 (1) 6 (1) 7 (1Hox) 11 (1Hox)
2	1	3 (1) 4 (1)
3	1	1 (1) 16 (1Hox)
4	1	3 (1)
5	1	2 (1)

Figure 3 Patterns of homeodomain sequence distribution into subclasses. For each sequence class, the number of its gene members in each clade was determined. The occurrence of a given pattern was counted, and patterns are depicted here, in which each clade is represented by at least one gene. Seventeen such patterns were found (five patterns within the Hox subclass). In addition (not depicted here), five distinct distribution patterns with counterparts missing or unidentifiable in one clade were found seven times; thirteen patterns with one or more genes in only one clade accounted for the remainder, a large part constituted by the unique duplicated homeodomains in zinc-finger proteins in vertebrates. It is obvious that no single pattern dominates

or contraction. The results shown here demonstrate that there was no single mode of gene number expansion. Rather, it appears that each sequence class evolved by itself, suggesting independent mechanisms of selection.

2.4. Homeodomain subclasses evolved independently

This conclusion is supported by quantitative analyses. I performed two assays to define the degree of relationship of genes within the same sequence class: (1) pairwise distance analysis and (2) cladistic analyses.

For the distance analyses, I performed pairwise comparisons of sequences within each subclass, extending my earlier approach developed for individual species (Kappen, 2000a). Each amino acid difference was counted as 1 regardless of the type of amino acid residue found in either position. I have previously shown that the resolution achieved by this simple method is in good agreement with methods that use substitution matrices (Kappen, 2000a,b). Where a sequence in one clade corresponded to more than one in the other; all possible pairwise combinations were scored. Data are depicted separately for pairwise comparisons of clades. It is evident from the results for the worm-fly comparison (Figure 4a) that the distances between pairs of sequences vary from closely related (5 amino acid differences/60aa) to divergent. In fact, three major peaks are found, at 9, 14–16, and 29. Interestingly, the same patterns of distance distribution were found for worm-mammal comparisons (Figure 4b) and for fly-mammal comparisons (Figure 4c),

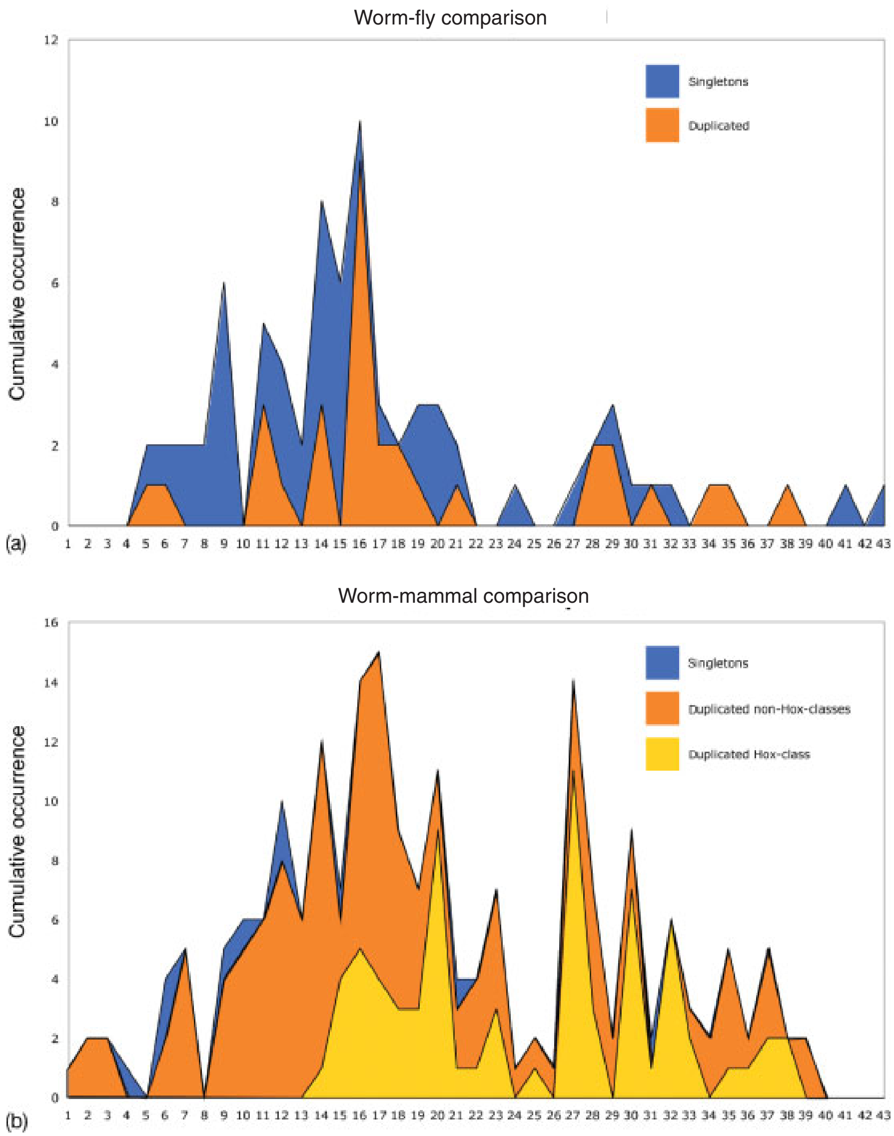


Figure 4 Distances between homeodomains within the same subclass. Pairwise comparisons were performed independently within each subclass and for each clade combination: (a) worm-fly pairwise comparisons; (b) worm-mammal pairwise comparisons; (c) fly-mammal pairwise comparisons. Datapoints contributed from sequences that have undergone duplication are added to single-gene (ortholog) comparisons. The cumulative height of each peak is the sum of all datapoints for a given distance. Color fills indicate blue: pairwise comparisons of corresponding homeodomains for which no duplicates exist in either clade (singletons); orange: pairwise comparisons of class members with each of their corresponding duplicate sequences from the other genome; yellow: pairwise comparisons of duplicate members of the Hox class

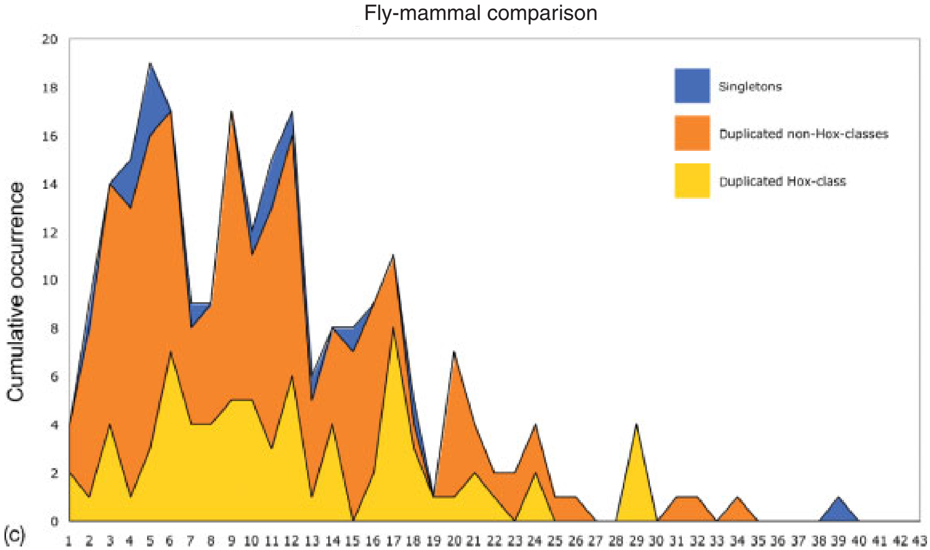


Figure 4 (*continued*)

although the distances were generally larger for worm-mammal comparisons and smaller for fly-mammal comparisons. The major peaks of worm-mammal distances are at 16–17, with shoulders at 20 and 23, and at 27; in contrast, fly-mammal distances peak at 4–5, 9–12, and 17. The difference in peak distances is largely contributed by the Hox subclass data, due to the divergence of worm Hox homeodomains. Accordingly, the similarity of Hox homeodomains between flies and mammals contributes to a shift toward shorter distances. The fact that distances distribute into recognizable peaks indicates that the sequences within some subclasses are more, or less, divergent than those in other subclasses. Finding this pattern for worm fly, fly-mammal, and worm-mammal comparisons suggests that the degree of divergence is a feature associated with a given subclass.

Several possible explanations can account for these data: Where subclasses have duplicates, these duplicates could have arisen at different times, with older duplicates allowing for more divergence than younger duplicates. About half (35/67 = 52%) of the data points of the worm-fly comparisons are for single orthologous sequence pairs, 47.8% from duplications. However, as the cumulative plot in Figure 4(a) shows, the distributions of data for singletons and duplicates are very similar, making it unlikely that duplicates would have a strong influence on the results. The more plausible scenario is that differential divergence reflects different selective pressure for distinct subclasses. This interpretation provides a scenario in which homeodomain sequence subclasses were under different evolutionary constraints – most likely related to function, although chromosomal structure has also been implicated (Wagner *et al*., 2003) as a selective force – and thus, subclasses evolved independently from each other during the divergence of worm and fly from their last common ancestor.

Interestingly, the same general pattern is found for the worm-mammal (Figure 4b) and fly-mammal comparisons (Figure 4c). Here, duplications in either clade after divergence from the last common ancestor account for 94.8% (worm-mammals) and 93.6% (fly-mammals) of the data points, and would be consistent with different time points for clade-specific duplications. However, since differential selection pressures in the different clades are also possible, distance analyses alone are not sufficient to resolve this aspect. What these results do show, however, is clear support for the conclusion that homeodomain subclasses evolved independently from each other.

The prediction from these results is that phylogenetic analyses would reveal the divergence of sequences within a subclass both with regard to sequence and timing of duplications, and with regard to degree of divergence. Therefore, we subjected separate subclasses to phylogenetic analyses using cladistics.

For the cladistic analyses, only those subclasses are informative where multiple sequences exist in at least two clades. From the data in Figure 3, 15 such cases are available. Gene trees were generated in PAUP in exhaustive (for <12 sequences) or "branch-and-bound" searches with Yeast (*Saccharomyces cerevisiae*) Matα2 as the outgroup. While this may bias the position of the root relative to each tree, it will have no influence on topology and branch length in the tree itself. Such trees are therefore suitable to assess relative divergence of subgroup members in comparison to other subgroups. The prediction here is that if all subclasses were under similar evolutionary pressure, the branching topologies and branch lengths should be largely congruent. At least for each subset of distinct distribution patterns, one would expect highly similar tree topologies and branch lengths. Alternatively, different topologies and varying branch lengths would further support the notion that subclasses evolved independently.

This latter scenario is indeed advanced by the results (Figure 5): Two cases exist of the distribution pattern 1,2,2 (worm, fly, mammals), the Barx and the En subclass, respectively (Figure 5a and b). In both, the topologies of the gene trees are similar, but branch lengths are quite different in that the En class mammalian sequences are closely related, while this is the case only for the fly sequences in the Barx class. Another group that has two vertebrate members is the prd subclass (Figure 5c). Here, the fly genes – although recognizable duplicates of each other – are not closely related, and the worm member is also quite divergent. The other subclasses also exhibit distinct topologies and divergence patterns.

An interesting corollary of the results in Figure 5 is that fly sequences show, in the majority of cases, greater affinity to vertebrate sequences than to worm sequences, highlighting what is believed to be the very derived nature of the worm genome.

Furthermore, the overall tree length is not directly correlated to number of members in the class: the onecut (Figure 5f) and Pbx (Figure 5h) subclasses each have eight members, with the latter tree being shorter than the former. Rather, tree length is greater with greater divergence of worm sequences, most prominently illustrated by the Pbx (Figure 5h), Meis (Figure 5i) and Irx (Figure 5j) subclasses: each of these is very compact with short branch lengths for mammalian and fly sequences. The comparison of these subclasses to the other groups also shows that duplicated homeodomains in vertebrates are divergent from each other to varying degrees,

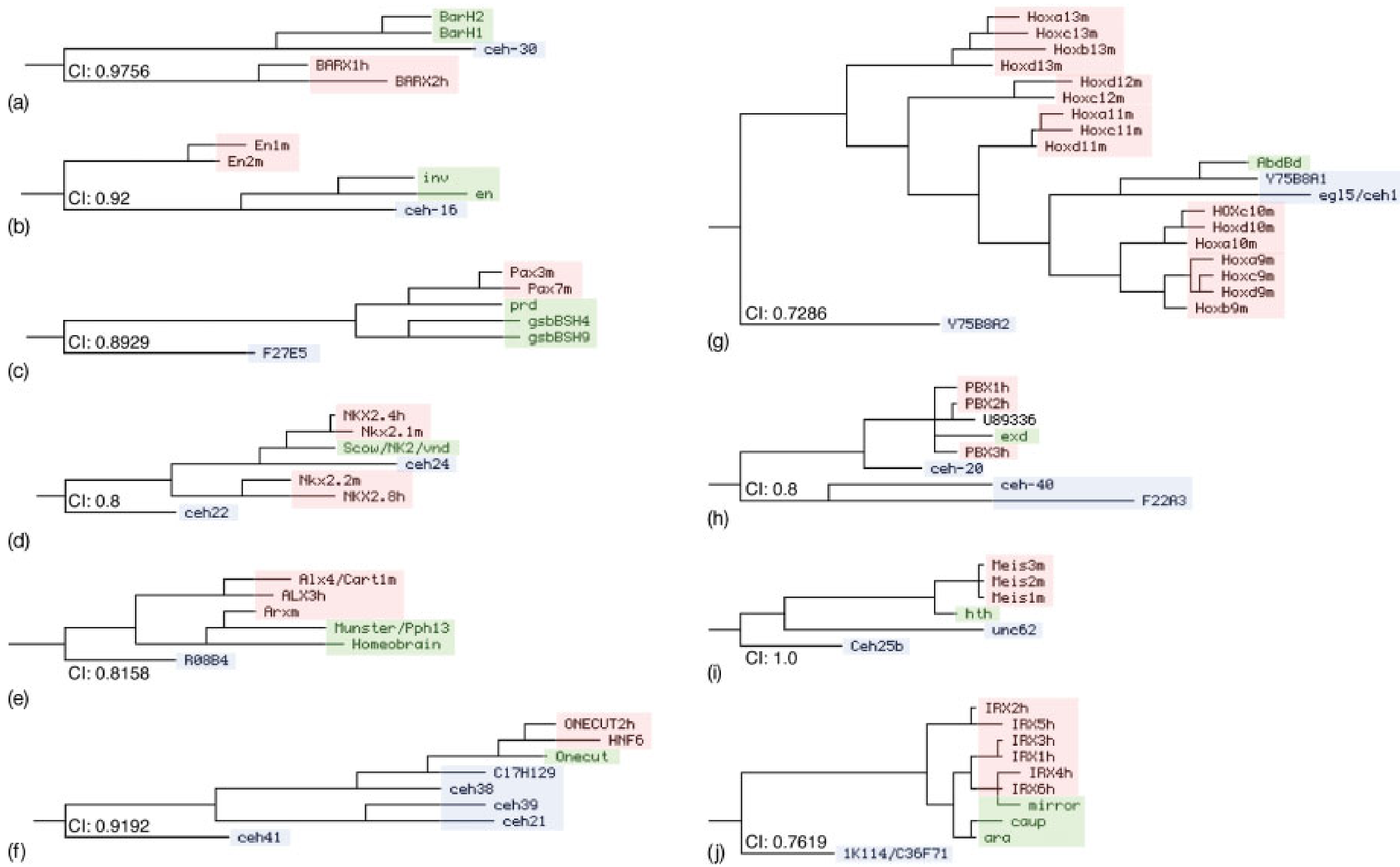


Figure 5 Phylogenetic patterns of homeodomain subclass evolution. Subclasses with multiple sequences in at least two clades were analyzed independently in exhaustive searches with either human or mouse serving as mammalian representatives. The consistency indices (excluding uninformative characters) were uniformly high, providing strong support for each tree. The somewhat reduced consistency indices for (g) and (j) reflect equally possible topologies of the same overall tree length with variable assignment of terminal branches due to a low number of parsimony-informative characters in the vertebrate homeodomain sequences of these subclasses. All trees were rooted to the same outgroup (Yeast Matα2) and scaled to the same scale for branch length. Shading denotes: red: mammalian sequences; green: fly sequences; and blue: worm sequences

from almost identical in the Meis group to greatest divergence in the Barx and Hox (Figure 5g) subclasses. These results are very consistent with the distance measurements. Thus, sequences in different subclasses evolved at different rates, arguing for distinct evolutionary pathways for each subclass.

2.5. *Duplication of homeodomains through genome-wide duplications: reconciliation of a controversy*

These types of results have – by others (Friedman and Hughes, 2003; Hughes, 1999) – been taken as arguments against the occurrence of genome-wide duplications in vertebrates. Indeed, the varying distances relative to the root of the branch points for mammalian homeodomains can be interpreted to mean that duplications of vertebrate homeodomains occurred at different points in time. One implication of this notion is that branch length is proportional to time; however, the scoring here was for replacement changes, not for silent nucleotide differences. Amino acid replacements alter protein structure, which in turn with altered function is subject to selection. Certainly, the distinct functional roles of the homeodomains in different subclasses provide the possibility that they could have been subjected to distinct selection pressures. Unfortunately, extending the analyses beyond the homeodomains does not improve resolution owing to the long divergence time since the postulated duplications. Noncoding sequences are also so divergent that no signatures of time elapsed since duplication remains (Kappen, unpublished). This is even more so for comparisons of duplicates in fly or worm, respectively. Rather than invoking consecutive independent duplications, different selection modes would also account for the three peaks of divergence found in the quantitative analysis (see Figure 4). As such, the different divergence rates at the protein level are not inherently inconsistent with genome duplications if independent evolutionary scenarios are applied to each class after duplication. If the duplications that created the vertebrate homeodomains indeed all happened at the same time, distances at silent nucleotide positions should be comparable between different subclasses.

In order to attempt to resolve the possible time points of duplication for the vertebrate homeodomains in different subclasses, I measured the distances between the vertebrate homeoboxes at the nucleotide level, taking into account only the silent sites. We have previously used this approach to determine the time point of Hox cluster duplication at 450 million years ago (Kappen *et al*., 1989), an estimate now well supported by experimental surveys of various chordate genomes. The prediction for these analyses is that equal rates of divergence at silent sites for independent subclasses predict the same evolutionary time point for duplication – such as through whole-genome duplication – while different rates have to be interpreted as reflecting independent duplications successive in time. This latter mode of duplication would indeed argue against whole-genome duplication in the origin of vertebrate evolution.

Figure 6 summarizes the distance measurements between duplicate mammalian homeobox nucleotide sequences. In all classes, numerous nucleotide differences were found, with measurements for the Engrailed and Nkx subclasses being

Subclass	Total changes	Replacement changes	Proportion of replacements/changes (%)	Proportion of silent changes/ possible silent sites (%)
Barx1/2	38	9	23.7	46.8
En1/2	26	3	11.5	37.1
Pax3/7	35	4	11.4	53.0
Nkx2 cognate pairs	26.5 ± 17.5	6.5 ± 4.5	17.8 ± 7.2	38.3 ± 11.7
Alx	46.7 ± 5.9	12.3 ± 3.8	26.4 ± 7.96	54.0 ± 9.0
Onecut1/2	47	15	31.9	55.2
Hox cognate pairs	28.7 ± 6.6	3.8 ± 2.2	13.2	57.1
Pbx	34 ± 3.6	2	5.95 ± 0.67	50.0 ± 5.9
Meis	34 ± 4.3	1.7 ± 0.5	4.8 ± 0.88	53.9 ± 6.4
Irx cognate pairs	11.3 ± 3.4	1 ± 0.8	7.5 ± 5.4	16.2 ± 4.1
Irx all	14.4 ± 4.8	5.6 ± 3.5	37.4 ± 19.3	15.4 ± 4.6

(a)

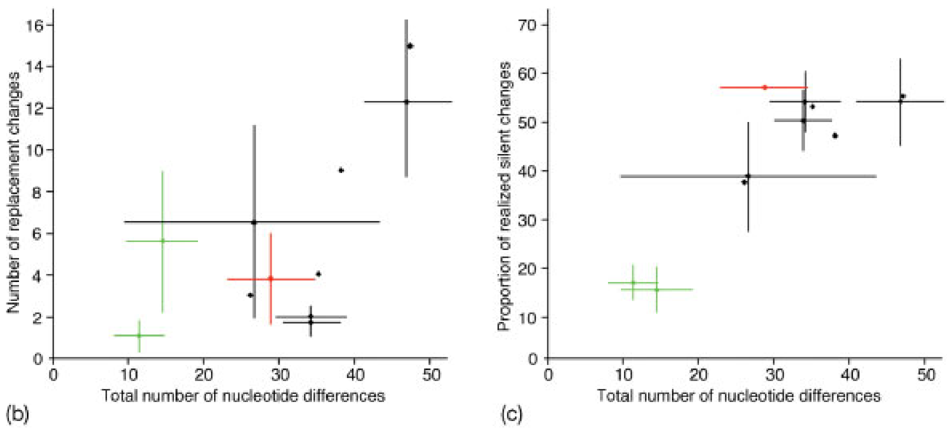


Figure 6 Analysis of nucleotide differences between homeobox sequences of different subclasses. (a) Pairwise sequence comparisons for the nucleotide sequences of the subclasses depicted in Figure 5. For all pairwise comparisons, the number of possible silent sites was very similar (between 57 and 66, with a mean of 62.1 + 2.5). Where genomic organization in clusters occurs (Nkx, Hox, Irx), pairwise comparisons were performed for the cognate paralogs. (b) Relationship of nucleotide differences to amino acid differences. (c) Relationship of nucleotide differences to silent changes. The data in (b) and (c) are plotted from (a). The data for the Hox subclass were taken from Kappen C, Schughart K and Ruddle FH (1989) Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 5459–5463

comparable to previous results for the Hox subclass. The other classes, with the exception of the Irx subclass (discussed below), exhibit even greater distances. This corresponds to a larger number of replacements that change amino acid sequence of the homeodomain, although there was not a strict correlation (Figure 6b). The proportion of replacement changes ranged from about 5% (Meis and Pbx subclasses) to over 30% (onecut subclass). These results are reflected in the different terminal branch lengths in the amino acid–based gene trees for these subclasses (see Figure 5j, f, and i respectively), and underscore the differential degree of protein sequence conservation in distinct subclasses. However, the degree of sequence divergence at the protein levels is selection-dependent.

A selection-independent measure of sequence divergence is presented by silent nucleotide differences. The total number of silent sites was consistently around 60 in all subclasses, and the degree of nucleotide differences at silent sites was

quite similar for all subclasses (between 37.1 and 57.1%; except for the Irx class, discussed below). These data are most consistent with the interpretation that duplicated homeodomains originated at about the same time, having been subject to the same degree of divergence. Thus, while selective pressure on the protein level may have been different for homeodomain subclasses, the rate of silent change accumulation indicates a common time point for the origin of the duplications. Of all potential silent changes, approximately 50% are realized in any given pairwise comparison, arguing against saturation of the silent sites. By the same logic, the fraction of multiple hits at the same site – which I did not correct for in the present analysis – is expected to be too low to substantially affect the major conclusions. Thus, with a similar rate of silent differences, the divergence time since duplication must have essentially been the same, presumably because all duplicates arose at the same time point.

The fact that the rate of silent changes is very similar to that found previously for the Hox subclass (Kappen *et al*., 1989) allows us to place the duplication time point coincident with the Hox cluster duplications, with the emergence of the vertebrate lineage. After duplication, different subclasses then were subject to distinct selective forces. Thus, using a combination of nucleotide and amino acid sequence analyses and employing cladistic as well as distance analyses, I propose to reconcile the existing controversies and the apparent contradiction of conclusions from protein sequence and nucleotide sequences, respectively.

2.6. *The exception to the rule: evidence for an unidentified mechanism that constrains divergence*

However, a glaring contradiction to my conclusions is constituted by the Irx sequence subclass: The mammalian Irx genes are localized on two genomic clusters and are coordinately regulated with their position in these clusters (Peters *et al*., 2000). There are clearly fewer nucleotide differences between cognate members of the two Irx gene clusters than for any other gene duplicates (line 10 in Figure 6a), even compared to cognate paralogous genes within the Hox clusters (line 7 in Figure 6a). The nucleotide differences that are present account for amino acid changes to approximately the same degree as in Hox genes (13.2% vs. 7.5 ± 5.4%). However, the major discrepancy is constituted by far fewer silent differences between Irx cognate genes. One theoretically possible scenario would be that the Irx gene cluster duplicated independently from all the other subclasses, and later in evolution. Yet, the three ancestral Irx genes would have arisen before the evolution of the vertebrates (since cognates are present in all vertebrates sampled to date) and therefore could have diverged substantially, particularly at silent sites. This would be apparent in large silent nucleotide distances between noncognate pairs of Irx sequences. However, when all possible pairwise comparisons of Irx homeoboxes are performed (line 11, Figure 6a), the fraction of replacement changes increases (as expected), yet the rate of silent change accumulation is still exceptionally low (see red data points in Figure 6c).

There are several possible explanations for this finding: (1) The ancestral Irx genes arose very late in vertebrates – this is clearly contradicted by the presence of

multiple Irx sequences in chicken and zebrafish; sufficient time would have elapsed for the accumulation of silent changes. (2) The origin of three precursor Irx-like genes was independent in fly and vertebrate lineage, and preceded the Irx cluster duplications – the low number of silent differences between Irx noncognates is inconsistent with this scenario; the proportion of silent changes should be higher for noncognate pairs than for cognates. (3) Irx genes are subject to higher order "repair" processes or a fundamentally lower mutation rate than other homeobox genes. This could have to do with their unique organization in coordinately regulated clusters, but would differ from the Hox genes, which are also located in gene clusters. There is currently no evidence available, but chromatin structure, gene conversion, recombination events between Irx clusters, template-mediated repair, or constraint from functional sequences on the opposite strand would be plausible mechanisms. Then, unless one dismisses the entire methodology of silent site analysis as invalid, the results for the Irx homeobox subclass suggest the presence of yet undiscovered mechanisms that limit sequence divergence.

Taken together, these considerations suggest a unique modus of evolution for the Irx homeobox subclass that remains to be investigated. The evolutionary history of the majority of homeodomain subclasses, nevertheless, is consistent with prior hypotheses, most notably with the notion of genome duplications at the origin of the vertebrate clade.

3. Implications for the evolution of diversity

In response to the major questions asked at the outset of this study, the following conclusions emerge:

3.1. On relationship of genomic repertoire and organismal complexity

Does the complement of homeobox genes in a species relate to its complexity in body patterning?

At the gross level, the number of homeobox genes increases in vertebrates who have more complex body plans relative to flies and worms with simpler anatomy (Marshall *et al*., 1999). The repertoire of shared sequence classes also, at gross level, is very comparable between clades. Yet, each clade also harbors clade-specific sequence classes, and their relative contribution to the repertoire varies: clade-specific sequence classes make up 23.6% of all sequence classes in the worm, 12.3% in the fly, and 43% in mammals. Thus, in addition to strong conservation of shared repertoires, unique features are present and contribute to overall complexity, at least based upon sequence numbers and their distribution into subclasses. Yet, well-defined measures for complexity of gene repertoires are needed to unequivocally answer the fundamental question. From a sequence-focused viewpoint, parameters of divergence, repertoire composition, representation of duplicates and so on would all have to be considered in a formal definition of complexity.

Measures of biological complexity are even more difficult. Not only does tissue architecture vary, but even with apparently conserved molecular networks, flies and

worms undergo principally determined developmental processes that, in contrast, are subject to regulation in vertebrates (Davidson and Ruvkun, 1999). Certainly, vertebrate innovations have increased body plan complexity, and homeobox genes feature prominently in the development of these structures (Shimeld and Holland, 2000; Wagner *et al*., 2003), but which specific changes in gene repertoires are relevant to these innovations is less well understood (Szathmary *et al*., 2001).

In addition, more detailed studies are needed to elucidate the functional roles of clade-specific homeobox subclasses and subclasses with clade-specific expansions. It would be expected that such classes control the particular features of body plans in a given animal clade. Indeed, rapid divergence at the OdsH locus has been implicated in fly speciation events (Ting *et al*., 1998, 2001), but evidence for additional such examples is currently lacking. For vertebrates, it has been speculated that the expansion of the AbdB-like Hox subclass enabled limb development (Izpisua-Belmonte and Duboule, 1992); but the roles of these genes in primary axial development (Favier *et al*., 1996; Fromental-Ramain *et al*., 1996; Zakany *et al*., 1996) and cell-type-specific differentiation (Godwin and Capecchi, 1998; Rijli *et al*., 1995; Satokata *et al*., 1995) complicate the determination of which of their functions was primordial in evolution. Furthermore, the functional significance of clade-specific gene duplications and expansions in the homeodomain repertoire remains to be investigated.

3.2. On the mechanisms of homeodomain subclass expansion

Is there an identifiable pattern of gene multiplication as species with increased homeobox gene number arise?

The answer here is an unequivocal “no”. On the basis of qualitative, distance, and cladistic analyses, it becomes clear that the various homeodomain subclasses evolved independently of each other. Most notably, some subclasses enjoyed expansion in clade-specific fashion, suggesting biological roles in processes particular to this clade. Within the vertebrates, many subclasses contain duplicates that appear to have emerged at about the same time point as the Hox cluster duplications. This is consistent with the hypothesis of genome duplications at the origin of the vertebrates (Bailey *et al*., 1997; Friedman and Hughes, 2003; Holland *et al*., 1994; Pebusque *et al*., 1998). Independent analyses at level of nucleotide sequences for nine distinct homeobox subclasses strongly support the conclusion that expansion of vertebrate homeodomain subclasses could have been achieved by genome duplications.

Deviation from this scenario is found for the Irx subclass, either because of peculiar evolutionary history or because of particular chromatin features. The relevance of clustered organization for function of the Irx homeobox gene subclass remains to be established experimentally. Regardless of the outcome of these future studies, my complementary analyses of protein and nucleotide sequences demonstrate that conclusions based entirely on only one of these approaches may be biased and incomplete (Agosti *et al*., 1996; Marshall *et al*., 1999). In this regard, a thorough treatment of the controversy surrounding genome duplications will require parallel nucleotide and amino acid sequence analyses for multiple gene classes. This

may still be restricted to the most conserved domains of a gene, but the extension to multiple diverse gene families (Van de Peer *et al*., 2001) would likely overcome features peculiar to any given family, such as chromosomal position, clustering, chromatin structure, differential recombination rates of loci and the like. While an absolute estimate in terms of a molecular clock may be impossible to achieve, the relative rates of evolution at silent sites in various multigene families should be at least as informative as the analyses performed here for the homeobox gene sequences.

3.3. On the evolutionary trajectories of homeodomains

Does the diversification of the homeobox gene repertoire follow a common trend along the evolutionary trajectory; are homeobox genes generally under common evolutionary constraints?

The answer here again is an unequivocal "no". Qualitative analyses reveal different compositions of homeodomain subclasses in patterns that are not generalizable but appear specific to a given sequence subclass. Distance as well as cladistic analyses also establish distinct patterns and degrees of divergence. These distinct evolutionary trajectories argue for independent selection mechanisms for each subclass.

Relative to postulated ancestral genes (or, where available, sequences in early chordates), it is not possible to ascertain which of the two vertebrate duplicates can be considered "ancestral" or "derived" copy. This is also the case in worm, where the divergence of duplicates is greatest. The observed patterns of sequence divergence are thus not easily compatible with the idea of "freeze one – diverge the other" (Bailey *et al*., 1997; Ruddle *et al*., 1994; Van de Peer *et al*., 2001). On the other hand, the large distances of the clades included in the analyses presented here may obscure events on finer evolutionary time scale. Nevertheless, for vertebrates, the data do not provide much support for the "gene freeze" hypothesis.

Yet, at a time of rapid radiation and extensive diversification of species, such as in the pre-Cambrian prior to emergence of the chordate lineage, divergence may have been possible without immediate compromise of successful organismal development. In fact, it is intriguing that replacement of the majority of residues in the homeodomain with a uniform sequence of Alanines does not change regulatory specificity (in vitro!) as long as the N-terminal arm and the DNA binding helix 3 remains unaltered (Shang *et al*., 1994). Given that regions outside of the homeodomains are vastly different even in the corresponding (believed to be orthologous) genes in different clades, it remains to be investigated to which extent amino acid replacements account for changes in homeodomain function. Where attempted, replacement – in genomo – of one homeobox gene with a duplicate often rescues the majority of functions (Greer *et al*., 2000; Hanks *et al*., 1995; Zakany *et al*., 1996; Zhao and Potter, 2002), arguing for extended functional redundancy even for divergent nonhomeodomain sequences. The ability of fly and vertebrate homeodomain proteins to functionally substitute for each other (Malicki *et al*., 1990, 1993; McGinnis *et al*., 1990; Zhao *et al*., 1993) also argues against significant functional involvement of nonconserved protein domains. These

studies underscore the importance of the homeodomain – and, implicitly, of DNA binding – as its major function. The significance of amino acid residue changes in the evolution of a given homeodomain in different clades remains to be elucidated. One prediction from these considerations for human disease is that the most phenotypically relevant mutations would be expected to occur in or affect the function of the homeodomain.

3.4. *On the predictive aspects of molecular evolutionary analyses*

Can the patterns of history in homeobox gene evolution inform us about possible future trends?

Here, the limitations of focusing on major clades become most apparent: Clearly, in order to define evolutionary trajectories of specific homeodomain subclasses in any detail, it would be helpful to obtain the sequence of a given homeobox gene from as many species as possible. Then, it becomes possible to define changes that are unique to a given species, and, thus, uninformative unless associated with functional alterations. Conversely, changes common to multiple species can at least serve as a sequence trace of common ancestry, and possibly, of shared functional changes. Eventually, all manifest sequence differences in extant species can be classified as (1) tolerable and viable variations or (2) viable but functionally relevant replacements; detrimental changes will not be fixed in any population. Whether or not enough natural variations will be found to resemble a mutational series – in realo – remains to be seen, but such a dataset would clearly help in assigning functional relevance to particular residues.

The de facto possible variations resemble potential directions of future change for existing homeodomains, and structure predictions would help in eliminating – at least theoretically – from this vast reservoir those changes that will disrupt homeodomain structure. Knowing evolutionary trajectories would allow to map a pathway in sequence space, and the projection of this pathway would indicate potential future directions (Kappen, 2000a). This should presently be feasible for changes at single residues, but even less evidence is available on coselected residues; yet, cooperation of distant residues within the homeodomain for determining specificity is well established (Billeter *et al*., 1996; Gehring *et al*., 1994). Extended evolutionary analyses will thus complement and provide predictive power to experimental approaches.

4. Conclusion

With regard to organismal diversity, the present study advances a picture characterized by multiple distinct evolutionary trajectories for subclasses of homeodomains. Under this scenario, the different biological processes utilizing these homeodomain transcription factors are each themselves subject to distinct evolutionary forces. To determine how these forces integrate at the level of gene repertoires and individual organisms remains the challenge for future experimental and molecular evolutionary studies.

The present analyses also advance a theoretical framework that can be extended to virtually any genome and multiple gene families. My studies demonstrate the power of results generated from a combination of – applied in parallel to the same datasets – qualitative and quantitative methods. Through such combinations of complementary approaches, it should be possible in the future to precisely delineate the mechanisms in molecular evolution that produce changes at the level of individual development and organismic evolution.

Acknowledgments

I wish to dedicate this article to Dr Frank H. Ruddle, my postdoctoral supervisor, on the occasion of his retirement. His encouragement has been invaluable. I also thank my parents for fostering my curiosity and adventurous spirit, and Dr J. Michael Salbaum for continuing critical discussions and companionship in real-life adventures. Saralyn Fisher provided expert assistance in preparation of the manuscript.

References

Abate-Shen C (2002) Deregulated homeobox gene expression in cancer: cause or consequence? *Nature Reviews. Cancer*, **2**, 777–785.

Abate-Shen C (2003) Homeobox genes and cancer: new OCTaves for an old tune. *Cancer Cell*, **4**, 329–330.

Agosti D, Jacobs D and DeSalle R (1996) On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics*, **12**, 65–82.

Akam M, Averof M, Castelli-Gair J, Dawes R, Falciani F and Ferrier D (1994) The evolving role of Hox genes in arthropods. *Development Supplement*, 209–215.

Averof M and Patel NH (1997) Crustacean appendage evolution associated with changes in Hox gene expression. *Nature*, **388**, 682–686.

Bailey WJ, Kim J, Wagner GP and Ruddle FH (1997) Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Molecular Biology and Evolution*, **14**, 843–853.

Banerjee-Basu S and Baxevanis AD (2001) Molecular evolution of the homeodomain family of transcription factors. *Nucleic Acids Research*, **29**, 3258–3269.

Bayarsaihan D, Enkhmandakh B, Makeyev A, Greally JM, Leckman JF and Ruddle FH (2003) Homez, a homeobox leucine zipper gene specific to the vertebrate lineage. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 10358–10363.

Billeter M, Guntert P, Luginbuhl P and Wüthrich K (1996) Hydration and DNA recognition by homeodomains. *Cell*, **85**, 1057–1065.

Bürglin TR (1995) The evolution of homeobox genes. *Biodiversity and Evolution*, The National Science Museum Foundation: Japan, pp. 291–336.

Bürglin TR (1998) The PBC domain contains a MEINOX domain: coevolution of Hox and TALE homeobox genes? *Development, Genes and Evolution*, **208**, 113–116.

Bürglin TR and Cassata G (2002) Loss and gain of domains during evolution of cut superclass homeobox genes. *International Journal of Developmental Biology*, **46**, 115–123.

Chow RL, Snow B, Novak J, Looser J, Freund C, Vidgen D, Ploder L and McInnes RR (2001) Vsx1, a rapidly evolving paired-like homeobox gene expressed in cone bipolar cells. *Mechanisms of Development*, **109**, 315–322.

Cillo C (1994) HOX genes in human cancers. *Invasion and Metastasis*, **14**, 38–49.

Cillo C, Faiella A, Cantile M and Boncinelli E (1999) Homeobox genes and cancer. *Experimental Cell Research*, **248**, 1–9.

Davidson EH and Ruvkun G (1999) Themes from a NASA workshop on gene regulatory processes in development and evolution. *Journal of Experimental Zoology*, **285**, 104–115.

Favier B, Rijli FM, Fromental-Ramain C, Fraulob V, Chambon P and Dolle P (1996) Functional cooperation between the nonparalogous genes Hoxa-10 and Hoxd-11 in the developing forelimb and axial skeleton. *Development*, **122**, 449–460.

Friedman R and Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Molecular Biology and Evolution*, **20**, 154–161.

Fromental-Ramain C, Warot X, Lakkaraju S, Favier B, Haack H, Birling C, Dierich A, Dolle P and Chambon P (1996) Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning. *Development*, **122**, 461–472.

Galliot B, de Vargas C and Miller D (1999) Evolution of homeobox genes: Q50 Paired-like genes founded the Paired class. *Development, Genes and Evolution*, **209**, 186–197.

Gehring WJ, Müller M, Affolter M, Percival-Smith A, Billeter M, Qian YQ, Ötting G and Wüthrich K (1990) The structure of the homeodomain and its functional implications. *Trends in Genetics*, **6**, 323–329.

Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Ötting G and Wüthrich K (1994) Homeodomain-DNA Recognition. *Cell*, **78**, 211–223.

Gibert JM (2002) The evolution of engrailed genes after duplication and speciation events. *Development, Genes and Evolution*, **212**, 307–318.

Godwin AR and Capecchi MR (1998) Hoxc13 mutant mice lack external hair. *Genes and Development*, **12**, 11–20.

Greer JM, Puetz J, Thomas KR and Capecchi MR (2000) Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature*, **403**, 661–665.

Hanks M, Wurst W, Anson-Cartwright L, Auerbach AB and Joyner AL (1995) Rescue of the En-1 mutant phenotype by replacement of En-1 with En-2. *Science*, **269**, 679–682.

Holland PW and Garcia-Fernandez J (1996) Hox genes and chordate evolution. *Developmental Biology*, **173**, 382–395.

Holland PW, Garcia-Fernandez J, Williams NA and Sidow A (1994) Gene duplications and the origins of vertebrate development. *Development Supplement*, 125–133.

Hughes AL (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of Molecular Evolution*, **48**, 565–576.

Izpisua-Belmonte JC and Duboule D (1992) Homeobox genes and pattern formation in the vertebrate limb. *Developmental Biology*, **152**, 26–36.

Kappen C (1995) Ancient evolutionary origin of homeodomains: studies on yeasts, plants and animals. In *Current Topics on Molecular Evolution*, Nei M and Takahata N (Eds.), Pennsylvania State University: College Park, pp. 199–212.

Kappen C (2000a) Analysis of a complete homeobox gene repertoire: implications for the evolution of diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4481–4486.

Kappen C (2000b) The homeodomain: an ancient evolutionary motif in animals and plants. *Computers and Chemistry*, **24**, 95–103.

Kappen C, Schughart K and Ruddle FH (1989) Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 5459–5463.

Kappen C, Schughart K and Ruddle FH (1993) Early evolutionary origin of major homeodomain sequence classes. *Genomics*, **18**, 54–70.

Kenyon C (1994) If birds can fly, why can't we? Homeotic genes and evolution. *Cell*, **78**, 175–180.

Kourakis MJ and Martindale MQ (2000) Combined-method phylogenetic analysis of Hox and ParaHox genes of the metazoa. *Journal of Experimental Zoology*, **288**, 175–191.

Maiti S, Doskow J, Sutton K, Nhim RP, Lawlor DA, Levan K, Lindsey JS and Wilkinson MF (1996) The Pem homeobox gene: rapid evolution of the homeodomain, X chromosomal localization, and expression in reproductive tissue. *Genomics*, **34**, 304–316.

Malicki J, Bogarad LD, Martin MM, Ruddle FH and McGinnis W (1993) Functional analysis of the mouse homeobox gene HoxB9 in Drosophila development. *Mechanisms of Development*, **42**, 139–150.

Malicki J, Schughart K and McGinnis W (1990) Mouse Hox-2.2 specifies thoracic segmental identity in Drosophila embryos and larvae. *Cell*, **63**, 961–967.

Marshall CR, Orr HA and Patel NH (1999) Morphological innovation and developmental genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9995–9996.

McGinnis N, Kuziora MA and McGinnis W (1990) Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. *Cell*, **63**, 969–976.

Meyerowitz EM (2002) Plants compared to animals: the broadest comparative study of development. *Science*, **295**, 1482–1485.

Pebusque MJ, Coulier F, Birnbaum D and Pontarotti P (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution*, **15**, 1145–1159.

Peters T, Dildrop R, Ausmeier K and Rüther U (2000) Organization of mouse Iroquois homeobox genes in two clusters suggests a conserved regulation and function in vertebrate development. *Genome Research*, **10**, 1453–1462.

Reiser L, Sanchez-Baracaldo P and Hake S (2000) Knots in the family tree: evolutionary relationships and functions of knox homeobox genes. *Plant Molecular Biology*, **42**, 151–166.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, *et al*. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Rijli FM, Matyas R, Pellegrini M, Dierich A, Gruss P, Dolle P and Chambon P (1995) Cryptochordism and homeotic transformations of spinal nerves and vertebrae in Hoxa-10 mutant mice. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 8185–8189.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al*. (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Ruddle FH, Bentley KL, Murtha MT and Risch N (1994) Gene loss and gain in the evolution of the vertebrates. *Development Supplement*, 155–161.

Satokata I, Benson G and Maas R (1995) Sexually dimorphic sterility phenotypes in Hoxa-10-deficient mice. *Nature*, **374**, 460–463.

Schmid KJ and Tautz D (1997) A screen for fast evolving genes from Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9746–9750.

Shang Z, Isaac VE, Li H, Patel L, Catron KM, Curran T, Montelione GT and Abate C (1994) Design of a "minimAl" homeodomain: the N-terminal arm modulates DNA binding affinity and stabilizes homeodomain structure. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 8373–8377.

Shimeld SM and Holland PW (2000) Vertebrate innovations. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4449–4452.

Steelman S, Moskow JJ, Muzynski K, North C, Druck T, Montgomery JC, Huebner K, Daar IO and Buchberg AM (1997) Identification of a conserved family of Meis1-related homeobox genes. *Genome Research*, **7**, 142–156.

Sutton KA and Wilkinson MF (1997) Rapid evolution of a homeodomain: evidence for positive selection. *Journal of Molecular Evolution*, **45**, 579–588.

Szathmary E, Jordan F and Pal C (2001) Molecular biology and evolution. Can genes explain biological complexity? *Science*, **292**, 1315–1316.

Tautz D (1996) Selector genes, polymorphisms, and evolution. *Science*, **271**, 160–161.

Ting CT, Takahashi A and Wu CI (2001) Incipient speciation by sexual isolation in Drosophila: concurrent evolution at multiple loci. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6709–6713.

Ting CT, Tsaur SC, Wu ML and Wu CI (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science*, **282**, 1501–1504.

Van de Peer Y, Taylor JS, Braasch I and Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *Journal of Molecular Evolution*, **53**, 436–446.

Wagner GP, Amemiya C and Ruddle F (2003) Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14603–14606.

Zakany J, Gerard M, Favier B, Potter SS and Duboule D (1996) Functional equivalence and rescue among group 11 Hox gene products in vertebral patterning. *Developmental Biology*, **176**, 325–328.

Zhang J and Nei M (1996) Evolution of Antennapedia-class homeobox genes. *Genetics*, **142**, 295–303.

Zhao JJ, Lazzarini RA and Pick L (1993) The mouse Hox-1.3 gene is functionally equivalent to the Drosophila Sex combs reduced gene. *Genes and Development*, **7**, 343–354.

Zhao Y and Potter SS (2002) Functional comparison of the Hoxa 4, Hoxa 10, and Hoxa 11 homeoboxes. *Developmental Biology*, **244**, 21–36.

Specialist Review

Geographic structure of human genetic variation: medical and evolutionary implications

Giorgio Bertorelle and Guido Barbujani
University of Ferrara, 44100 Ferrara, Italy

1. Introduction

This article is divided into three sections. In the first section we shall describe what we consider the main features of human geographic structure, that is, the fact that, on the average, two individuals living in the same geographic area are genetically more similar than two individuals living in different areas. Some common patterns observed at the global or local scale will be considered, and their possible origin will be discussed. In the second section, we shall summarize the consequences of the genetic structure currently observed in our species. Particular emphasis will be given to the known biomedical aspects, but the potential evolutionary implications will be also analyzed. Finally, in the third section, we shall discuss the possible future evolution of the human population structure, and how to predict it.

2. Human populations are genetically structured: how much, how, and why

Despite the recent appearance of our species and the ability and tendency to move not only to colonize new empty environments but also to exchange migrants among populations (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1, Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1, Article 2, **Modeling human genetic history**, Volume 1, Article 71, **SNPs and human history**, Volume 4) human groups from different geographic areas are often genetically distinct. The levels and patterns of this differentiation vary widely depending on the population, the genetic marker, and the sexes we consider. In humans, more than in other species, genetic structure is also unstable with time. Since the first colonization of the world by *Homo sapiens*, we probably never had a period of time long enough to establish a stable equilibrium of gene flow between local groups. Cultural and social changes, almost always associated with demographic effects, modified and continue to modify our population-genetic structure. But what is the situation today?

Roughly speaking, about 85% of the human genetic variation analyzed so far at the worldwide scale can be found, on the average, within a single population (Lewontin, 1972; Jobling *et al*., 2004; Barbujani, 2005; *see also* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1). This figure is much higher than observed in several terrestrial mammals even at smaller geographic scales, where values smaller than 50% are not uncommon (e.g., Templeton, 1999). In other words, considering just this figure, population structure in humans appears limited: on the average, only about 15% of the genetic variation in our species occurs between populations, much less than observed, for example, in roe deer populations in Europe (Vernesi *et al*., 2002) or in common chimpanzee groups from different African forests (Kaessmann *et al*., 1999). This overall high level of population homogeneity can be explained by the history and the demography of humans: recent evolution, and high mobility. We all descend from some group of Eastern Africans, probably in the order of magnitude of 10^4 individuals (Takahata *et al*., 1995; Jobling *et al*., 2004) who dispersed and started to evolve in partially independent groups only in the last 60 000 years or so, and from which the current 6 billions of humans inherited their genes. However, the genetic structure of human populations would probably be stronger in a less mobile species. Genetic drift of allele frequencies due to founder effects during colonization or, in general, to small population size was probably the rule for much of our history. Only recurrent gene flow (mainly between neighboring groups) and massive migration processes (e.g., during the Neolithic transition) probably prevented a higher average divergence (at least in terms of allele frequencies) between human groups.

F_{st}, an index of genetic distance, can express differences between populations that measure the fraction of between-population variation over the total genetic variance. An average F_{st} of 15% can be considered low compared to other species, but still substantial since it indicates that human populations are, in fact, genetically differentiated. Only looking at specific populations, cultures, geographic areas, genes or markers, and time frames in more detail, we can try to understand the fine structure of human genetic variation and its implications.

There are many exceptions, but, as a rule, genetic variation tends to be geographically structured following a clinal or an isolation-by-distance (IBD) pattern. This means that the major patterns observed at large geographic scales can be described as a simple increase of F_{st} with geographic distance (Cavalli-Sforza *et al*., 1994; Relethford, 2004; Serre and Pääbo, 2004). The major mass migration processes responsible for the clines are believed to be the Paleolithic out-of-Africa dispersal of anatomically modern humans, and the dispersal of Neolithic farmers from areas of origin of agriculture in all continents but Australia (Bellwood, 2004). On the other hand, local gene flow between neighboring populations can produce the IBD pattern, that is, a decrease of F_{st} with increasing geographic distance in any direction, up to a certain distance where genetic exchange is minimal. Isolation by distance is probably a continuing process in virtually all-human populations. Both IBD and directional dispersal produce continuous genetic change (CGC), which implies that (1) F_{st} values can be much higher than 15% when geographically distant populations are compared, and (2) the definition of a higher hierarchical classification, that is, the identification of distinct groups of genetically similar

populations, is hard or not possible (Cooper *et al*., 2003). This last point, of course, does not mean that natural (sea, mountains, deserts, etc.) and cultural (language, tradition, behavior, etc.) barriers to gene flow do not create some level of genetic discontinuity between specific groups of populations. However, when different sets of markers are used to estimate the number of such groups or clusters, and to correctly reallocate genotypes to the cluster of origin, only broad continental groups (Africa, Eurasia, Oceania, and America) appear consistent across studies (Romualdi *et al*., 2002; Rosenberg *et al*., 2002; Bamshad *et al*., 2003). A recent study (Serre and Pääbo, 2004) even suggests that the sampling bias (populations are usually sampled on the basis of an a priori idea about the groups) could explain better this result than the real presence of continental genetic barriers. Additional samples more evenly distributed in the geographic space are needed, but present data suggest that CGC is the rule, and genetic discontinuities in space are the exception: major genetic grouping, including races (Barbujani, 2005), is not an accurate concept to classify humans. CGC does not mean that Africans and Asian are not genetically different, but it does mean that this difference can be low or high, depending on the geographic distance of the populations considered and on the set of loci analyzed. Therefore, the genetic and evolutionary meaning of these groups is limited.

An additional feature of the human geographic structure is the occurrence of single populations that strongly deviate from the general patterns reflecting the 85/15 apportionment of diversity and the effects of CGC. Geographic and cultural barriers, which strongly reduce reproductive contact with neighboring groups, usually coupled with small population sizes, can result in highly differentiated populations, or genetic isolates. A genetic threshold to identify such specific population histories has not been defined. However, populations which clearly emerge from the 85/15 + CGC patterns when several loci are considered (e.g., in a multivariate analysis), and have therefore pairwise F_{st} values higher than 15% in comparison with most of the other populations, can be considered genetic isolates. Typical barriers implied in the known cases (Cavalli-Sforza *et al*., 1994; Jobling *et al*., 2004) are the sea (e.g., Sardinians, Papua New Guineans), language (e.g., the Basques, who are possibly a Paleolithic non-Indo-European relic), life style (e.g., hunter-gatherer as African Pygmies) sometimes combined with high geographic isolation (e.g., the Lapps). Other groups, initially identified as genetic isolates on the basis of single marker analyses or high incidence of genetically inherited diseases (e.g., Ladin speakers, or Ashkenazy Jews, or Finns), revealed only an average degree of differentiation when investigated at the genomic level. Only multilocus analyses can estimate the real level of genetic isolation, since stochastic errors or gene-specific selection processes may confound the pattern in single-locus studies. For example, Ashkenazy Jews were recently classified together with Norwegians and Armenians when 39 markers were considered (Wilson *et al*., 2001) and Finns, who are characterized by a specific set of disease alleles, do not appear to have a specific genomic make-up (Cavalli-Sforza *et al*., 1994; Jobling *et al*., 2004). Similarly, when genetic isolates are identified, it is important to distinguish between "pure drift" isolates, that is, those recently diverged small populations with reduced variability and almost no specific alleles or mutations, from populations with a long independent accumulation of specific molecular variation. The former can be very

useful in mapping studies, but the latter contain certainly more information about the evolution of our species and our genomes.

The reduced fraction of genetic variation between populations, only about 15% on the average, is thus structured in a CGC pattern with outliers. This general view is mainly based on genetic polymorphisms such as mtDNA sequences, microsatellites, SNPs (*see* Article 71, **SNPs and human history**, Volume 4), Alu insertions, and classical electrophoretic markers, that is, on genome fragments, where selection is commonly believed to be either weak or absent. Most of the genome is constituted by regions of this kind (even though their possible regulatory functions are not completely understood). However, the geographic distribution of positively selected genes can be very different from the 85/15 + CGC + outliers predictions. Local adaptation and widespread (and similar) selective pressures produce, in fact, stronger and weaker geographic structure, respectively, and the genetic effects of selection can be very rapid (Hartl and Clark, 1997; Hendry and Kinnison, 2001). There is evidence, for example, that the mutation associated with the lactase persistence (that allows lactose digestion) in adulthood was selected in the last 10 000 years or so in pastoralist societies where milk drinking was an important part of the diet (Beja-Pereira *et al*., 2003; Coelho *et al*., 2005). This process produced an increase of frequency of the mutations associated with lactase persistence in these groups, and we observe now a higher population divergence (at this and probably at physically linked loci) when their descendents are compared with nonmilk drinkers. On the contrary, we expect that directional selection for the same allele in different populations, or balancing selection maintaining the same set of alleles in different groups, would produce lower-than-average population divergence. Genes showing low geographic structure, possibly because of balancing selection, includes *HLA*, *CCR5*, *PTC* genes (Cavalli-Sforza *et al*., 1994; Bamshad *et al*., 2002; Wooding *et al*., 2004). It is interesting to note that recent methods (Luikart *et al*., 2003), based on earlier ideas (Lewontin and Krakauer, 1973) reverse this approach; F_{st} is estimated for many DNA regions, and values falling in the upper and lower tail of the distribution are considered as suggestive of selective processes.

Finally, the fact that we have separate sexes has also an impact on human geographic structure. In a population with the same $N/2$ number of men and women, there are $2N$ copies of each autosomal DNA fragment, $N/2$ copies of each X-linked fragment, and $N/2$ copies of Y-linked and mtDNA fragments that can be transmitted to the next generation. In other words, random drift, which is negatively correlated with population size, is working at three different speeds in the same populations at these three classes of markers. Consequently, three different levels of population structure should be expected. In addition, the migration rate and the effective population size may differ in men and women. Most human populations are patrilocal, meaning that after marriage the men tend to stay in their birth place more than women, and our species was probably polygynous until recent times (Dupanloup *et al*., 2003), meaning that the number of transmitted Y chromosomes every generation was lower (and drift higher) than expected. Experimental data suggest that all these factors have been important, because the average level of geographic structure (say, F_{st}) in humans increases from autosomal to mtDNA markers (probably due to the different population size of the markers), and from

mtDNA markers to Y-linked markers (presumably due to different male and female behavior at reproduction) (Jobling *et al*., 2004). The relevance of the sex-specific migration patterns in shaping the genetic structure of human populations has been recently confirmed comparing patrilocal and matrilocal tribes of Thailand, where reverse patterns in mtDNA and Y-chromosome markers have been found (Oota *et al*., 2001; Hamilton *et al*., 2005).

In conclusion, the present structure of human populations is characterized not only by some general rules and established patterns, which tell us something about the major processes that shaped genetic variation, but also by several exceptions. The goal for the future is to increase our knowledge and understanding of the fine structure studying isolated groups and selected genes.

3. Population structure in humans: what does this imply?

We discuss now some biomedical and evolutionary implications of the genetic structure described in the previous section.

3.1. Biomedical implications

The geographic patterns at neutral markers, mainly affected by demographic and historical processes, are in part known and understood (see previous section), and can be easily investigated more deeply typing more populations and more markers. But what do we know about the genetic structure of disease alleles? Theoretical predictions and available data can be helpful.

Dominant deleterious mutations are rare, because their expression in all individuals with at least one affected chromosome rapidly drive them to extinction. Simple monogenic diseases with this type of inheritance are therefore rare, with a similar incidence in different populations simply regulated by mutation-selection equilibrium. Similarly, infrequent are the simple genetic diseases whose heterozygous carriers enjoy some selective advantage also over the healthy homozygotes (e.g., sickle cell anemia). In the few clear situations of this type, the population structure at these genes can be decoupled from the structure observed at neutral markers. Environmental differences in the selective pressure (e.g., the geographic distribution of malaria for the sick cell anemia) are, in fact, responsible of the observed population structure. On the other hand, when the fitness effects of mutations are expressed only in the homozygotes, or after the reproductive age, or are, in general, very limited as in the complex multigenic diseases (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2; Article 58, **Concept of complex trait genetics**, Volume 2), the major factors acting on their frequencies are the same that affect neutral markers: demographic and historical processes. For example, the high frequency of several genetic diseases in the Jews is a consequence of their isolation and small effective population size. In other words, the major patterns of population structure described in the previous section are expected to be very similar also in most simple genetic disorders and in the numerous susceptibility alleles involved in complex genetic diseases.

There seem to be six main reasons why population structure should not be ignored in biomedical studies:

1. Priorities in genetic testing for different diseases (*see* Article 69, **Current approaches to prenatal screening and diagnosis**, Volume 2; Article 83, **Carrier screening: a tutorial**, Volume 2) or in treatments should consider the differences among populations. Given certain symptoms, the most likely disease depends also on the population affiliation of the affected individual. If data on the disease incidence are not known, genetically similar population at neutral markers can be used as proxy. If a population is a genetic isolate, specific disease risk alleles are expected.
2. Isolated and recently founded populations (e.g., Finns, possibly Sardinians or Icelanders) should be preferred in mapping studies by linkage disequilibrium (LD) in unrelated individuals and for studies of multigenic diseases in general. These populations are, in fact, comparable to large families, where genetic heterogeneity is rare (and so affected individuals carry the same mutation) and recombination did not have the time to disrupt the statistical association between the hunted gene and the flanking markers. The possibility to find such situations is clearly related to the fact that human populations are, at least to a certain degree, geographically structured, and in particular, genetic isolates exist (see Section (2)).
3. The common disease – common variant hypothesis (*see* Article 59, **The common disease common variant concept**, Volume 2) suggests that only few susceptibility alleles, at high frequencies in different populations, are responsible for common diseases (Chakravarti, 2001; Reich and Lander, 2001). However, the empirical evidence for this hypothesis is scant (Jobling *et al*., 2004). Several causal alleles are probably rare and population specific (Kittles and Weiss, 2003; Tishkoff and Kidd, 2004), and can also be embedded in a block structure (the partition of the genome in high LD regions, see for example Daly *et al*., 2001; *see* Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4) specific of a population (Verhoeven and Simonsen, 2004). Both these factors indicated that LD mapping should not be performed in pooled samples of individuals of different origins, and we expect that different studies in single populations should be able to better identify different loci with higher power.
4. Spurious disease-marker associations are expected if sampling is not stratified by populations. If the frequency of a trait (the disease) is higher in a population, which is also genetically distinct at some neutral markers, and the population is jointly analyzed with the others (either because sample sizes are limited, or because the two populations coexist in the same area), the trait and the markers will result statistically, but not genetically, associated. For instance, a wide-genome scan carried out in the United States for the genetic determinant of β-thalassemia, a well-understood monogenic disease typical of Mediterranean Europe, might lead to the conclusion that many genes cause the disease. This is because the affected children (most of them from Greek and Italians families) might differ at several genome regions from the controls (representing the general US population).

5. The presence of genetically different groups implies that admixture, if occurs, can produce several consequences. This process was not uncommon during the history and prehistory of our species, for example, between hunter-gatherer residents and immigrating farmers in the Neolithic, or more recently between Native Americans, Africans, and Europeans in the Americas. One practical consequence is that mapping by admixture linkage disequilibrium (MALD; *see* Article 76, **Mapping by admixture linkage disequilibrium (MALD)**, Volume 4) is possible. This recently resurrected approach (McKeigue, 2005) exploits the transitory LD arising when distinct populations come together (of course trying to avoid the initially spurious correlations). From the point of view of the admixed population and individuals, genetic admixture produces a beneficial initial effect of heterosis (population-specific recessive deleterious alleles are not expressed in heterozygotes), and in general a decrease of the fraction of affected individuals at recessive disease alleles (since different allele frequencies result in less than average frequencies of homozygotes). On the other hand, outbreeding depression effects, although never demonstrated in humans, are possible in admixed populations. The frequency reduction of alleles positively and differentially selected in different environments (e.g., at the *MC1R* gene involved in pigmentation, or at the HLA system involved in immune response) may produce an increase of disease incidence in the admixed group. For example, skin cancer or vitamin D deficiency problems can increase when Caucasians and Africans mix, or infectious diseases can spread when populations adapted to different pathogen community come together. Also, coadapted combination of alleles at polygenic disease loci might be disrupted by admixture.
6. Exogenous molecules such as drugs should be absorbed and transported before acting, and metabolized and excreted after that. Several proteins are implicated during these phases, and their gene polymorphisms are related to differential efficiency (e.g., Meyer, 2004). On of the best-known examples is the gene *CYP2D6*, which encodes an enzyme that metabolize about 20% of the drugs in the market. Different genetic variants (that includes also variation in copy number) have different efficiency, where high and low efficiency result in reduced or toxic effects of the drug, respectively. The development of individual forms of pharmacological treatment based on the individuals' genotype is clearly the long-term goal of applied pharmacogenomic studies, but it will not be a possibility in the immediate future. In the short term, it is unclear whether priority in drug treatments could be based on population affiliation, because different studies yielded very different results as for the degree of geographical structuring among populations for drug-metabolizing genes (Bradford, 2002; Shimizu *et al*., 2003).

3.2. *Evolutionary implications*

Adaptive genetic change is the ultimate consequence of environmental pressures affecting the individuals' fitness, but often organisms respond to environmental change by physiological or cultural adaptation, the latter being very important in

primates, and especially for humans. Cultural adaptation can allow reproduction of individuals who otherwise would have reduced or zero fitness, and can drive to accumulation of deleterious or maladaptive mutations. However, (1) fitness is still reduced in individuals affected by several inheritable disorders, or in individuals with maladaptive allelic combinations at important genes such as HLA, especially in underdeveloped countries; (2) only few genes responsible for local adaptations are known, but we expect that more, still under selection in some human groups, will be discovered in the future; (3) human phenotypes selected in mating, cultural and social contexts, for example "attractiveness" or "trust", or "cooperation", or "intellectual abilities", or "novelty seeking", may have a genetic component (see e.g., Ding *et al*., 2002; Evans *et al*., 2004; Kosfeld *et al*., 2005), and are probably still under selection; (4) evolution is not only natural selection, culture might reduce selective pressures, but might increase other evolutionary processes; for example, linguistic barriers decrease the level of gene flow thus favoring the genetic divergence between populations. In addition, massive migration is bringing new alleles into new areas, and causing either admixture or the onset of local reproductive barriers, depending on the circumstances.

It is clear, therefore, that the genetic composition of our species is still subjected to evolutionary change, and it makes sense to ask what are the predictable evolutionary consequences of the current geographic structure.

High genetic variation is regarded as a positive asset of a species, both because it is associated with low inbreeding depression effect, and because the evolutionary potential is maintained. Does the global level of genetic variation in our species depend on the fact that we are not a single panmictic group? Surprisingly, population genetics theory has no unequivocal answer to this apparently simple question. Different results are obtained with different assumptions on the population structure model and the demographic parameters, and using different metrics to quantify genetic variation (Slatkin, 1987; Strobeck, 1987; Whitlock and Barton, 1997; Wakeley and Aliacar, 2001). Nobody can tell to what extent human populations were subjected to processes of local extinction and recolonization (that are expected to reduce genetic variation) or alternatively represented rather stable demographic units (whose fragmentation is expected to enhance genetic variation). However, if the extinction of differentiated groups is not frequent, as expected in a recent and expanding species (Foley, 1998), geographic structuring should result in the existence of relatively high genetic divergence. If this prediction is appropriate for humans, the overall geographic structure, limited but significant, positively affected the level of genetic variation in our species. Partly different gene pools would have evolved in diverging groups, and only rarely would such gene pools disappear by population extinction. The expected end result of a process of this kind is a global variation larger than in a panmictic population of the same size.

Be that as it may, the current geographic patterning of genetic variation suggests at least two additional considerations: (1) the widespread presence of clines (CGC), and the consequent absence of major genetic barriers, indicate that a process of speciation, clearly unlikely today, was probably not a possibility also during the whole history of our species; and (2) the presence of highly differentiated groups, either outliers in clinal pattern or populations at the extremes of the

clines, resulted in the localized occurrence of population-specific alleles. As a consequence, recent secondary contacts might result in the increase of infectious diseases, through a genetic out breeding depression effect (that can affect of course disease susceptibility as well as many other locally adapted traits) or in the diffusion of pathogens in immunologically naïve populations. However, similar to what happened in local breeds of domestic animals or local varieties of plants, differentiated groups might have preserved specific and unique portions of adaptive genetic variation, which might be useful not only in evolutionary but also medical terms. An example of that is the small community of Limone sul Garda where most individuals are genetically protected against heart diseases (Bielicki and Oda, 2002).

Finally, it is interesting to note that genetic and cultural factors can reinforce each other. Linguistic and genetic diversity, for example, appear broadly correlated over much of the planet (Cavalli-Sforza *et al*., 1988), but this is not due to the existence of genetic factors predisposing people to speak certain languages. On the contrary, there is evidence that language barriers, much like geographic barriers, reduce gene flow, so that the existence of cultural differences results in an increased genetic divergence (Barbujani and Sokal, 1990). In a sense, cultural differences (language differences being just the simplest traits to analyze) would create an opportunity for sexual selection to operate (Harpending and Rogers, 2000). Thus, in various contexts, cultural difference might additionally increase because contacts are limited by differences in traits related with mate choice. If this is true, we expect that the enormous importance of culture in our specie could have resulted in some populations in a species-specific runaway process of cultural-genetic divergence favoring extreme cultural, and genetic, characteristics.

4. What is the future of human geographic structure?

Modern humans dispersed from Africa into Europe, Asia, and Australia (probably between 60 and 40 Ky ago), colonized the Americas (between 15 and 10 ky ago), and more recently (<3.5 ky) reached remote Oceania. All these events, with few exceptions, put the bases for the establishment of the geographic structure, in populations descended from small numbers of founders and hence highly subjected to random drift.

This process of genetic divergence was counterbalanced by the continuous local gene flow and by major dispersal processes, among which those associated to the agricultural transition (Bellwood, 2004). In different areas of Eurasia and of the Americas, technologies for food production were developed. The resulting increase in the farmers' population size caused dispersal, which favored genetic mixing with previously, settled hunter-gatherer, and the reduction of drift effects. Of course, this process could have also increased the genetic divergence between human groups adopting different subsistence strategies, limiting the levels of gene flow among them. This might have contributed to the genetic isolation of some populations (e.g., Andaman Islanders, Pigmy), but the major effect of farming at global scale was probably the increase of genetic homogeneity (Excoffier and Schneider, 1999).

In very recent times, the level of population structure probably declined even more, because of the continuous but differential increase of the census size (that

produced emigration from countries like Europe in the past or Africa now), the development of transportation means, and more in general the socioeconomical changes. Only in the last few centuries, millions of people from Africa, Europe, and America, relocated in different geographic regions the genomes inherited from their parents, a certain level of mixing with the local genetic pools occurred. In addition, contacts between populations with locally adapted variation for pathogen resistance (such as, e.g., HLA variants) might have produced global selective sweeps, decreasing the population structure also at these genes. Future advances in the biochemical methods to characterize the genes of past populations (e.g., Vernesi *et al*., 2004), and in the statistical methods necessary to analyze complex demographic scenarios (e.g., Beaumont *et al*., 2002; Marjoram *et al*., 2003) will definitely help us to better understand past evolutionary processes leading to the current structuring of the human populations.

If this reconstruction of the evolution of the human population structure is at least approximately accurate, should we simply predict that the "defragmentation" phase that followed the initial "structuring" process would continue? It is hard to imagine what will happen in the future, but a few educated guesses are possible. One is, large numbers of individuals are currently migrating to various areas of the world. However, it is not necessary to assume that large genetic changes will follow. Indeed, most documented historical migration processes that are considered to have contributed to homogenizing genetically human populations, occurred at low population densities. That is the case for all the main dispersal processes leading to the farming expansions, as well as for expansions occurring in the last few centuries. However, at the high densities, typical of modern urban societies, very high numbers of immigrants are necessary to produce significant changes in allele frequencies. In addition, the local effect on population structure will depend not only on the rates of gene flow, but also on the tendency of people to admix, once they are settled in a new region. Examples from the past range from extensive admixture in countries such as Brazil (see e.g., Parra *et al*., 2003) to long-term coexistence of reproductively isolated communities in countries such as India (see e.g., Bamshad *et al*., 2001) or the US (see e.g., Shriver *et al*., 1997), with many intermediate situations. In some cases, a high level of structuring within metropolitan areas is to be expected, with ethnic or linguistic communities living within a few hundred meters of distance showing levels of genetic differentiation comparable to those between subcontinents. Cultural barriers, and not mountain chains or seas, might thus lead to the evolution of a secondary phase of structuring in humans.

Related articles

Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1; Article 2, **Modeling human genetic history**, Volume 1; Article 6, **The genetic structure of human pathogens**, Volume 1; Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1; Article 7, **Genetic signatures of natural selection**, Volume 1; Article 5, **Studies of human genetic**

history using mtDNA variation, Volume 1; Article 10, **Measuring variation in natural populations: a primer**, Volume 1; Article 59, **The common disease common variant concept**, Volume 2; Article 58, **Concept of complex trait genetics**, Volume 2; Article 32, **Comparisons with primate genomes**, Volume 3; Article 68, **Normal DNA sequence variations in humans**, Volume 4; Article 71, **SNPs and human history**, Volume 4; Article 72, **Evolutionary modeling in haplotype analysis**, Volume 4

References

Bamshad MJ, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A *et al*. (2001) Genetic evidence on the origins of Indian caste populations. *Genome Research*, **11**, 994–1004.

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB *et al*. (2002) A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. *Proceedings of the National Academy of Sciences, USA*, **99**, 10539–10544.

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA and Jorde LB (2003) Human population genetic structure and inference of group membership. *American Journal of Human Genetics*, **72**, 578–589.

Barbujani G (2005) Human races: classifying people vs understanding diversity. *Current Genomics*, **4**, 215–226.

Barbujani G and Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences USA*, **87**, 1816–1819.

Beaumont MA, Zhang W and Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N *et al*. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics*, **35**, 311–313.

Bellwood P (2004) *First Farmers*, Blackwell: Oxford.

Bielicki JK and Oda MN (2002) Apolipoprotein A-I Milano and apolipoprotein A-I Paris exhibit an antioxidant activity distinct from that of wild-type apolipoprotein A-I. *Biochemistry*, **41**, 2089–2096.

Bradford LD (2002) CYP2D6 allele frequency in European Caucasians, Asian, Africans and their descendants. *Pharmacogenomics*, **3**, 229–243.

Cavalli-Sforza LL, Piazza A and Menozzi P (1994) *The History and Geography of Human Genes*, Princeton University Press: Princeton, NJ.

Cavalli-Sforza LL, Piazza A, Menozzi P and Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences USA*, **85**, 6002–6006.

Chakravarti A (2001) Single nucleotide polymorphisms ... to a future of genetic medicine. *Nature*, **409**, 822–823.

Coelho M, Luiselli D, Bertorelle G, Lopes IA, Seixas S, Destro-Bisol G and Rocha J (2005) Microsatellite variation and evolution of human lactase persistence. *Human Genetics*, **117**, 329–339.

Cooper RS, Kaufman JS and Ward R (2003) Race and genomics. *New England Journal of Medicine*, **348**, 1166–1170.

Daly MJ, Rioux JD, Schaffner SE, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Ding Y, Chi H-C and Grady Dl (2002) Evidence of positive selection acting at the human dopamine receptor D4 gene locus. *Proceedings of the National Academy of Sciences, USA*, **99**, 309–314.

Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, Amorim A and Barbujani G (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *Journal of Molecular Evolution*, **57**, 85–97.
Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, Dorus S and Lahn BT (2004) Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. *Human Molecular Genetics*, **13**, 489–494.
Excoffier L and Schneider S (1999) Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proceedings of the National Academy of Sciences, USA*, **96**, 10597–10602.
Foley R (1998) The context of human genetic evolution. *Genome Research*, **8**, 339–347.
Hamilton G, Stoneking M and Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences, USA*, **102**, 7476–7480.
Harpending H and Rogers A (2000) Genetic perspective on human origin and differentiation. *Annual Review of Genomics and Human Genetics*, **1**, 361–385.
Hartl DL and Clark AG (1997) *Principles of Population Genetics*, Sinauer: Sunderland, Massachusetts.
Hendry AP and Kinnison MT (2001) An introduction to microevolution: rate, pattern, process. *Genetica*, **112-113**, 1–8.
Jobling MA, Hurles ME and Tyler-Smith C (2004) *Human Evolutionary Genetics*, Garland Science: New York.
Kaessmann H, Wiebe V and Paabo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286**, 1159–1162.
Kittles RA and Weiss KM (2003) Race, ancestry and genes: implications for defining disease risk. *Annual Review of Genomics and Human Genetics*, **4**, 33–67.
Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U and Fehr E (2005) Oxytocin increases trust in humans. *Nature*, **435**, 673–676.
Lewontin RC (1972) The apportionment of human diversity. *Evolutionary Biology*, **6**, 381–398.
Lewontin RC and Krakauer JK (1973) Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.
Luikart G, England PR, Tallmon D, Jordan S and Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
Marjoram P, Molitor J, Plagnol V and Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences USA*, **100**, 15324–15328.
McKeigue PM (2005) Prospects for admixture mapping of complex traits. *American Journal of Human Genetics*, **76**, 1–7.
Meyer U (2004) Pharmacogenetics. Five decades of therapeutic lessons from genetic diversity. *Nature Reviews Genetics*, **5**, 669–676.
Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T and Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, **29**, 20–21.
Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM and Pena SDJ (2003) Color and genomic ancestry in Brazilians. *Proceedings of the National Academy of Sciences, USA*, **100**, 177–182.
Reich DE and Lander ES (2001) On the allelic spectrum of human disease. *Trends in Genetics*, **17**, 502–510.
Relethford JH (2004) Global patterns of isolation by distance based on genetic and morphological data. *Human Biology*, **76**, 499–513.
Romualdi C, Balding D, Nasidzr IS, Risch G, Robichaux M, Sherry ST, Stoneking M and Batzer MA (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Research*, **12**, 602–612.
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW 2002 Genetic structure of human populations. *Science* **298**, 2381–2385.
Serre D and Pääbo S (2004) Evidence for gradients of human genetic diversity, within and among continents. *Genome Research*, **14**, 1679–1685.

Shimizu T, Ochiai H, Asell F, Yokono Y, Kikuchi Y, Nitta M, Hama Y, Yamaguchi S, Hashimoto M, Taki K, *et al*. (2003) Bioinformatics research on inter-racial differences in drug metabolism. I. Analysis on frequencies of mutant alleles and poor metabolizers on CYP2D6 and CYP2C19. *Drug Metabolism and Pharmacokinetics*, **18**, 48–70.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R and Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, **60**, 957–964.

Slatkin M (1987) The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology*, **32**, 42–49.

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, **117**, 149–153.

Takahata N, Satta Y and Klein J (1995) Divergence time and population size in the lineage leading to modern humans. *Theoretical Population Biology*, **48**, 198–221.

Tishkoff SA and Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics Supplement*, **11**, S21–S27.

Templeton AR (1999) Human races: a genetic and evolutionary perspective. *American Journal of Anthropology*, **100**, 632–650.

Verhoeven KJF and Simonsen KL (2004) Genomic haplotype blocks may not accurately reflect spatial variation in historic recombination intensity. *Molecular Biology and Evolution*, **22**, 735–740.

Vernesi C, Caramelli D, Dupanloup I, Bertorelle G, Lari M, Cappellini E, Moggi-Cecchi J, Chiarelli B, Castri L, Casoli A *et al*. (2004) The Etruscans: a population-genetic study. *American Journal of Human Genetics*, **74**, 694–704.

Vernesi C, Pecchioli E, Caramelli D, Tiedemann R, Randi E and Bertorelle G (2002) The genetic structure of natural and reintroduced roe deer (*Capreolus capreolus*) populations in the Alps and Central Italy, with reference to the mitochondrial DNA phylogeography of Europe. *Molecular Ecology*, **11**, 1285–1297.

Wakeley J and Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.

Whitlock MC and Barton NH (1997) The effective size of a subdivided population. *Genetics*, **146**, 427–441.

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N and Goldstein DB (2001) Population genetic structure of variable drug response. *Nature Genetics*, **29**, 265–269.

Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB and Drayna D (2004) Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. *American Journal of Human Genetics*, **74**, 637–646.

Short Specialist Review

Studies of human genetic history using the Y chromosome

Mark A. Jobling
University of Leicester, Leicester, UK

The human Y chromosome's raison d'être is sex (in particular, male sex determination), and this louche connection has made it perhaps the strangest segment of our genome. It is specific to males and constitutively haploid in a diploid organism, and therefore escapes from recombination for most of its length, apart from the two pseudoautosomal regions in which crossing-over with the X chromosome occurs. This absence of intergenerational reshuffling means that it passes from father to son as an immense haplotype, changing only by mutation, and so contains within it a relatively straightforward record of its past. Over the last few years, many useful DNA polymorphisms (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4) have been discovered in the 23 Mb of nonrecombining euchromatin, and this has led to an explosion in studies of human genetic history using the Y chromosome.

A set of over 200 well-characterized binary polymorphisms representing unique or rare events in human evolution, mostly SNPs (single nucleotide polymorphisms), has been used to define Y-chromosomal haplotypes, known as *haplogroups*. These can be arranged into a unique phylogeny (Figure 1a) rooted by comparison to chimpanzee sequences (Underhill *et al*., 2001; Y Chromosome Consortium, 2002).

The Y chromosome also carries faster mutating markers, in particular microsatellites. These can be used to examine diversity within haplogroups, allowing time to most recent common ancestor (TMRCA) to be estimated, given an estimate of mutation rate (typically ∼0.2% per microsatellite per generation). Intrahaplogroup diversity can be compared between populations, allowing deductions about the geographical origins of population expansions. Typically, most studies analyze a heterogeneous set of 6–20 microsatellites, but recently (Kayser *et al*., 2004), the entire chromosome has been searched for new markers, resulting in the isolation of 172 novel microsatellites to add to the known set of ∼50. This resource should allow new accuracy in the estimation of TMRCAs, and the choice of sets of markers with relatively uniform and predictable mutational properties.

How are Y haplogroups distributed in different populations? The deepest-rooting clades within the Y phylogeny (haplogroups A and B) are almost entirely confined to sub-Saharan Africa (Figure 1b), and one estimate for TMRCA based on SNPs and a model of exponential population growth (Thomson *et al*., 2000) is 59 thousand

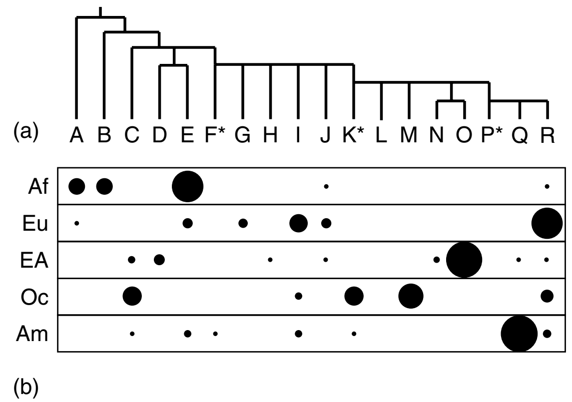


Figure 1 Frequencies of the major Y haplogroups in five continental populations. (a) The Y phylogeny, showing the 18 major haplogroups, A to R. (b) Relative haplogroup frequencies (represented by area of filled circles) in indigenous populations of sub-Saharan Africa (Af), Europe (Eu), East Asia (EA), Oceania (Oc), and the Americas (Am). Data from Underhill *et al*. (2001)

years ago (KYA), with 95% confidence interval limits of 40–140 KYA. These observations are compatible with a recent out-of-Africa origin for modern humans.

The distribution of Y haplogroups shows a generally high degree of geographical differentiation, with, in one study (Seielstad *et al*., 1998), 65% of genetic variance existing between populations, and 35% within. This contrasts starkly with an autosomal figure of 15% between- and 85% within-population variance (Barbujani *et al*., 1997). The difference has been ascribed (Seielstad *et al*., 1998) to the predominant practice of patrilocality, in which women tend to move to the husband's birthplace upon marriage, although this effect may only be appreciable at the local rather than global level (Wilder *et al*., 2004). Another contributory factor is genetic drift: stochastic variation in haplotype frequency from one generation to the next due to variance in offspring number, which is particularly marked on the Y because its effective population size is one-quarter of that of any autosome.

High geographical differentiation makes the Y a powerful system for the study of past population movements, including colonization and admixture. In the case of the Lemba, a Bantu-speaking South African tribe, oral history telling of descent from Jews who came from the north by boat finds support in the fact that they carry a high frequency of a Y-chromosomal haplotype that is typical of Jewish Middle Eastern populations (Thomas *et al*., 2000). Combining Y studies with analysis of maternally inherited mitochondrial DNA (mtDNA; *see* Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1) has revealed evidence for sex-biased admixture, in which men, but not women, from one population have contributed genes to another population. Examples are seen in Greenland, where all mtDNAs analyzed are of Native American origin, while up to 64% of Y chromosomes are European (Bosch *et al*., 2003); and in Brazil, where European Y chromosomes are close to 100% (Carvalho-Silva *et al*., 2001) but mtDNAs originate approximately equally from Europeans, Africans (contributed by imported slaves), and indigenous Native Americans.

High-resolution Y haplotype analysis has revealed the extraordinary extent to which social structures can allow a single individual to contribute disproportionately to future populations (Zerjal *et al*., 2003). About 8% of the Y chromosomes from a large region of CentralAsia (~0.5% of the global total) belong to a very closely related lineage cluster with an estimated TMRCA of only ~1000 years, spread across 16 different populations from the Pacific to the Caspian Sea. The age of the cluster, its distribution, and its probable origin in Mongolia are all consistent with Genghis Khan and his dynasty being responsible for its unprecedented spread.

The discussion above assumes that natural selection is not influencing Y haplotype distributions. Under positive selection, a beneficial mutation could have led to the spread or fixation of a particular Y haplotype in the past. There is no convincing evidence that such a "selective sweep" has occurred, and apparent evidence from diversity data could always be explained as the effect of population expansion (*see* Article 7, **Genetic signatures of natural selection**, Volume 1). Deleterious mutations can occur in essential genes, including those involved in sperm production, but purifying selection acts continually to weed these out; also, the multicopy nature of many spermatogenesis genes may allow correction of mutations through intrachromosomal recombination (gene conversion) with unmutated copies (Rozen *et al*., 2003). Some studies of selection have focused on particular deleterious phenotypes known (or suspected) to be associated with the Y chromosome. Significant effects of Y haplotype on the probability of, for example, having low sperm count (Krausz *et al*., 2001) or prostate cancer (Paracchini *et al*., 2003) have been found, but are likely to have only minor effects on haplotype frequencies in populations. With current knowledge, it therefore seems reasonable to treat haplotype frequencies as an outcome of past population processes. However, advocates of the Y chromosome must not forget that, whatever its allure, it represents but a single "run" of the genealogical process, and that combination with other loci in the genome is necessary to provide a balanced view of the human past from DNA evidence.

Related articles

Article 2, **Modeling human genetic history**, Volume 1; Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1; Article 71, **SNPs and human history**, Volume 4

Acknowledgments

I thank Matt Hurles for assistance with the figure, and the Wellcome Trust for support under a Senior Fellowship in Basic Biomedical Science.

Further reading

Jobling MA, Hurles ME and Tyler-Smith C (2003) *Human Evolutionary Genetics: origins, peoples and disease*, Garland Science: New York/Abingdon.

Jobling MA and Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews. Genetics*, **4**, 598–612.

References

Barbujani G, Magagni A, Minch E and Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 4516–4519.

Bosch E, Calafell F, Rosser ZH, Nørby S, Lynnerup N, Hurles ME and Jobling MA (2003) High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. *Human Genetics*, **112**, 353–363.

Carvalho-Silva DR, Santos FR, Rocha J and Pena SDJ (2001) The phylogeography of Brazilian Y-chromosome lineages. *American Journal of Human Genetics*, **68**, 281–286.

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA, *et al*. (2004) A comprehensive survey of human Y-chromosomal microsatellites. *American Journal of Human Genetics*, **74**, 1183–1197.

Krausz C, Quintana-Murci L, Rajpert-De Meyts E, Jorgensen N, Jobling MA, Rosser ZH, Skakkebaek NE and McElreavey K (2001) Identification of a Y chromosome haplogroup associated with reduced sperm counts. *Human Molecular Genetics*, **10**, 1873–1877.

Paracchini S, Pearce CL, Kolonel LN, Altshuler D, Henderson BE and Tyler-Smith C (2003) A Y chromosomal influence on prostate cancer risk: the multi-ethnic cohort study. *Journal of Medical Genetics*, **40**, 815–819.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK and Page DC (2003) Abundant gene conversion between arms of massive palindromes in human and ape Y chromosomes. *Nature*, **423**, 873–876.

Seielstad MT, Minch E and Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, **20**, 278–280.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N and Goldstein DB (2000) Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba - the "Black Jews of Southern Africa". *American Journal of Human Genetics*, **66**, 674–686.

Thomson R, Pritchard JK, Shen P, Oefner PJ and Feldman MW (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 7360–7365.

Underhill PA, Passarino G, Lin AA, Shen P, Lahr MM, Foley RA, Oefner PJ and Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics*, **65**, 43–62.

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM and Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genetics*, **36**, 1122–1125.

Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research*, **12**, 339–348.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, *et al*. (2003) The genetic legacy of the Mongols. *American Journal of Human Genetics*, **72**, 717–721.

Short Specialist Review

Studies of human genetic history using mtDNA variation

Antonio Torroni and Chiara Rengo
University of Pavia, Pavia, Italy

At fertilization, the genetic contribution of the oocyte to the zygote differs from that of the spermatozoon since the latter does not contribute viable mitochondria. These cytoplasmic organelles harbor numerous copies of a circular genome (~16 570 bp), which is characterized by a much higher evolution rate (10–20 folds) than that of the average nuclear gene. Thus, human mitochondrial DNA (mtDNA) is maternally transmitted and its sequence variation, which cannot be reshuffled by recombination as in autosomal genes, has been generated exclusively by the sequential accumulation of new mutations along radiating maternal lineages. Because this process of molecular differentiation is relatively fast and occurred mainly during and after the recent process of human colonization and diffusion into different regions and continents, the different subsets of mtDNA variation tend to be restricted to different geographic areas and population groups.

As a result of these peculiar features, the human mtDNA has a single genealogical history, which can be reconstructed as a gene tree (or network) (Bandelt *et al.*, 1995); migrations (instances of gene flow between regions) can be detected by incorporating the geographical origin of subjects into the tree; and time depth within the tree can be estimated using a molecular clock (Richards and Macaulay, 2001). The application of these principles, now referred to as the *phylogeographic approach* (Avise, 2000) was pioneered in humans by Douglas C. Wallace in the early 1980s.

Before providing some examples of the results obtained in the last 20 years by mtDNA studies, it should be pointed out that, despite its unique features for studying human genetic history, mtDNA remains a *single* locus – prone to genetic drift and possibly under selection – thus, it should not be used as the only tool for inference in human evolution. Conclusions based on mtDNA data require corroboration by other genetic systems (Y chromosome (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1) and autosomal DNA) and other disciplines (for instance, linguistics, archaeology, and even climatology).

The earliest mtDNA work began by digesting the entire genome with a few restriction enzymes (often five or six) on fairly large sample sets (Johnson *et al.*, 1983). However, this approach attracted public attention only in 1987, when

Rebecca Cann, Mark Stoneking, and Allan Wilson using high-resolution restriction analysis (12 enzymes) obtained a much more detailed mtDNA phylogeny. Their analysis of 147 mtDNAs from the different continents identified what was improperly defined as "the mitochondrial Eve" and led to the hypothesis that all modern mtDNAs descend from a woman who lived in Africa about 200 000 years ago (Cann *et al*., 1987). This proposal generated a fierce debate between the proponents of a recent African origin of anatomically modern humans, and those who favored multiregional alternatives. Although the African root was also supported by the sequencing data from the fast-evolving first and second hypervariable segments (HVS-I and HVS-II) of the mtDNA control region (Vigilant *et al*., 1991), the debate lasted for almost a decade, and was for the most part resolved only when the control region of the first Neanderthal specimen was finally PCR amplified and sequenced (Krings *et al*., 1997), revealing that its sequence fell outside the variation of modern humans and did not contribute to the current mtDNA pool.

During the early 1990s, in parallel with the studies addressing the origin of *Homo sapiens sapiens*, the high-resolution restriction analysis also began to be applied to individual continents on large numbers of samples, with the objective of determining human origins in each major geographic area. This resulted in a much more refined picture of the mtDNA world phylogeny – one in which the haplotype clades, or haplogroups, characterized by one or several diagnostic restriction markers, were usually restricted geographically, some to sub-Saharan Africans, others to Europeans, and yet others to East Asians. Many of these haplogroups could not be distinguished by control-region data alone (despite the popularity of this approach in the 1990s and its wide application in forensics), although they could often be picked out from control-region data after an exploratory combined control-region/restriction analysis (Torroni *et al*., 1993, 1996). The haplogroups could, moreover, be subdivided into smaller evolutionary units by including control-region information (Macaulay *et al*., 1999).

The first large comprehensive population studies were carried out in Native Americans and focused on the origin, timing, and numbers of ancestral migrations from Asia. These revealed that virtually all Native American mtDNAs belonged to haplogroups A, B, C, and D (Torroni *et al*., 1993) – later joined by the uncommon haplogroup X – and that only one or two founder haplotypes for each haplogroup were shared between the Native Americans and their Asian counterparts. This indicated that a limited number of mtDNAs arrived in the Americas, in one or two population expansions from Siberia/Beringia. A conclusion that appears to be entirely supported by the data recently obtained from the analysis of entire mtDNA sequences (Bandelt *et al*., 2003).

In Europe, mtDNA variation was studied for the first time by a number of groups in the early 1990s, mostly focusing on the HVS-I region. Initially, it seemed that the European mtDNA landscape might be so flat as to be almost entirely uninformative with respect to European prehistory suggesting that mtDNA may not be a useful demographic marker system (perhaps due to selection or high rates of female gene flow in recent times). However, high-resolution restriction analysis studies showed that this was not true. Indeed, by supplementing HVS-I data with additional informative variants from the coding region, mtDNA variation

was dissected into these major haplogroups (H, I, J, K, T, U, V, W, and X) (Torroni *et al*., 1996; Macaulay *et al*., 1999), which are now fully supported by the sequence analysis of entire mtDNAs (Finnila *et al*., 2001). The first results from European mtDNA (Richards *et al*., 1998) suggested that only a small minority of lineages dated to the Neolithic, with the remainder dating back to between 15 000 and about 50 000 years ago. The majority appeared to descend from founders of Middle or Late Upper Paleolithic origin. These clades were strikingly starlike, indicating dramatic population expansions, which suggested that they were mainly the result of reexpansions in the Late Glacial or Postglacial period. The results were, however, rather tentative, because they were reliant on comparisons with a very small and inadequate sample from the Near East. However, further work (Richards *et al*., 2000, 2002) showed that, with a sufficiently large sample size and a better resolved phylogeny, clades of mtDNA do indeed exhibit gradients similar to those of other marker systems (Cavalli-Sforza *et al*., 1994), and provided evidence that more than three-fourths of the present-day European mtDNAs could be from indigenous Paleolithic ancestors. This conclusion is supported by some analyses of the paternally transmitted Y chromosome (Semino *et al*., 2000; Rootsi *et al*., 2004), but overall rejected by other studies that infer admixture coefficients considering different potential parental populations (Dupanloup *et al*., 2004). In this debated context, first the molecular dissection of one autochthonous European mtDNA clade (haplogroup V), and much more recently that of haplogroup H – the most common haplogroup in Europe (40–50%) – were particularly informative in revealing that there was indeed a dramatic Late Glacial expansion from the Franco-Cantabrian glacial refuge that repopulated much of the western part of the continent of Paleolithic mtDNAs from about 11 000–15 000 years ago (Achilli *et al*., 2004). To determine whether the refuge area(s) of Eastern Europe played a similar role on the other side of Europe would require, in phylogeographic studies, the identification of similarly informative mtDNA subhaplogroups, and, in admixture studies, the evaluation of Eastern Europe as a potential homeland for a parental population.

After having passed through a number of technological and methodological stages, the analysis of mtDNA variation is now in the era of complete mtDNA sequences, and this has opened new, interesting perspectives. This procedure is allowing the identification of new subhaplogroup markers that can be very informative also at the microgeographical level. In addition, the phylogeny of complete sequences has shown that some internal clades are disproportionately derived, compared with others – a result not consistent with a simple model of neutral evolution with a uniform molecular clock – hinting at a role for selection in the evolution of human mtDNA (Torroni *et al*., 2001). This notion has been further developed by Ruiz-Pesini *et al*. (2004) who proposed that the relative frequency and amino acid conservation of internal branch replacement mutations has increased from tropical Africa to temperate Europe and Siberia, and the same haplogroups correlate with increased propensity for energy deficiency diseases as well as longevity. Thus, specific mtDNA replacement mutations permitted our

ancestors to adapt to more northern climates, and these same variants influence our health today.

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, *et al*. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *American Journal of Human Genetics*, **75**, 910–918.

Avise JC (2000) *Phylogeography*, Harvard University Press: Cambridge.

Bandelt H-J, Forster P, Sykes BC and Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–753.

Bandelt H-J, Herrnstadt C, Yao Y-G, Kong Q-P, Kivisild T, Rengo C, Scozzari R, Richards M, Villems R, Macaulay V, *et al*. (2003) Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. *Annals of Human Genetics*, **67**, 512–524.

Cann RL, Stoneking M and Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Cavalli-Sforza LL, Menozzi P and Piazza A (1994) *The History and Geography of Human Genes*, Princeton University Press: Princeton.

Dupanloup I, Bertorelle G, Chikhi L and Barbujani G (2004) Estimating the impact of prehistoric admixture on the genome of Europeans. *Molecular Biology and Evolution*, **21**, 1361–1372.

Finnila S, Lehtonen MS and Majamaa K (2001) Phylogenetic network for European mtDNA. *American Journal of Human Genetics*, **68**, 1475–1484.

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC and Cavalli-Sforza LL (1983) Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *Journal of Molecular Evolution*, **19**, 255–271.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M and Paabo S (1997) Neandertal DNA sequences and the origin of modern humans. *Cell*, **90**, 19–30.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B and Torroni A (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *American Journal of Human Genetics*, **64**, 232–249.

Richards M and Macaulay V (2001) The mitochondrial gene tree comes of age. *American Journal of Human Genetics*, **68**, 1315–1320.

Richards MB, Macaulay VA, Bandelt HJ and Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. *Annals of Human Genetics*, **62**, 241–260.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, *et al*. (2000) Tracing European founder lineages in the near Eastern mtDNA pool. *American Journal of Human Genetics*, **67**, 1251–1276.

Richards M, Macaulay V, Torroni A and Bandelt HJ (2002) In search of geographical patterns in European mitochondrial DNA. *American Journal of Human Genetics*, **71**, 1168–1174.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O, *et al*. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *American Journal of Human Genetics*, **75**, 128–137.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V and Wallace DC (2004) Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, **303**, 223–226.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, *et al*. (2000) The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science*, **290**, 1155–1159.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM and Wallace DC (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. *American Journal of Human Genetics*, **53**, 563–590.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML and Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics*, **144**, 1835–1850.

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, *et al*. (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *American Journal of Human Genetics*, **69**, 1348–1356.

Vigilant L, Stoneking M, Harpending H, Hawkes K and Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–1507.

Short Specialist Review

The genetic structure of human pathogens

Daniel J. Wilson and Daniel Falush
University of Oxford, Oxford, UK

The contemporary genetic structure and the evolutionary history of human pathogens are important to control and prevention strategies, as well as providing interesting case studies in evolutionary biology. Recently, large-scale DNA sequence data have become available for a wide variety of pathogens, and have revealed a variety of structures as diverse as the pathogens themselves. We describe the patterns observed in several of the most notorious killers, and ask what makes them so different from each other.

The genetic structuring that a human pathogen exhibits not only varies from species to species but also from one place to another. A good example is the protozoan parasite *Plasmodium falciparum*, responsible for the most virulent form of human malaria. *Plasmodium falciparum* populations exhibit a wide range of genetic structures that highlight the importance of local epidemiology. Unlike many human pathogens (viruses and bacteria in particular), *P. falciparum* has an obligate sexual stage in its life cycle. Male and female gametes fuse in the mosquito vector to form a short-lived diploid stage. In general, recombination breaks down associations between particular alleles at different loci (linkage disequilibrium). Nevertheless, strong linkage disequilibrium caused by inbreeding is typical in areas of low prevalence where self-fertilization is commonplace, such as South America. In contrast, *P. falciparum* populations in Africa exhibit high heterozygosity and low linkage disequilibrium, more typical of breeding populations, while populations in Southeast Asia show intermediate levels (Anderson *et al*., 2000).

Genetic structuring can persist stably in a pathogen population in spite of transient epidemic waves, a pattern that is seen in *Neisseria meningitidis*. These waves can often be attributed to the transmission of specific genotypes (Zhu *et al*., 2001). The agent of meningococcal meningitis and septicemia, *N. meningitidis*, is a common, diverse bacterium (Jolley *et al*., 2000), which undergoes recombination at a rate sufficiently high to render the phylogenetic signal between loci incongruent (Holmes *et al*., 1999). Zhu *et al*. (2001) showed how the importation of gene fragments from other *Neisseria* species – which might actually be more frequent than within-species recombination (Linz *et al*., 2000) – can drive these waves. Since the 1960s, subgroup III *N. meningitidis* has caused three successive waves of meningococcal disease worldwide. Within each wave, interspecific recombination

can lead to escape variants with novel antigenic properties. They enjoy transient superiority and rise in frequency, but do not subsequently persist because they do not successfully colonize new host populations. So, although immunological novelty is an important advantage, stabilizing selection may account for the persistence of lineages over time.

Not just recent epidemic waves but the origin and subsequent spread of the entire species may be recorded in the population structure of a recently evolved pathogen. One such historically important pathogen is *Yersinia pestis*, the highly virulent bacterium responsible for plague. The first recorded plague pandemic dates to 541–767 A.D., and is thought to have been imported from East or Central Africa and spread from Egypt to Mediterranean countries. Its extremely limited nucleotide sequence diversity makes genetic analysis difficult because of the lack of polymorphism. A study by Achtman *et al*. (1999) using restriction fragment length polymorphisms revealed that *Y. pestis* is a highly uniform clone of *Y. pseudotuberculosis*, a global pathogen of wild and domestic animals, which infrequently causes a self-limiting infection in humans. On the basis of a comparison of the genome sequences of *Y. pestis* and *Y. pseudotuberculosis*, Prentice *et al*. (2001) have postulated that the critical event in the emergence of *Y. pestis* was the acquisition of a plasmid conferring pathogenicity by horizontal gene transfer. Achtman *et al*. (1999) fitted a model of rapid population expansion to date the origin of the pathogen at 1500–20 000 years ago, consistent with the date of the first pandemic.

It is somewhat counterintuitive to think that a microbe as strictly clonal as *Y. pestis* might have originated from a recombination event in which it acquired a pathogenicity-conferring plasmid. The emergence of the apparently nonrecombining (McVean *et al*., 2002) hepatitis C virus (HCV) in Egypt has been attributed to human medical intervention. HCV, in contrast to *Y. pestis*, exhibits high levels of sequence diversity despite its clonal nature. A leading cause of liver cancer and cirrhosis, HCV is genetically structured into various subtypes, of which 4a is prevalent in Egypt. Pybus *et al*. (2003) performed statistical analysis of the genetic structure within subtype 4a. Their coalescent approach was based on an epidemiological model that dated the period of expansion of subtype 4a to between 1930 and 1955, coinciding with an extensive anti-schistosomiasis vaccination campaign in the country. Moreover, the high rate of spread estimated by the analysis is consistent with existing hypotheses that implied spread by unsterile equipment.

There have been attempts to extract information about historic changes in prevalence even from highly recombining pathogens, the epitome of which must be the human immunodeficiency virus (HIV), which, according to population genetic estimates, is perhaps the most highly recombining of all microorganisms (McVean *et al*., 2002). Lemey *et al*. (2003) fitted a similar model to that of Pybus *et al*. (2003) to HIV-2 gene sequences obtained from patients in Guinea-Bissau. They found that the transmission of HIV-2 increased around 1955–1970, overlapping with the 1963–1974 war of independence in the country. Their analysis hints at the role of social changes, such as wartime, in the establishment of emergent infections.

If the driving force in the historic spread and subsequent globalization of a pathogen was the migration of its host populations, then a pathogen's contemporary genetic structure can contain information about its host past, as has been seen

in *Helicobacter pylori*, a common bacterium that colonizes the human gut and, through a process of constant irritation, is apparently responsible for most of the stomach maladies found in man, including peptic ulcer and gastric cancer. Recombination in *H. pylori* is so high that different loci, and polymorphisms within loci, appear to be in linkage equilibrium (Suerbaum *et al*., 1998), the result of very frequent recombination during mixed colonization by unrelated strains (Falush *et al*., 2001). Nevertheless, Falush *et al*. (2003) found that contemporary *H. pylori* could be divided into seven populations with distinct geographical distributions. The high sequence diversity and residual linkage disequilibrium in modern populations allowed the identification of ancestral populations from Africa, Central Africa, and East Asia. By reconstructing the genetic profiles of these ancient populations, they were able to produce putative migration routes for the ancient populations, migration routes that parallel hypothesized ancient human migration routes. Thus, analysis of *H. pylori* from human populations appears to provide independent insight into the details of human migrations.

The exceptional population structure of *H. pylori* allows reconstruction of the host evolutionary history, highlighting the intimate association of the host and parasite over thousands of years. But why does diversity within *H. pylori* correlate so well with human migrations, when others do not?

In fact, a lack of correlation is generally much easier to explain than a good correlation. Both HIV and *N. meningitidis* spread epidemically, with a transmission time measured in weeks or years rather than decades. Although geographical and cultural barriers can slow down their spread, current levels of human movement have proved sufficient to allow their global dissemination within years or decades. *Mycobacterium tuberculosis*, the cause of TB and human suffering that has inspired many literary works, spreads more slowly, and does show clear patterns of geographic structuring (Hirsh *et al*., 2004). However, because TB is apparently entirely clonal (Supply *et al*., 2003), any selective events associated with the adaption of TB to the human host, or host-pathogen coevolution, will affect the population structure across the entire genome, an example of which may be the global emergence of the hypervirulent Bejing genotype (Lillebaek *et al*., 2003).

The population structure of *H. pylori* is exceptional because its biology includes several key properties. These include slow transmission (which typically seems to occur in families), and, in the absence of antibiotics, typically life-long chronic infection. Further, it mutates at very high rates, which results in informative signals at each gene, while frequent recombination uncouples the signal observed at most genes from selective events that occur at antigenically important loci. Are there other pathogens with a similar suite of properties? Despite a few candidates (e.g., JC virus, Agostini *et al*., 1997), none really seems to fit the bill, but only time will tell which of our unwanted but intimate companions has most to reveal about our own history.

References

Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A and Carniel E (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14043–14048.

Agostini HT, Yanagihara R, Davis V, Ryschkewitsch CF and Stoner GL (1997) Asian genotypes of JC virus in Native Americans and in a Pacific Island population: markers of viral evolution and human migration. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 14542–14546.

Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, Bockarie M, Mokili J, Mharakurwa S, French N, *et al*. (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution*, **17**, 1467–1482.

Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M and Suerbaum S (2001) Recombination and mutation during long-term gastric colonization by Helic*obacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 15056–15061.

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, *et al*. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.

Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW and Small PM (2004) Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4871–4876.

Holmes EC, Urwin R and Maiden MC (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Molecular Biology and Evolution*, **16**, 741–749.

Jolley KA, Kalmusova J, Feil EJ, Gupta S, Musilek M, Kriz P and Maiden MC (2000) Carried meningococci in the Czech Republic: a diverse recombining population. *Journal of Clinical Microbiology*, **38**, 4492–4498.

Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M and Vandamme AM (2003) Tracing the origin and history of the HIV-2 epidemic. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6588–6592.

Lillebaek T, Andersen AB, Dirksen A, Glynn JR and Kremer K (2003) Mycobacterium tuberculosis Beijing genotype. *Emerging Infectious Diseases*, **9**, 1553–1557.

Linz B, Schenker M, Zhu P and Achtman M (2000) Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. *Molecular Microbiology*, **36**, 1049–1058.

McVean G, Awadalla P and Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Prentice MB, James KD, Parkhill J, Baker SG, Stevens K, Simmonds MN, Mungall KL, Churcher C, Oyston PC, Titball RW, *et al*. (2001) *Yersinia pestis* pFra shows biovar-specific differences and recent common ancestry with a *Salmonella enterica* serovar Typhi plasmid. *Journal of Bacteriology*, **183**, 2586–2594.

Pybus OG, Drummond AJ, Nakano T, Robertson BH and Rambaut A (2003) The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular Biology and Evolution*, **20**, 381–387.

Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E, Dyrek I and Achtman M (1998) Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 12619–12624.

Supply P, Warren RM, Banuls AL, Lesjean S, Van der Spuy GD, Lewis LA, Tibayrenc M, Van Helden PD and Locht C (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Molecular Microbiology*, **47**, 529–538.

Zhu P, van der Ende A, Falush D, Brieske N, Morelli G, Linz B, Popovic T, Schuurman IG, Adegbola RA, Zurth K, *et al*. (2001) Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5234–5239.

Short Specialist Review

Genetic signatures of natural selection

Molly Przeworski
Brown University, Providence, RI, USA

Carlos D. Bustamante
Cornell University, Ithaca, NY, USA

1. Introduction

Identifying genomic regions involved in the adaptive divergence of closely related species or populations is a major goal of evolutionary genetics research. Adaptations come about through a variety of modes, including global selection for a particular allele, local adaptation, where alternative alleles are favored in different environments, and balancing selection (heterozygote advantage), where increased diversity is favored at a particular locus. These modes of selection make distinct predictions regarding the distribution of inter- and intraspecific genetic variation near the selected region, thus leaving a distinguishing "signature". For example, repeated favorable changes to a protein lead to accelerated rates of amino acid evolution between species relative to synonymous substitution rates, while long-term heterozygote advantage increases intraspecies diversity around the selected site (Hudson and Kaplan 1988).

2. Using interspecies divergence

Over the past three decades, mathematical and computer models have helped to characterize the signature of natural selection on nucleotide sequence variation (e.g., Maynard Smith and Haigh 1974; Hudson and Kaplan 1988; Takahata 1990). On the basis of these characterizations, statistical approaches have been developed to identify genomic targets of adaptation and make inferences about current and past selective pressures. An important class of methods aims to detect nonneutral evolution of amino acid sequences using phylogenetic patterns of synonymous and nonsynonymous variation (for a review, see Yang and Bielawski 2000). These approaches have been successful in identifying a subset of proteins with an accelerated rate of amino acid evolution (e.g., Clark *et al*., 2003), but cannot

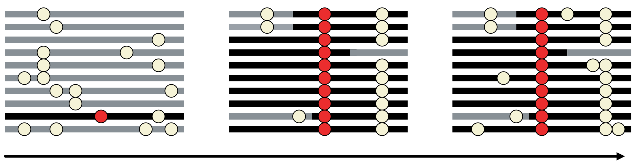


Figure 1 The effect of a "selective sweep" on linked neutral variability. Each line represents a chromosome and the circles the nonancestral allele at a site. Light yellow circles are mutations with no fitness effect, i.e., which evolve neutrally. A favorable mutation arises (shown in red) and, because it is favored, increases rapidly in frequency in the population. It recombines onto a limited number of haplotypes during its ascent; how much recombination occurs depends on the recombination rate and the strength of selection. After fixation, variability is replenished by mutation. However, for a while, most mutations are new and therefore tend to be at low frequency

detect adaptations consisting of few substitutions to a given protein or changes that occur outside of the coding region.

3. Using polymorphism data

Regulatory adaptations or subtle adaptations in proteins may be detectable from patterns of polymorphism in samples of extant individuals, so long as the selective events are recent. Indeed, when a rare allele arises and rapidly fixes in the population, it distorts patterns of variation at linked neutral sites relative to the expectation in the absence of natural selection (Maynard Smith and Haigh 1974) (see Figure 1). Under simplifying assumptions, this signature of a "selective sweep" can be exploited to find regions that have been the target of positive selection in the past $\sim N_e$ generations, where N_e is the diploid "effective population size" of the species (Przeworski 2002). In humans, estimates of N_e are on the order of 10 000, suggesting that polymorphism-based approaches can be informative about selective events over the past ∼250 000 years, including those associated with the emergence of anatomically modern humans.

The most common methods used to identify adaptive genetic changes from polymorphism data are the so-called tests of neutrality. These tests begin by assuming a neutral null model of a random-mating population of constant size (the "standard neutral model") and then assess whether the value of some summary of the data (e.g., the number of segregating sites) is unexpected under this model. A poor fit of the standard neutral model is interpreted as evidence for the action of natural selection. Commonly used test statistics are summaries of the allele frequencies (e.g., Tajima 1989) or of linkage disequilibrium (e.g., Hudson *et al*., 1994). The power of these tests varies widely (e.g., Simonsen *et al*., 1995).

Although tests of neutrality have been used extensively to identify genes that may have experienced recent adaptations, they suffer from a number of drawbacks. For one, there is no obvious way to combine tests performed on different but correlated aspects of the data. In addition, the approach does not allow one to estimate the parameters of interest, such as the timing of the selective event or the strength of selection. Moreover, while a significant result suggests that the data are unlikely

under the null hypothesis, it says nothing about whether the data are more likely under a model of natural selection.

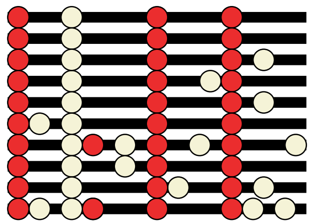
Some of these shortcomings can be addressed by the use of likelihood-based methods that explicitly compare the probability of the data under a model with selection versus without and allow one to estimate parameters (Kim and Stephan 2002; Przeworski 2003). At present, it is not computationally feasible to make these inferences using all the polymorphism data, even for simpler models. Instead, these methods base themselves on (multiple) summaries of the data that are known to be sensitive to selection at a linked site. Although the methods sacrifice some of the information carried by polymorphism data, they have been shown to produce reliable results on data simulated under standard assumptions.

What is less clear is how these methods perform on data generated under more realistic models. Like tests of neutrality, they rely on highly simplified assumptions about the population history. Since allele frequencies are known to be sensitive to demographic assumptions (e.g., Tajima 1989; Simonsen *et al*., 1995), as are other aspects of the data, methods based on these may not be reliable when applied to real data. In particular, when some of these statistics are used as tests of neutrality, they have high rates of Type I error (probability of rejecting a true null hypothesis) if unknown population structure is present (Nielsen 2001; Przeworski 2002). Thus, the interpretation of significant results is not straightforward. In that respect, a potential benefit of using multiple aspects of the data (or all the data) is that the method may be more robust to the misspecification of the model.

Ideally, of course, one would like to incorporate more realistic features of the demographic history. Unfortunately, little is known about appropriate demographic models for species of interest. As an illustration, although the human population has obviously increased in size, we currently have little idea of when or where population growth occurred. An alternative would be to allow for a model of demography and selection and coestimate all demographic and selective parameters. At present, however, this is not computationally feasible. When the aim is to detect selective pressures acting on a particular class of mutations (e.g., amino acid replacements), a third approach is to employ statistical tests that do not rely on an explicit population genetic model but rather look for discordance between distinct functional classes of sites. For example, Akashi (1997) proposed comparing the distribution of allele frequencies at preferred and unpreferred codons as a test of neutrality, and the same can be done for silent versus replacement sites. If a subset of sites are invariant, while the remaining silent and replacement sites are evolving neutrally, the proportion of variable sites that are silent versus replacement should be the same across all frequency ranges – irrespective of the underlying demography of the species. This neutral expectation can be evaluated by a simple goodness of fit test (by conditioning on the total number of polymorphic sites in a set of nonoverlapping frequency ranges).

4. Combining polymorphism and divergence data

In the same spirit, levels of variation within and between species can be compared for silent versus amino acid replacement changes (e.g., McDonald and Kreitman



	Polymorphism	divergence
Amino acid replacement	1	3
Silent replacement	9	1

Figure 2 Effect of repeated beneficial changes in a protein.
Amino acid changes are shown in red and silent changes in yellow. The data in the sample are summarized in the McDonald–Kreitman table shown to the right. Alleles carried by every chromosome in the sample are assumed to be fixed in the species. If mutations at synonymous sites are neutral, a larger ratio of amino acid to silent fixations relative to amino acid to silent polymorphisms suggests that the protein is evolving under positive selection (in this example, P = 0.041, as assessed by a Fischer's Exact Test)

1991). The McDonald–Kreitman (MK; described in Figure 2) and Akashi tests are fully robust to demography since they depend only on differences in patterns of variation among classes of sites that share the same evolutionary histories, that is, no demographic assumptions need to be specified to test the null hypothesis of neutrality of mutations. In their current formulation, MK and Akashi tests are inapplicable to changes that occur outside of the coding region. In principle, similar methodology could be extended to noncoding regions; however, the utility of the approach will depend on the ability to accurately dichotomize noncoding sites as functional or nonfunctional.

MK data can also be used to infer the magnitude and directionality of natural selection if one is willing to specify a demographic model and a selective alternative to neutrality, such as genic selection (Sawyer and Hartl 1992). An advantage of such a model-based approach is that one can directly compare the inferred strength of selection across genes and classify loci into those that show evidence of being subject to deleterious mutations, neutrally evolving, or putatively positively selected. The parameterized test of neutrality has the same Type I error and power as the usual MK table (described in Figure 2). In contrast to significance testing, however, parameter estimation does depend on the specifics of demography and recombination used to estimate the strength, and possibly the directionality, of selection. One way to increase the reliability of these model-based approaches may be to combine genomic patterns of polymorphism and divergence for different classes of sites *across the genome* in order to detect selection in a given gene. Such an approach has been shown to increase power while maintaining a relatively high robustness to the misspecification of the null model (Bustamante *et al*., 2002).

5. Empirical approaches

The last few years have also seen the use of observed distributions of human genetic variation to identify loci under selection. To date, researchers have taken

one of two approaches: in the first, they have contrasted a set of candidate loci to putatively neutrally evolving regions in an attempt to "control" for genome-wide demographic effects (e.g., Gilad *et al*., 2003). In the second, they have ranked loci according to some summary of the data and focused on the subset in the tails of the empirical distribution (e.g., Akey *et al*., 2002). These empirical approaches circumvent some of the difficulties involved in specifying a demographic model but at a cost, since little can be learned about a region once it has been identified as unusual. Moreover, although these approaches do not rely on a model, they make a number of implicit assumptions: in particular, the second approach is only practical if most loci in the tails of the distribution are indeed targets of selection. Whether this condition is likely to be met depends on the nature and strength of selection as well as on the demographic parameters (cf. Beaumont and Balding 2004).

6. Ascertainment bias

A final caveat pertains to ascertainment bias in population genetic data. Almost all of population genetic theory for detecting natural selection assumes that the researcher has full sequence for all individuals in the sample. Yet, there is growing popularity of schemes where one uses a small "discovery" panel to identify common SNPs (single nucleotide polymorphisms) and then types these markers in a much larger sample to obtain gene frequency estimates across populations. These schemes violate the assumptions of the methods, leading to invalid analyses unless the ascertainment process is explicitly and accurately modeled (cf. Nielsen and Signorovitch 2003). This is particularly problematic for population genetic analyses based on pooling ascertained data (such as that found in large public databases e.g., dbSNPs). Such analyses run the risk of detecting patterns that do not result from natural selection but instead from inherent biases in the data.

7. What have we learned thus far?

Recent analyses of genomic patterns of variation have helped to discern the principal modes of adaptation. Notably, the use of MK tables (and related approaches) has suggested that many coding region mutations are weakly deleterious, such that they can persist as polymorphisms in the population but rarely contribute to divergence (e.g., Akashi 1995; Weinreich and Rand 2000; Hellmann *et al*., 2003). Similar approaches also indicate that most amino acid changes fixed between *Drosophila melanogaster* and *Drosophila simulans* are adaptive (e.g., Sawyer *et al*., 2003).

In addition, surveys of noncoding variation in *D. melanogaster* and other species have revealed decreased polymorphism, but not decreased divergence, in regions of low recombination (Begun and Aquadro 1992). This observation has no clear neutral explanation and is likely due to repeated episodes of directional selection at linked sites (Andolfatto and Przeworski 2001). From these lines of inquiry, it is apparent that natural selection shapes the patterns of genetic variation; the challenge is to harness this signature to help identify specific regions that underlie adaptations.

We already have a number of convincing examples: balancing selection at Adh in *D. melanogaster* (McDonald and Kreitman 1991) and the MHC complex in mammals (Hughes and Nei 1988), or repeated directional selection on reproductive proteins in mammals (e.g., Swanson *et al*., 2001) and on the HIV envelope gene (Nielsen and Yang 1998). What strengthens the interpretation in these cases is that the finding makes biological sense in light of the phenotype. When such independent knowledge is not available, statistical methods are only a starting point for further functional analyses. As an illustration, both polymorphism and divergence data strongly suggest that directional selection has shaped the evolution of *FOXP2*, a gene involved in speech and language in humans (Enard *et al*., 2002). However, to substantiate this hypothesis, we need a better understanding of the function of *FOXP2*.

8. Outlook

The current wave of comparative evolutionary genomics promises to identify many loci that show evidence of positive selection based on patterns of DNA sequence variation. While these approaches will generate a number of interesting hypotheses, it is important to realize that they are not ends in themselves and that the ultimate utility of this research program depends on experimental confirmation. We therefore look forward to efforts to verify (or disprove) these hypotheses via comparative functional analysis of molecular, cellular, and organismal phenotypes linked to putatively adaptive genotypic variation.

References

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics*, **139**, 1067–1076.

Akashi H (1997) Codon bias evolution in Drosophila. Population Genetics of Mutation-selection Drift. *Gene*, **205**, 269–278.

Akey M, Zhang G, Zhang K, Jin L and Shriver D (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Andolfatto P and Przeworski M (2001) Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics*, **158**, 657–665.

Beaumont A and Balding J (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Begun J and Aquadro F (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Bustamante D, Nielsen R, Sawyer A, Olsen M, Purugganan D and Hartl L (2002) The cost of inbreeding in Arabidopsis. *Nature*, **416**, 531–534.

Clark G, Glanowski S, Nielsen R, Thomas D, Kejariwal A, Todd A, Tanenbaum M, Civello D, Lu F, Murphy B, *et al*. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.

Enard W, Przeworski M, Fisher E, Lai S, Wiebe V, Kitano T, Monaco P and Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Gilad Y, Bustamante C, Lancet D and Paabo S (2003) Natural selection on the OR gene family in humans and chimpanzees. *American Journal of Human Genetics*, **73**, 489–501.

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B and Paabo S (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Research*, **13**, 831–837.

Hudson R, Bailey K, Skarecky D, Kwiatowski J and Ayala J (1994) Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics*, **136**, 1329–1340.

Hudson R and Kaplan L (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Hughes L and Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Kim Y and Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**, 765–777.

Maynard Smith M and Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

McDonald H and Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.

Nielsen R and Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, **63**, 245–255.

Nielsen R and Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics*, **160**, 1179–1189.

Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.

Sawyer A and Hartl L (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

Sawyer A, Kulathinal J, Bustamante D and Hartl L (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *Journal of Molecular Evolution*, **57** (Suppl 1), S154–S164.

Simonsen L, Churchill A and Aquadro F (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413–429.

Swanson J, Clark G, Waldrip-Dail M, Wolfner F and Aquadro F (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7375–7379.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2419–2423.

Weinreich M and Rand M (2000) Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics*, **156**, 385–399.

Yang Z and Bielawski P (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, **15**, 496–503.

The role of gene regulation in evolution

Matthew V. Rockman
Duke University, Durham, NC, USA

Evolutionary biologists have long sought to understand the genetic basis of evolutionary change. For theoretical reasons, biologists have predicted that mutations affecting the regulation of gene expression, rather than the physical structure of gene products, will play a central role in phenotypic evolution. Recent empirical work has begun to nail down the molecular details underpinning phenotypic evolution, and in a remarkable number of cases, the evidence strongly supports the theoretical models.

Two disparate schools of theory have predicted the central role for gene regulation in evolutionary change. The critical parameter for both accounts is pleiotropy, the phenomenon whereby a single mutation has multiple phenotypic effects.

One school of thought, growing out of mathematical population genetics and quantitative genetics, holds that mutations fixed by evolution are likely to be those with very limited pleiotropic effects (Fisher, 1930). Because extant organisms must be in some sense fit for survival, a mutation that alters many features at once is likely to change at least some of them for the worse. A mutation that alters only a single feature stands a better chance of improving the organism's prospects, and therefore spreading through the populations.

Regulatory mutations, particularly those that affect transcriptional regulation, are ideal candidates for mutations with restricted pleiotropy. The temporal and spatial pattern of a gene's expression, the quantitative profile of its transcription rate, and its responsiveness to a range of inductive cues are influenced by the binding of transcription factors to cis-regulatory DNA elements. Because the array of transcription factors differs among cells, each piece of the total expression pattern may be regulated independently by a unique complement of transcription factors and a discrete segment of regulatory DNA (Arnone and Davidson, 1997). Cis-regulatory mutations that affect expression in one part of the organism may leave expression in other parts unchanged. This modular property of transcriptional regulation makes cis-regulatory DNA conducive to evolutionarily important mutation (Gerhart and Kirschner, 1997; Davidson, 2001).

A second school of thought, growing out of evolutionary developmental biology and molecular evolution, holds that the mutations responsible for the conspicuous phenotypic transformations that account for the diversity of organismal form are

likely to be those with quite radical pleiotropic effects. More specifically, a mutation with coordinated pleiotropic effects, radically altering form in a coherent and systemic fashion, is a mutation that could account for a dramatic change in form while retaining functional integration.

While evolutionary change affecting a suite of integrated traits could be due to a large number of separate mutations, a single mutation influencing the whole suite of traits is more likely to preserve functional integration, particularly when the traits are genetically correlated by a shared developmental basis. Even before the molecular mechanisms of development were well understood, this notion found footing in the concept of heterochrony (e.g., Gould, 1977), whereby simple changes in the timing of developmental events resulted in dramatic, but functionally integrated, changes in form. Of course, heterochronic changes in timing are likely to result from changes in the temporal or quantitative regulation of gene expression. The argument for coordinated pleiotropy is famously realized in King and Wilson's (1975) model for human origins. Because humans and chimpanzees are so similar at the level of structural genes, they argued, the striking phenotypic differences between the species must be due to a small number of mutations with systemic effects.

At the same time, developmental biologists have found that the set of genes regulating development is largely shared among animals (and similarly among plants), a finding confirmed by the proliferation of whole-genome sequences and EST projects. Given a shared "toolkit" of developmentally important proteins, evolutionary developmental biologists have embraced the notion that changes in form are due to changes in the pattern of expression of the shared genes, and the recruitment of whole developmental networks to evolutionarily novel tasks (e.g., Lowe and Wray, 1997). Studies correlating changes in patterns of gene expression to changes in phenotypic form have become a mainstay of evolutionary developmental biology.

The theoretical commitments to limited pleiotropy on the one hand, and to pervasive, integrated pleiotropy on the other hand, have been brought together and reinforced by recent empirical work, which bridges the divide between the developmental biologists, who have focused on comparisons among species, and quantitative geneticists, who have been examining variation within species. Although case studies documenting the role of regulatory mutations in the genetic basis of phenotypic evolution are now numerous (e.g., Sucena *et al*., 2003; Beldade *et al*., 2002; Wang and Chamberlin, 2004; Wang *et al*., 1999; Enattah *et al*., 2002), an extraordinary recent study of stickleback fish (*Gasterosteus aculeatus*) exemplifies the role of gene regulation in meeting the competing needs of limited pleiotropy and systemic effects.

Shapiro *et al*. (2004) genetically mapped the loci underlying the dramatic and very recent evolutionary reduction of the pelvic skeleton in lake populations of sticklebacks. A virtue of this study is that the mapping strategy makes no a priori assumption about the cis-regulatory nature of the underlying locus. Shapiro *et al*. found that a single major locus accounts for a suite of pelvic traits, consistent with the coordinated pleiotropy model. The locus was identified as *Pitx1*, a transcription factor, but no differences were seen in the protein sequence. Instead, the fish with and without pelvic skeletons differ in the spatial pattern of *Pitx1* expression during development. Fish with reduced skeletons show a loss of expression at the site

of the presumptive pelvic region. At the same time, expression of the gene in a variety of other developing organs is retained, consistent with the restricted pleiotropy model. Because the complete knockout of *Pitx1* (in mice) is lethal, the evolutionary importance of the spatially limited mutational effect can hardly be doubted.

Regulatory mutations fit the theoretical bill for evolutionarily important mutations. But phenotypic evolution certainly involves mutations of all sorts – indeed, recent work revealing the extent of positive selection on amino acid substitution in human evolution (Clark *et al*., 2003) shows that structural mutations are essential contributors to the origin of our own species, despite the predictions of King and Wilson (1975). If regulatory mutations are rare, then their importance to evolution, regardless of their suitability, may be minimal. Three recent findings suggest that regulatory mutations may actually constitute the majority of phenotypically relevant variation in populations. First, comparative genomics has revealed that a surprising fraction of noncoding DNA is conserved among divergent species (e.g., Rat Genome Sequencing Project Consortium, 2004). Because evolutionary conservation implies functional importance, it seems that most of the functional nucleotides (in mammalian genomes, at least) are noncoding and potentially cis-regulatory. These nucleotides, therefore, represent a bigger target for mutation than do coding sequences. Second, surveys of noncoding variants (again, in humans) suggest that a large fraction affect transcription (Rockman and Wray, 2002; Buckland *et al*., 2004). Because we lack a means to infer the functional import of cis-regulatory DNA directly from sequence, however, our understanding of the extent of regulatory variation remains limited. Third, genome-wide analyses of gene expression variation in a range of species have shown that such phenotypic variation is ubiquitous (e.g. Jin *et al*., 2001; Cavalieri *et al*., 2000). Although efforts to map the loci responsible for the variation are highly sensitive to experimental design and statistical analysis issues, one of the best studies to date found that the majority of such loci map to the same genetic location as the variable transcript, implying that the loci may be cis-acting (e.g., Brem *et al*., 2002). Allele-specific measurement of gene expression, a technique that exclusively discovers the effects of cis-acting variation, also reveals the presence of abundant variation in humans (Lo *et al*., 2003). Consequently, putting aside all theoretical commitments, we should not be surprised to find regulatory variation underlying phenotypic variation.

Empirical and theoretical results point to changes in gene regulation as a major factor in phenotypic evolution. As molecular quantitative genetics continues to make strides in connecting phenotype and genotype, our understanding of the molecular basis for evolutionary change will grow even richer.

Further reading

Caroll S, Grenier J and Weatherbee S (2001) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*, Blackwell Scientific: Boston.

Stern DL (2000) Evolutionary developmental biology and the problem of variation. *Evolution*, **54**, 1079–1091.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV and Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377–1419.

Wilkins AS (2001) *The Evolution of Developmental Pathways*, Sinauer Associates: Sunderland.

References

Arnone MI and Davidson EH (1997) The hardwiring of development: Organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.

Beldade P, Brakefield PM and Long AD (2002) Contribution of *Distal-less* to quantitative variation in butterfly eyespots. *Nature*, **415**, 315–318.

Brem RB, Yvert G, Clinton R and Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Buckland PR, Coleman SL, Hoogendoorn B, Guy C, Smith SK and O'Donovan MC (2004) A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. *Gene Expression*, **11**, 233–239.

Cavalieri D, Townsend JP and Hartl DL (2000) Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12369–12374.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, *et al*. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.

Davidson EH (2001) *Genomic Regulatory Systems: Development and Evolution*, Academic Press: San Diego.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L and Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, **30**, 233–237.

Fisher RA (1930) *The Genetical Theory of Natural Selection*, Clarendon Press: Oxford.

Gerhart J and Kirschner M (1997) *Cells, Embryos, and Evolution*, Blackwell Science: Boston.

Gould SJ (1977) *Ontogeny and Phylogeny*, Belknap Press: Cambridge.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G and Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics*, **29**, 389–395.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH and Lee MP (2003) Allelic variation in gene expression is common in the human genome. *Genome Research*, **13**, 1855–1862.

Lowe CJ and Wray GA (1997) Radical alterations in the roles of homeobox genes during echinoderm evolution. *Nature*, **389**, 718–721.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Rockman MV and Wray GA (2002) Abundant raw material for *cis*-regulatory evolution in humans. *Molecular Biology and Evolution*, **19**, 1991–2004.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D and Kingsley DM (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**, 717–723.

Sucena E, Delon I, Jones I, Payre F and Stern DL (2003) Regulatory evolution of *shavenbaby/ovo* underlies multiple cases of morphological parallelism. *Nature*, **424**, 935–938.

Wang RL, Stec A, Hey J, Lukens L and Doebley J (1999) The limits of selection during maize domestication. *Nature*, **398**, 236–239.

Wang X and Chamberlin HM (2004) Evolutionary innovation of the excretory system in *Caenorhabditis elegans*. *Nature Genetics*, **36**, 231–232.

Short Specialist Review

Modeling protein evolution

David D. Pollock
Louisiana State University, Baton Rouge, LA, USA

Richard A. Goldstein
National Institute for Medical Research, London, UK

Proteins are the biological macromolecular entities most closely and directly related to organismal function, and knowledge of protein structure and function is critical to understanding biological organization. The evolution of protein sequence is molded by the specific requirements of structure and function, and thus an important avenue for predicting features of protein structure and function is to model the evolutionary process in proteins. Despite this, models of protein evolution have lagged far behind models of DNA evolution. An obvious and important reason for this is that proteins are composed of 20 amino acids, while DNA is composed of only four nucleotides; it is both harder to make calculations with amino acids and harder to find data sets sufficiently large to accurately estimate substitution probabilities between so many states. These problems become compounded when the evolutionary process occurring at the DNA and protein levels are combined to model evolution of amino acid coding units (codons, which have 64 states). A somewhat more subtle point is that DNA models have progressed partly because they usually assume, implicitly or explicitly, that nucleotide substitutions during the course of evolution occur in a random, or neutral fashion, with all positions at all times evolving independently in the same manner and rate. In contrast, experimental evidence (e.g., mutagenesis and functional analysis) has long shown that different positions in proteins can have wildly different tolerances for different amino acids and that nonadditive or interactive functional relationships among positions abound.

Although it causes computational difficulties, the complexity of protein evolution also provides useful opportunities. Because proteins are under higher degrees of selective pressure, they often change more slowly, and their evolution can allow us to look back further in evolutionary time. The dependence of the evolutionary process on a protein's structure and function means that we can use the evolutionary data to gain insight and understanding of these fundamental characteristics. Studying changes in the evolutionary process of individual proteins can tell us about how protein structure and function change over time. Finally, the sequence databases represent the record of eons of natural mutagenesis and selection experiments, allowing us to probe the relationship between protein sequence and resultant protein properties.

A number of recent developments have allowed the modeling of protein evolution to become a more accurate, diversified, and broadly useful pursuit. An increase in computational power is perhaps foremost among these developments, but not simply because the same old calculations can be performed faster. In addition to increasing speed, computational advances have spurred and made practical the development of novel and sophisticated statistical methodologies using complex models that were unthinkable when computers were slower. These fundamentally model-based approaches allow incorporation of biochemical knowledge and testing of evolutionary hypotheses in a flexible and statistically sound manner. They also allow the incorporation of hypotheses concerning the phylogenetic relationships among genes or species, a component that is essential for reducing noise and spurious correlations in evolutionary analyses. The simple but obviously incorrect assumption of treating sequences as though they are independent (unrelated) entities is no longer necessary or advisable.

Another important development, the same that spurred the creation of the genomics and bioinformatics fields, was the advent of rapid, cheap, and large-scale DNA sequencing capabilities. Along with the sequencing efforts focused on obtaining large quantities of sequence from one or a few species (which except for large multigene families are relatively useless for studying evolutionary processes), there has also been considerable production of homologous sequences from divergent organisms. This sampling of many taxa, or "genomic biodiversity", is essential to the development of sophisticated and realistic models of protein evolution, since discerning the evolutionary behavior at individual sites requires enough biodiversity data such that many substitutions will have occurred at each individual site over the evolutionary time during which the sequences diverged.

Finally, the newest and potentially most revolutionary developments have been in the fields of protein structure prediction, experimental determination, and design. Despite a tradition of unbridled optimism, protein folding and structure prediction has long been problematical and even less tractable than studying protein evolutionary dynamics. There has been more success with the inverse problem, that of finding sequences that conform to a particular fold. A dramatic recent success in the inverse folding field used a heuristically optimized combination of simple energy potentials and observed distributions of conformations for both amino acid side chains and main chain oligomers. Calculations for these methods are relatively rapid, and are in the range of being useful for evolutionary studies and predictions. In addition, observations of adjacency statistics between pairs of residues in an enlarged database of three-dimensional protein structures (from X-ray crystallography or NMR), although obviously crude, have provided useful information for modifying amino acid substitution probabilities, and have provided semirealistic complex models for simulating long-term evolutionary processes and discovering consequences of these processes that may be reflected in real proteins.

How then have models of protein evolution been developed to incorporate biological realism? For a long time, substitution models were never inferred from an individual data set of interest, but were instead obtained from observations of differences between many closely related protein pairs or sets of sequences. Since these observations were averaged over protein positions, evolutionary time, and many unrelated proteins, they necessarily assumed an unwarranted consistency.

Relatively early modifications included specific focus on slowly evolving sites, particular genome types (e.g., mitochondria or viral genomes), or residues involved in particular structural features (e.g., α-helices, β-sheets, or buried residues and residues exposed to solvent), as well as collecting observations for different amounts of evolutionary divergence. Eventually, models of transition between these different contexts along the sequence (hidden Markov models) were used to allow incorporation of these specialized observations into phylogenetic likelihood analysis. Some success in predicting secondary structure was achieved, along with more probable reconstructions of protein phylogenies. Still, though, there was not great reason to believe that these specialized contexts, defined on the basis of preconceived conceptions of what might be most important in determining evolutionary processes, did not represent unknown amalgams of heterogeneous processes that were yet to be deciphered. Although incorporation of rate variation among sites, a technique used in DNA evolutionary analysis, was used to address some of this hidden variation, an important advance was the use of mixture models and substitution matrices derived as functions of the chemical properties of amino acids. These mixture models allow the associations of substitution matrices with positions in an alignment to arise freely during the course of analysis, and thus can obtain novel information and produce inferences that are not possible when the substitution classes are predefined. The evolutionary process may also change over time, but evolutionary models incorporating such change are still in their infancy, and still limited by the size of current data sets.

One of the more important reasons for improving models of evolution in proteins is to understand the forces of selection that act on proteins, and to separate these forces from stochastic processes that affect substitution probabilities (random drift). Although it cannot be said that these two processes of change have been cleanly separated, in many cases statistical features have been evaluated that strongly indicate selection. The most clear-cut cases are those of diversifying selection, in which a sort of molecular cat-and-mouse scenario emerges that drives amino acid substitution at an accelerated rate that is greater than neutral expectations. Such scenarios may occur commonly in situations involving, for instance, a pathogen avoiding a host immune response, but still represent special cases and are not observed in most proteins. It is much more common that a burst of amino acid substitution might occur on a particular branch, possibly as a result of a change of function or specificity, and this may be detected as a brief elevated rate of amino acid versus nucleotide substitution. It is also possible to detect coevolution between residues in proteins, in which case substitutions at one position alter substitution probabilities at other positions. When genes have duplicated, it is possible to detect changes in rates or patterns of substitutions at individual sites, and thus to identify changes that may have been due to functional divergence. Finally, it is possible to use evolutionary analysis to predict ancestral sequences; these can then be resurrected and analyzed to infer patterns of functional change along the phylogeny (a process sometimes referred to as "paleobiochemistry").

Despite the successes of these types of analyses, it may still be argued that we are far from a complete and realistic model of protein evolution. Many of the techniques used simply detect extreme evolutionary processes or unexplained changes in the evolutionary process, while the causal mechanism (adaptive burst,

functional divergence, coevolution) is more a matter of perspective and hope than of direct evidence. Many of these mechanisms may be related and interact in ways that have yet to be deciphered. Evolutionary analyses have also not yet made dramatic progress in prediction of protein structure and function, although preliminary results are promising. The advances in protein engineering are therefore quite exciting, in that they may provide an avenue for integrating models of evolutionary change with realistic biophysical models that predict sequence compatibility with specific structures. Simple pairwise contact probabilities have already been used to model sequence evolution in structures from the protein data bank, and the full integration of structure and function prediction methods with models of protein sequence evolution may soon provide more accurate and useful tools for both purposes.

Measuring variation in natural populations: a primer

Henry Harpending
University of Utah, Salt Lake City, UT, USA

1. Sources of data

Most measures of variation (synonymous with diversity) in natural populations quantify the average amount of difference between two entities from a defined population. For example, biologists may be interested in differences in size or length among organisms in a population, differences among DNA or protein sequences from those organisms, and so on.

Variation in quantitative traits (synonymous with metric traits) such as mass, length, or color is quantified with standard statistics such as means, variances, and covariances. Many work with the logarithms of measurements rather than with the measurements themselves. The statistics have useful properties when the measured quantities are normally distributed, and the distributions of logarithms of metric traits are often more like normal distributions.

Classical markers are polymorphisms where DNA variation is detected indirectly rather than by sequencing. The first to be studied systematically were immunological markers such as the ABO and Rh loci. At the ABO locus, for example, a chromosome could code for A-substance, B-substance, or no substance. Over time, technologies appeared that increased the number of markers: electrophoresis allowed protein variants to be distinguished, and restriction enzymes that recognize and cut a specific short nucleotide sequence led to a large number of markers for restriction fragment length polymorphisms (RFLPs) where the alleles were either cut or not cut by a specific enzyme. Methods of measuring diversity for all these are essentially the same, computation of expected heterozygosity under random mating. As large-scale sequencing of DNA has become feasible, these classical markers have fallen into disuse.

Classical markers, electrophoretic loci, and RFLPs are compromised and not so useful for the assessment of population differences in diversity on a global scale. A landmark study of classical markers in humans revealed essentially no interesting patterns nor striking population differences (Cavalli-Sforza *et al*., 1994), while subsequent work on DNA sequences and on repeat polymorphisms invariably shows that diversity within populations is greater in African populations. The problem is ascertainment bias: the markers were discovered mostly in European populations, so the sample of marker loci available is biased toward loci that are

most variable in Europeans. This skews population diversity comparisons, and excess diversity in African populations is obscured. The human genome project has led to the discovery of a large number of single nucleotide polymorphisms (SNPs), millions of them, but the usefulness of these for population studies is again compromised by the ascertainment problem. The origins of the chromosomes used for discovery are not publicly known.

Some of this difficulty with ascertainment is overcome with the use of variable number of tandem repeat (VNTR) markers. VNTR loci are places on the chromosome where there are long stutters of a motif. For example, a locus where there are repeats of the nucleotide sequence ATAG would be a tetranucleotide polymorphism because the motif is four bases long. Short VNTRs, also called short tandem repeats (STRs) or microsatellites, with motifs of two to four repeats have been discovered and published in large numbers. Because of the high diversity of these markers, the ascertainment problem is not great, and they are useful for population comparisons (Rogers and Jorde, 1996).

In recent decades, technology has become available to make direct study of DNA sequences from populations possible. The first sequence collections published were from mitochondrial DNA because it is available at high concentrations in cells and is haploid (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1 and Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1). More recently, nuclear autosomal sequences have become widely available. In general, sequence data do not suffer from the ascertainment bias that interferes with interpretations of classical markers, RFLPs, and SNPs.

2. Measuring and describing variation

It is important to be aware of the distinction between purely descriptive statistics and interpretations of those statistics in terms of a model of evolution. Most of the statistics I discuss below can be justified in terms of an evolutionary model but they are perfectly useful and routinely used when the conditions of the model are violated. When they are model-bound, it is usually appropriate to make small corrections to account for estimation bias: these issues are treated well in Nei (1987).

Almost all descriptive statistics about variation measure average differences between entities, whether organisms or DNA sequences or repeat lengths. A common justification is that differences accumulate since the separation time of the entities and the statistics are estimates of that time. The time may be an average coalescence time between DNA sequences or, in the case of organisms, an average coalescence time over all the loci in those organisms that contribute to variation in the trait.

Much theory is based on the special case of neutral genes in a panmictic population that has been of constant size for a long time, 4 to 8 G generations where G is the number of genes transmitted in the population each generation. For quantitative traits of whole organisms, the assumption is that a large number of neutral mutations of small effect accumulate along the lines of descent of loci that affect the trait. For VNTRs, the assumption is that mutations cause alleles to change in length by single steps, that loss or gain of a motif are equally likely, and that allele

length is selectively neutral. For DNA sequences, the assumption is that mutations occur along the sequence according to a Poisson process, that the mutations are neutral, and that there is such a large number of nucleotide positions relative to the number of mutations that any single nucleotide position never experienced more than a single mutation event: this is the infinite sites assumption.

In the case of quantitative traits, the means, variances, and covariances among the traits, or, the logarithms of the traits, are the ordinary description of variation. If the heritabilities of the traits are known, then the covariance matrix among the traits can be transformed to be proportional to variation of genes underlying the traits. The heritability of a trait is the fraction of the variance that is due to variation in the additive effects of the underlying genes. When there are several traits, the heritability is a matrix, but in practice, detailed knowledge of heritabilities in natural populations is not available. Workers generally choose a plausible figure, perhaps 30 to 50%, and assume it applies to all the traits. An example of this approach is found in Relethford and Harpending (1994), who compared world human variation in craniometric traits with variation in classical markers and found the two data sets to agree closely.

Variation in a set of classical markers or RFLPs or SNPs is measured simply by averaging the expected heterozygosity over all the loci, where expected heterozygosity is computed by computing allele frequencies, then summing the product of each frequency and its complement over all the alleles. This sum has no theoretical interpretation because of ascertainment issues. It is possible to compare closely related populations but, as I mentioned above, global comparisons are apparently meaningless. As an example as part of a study of Ashkenazi Jewish genetics (Cochran and Harpending, 2004), we collected marker frequencies at 144 autosomal loci from several European populations from the excellent ALFRED website (ALFRED, 2004) maintained by Kenneth and Judy Kidd at Yale University. Relative to a baseline of 1.0 for mixed Europeans, the average heterozygosity of Ashkenazi was 0.988, of Russians 0.988, and of Samaritans 0.846. The reduced heterozygosity of Samaritans is a signature of the population bottleneck in their history, while the figure for Ashkenazi Jews, comparable to that of Russians, shows that there was no detectable bottleneck at all in Ashkenazi history. This is of interest because the high frequency of certain inherited disorders in Ashkenazi is often attributed to genetic drift during a bottleneck, but the data show no sign of a bottleneck.

Variation in a sample of STR data can be measured as heterozygosity, the probability that two alleles are not the same, but a better measure is allele size variance. The single-step mutation model of STRs leads to the expectation that the mean squared difference in size between two alleles is proportional to their coalescence time. The mean squared allele size difference is just twice the variance of allele size, so overall variation is computed as the average variance of all loci. Empirically this is not satisfactory because there are usually outlier loci with very large variances that dominate the results. In comparisons of variation between populations, it is better to work with rank-ordered differences in variance rather than with variances themselves. (Jorde *et al*., 1997).

There are two standard ways to measure variation in a collection of DNA sequences. One is the mean pairwise sequence difference (MPD), the average

number of nucleotide differences between all pairs of sequences. This is equivalent to heterozygosity computed at each nucleotide position and summed over all sites in the sequence. Since the number of mutations along a genealogy of a pair of sequences under a simple model is on average proportional to the length of the genealogy, MPD is another estimator of average separation time of a set of sequences. As an example, there is on the order of a single nucleotide difference between two human DNA sequences per 1000 nucleotide positions, so the MPD would be equal to 1 for kilobyte sequences, 5 for five kilobyte sequences, and so on.

Another statistic appropriate for sequences is the normalized number of segregating sites in the sample: under standard simplifying assumptions, the number of segregating sites divided by $\sum_{i=1}^{n-1} 1/i$, where n is the number of sequences, should be equal to the MPD. The difference between the two numbers is the basis of the Tajima statistic (Tajima, 1989) used to test for selection or demographic change in a population. This is the only statistic discussed in this article that is not a measure of the average separation time of chromosomes.

All the above statistics describe variation within a population. It is often of interest to describe how different populations are from each other, variation among populations. The simple way to do this is to use any of the above statistics to compute variation separately within each of the subpopulations, take the average, and call it V_s. Then pool the subpopulations as if they were all members of the same population, compute variation in this pooled sample, and call it V_t. From these, compute the fraction $F_{st} = (V_t - V_s)/V_t$ to describe the fraction of total variation that is between populations. This is invariably 10 to 15% among major human populations whether measured on metric traits, classical markers, VNTRs, or DNA sequences. Therefore, at neutral markers, the differences among major human populations correspond roughly to the differences among sets of half siblings from a random-mating population (12.5%) (*see* Article 2, **Modeling human genetic history**, Volume 1).

3. Interpreting pattern in diversity

So far I have discussed statistics that are single numbers that quantify the amount of diversity in a population. With many kinds of genetic data, there is more information to be gotten by looking at the patterning of diversity. With a collection of DNA sequences we can quantify diversity by the average pairwise sequence difference, but we can also look beyond the average at the distribution of all pairwise sequence differences. Similarly, the number of segregating sites is the basis of another number that quantifies diversity, as described above. Some segregating sites are present as singletons, sites where there is one nucleotide in a single sequence and another nucleotide in the $n - 1$ other n sequences, while other segregating sites may have each nucleotide present in several of the sequences. Sequences in a sample are tips of a tree of descent, and characteristics of the sample of sequences may allow us to infer characteristics of that tree and from there something about population history or natural selection at the locus.

Figure 1 shows the history of a sample of seven sequences from a population. As we go backward in time, the number of ancestors of the sample decreases as pairs

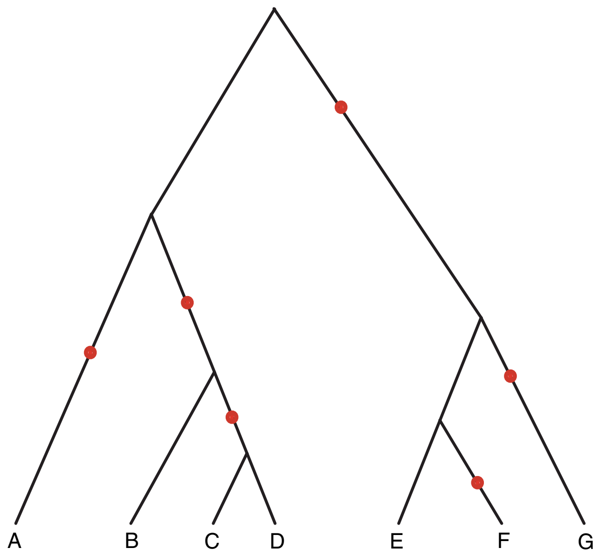


Figure 1 Gene tree showing the history of seven genes sampled from a population. The top of the tree, the coalescence of the sample, is usually 1 to 2 million years in the past for most human genes. Red circles show where mutations occurred

of sequences coalesce in common ancestors. For example, if there were two cousins in a sample of mtDNA sequences, daughters of sisters, then two generations ago their mtDNAs coalesced into the mtDNA carried by their maternal grandmother. We do not usually know what the tree that generated a set of data was like, but several simple statistics are useful to infer properties of the tree.

The tree in Figure 1 is typical of gene trees of neutral genes in populations that have not undergone drastic changes in size: the top of the tree for many human nuclear regions is on the order of 1.5 to 2 million years old. In the history of these sequences, six mutations occurred, indicated by circles on the tree. Mutations occur at random along the branches, and the probability of a mutation in any branch is proportional to the length of the branch. The oldest mutation in the tree is present in sequences E, F, and G, splitting the sample into 4 without and 3 with the mutation. Sequences B, C, and D share a mutation that also splits the sample into 3 with and 4 without. Sequences F, G, and A each carry a unique mutation, a singleton. Finally, sequences C and D share a mutation. If we tabulate segregating sites according to the number of copies of the mutation, we have 3 mutations that occur in 1 sequence, 1 that occurs in 2 sequences, and 2 that occur in 3 sequences. This tabulation is called a frequency spectrum, and we will see below that different demographic or selective histories can lead to different and distinctive frequency spectra.

Figure 2 shows a coalescence tree that looks very different from that in Figure 1: it is shaped more like a comb or a star. There are at first few coalescence events, going backward in time, but then coalescences occur rapidly. This tree might represent human mitochondrial DNA, with the top of the tree at about 200 000 years ago. Since the rate of coalescence at any time is proportional to the reciprocal of the population size, Figure 2 suggests that the population from

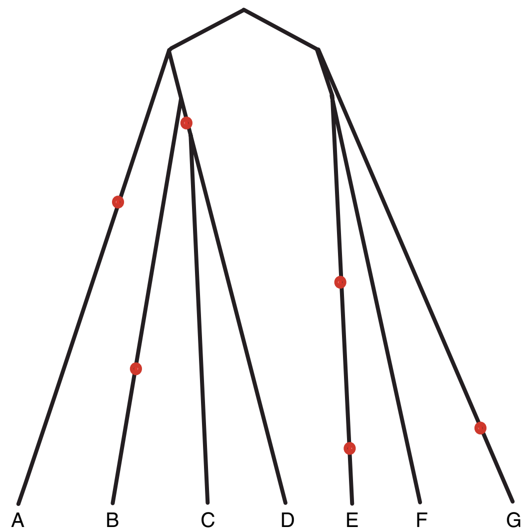


Figure 2 Gene tree showing the history of seven genes sampled from a population as in Figure 1. The population underwent population growth from a small number in the past, just after the coalescence. Human mitochondrial DNA has a history like this, with the top of the tree about 250 000 years ago

which we drew our samples was large until some time in the past, before which it was small so that coalescence events happened fast. In our sample of seven sequences from the tree in Figure 2, six are singletons while only one, that shared by sequences C and D, occurs in two of the seven sequences. This excess of singletons is a signature of a star-like gene genealogy, suggesting that an episode of rapid population growth generated the data. Another interpretation is that a new selectively advantageous variant occurred about 200 000 years ago and spread rapidly and that today's sequences are all derived from it. In general, we cannot distinguish between an episode of rapid population growth and a selective sweep of a new advantageous variant: both of these have been proposed to account for the clear star-like genealogical history of human mitochondrial DNA.

There is another property of the diversity portrayed in Figure 2 that is useful to compute. Imagine that there were many more mutations on the tree of Figure 2 that occurred at random along the branches. If we were to tabulate all possible pairwise sequence differences, they would all be similar because of the star-like structure of the tree. In Figure 1, sequences C and D are similar since not much time separates them, while in Figure 2 the difference between C and D is not very different from other pairwise differences. A histogram of pairwise differences is called a mismatch distribution, and a smooth unimodal mismatch distribution is a signature of a comb or star-like tree as in Figure 2.

Figure 3 shows a coalescence tree from a locus that has been under balancing selection. Selection that somehow favors a variant as it becomes less common can maintain separate gene lineages for much longer than they would persist if they were neutral. Figure 3 might represent, for example, the history of several human

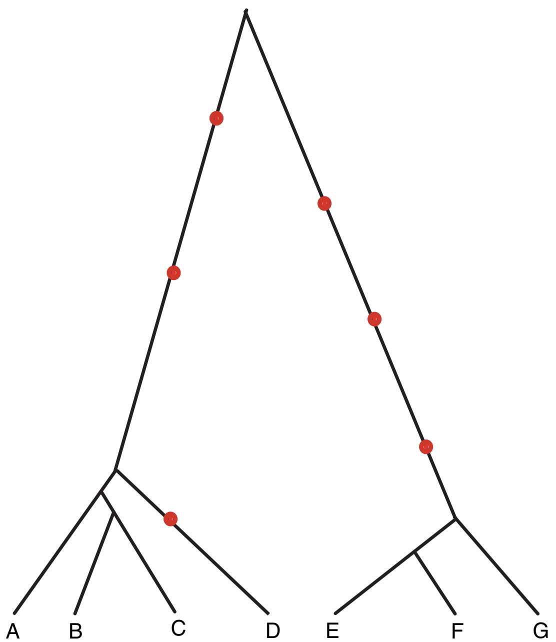


Figure 3 Gene tree showing the history of seven genes sampled from a population as in Figure 1. At this locus balancing, selection maintains both lineages in the population so the coalescence is very old. For human HLA system genes, for example, the top of the tree may be several tens of millions of years old

HLA alleles: in this system, uncommon variants are favored and alleles have deep gene trees, on the order of tens of millions of years.

Six random mutations are shown on this coalescence tree. Most of them are on the long deepest branches and divide the sample into 3 and 4: one mutation is present as a singleton, in chromosome D. The frequency spectrum has a "hump in the middle" in this case since there are more segregating sites with alleles of intermediate frequency than standard neutral theory predicts (*see* Article 7, **Genetic signatures of natural selection**, Volume 1).

There are standard statistics, provided by well-tested computer packages, that describe and quantify these insights, but their evaluation often involves complicated numerical calculations or even simulations. The simple graphical arguments given here are the bases for the more formal statistics.

References

ALFRED (2004) **http://alfred.med.yale.edu**.

Cavalli-Sforza LL, Menozzi P and Piazza A (1994) *The History and Geography of Human Genes*, Princeton University Press: Princeton.

Cochran G and Harpending H (2004) *The Natural History of Ashkenazi Intelligence*, **http://harpend.dsl.xmission.com/Documents/Ashkenazi.IQ.pdf**.

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J and Harpending HC (1997) Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 3100–3103.

Nei M (1987) *Molecular Evolutionary Genetics*, Columbia University Press: New York.

Relethford JH and Harpending HC (1994) Craniometric variation, genetic theory, and modern human origins. *American Journal of Physical Anthropology*, **95**, 249–270.

Rogers AR and Jorde LB (1996) Ascertainment bias in estimates of average heterozygosity. *American Journal of Human Genetics*, **58**, 1033–1041.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Introductory Review

Human cytogenetics and human chromosome abnormalities

Terry Hassold
Washington State University, Pullman, WA, USA

1. Introduction

Modern human cytogenetics can be divided into three distinct eras, on the basis of the technologies that were available at the time. The first of these, the "golden era" extended from the late 1950s to about 1970. During this time, methodology was introduced, that made it possible to obtain karyotypes from a variety of tissues; most importantly, peripheral blood became amenable for study, making it possible to rapidly obtain cytogenetic information from virtually anyone. This quickly led to the identification of several common chromosomal syndromes, including Down syndrome, Klinefelter syndrome, and Turner syndrome.

These early studies were hindered by the inability to unequivocally identify each human chromosome, a limitation that was overcome by the introduction of chromosome banding techniques in the early 1970s. Thus, the second era of human cytogenetics, the "era of banding", began. Banding techniques made it possible to detect small structural chromosome rearrangements, such as deletions, duplications, or translocations that previously would have gone undetected. As a result, several new syndromes associated with deletions or partial trisomies were identified in the 1970s. Further, the ability to band chromosomes resulted in the emergence of important clinical applications of chromosome testing, for example, prenatal cytogenetic analysis and cancer cytogenetics. Initially restricted to laboratories in a few academic institutions, these fields quickly grew and are now common practice in clinical cytogenetic laboratories throughout the world.

The third era, "the era of molecular cytogenetics", began in the 1980s and continues today. It represents the fusion of conventional cytogenetic with molecular methodologies, and has led to the precise characterization of a number of new chromosomal disorders, including imprinting and other genomic disorders. The techniques that comprise molecular cytogenetics are discussed in detail in other chapters in this section (*see* Article 12, **The visualization of chromosomes**, Volume 1, Article 22, **FISH**, Volume 1, and Article 23, **Comparative genomic hybridization**, Volume 1). In the remainder of this chapter, we summarize the incidence and types of chromosome abnormalities that have been identified using these methodologies. We will focus on the first types of abnormalities

that were identified (i.e., conventional numerical and structural abnormalities), as several of the more recently identified types of abnormalities (microdeletion syndromes, imprinting disorders) are discussed in later chapters (*see* Article 17, **Microdeletions**, Volume 1 and Article 19, **Uniparental disomy**, Volume 1).

2. The incidence and types of human chromosome abnormalities

The normal diploid number of chromosomes in humans is 46, consisting of 22 pairs of autosomes and one pair of sex chromosomes (XX for females, XY for males). Because even the smallest autosome contains at least 200–300 genes, and imbalance of even a small number of genes is known to have profound effects on normal development, it might be expected that any kind of chromosome abnormality is problematic. Indeed, with rare exceptions this is the case and, considered as a class, chromosome abnormalities are the leading known cause of pregnancy loss and congenital birth defects in our species. In the following summary, we discuss the incidence and types of chromosome abnormalities, first focusing on abnormalities detectable on conventional cytogenetic analysis and subsequently briefly considering those detectable with specialized methodology.

2.1. "Old" chromosome abnormalities: numerical and "conventional" structural abnormalities

Early cytogenetic studies made it clear that a surprisingly high proportion of human conceptions are chromosomally abnormal. The vast majority of these are *numerical*, that is, involving too many or too few chromosomes, while a smaller proportion are *structural*, that is, resulting from breakage and reunion of chromosome segments.

There are four common types of numerical abnormalities in humans: trisomy, monosomy, triploidy, and tetraploidy, which cumulatively represent the leading genetic cause of mental retardation, congenital births defects, and pregnancy loss in our species.

Trisomy (the presence of an additional chromosome, leading to a diploid number of 47 rather than 46) is clinically the most important of these categories. Approximately, 0.3% of all liveborn infants are trisomic, with the most common single abnormality being trisomy 21, the chromosome complement responsible for most cases of Down syndrome. Occurring in approximately 1/600–1/800 liveborn infants, Down syndrome is associated with moderate to severe mental retardation and with a constellation of characteristic phenotypic features; by itself, it is the leading genetic cause of mental retardation in humans.

Sex chromosome trisomies are the next most common trisomies, resulting in individuals with 47,XXX, 47,XYY, or 47,XXY (Klinefelter syndrome) chromosome complements. Of these abnormalities, Klinefelter syndrome is the most serious clinically, as it is associated with infertility and frequently, mild cognitive or behavioral abnormalities.

Other trisomies are rare, owing to the fact that the phenotypic abnormalities prevent survival to term. However, they are extremely common among clinically recognized spontaneous abortions (miscarriages), accounting for at least 25% of all such conceptions. Trisomies for most chromosomes have been identified in spontaneous abortions, although there is enormous variation in the frequency of the different trisomies. For example, trisomy 16, the most commonly identified trisomy, accounts for approximately one-third of all trisomies identified in abortions and, together with trisomies 21 and 22, for one-half of all trisomies. Other commonly observed trisomies include trisomies 2, 13, 14, and 15. In contrast, certain trisomies (e.g., 1, 5, 11, and 19) are virtually nonexistent, presumably because they house genes that are vital for embryonic/fetal development.

Because of their clinical importance, a great deal of attention has been given to the genesis of human trisomies – how they originate and factors that may influence their likelihood of occurrence. Using DNA polymorphisms to trace the inheritance of the extra chromosome, studies of the last decade have made it clear that most trisomies originate from meiotic nondisjunction – that is, errors in cell division that lead to the presence of an extra (or missing) chromosome in the gametes (eggs or sperm). While errors in both oogenesis and spermatogenesis have been identified, meiotic divisions in the development of the egg appear to be the most susceptible to nondisjunction. Indeed, over 90% of all human trisomies appear to be attributable to maternal meiotic nondisjunction, with errors at the first meiotic division (MI) being the predominant source of trisomies. While the reason for this propensity to error remains unknown, it is likely due to the unusual biology of MI in the egg: the division is initiated before birth in the fetal female ovary, but arrests until the time of ovulation some 10–15 to 40–50 years later, at which time it is finally reinitiated and completed.

While a great deal is now known about the parent and meiotic origin of human trisomies, much less is known about factors that influence their frequency. One factor stands out: the association between increasing maternal age and trisomy, which is arguably the most important etiologic factor in human genetic disease. Among women under the age of 25, approximately 2% of all clinically recognized pregnancies are trisomic; by the age of 36, however, this figure increases to 10% and by the age of 42, to over 33%. This association between maternal age and trisomy is exerted without respect to race, geography, or socioeconomic factors. Thus it appears to be "hard-wired" into our species, although the evolutionary benefit that this provides, and its underlying molecular basis, remain obscure.

Until recently, increasing maternal age nondisjunction was the only factor known to influence the frequency of nondisjunction. However, studies of the past decade have finally detected a second contributing factor: abnormal meiotic recombination. Recombination is a process that occurs during the MI, involving pairing, synapsis, and exchange (crossing-over) between homologous pairs of chromosomes; it is discussed in more detail in another chapter (*see* Article 13, **Meiosis and meiotic errors**, Volume 1). Recombination results in the exchange of genetic material between homologs, thus acting to increase genetic diversity within a species. However, it has a second, sometimes overlooked, function. Specifically, the physical exchange of material serves to tether the homologs together during prophase of

meiosis I, thus facilitating their alignment on the metaphase I plate and increasing the chances of proper segregation at the first division.

From studies of model organisms, it had long been known that alterations in recombination increased the likelihood of nondisjunction. With the advent of DNA polymorphism analysis in the 1990s, it became possible to study the effects of recombination on human nondisjunction, and to ask whether, as in model organisms, human trisomies were associated with altered recombination. The results have been compelling. For all human trisomies yet studied, aberrant meiotic recombination is a major contributing factor. To date, three types of abnormal recombination patterns have been identified: (1) complete failure of recombination between homologs, leaving the homologs to wander at random at MI, (2) recombinational events in which the homologs are held together only at the tips of the chromosomes, presumably leading to "weak" connections that are unable to sufficiently bind the homologs together, and (3) recombinational events that occur close to the centromeres of the chromosomes, possibly leading to inter-homolog connections that are too "strong", preventing them from separating from one another at the first division.

The demonstration of a link between abnormal meiotic recombination and trisomies has provided the first molecular correlate for human nondisjunction. Intensive efforts are now underway to determine why such abnormal recombinational patterns occur in the first place, and whether these patterns may be responsible for some proportion of maternal age-dependent human trisomies.

Monosomy (the absence of a single chromosome, leading to a diploid number of 45 rather than the normal 46) is sometimes thought of as the "flip side" of trisomy. That is, a nondisjunctional event should lead to a gamete with a missing chromosome, as well as a gamete with an additional chromosome. Thus, it might be expected that monosomies and trisomies would occur in equal frequency. However, this turns out not to be the case, as monosomies are exceedingly rare. Indeed, only one monosomy, sex chromosome monosomy (45,X; the chromosome complement associated with Turner syndrome), is identified in any frequency. It is identified in approximately 1/20 000 newborns, who develop as phenotypic females, but with short stature, streak gonads, and infertility. For reasons that are not clear, the 45,X chromosome complement is much more frequent in miscarriages where it accounts for nearly 10% of all such conceptions.

Why are monosomies so rare? Likely, they occur at appreciable frequency but are eliminated very early in gestation because of the detrimental effects of haplosufficiency, that is, the presence of only one, rather than two, copies of a gene(s). Haploinsufficiency for a single gene is frequently associated with fetal demise or severe birth defects; therefore, it is not surprising that haploinsufficiency for an entire chromosome would be incompatible with embryonic development. Indeed, the fact that 45,X conceptions frequently survive long enough to be detected as clinically recognizable miscarriages or, much more rarely, as liveborn infants, is presumably attributable to X-chromosome inactivation (*see* Article 15, **Human X chromosome inactivation**, Volume 1). This is the process whereby one of the two X chromosomes in normal female cells is silenced, leaving only one functional, expressed copy for most X-linked genes. As no such process exists for autosomes, autosomal monosomies are much more lethal than is the 45,X condition.

Triploid conceptions have an additional whole haploid set of chromosomes, that is, 69 rather than 46 chromosomes. They are almost never identified in liveborn infants but are extremely common in miscarriages, accounting for nearly 10% of such conceptions. Triploidy can result from one of two processes: (1) failure of a meiotic division, resulting in a gamete with 46 chromosomes, for example, failure of MI in oogenesis, followed by fertilization by a normal sperm, leads to triploidy of maternal origin or (2) penetration of a normal egg by two sperm ("dispermy"), yielding triploidy of paternal origin.

Both types of triploidy have been identified but, intriguingly, the phenotypes are different: for example, paternal triploids are characterized by the overgrowth of the placenta, while maternal triploids are not. This "parent-of-origin" difference in phenotype is an example of an imprinting effect, which is discussed in more detail later (*see* Article 19, **Uniparental disomy**, Volume 1).

Tetraploid conceptions have two extra sets of chromosomes, that is, 92 rather than 46 chromosomes. These are identified only in miscarriage samples and apparently almost always result from failure of an early mitotic cleavage division.

2.1.1. Structural chromosome abnormalities

Structural rearrangements involve breakage and reunion of chromosomal material. Although much less common than numerical abnormalities, they have a special clinical significance since they can be inherited through generations. That is, for some types of structural rearrangements, clinically normal individuals may carry the rearrangement in "balanced" form but transmit an "unbalanced" form to their progeny, leading to phenotypic abnormalities. This is in contrast to numerical abnormalities, which arise de novo in meiosis or in an early mitotic division, and thus are not passed on from generation to generation.

Rearrangements may be interchromosomal (involving exchange of material between different chromosomes) or intrachromosomal (involving the reordering, loss or gain of genetic material on a single chromosome).

The most important classes of interchromosomal rearrangement are translocations, which are categorized as reciprocal or Robertsonian. Reciprocal translocations simply refer to situations in which material is exchanged between different chromosomes, resulting in "hybrid" chromosomes containing segments of each of the original nontranslocated chromosomes. By themselves, they do not cause problems, since the individual carrying them may have the same genetic material as a chromosomally normal individual. However, there are at least two ways in which translocations may result in phenotypic abnormalities: (1) the breaks that led to the translocation events may be complex, leading to loss of material surrounding the break points and consequently to genetic imbalance and (2) during meiosis in individuals heterozygous for translocations, complex structures link the translocated chromosomes with their structurally normal homologous partners, making it possible that the wrong combination of translocated and nontranslocated chromosomes will pass together into the gametes. This results in genetic imbalance for the translocated portions, which, depending on the length of the translocated segments, can have disastrous phenotypic consequences.

For the most part, reciprocal translocations occur randomly throughout the genome, that is, there are no “hot spots” for reciprocal exchanges. However, there is one notable exception to this rule: the 11;22 translocation. First described in 1980, the translocation involves the exchange of material between the long arm of chromosome 11 (at q23) and the long arm of chromosome 22 (at q11). Balanced carriers of the 11;22 translocation are themselves clinically normal, but if their progeny inherit the translocation chromosome 22 as an “extra” chromosome, they manifest the “der 22 syndrome”, with a distinctive set of abnormal phenotypic features (e.g., with abnormalities of the ear, heart defects, cleft palate, and mental retardation).

Unlike other reciprocal translocations, the points of exchange are identical among the several hundred different families in which the 11;22 translocation has been identified. Thus, the 11;22 translocation is a rare example of a recurrent reciprocal translocation, involving regions of chromosomes 11 and 22 that are susceptible to rearrange with one another. Recent molecular analyses of these regions has elucidated the reason for this: the breakpoints on each chromosome are located within similar AT-rich palindromic repeats, creating cruciform DNA structures that presumably facilitate the illegitimate transfer of material between the nonhomologous chromosomes during meiosis.

Robertsonian translocations are a special class of reciprocal translocation, in which the long arms of two acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22) are fused, creating a chromosome that contains virtually all of the genetic material of the original two chromosomes. If the Robertsonian translocation is present in unbalanced form, a monosomic or trisomic conception results. For example, approximately 3% of cases of Down syndrome are attributable to unbalanced Robertsonian translocations, most often involving chromosomes 14 and 21. In this instance, the affected individual has the normal number of chromosomes (46) but in addition to two structurally normal chromosomes 21, carries a translocation 14/21 chromosome. This results in triplication for genes on the long arm of chromosome 21, leading to phenotypic abnormalities identical to those associated with conventional trisomy 21. Similarly, unbalanced Robertsonian translocations involving chromosome 13 account for a small proportion of individuals with trisomy 13 syndrome.

In addition to interchromosomal rearrangements, there are several types of intrachromosome structural abnormalities, including inversions (in which a segment of the chromosome is “flipped”, so that the orientation of genes is in the opposite orientation from that on a structurally normal chromosome), isochromosomes (in which one part of the chromosome is a mirror image of another), rings (in which breaks at both ends of the chromosome are sealed together, forming a circular chromosome), duplications (in which a segment of the chromosome is present in two or more copies), and deletions (in which a segment is lost). Of these, deletions are clinically the most significant. Two of the best characterized deletions, involving loss of material on chromosome 4p and 5p, result in Wolf–Hirschhorn syndrome and cri-du-chat syndromes, respectively. In both instances, the deletion involves the loss of a relatively small chromosomal segment; nevertheless, both are associated with multiple congenital anomalies, profound mental retardation, and reduced life span.

2.2. *"New" chromosome abnormalities: microdeletion and imprinting syndromes*

In 1986, Roy Schmickel coined the term "contiguous gene syndromes" to describe genetic disorders that mimicked single-gene disorders but that resulted from the deletion of a small number of tightly clustered genes. It was proposed that many such deletions would be so small as to be undetectable cytogenetically, that is, they would be "microdeletions". Over the past two decades, the increasing application of molecular techniques has led to the identification of over 20 of these microdeletion syndromes; many of these have parent-of-origin effects ("imprinting disorders"). Several of these are discussed in later chapters (*see* Article 17, **Microdeletions**, Volume 1 and Article 19, **Uniparental disomy**, Volume 1), and will not be considered in detail here. However, in the following discussion, we focus on two types of these disorders, emphasizing some of the general principles that have emerged from their analysis: a microdeletion syndrome (velocardiofacial syndrome, VCF) and two complementary disorders exhibiting parent-of-origin effects (Angelman and Prader–Willi syndromes).

Deletions of chromosome 22 (at 22q11) are among the most common microdeletions yet studied, as they are detected in approximately 1/3000 newborns. The most common syndrome associated with these deletions, VCF syndrome, includes a number of phenotypic defects, for example, learning disabilities or mild mental retardation, palatal defects, and congenital heart defects. Occasionally, individuals with a microdeletion of 22q11 deletion are more severely affected and are diagnosed as having DiGeorge syndrome, which involves abnormalities in the development of the third and fourth branchial arches, leading to thymic hypoplasia, parathyroid hypoplasia, and conotruncal heart defects. In approximately 30% of these cases, a deletion at 22q11 can be detected with high-resolution banding; by combing conventional cytogenetics, fluorescence in situ hybridization (FISH) (*see* Article 12, **The visualization of chromosomes**, Volume 1 and Article 22, **FISH**, Volume 1) and molecular detection techniques (i.e., Southern blotting or polymerase chain reaction analyses), these rates improve to over 90%. Additional studies have demonstrated a surprisingly high frequency of 22q11 deletions in individuals with nonsyndromic conotruncal defects. Approximately 10% of individuals with a 22q11 deletion inherited it from a parent with a similar deletion.

Prader–Willi syndrome (PWS) and Angelman syndrome (AS) both involve abnormalities of chromosome 15. In 1981, David Ledbetter described a series of patients with PWS, a proportion of whom had a cytogenetically detectable deletion between 15q11 and 15q13. Subsequently, a number of AS patients were observed to have a deletion in the same region. This was curious, because the syndromes are very dissimilar – PWS is characterized by obesity, hypogonadism, and mild to moderate mental retardation, while AS is associated with microcephaly, ataxic gait, seizures, inappropriate laughter, and severe mental retardation. The resolution to this question came some years later, when it was recognized that the parental origin of the deletion determines which phenotype ensues, that is, if the deletion is paternal, the result is PWS and if the deletion is maternal, the result is AS. However, an additional complication to this story arose when it was recognized that not all individuals with PWS or AS carry the deletion; in fact, in both syndromes, many

of the patients have two normal chromosomes 15. For most such individuals, it turns out that the parental origin of the chromosomes 15 is again the important determinant. In PWS, nondeletion patients invariably have two maternal and no paternal chromosomes 15, while for most nondeletion AS patients the reverse is true. This indicates that at least some genes on chromosome 15 are differentially expressed, depending on which parent contributed the chromosome. Additionally, this means that normal fetal development requires the presence of one maternal and one paternal copy of chromosome 15. Errors resulting in a deviation from this situation (e.g., fertilization of an egg with a 24, X,+15 karyotype by a sperm with a 22, X,−15 karyotype) will result in an abnormal phenotype, either PWS or AS.

Chromosomes that behave in this manner are said to be imprinted, and it is likely that several – although certainly not all – human chromosomes are imprinted. Chromosome 11 must be one of these, since it is now known that a proportion of individuals with the overgrowth syndrome Beckwith–Wiedeman syndrome have two paternal, but no maternal, copies of this chromosome.

In addition to microdeletion and imprinting syndromes, there is now at least one well-described microduplication syndrome. This is Charcot-Marie-Tooth syndrome Type 1A (CMT1A), a nerve conduction disease previously thought to be inherited as a simple autosomal dominant disorder. Molecular studies conducted in the 1990s demonstrated that affected individuals are heterozygous for a duplication of a small region of chromosome 17 (17p11.2-12). This makes the reason for the inheritance pattern clear: one-half of the offspring of affected individuals would inherit the duplication-carrying chromosome.

3. Summary

Chromosome abnormalities occur with astonishing frequency in our species, with an estimated 20–30% of all fertilized eggs having too many or too few chromosomes, or carrying structurally abnormal chromosomes. In the following chapters, we discuss the techniques that have led to the identification of these disorders, and describe the phenotypic outcomes associated with different classes of chromosome abnormality. Additionally, for those abnormalities in which the information is available, the molecular basis for the abnormalities is described. However, it will quickly become apparent that for most abnormalities, this information is simply not there. One of the major challenges in human genetics is the identification of the molecular bases for these conditions, and the identification of etiological factors that influence their occurrence.

Introductory Review

The visualization of chromosomes

Stuart Schwartz
University of Chicago, Chicago, IL, USA

Chromosomes have been both understood and misunderstood for over the past century. As early as 1882, Walther Flemming was believed to have started the field of human cytogenetics by describing chromatin (chromosome) as the rod-shaped bodies that he microscopically saw in the nucleus of cells. He also coined the term mitosis as the division of cells. Both Theodor Boveri and Walter Sutton in 1903 independently hypothesized that units of heredity were carried on these (yet undefined) chromosomes. Knowing the importance of the human chromosome, the next logical step was to understand the true chromosome number in man. In 1912, Von Winiwater stated that he believed that women had 47 chromosomes and men had 48 chromosomes. And T. S. Painter, in 1923, showed that the "true chromosome" number was indeed 48. Given the crude equipment and methodologies that these investigators had at their disposal, it is laudatory that they were indeed so close.

The chromosome number in humans remained at 48 from 1923 to 1956; however, even during that time it became recognized that changes in chromosome number was probably responsible for phenotypic abnormalities. Several authors in the 1930s, including L. S. Penrose, believed that a change in chromosome number was the underlying cause of Down syndrome. In 1956, Tjio and Levan eloquently described the chromosome number in humans as 46 (Tjio and Levan, 1956). In their paper in *Hereditas*, they stated that "Before a renewed, careful control has been made of the chromosome number in spermatogonial mitoses of man we do not wish to generalize our present findings into a statement that the chromosome number of man is $2n = 46$, but it is hard to avoid the conclusion that this would be the most natural explanation of our observations." – a remarkable understatement for such a landmark finding. This finding did not occur in a vacuum as many advances in cytogenetics were used to make the chromosomes more readable. This included the use of colchicine as a mitotic inhibitor, the utilization of a fixative and, most notably, the development of a hypotonic solution, by T. C. Hsu in 1956, to help spread out chromosomes, making them more visible.

Up until this time, chromosome analysis was done using tissue preparations, such as testicular material from deceased prisoners that was used by Painter, or fetal lung fibroblasts as used by Tjio and Levan. Most of the work in 1956 consisted of cell culture establishment, colchicine pretreatment, hypotonic solution pretreatment, and

squashing. However, the cell culture technology was an impediment to the better visualization of chromosomes, and as pointed out in 1932 by J. B. S. Haldane, “A technique for the counting of human chromosomes without involving the death of the person concerned is greatly to be desired”. In 1960, Nowell and Hungerford published their work on short-term cultures using lymphocyte cultures, which were preferable to long-term cultures. These were effective due to the discovery of phytohemagglutinin, extracted from the common navy bean, which allowed for the stimulation and growth of lymphocytes in culture. This work allowed for a glorious period of human cytogenetics in which a number of syndromes were shown to be chromosomal in origin including the cause of Down syndrome by Lejeune, Trisomy 13 by Patau, Trisomy 18 by Edwards, as well as Turner syndrome (Ford), and Klinefelter syndrome (Jacobs and Strong) (Lejeune *et al*., 1959; *see also* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1). However, at this point, all of the analysis and visualization was done using solidly stained chromosomes and their limited differentiation through the differences and size and centromere placement.

The next major breakthrough in the visualization of chromosomes came from the differential staining of chromosomes. In 1970, Caspersson and his colleagues published a landmark paper describing the identification of human chromosomes by DNA-binding fluorescent agents. Caspersson was not a cytogeneticist, but used his knowledge of chemistry to ultimately develop the first method of identifying each individual chromosome (Caspersson *et al*., 1970; Caspersson, 1989). This work utilized a fluorescent microscope and was difficult for many labs to utilize. Therefore, when Seabright (1972) published her paper on the rapid use or trypsin for producing a banding pattern that could be visualized by light microscopy, this technology was readily used by most laboratories to both visualize and characterize chromosomes. Many other banding technologies were developed and utilized to better understand chromosome structure and to define chromosome abnormalities. These include, but are not limited to, C-banding, G11-banding, NOR staining, R-banding, T-banding, and Cd-banding. While these were highly useful when developed, most have been superceded by a variety of DNA probes used with fluorescence in situ hybridization (FISH), as will be described below. The development of banding techniques again revitalized the field of cytogenetics and allowed for the visualization and characterizations of syndromes involving deletions (Cri-du-Chat; Wolf–Hirschhorn), duplications, reciprocal translocations (9/22 translocations), Robertsonian translocations, and inversions (recombinant 8 syndrome).

Since all of the individual chromosomes could now be visualized, it became obvious that at times two pairs of chromosomes showed different numbers of bands, which could not be accounted for the differential condensation of the chromosomes. If chromosomes were visualized in prometaphase rather than metaphase, it produced a higher resolution and allowed for the detection of abnormalities not seen in routine chromosomal analysis. Yunis first described this methodology in 1976 and continued to illustrate its usefulness in clinical cytogenetics (Yunis, 1976; Yunis *et al*., 1978; Yunis and Chandler, 1978). While the initial methodology focused on synchronizing the cell cycle in cultures using methotrexate, many different methodologies were used to either synchronize cultures (e.g., use of

excess TdR) or to keep the chromosomes from condensing (e.g., the addition of ethidium bromide). Many submicroscopic deletions that are studied extensively today using FISH were initially described using high-resolution methodologies. These include the deletions seen in Prader–Willi and Angelman syndrome, the deletion in DiGeorge syndrome (*see* Article 17, **Microdeletions**, Volume 1), and the duplication in Beckwith–Wiedermann syndrome. This new methodology allowed for a higher resolution analysis of chromosomes, which most laboratories ultimately developed. New abnormalities and new syndromes were described, but cytogenetics mostly went into a period of senescence after the adaptation and utilization of this methodology.

Even after the introduction of high-resolution methodologies, it had become more apparent that there was a limitation to the resolution of chromosomes microscopically. In order to produce more resolution, many investigators tried to utilize a technique pioneered by Pardue and Gall in 1969, in which they utilized radioactive DNA for direct hybridization to chromosomes (Pardue and Gall, 1969). Their work was very successful as they utilized repetitive DNA for this hybridization. Other workers had varying success, and the technology was used in many instances for the chromosomal localization of known genes. However, this work was very time-consuming and tedious and was not ready to be used routinely by clinical laboratories. Simultaneously, both Pinkel and his colleagues and Ward and his coworkers were developing similar nonradioactive fluorescence technology. Pinkel *et al*.'s initial paper in 1986 opened the doors for the use of FISH for the study of chromosomes, although it took many years for laboratories to incorporate this technology into their routine work (Pinkel *et al*., 1986). Since its initial use, FISH has changed the face of cytogenetics, invigorated the field, and no laboratory can be considered complete without using the technology routinely (*see* Article 22, **FISH**, Volume 1).

Briefly speaking, FISH utilizes a DNA probe that has been labeled to hybridize to denature DNA on a slide. The initial probes were indirectly labeled with a hapten and then detected fluorescently. However, with advances in technology, almost all FISH is currently done using direct fluorescent labels. FISH can be utilized to study metaphase chromosomes or DNA in interphase cells whether from cultures or paraffin sections. As long as the DNA, that is to be studied, is not degraded, it can be analyzed using FISH. This technology has a phenomenal range in that not only can metaphase and interphase cells be studied but essentially any tissue with a nucleus can also be analyzed. Constitutionally, FISH is routinely used in interphase cells to rapidly detect aneuploidy and in metaphase cells to detect subtle deletions. The microdeletions (e.g., Prader–Willi syndrome), described above, that could previously only be detected with good high-resolution banding can now easily be detected with appropriate FISH probes. FISH has also been extremely useful in the delineation of structural abnormalities such as marker chromosomes, duplications, or derivative chromosomes by the simple use of or combination of alpha-satellite probes and/or chromosomes paints. More recently, a set of chromosome-specific subtelomere probes have been developed and used to detect cryptic rearrangements that have been missed with routine banding. Many of these rearrangements (either simple deletions or derivative chromosomes) have been too small to detect with routine banding, but contain enough loss or gain of genetic material to lead to

phenotypic anomalies. It is estimated that between 2 and 6% of individuals with mental retardation/dysmorphic features without a recognizable G-banded anomaly may have a cryptic rearrangement. Thus, it is apparent that our resolution (and visualization) of chromosomes have been increased 20-fold from about 3 Mb with the analysis of banded chromosomes to approximately 150 kb with directed FISH analysis looking for specific DNA alterations.

This increased resolution is also important and easily seen with somatic changes in chromosomes associated with cancer (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1). For these specimens, the development of probes has mainly involved resources for the detection of specific rearrangements associated with a variety of leukemias and lymphomas (*see* Article 24, **Cytogenetic analysis of lymphomas**, Volume 1). Again, as for constitutional abnormalities, these probes can be utilized in both metaphase and interphase preparations. Thus, while a bcr-abl fusion in chronic myelogenous leukemia can easily be detected routinely in metaphase G-banded chromosomes, FISH can be utilized in interphase cells (when appropriate metaphase preparations are not available) to detect the fusion. Similarly, probes for the TEL-AML1 fusion are routinely used to detect this change as the underlying 12/21 translocation cannot usually be detected analyzing the banded chromosomes. These are just two examples showing how FISH has changed, how chromosomes are studied, and what can be detected. More specific examples of its use can be seen throughout this book.

Most of the FISH analysis described above and throughout the rest of the chapters in this book has been possible because of the advances in technologies and the development of a series of commercial probes for these FISH studies. Most work in clinical laboratories, whether doing subtelomeric analysis or studying specific translocations in cancer, is done with prepared commercial probes. However, the study of the Human Genome has been a veritable boom for cytogeneticists. The resources produced in the elucidation of the genome allows for the study of any region of chromosomes. As almost the entire genome has been sequenced, there are BAC probes readily available, for any location in the genome, to delineate any structural change or better describe a chromosome anomaly. These probes have now easily been used to show cryptic deletions and the complexity of changes in chromosomes and have immensely improved our detection of chromosome changes. Again, however, most of these analyses need to be used as a directed study of abnormalities in that specific areas can be studied as desired. However, it is the development of these resources that has led to the next milestone in cytogenetics, the utilization of comparative genomic hybridization with BAC arrays to detect and study chromosome anomalies. Comparative genomic hybridization has been utilized for many years (*see* Article 24, **Cytogenetic analysis of lymphomas**, Volume 1). In its inception, it is a technology where you take DNA from a specimen you are interested in studying, label that DNA, and then hybridize to normal metaphase chromosomes. Utilizing this technology, you could determine if there were any excess or deficiencies in the DNA sample studied. While this is a good technique in most hands, its resolution was limited to about 10 Mb, not necessarily adding any resolution to routine banding. However, with the delineation of the genome, the DNA of interest can be labeled and hybridized against a BAC array. This array can

contain BACs that are 1 MB apart, allowing for a resolution of this magnitude, not just for an area of interest but for the entire human karyotype (Fiegler *et al*., 2003; Greshock *et al*., 2004; Bignell *et al*., 2004; Ishkanian *et al*., 2004). While this work is still in the research stage, the potential it can provide is enormous and soon will allow visualization of the chromosomes past the bands and to the DNA backbone. These advances will not only allow us to have greater visualization of chromosomes but also allow for a better understanding of chromosome abnormalities over the next 100 years.

References

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, **14**(2), 287–295.

Caspersson TO (1989) The William Allan memorial award address: the background for the development of the chromosome banding techniques. *American Journal of Human Genetics*, **44**(4), 441–451.

Caspersson T, Zech L, Johansson C and Modest EJ (1970) Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma*, **30**(2), 215–227.

Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IP, *et al*. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes, Chromosomes & Cancer*, **36**(4), 361–374.

Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M and Weber BL (2004) 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Research*, **14**(1), 179–187.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**(3), 299–303.

Lejeune J, Turpin R and Gautier M (1959) Chromosomic diagnosis of mongolism. *Archives Francaises de Pediatrie*, **16**, 962–963.

Pardue ML and Gall JG (1969) Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, **64**(2), 600–604.

Pinkel D, Straume T and Gray JW (1986) Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **83**(9), 2934–2938.

Seabright M (1972) Human chromosome banding. *Lancet*, **7757**, 967.

Tjio JH and Levan A (1956) The chromosome number in man. *Hereditas*, **42**, 1–6.

Yunis JJ (1976) High resolution of human chromosomes. *Science*, **191**(4233), 1268–1270.

Yunis JJ and Chandler ME (1978) High-resolution chromosome analysis in clinical medicine. *Progress in Clinical Pathology*, **7**, 267–288.

Yunis JJ, Sawyer JR and Ball DW (1978) The characterization of high-resolution G-banded chromosomes of man. *Chromosoma*, **67**(4), 293–307.

Specialist Review

Meiosis and meiotic errors

Maj Hultén and Maira Tankimanova

University of Warwick, Coventry, UK

Hazel Baker

University of Warwick, Coventry, UK

University Hospitals Coventry, and Warwickshire NHS Trust, Coventry, UK

1. The general scheme of meiosis

The meiotic process involves two cell divisions, and is accordingly divided into two parts, meiosis I and meiosis II (Figure 1). As in mitosis, these divisions are subdivided into the stages Prophase, Metaphase, Anaphase, and Telophase, and thus a full meiotic process includes two of each of these stages. The meiotic process has been recently reviewed by Page and Hawley (2003) and Tease and Hultén (2003). The salient points of each stage are highlighted in Table 1.

Following DNA replication, cells enter meiosis, with each chromosome being composed of two chromatids held together at the centromere. Meiosis I includes a prolonged Prophase in preparation for the halving of chromosome number, this "reductional" division taking place at Anaphase I (AI). Prophase I (PI) is by tradition subdivided into Leptotene, Zygotene, Pachytene, and Diplotene, with an intermediate Diakinesis stage before Metaphase I (MI).

The preparation for the MI/AI spindle to carry its cargo of whole chromosomes (rather than chromatids as in mitosis) takes place at Prophase I and involves intimate pairing of all four chromatids of the maternal and paternal homologs. This intimate pairing (synapsis) is mediated by a proteinaceous structure, called the Synaptonemal Complex (SC). Breakage and reunion (crossing-over) between nonsister chromatids gives rise to configurations known as chiasmata. The chiasmata hold homologous chromosomes together locally during the next phase, when they separate at Diplotene. Following chromatin condensation during the intermediate Diakinesis stage, bivalents align on the MI plate.

It is generally accepted that at least one so-called obligate chiasma is necessary for a normal segregation to take place. The number of additional chiasmata required to be formed to ensure mechanical stability is associated with the length of the chromosomes (Hultén, 1974; Lynn *et al*., 2002; Tease *et al*., 2002; Tease and Hultén, 2004). Most importantly, mechanical stability in this context also involves appropriate orientation of maternal and paternal kinetochores (proteinaceous spindle

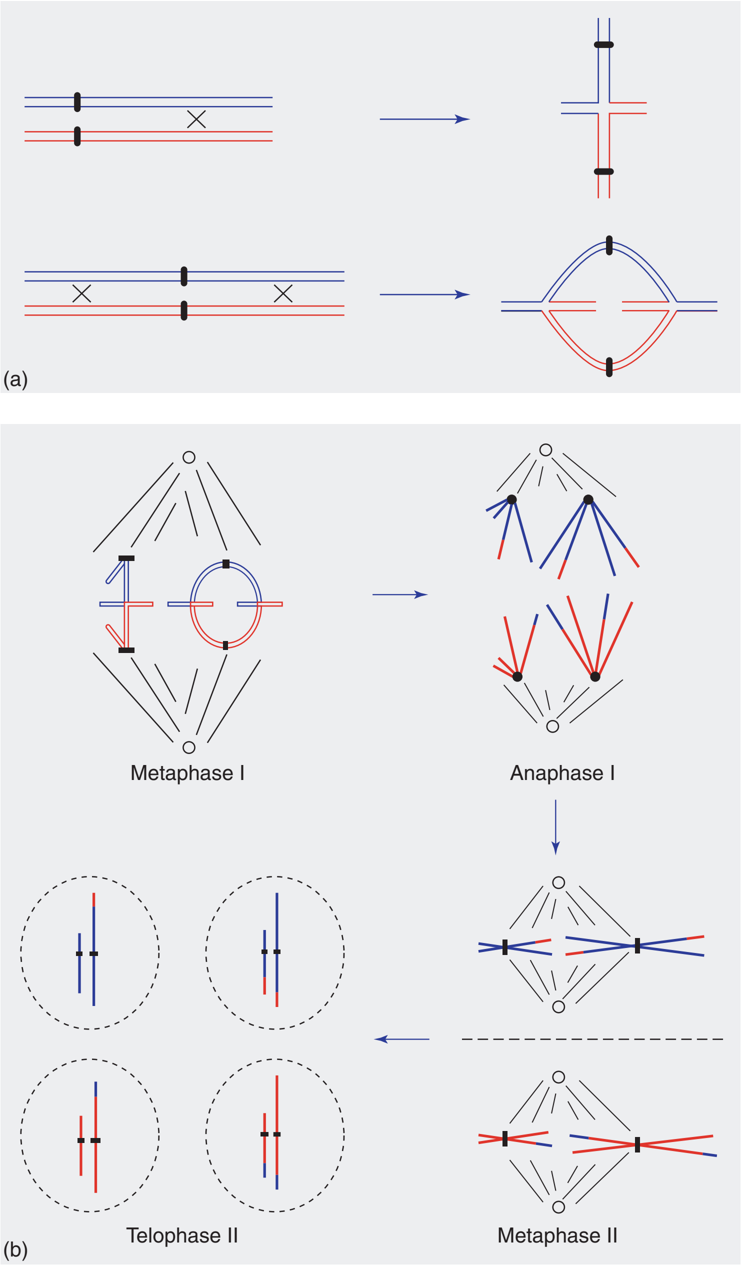


Figure 1 Schematic illustration of meiosis with (a) homologous chromosome synapsis, and crossing-over at the pachytene stage of prophase I and the derivative bivalents at MI, and (b) progression through metaphase I to anaphase I, metaphase II to anaphase II, and telophase II showing the four potential haploid gametes

Table 1 Salient points of meiotic stages

Stage	Subdivision	Features
Prophase 1		DNA replication of each chromosome to give sister chromatids held together at the centromere
	Leptotene	Start of chromosome condensation, formation of the lateral elements of the synaptonemal complex
	Zygotene	Chromosome pairing; completion of the formation of the synaptonemal complex
	Pachytene	Chromosomes fully paired, crossovers/chiasmata fully established
	Diplotene	Initiation of separation of the synapsed chromosomes
	Diakinesis	Chromosomes condense and become fully separated, except at points of crossing-over/chiasma formation.
Metaphase 1		Homologous chromosomes align on the equatorial plate with the kinetochores being the point of attachment to the spindles
Anaphase 1		Reductional division; homologous pairs separate, but sister chromatids remain together
Telophase 1		Formation of two daughter cells
Prophase 2		Nuclear envelope dissolves; creation of a new spindle
Metaphase 2		Chromosomes align on the spindle
Anaphase 2		Separation of centromeres; migration of sister chromatids to opposite poles
Telophase 2		Further cell division resulting in four potential haploid gametes from each parent cell

attachments at centromeres) normally facing opposite spindle poles at MI. This is facilitated by the topology of bivalents, where, for example, single-chiasma bivalents appear as cross configurations, doubles as rings, and triples as figures-of-eight. It is important to note that at AI (in contrast to the situation at mitosis) centromeres of homologs are kept together. The chance selection of the paternal and maternal kinetochores of any one bivalent to be presented to the opposite spindle poles leads to their random assortment in daughter cells. Each crossover event and chiasma formation give rise to a recombinant together with a nonrecombinant chromatid, which together are passed on to the daughter cells at AI, providing the basis for additional genetic variation in offspring.

After a brief PII without DNA replication, the two daughter cells complete meiosis II to produce four end products. This "equational" division is similar to mitosis. It should be recognized though that normally at least one of the sister chromatids (potential gametes) of each chromosome now is recombinant, containing a combination of variant genes (alleles) from the maternal and paternal homologs (Figure 2).

2. Variation between the two sexes

Interestingly, there are many differences in behavior between male and female germ cells during gametogenesis and meiosis. This includes the timing and continuity of events, the pairing and crossover processes, chromosome segregation, and the actual

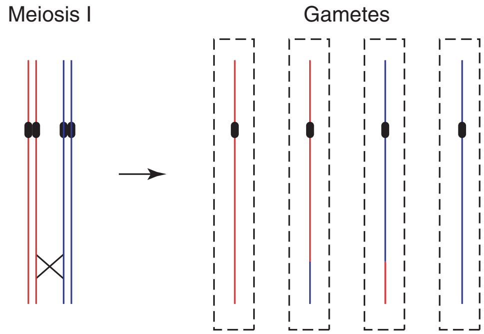


Figure 2 Schematic illustration of the relationship of a crossover, as viewed directly (as a chiasma or MLH1 focus) and the meiotic products (gametes). There are four chromatids, each a potential gamete, only two of which are recombinant

gamete produced – two cells as diverse as the oocyte and the sperm could hardly be imagined to derive from the same process! These differences are summarized in Table 2, and some of the more interesting aspects are discussed in more detail below.

2.1. *Timing of meiosis*

Meiosis in males starts after puberty and continues throughout life. On the other hand, meiosis in females begins around the 12th week of fetal life with homolog pairing, crossing-over, and chiasma formation up until around week 20. Oocytes arrest at the Diplotene stage, and are generally thought to undergo extensive cell death with numbers dropping from around 7 million to around 2 million at birth (Baker, 1963; Tilly, 2001). Meiosis resumes after puberty, when after the midcycle Luteinizing Hormone surge, chromosomes line up on the MI plate and undergo the reductional division at AI just before ovulation. This division is asymmetrical, where most of the cytoplasm is retained by one of the daughter cells, which will form the future mature oocyte. The other daughter cell forms a small polar body, which soon degenerates. Following progression through the second prophase, oocytes then arrest at MII until fertilization occurs. At fertilization, once the sperm has entered the oocyte and caused activation, oocyte meiosis continues through AII and TII, resulting in the ejection of the second polar body.

2.2. *The synaptonemal complex*

Meiotic chromosome pairing at PI involves three successive developmental stages, homolog recognition, presynaptic alignment, and intimate synapsis. This precisely timed process is mediated by the proteinaceous SC. The SC forms a protein scaffold for meiotic chromosome pairing taking place during PI, that is, during the substages Leptotene, Zygotene, and Pachytene. The SC then breaks down at Diplotene.

Table 2 Summary of the main differences between male and female meiosis

Male	Female
Meiosis is a continuous process from puberty throughout life, and can result in the production of around 200–300 million spermatozoa daily. A full cycle of spermatogensis takes approximately 120 days, of which 72–74 days are spent during meiosis.	Meiosis can take over 40 years from start to finish and only a few oocytes actually progress to the final stages, most being lost before birth.
Each parent cell produces 4 gametes (spermatozoa), all divisions being equal.	The two cell divisions are unequal with most cytoplasm retained in the oocyte and only a minor part forming the first and second polar bodies, thus only one actual gamete is produced from each parent cell.
The testis contains a population of stem cells that give rise to the continual supply of gametes.	Oocyte numbers appear to be limited to those present at birth with around 350 ovulating between puberty and the menopause. However in mice, recent experimentation has suggested the existence of stem cells that may constitute a reserve (Johnson *et al*., 2004).
Chromosome synapsis is initiated very near the ends of chromosomes in karyotypically normal, fertile men.	Interstitial initiation of synapsis is common in females.
Obligate chiasma formation is an efficient process.	Poor efficiency of obligate chiasma formation.
Lower overall chiasmata frequency.	Higher overall chiasma frequency.
Synaptonemal complexes are condensed.	Synaptonemal complexes are relatively decondensed, having almost twice the total length per cell compared to that in males.
Tendency of chiasmata to occupy preferential positions with hotspots near the ends of chromosomes.	Chiasmata occupy more interstitial positions.

The SC is a protein lattice that resembles railroad tracks and connects paired homologous chromosomes in most meiotic systems. The two side rails of the SC, known as lateral elements, are connected by proteins known as transverse filaments (reviews in Moens *et al*., 2002; Page and Hawley, 2004).

The SC was first discovered in mice and was later found to be evolutionarily conserved (albeit with species- and sex-specific modifications). Many antibodies against the proteins involved are available, providing tools for SC research using immunofluorescence techniques. Fluorescence in situ hybridization (FISH) methods have also been used sequentially to highlight the behavior of some of the SC proteins involved and their interaction with chromatin components during the meiotic chromosome pairing stages in human spermatocytes and oocytes (Barlow and Hultén, 1996; Barlow and Hultén, 1997; Barlow and Hultén, 1998a; Codina-Pascual *et al*., 2004; Hartshorne *et al*., 1999; Prieto *et al*., 2004; Sun *et al*., 2004; Lenzi *et al*., 2005).

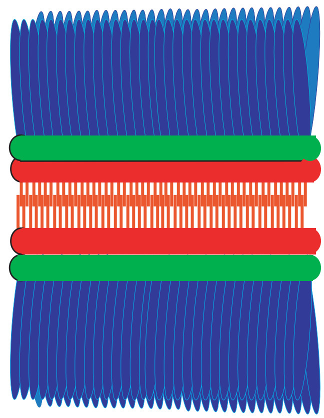


Figure 3 Schematic illustration of the SC

Even before meiosis is initiated, some "meiosis-specific" proteins, members of the cohesin complex (such as REC8 and STAG3) appear as diffuse aggregations within the cell nuclei. Subsequently, these cohesins function to stick chromatids together by interacting with the chromatin (DNA and proteins) of chromosomes to form initial cohesion axes.

During the initial development of the SC, most of the DNA form large loops emanating from the sides of the SC in a "chromatin cloud" (Figure 3). The chromatin is attached to the initially unpaired axial, and later, paired lateral elements of the SC by the cohesin proteins and the protein SYCP3 acting as "glue" between chromatids. These cohesion axes become "sandwiched" between chromatin and the SC axial elements marked by SYCP3. Pairing is subsequently stabilized at intimate synapsis, with transverse filaments developing between the lateral elements of the SC, forming a central element composed of proteins including SYCPI.

2.3. Chiasma formation

It is now generally accepted that initiation of crossing-over and chiasma formation in mammals takes place via double strand breaks (DSBs) induced by the protein SPO11 at Leptotene; and a range of proteins are involved in subsequent synapsis and recombination events (Bannister and Schimenti, 2004; Roig *et al*., 2004; Svetlanov and Cohen, 2004; Lenzi *et al*., 2005). Numerous DSBs (visualized by the γ-H2XA antibody) are initially created in DNA loops located outside the SC. Only a minority of this DNA subsequently becomes tethered to the core of the SC, exposing them to the recombinatorial machinery, associated with successive conversion and crossover events mediated by proteins such as RAD51 and MLH1 (Barlow *et al*., 1997).

The application of anti-MlH1 (MutL homolog1) to human SC preparations shows a labeling pattern consisting of distinct foci, always precisely associated with the SC and never in closely juxtaposed pairs (Figure 4). The frequency distribution of male MLH1 foci agrees with that of chiasmata at Diakinesis/MI (Figure 5); and

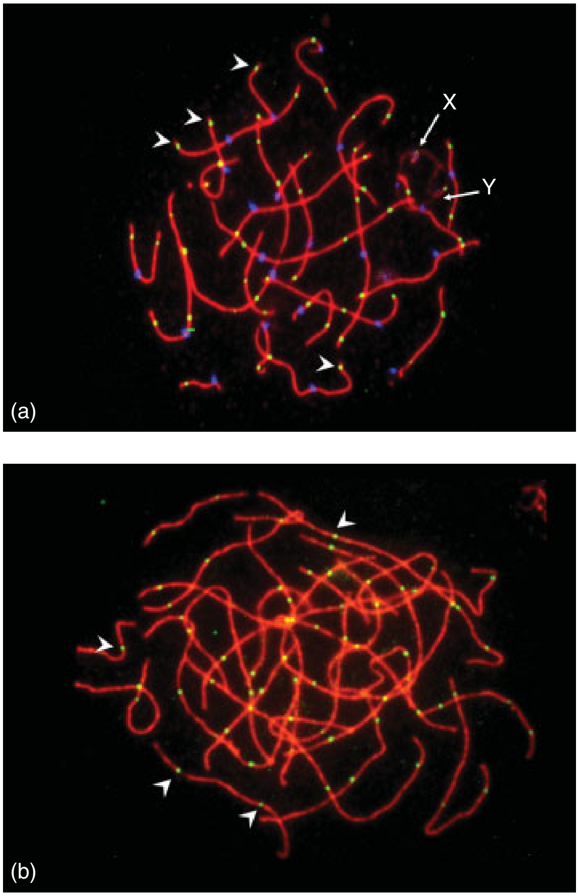


Figure 4 MLH1 foci on SCs of spermatocyte (left) and oocyte (right). Note the near-telomere preference in the spermatocyte and the higher number of foci in the oocyte

there is now no doubt that MLH1 is an appropriate marker for chiasma formation (Marcon and Moens, 2003).

The average crossover/recombination frequency estimated from MLH1 analysis in human males is around 50 with a range of 40–60 (Barlow and Hultén, 1998b; Lynn *et al*., 2002; Tease *et al*., 2002; Hultén and Tease, 2003a; Sun *et al*., 2004; Tease and Hultén, 2004), while that of females is around 70 but with a much larger variation between individuals than in males (Barlow and Hultén, 1998b; Tease *et al*., 2002; Roig *et al*., 2004; Tease and Hultén, 2004; Lenzi *et al*., 2005).

2.4. *Patterns of chiasma distribution*

A large database of patterns of chiasmata has accumulated over the last three decades by the study of Diakinesis/MI spermatocytes in preparations from testicular biopsy samples in human males, demonstrating that chiasma formation is a dynamic

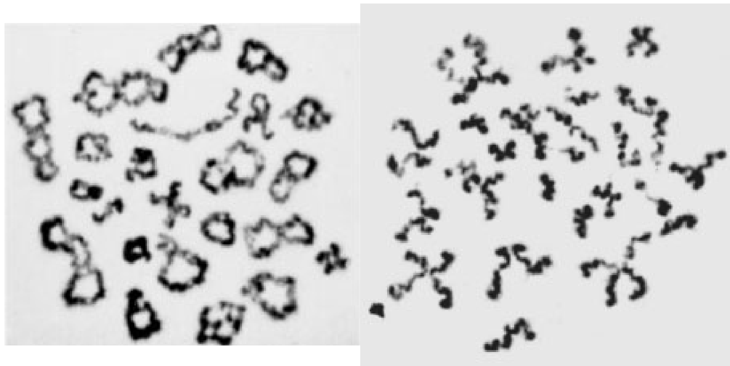


Figure 5 The morphology of MI and MII chromosomes is very different

process, varying between chromosomes, cells, and subjects (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1 and Article 20, **Cytogenetics of infertility**, Volume 1). We have quite a good knowledge on the basic patterns and its variation between individual chromosomes, cells, and subjects in normal fertile men, as well as changes associated with infertility, and in carriers of structural chromosome rearrangements (review in Hultén and Tease, 2003a,b; Oliver-Bonet *et al*., 2004). On the other hand, our knowledge on the underlying mechanisms of regulation of this basic pattern and its variations is still very poor. For technical reasons, the same type of information on chiasmata at the Diakinesis/MI stage is not available in human females.

One important role of chiasma formation is to link chromosome pairs in such a way that whole chromosomes are conveniently transported to daughter cells at AI. Chiasma formation allows maternal and paternal homologs to be presented together to the MI-AI spindle so that daughter cells may receive a combination of these chromosomes rather than exclusively maternal or paternal homologs. Together with the random presentation of paternal and maternal kinetochores to opposite spindle poles, this provides the basis for the "infinite" genetic variability of offspring (*see* Article 16, **Nondisjunction**, Volume 1). The 3D topology of bivalents at this stage means that chiasmata are locked in their original positions owing to the behavior of bivalents during the transition from Pachytene through Diplotene and Diakinesis to MI (Hultén, 1990). This means that chiasmata do not need to "terminalize" in order to resolve. Thus, the long-held belief, originally proposed by Darlington, that chiasmata must terminalize in order to resolve is not a valid proposition.

3. Obligate chiasma formation and implications of chiasma failure

In the normally fertile human male, one chiasma is generally formed per bivalent, irrespective of its length; this chiasma is considered to be obligate to secure "regular" segregation of the chromosomes involved, by allowing the

maternal and paternal chromosomes of the bivalent to be properly orientated on the MI spindle. Failure of chiasma formation means that maternal and paternal homologs appear as disoriented univalents, leading to random rather than regular segregation. Daughter cells may thus receive a paternal, a maternal, both, or none of these (nonexchanged) chromosomes. One other complication of this situation is the potential for chromatids of univalent chromosomes to undergo precocious separation followed by disoriented segregation, leading to daughter cells receiving one, two, or no chromatids, instead of a whole (maternal, paternal, or recombinant maternal–paternal) chromosome. Univalent chromosomes are very rarely seen in male meiosis either during Pachytene (Barlow and Hultén, 1998b; Lynn *et al*., 2002) or at Diakinesis/MI (Hultén, 1974; Laurie and Hultén, 1985).

Efficient obligate chiasma formation in the human male stands in sharp contrast to the situation in females. Little information exists on patterns of chiasmata at the MI stage in human oocytes, however, recent SC and MLH1 observations on human fetal oocytes quite clearly demonstrate failure of synapsis and chiasma formation in a large proportion at the Pachytene stage (Tease *et al*., 2002).

4. Additional chiasmata, interference, and crossover hotspots

Numbers of additional chiasmata over and above the obligate are dependent on chromosome length (Hultén, 1974; Lynn *et al*., 2002; Tease *et al*., 2002; Tease and Hultén, 2004). Successive chiasmata are separated by large (many Mb) chromosome segments, a phenomenon termed chiasmata interference (Figure 6). Measurements of distances between MLH1 crossover foci on the SCs at Pachytene indicate interference distances in terms of physical SC length to be basically the same in human spermatocytes and oocytes (Tease and Hultén, 2004). This may "explain" the higher chiasma frequency in human females in comparison to males because the SCs in oocytes are relatively "decondensed" and much longer than in spermatocytes (Wallace and Hultén, 1985).

Chiasmata have a tendency for sex-specific preferential positioning (Figure 7) at crossover hotspots, occupying large (many Mb) stretches of chromosomes, which we may term megahotspots. A different type of preferential crossover positioning

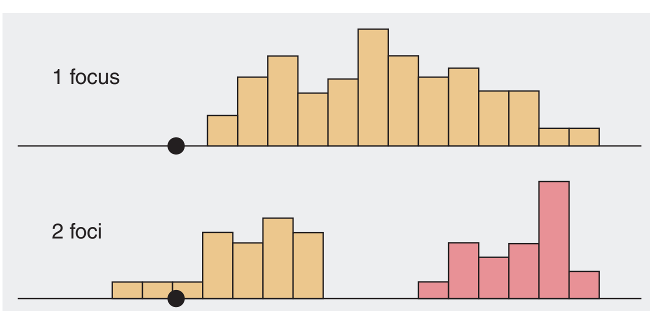


Figure 6 The distribution patterns of MLH1 foci on chromosome 21 SCs in oocytes illustrating the difference with one in comparison to two foci, which are widely spaced apart

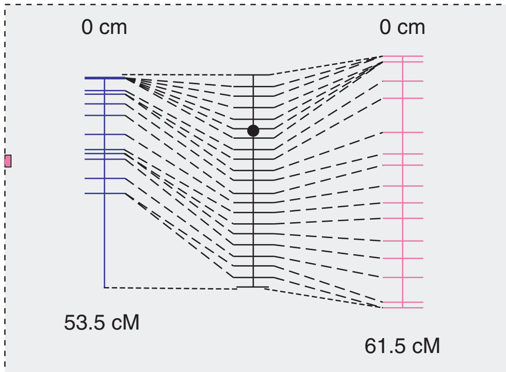


Figure 7 Genetic maps of chromosome 21 illustrating high amount of distal crossovers and the corresponding expansion of genetic length distally in the male (blue) in comparison to that in the female (red)

has more recently been revealed by DNA analysis of sperm. Comparison of certain DNA markers in somatic tissue (using e.g., blood samples) and sperm has demonstrated crossover hotspots stretching over a few kilobasepairs separated by longer noncrossover, so-called haploblock intervals, in the order of 50–100 kb. Such fine-scale hotspots were originally discovered because they may occur in the vicinity of hypervariable DNA sequences (minisatellites) that have been used extensively for "DNA fingerprinting" in forensics (Jeffreys *et al*., 1985). It has been suggested that these types of hotspots, which we may refer to as minihotspots, represent a general situation of "punctuate" preferential crossover positioning across the whole genome (review in Jeffreys *et al*., 2004). The relationship of these minihotspots to the megahotspots identified by MLH1 and chiasma analysis remains unknown, and requires further study.

5. Construction of genetic maps

Genetic maps illustrate the frequency of crossovers that on average take place along the length of each chromosome arm. The construction of genetic maps from crossover data obtained by cytogenetic analysis in individual cells has many advantages (review in Hultén and Tease, 2003b). First of all they allow sex-specific estimates of both intra- (intercellular) and interindividual variation in patterns of meiotic recombination with respect to the whole genome, individual chromosomes, and chromosome segments. They also distinguish between reciprocal recombination (crossing-over/chiasma formation) and other types of recombination events (such as sister chromatid exchange and conversion). In addition, the direct meiotic crossover analysis allows estimates of genetic map distances as well as recombination fractions from the same observed raw data.

Haldane (1919) originally defined the Morgan unit of genetic map distance as that length of a chromatid that has experienced on average one crossover per chromatid. Each chromatid in this situation corresponds to a potential gamete

(Figure 2). As each crossover event (MLH1 focus/chiasma) may give rise to two recombinant as well as two nonrecombinant gametes, the genetic map distance (Morgan, M) is calculated as half the average number of MLH1 foci or chiasmata in the interval concerned. In other words, the genetic map distance in centimorgans (cM) is obtained by multiplying the average number by 50, thus an average of 50 autosomal chiasmata in spermatocytes corresponds to a male genetic map length of 2500 cM, while 70 MLH1 foci in oocytes translate to a female genetic map length of 3500 cM. As already mentioned, these averages, however, hide a substantial interindividual variation in crossover frequency. Effectively, we all have our own average genetic maps with further individual specificity of the particular sperm and egg transmitted to offspring.

Assuming there is no chromatid interference but successive crossover chromatids are chosen by chance, then the corresponding recombination fraction is calculated as half the proportion of cells having one or more MLH1 focus/chiasma in the interval. Most human genetic maps have been constructed by tracing DNA markers through generations where the raw data are recombination fractions between markers (e.g., Kong *et al*., 2002; Cullen *et al*., 2002). This approach has the advantage that recombination data are linked to genes or DNA sequences. On the other hand, relatively large sample sizes of families/sibships are required, and importantly the analysis recovers only 50% of recombinants from each parental crossover event and only 25% of doubles. Assumptions are also required on the risk of double crossovers within the interval studied. Double crossovers involving the same DNA strand would cancel each other out, a phenomenon that is taken into account and compensated for by the use of the so-called mapping functions, estimating crossover/chiasma interference and thus the risk of double crossovers. Earlier linkage-based genetic maps showed large discrepancies in relation to those based on cytogenetic data with a general tendency for inflation of linkage-based genetic map lengths (likely due to overcompensation for presumed double recombinants) (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1, Article 9, **Genome mapping overview**, Volume 3, Article 15, **Linkage mapping**, Volume 3 and Article 67, **History of genetic mapping**, Volume 4). The most recent ones, however, are very similar, at least as regards total genetic map length for each individual chromosome (Kong *et al*., 2002; Cullen *et al*., 2002; review in Hultén and Tease, 2003b). The high density of DNA markers now available should in our view make traditional mapping processes redundant. Previous complications in estimating genetic map distance by family studies due to misinformation as regards order of loci, should also now be much less of a problem, due to the detailed information on DNA marker order available from the Human Genome Project (Collins *et al*., 2004; *see also* Article 24, **The Human Genome Project**, Volume 3).

6. The transition from metaphase I to anaphase I

From Pachytene until MI, all four chromatids are “glued” together along their entire length (Figure 1). At MI, each pair of sister chromatids presents as a unit with both sister kinetochores facing in the same direction. This means that both sister

chromatids of a homolog are pulled to the same pole, once AI is initiated (review in Hauf and Watanabe, 2004). It appears from animal experimentation that AI chromosome segregation cannot start until all bivalents have orientated themselves on the MI plate with kinetochores of homologs facing opposite spindle poles, thus there may be a long delay before the onset of AI takes place. Once initiated however, this is seemingly a very rapid process, as indicated by, for example, the absence of any AI spermatocytes in preparations from human testicular biopsy samples.

A wealth of information about the complex interaction of proteins between kinetochores and spindle microtubules at the MI to AI transition has recently accumulated (reviewed in Nasmyth, 2002; Maiato *et al*., 2004; Hauf and Watanabe, 2004; Firooznia *et al*., 2005). A range of chromosome-associated proteins (either meiosis-specific or showing meiosis-specific behavior) have been identified including meiosis-specific cohesins, and centromere/kinetochore-associated proteins. The exactly timed behavior of these proteins ensures that normally all four chromatids are "glued together" until the onset of Anaphase I and retained within the centromere/kinetochore region until Anaphase II. It is also known that unattached kinetochores cause the delay of AI initiation through the action of several "wait" signals including Mad1/Mad2, Cdc20 and Bub1, localized to kinetochores (Howell *et al*., 2004; Vigneron *et al*., 2004; Shah *et al*., 2004; Taylor *et al*., 2004); subsequent attachment of these kinetochores then involves the activation of proteins forming the Anaphase Promoting Complex (APC). On the other hand, lapse of cohesion between chromatids, located peripheral to the chiasma, is obligatory to allow regular segregation of homologs at Anaphase I; and lapse of centromere/kinetochore-specific proteins is essential for regular Anaphase II segregation.

Activation of the APC in turn leads to Separase activation (Terret *et al*., 2003). Once activated, Separase is able to cleave cohesin complexes, allowing chromatids to separate along chromosome arms, initiating the MI-to-AI transition. It is important to note that only chromatid arm cohesion, necessary for separation of homologs is lost at AI. Sister chromatids remain connected via the action of amongst others, the cohesin REC8, STAG3, and Shugoshin at the centromere up until MII (Lee *et al*., 2003; Watanabe, 2004). It is not yet clear how the absolutely crucial protection of these cohesins from cleavage at centromeres during Meiosis I is regulated. Neither is it known why Meiosis I chromosome segregation is so error-prone in human females (see further below) in relation to males (and in relation to other mammalian species such as mice).

7. MII analysis for evaluation of AI segregation

Evaluation of the efficacy of AI segregation can be made by chromosome analysis of cells at the MII stage. Information in the male may be obtained by investigation of spermatocytes at MII, prepared from testicular biopsies in a way very similar to that used for making preparations from blood lymphocytes for somatic chromosome analysis. Cells are initially exposed to a hypotonic solution and then fixed in a (3:1) mixture of acetic acid and alcohol (Hultén *et al*., 2001). One particular problem here is that chromatids of individual MII chromosomes in spermatocytes prepared

this way often splay to the extreme and chromatids are often precociously separated. In addition they are loosely coiled and have a tendency to hook into each other (Figure 5). This makes even the counting of chromosomes problematic. Laurie *et al*. (1985) using stringent criteria to avoid artifacts did not detect any numerical abnormalities in 200 MII spermatocytes from six normal men with apparently normal mitotic karyotypes. Thus, there was no indication of either extra or missing chromosomes or indeed any extra or missing chromatids. To our knowledge no other studies have to date been presented on this issue. We may tentatively conclude that AI malsegregation in the human male is rare, and occurs with an incidence of less than 0.5–1% of cells.

This situation stands in sharp contrast to that in the human female. Chromosome analysis of a large population of oocytes at the MII stage has been performed. These studies demonstrate quite clearly that segregation errors involving both whole chromosomes and chromatids are common (review in Pellestor *et al*., 2003). It should be noted, however, that the material investigated so far consists almost exclusively of MII oocytes obtained at infertility (IVF) treatment, where spare oocytes, spontaneously arrested at MII, have been donated for research and used for karyotyping.

The morphology of oocyte MII chromosomes makes them much more amenable to karyotyping than is the case for spermatocytes; chromatids do not normally splay to the extent that they do in spermatocytes (Figure 5). Oocyte MII chromosomes are also more condensed and therefore more easily spread and separated from each other. A total of nearly 5000 human oocytes at the MII stage have been investigated with a view to obtain information on the efficacy of AI segregation as well as the occurrence of structural chromosome aberrations. Two types of methods have been used for making preparations including that of Kamiguchi *et al*. (1993), involving a gentle gradual fixation. Chromosome analysis has been performed by conventional (block staining as well as R-banding) methods, FISH for selected chromosomes and spectral karyotyping.

The largest study so far is that of Pellestor *et al*. (2002) involving 1397 MII oocytes, fixed by the Kamiguchi *et al*. (1993) technique and analyzed by R-banding. Numerical abnormalities were detected in around one-fifth (20.1%), mainly composed of extra or missing whole chromosomes or chromatids, while structural abnormalities (breaks, deletions and, acentric fragments) were much more rare, seen in 2.1% of cells. Numerical abnormalities caused by extra or missing chromatids were more common than aneuploidy of whole chromosomes, confirming the proposal by Angell that AI segregation errors of chromatids are common in human oocytes obtained through IVF studies (Angell, 1997). One very interesting observation in the Pellestor series is the strong positive correlation between maternal age and rate of aneuploidy, most pronounced as regards single-chromatids. The same has been seen in a small sample of fresh oocytes (not stimulated by hormone injections as part of IVF treatment) by spectral analysis (Sandalinas *et al*., 2002).

Another way of assessing AI segregation errors is by FISH analysis of the first polar body (review in Kuliev and Verlinsky, 2004). This has been achieved by micromanipulation of the polar bodies with subsequent FISH with 3–5 chromosome-specific probes as part and parcel of preimplantation genetic diagnosis (PGD) (*see* Article 21, **Preimplantation genetic diagnosis for chromosome**

abnormalities, Volume 1 and Article 22, **FISH**, Volume 1). Only the "aneuploidy-free" embryos have been transferred (following FISH analysis of the second polar body as well, as described further below). The largest study here is that summarized by Kuliev *et al*. (2003) involving 6733 oocytes from women above the age of 35 (average 38.5 years). In this series, a large proportion of oocytes were found to be aneuploid with respect to the chromosomes tested (chromosomes 21, 22, 13, 15, and 18). Thus, 41.7% of oocytes were considered to be aneuploid because of AI malsegregation, the majority again involving chromatids rather than whole chromosomes.

8. Gametic output

Investigations of chromosomes in mature gametes allow information to be obtained on the sum total of segregation errors including both AI and AII, as well as any structural chromosome aberrations that have occurred.

Chromosome analysis of mature gametes has primarily involved sperm (*see* Article 25, **Human sperm–FISH for identifying potential paternal risk factors for chromosomally abnormal reproductive outcomes**, Volume 1). Earlier studies were performed by the so-called humster technology (human sperm and hamster ovum pseudofertilization). This type of investigation is laborious and expensive, and has been restricted to a relatively few research groups. Guttenbach *et al*. (1997) reviewed observations by eight groups, involving over 20 000 sperm karyotypes from healthy donor men. Aneuploidy was detected in 1–3% of sperm. Another 5–10% showed structural chromosome abnormalities, many of which are presumed to have arisen postmeiotically, during differentiation of spermatids to mature sperm.

More recently FISH has been applied for analysis of selected chromosomes on very large populations of sperm from both chromosomally normal men and carriers of structural chromosome rearrangements. Shi and Martin (2000) have reviewed the experience gained on an impressive sample of more than 5 million sperm from around 500 normal fertile men, using one-, two-, or three-probe FISH. By and large, these studies have confirmed the findings using the humster technique that around 2–3% of mature spermatozoa are aneuploid.

Investigations of individual chromosomes indicate that their average numerical abnormality rate is around 0.1–0.2%. Only chromosomes 21 and 22 and the XY pair show slightly increased rates of sperm aneuploidy. The question whether some men may be predisposed to aneuploid offspring has not yet been answered with any certainty (Martinez-Pasarell *et al*., 1999; Soares *et al*., 2001a,b; Shi *et al*., 2001). On the other hand, tracing of DNA markers between fathers and XXY Klinefelter sons has demonstrated a reduced XY recombination, where the implication is that this has led to XY disomy in sperm (Thomas and Hassold, 2003; *see also* Article 19, **Uniparental disomy**, Volume 1).

FISH analysis of sperm from carriers of structural chromosome rearrangements, such as translocations, has shown a high rate of imbalance for the chromosomes involved (review in Shi and Martin, 2001; Oliver-Bonet *et al*., 2004).

As might be expected, there is much less direct cytogenetic information on gametic output including AII segregation errors in females. One intriguing way

of approaching this question is FISH analysis of not only the first polar body (as described above) but also the second polar body, obtained for analysis by micromanipulation at PGD treatment in pregnancies (review in Kuliev and Verlinsky, 2004). One example of this approach concerns the FISH study, summarized by Kuliev *et al*. (2003) on 6733 oocytes/fertilized eggs, where the conclusion was that AII segregation errors, as regards the chromosomes tested (chromosomes 21, 22, 13, 15, 16, and 18) had occurred in a slightly lower proportion of cases than AI errors, that is, in 30.7%. Remarkably, however, more than one-quarter were found to have both AI and AII segregation errors (27.6%).

9. Mechanisms of meiotic errors

Hardly anything is so far known from direct meiotic studies on the mechanisms of origin of structural chromosome rearrangements, including, for example, extra marker chromosomes, Robertsonian and reciprocal translocations, inversion, insertions, deletions, and duplications. Some of these may arise from breakage and repair processes taking place either pre- or postmeiosis. Special attention has been paid to the origin of de novo deletions and duplications, comprising the so-called genomic disease (Shaw and Lupski, 2004). The suggestion from somatic DNA marker investigations of children and their parents is that these chromosome disorders originate by misalignment of homologs at Meiosis I Prophase, followed by "ectopic" crossing-over/chiasma formation/recombination between similar (paralogous) DNA sequences, located at a distance from each other, either within the same homolog or indeed on different homologs. To our knowledge, no investigation has yet been performed to identify such ectopic crossovers directly by investigation of SCs. However, misalignment has been clearly documented in human oocytes, in which centromere signals (by the CREST antibody) are often staggered (Hartshorne *et al*., 1999).

Attention has focused on aneuploidy, which is a major cause of reproductive failure and congenital disease with special reference to common chromosome disorders such as Trisomy 21 Down, Trisomy 13 Patau and Trisomy 18 Edward Syndromes as well as the sex chromosome aneuploidies XXY Klinefelter, XYY, XXX, and X Turner syndromes. Much effort has been devoted to explaining in particular the maternal age effect especially noticeable as regards Trisomy 21 Down syndrome, this being the most common genetic disorder in humans.

It is known from previous indirect studies, tracking DNA polymorphisms between parents and children, that patterns of meiotic recombination (deciding the formation of bivalents and their shape) play an important role (Hassold and Hunt, 2001). However, it is not yet clear how this may work in detail. Most segregational errors underlying these aneuploidies have been found to take place at maternal AI; and it has, for example, been suggested that certain chiasma positions, such as a single distal chiasma of bivalent 21 may constitute a specific risk factor. On the other hand, the very same position is one of the most common in spermatocytes (Hultén, 1974; Laurie *et al*., 1985; Hultén and Tease, 2003b), where AI segregation errors are uncommon, and there is no known paternal age effect in this respect. Clearly a so-called second hit must be of paramount importance in handling such

oocyte bivalents differently from spermatocytes at a later stage. What would this be is yet unclear.

More detailed information on the intricate interaction between chromosomal and spindle proteins in oocytes at MI and AI in particular may provide some answers including an explanation for the well-known maternal age effect. Recent work indicates that knowledge obtained from the mitotic cell cycle may to some extent be extrapolated to meiosis, which opens up new avenues here. It has been suggested that age-related reduction in some proteins (such as the kinetochore-bound spindle checkpoint proteins Mad1 and Bub2 and/or the meiosis-specific cohesins) in combination with the specific spindle morphology and large size of the oocyte nucleus may play specific roles (review in Eichenlaub-Ritter, 2004). Quantitative analysis of the mRNA of some of these proteins has recently revealed an age-related degradation in human oocytes (Steuerwald *et al*., 2000). There is also a possibility that the morphology of the centromeres may contribute, where those with small amounts of specific (alphoid) centromeric DNA may be disadvantaged (Maratou *et al*., 2000). Furthermore, oocyte Anaphase I spindles have been shown to be barrel shaped rather than triangular as in spermatocytes. This may make the task of chromosome segregation more difficult. Remarkably, there are clear indications that Meiosis I spindles in oocytes obtained from older women are much more irregular than those from younger women (Battaglia *et al*., 1996; Volarcik *et al*., 1998).

Many factors may contribute to the formation of spindles including a range of spindle proteins, mitochondrial and hormonal status, follicle maturation in relation to the oocyte pool and peri-follicular microcirculation, and age-related changes in any of these factors may play a role.

Further reading

Hassold T, Judis L, Chan ER, Schwartz S, Seftel A and Lynn A (2004) Cytological studies of meiotic recombination in human males. *Cytogenetic and Genome Research*, **107**, 249–255.

Lynn A, Ashley T and Hassold T (2004) Variation in human meiotic recombination. *Annual Review of Genomics and Human Genetics*, **5**, 317–349.

Marston AL and Amon A (2004) Meiosis: cell-cycle controls shuffle and deal. *Nature Reviews. Molecular Cell Biology*, **5**, 983–997.

Pellestor F, Anahory T and Hamamah S (2005) The chromosomal analysis of human oocytes. An overview of established procedures. *Human Reproduction Update*, **11**, 15–32.

Wantanabe Y and Kitajima TS (2005) Shugoshin protects cohesin complexes at centromeres. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **360**, 515–521.

References

Angell R (1997) First-meiotic-division nondisjunction in human oocytes. *American Journal of Human Genetics*, **61**, 23–32.

Baker TG (1963) A quantitative and cytological study of germ cells in human ovaries. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **158**, 417–433.

Bannister LA and Schimenti JC (2004) Homologous recombinational repair proteins in mouse meiosis. *Cytogenetic and Genome Research*, **107**, 191–200.

Barlow AL, Benson FE, West SC and Hulten MA (1997) Distribution of the Rad51 recombinase in human and mouse spermatocytes. *The EMBO Journal*, **16**, 5207–5215.

Barlow AL and Hultén MA (1996) Combined immunocytogenetic and molecular cytogenetic analysis of meiosis I human spermatocytes. *Chromosome Research*, **8**, 562–573.

Barlow AL and Hultén MA (1997) Sequential immunocytogenetics, molecular cytogenetics and transmission electron microscopy of microspread meiosis I oocytes from a human fetal carrier of an unbalanced translocation. *Chromosoma*, **106**, 293–303.

Barlow AL and Hultén MA (1998a) Combined immunocytogenetic and molecular cytogenetic analysis of meiosis I oocytes from normal human females. *Zygote*, **6**, 27–38.

Barlow AL and Hultén MA (1998b) Crossing over analysis at pachytene in man. *European Journal of Human Genetics*, **6**, 350–358.

Battaglia DE, Goodwin P, Klein NA and Soules MR (1996) Influence of maternal age on meiotic spindle assembly in oocytes from naturally cycling women. *Human Reproduction*, **11**, 2217–2222.

Codina-Pascual M, Kraus J, Speicher MR, Oliver-Bonet M, Murcia V, Sarquella J, Egozcue J, Navarro J and Benet J (2004) Characterization of all human male synaptonemal complexes by subtelomere multiplex-FISH. *Cytogenetic and Genome Research*, **107**, 18–21.

Collins A, Lau W and De La Vega FM (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Human Heredity*, **58**, 2–9.

Cullen M, Perfetto SP, Klitz W, Nelson G and Carrington M (2002) High-resolution patterns of meiotic recombination across the major histocompatibility complex locus. *Human Genetics*, **71**, 759–776.

Eichenlaub-Ritter U, Vogt E, Yin H and Gosden R (2004) Spindles, mitochondria and redox potential in ageing oocytes. *Reproductive Biomedicine Online*, **8**, 45–58.

Firooznia A, Revenkova E and Jessberger R (2005) From the XXVII North American Testis Workshop: the function of SMC and other cohesin proteins in meiosis. *Journal of Andrology*, **26**, 1–10.

Guttenbach M, Engel W and Smith M (1997) Analysis of structural and numerical abnormalities in sperm of normal men and carriers of constitutional chromosome aberrations. *Human Genetics*, **100**, 1–21.

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked characters. *Journal of Genetics*, **8**, 299–309.

Hartshorne GM, Barlow AL, Child TJ, Barlow DH and Hultén MA (1999) Immunocytogenetic detection of normal and abnormal oocytes in human fetal ovarian tissue in culture. *Human Reproduction*, **14**, 172–182.

Hassold T and Hunt P (2001) To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews. Genetics*, **2**, 280–291.

Hauf S and Watanabe Y (2004) Kinetochore orientation in mitosis and meiosis. *Cell*, **119**, 317–327.

Howell BJ, Moree B, Farrar EM, Stewart S, Fang G and Salmon ED (2004) Spindle checkpoint protein dynamics at kinetochores in living cells. *Current Biology*, **14**, 953–964.

Hultén M (1974) Chiasma distribution at diakinesis in the normal human male. *Hereditas*, **76**, 55–78.

Hultén M (1990) The topology of meiotic chiasmata prevents terminalization. *Annals of Human Genetics*, **54**, 307–314.

Hultén MA, Barlow AL and Tease C (2001) Meiotic studies in humans. In *Human Cytogenetics: Constitutional Analysis (Third Edition): A Practical Approach*, Rooney DE (Ed.), Oxford University Press, pp. 211–236.

Hultén MA and Tease C (2003a) Genetic mapping: direct meiotic analysis. In *Nature Encyclopedia of the Human Genome 2003*, Vol. 2, Cooper DN (Ed.), Nature Publishing Group: London, pp. 882–887.

Hultén MA and Tease C (2003b) Genetic mapping: comparison of direct and indirect approaches. In *Nature Encyclopedia of the Human Genome 2003*, Vol. 2, Cooper DN (Ed.), Nature Publishing Group: London, pp. 876–881.

Jeffreys AJ, Holloway JK, Kauppi L, May CA, Neumann R, Slingsby MT and Webb AJ (2004) Meiotic recombination hot spots and human DNA diversity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **359**, 141–152.

Jeffreys AJ, Wilson V and Thein SL (1985) Hypervariable "minisatellite" regions in human DNA. *Nature*, **314**, 67–73.

Johnson J, Canning J, Kaneko T, Pru JK and Tilly JL (2004) Germline stem cells and follicular renewal in the postnatal mammalian ovary. *Nature*, **428**, 145–150.

Kamiguchi Y, Rosenbusch B, Sterzik K and Mikamo K (1993) Chromosomal analysis of unfertilized human oocytes prepared by a gradual fixation-air drying method. *Human Genetics*, **90**, 533–541.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, **31**, 241–247.

Kuliev A, Cieslak J, Ilkevitch Y and Verlinsky Y (2003) Chromosomal abnormalities in a series of 6,733 human oocytes in preimplantation diagnosis for age-related aneuploidies. *Reproductive Biomedicine Online*, **6**, 54–59.

Kuliev A and Verlinsky Y (2004) Meiotic and mitotic nondisjunction: lessons from preimplantation genetic diagnosis. *Human Reproduction Update*, **10**, 401–407.

Laurie DA, Firkett CL and Hultén MA (1985) A direct cytogenetic technique for assessing the rate of first meiotic non-disjunction in the human male by the analysis of cells at metaphase II. *Annals of Human Genetics*, **49**, 23–29.

Laurie DA and Hultén MA (1985) Further studies on bivalent chiasma frequency in human males with normal karyotypes. *Annals of Human Genetics*, **49**, 189–201.

Lee J, Iwai T, Yokota T and Yamashita M (2003) Temporally and spatially selective loss of Rec8 protein from meiotic chromosomes during mammalian meiosis. *Journal of Cell Science*, **116**, 2781–2790.

Lenzi ML, Smith J, Snowden T, Kim M, Fishel R, Poulos BK and Cohen PE (2005) Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis in human oocytes. *American Journal of Human Genetics*, **2005**, 112–127.

Lynn A, Koehler KE, Judis L, Chan ER, Cherry JP, Schwartz S, Seftel A, Hunt PA and Hassold TJ (2002) Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science*, **296**, 2222–2225.

Maiato H, Deluca J, Salmon ED and Earnshaw WC (2004) The dynamic kinetochore-microtubule interface. *Journal of Cell Science*, **117**, 5461–5477.

Maratou K, Siddique Y, Kessling AM and Davies GE (2000) Variation in alphoid DNA size and trisomy 21: a possible cause of nondisjunction. *Human Genetics*, **106**, 525–530.

Marcon E and Moens P (2003) MLH1p and MLH3p localize to precociously induced chiasmata of okadaic-acid-treated mouse spermatocytes. *Genetics*, **165**, 2283–2287.

Martinez-Pasarell O, Nogues C, Bosch M, Egozcue J and Templado C (1999) Analysis of sex chromosome aneuploidy in sperm from fathers of Turner syndrome patients. *Human Genetics*, **104**, 345–349.

Moens PB, Kolas NK, Tarsounas M, Marcon E, Cohen PE and Spyropoulos B (2002) The time course and chromosomal localization of recombination-related proteins at meiosis in the mouse are compatible with models that can resolve the early DNA-DNA interactions without reciprocal recombination. *Journal of Cell Science*, **115**, 1611–1622.

Nasmyth K (2002) Segregating sister genomes: the molecular biology of chromosome separation. *Science*, **297**, 559–565.

Oliver-Bonet M, Navarro J, Codina-Pascual M, Abad C, Guitart M, Egozcue J and Benet J (2004) From spermatocytes to sperm: meiotic behaviour of human male reciprocal translocations. *Human Reproduction*, **19**, 2515–2522.

Page SL and Hawley RS (2003) Chromosome choreography: the meiotic ballet. *Science*, **301**, 785–789.

Page SL and Hawley RS (2004) The genetics and molecular biology of the synaptonemal complex. *Annual Review of Cell and Developmental Biology*, **20**, 525–558.

Pellestor F, Andreo B, Arnal F, Humeau C and Demaille J (2002) Mechanisms of non-disjunction in human female meiosis: the co-existence of two modes of malsegregation evidenced by the karyotyping of 1397 in-vitro unfertilized oocytes. *Human Reproduction*, **8**, 2134–2145.

Pellestor F, Andreo B, Arnal F, Humeau C and Demaille J (2003) Maternal aging and chromosomal abnormalities: new data drawn from in vitro unfertilized human oocytes. *Human Genetics*, **112**, 195–203.

Prieto I, Tease C, Pezzi N, Buesa JM, Ortega S, Kremer L, Martinez A, Martinez-A C, Hultén MA and Barbero JL (2004) Cohesin component dynamics during meiotic prophase I in mammalian oocytes. *Chromosome Research*, **12**, 197–213.

Roig I, Liebe B, Egozcue J, Cabero L, Garcia M and Scherthan H (2004) Female-specific features of recombinational double-stranded DNA repair in relation to synapsis and telomere dynamics in human oocytes. *Chromosoma*, **113**, 22–33.

Sandalinas M, Marquet C and Munne S (2002) Spectral karyotyping of fresh, non-inseminated oocytes. *Molecular Human Reproduction*, **8**, 580–585.

Shah JV, Botvinick E, Bonday Z, Furnari F, Berns M and Cleveland DW (2004) Dynamics of centromere and kinetochore proteins; implications for checkpoint signaling and silencing. *Current Biology*, **14**, 942–952.

Shaw CJ and Lupski JR (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human Molecular Genetics*, **13**, Spec No 1, R57–R64.

Shi Q and Martin RH (2000) Aneuploidy in human sperm: a review of the frequency and distribution of aneuploidy, effects of donor age and lifestyle factors. *Cytogenetics and Cell Genetics*, **90**, 219–226.

Shi Q and Martin RH (2001) Aneuploidy in human spermatozoa: FISH analysis in men with constitutional chromosomal abnormalities, and in infertile men. *Reproduction*, **121**, 655–666.

Shi Q, Spriggs E, Field LL, Ko E, Barclay L and Martin RH (2001) Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24,XY human sperm. *American Journal of Medical Genetics*, **99**, 34–38.

Soares SR, Vidal F, Bosch M, Martinez-Pasarell O, Nogues C, Egozcue J and Templado C (2001a) Acrocentric chromosome disomy is increased in spermatozoa from fathers of Turner syndrome patients. *Human Genetics*, **108**, 499–503.

Soares SR, Templado C, Blanco J, Egozcue J and Vidal F (2001b) Numerical chromosome abnormalities in the spermatozoa of the fathers of children with trisomy 21 of paternal origin: generalised tendency to meiotic non-disjunction. *Human Genetics*, **108**, 134–139.

Steuerwald N, Cohen J, Herrera RJ and Brenner CA (2000) Quantification of mRNA in single oocytes and embryos by real-time rapid cycle fluorescence monitored RT-PCR. *Molecular Human Reproduction*, **6**, 448–453.

Sun F, Oliver-Bonet M, Liehr T, Starke H, Ko E, Rademaker A, Navarro J, Benet J and Martin RH (2004) Human male recombination maps for individual chromosomes. *American Journal of Human Genetics*, **74**, 521–531.

Svetlanov A and Cohen PE (2004) Mismatch repair proteins, meiosis, and mice: understanding the complexities of mammalian meiosis. *Experimental Cell Research*, **296**, 71–79.

Taylor SS, Scott MI and Holland AJ (2004) The spindle checkpoint: a quality control mechanism which ensures accurate chromosome segregation. *Chromosome Research*, **12**, 599–616.

Tease C, Hartshorne GM and Hultén MA (2002) Patterns of meiotic recombination in human fetal oocytes. *American Journal of Human Genetics*, **70**, 1469–1479.

Tease C and Hultén MA (2003) Meiosis. In *Nature Encyclopedia of the Human Genome 2003*, Vol. 3, Cooper DN (Ed.), Nature Publishing Group: London, pp. 865–873.

Tease C and Hultén MA (2004) Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenetic and Genome Research*, **107**, 208–215.

Terret ME, Wassmann K, Waizenegger I, Maro B, Peters JM and Verlhac MH (2003) The meiosis I-to-meiosis II transition in mouse oocytes requires separate activity. *Current Biology*, **13**, 1797–1802.

Thomas NS and Hassold TJ (2003) Aberrant recombination and the origin of Klinefelter syndrome. *Human Reproduction Update*, **9**, 309–317.

Tilly JL (2001) Commuting the death sentence: how oocytes strive to survive. *Nature Reviews. Molecular Cell Biology*, **2**, 838–848.

Vigneron S, Prieto S, Bernis C, Labbe JC, Castro A and Lorca T (2004) Kinetochore localization of spindle checkpoint proteins: who controls whom? *Molecular Biology of the Cell*, **15**, 4584–4596.

Volarcik K, Sheean L, Goldfarb J, Woods L, Abdul-Karim FW and Hunt P (1998) The meiotic competence of in-vitro matured human oocytes is influenced by donor age: evidence that folliculogenesis is compromised in the reproductively aged ovary. *Human Reproduction*, **1**, 154–160.

Wallace BM and Hultén MA (1985) Meiotic chromosome pairing in the normal human female. *Annals of Human Genetics*, **49**, 215–226.

Watanabe Y (2004) Modifying sister chromatid cohesion for meiosis. *Journal of Cell Science*, **117**, 4017–4023.

Specialist Review

Acquired chromosome abnormalities: the cytogenetics of cancer

Susanne M. Gollin
University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA

1. Introduction

Normal cellular growth is a delicate balance between cellular proliferation and cell death. Cancer is a disruption in this homeostasis, resulting in an imbalance favoring cellular proliferation over cell death. In the last several years, it has become clear that this homeostasis is controlled by genes (*see* Article 65, **Complexity of cancer as a genetic disease**, Volume 2). Thus, although some cases of cancer result from genetic predisposition due to germ-line genetic alterations that increase an individual's risk of developing cancer, cancer is a genetic disease of somatic cells (*see* Article 88, **Cancer genetics**, Volume 2). Both activation of oncogenes and loss (deletion, mutation, or inactivation) of tumor suppressor genes appear to be common in solid tumors, and altered regulation of oncogenes is prevalent in hematologic malignancies (Cavenee and White, 1995; Weinberg, 1996). This dysregulation is frequently caused by chromosomal abnormalities.

To describe the dysregulation of cancer genes, an automobile analogy serves us well. Oncogenes are dominant-acting cellular genes normally involved in the processes of cellular growth and proliferation. When one copy of an oncogene is mutated or otherwise deregulated, it functions in a manner similar to a "stuck accelerator pedal", resulting in uncontrolled cellular proliferation. Examples of oncogenes that are dysregulated in cancer cells include *CCND1* and *MYC* (Schwab, 1999). Tumor suppressor genes are recessive-acting "gatekeepers" of cellular proliferation; they may inhibit cell growth or promote cell death (Kinzler and Vogelstein, 1997). When both parental alleles of a tumor suppressor gene are mutated, lost, or otherwise inactivated in a somatic cell, it is as if the cell has "lost its brakes". Examples of such tumor suppressor genes altered in cancer cells are *TP53*, *RB1*, and *CDKN2A* (Sidransky, 1995). An additional family of genes critical to carcinogenesis is the DNA repair genes or "caretakers" of the genome (Kinzler and Vogelstein, 1997). These genes, including *ATM*, *MSH2*, and *MLH1*, maintain the "integrity" of the genome, playing key roles in DNA repair processes. Their role in cancer is comparable to experiencing a car problem when your mechanic was unavailable to repair it, thus leading to additional engine (DNA) damage.

Tumors develop as a result of deregulation of not one but multiple cancer-related genes. Alteration of three to six genes is thought to be necessary for tumor development. Mutations in additional genes result in increased cellular proliferation, disorganization, tumor progression, invasion, and metastasis (Vogelstein and Kinzler, 1993). Fearon and Vogelstein (1990) first proposed a linear pathogenetic schema for colon cancer, based on the concept that tumors develop and progress as a result of a distinctly ordered accumulation of tumor suppressor gene losses and oncogene activations, which endow a clonal population of cells with a proliferative advantage. The genetic changes in tumors may be conceptualized as a series of evolutionary events, which may have neutral, deleterious, or advantageous effects on the proliferation of a clone of cells. Neutral or deleterious genetic changes may result in stagnation or cell death, whereas advantageous events may result in a proliferative advantage, an increase in recruitment of blood vessels to the developing tumor, and/or the ability to metastasize. The Darwinian concept of "survival of the fittest" (Darwin, 1860) may be modified to "survival of the baddest", that is, those cells that can proliferate the most efficiently, vascularize, and metastasize, but they eventually defeat themselves by killing the host.

Cancer is a disease in which genetic alterations, most frequently chromosomal rearrangements, are acquired in somatic cells, resulting in loss of critical cellular checks and balances, resulting in dysregulation, proliferation, and ultimately, malignant transformation. According to Hanahan and Weinberg (2000), the "hallmarks of cancer" are defects in cell growth and proliferation, evasion of apoptosis, genome instability, sustained angiogenesis, and invasion and metastasis. Human cancer cells are characterized by cytogenetic alterations: numerical chromosomal alterations (aneuploidy), structural chromosomal alterations, and more frequently in carcinomas than hematologic malignancies or sarcomas, chromosomal instability. Aneuploidy results from the numerical chromosomal alterations. Further segmental chromosomal gains and losses come from structural chromosomal alterations, including reciprocal and nonreciprocal translocations, homogeneously staining regions, amplifications, and deletions. The cytogenetic aberrations in hematologic malignancies and sarcomas are frequently clonal, that is, the same abnormalities are present in multiple cells, all the progeny of one cell with a primary genetic alteration. In carcinomas, it is not unusual to find that each cell within a tumor has a different karyotype owing to chromosomal instability, which can be defined simply as numerical and structural chromosomal alterations that vary from cell to cell. Chromosomal instability is thought to be the means by which cells develop the features that enable them to become cancer cells (Hanahan and Weinberg, 2000). In spite of chromosomal instability, cancer cell karyotypes remain stable over time, probably because their genotypes have been optimized over their evolution for growth, making it less likely that additional genetic alterations will confer an additional growth advantage (Albertson *et al*., 2003).

2. Historical perspective and technical background

Chromosomal alterations and karyotypic instability in human tumor cells have been investigated for more than a century. In 1890, David Hansemann reported

that tumor cells with an abnormal chromosome constitution (aneuploidy) had abnormal mitotic spindles with extra poles and other chromosome segregational abnormalities (Hansemann, 1890). This aneuploidy theory was formalized in the early 1900s when Theodor Boveri, while studying chromosomal segregation in *Ascaris* worms and *Paracentrotus* sea urchins, suggested that malignant tumors arise from a single cell that has acquired an abnormal chromosome constitution (Boveri, 1929). Chromosomal cytology was advanced in the 1920s and 1930s by two technical improvements by plant cytologists, the squash technique that eliminated the difficulty in reconstructing and analyzing chromosomes from serially sectioned preparations and the application of the mitotic inhibitor, colchicine, derived from the Mediterranean *Colchicum* plant, to disrupt the mitotic spindle and increase the yield of mitotic cells available for chromosome analysis. In spite of these improvements, mammalian cytogenetics did not blossom until the 1950s, when hypotonic treatment was found to facilitate chromosome spreading, enabling visualization, counting, and analysis of individual chromosomes. In 1955, Tjio and Levan applied these new advances in cytogenetics to their cultures of human fetal lung tissues and discovered that the human chromosome number is 46 (23 pairs of autosomes plus the two sex chromosomes, XX in females or XY in males), rather than the previously reported 48. During this time period, Theodore Hauschka, Sajiro Makino, and Albert Levan each observed that tumor cell lines expressed both numerical and structural chromosomal aberrations. Soon afterward, Jerome Lejeune founded the field of clinical cytogenetics, by reporting that children with Down syndrome (then called mongolism) had 47 chromosomes due to trisomy 21. Cancer cytogenetics advanced in 1960, when Nowell and Hungerford reported the first chromosome abnormality in malignant human cells, the Philadelphia (Ph) chromosome, in chronic myelogenous leukemia (CML). Their finding supported Boveri's theory that cancer cells express chromosomal alterations. Major breakthroughs in cytogenetics occurred in the 1970s as a result of the development of chromosomal banding methods that enabled the identification of individual chromosomes as unique entities, based on their staining patterns (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1 and Article 71, **Advances in cytogenetic diagnosis**, Volume 2). Classical karyotyping was formalized at the Paris conference in 1971 (and a subsequent meeting in Edinburgh early in 1972), when a uniform system of human chromosome identification based on the new chromosome banding methods was established. Current karyotyping practice involves analysis of 20 trypsin-Giemsa banded metaphase cells and karyotyping of at least two per clone. Cytogenetic nomenclature divides chromosomes into arms (p for petit and q, the next letter in the alphabet after p), and arms into regions, and regions into bands and subbands (ISCN, 1995). Thus, chromosomal locations are pronounced by chromosome number, arm, region, and band. For example, chromosomal band 11q13 is pronounced, "eleven q one three" rather than "eleven q thirteen", since 13 refers to region one band three and there are not 13 bands on the long arm of chromosome 11. Numerous additional classical cytogenetic banding methods have been developed to define chromosome structure and specific chromosome abnormalities.

Cytogenetic techniques critical to our understanding of genetic alterations in cancer cells are listed in Table 1. Molecular cytogenetic analysis, which blossomed

Table 1 Techniques useful in cancer cytogenetics

Cytogenetic technique	Application	Tissue preparation
Trypsin-Giemsa banding (G-banding)	Classical karyotyping	Metaphase spreads from cultured cancer cells
Fluorescence in situ hybridization (FISH)	Molecular cytogenetic analysis of targets defined by the DNA probe utilized, e.g., centromeres by α-satellite DNA, cosmid or bacterial artificial chromosome (BAC) clones of genes, etc. for detection of chromosome or gene copy number or gene mapping	Metaphase spreads or interphase nuclei from cultured cells, paraffin or frozen sections, touch preparations, cytology brush preparations, or dissociated tissue biopsies
Chromosome painting	Molecular cytogenetic analysis of specific chromosomes or chromosome segments	Metaphase spreads from cultured cancer cells
Spectral karyotyping or multiplex FISH	Molecular karyotyping using a set of chromosome-specific DNA probes, each labeled with a different dye or dye combination, to distinguish each of the human (or mouse) chromosomes as unique	Metaphase spreads from cultured cancer cells
Comparative genomic hybridization (CGH)	Comprehensive screen of the genome for segmental gains and losses using DNA isolated from test tissue (e.g., tumor cells; labeled in green) and control DNA (labeled in red)	Normal control male metaphase chromosomes; high molecular weight DNA isolated from cancer cells

in the late 1980s and continues to advance, has revolutionized the field of cytogenetics, particularly, cancer cytogenetics. The primary technique used for molecular cytogenetic characterization of chromosomes is fluorescence in situ hybridization (FISH) (Landegent *et al*., 1987; Cremer *et al*., 1988; Trask, 1991; Lichter and Ried, 1994; *see also* Article 22, **FISH**, Volume 1). This method involves hybridization of one or more fluorescently labeled DNA probes to metaphase chromosomes and/or nuclei. The most common form of FISH involves hybridizing a fluorescently labeled DNA probe (or a hapten-labeled DNA probe that can later be detected by a fluorescent tag) to a gene or DNA segment of interest and a different colored fluorochrome-labeled control probe (that maps to the same or a different chromosome) to metaphase spreads from a tissue or cells of interest. Cancer-related applications include mapping tumor-related genes or enumeration of gene or chromosome copy number in the context of deletions, amplifications, or ploidy changes. FISH hybridization to interphase nuclei (in paraffin or frozen sections, touch preparations, cytology brush preparations, or from harvested, dissociated tissue) is quite useful in cancer cytogenetics. Optimally, 200 cells are examined for this type of analysis. The results might also provide an indication of ploidy.

Interphase FISH eliminates the necessity and time of cell culture and enables enumeration of gene copy number compared to a control probe on the same chromosome in large populations of cells by molecular karyotyping of interphase nuclei even in the absence of metaphase chromosomes. However, FISH analyses only identify alterations of the targeted DNA probe(s), and do not detect other alterations, such as chromosomal abnormalities that may be important for a particular diagnosis or prognosis in a cancer patient. Therefore, at the present time, it is critically important that FISH be used as an adjunct to classical cytogenetic analysis and not as a "stand-alone" test. Another FISH method, chromosome painting, utilizes a cocktail of fluorochrome-labeled DNA probes from a chromosome or selected chromosomal region to localize those sequences in metaphase spreads from a tissue of interest (Pinkel *et al*., 1988; Ried *et al*., 1998). Chromosome painting enables identification of the origin of chromosomal segments involved in translocations and characterization of marker chromosomes, the origin of which is refractory to definition by classical cytogenetic methods. When examined using standard fluorescence microscopy, painting of two or more differentially labeled chromosomes in interphase nuclei results in beautiful multicolored nuclei, but provides little specific structural information owing to overlap of the chromosomal domains, which can only be resolved by computerized optical sectioning using confocal or other sophisticated computer-controlled microscopy.

Two new variations of chromosome painting that are extremely useful in cancer cytogenetics are spectral karyotyping (SKY) and multiplex FISH (M-FISH) (LeBeau, 1996; Schröck *et al*., 1996; Speicher and Ward, 1996; Speicher *et al*., 1996). These methods distinguish the 24 different human chromosomes (22 autosomes and two sex chromosomes) using a set of chromosome-specific DNA probes, each labeled with a different array of one or more fluorescent dyes and FISH hybridization to metaphase chromosomes. The detection and classification of the rainbow of probes utilizes different strategies, both involving CCD cameras and computer algorithms, with SKY taking advantage of Fourier spectroscopy to resolve the multicolored probes and M-FISH using multiple optical filters and image superimposition. Combinations of these classical and cytogenetic methods are essential to fully characterize the karyotypic alterations in cancer cells.

A common problem in cancer cytogenetics is a low mitotic index and the consequent lack of metaphase chromosomes to analyze and karyotype. Comparative genomic hybridization (CGH) is a form of FISH that comprehensively screens the genome for gains and losses of DNA segments (Kallioniemi *et al*., 1992; Du Manoir *et al*., 1993). The advantages of CGH are that it provides an overall picture of genomic imbalances in a tumor, eliminates the need for cell culture, and thus overcomes the primary difficulty of karyotyping tumors. The disadvantages of CGH are that it cannot detect balanced chromosomal rearrangements, such as translocations or inversions, and it cannot suggest the manner of loss in terms of chromosome rearrangements (i.e., whether loss of a segment is due to a deletion or an unbalanced translocation). Further, CGH can only detect relatively large genetic aberrations (such as gains of at least 2 mB and losses on the order of 10 mB). Moreover, the smallest abnormalities are only detectable by CGH when present in the vast majority of the cells (∼85%) in the population(s) from which the DNA was isolated, and larger abnormalities must be present in the majority of cells

(∼70%). Like other molecular methods that average the results, CGH analysis of tumor specimens is hampered significantly by the genetic heterogeneity seen within tumors and the frequent admixture with normal cells. This procedure is facilitated by commercially available imaging instrument and computer systems combined with sophisticated proprietary software. Newer advances include array CGH, a method involving the use of DNA microarrays containing specifically spaced, mapped markers distributed throughout the genome that are then hybridized and the ratios are linked to segmental gains and losses (Pinkel *et al*., 1998). The advantage of array CGH over cytogenetic CGH is higher resolution. The significant disadvantage at present is that high-resolution arrays are not readily available to the scientific community.

Advances in cancer cytogenetics have led from the discovery of the Ph chromosome to the characterization by Janet Rowley of the Ph chromosome as a reciprocal translocation between chromosomes 9 and 22, t(9;22)(q34;q11.2) (Figure 2). Identification of the Ph translocation, whether by classical or molecular cytogenetic methods or molecular genetic techniques, is now essential for diagnosis of patients with CML, although the Ph chromosome may also be observed in some patients with acute lymphocytic leukemia (ALL) and less frequently, acute myeloid leukemia (AML). Characterization of the Ph chromosome in CML led to cloning of the Philadelphia translocation breakpoint and the discovery that the translocation creates a hybrid gene consisting of 5′ regulatory and coding sequences of the *BCR* gene on chromosome 22 joined with 3′ coding, polyadenylation, and termination sequences from the *ABL* proto-oncogene on chromosome 9. This BCR/ABL rearrangement may be identified by FISH, a molecular cytogenetic method described earlier (Figure 5). The new chimeric gene formed by the Philadelphia translocation codes for a protein with enhanced tyrosine kinase activity. Identification of this protein led to the development of an effective new drug, Gleevec®, which inhibits this tyrosine kinase, blocks cellular proliferation, and induces apoptosis in cells expressing the Ph chromosome. Thus, clarification of the cytogenetic aberrations in a cancer enables not only diagnosis, but molecular characterization of the breakpoints, identification of the genes involved, description of the molecular pathology of the malignancy, and development of targeted therapies. As noted recently by Rowley (2001), cytogenetic identification and subsequent molecular cloning of translocation breakpoints has become one of the most efficient tools for identifying genes involved in malignant transformation.

3. Clinical applications of cancer cytogenetics

Over the past four decades, it has become clear that acquired, nonrandom chromosome abnormalities are associated with specific malignancies, particularly hematologic malignancies and sarcomas, although other alterations are characteristic of carcinomas, either individually or as a group (Table 2; Figures 1–5). The diagnosis of hematopoietic and lymphoid malignancies has been codified recently by the World Health Organization (WHO) (Jaffe *et al*., 2001). Cytogenetic studies are valuable for determining the diagnosis and prognosis of specific cancers and serve as independent etiologic or prognostic risk factors even when age, white blood

Table 2 Most common clonal chromosomal abnormalities in malignancies[a,b]

Malignancy	Consistent chromosomal abnormalities	Specific genes involved
AML	t(8;21)(q22;q22)	*AML1/ETO*
Acute promyelocytic leukemia	t(15;17)(q22;q12)	*PML/RARα*
AML with abnormal bone marrow eosinophils	inv(16)(p13q22)[c] or t(16;16)(p13;q11)	*CBFB/MYH11X*
AML	11q23 rearrangements	*MLL*
Chronic myelogenous leukemia (CML)	t(9;22)(q34;q11.2)[d]	*BCR/ABL*[g]
Precursor B-cell acute lymphoblastic leukemia	t(9;22)(q34;q11.2)	*BCR/ABL*
	t(4;11)(q21;q23)[e]	*MLL2(AF4)/MLL*
	t(1;19)(q23;p13)	*PBX/E2A*
	t(12;21)(p12;q22)	*ETV/CBFA*
Burkitt lymphoma	t(8;14)(q24;q32)	*MYC/IGH@*
Follicular lymphoma	t(14;18)(q32;q21)	*IGH@/BCL2*
Mantle cell lymphoma	t(11;14)(q13;q32)	*CCND1/IGH@*
Anaplastic large cell lymphoma	t(2;5)(p23;q35)[f]	*NPM1/ALK*
Alveolar rhabdomyosarcoma	t(2;13)(q35;q14)	*PAX3/FOXO1A*
Ewing sarcoma	t(11;22)(q24;q12)	*EWSR1/FLI1*
Synovial sarcoma	t(X;18)(p11;q11)	*SS18/SSX1*, *SS18/SSX2*, *SS18/SSX4*

[a]AML: acute myeloid leukemia; CLL: chronic lymphocytic leukemia.
[b]Based on Mitelman Database of Chromosome Aberrations in Cancer (2003).
[c]Figure 1.
[d]Figure 2.
[e]Figure 3.
[f]Figure 4.
[g]Figure 5.

cell count, classification, and cell surface markers are evaluated. The new WHO classification of hematopoietic malignancies is the first to formally recognize the importance of genetic alterations in the diagnosis of leukemias, especially with regard to AMLs. Several types of AML are classified specifically on the basis of the cytogenetic aberrations present in the cells (Table 2; Figure 1). A cytogenetic alteration may be used to diagnose a particular malignancy on the basis of the specific association of the aberration with the disorder even when other diagnostic criteria are equivocal. Moreover, specific chromosomal abnormalities in malignancies may be associated with specific etiologies, such as loss of one chromosome 5, one chromosome 7, or deletion of the long arm of one chromosome 5 or one chromosome 7 in patients with AML after exposure to organic chemicals (such as dry cleaning or furniture refinishing solvents) or mutagenic therapy. Specific chromosomal alterations are often associated with a good prognosis. In AML, the t(8;21), t(15;17), and inv(16) rearrangements (Table 2) are usually associated with a favorable response to therapy and long-term survival (Grimwade *et al*., 1998). A complex karyotype, loss of one chromosome 5 or one chromosome 7, deletion

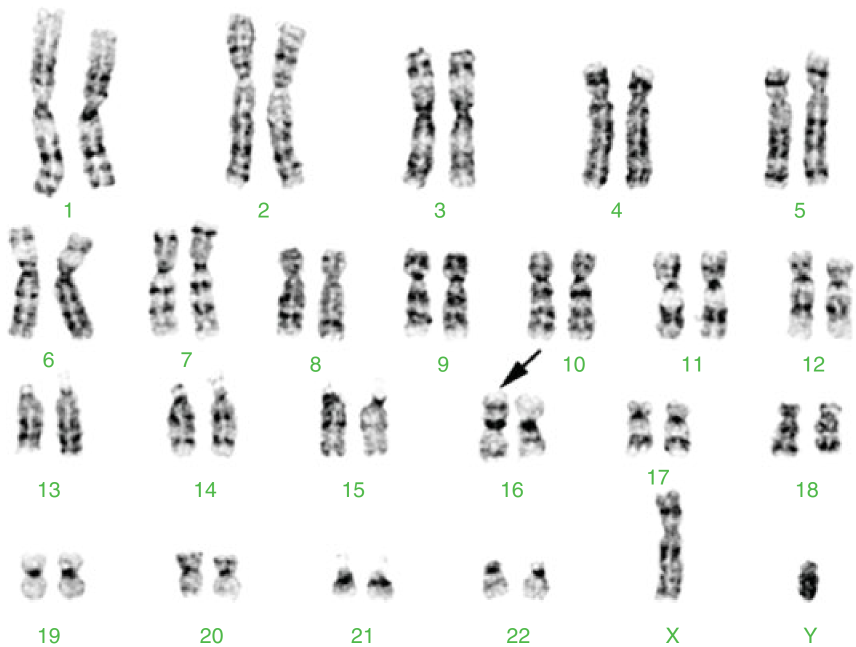


Figure 1 Karyotype from a 10-year-old female with acute myeloid leukemia (AML) with abnormal bone marrow eosinophils and the characteristic inverted chromosome 16 (arrow). This inversion results in a fusion between *CBFB* on the long arm that codes for a subunit of a DNA binding protein complex and the *MYH11* gene on the short arm that codes for the smooth muscle form of myosin heavy chain protein. The fusion protein is thought to interact with the AML1 transcription factor and increase its ability to bind to DNA and regulate transcription. 46,XX,inv(16)(p13q22)

of the long arm of one chromosome 5 or one chromosome 7, and/or abnormalities involving the long arm of chromosome 3 are associated with a relatively poor prognosis in AML. In addition to the diagnostic and prognostic usefulness of cytogenetics, a change in the karyotype of the abnormal cells from a patient often precedes progression of their malignancy. Disease relapse in a patient following bone marrow transplantation may also be heralded by the reappearance of host cells with acquired chromosomal abnormalities that may have been observed prior to transplantation. Thus, chromosomal abnormalities are common in human cancer cells and may provide useful diagnostic and prognostic information.

Cytogenetic alterations often point to the genes involved in the development and/or progression of malignancies. Numerous chromosomal translocation breakpoints have been cloned, and the genes, regulation of which has been altered by those translocations identified. These genes belong to specific functional families that are involved in the regulation of cellular growth (see Rowley, 2001, Table 1). Genes altered in cancer cells are important targets for the development of designer drugs to limit aberrant cellular proliferation. The example of the Ph chromosome was discussed earlier. Identification of the altered gene product in CML, a novel tyrosine kinase resulting from this translocation led to Gleevec®, a targeted

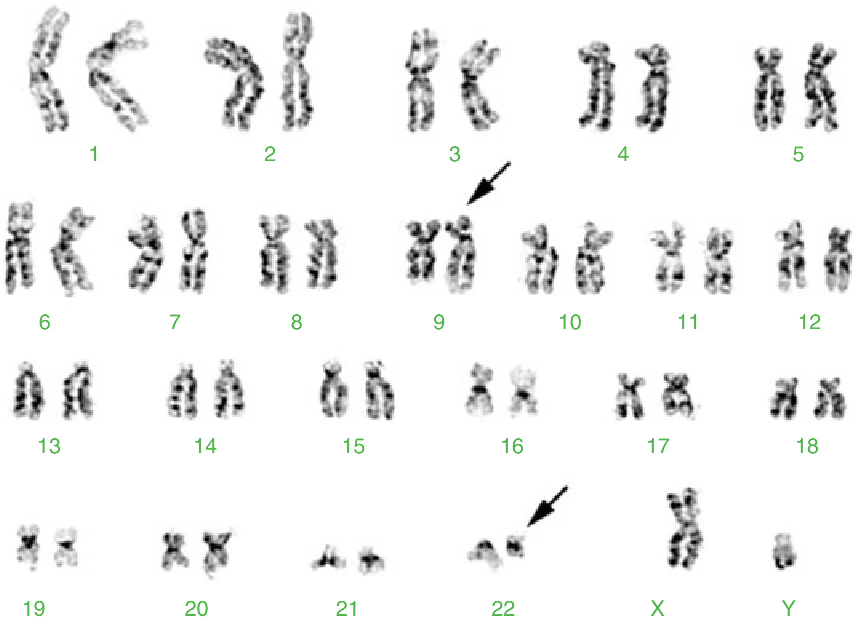


Figure 2 Karyotype from a 33-year-old male presenting with chronic myelogenous leukemia (CML). 46,XY,t(9;22)(q34;q11.2). See text for details about this translocation. The derivative chromosomes 9 and 22 are marked by arrows

(designer) therapy, with significant efficacy in CML and several solid tumors. Targeting of additional genes or gene families may lead to effective therapies for other malignancies.

4. On the origin of chromosomal abnormalities in human cancers

Although specific numerical and/or structural cytogenetic alterations have been associated with particular malignancies, the mechanisms by which those abnormalities occur have yet to be delineated.

Numerical chromosomal alterations can arise by random loss or gain of chromosomes, but also as a result of defects in the process by which equal numbers of chromosomes are distributed to daughter cells at mitosis, that is, chromosome segregation. Recent studies have shown that several factors can result in segregation defects, including abnormal kinetochore-spindle interactions, premature chromatid separation, centrosome amplification, multipolar spindles, and abnormal cytokinesis. Research by a number of investigators has recently advanced our understanding of the relationship between centrosomal defects, multipolar spindles, unequal chromosome distribution, and chromosomal instability. Investigations of several centrosomal proteins have been important in our understanding of segregational defects in human cells. Examination of Aurora A kinase and other proteins



Figure 3 Karyotype from a young boy with acute lymphoblastic leukemia. This translocation results in fusion of the *MLL* gene at 11q23 and the *MLL2/AF4* gene at 4q21, resulting in an abnormal chimeric transcript. The leukemia in these patients usually has aggressive clinical features, hyperleukocytosis, a B-precursor immunophenotype with coexpression of myeloid markers, and a poor response to conventional chemotherapy. 46,XY,t(4;11)(q21;q23). The derivative chromosomes 4 and 11 are marked by arrows

(pericentrin, NuMA, γ-tubulin, and polo kinase) involved in centrosome function and the spindle checkpoint are in progress (reviewed in Gollin, 2004). Their role in chromosomal instability and implications in the diagnosis, prognosis, and therapy of human tumors will be revealed in the coming years. Further, these proteins are candidate targets for drug development.

Structural aberrations in cancer cells include structural chromosomal alterations, including reciprocal and nonreciprocal translocations, homogeneously staining regions, amplifications, and deletions. The etiology of structural chromosomal abnormalities is not clear. However, we do know that (1) environmental agents, such as cigarette smoking, benzene, and other chemicals, cause DNA double-strand breaks; (2) repeated sequences, including duplicons and other genomic repeats may play a role in chromosomal rearrangements in cancer (Saglio *et al*., 2002; Lupski, 2003; and discussed in Rowley, 2001); and (3) patients with DNA repair defects are prone to develop cancer (Shiloh, 2003; Gollin, 2004). Thus, structural abnormalities in cancer may be related to one or a combination of factors, including the formation of DNA double strand breaks, and then either recombination between homologous genomic motifs and/or defects in the DNA damage response. Structural chromosomal abnormalities in cancer cells often result



Figure 4 Karyotype from a male with anaplastic large cell lymphoma with the t(2;5)(p23;q35) and trisomy 7. This translocation results in fusion between the anaplastic lymphoma kinase (*ALK*) gene on 2p23 and the nucleophosmin (*NPM*) gene on 5q35. 47,XY,t(2;5)(p23;q35),+7. The derivative chromosomes 2 and 5 are marked by arrows

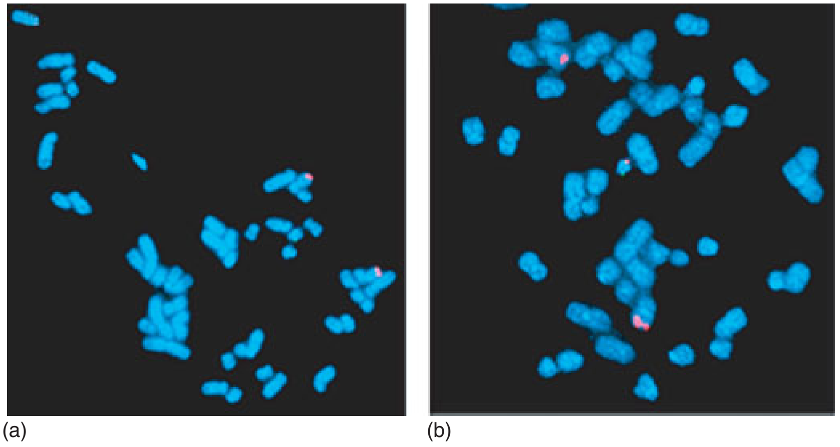


Figure 5 FISH for the *BCR/ABL* rearrangement, usually seen in chronic myelogenous leukemia. The cells are hybridized with the LSI *BCR/ABL* ES (extra signal) probe (Vysis, Inc., Downers Grove, IL). (a) Negative control (no rearrangement). Two red signals for the *ABL* oncogene (9q34) on distal 9q. Two green signals for the *BCR* gene at 22q11.2. (b) Metaphase from a CML patient with the t(9;22)(q34;q11.2). *BCR/ABL* rearrangement (adjacent red and green signals on the small Philadelphia chromosome in the center of the metaphase spread), one red signal on the normal chromosome 9, a residual red signal on the rearranged chromosome 9, and one green signal on the normal chromosome 22

in further structural alterations and instability due to chromosomal breakage. Factors involved in structural and numerical chromosomal instability in cancer cells are reviewed in more detail in Gollin (2004). Further studies on the mechanisms behind chromosomal alterations in human tumors are underway in many labs around the world.

5. Conclusion

Cancer cytogenetics has become a very important area of clinical cytogenetics, critical to the diagnosis of hematologic malignancies and helpful in the differential diagnosis of solid tumors. Research in cancer cytogenetics will continue to play an important role in the identification of oncogenes and tumor suppressor genes and examination of their function(s) in the cellular regulatory and signaling pathways. With the results of the Human Genome Project and the advances in molecular biology laboratory methodology and drug development, our prospects for significant advances in cancer diagnostic and prognostic testing and therapy based on the results of cancer cytogenetic studies are great. Yet, it is likely that classical as well as molecular cytogenetic studies will continue to be carried out on malignancies, as they provide a relatively rapid, inexpensive overall view of genomic alterations in a malignancy, useful for diagnosis, prognosis, and choice of therapeutic regimen.

Acknowledgments

The author is grateful to Ms. Maureen Sherer for preparing the illustrations for this contribution, Drs. Janet Rowley and Sofia Shekhter-Levin for helpful discussions, and the National Institute for Dental Research for continued research support. This article is dedicated to Professor Robert C. King on the occasion of his 75th birthday. He believed that the author could become a successful scientist and writer, and it happened because of his nurturing.

References

Albertson DG, Collins C, McCormick F and Gray JW (2003) Chromosome aberrations in solid tumors. *Nature Genetics*, **34**, 369–376.

Boveri T (1929) *The Origin of Malignant Tumors*, translated by Boveri Marcella, The Williams & Wilkins Company: Baltimore.

Cavenee WK and White RL (1995) The genetic basis of cancer. *Scientific American*, **272**, 72–79.

Cremer T, Lichter P, Borden J, Ward DC and Manuelidis L (1988) Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Human Genetics*, **80**, 235–246.

Darwin C (1860) *On The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murray: London.

Du Manoir S, Speicher MR, Joos S, Schröck E, Popp S, Döhner H, Kovacs G, Robert-Nicoud M, Lichter P and Cremer T (1993) Detection of complete and partial chromosome gain and losses by comparative genomic in situ hybridization. *Human Genetics*, **90**, 590–610.

Fearon ER and Vogelstein B (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.

Gollin S (2004) Chromosomal instability. *Current Opinion in Oncology*, **16**, 25–31.

Grimwade D, Walker H, Oliver F, Wheatley K, Harrison C, Harrison G, Rees J, Hann I, Stevens R, Burnett A, *et al*. (1998) on behalf of the Medical Research Council Adult and Children's Leukaemia Working Parties, The importance of diagnostic cytogenetics on outcome in AML: Analysis of 1,612 patients entered into the MRC AML 10 trial. *Blood*, **92**, 2322–2333.

Hanahan D and Weinberg RA (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hansemann D (1890) Ueber asymmetrische celltheilung in epithel krebsen und deren biologische bedeutung. *Virchows Archiv für Pathologische Anatomie*, **119**, 299–326.

ISCN (1995) In *An International System for Human Cytogenetic Nomenclature*, Mitelman F (Ed.), S. Karger: Basel.

Jaffe ES, Harris NL, Stein H and Vardiman JW (Eds.) (2001) *Tumours of Haematopoietic and Lymphoid Tissues, Who Classification of Tumors*, IARC Press: Lyon.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kinzler KW and Vogelstein B (1997) Cancer-susceptibility genes: gatekeepers and caretakers. *Nature*, **386**, 761–763.

Landegent JE, Jansen in de Wal N, Dirks RW, Baas F and van der Ploeg M (1987) Use of whole cosmid cloned genomic sequences for chromosomal localization by non-radioactive in situ hybridization. *Human Genetics*, **77**, 366–370.

LeBeau MM (1996) One FISH, two FISH, red FISH, blue FISH. *Nature Genetics*, **12**, 341–344.

Lichter P and Ried T (1994) Molecular analysis of chromosome aberrations-in situ hybridization. *Methods in Molecular Biology*, **29**, 449–478.

Lupski JR (2003) 2002 Curt stern award address. Genomic disorders recombination-based disease resulting from genomic architecture. *American Journal of Human Genetics*, **72**, 246–252.

Mitelman Database of Chromosome Aberrations in Cancer, Mitelman F, Johansson B and Mertens F (Eds.) (2003) **http://cgap.nci.nih.gov/Chromosomes/Mitelman**.

Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J and Gray J (1988) Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 9138–9142.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W-L, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Ried T, Schrock E, Ning Y and Wienberg J (1998) Chromosome painting: a useful art. *Human Molecular Genetics*, **7**, 1619–1626.

Rowley JD (2001) Chromosome translocations: dangerous liaisons revisited. *Nature Reviews Cancer*, **1**, 245–250.

Saglio G, Storlazzi CT, Giugliano E, Surace C, Anelli L, Rege-Cambrin G, Zagaria A, Jimenez Velasco A, Heiniger A, Scaravaglio P, *et al*. (2002) A 76-kb duplicon maps close to the BCR gene on chromosome 22 and the ABL gene on chromosome 9: possible involvement in the genesis of the Philadelphia chromosome translocation. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9882–9887.

Schröck E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, *et al*. (1996) Multicolor spectral karyotyping of human chromosomes. *Science*, **273**, 494–497.

Schwab M (1999) Oncogene amplification in solid tumors. *Seminars in Cancer Biology*, **9**, 319–325.

Shiloh Y (2003) ATM and related protein kinases: safeguarding genome integrity. *Nature Reviews Cancer*, **3**, 155–168.

Sidransky D (1995) Molecular genetics of head and neck cancer. *Current Opinion in Oncology*, **7**, 229–233.

Speicher MR and Ward DC (1996) The coloring of cytogenetics. *Nature Medicine*, **2**, 1046–1048.

Speicher MR, Ballard SG and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, **12**, 368–375.

Trask B (1991) DNA sequence localization in metaphase and interphase cells by fluorescence in situ hybridization. *Methods in Cell Biology*, **35**, 3–35.

Vogelstein B and Kinzler KW (1993) The multistep nature of cancer. *Trends in Genetics: TIG*, **9**, 138–141.

Weinberg RA (1996) How cancer arises. *Scientific American*, **275**, 62–70.

Specialist Review

Human X chromosome inactivation

Samuel C. Chang and Carolyn J. Brown
University of British Columbia, Vancouver, BC, Canada

1. X chromosome inactivation

1.1. The mammalian sex chromosomes

Dioecy, the presence of two distinct sexes, is common amongst animals. In mammals, sex determination is controlled by distinct sex chromosomes (the X and Y), which are believed to have originated from an autosomal chromosome pair by the acquisition of a major sex-determining locus and subsequent divergence (Ayling and Griffin, 2002). Suppression of recombination near the sex-determining locus would be evolutionarily favored; however, the absence of exchange between the homologs would also lead to the degeneration of the sex-determining Y over evolutionary time (Ohno, 1967). The human X and Y maintain identity in only a small region of homology at the tips of both the long and short arms, referred to as the "pseudoautosomal" region (PAR), which undergoes recombination during male meiosis. The male-specific region of the Y also retains 20 genes or gene families derived from X chromosomal sequences in addition to testis-specific multicopy sequences, many of which are arrayed in palindromic segmental duplications (Skaletsky *et al*., 2003). The presence of such repeats has been proposed as a mechanism allowing gene conversion to protect haploid Y genes against their inevitable decay due to acquisition of mutations (Rozen *et al*., 2003). The X retains much of its old autosomal information, including numerous important housekeeping genes, despite being a relatively gene-poor chromosome containing approximately 5% of the haploid genome but only about 3% of the genes (Graves *et al*., 2002).

1.2. Dosage compensation in mammals and other species

In response to the degeneration of the Y, there would have been adaptive pressure to increase expression from the single expressed X-linked copy of genes. Indeed, dosage of X-linked genes, despite their monoallelic expression, seems comparable to autosomal genes, as determined by comparison of expression levels of a variably X-linked gene in different species (Adler *et al*., 1997), microarray analysis (Disteche *et al*., 2001), or impact of aneuploidy (Mizuno *et al*., 2002). In addition,

excessive X-linked gene expression in females relative to males established the need for dosage compensation, which is achieved by X chromosome inactivation, whereby a single X is inactivated in female cells early in mammalian development. In individuals with an abnormal number of X's (e.g., 45,X; 47,XXX; 47,XXY, etc.), a single X remains active no matter how many are present (the "n−1" rule). In triploids, two X's can be active, implicating autosomal copy number in determining the number of X's that remain active (Migeon, 2002).

Dosage compensation also occurs in other dioecous organisms, utilizing a variety of mechanisms. In the fruit fly, *Drosophila melanogaster* (males XY, females XX), hypertranscription of most genes along the single male X matches the output of the female's two X's, and in the nematode, *Caenorhabditis elegans* (males XO, hermaphrodites XX), dosage compensation is accomplished in hermaphrodites by downregulation of X-linked gene transcription from both X's (reviewed in Marin *et al*., 2000). In birds (males ZZ, females ZW), there is some evidence for dosage compensation; however, it appears that this is not accomplished by the silencing of one Z chromosome as biallelic expression in the male is observed (Kuroiwa *et al*., 2002). Despite their substantial differences, dosage compensation in mammals, flies, and worms is thought to involve the control of chromatin structure, and in flies and mammals, it apparently involves functional RNAs (XIST in mammals, see below; and roX in flies (Meller *et al*., 1997; *see also* Article 27, **Noncoding RNAs in mammals**, Volume 3)). Mammalian X chromosome inactivation is unique amongst these mechanisms in that it requires differentiation between an active and inactive X within the same nuclear environment.

2. Features of the inactive X

The inactive X (Xi), as befits facultative heterochromatin, shares general features with constitutive heterochromatin. These include suppression of gene expression, condensation of chromatin (forming the Barr body), late replication, nuclease insensitivity, hypoacetylation of histones, and hypermethylation of DNA. However, several unique features characterize heterochromatin of the Xi, including the association of a large untranslated RNA, the X-inactive specific transcript (*XIST* in human or *Xist* in mice), and a nonrandom distribution of variants of the core histone H2A. Once the inactive state is established, silencing is faithfully maintained throughout subsequent mitotic cell divisions, resulting in clonal populations, that, in the case of some tumors, could include 10^9 cells (Linder and Gartler, 1965). Thus, X inactivation is a paradigm for both the establishment of a differentiated chromatin configuration and cellular memory of that differentiated state. Figure 1 summarizes some of these features of the Xi, which are discussed below.

2.1. *Gene silencing*

Genes that continue to be shared between the X and Y, notably those in the PARs, would have no need for dosage compensation, as they would continue to

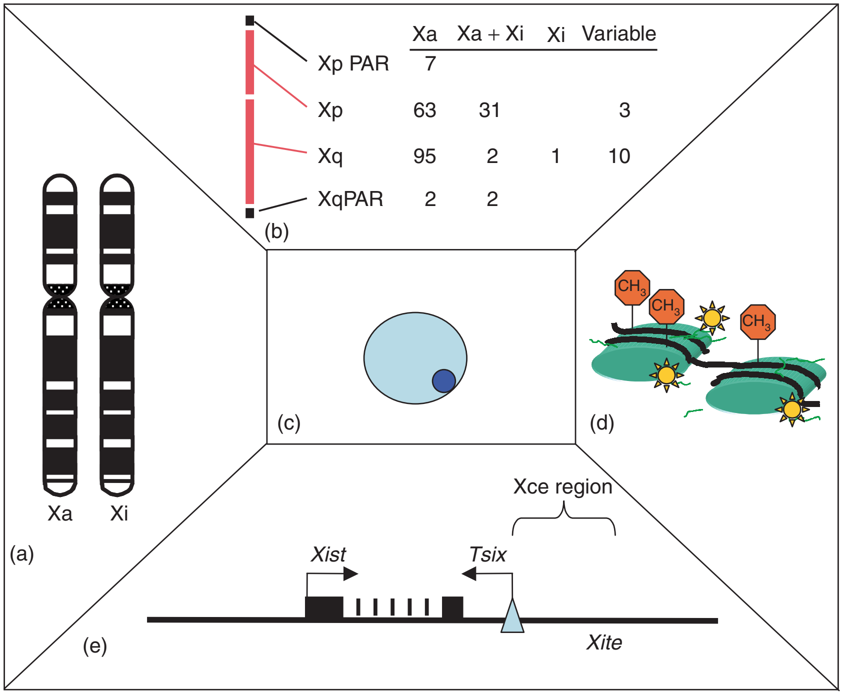


Figure 1 Facets of X inactivation. (a) In mammals, dosage compensation must differentiate between the active and inactive X chromosome in the same nuclear environment. Shown here are ideograms of the human X chromosomes. (b) Furthermore, in humans, approximately 15% of genes escape inactivation, being expressed from the Xa and Xi. As tabulated in this drawing, the majority of genes escaping inactivation are located on the short arm of the X (data from Carrel L, Cottle AA, Goglin KC and Willard HF (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14440–14444). (c) In interphase nuclei, the Xi can manifest as the Barr body, which stains darkly with DAPI, as diagrammed in this cartoon. (d) Modifications of the chromatin on the Xi include methylation of the DNA at the promoters of X-linked genes (shown as red stop signs) and hypoacetylation of histones as well as hypermethylation at lysine 9 and 27 of histone H3 (represented as stars). (e) The *Xist* gene is necessary for the establishment of inactivation, and in mice, *Xist* expression is regulated by the antisense *Tsix* transcript that initiates transcription at a CpG island (blue triangle) shown to bind CTCF. The choice of X to inactivate is influenced by the *Xce* locus and the recently defined *Xite* region (see text for references)

be present in two copies in both males and females, and thus it was anticipated that some genes would "escape" inactivation and continue to be expressed from the Xi (Lyon, 1962). What is surprising, however, is the number of genes escaping inactivation in humans. Over 15% of human genes continue to be expressed from the inactive X chromosome (Carrel *et al*., 1999), and clearly, not all of these have Y homologs. Thus, a dosage imbalance may continue to exist, whereby human females express more X-linked gene product than males. Interestingly, in mice, fewer genes have been identified to escape inactivation (reviewed in Disteche *et al*., 2002). Genes that escape inactivation are not all "fully" expressed from the Xi, and several human genes have been shown to have variable inactivation, escaping inactivation in some females but not in others (e.g., *TIMP1* and *CHM1* (Anderson

and Brown, 1999; Carrel and Willard, 1999)). The question of how some genes on the chromosome continue to be expressed while others are being silenced by a cis-acting mechanism is a fascinating biological question, and may also have clinical implications for disease, as discussed below. For many of the genes that have been examined, the general features of silent chromatin (e.g., replication timing, DNA methylation, and histone modifications) are not associated with an active gene on the Xi (see Disteche *et al*., 2002 for a review). A clue to the mechanism by which genes escape silencing may lie in the fact that many, although not all, of the genes expressed from the Xi are located in clusters, implicating a regional mechanism in their failure to be silenced (e.g. Tsuchiya and Willard, 2000). It is not clear whether expression from the Xi results from an initial inability to undergo inactivation or instead reflects an inability to maintain the silent state. The mouse *Smcx* gene is subject to inactivation in a proportion of embryonic cells; however, reactivation of the gene occurs very early in development (Lingenfelter *et al*., 1998). Such an early reactivation could be difficult to distinguish from the failure of inactivation spreading. Given the large number of genes escaping inactivation in humans, it is likely that there is more than a single mechanism by which expression from the Xi is acquired.

2.2. Barr body formation

During interphase, the human Xi forms a darkly staining body, the Barr body, which is seen associated with the nucleolus (Barr and Bertram, 1949) or at the nuclear periphery (Dyer *et al*., 1989), the usual sites for heterochromatin. Three-dimensional reconstructions of the Xi territory suggest that rather than being substantially more condensed than the active X (Xa), the principal difference between the two chromosomes is in shape (Eils *et al*., 1996). Thus, the intense staining of the Barr body with DNA dyes likely reflects the distinctive packaging of DNA on the Xi; however, to date, only a limited number of proteins have been shown to associate specifically with the Xi.

The Xi is characterized by the accumulation of variant histone H2A isoforms termed *macrohistone H2A1 and 2* (Costanzi and Pehrson, 1998; Chadwick and Willard, 2001). These proteins are very similar to H2A, but include a large carboxy-terminal domain, and their enrichment at the Xi forms a characteristic structure in the female nucleus, referred to as a *macrochromatin body* (*MCB*). MacroH2A has been shown to have a transcriptional repressive effect (Perche *et al*., 2000), potentially caused by both interfering with transcription factor binding and chromatin remodeling (Angelov *et al*., 2003). Deletion of *Xist* after inactivation results in loss of macroH2A association with the Xi (Csankovszki *et al*., 1999), while induction of *Xist* is sufficient to induce MCB formation (Rasmussen *et al*., 2001), suggesting a direct or indirect association between the Xist RNA and macroH2A. Such an association is further supported by coimmunoprecipitation of Xist by macroH2A antisera (Gilbert *et al*., 2000).

Chadwick and Willard (2003b) recently examined a variety of chromatin proteins for enrichment on the Xi and demonstrated elevated levels of heterochromatin protein-1 (HP1), histone H1, and the high mobility group protein HMG-I/Y at

the territory of the Xi in interphase human cells. The association, or at least the visualization of the association, of these various proteins with the Xi seems to be cell-cycle and perhaps cell-type specific, and often correlates with visualization of the Barr body (Chadwick and Willard, 2003a). Surprisingly, another protein shown to be associated with the inactive X and required for XIST localization is the BRCA1 protein (Ganesan *et al*., 2002), which has multiple cellular functions, including chromatin remodeling (Bochar *et al*., 2000). Furthermore, transient association of the polycomb group complex proteins Eed and Enx1 is seen in the earliest stages of X inactivation (Wang *et al*., 2001; Mak *et al*., 2002; Silva *et al*., 2003). Interestingly, these proteins are also required early in development for the monoallelic expression of some imprinted loci (Mager *et al*., 2003). Monoallelically expressed genes (both imprinted and X-linked) seem to have dimethylation of histone H3 at lysine (K) 4 restricted to the promoter and not the body of the gene, a mark that is already present in totipotent embryonic stem (ES) cells (Rougeulle *et al*., 2003). Intriguingly, this mark is not seen for the *Smcx* gene that shows rapid reactivation from the Xi.

2.3. Replication timing

Synchrony of replication origin firing in mammalian cells results in megabase-sized domains that replicate at a similar time during S phase (see Goren and Cedar, 2003 for a review). In general, expressed genes tend to replicate earlier in S phase than silenced genes (Woodfine *et al*., 2004). Late replication of the Xi is observed after pulse labeling of cells with bromodeoxy-uridine, and while there is some cell and tissue specificity to the replication patterns (e.g., Willard, 1977), there are regions of the inactive X that replicate earlier than the rest of the chromosome (e.g., XpPAR), and thus have been hypothesized to contain genes that do not undergo inactivation (Schempp and Meer, 1983). Assessment of replication timing for individual genes supports this suggestion as genes escaping inactivation show synchronous DNA replication between the Xa and Xi (e.g., Boggs and Chinault, 1994; Hansen *et al*., 1996). The origins of replication for the *HPRT* and *G6PD* loci are the same on the Xa and Xi (Cohen *et al*., 2003), implicating the yet-to-be-elucidated epigenetic determinants that delay replication firing on the Xi. A long-standing question is whether replication timing dictates the structure of chromatin or vice versa, and recent DNA injection experiments suggest that timing of replication is sufficient to alter transcriptional activity (Zhang *et al*., 2002). However, the transition from synchronous to asynchronous replication of the two X's in early embryos appears to follow alterations in histone modifications, suggesting that chromatin changes may establish the late replication of the Xi (Chaumeil *et al*., 2002).

2.4. Histone modifications

In addition to variants of the core histones, chromatin can vary in the modifications of histones (*see* Article 27, **The histone code and epigenetic inheritance**, Volume 1). Many amino acid residues of core histones, located predominantly within

the tail regions, have the potential to acquire a range of covalent modifications, including acetylation, methylation, ubiquitinylation, and phosphorylation. It has been hypothesized that a specific combination of modifications confers a particular property to the local chromatin, referred to as the "histone code" (Jenuwein and Allis, 2001), and the Xi shows modifications common to other regions of heterochromatin. Specifically, the Xi shows hypomethylation at K4 and 36, and arginine (R) 17 of histone H3; hypermethylation at K9 and K27 of histone H3; hypoacetylation at R2, R12, and R26 of histone H3 and K5, K8, K12, and K16 of histone H4 (see Chaumeil *et al*., 2002; Chadwick and Willard, 2003a). How the histone modifications associated with the Xi determine or influence interactions with other chromatin proteins that are involved in X inactivation is largely unknown, however, acquisition of H3 modifications is one of the first detectable events after Xist RNA coating of the presumptive Xi (Chaumeil *et al*., 2002). The Suv39h histone methyltransferase responsible for constitutive heterochromatin histone K9 methylation does not seem to be necessary for Xi histone methylation (Peters *et al*., 2002), while the Eed-Enx1 histone methyltransferase complex shows only a transient association with the Xi (Silva *et al*., 2003), implicating the involvement of additional chromatin-modifying enzymes in X inactivation.

2.5. *DNA methylation*

In mammals, symmetrical DNA methylation at CpG dinucleotides is a stable and clonally heritable epigenetic modification that is maintained by the action of a maintenance methyltransferase acting preferentially on hemimethylated substrates that result from DNA replication (Riggs, 1975). DNMT1 is generally ascribed this maintenance methylase activity, while DNMT3A and B are considered de novo methylases; however, understanding of how methylation patterns are established is limited (see Meehan, 2003 for a review). DNA methylation can regulate gene expression either by directly blocking transcription regulatory factors from binding to their target sequences or through several methyl-CpG-binding proteins that "read" DNA methylation patterns and form complexes with transcriptional repressor proteins including histone-modifying enzymes (Jones *et al*., 1998; Fuks *et al*., 2003). The CpG islands in the promoter regions of X-linked genes are heavily methylated on the Xi, in contrast to their unmethylated state on the Xa and the general state for autosomal genes. This epigenetic modification is one of the latest to be acquired during the process of X inactivation, following inactivation by several days and implicating methylation in maintenance rather than initiation of X-linked gene silencing (Keohane *et al*., 1996). X inactivation in marsupials and eutherian extraembryonic tissues, where there is reduced DNA methylation, is less stable (e.g., Kaslow and Migeon, 1987; Migeon *et al*., 1985). Similarly, patients with immunodeficiency-centromeric instability-facial anomalies syndrome (ICF, MIM 242860), which results from mutations in *DNMT3b*, have some hypomethylated CpG islands, and show reactivation of some X-linked loci in some cells (Hansen *et al*., 2000). Interestingly, they also show hypomethylation of LINE elements on only the Xi, suggesting that another methyltransferase is involved in LINE element methylation on the Xa and autosomes (Hansen, 2003). Furthermore, inhibition or

loss of DNA methyltransferase leads to derepression of X-linked genes, providing experimental evidence for the importance of DNA methylation in maintenance of X inactivation (Venolia *et al*., 1982; Sado *et al*., 2000). Synergism between XIST localization, DNA methylation, and histone modifications results in extremely stable silencing of the Xi (Csankovszki *et al*., 2001).

3. Process of inactivation

X inactivation can be envisioned as a cycle of events. Reactivation of the entire Xi occurs normally during oogenesis, while the X chromosome is transcriptionally inactivated together with the Y chromosome in spermatogenic cells shortly before or during early meiotic prophase. Such male meiotic sex chromosome inactivation (MSCI) has been suggested to mask the nonsynapsed regions of the sex chromosomes, possibly in order to avoid the checkpoint that causes germ cell death in response to defective chromosome synapsis; prevent the initiation of recombination events in the nonsynapsed regions of the sex chromosomes; or ensure efficient sex chromosome synapsis (reviewed in Turner *et al*., 2002). Inactivation occurs in female somatic cells at approximately the time of implantation during the late blastula stage, and is observed first in the extraembryonic tissues, where it is imprinted in mice (see below, and Article 41, **Initiation of X-chromosome inactivation**, Volume 1). Recent results suggest that the mouse X may be partially inactivated at the 2-cell stage (Huynh and Lee, 2003), and that early imprinted silencing is reversed in cells of the inner cell mass prior to random inactivation (Okamoto *et al*., 2003). As human X inactivation is not imprinted, it is not known to what extent events in very early development will be similar between humans and mice.

Much of our current understanding about the molecular mechanisms involved in the regulation of X inactivation is derived from the studies of mice, particularly ES cells that can undergo inactivation upon differentiation. A time line of inactivation suggests that first XIST expression is stabilized, coating the future Xi, and then histone modifications, gene silencing, asynchronous replication, and macroH2A accumulations occur (Chaumeil *et al*., 2002). The mechanism by which inactivation spreads in cis along the 160-megabase chromosome remains unsolved (*see* Article 40, **Spreading of X-chromosome inactivation**, Volume 1); however, it has been hypothesized that there are “way-stations” or booster elements distributed along the chromosome that serve to assist in the propagation of inactivation (Riggs *et al*., 1985). In X/autosome translocations, silencing can spread into the autosomal region; however, spread within the autosome does not seem to be as effective as on the X (White *et al*., 1998). Thus, way-stations are likely concentrated on the X, and Lyon has hypothesized that LI elements, which are enriched on the X, could be the way-stations (Lyon, 1998). Several bioinformatic studies have examined the characteristics of regions that escape inactivation (Bailey *et al*., 2000; Friel *et al*., 2002). However, as mentioned above, such regions may escape inactivation either because of failure to respond to the initial signal or inability to maintain silencing; and thus, the enrichment of particular elements

in these regions may reflect a role in either the spread or the maintenance of silencing.

3.1. The X inactivation center

A critical region necessary for inactivation of the X was defined by the analysis of the ability of rearranged X's to undergo inactivation. This region, the X inactivation center (*XIC*), was refined to 1 Mb of human Xq13 (Leppig *et al*., 1993), and more recent mouse transgenic studies suggest that the *Xic* is no more than 450-kb long (Lee *et al*., 1996). This region includes the *XIST/Xist* locus, as well as several cis-acting loci that influence the random choice of X to inactivate (reviewed in Rougeulle and Avner, 2003; and discussed below).

3.1.1. XIST

XIST/Xist, a large (>15 kb) alternatively processed, noncoding RNA, is the only gene known to be transcribed from the Xi but not from the Xa in somatic cells, and the RNA associates in cis with the Xi as part of the Barr body (Brown *et al*., 1991; Brockdorff *et al*., 1991; Clemson *et al*., 1996). The *Xist* gene is required for inactivation, and transgenes of *Xist* are able to induce inactivation of autosomes, identifying *Xist* as the principal component of the *Xic* (see Rougeulle and Avner, 2003). Prior to the onset of inactivation, *Xist* is unstable and expressed at a low level from all X chromosomes. Upon differentiation, *Xist* RNA is stabilized and coats the presumptive Xi, while the low-level pinpoint of *Xist* RNA on the Xa is subsequently extinguished (Panning *et al*., 1997; Sheardown *et al*., 1997).

Xist is known to be essential for the initiation of X inactivation based on targeted deletions in the mouse. Specifically, an *Xist* deficiency in female ES cells abolishes X inactivation potential in cis, enabling inactivation only of the chromosome carrying the wild-type *Xist* allele. Subsequent experiments have suggested that the *Xist* RNA levels determine which X undergoes inactivation (Nesterova *et al*., 2003). The continued expression and association of the transcripts with the Xi in somatic cells also implicated the RNA in the maintenance of the Xi; however, deletion of *XIST/Xist* after inactivation does not eliminate gene silencing (Brown and Willard, 1994; Rack *et al*., 1994; Csankovszki *et al*., 1999), although it does eliminate macroH2A association and increases the rate of reactivation of individual genes (Csankovszki *et al*., 2001). Reactivation in the absence of *Xist* is synergistically enhanced by treatment with demethylating or deacetylase-blocking agents, emphasizing the highly redundant nature of the maintenance of inactivation (Csankovszki *et al*., 2001).

Xist by itself, however, is not sufficient to induce chromosome-wide gene silencing in rodent somatic cells where reactivation of the endogenous *Xist* or induction of Xist-encoding transgenes produce localized Xist that does not inactivate the X (Clemson *et al*., 1998; Hansen *et al*., 1998; Rasmussen *et al*., 2001). Nevertheless, in human-transformed cells, it has been demonstrated that *XIST* expression can induce hallmarks of silencing, suggesting that the ability of *XIST/Xist* to induce inactivation may be species or cell-type specific (Hall *et al*.,

2002). Unexpectedly, in human/mouse hybrid cells, *XIST* is unable to localize to the human X, implicating species-specific factors in the chromatin recognition of *XIST* (Clemson *et al*., 1998; Hansen *et al*., 1998).

Xist RNA consists of separable domains for silencing and coating the chromosome in cis (Wutz *et al*., 2002). The transcriptional silencing activity of the transcript can be attributed to a domain at the 5′ end of the RNA that shows sequence conservation in all species in which *Xist* has been analyzed. An *Xist* RNA deleted for this 5′ domain coats, but does not silence, the X. Furthermore, only RNAs that were capable of coating the chromosome were functional in silencing. The separable motifs within *Xist* RNA probably recruit protein complexes to mediate coating and/or silencing of the X.

3.1.2. Tsix

A gene overlapping and antisense to *Xist*, named *Tsix*, has been identified in mouse ES cells; it is expressed from both X's before and during, but not after, differentiation (Lee *et al*., 1999; Debrand *et al*., 1999). *Tsix* initiates at a major transcription start site 13 kb downstream of the *Xist* 3′ end and includes four exons that undergo alternative splicing (Mise *et al*., 1999; Shibata and Lee, 2003). Like *Xist*, *Tsix* has no open reading frames, and its RNA is found in the nucleus. *Tsix* expression seems opposed to *Xist* expression, and truncation of the *Tsix* transcript led to complete nonrandom inactivation of the targeted X chromosome (Luikenhuis *et al*., 2001). Induction of *Tsix* transcription during ES cell differentiation, on the other hand, caused the targeted chromosome always to be chosen as the active chromosome (Stavropoulos *et al*., 2001). These results suggest that *Tsix* is a critical negative regulator of *Xist*. *Tsix* knockout mice show a dramatic parent-of-origin phenotype consistent with *Tsix* being a central player in imprinted inactivation (Lee, 2000). Although an antisense reminiscent of *Tsix* has been reported in humans, the roles of the human *TSIX* remain controversial (Migeon *et al*., 2001; Chow *et al*., 2003).

3.1.3. Xce and random versus imprinted X chromosome inactivation

In eutherian embryonic cells, the choice between inactivation of the maternally or paternally derived X is random. However, in marsupials, the paternal X is inactivated in all cells (Cooper *et al*., 1971), and a similar imprinted inactivation is found in the extraembryonic tissues of mice in which the paternal X is always inactive (Takagi and Sasaki, 1975). *Tsix* has been suggested to be the cis-acting imprinting factor that protects the maternal X from silencing, thus playing a critical role in imprinted X inactivation within developing mouse extraembryonic tissues (Lee, 2000; Sado *et al*., 2001). Preferential paternal inactivation in the extraembryonic tissues is also observed in cows (Xue *et al*., 2002). While there have been some conflicting results, it appears that despite some limited bias toward paternal inactivation in early trophoblast, inactivation in the extraembryonic tissues of humans is not imprinted (see review by Migeon, 2002).

In mice, the X controlling element (*Xce*), biases inactivation toward one X, with 20–30% distortions from random inactivation being observed in mice with

different alleles of the *Xce* locus (Cattanach and Issacson, 1967). *Xce* maps 3′ to *Xist* (Simmler *et al*., 1993), and the *Xite* locus has been proposed as a candidate for *Xce* activity. *Xite* is a region of intergenic transcription beyond *Tsix* that is associated with DNaseI hypersensitivity (Ogawa and Lee, 2003). A human equivalent to *Xce* has not been identified, although a screen of 38 families identified several families with a statistically significant skewing of X inactivation (Naumova *et al*., 1998). Mutations in the *XIST* promoter have been associated with deviations in X inactivation, but this strong skewing of inactivation was not consistent throughout families (Plenge *et al*., 1997; Tomkins *et al*., 2002). A mutational screen for modifiers of the X inactivation ratio in mouse identified two autosomal dominant loci, which have yet to be characterized (Percec *et al*., 2002).

4. Clinical consequences and considerations

X inactivation results in mammalian females being mosaics, composed of cell lines in which either the maternally or the paternally inherited X is inactivated. This mosaicism is generally sufficient to protect females from X-linked disease, but females may show "patchy" expression if the gene product is cell restricted, or disease, depending upon the randomness of their X inactivation pattern. There are also X-linked dominant disorders, many of which are believed to be male lethal (e.g., Rett syndrome, MIM 312750 or incontinentia pigmenti, MIM 308300), although others seem to manifest predominantly in females, possibly due to metabolic interference (e.g., craniofrontonasal syndrome MIM 304110 (Wieland *et al*., 2002)).

The process of inactivation also minimizes the impact of X aneuploidy, as all but one X will be inactivated. However, the lack of an Xi in females results in Turner syndrome, demonstrating that the Xi provides critical gene products for normal development, which are presumably provided by the Y in males. Female mice lacking an Xi show less phenotypic consequence, likely reflecting the fact that fewer genes are expressed from the Xi in mice (Disteche *et al*., 2002). The X/Y homologous genes represent only a small proportion of the genes that are expressed from the Xi in humans (Carrel *et al*., 1999), and thus, there is the possibility for considerable differential expression between males and females. The role of differential X-linked gene expression is quite difficult to separate from the substantial hormone-induced differences between males and females; however, there is a developing literature supporting difference in sex-chromosome gene expression between the sexes, particularly in brain (e.g., Xu *et al*., 2002; Good *et al*., 2003).

Deviations from the normal random inactivation pattern can occur either owing to primary influences upon the choice of X to inactivate or more likely owing to postinactivation selection for or against a population of cells with one X active (Migeon, 1998). A large number of assays have been developed that combine an X-linked polymorphism with a distinguishable feature of the Xi or Xa (generally methylation or expression) to allow determination of the extent of skewing of inactivation. In addition to being used to determine skewing in females suspected of carrying X-linked defects, these assays are often used for the analysis of clonality

in tumor tissue. If a tumor originates from a single cell, tumor tissues will have the same inactivation pattern in all cells. This nonrandom pattern of inactivation has been observed in a wide range of neoplastic tissue, demonstrating the monoclonal origin of the neoplasia (reviewed in Busque and Gilliland, 1998).

Although the analysis of inactivation patterns in females has been valuable in studies of carriers of X-linked disease and for determining the clonality of tumors, our limited understanding of the contribution of different causes of skewing confounds the utility of such studies. Skewed inactivation patterns have been reported in association with females with breast cancer, ovarian cancer, or females with recurrent spontaneous abortions, and these associations have yet to be satisfactorily explained (Buller *et al*., 1999; Kristiansen *et al*., 2002; Sangha *et al*., 1999, Lanasa *et al*., 1999, Uehara *et al*., 2001). One possibility is that X-linked mutations underlie such associations (e.g., Lanasa *et al*., 2001). Alternatively, skewing of inactivation may reflect a limitation in the number of viable cells that contribute to an embryo, as is the case for the skewed inactivation frequently observed in fetuses or newborns associated with confined placental mosaicism, in which a trisomic conceptus likely underwent a rescue event leading to a limited population of disomic cells (Lau *et al*., 1997). Thus, further studies are necessary for the determination of the cause or causes of these associations.

References

Adler DA, Rugarli EI, Lingenfelter PA, Tsuchiya K, Poslinski D, Liggit HD, Chapman VM, Elliott RW, Ballabio A and Disteche CM (1997) Evidence of evolutionary up-regulation of the single active X chromosome in mammals based on *Clc4* expression levels in *Mus spretus* and *Mus musculus*. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9244–9248.

Anderson C and Brown CJ (1999) Polymorphic X-chromosome inactivation of the human TIMP1 gene. *American Journal of Human Genetics*, **65**, 699–708.

Angelov D, Molla A, Perche PY, Hans F, Cote J, Khochbin S, Bouvet P and Dimitrov S (2003) The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. *Molecular Cell*, **11**(4), 1033–1041.

Ayling LJ and Griffin DK (2002) The evolution of sex chromosomes. *Cytogenetic and Genome Research*, **99**, 125–140.

Bailey JA, Carrel L, Chakravarti A and Eichler E (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6634–6639.

Barr ML and Bertram EG (1949) A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, **163**, 676–677.

Bochar DA, Wang L, Beniya H, Kinev A, Xue Y, Lane WS, Wang W, Kashanchi F and Shiekhattar R (2000) BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell*, **102**(2), 257–265.

Boggs BA and Chinault AC (1994) Analysis of replication timing properties of human X-chromosomal loci by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 6083–6087.

Brockdorff N, Ashworth A, Kay GF, Cooper P, Smith S, McCabe VM, Norris DP, Penny GD, Patel D and Rastan S (1991) Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome. *Nature*, **351**, 329–331.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R and Willard HF (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38–44.

Brown CJ and Willard HF (1994) The human X inactivation center is not required for maintenance of X inactivation. *Nature*, **368**, 154–156.

Buller RE, Sood AK, Lallas T, Buekers T and Skilling JS (1999) Association between nonrandom X-chromosome inactivation and BRCA1 mutation in germline DNA of patients with ovarian cancer. *Journal of the National Cancer Institute*, **91**, 339–346.

Busque L and Gilliland DG (1998) X-inactivation analysis in the 1990s: promise and potential problems. *Leukemia*, **12**, 128–135.

Carrel L, Cottle AA, Goglin KC and Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14440–14444.

Carrel L and Willard HF (1999) Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 7364–7369.

Cattanach BM and Issacson JH (1967) Contolling elements in the mouse X chromosome. *Genetics*, **57**, 331–346.

Chadwick BP and Willard HF (2001) Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant. *Human Molecular Genetics*, **10**, 1101–1113.

Chadwick BP and Willard HF (2003a) Barring gene expression after XIST: maintaining facultative heterochromatin on the inactive X. *Seminars in Cell & Developmental Biology*, **14**, 359–367.

Chadwick BP and Willard HF (2003b) Chromatin of the Barr body: histone and non-histone proteins associated with or excluded from the inactive X chromosome. *Human Molecular Genetics*, **12**, 2167–2178.

Chaumeil J, Okamoto I, Guggiari M and Heard E (2002) Integrated kinetics of X chromosome inactivation in differentiating embryonic stem cells. *Cytogenetic and Genome Research*, **99**(1–4), 75–84.

Chow JC, Hall LL, Clemson CM, Lawrence JB and Brown CJ (2003) Characterization of expression at the human XIST locus in somatic, embryonal carcinoma, and transgenic cell lines. *Genomics*, **82**(3), 309–322.

Clemson CM, Chow JC, Brown CJ and Lawrence JB (1998) Stabilization and localization of *Xist* RNA are controlled by separate mechanisms and are not sufficient for X inactivation. *Journal of Cell Biology*, **142**, 13–23.

Clemson CM, McNeil JA, Willard HF and Lawrence JB (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology*, **132**(3), 259–275.

Cohen SM, Brylawski BP, Cordeiro-Stone M and Kaufman DG (2003) Same origins of DNA replication function on the active and inactive human X chromosomes. *Journal of Cellular Biochemistry*, **88**(5), 923–931.

Cooper DW, VandeBerg JL, Sharman GB and Poole WE (1971) Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. *Nature: New Biology*, **230**, 155–157.

Costanzi C and Pehrson JR (1998) Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature*, **393**, 599–601.

Csankovszki G, Nagy A and Jaenisch R (2001) Synergism of *Xist* RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *Journal of Cell Biology*, **153**, 773–783.

Csankovszki G, Panning B, Bates B, Pehrson JR and Jaenisch R (1999) Conditional deletion of *Xist* disrupts histone macroH2A localization but not maintenance of X inactivation. *Nature Genetics*, **22**, 323–324.

Debrand E, Chureau C, Arnaud D, Avner P and Heard E (1999) Functional analysis of the DXPas34 locus, a 3′ regulator of Xist expression. *Molecular and Cellular Biology*, **19**, 8513–8525.

Disteche CM, Filippova GN and Tsuchiya K (2002) Escape from X inactivation. *Cytogenetic and Genome Research*, **99**, 36–43.

Disteche CM, Holzman T, Nguyen DK and Bumgarner R (2001) Microarray expression analysis supports upregulation of genes on the single active X chromosome, as compared to autosomal genes. *American Journal of Human Genetics*, **69**, S472.

Dyer KA, Canfield TK and Gartler SM (1989) Molecular cytological differentiation of active from inactive X domains in interphase: implications for X chromosome inactivation. *Cytogenetics and Cell Genetics*, **50**, 116–120.

Eils R, Dietzel S, Bertin E, Schrock E, Speicher MR, Ried T, Robert-Nicoud M, Cremer C and Cremer T (1996) Three-dimensional reconstruction of painted human interphase chromosomes: active and inactive X chromosome territories have similar volumes but differ in shape and surface structure. *Journal of Cell Biology*, **135**, 1427–1440.

Friel AM, Tsuchiya K, Ioshikhes IP, Fazzri M and Greally JM (2002) Evolutionary history and promoter characteristics interact to determine susceptibility to X chromosome inactivation. *American Journal of Human Genetics*, **71**, 184.

Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP and Kouzarides T (2003) The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *Journal of Biological Chemistry*, **278**(6), 4035–4040.

Ganesan S, Silver DP, Greenberg RA, Avni D, Drapkin R, Miron A, Mok SC, Randrianarison V, Brodie S, Salstrom J, *et al*. (2002) BRCA1 supports XIST RNA concentration on the inactive X chromosome. *Cell*, **111**(3), 393–405.

Gilbert SL, Pehrson JR and Sharp PA (2000) XIST RNA associates with specific regions of the inactive X chromatin. *Journal of Biological Chemistry*, **275**(47), 36491–36494.

Good CD, Lawrence K, Thomas NS, Price CJ, Ashburner J, Friston KJ, Frackowiak RS, Oreland L and Skuse DH (2003) Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. *Brain*, **126**(Pt 11), 2431–2446.

Goren A and Cedar H (2003) Replicating by the clock. *Nature Reviews. Molecular Cell Biology*, **4**(1), 25–32.

Graves JA, Gécz J and Hameister H (2002) Evolution of the human X - a smart and sexy chromosome that controls speciation and development. *Cytogenetic and Genome Research*, **99**, 141–145.

Hall LL, Byron M, Sakai K, Carrel L, Willard HF and Lawrence JB (2002) An ectopic human XIST gene can induce chromosome inactivation in postdifferentiation human HT-1080 cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(13), 8677–8682.

Hansen RS (2003) X inactivation-specific methylation of LINE-1 elements by DNMT3B: implications for the Lyon repeat hypothesis. *Human Molecular Genetics*, **12**(19), 2559–2567.

Hansen RS, Canfield TK, Fjeld AD and Gartler SM (1996) Role of late replication timing in the silencing of X-linked genes. *Human Molecular Genetics*, **5**, 1345–1353.

Hansen RS, Canfield TK, Stanek AM, Keitges EA and Gartler SM (1998) Reactivation of XIST in normal fibroblasts and a somatic cell hybrid: abnormal localization of XIST RNA in hybrid cells. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5133–5138.

Hansen RS, Stoger R, Wijmenga C, Stanek AM, Canfield TK, Luo P, Matarazzo MR, D'Esposito M, Feil R, Gimelli G, *et al*. (2000) Escape from gene silencing in ICF syndrome: evidence for advanced replication time as a major determinant. *Human Molecular Genetics*, **9**(18), 2575–2587.

Huynh KD and Lee JT (2003) Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature*, **426**(6968), 857–862.

Jenuwein T and Allis CD (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Jones PL, Veenstra GJC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J and Wolffe AP (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, **19**, 187–191.

Kaslow DC and Migeon BR (1987) DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X

dosage compensation. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 6210–6214.

Keohane AM, O'Neill LP, Belyaev ND, Lavender JS and Turner BM (1996) X-inactivation and histone H4 acetylation in embryonic stem cells. *Developmental Biology*, **180**, 618–630.

Kristiansen M, Langerod A, Knudsen GP, Weber BL, Borresen-Dale A-L and Orstavik KH (2002) High frequency of skewed X inactivation in young breast cancer patients. *Journal of Medical Genetics*, **39**, 30–33.

Kuroiwa A, Yokomine T, Sasaki H, Tsudzuki M, Tanaks K, Namikawa T and Matsuda Y (2002) Biallelic expression of Z-linked genes in male chickens. *Cytogenetic and Genome Research*, **99**, 310–314.

Lanasa MC, Hogge WA, Kubik C, Blancato J and Hoffman EP (1999) Highly skewed X-chromosome inactivation is associated with idiopathic recurrent spontaneous abortion. *American Journal of Human Genetics*, **65**(1), 252–254.

Lanasa MC, Hogge WA, Kubik CJ, Ness RB, Harger J, Nagel T, Prosen T, Markovic N and Hoffman EP (2001) A novel X chromosome-linked genetic cause of recurrent spontaneous abortion. *American Journal of Obstetrics and Gynecology*, **185**(3), 563–568.

Lau AW, Brown CJ, Penaherrera M, Langlois S, Kalousek DK and Robinson WP (1997) Skewed X-chromosome inactivation is common in fetuses or newborns associated with confined placental mosaicism. *American Journal of Human Genetics*, **61**, 1353–1361.

Lee JT (2000) Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell*, **103**, 17–27.

Lee JT, Davidow LS and Warshawsky D (1999) *Tsix*, a gene antisense to *Xist* at the X-inactivation center. *Nature Genetics*, **21**, 400–404.

Lee JT, Strauss WM, Dausman JA and Jaenisch R (1996) A 450 kb transgene displays properties of the mammalian X-inactivation center. *Cell*, **86**, 83–94.

Leppig KA, Brown CJ, Bressler SL, Gustashaw K, Pagon RA, Willard HF and Disteche CM (1993) Mapping of the distal boundary of the X-inactivation center in a rearranged X chromosome from a female expressing XIST. *Human Molecular Genetics*, **2**(7), 883–888.

Linder D and Gartler SM (1965) Glucose-6-phosphate dehydrogenase mosaicism: Utilization as a cell marker in the study of leiomyomas. *Science*, **150**, 67–69.

Lingenfelter PA, Adler DA, Poslinski D, Thomas S, Elliot RW, Chapman VM and Disteche CM (1998) Escape from X inactivation of Smcx is preceded by silencing during mouse development. *Nature Genetics*, **18**, 212–213.

Luikenhuis S, Wutz A and Jaenisch R (2001) Antisense transcription through the *Xist* locus mediates *Tsix* function in embryonic stem cells. *Molecular and Cellular Biology*, **21**, 8512–8520.

Lyon MF (1962) Sex chromatin and gene action in the mammalian X-chromosome. *American Journal of Human Genetics*, **14**, 135–145.

Lyon MF (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenetics and Cell Genetics*, **80**, 133–137.

Mager J, Montgomery ND, de Villena FP and Magnuson T (2003) Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nature Genetics*, **33**(4), 502–507.

Mak W, Baxter J, Silva J, Newall AE, Otte AP and Brockdorff N (2002) Mitotically stable association of polycomb group proteins eed and enx1 with the inactive X chromosome in trophoblast stem cells. *Current Biology*, **12**(12), 1016–1020.

Marin I, Siegal ML and Baker BS (2000) The evolution of dosage-compensation mechanisms. *Bioessays*, **22**(12), 1106–1114.

Meehan RR (2003) DNA methylation in animal development. *Seminars in Cell & Developmental Biology*, **14**(1), 53–65.

Meller VH, Kwok HW, Roman G, Kuroda MI and Davis RL (1997) roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell*, **86**, 445–457.

Migeon BR (1998) Non-random X chromosome inactivation in mammalian cells. *Cytogenetics and Cell Genetics*, **80**, 142–148.

Migeon BR (2002) X chromosome inactivation: theme and variations. *Cytogenetic and Genome Research*, **99**, 8–16.

Migeon BR, Chowdury AK, Dunston JA and McIntosh I (2001) Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *American Journal of Human Genetics*, **69**, 951–960.

Migeon BR, Wolf SF, Axelman J, Kaslow DC and Schmidt M (1985) Incomplete X chromosome dosage compensation in chorionic villi of human placenta. *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 3390–3394.

Mise N, Goto Y, Nakajima N and Takagi N (1999) Molecular cloning of antisense transcripts of the mouse *Xist* gene. *Biochemical and Biophysical Research Communications*, **258**, 537–541.

Mizuno H, Okamoto I and Takagi N (2002) Developmental abnormalities in mouse embryos tetrasomic for chromosome 11: apparent similarity to embryos functionally disomic for the x chromosome. *Genes & Genetic Systems*, **77**(4), 269–276.

Naumova AK, Olien L, Bird LM, Smith M, Verner AE, Leppert M, Morgan K and Sapienza C (1998) Genetic mapping of X-linked loci involved in skewing of X chromosome inactivation in the human. *European Journal of Human Genetics*, **6**, 552–562.

Nesterova TB, Johnston CM, Appanah R, Newall AE, Godwin J, Alexiou M and Brockdorff N (2003) Skewing X chromosome choice by modulating sense transcription across the *Xist* locus. *Genes & Development*, **17**(17), 2177–2190.

Ogawa Y and Lee JT (2003) *Xite*, X-inactivation intergenic transcription elements that regulate the probability of choice. *Molecular Cell*, **11**, 731–743.

Ohno S (1967) Sex chromosomes and sex-linked genes. In *Monographs on Endocrinology*, Vol. 1, Labhart A, Mann T, Samuels LT and Zander J (Eds.), Springer-Verlag: Berlin, p. 192.

Okamoto I, Otte AP, Allis CD, Reinberg D and Heard E (2003) Epigenetic Dynamics of Imprinted X Inactivation During Early Mouse Development. *Science*, **303**, 644–649.

Panning B, Dausman J and Jaenisch R (1997) X chromosome inactivation is mediated by RNA stabilization. *Cell*, **90**, 907–916.

Percec I, Plenge RM, Nadeau JH, Bartolomei MS and Willard HF (2002) Autosomal dominant mutations affecting X inactivation choice in the mouse. *Science*, **296**(5570), 1136–1139.

Perche P-Y, Vourc'h C, Konecny L, Souchier C, Robert-Nicoud M, Dimitrov S and Khochbin S (2000) Higher concentrations of histone macroH2A in the Barr body are correlated with higher nucleosome density. *Current Biology*, **10**, 1531–1534.

Peters AH, Mermoud JE, O'Carroll D, Pagani M, Schweizer D, Brockdorff N and Jenuwein T (2002) Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. *Nature Genetics*, **30**(1), 77–80.

Plenge RM, Hendrich BD, Schwartz C, Arena JF, Naumova A, Sapienza C, Winter RM and Willard HF (1997) A promoter mutation in the XIST gene in two unrelated families with skewed X-chromosome inactivation. *Nature Genetics*, **17**, 353–356.

Rack KA, Chelly J, Gibbons RJ, Rider S, Benjamin D, Lafreniere RG, Oscier D, Hendriks RW, Craig IW, Willard HF, *et al*. (1994) Absence of the XIST gene from late-replicating isodicentric X chromosomes in leukemia. *Human Molecular Genetics*, **3**(7), 1053–1059.

Rasmussen TP, Wutz A, Pehrson JR and Jaenisch R (2001) Expression of *Xist* RNA is sufficient to initiate macrochromatin body formation. *Chromosoma*, **110**, 411–420.

Riggs AD (1975) X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics*, **14**, 9–25.

Riggs AD, Singer-Sam J and Keith DH (1985) Methylation of the PGK promoter region and an enhancer way-station model for X-chromosome inactivation, in *Biochemistry and Biology of DNA Methylation*, Vol. 198, Alan R. Liss, Inc.: New York.

Rougeulle C and Avner P (2003) Controlling X-inactivation in mammals: what does the centre hold? *Seminars in Cell & Developmental Biology*, **14**, 331–340.

Rougeulle C, Navarro P and Avner P (2003) Promoter-restricted H3 Lys 4 di-methylation is an epigenetic mark for monoallelic expression. *Human Molecular Genetics*, **12**(24), 3343–3348.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK and Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, **423**(6942), 873–876.

Sado T, Fenner MH, Tan SS, Tam P, Shioda T and Li E (2000) X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Developmental Biology*, **225**(2), 294–303.

Sado T, Wang Z, Sasaki H and Li E (2001) Regulation of imprinted X-chromosome inactivation in mice by *Tsix*. *Development*, **128**, 1275–1286.

Sangha KK, Stephenson MD, Brown CJ and Robinson WP (1999) Extremely skewed X-chromosome inactivation is increased in women with recurrent spontaneous abortion. *American Journal of Human Genetics*, **65**, 913–917.

Schempp W and Meer B (1983) Cytologic evidence for three human X-chromosomal segments escaping inactivation. *Human Genetics*, **63**, 171–174.

Sheardown SA, Duthie SM, Johnston CM, Newall AET, Formstone EJ, Arkell RM, Nesterova TB, Alghisi G-C, Rastan S and Brockdorff N (1997) Stabilisation of *Xist* RNA mediates initiation of X chromosome inactivation. *Cell*, **91**, 99–107.

Shibata S and Lee JT (2003) Characterization and quantitation of differential *Tsix* transcripts: implications for *Tsix* function. *Human Molecular Genetics*, **12**, 125–136.

Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, Peters AH, Jenuwein T, Otte AP and Brockdorff N (2003) Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Developmental Cell*, **4**(4), 481–495.

Simmler M-C, Cattanach B, Rasberry C, Rougeulle C and Avner P (1993) Mapping the murine Xce locus with (CA)n repeats. *Mammalian Genome*, **4**, 523–530.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, *et al*. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, **423**(6942), 825–837.

Stavropoulos N, Lu N and Lee JT (2001) A functional role for *Tsix* transcription in blocking *Xist* RNA accumulation but not in X-chromosome choice. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10232–10237.

Takagi N and Sasaki M (1975) Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, **256**, 640–642.

Tomkins DJ, McDonald HL, Farrell SA and Brown CJ (2002) Lack of expression of XIST from a small ring X chromosome containing the XIST locus in a girl with short stature, facial dysmorphism and developmental delay. *European Journal of Human Genetics*, **10**(1), 44–51.

Tsuchiya KD and Willard HF (2000) Chromosomal domains and escape from X inactivation: comparative X inactivation analysis in mouse and human. *Mammalian Genome*, **11**, 849–854.

Turner JM, Mahadevaiah SK, Elliott DJ, Garchon HJ, Pehrson JR, Jaenisch R and Burgoyne PS (2002) Meiotic sex chromosome inactivation in male mice with targeted disruptions of *Xist*. *Journal of Cell Science*, **115**(Pt 21), 4097–4105.

Uehara S, Hashiyada M, Sato K, Sato Y, Fujimori K and Okamura K (2001) Preferential X-chromosome inactivation in women with idiopathic recurrent pregnancy loss. *Fertility and Sterility*, **76**(5), 908–914.

Venolia L, Gartler SM, Wassman ER, Yen P, Mohandas T and Shapiro LJ (1982) Transformation with DNA from 5-azacytidine-reactivated X chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 2352–2354.

Wang J, Mager J, Chen Y, Schneider E, Cross JC, Nagy A and Magnuson T (2001) Imprinted X inactivation maintained by a mouse Polycomb group gene. *Nature Genetics*, **28**(4), 371–375.

White WM, Willard HF, Van Dyke DL and Wolff DJ (1998) The spreading of X inactivation into autosomal material of an X; autosome translocation: evidence for a difference between autosomal and X-chromosomal DNA. *American Journal of Human Genetics*, **63**, 20–28.

Wieland I, Jakubiczka S, Muschke P, Wolf A, Gerlach L, Krawczak M and Wieacker P (2002) Mapping of a further locus for X-linked craniofrontonasal syndrome. *Cytogenetic and Genome Research*, **99**(1–4), 285–288.

Willard HF (1977) Tissue-specific heterogeneity in DNA replication patterns of human X chromosomes. *Chromosoma*, **61**, 61–73.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I and Carter NP (2004) Replication timing of the human genome. *Human Molecular Genetics*, **13**(2), 191–202.

Wutz A, Rasmussen TP and Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nature Genetics*, **30**(2), 167–174.

Xu J, Burgoyne PS and Arnold AP (2002) Sex differences in sex chromosome gene expression in mouse brain. *Human Molecular Genetics*, **11**(12), 1409–1419.

Xue F, Tian XC, Du F, Kubota C, Taneja M, Dinnyes A, Dai Y, Levine H, Pereira LV and Yang X (2002) Aberrant patterns of X chromosome inactivation in bovine clones. *Nature Genetics*, **31**(2), 216–220.

Zhang J, Xu F, Hashimshony T, Keshet I and Cedar H (2002) Establishment of transcriptional competence in early and late *S* phase. *Nature*, **420**(6912), 198–202.

Short Specialist Review

Nondisjunction

Neil E. Lamb
Emory University, Atlanta, GA, USA

Nondisjunction is the improper segregation of chromosomes during either meiotic or mitotic cell divisions. The result of nondisjunction is dosage imbalance – either one extra or one missing chromosome for some proportion of the resulting cells. Mitotic nondisjunction is commonly observed in cancer cells (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1), while nondisjunction during meiosis results in gametes that are chromosomally unbalanced (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1). If an unbalanced gamete participates in fertilization, the resulting embryo will be aneuploid, with either one chromosome too many (trisomy) or one chromosome too few (monosomy). Such embryos are generally inviable, and for humans, this has serious clinical consequences. As the most commonly identified chromosome abnormality, identified in at least 5% of all clinically recognized pregnancies, aneuploidy is the leading known cause of pregnancy loss and fetal wastage. Among conceptions that survive to term (primarily trisomy for chromosomes 13, 18, and 21 and various combinations of the sex chromosomes), aneuploidy is the leading genetic cause of mental impairment and developmental disabilities.

Unlike humans, nondisjunction in most organisms is a very rare occurrence. For example, chromosome missegregation is estimated to occur in only 1 of every 10 000 meiotic events for *Saccharomyces cerevisiae* (Sears *et al*., 1992). The underlying basis for the increased incidence in humans is unclear as the basic meiotic pathway is highly conserved across species. Meiosis generates haploid gametes through a process that consists of one round of DNA replication followed by two cellular divisions (*see* Article 13, **Meiosis and meiotic errors**, Volume 1). The first cell division (known as *meiosis I* or *MI*) separates homologous chromosomes (Figure 1), while meiosis II (MII) segregates the sister chromatids of each homolog. Nondisjunction can occur at either of these meiotic divisions, and the alternatives can often be distinguished using polymorphic genetic markers at or near the centromere of the nondisjoined chromosome. If both copies of the nondisjoined chromosome are heterozygous for alleles at these markers, it is likely that the error arose at MI. In contrast, homozygosity at the centromere of the nondisjoined chromosomes suggests an error at MII. In much the same way, marker studies can be used to determine in which parent the nondisjunction error occurred. As a result, numerous aneuploid conditions have been studied to

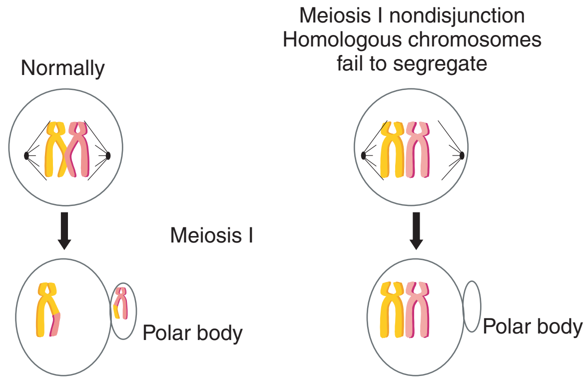


Figure 1 During the first stage of meiosis, homologous chromosomes segregate, traveling to opposite poles of the spindle. In humans, one homolog is placed in the polar body, while the other remains in the maturing oocyte. Nondisjunction at the first meiotic division results in both homologs traveling toward the same spindle pole. This is frequently associated with altered patterns of meiotic recombination along the nondisjoined homologs: generally either poor positioning of existing recombination or a lack of recombination altogether

determine the parental origin and meiotic stage of the nondisjunction error. Very little data has been identified for monosomic conditions, which appear to result in early embryonic lethality. In contrast, most trisomic conditions are compatible with at least some fetal development and results are available for several trisomies (reviewed in Hassold and Hunt, 2001). These data show remarkable variation with respect to parental origin and meiotic stage. Nonetheless, it is evident that maternal meiosis I nondisjunction errors predominate among nearly all trisomic conditions. This is perhaps not surprising, given that the first stage of meiosis in females is amazingly protracted: it is initiated prenatally in all oocytes but arrested shortly thereafter and is not resumed until just before the oocyte is ovulated, 15 to 50 years later. In addition, increasing evidence suggests that meiotic disturbances are handled differently between males and females (reviewed in Hunt and Hassold, 2002). Abnormalities that often lead to prophase or metaphase–anaphase arrest in male meiosis appear to escape detection in female meiosis, leading to nondisjunction. Consequently, it is clear that an understanding of the risk factors for human aneuploidy will require a greater understanding of the events of meiosis I in the human female.

Despite the high frequency and obvious clinical importance of nondisjunction, there is a lack of knowledge about the predisposing genetic and environmental factors. However, there is one factor incontrovertibly linked to human aneuploidy – increasing maternal age. Most, if not all, human trisomies are affected by increasing maternal age, although the magnitude of the effect varies among trisomies (Risch *et al*., 1986; Morton *et al*., 1988). Among women under the age of 25 years, ~2% of all clinically recognized pregnancies are trisomic but this frequency approaches 35% for women over the age of 40. Very little is understood regarding the mechanisms that underlie this age effect. It is thought to involve meiosis I, consistent with the studies described above examining the parent and meiotic stage

of the nondisjunction event. However, the specific timing of the event is unclear, and numerous models have been advanced to describe when in meiosis I the age effect develops (for review see Gaulden, 1992).

Recently, altered genetic recombination has been identified as the first molecular correlate of human nondisjunction. In model organisms, mutations that affect the amount and location of meiotic recombination are associated with nondisjunction. It appears that altered amounts and placement of recombination in humans also increases the risk for nondisjunction. It is possible to recapitulate the recombination patterns that occurred in meiotic events that ultimately lead to nondisjunction. Studies of this type have identified significant reductions in recombination for all meiosis I trisomies studied to date, including maternally derived cases of trisomies 15, 16, 18, 21, and sex-chromosome trisomies (Hassold *et al*., 1995; Lamb *et al*., 1996, 1997; Bugge, 1998; Robinson, 1998). Reduced recombination has also been identified in paternally derived cases of trisomy 21 and Klinefelter syndrome (47, XXY) (Savage, 1998; Thomas, 2000). For many of these trisomies, the reduction in recombination has been due to a proportion of cases that never engaged in genetic recombination. In addition, altered placement of meiotic recombination has been identified for a subset of maternally derived trisomies 21 and 16, with the exchange events occurring more distal than expected. These "distal only" exchange events appear to be less efficient at proper chromosome segregation.

Surprisingly, altered patterns of meiotic recombination are also associated with maternal meiosis II trisomy for chromosome 21 (Lamb, 1997). Specifically, the increase was most pronounced in the pericentromeric region of the chromosome. The connection between recombination (a meiosis I event) and the meiosis II nondisjunction may be explained if the meiosis II errors actually originated in meiosis I by chromosome entanglement due to the pericentromeric exchange or premature separation of sister chromatids at meiosis I followed by comigration of the sister chromatids due to chance at meiosis II. As a result, the nondisjoined chromosomes would have identical centromeres, leading to the "meiosis II" classification even though the event was initiated at meiosis I. Subsequent studies of trisomy 18 or sex-chromosome trisomy arising at maternal meiosis II have not, however, identified a significant effect for recombination. As a result, it appears likely that the liability posed by altered recombination varies from chromosome to chromosome.

Do other risk factors exist for nondisjunction? When compared to the maternal age effect, additional proposed risks have a much weaker or unsubstantiated influence. Many such factors have been proposed, including: the effects of diminished oocyte pool leading to increased risk of nondisjunction and earlier maternal menopause (Kline, 2000); a reduced ovarian complement, due to the congenital absence or surgical removal (Freeman, 2000); maternal smoking and oral contraceptive use around the time of conception (Yang, 1999); and maternal polymorphisms at genes in the folate pathway (van der Put, 1998; Hobbs, 2000). In addition, various chemical, drug, or irradiation exposures have also been evaluated (Hook, 1992). To date, however, none of these potential risk factors have been conclusively proven. This does not mean that there are no relevant risk factors – possibly the impact is so small that each risk escapes detection. On the other hand, the correct risk factors may have simply not been identified. Regardless,

an understanding of the molecular mechanisms for nondisjunction and the maternal age effect remains elusive.

Further reading

James SJ, Pogribna M, Pogribny IP, Melnyk S, Hine RJ, Gibson JB, Yi P, Tafoya DL, Swenson DH, Wilson VL, *et al*. (1999) Abnormal folate metabolism and mutation in the methylenetetrahydrofolate reductase gene may be maternal risk factors for Down syndrome. *American Journal of Clinical Nutrition*, **70**, 495–501.

References

Bugge M, Collins A, Petersen MB, Fisher J, Brandt C, Hertz JM, Tranebjaerg L, de Lozier-Blanchet C, Nicolaides P, Brondum-Nielsen K, *et al*. (1998) Nondisjunction of chromosome 18. *Human Molecular Genetics*, **7**, 661–669.

Freeman SB, Yang Q, Allran K, Taft LF and Sherman SL (2000) Women with a reduced ovarian complement may have an increased risk for a child with Down syndrome. *American Journal of Human Genetics*, **66**, 1680–1683.

Gaulden ME (1992) Maternal age effect: the enigma of Down syndrome and other trisomic conditions. *Mutation Research*, **296**, 69–88.

Hassold T and Hunt P (2001) To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews. Genetics*, **2**, 280–291.

Hassold T, Merrill M, Adkins K, Freeman S and Sherman S (1995) Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *American Journal of Human Genetics*, **57**, 867–874.

Hobbs CA, Sherman SL, Yi P, Hopkins SE, Torfs CP, Hine RJ, Pogribna M, Rozen R and James SJ (2000) Polymorphisms in genes involved in folate metabolism as maternal risk factors for Down syndrome. *American Journal of Human Genetics*, **67**, 623–630.

Hook EB. (1992) Chromosome abnormalities: Prevalence, risks and recurrence. In *Prenatal Diagnosis and Screening*, Brock DH, Rodeck CH, Ferguson-Smith MA (Eds.), Churchill Livingstone: Edinburgh, pp. 351.

Hunt PA and Hassold TJ. (2002) Sex matters in meiosis. *Science*, **296**, 2181–2183.

Kline J, Kinney A, Levin B and Warburton D (2000) Trisomic pregnancy and earlier age at menopause. *American Journal of Human Genetics*, **67**, 395–404.

Lamb NE, Feingold E, Savage-Austin A, Avramopoulos D, Freeman SB, Gu Y, Hallberg A, Hersey J, Pettay D, Saker D, *et al*. (1997) Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21. *Human Molecular Genetics*, **6**, 1391–1399.

Lamb NE, Freeman SB, Savage-Austin A, Avramopoulos D, Gu Y, Hallberg A, Hersey J, Pettay D, May KM, Saker D, *et al*. (1996) Susceptible chiasmate configurations of chromosome 21 predispose to nondisjunction in both maternal meiosis I and meiosis II. *Nature Genetics*, **14**, 400–405.

Morton NE, Jacobs PA, Hassold T and Wu D (1988) Maternal age in trisomy. *Annals of Human Genetics*, **52**, 227–235.

Risch N, Stein Z, Kline J and Warburton D (1986) The relationship between maternal age and chromosome size in autosomal trisomy. *American Journal of Human Genetics*, **39**, 68–78.

Robinson WP, Bernasconi F, Mutiranguara A, Ledbetter DH, Langlois S, Malcolm S, Morris MA and Schinzel AA (1998) Nondisjunction of chromosome 15: origin and recombination. *American Journal of Human Genetics*, **53**, 740–751.

Savage AR, Petersen MB, Pettay D, Taft L, Allran K, Freeman SB, Karadima G, Avramopoulos D, Torfs C, Mikkelsen M, *et al*. (1998) Elucidating the mechanisms of paternal non-disjunction of chromosome 21 in humans. *Human Molecular Genetics*, **7**, 1221–1227.

Sears DD, Hegemann JH and Hieter P (1992) Meiotic recombination and segregation of human-derived artificial chromosomes in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 5296–5300.

Thomas NS, Collins AR, Hassold TJ and Jacobs PA (2000) A reinvestigation of non-disjunction resulting in 47,XXY males of paternal origin. *European Journal of Human Genetics*, **8**, 805–808.

van der Put NM, Gabreels F, Stevens EM, Smeitink JA, Trijbels FJ, Eskes TK, van den Heuvel LP, Blom HJ (1998) A second common mutation in the methylenetetrahydrofolate reductase gene: an additional risk factor for neural-tube defects? *American Journal of Human Genetics*, **62**(5), 1044–1051.

Yang Q, Sherman SL, Hassold TJ, Allran K, Taft L, Pettay D, Khoury MJ, Erickson JD and Freeman SB (1999) Risk factors for trisomy: maternal cigarette smoking and oral contraceptive use in a population-based case-control study. *Genetics in Medicine*, **1**, 80–88.

Short Specialist Review

Microdeletions

John A. Crolla
Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, UK

In the 1980s, the development of in situ hybridization techniques, particularly those utilizing fluorescent markers (FISH) (*see* Article 22, **FISH**, Volume 1) (Pinkel *et al*., 1988), led to the discovery of the first cryptic deletion syndromes, that is, those in which the missing material was not visible using conventional microscopy. Discovery of most microdeletion syndromes resulted from collaborations between clinicians, cytogeneticists, and molecular geneticists. Chromosome markers such as translocation breakpoints found in patients with abnormal phenotypes were often pivotal in identifying the chromosomal region of interest following which positional cloning methods resulted in the identification and characterization of the gene(s) of interest (Tommerup, 1993). For example, the first direct evidence that hemizygosity at the *ELN* locus contributes to the Williams syndrome phenotype followed a report that a t(6;7)(p21.1;q11.23) translocation was segregating in a family with dominant supra valvular aortic stenosis (SVAS) (Ashkenas, 1996). Subsequent studies have identified a number of genes implicated in the William's contiguous gene phenotype, and in common with many other human microdeletion syndromes, hemizygosity at one or more loci leading to the disruption of expression of dosage-sensitive genes appears to be the principal mutational mechanism underlying the clinical phenotypes.

There have been over 60 reported microdeletion syndromes (*see* Article 87, **The microdeletion syndromes**, Volume 2) involving virtually all chromosomes (Table 1). Although haploinsufficiency of dosage-sensitive genes is thought to be the main mutational mechanism(s) underlying the associated clinical phenotype(s), there are rare examples where the deletion's breakpoints per se appear to be the mutational mechanism mediated by the breakpoint disrupting a regulatory element that controls a nearby gene (e.g, $3'$ interstitial deletions in cases with aniridia; Lauderdale *et al*., 2000; Crolla and van Heyningen, 2002).

Several terminal, or more correctly, subtelomeric deletions have been described following analysis of patients with normal karyotypes using chromosome-specific sub-telomere FISH probes (Knight *et al*., 1997). These studies were instrumental in defining the so-called terminal deletion syndromes that are distributed throughout the genome and were found in ~5% patients with idiopathic developmental delay with or without associated congenital abnormalities (de Vries *et al*., 2003). A proportion of the deletions observed in these studies were the unbalanced products

of balanced parental rearrangements (usually translocations), and reciprocal duplications have also been reported.

The molecular mechanisms underlying the formation of at least some of the cryptic rearrangements are being elucidated, some of which are thought to arise following unequal meiotic recombination mediated by the proximity of nonallelic segmental duplications particularly in the pericentromeric and telomeric regions (Emanuel and Shaikh, 2001). However, alternative mechanisms may operate. For example, the presence in some mothers with children with Angelman's syndrome and del(15)(q11-q13) of a submicroscopic heterozygous inversion at the regions defined by flanking segmental duplications has been proposed to represent "an intermediate estate" that facilitates the formation of a deletion in an offspring (Gimelli *et al*., 2003). Clearly, genomic organization has a pivotal role in determining the position and frequency of specific microdeletion syndromes and the several models proposed to date may not be exhaustive (Emanuel and Shaikh, 2001).

The discussion so far has focused on techniques used to identify imbalances at specific genomic locations. Leading up to the publication of the advanced draft assembly of the Human Genome Sequence in 2001, attention switched to the innovative use of fully sequenced clones (principally Bacterial or P1 Artificial Chromosomes, BAC/PACs). A method called "Array-CGH" was developed (*see* Article 55, **Polymorphic inversions, deletions, and duplications in gene mapping**, Volume 1) in which selected BAC/PAC clones (with an average insert size of ~150 Kb) were selected sequentially at contiguous genomic locations ~1 Mb apart and then spotted onto glass slides to act as targets for a comparative genomic hybridization (CGH) experiment using a probe mixture comprising differentially labeled DNA derived from both normal and clinically affected individuals (Fiegler *et al*., 2003). This approach opened up the possibility of detecting copy number changes either throughout the genome with a resolution only constrained by the density of the clones selected (i.e., at 1 Mb or complete coverage using tiling path clones (Ishkanian *et al*., 2004)) or at high density but focused on specific genomic regions.

A number of studies utilizing Array-CGH have illustrated the potential of this approach but has also provided unexpected insights into the number and distribution of large genomic imbalances distributed throughout the genome that may turn out to be nonpathogenic polymorphisms (see below). Array-CGH focused on specific chromosomal regions has been used to define critical deletion regions for specific phenotypes such as Congenital Aural Atresia in patients with 18q22.3-18q23 deletions (Veltman *et al*., 2003), 1p36 deletions (Yu *et al*., 2003), and imbalances involving 17p11-p12 (Shaw *et al*., 2004). Targeted Array-CGH led to the refinement of a critical deletion interval in 8q12 in some patients with CHARGE syndrome (a developmental disorder with a nonrandom pattern of congenital abnormalities). Sequence analysis of genes found in the deleted segment found mutations in *CHD7* in 10 out of 17 patients with CHARGE syndrome (Vissers *et al*., 2004), thereby illustrating the utility of Array-CGH for the identification of some novel disease-related gene loci.

Array-CGH has also been used in whole-genome scans using BACs at 1-Mb resolution on chromosomally normal patients with idiopathic developmental delay or mental retardation with and without associated congenital abnormalities. The

Table 1 Microdeletion syndromes

Chromosome	Locus/gene	Syndrome
1p36.3	*DVL1*	Monosomy 1p36
2p21	*SIX3*	Holoprosencephaly 2
2q13	*NPHP1*	Nephronophthisis 6
2q22.3	*ZFHX1B*	Mowat–Wilson
2q37	*HPE6*	Holoprosencephaly 6
2q37	*AHO*	Albright hereditary osteodystrophy
2q31-32	*HOXD9; HOXD13 EVX2*	Synpolydactyly
3p23	*VHL*	Von-Hippel Lindau
3q23	*BPES*	Blepharophimosis-ptosis-epicanthus-inversus
4p16	*WHS*	Wolf–Hirschorn
4q25	*PIX2*	Rieger
5p15	*CDC*	Cri-du-Chat
5q13.2	*NIPBL*	Cornelia de Lange
5q22.2	*APC*	Familial adenomatous polyposis
5q35.2	*MSX2*	Parietal Foramina
5q35	*NSD1*	Sotos
7p21	*TWIST1*	Saethre–Chotzen
7p13	*GLI3*	Greig cephalopolysyndactyly
7q11.23	*ELN*	Williams
7q36.3	*SHH*	Holoprosencaphaly 3
7q36.3	*HLXB9*	Sacral agenesis
8q12	*CHD7*	CHARGE
8q24.1	*EXT1*	Langer Giedion
8q24.1	*TRPS1*	Trichorhinophalangeal syndrome
9q22.3	*PTCH*	Holoprosencaphaly 7
10p15	*GATA3*	Hypoparathryoidism, deafness, renal dysplasia
10p13	*DGSII*	DiGeorge
11p15.5	*IGF2*	Beckwith–Wiedemann
11p13	*PAX6*	Aniridia
11p13	*WT1*	Wilms tumor
11p13	*WAGR*	Wilms, Aniridia, genital anomalies, retardation
11p11.2	*EXT2*	Multiple exostoses
11p11.2	*ALX4*	Parietal Foramina contiguous gene deletion
11q23-24.1	*MLL/ETS1*	Jacobsen
12q24.1	*PTPN11*	Noonan
13q14	*RB1*	Retinoblastoma
13q32	*ZIC2*	Holoprosencaphaly 5
14q12-13	*PAX9*	Autosomal dominant hypodontia
15q13	*SNRPN/UBE3A*	Prader–Willi/Angelman
16p13.3	*CREBBP*	Rubinstein–Taybi
16p13.3	*TSC2*	Tuberous sclerosis 2
16p13.3	*PKD1*	Polycystic kidney disease
17p13.3	*LIS1*	Miller–Dieker
17p11.2	*RAI1*	Smith Magenis
17p11.2	*PMP22*	Hereditary neuropathy with liability to pressure palsy
17q11.2	*NF1*	Neurofibromatosis 1
17q12-21	*SOST*	Van Buchem disease
18p11.3	*TGIF*	Holoprosencaphaly 4
20p11.23	*JAG1*	Alagille
21q22.3	*TMEM1*	Holoprosencaphaly 1

(*continued overleaf*)

Table 1 (*continued*)

Chromosome	Locus/gene	Syndrome
22q11.2	*TUPLE1;TBX1*	DiGeorge I
22q13.3	*ProSAP2*	Monosomy 22q13.3
Xp22.33Yp11.32	*SHOX*	Leri–Weill dyschondrosteosis, Madelung deformity, short stature
Xp22.33	*STS*	Steroid sulfatase
Xp22.33	*KAL1*	Kallmann
Xp22.33	*MLS*	Micopthalmia and linear skin defects
Xp21.2	*DAX1*	Adrenal hypoplasia congenita
Xp21	*GK*	Glycerol kinase deficiency
Xp21	*DMD*	Duchenne muscular dystrophy
Xq22	*PLP*	Pelizaeus–Merzbacher disease
Xq25	*XLP*	X-linked lymphoproliferative disease
Xq26.2	*ZIC3*	X-linked hererotaxy
Yp11.32	*SRY*	Ambiguous genitalia, sex reversal

first studies reported similar results (Table 2). Shaw-Smith *et al*. (2004) and Vissers *et al*. (2003) used 1-Mb Array-CGH on similar patient groups and showed that ∼25% carried genomic imbalances, that is, deletions or duplications. One-third of these imbalances were subsequently found to be carried by a clinically normal parent and were considered to be polymorphic copy number changes. However, the remaining two-thirds were novel de novo microdeletions/duplications (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1), which further studies will determine whether they are directly responsible for the clinical phenotypes observed.

Two subsequent papers reported large-scale copy number variations (deletions and duplications) ranging in size from 100 Kb to 2 Mb in two ethnically and geographically distinct populations of normal control individuals (Table 2). Iafrate *et al*. (2004), using Array-CGH, reported 255 variable loci in 55 individuals compared to Sebat *et al*. (2004) who used an oligonucleotide array and demonstrated 76 genomic changes in 20 individuals. Interpretation of what constitutes a pathogenic genomic imbalance compared to a polymorphism, particularly those involving microdeletions, will only be possible once the technical parameters, including the

Table 2 Number and distribution of large genomic imbalances in affected and control populations

A. Clinically affected populations

Study	Method	No. Del^n	No. Dups	Total	% imbalances
Vissers *et al*. (2003)	A-CGH	3	2	20	25
Shaw-Smith *et al*. (2004)	A-CGH	7	5	50	24
Total *(de novo)*		10* (14%) *(8)*	7* (10%) *(1)*	70	24

B. Clinically normal populations

Study	Method	No studied	Resolution (Kb)	Total	No (%) imbalances
Sebat *et al*. (2004)	ROMA	20	>100	221 CNPs (ave: 11.5)	70 (32%)
Iafrate *et al*. (2004)	Array-CGH	55	>150	255 LCVs (ave: 6.5)	142 (55%)

choice and density of clones used in the array, as well as an understanding of the sequences involved in "common" imbalances are fully understood (Carter, 2004).

In conclusion, our understanding of the clinical effects and mutational mechanisms associated with genomic imbalances, particularly deletions, has come a long way since the first cytogenetic karyotype/phenotype correlations in the 1960s and the descriptions of the first microdeletion syndromes in the 1980s. The future development of high-density genomic arrays promises further insights into the complex structure of the "normal" human genome, from which novel microdeletion syndromes will emerge providing new diagnostic strategies and insights into disease-associated human gene(s).

References

Ashkenas J (1996) Williams syndrome starts making sense. *American Journal of Human Genetics*, **59**, 756–762.

Carter NP (2004) As normal as normal can be? *Nature Genetics*, **36**, 931–932.

Crolla JA and van Heyningen V (2002) Frequent chromosome aberrations revealed bu molecular cytogenetic studies in patients with aniridia. *American Journal of Human Genetics*, **71**, 1138–1149.

de Vries BBA, Winter R, Schinzel A and van Ravenswaaij-Arts C (2003) Telomeres: a diagnosis at the end of the chromosomes. *Journal of Medical Genetics*, **40**, 385–398.

Emanuel BS and Shaikh TH (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature Reviews. Genetics*, **2**, 791–801.

Fiegler H, Douglas EJ, Carr P, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IP, *et al*. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes, Chromosomes & Cancer*, **36**, 361–374.

Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X, *et al*. (2003) Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Human Molecular Genetics*, **12**, 849–858.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **36**, 949–951.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

Knight SJL, Horsley SW, Regan R, Lawrie NM, Maher EJ, Cardy DLN, Flint J and Keamey L (1997) Development and clinical application of an innovative fluorescence in situ hybridization technique which detects submicroscopic rearrangements involving telomeres. *European Journal of Human Genetics*, **5**, 1–9.

Lauderdale J, Wilensky JS, Oliver ER, Walton DS and Glaser T (2000) 3′ deletions cause aniridia by preventing PAX6 gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 13755–13760.

Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J and Gray JW (1988) Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 9138–9142.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shaw CJ, Shaw CA, Yu W, Stankiewicz P, White LD, Beaudet AL and Lupski JR (2004) Comparative genomic hybridisation using a proximal 17p BAC/PAC array detects

rearrangements responsible for four genomic disorders. *Journal of Medical Genetics*, **41**, 113–119.

Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, *et al*. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *Journal of Medical Genetics*, **41**, 241–248.

Tommerup N (1993) Mendelian cytogenetics – chromosome rearrangements associated with Mendelian disorders. *Journal of Medical Genetics*, **30**, 713–727.

Veltman JA, Jonkers Y, Nuijten I, Janssen I, Van Der Vliet W, Huys E, Vermeesch J, Van Buggenhout G, Fryns JP, Admiraal R, *et al*. (2003) Definition of a critical region on chromosome 18 for congenital aural atresia by arrayCGH. *American Journal of Human Genetics*, **72**, 1578–1584.

Vissers LE, de Vries BB, Osoegawa K, Janssen IM, Feuth T, Choy CO, Straatman H, van der Vliet W, Huys EH, van Rijk A, *et al*. (2003) Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *American Journal of Human Genetics*, **73**, 1261–1270.

Vissers LE, van Ravenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC, *et al*. (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nature Genetics*, **36**, 955–957.

Yu W, Ballif BC, Kashork CD, Heilstedt HA, Howard LA, Cai WW, White LD, Liu W, Beaudet AL, Bejjani BA, *et al*. (2003) Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. *Human Molecular Genetics*, **12**, 2145–2152.

Short Specialist Review

Mosaicism

Wendy P. Robinson
University of British Columbia, Vancouver, BC, Canada

1. Mosaicism-overview

Chromosome mosaicism, the existence of two cell lines with differing chromosomal constitutions derived from a single fertilization, may be observed in amniotic fluid (AF) or chorionic villus (CVS) samples at prenatal diagnosis or in blood or skin samples of individuals referred for a variety of medical conditions. The diagnosis of mosaicism, particularly when ascertained prenatally, presents one of the most problematic genetic counseling situations as it is impossible to fully assess the level and distribution of the abnormal cells or to predict whether outcome of the pregnancy will be normal or not. In cases of mosaicism ascertained postnatally, the abnormal cells may be diagnostic for a specific syndrome, for example, the finding of mosaic tetrasomy 12p in a child with profound mental retardation and other features of Pallister–Killian syndrome (Schinzel, 1991), but can also be inconsequential, for example, a low level of 45,X cells in a woman experiencing recurrent miscarriage (Horsman *et al*., 1987). The interpretation of a mosaic finding thus needs to be carefully considered in terms of the patient phenotype, the type and origin of the abnormality, and the extent of abnormal cells.

To appreciate the effects of mosaicism, it is first important to consider that all human beings are very likely mosaics. The replication of the cellular genome, while reasonably accurate, is not foolproof and mutations arise and chromosomes segregate improperly with some low probability at each mitotic cell division. How low? That depends on the cell type and the type of abnormality. Studies of unused human embryos from in vitro fertilization (IVF) procedures have indicated that at least 70% are chromosomally abnormal, many of which may show mosaicism with normal diploid cells (Gianaroli *et al*., 2001; Magli *et al*., 2001; Ruangvutilert *et al*., 2000; Wells and Delhanty, 2000). This Figure may well be close to 100% if all cells and all chromosomes could be examined at once. There is presumably a high rate of chromosome missegregation, as well as misdivision of whole haploid complements in the first few postzygotic cell divisions (at least under IVF conditions). Very high rates of chromosome aneuploidy have also been found in normal neurons of the developing and mature mouse brain (Kaushal *et al*., 2003) and tetraploid cells are common in placental trophoblast of many mammals (Hoffman and Wooding, 1993), suggesting that the production of chromosomally "abnormal" cells may be part of the normal programmed development of some cell types. Aneuploid cells are also

found at increasing frequency in cultured lymphocytes as individuals age. The X chromosome seems particularly susceptible to nondisjunction and, on average, about 2–3% of cultured lymphocytes in young women (Fitzgerald and McEwan, 1977; Horsman *et al*., 1987; Nowinski *et al*., 1990) and 22% in female centenarians (Bukvic *et al*., 2001) exhibit X chromosome aneuploidy.

Most embryos with a high percentage of abnormal cells probably do not survive implantation or are aborted in early pregnancy, but it is not known how often low-level mosaics are rescued by the plasticity of early development and go on to produce term births. Mosaicism is detected in 1–2% of CVS and 0.1% of AF samples, mostly involving a trisomic (47 chromosomes) cell line. The lower rate of mosaicism in AF samples reflects both the fact that some abnormal pregnancies will be lost between the time of CVS (8–12 weeks gestation) and amniocentesis (15–20 weeks gestation) and that mosaicism is more commonly found in the placenta as compared to the fetus. However, the true frequency of fetal mosaicism is presumably higher, as AF and fetal blood analysis have been shown in several instances to fail to detect a trisomy that is later found in the fetus or newborn (Bruyere *et al*., 1999; Désilets *et al*., 1996; Hammer *et al*., 1991; Opstal *et al*., 1998). Chromosomal mosaicism has been observed in as many as 5–15% of placentae examined from healthy term pregnancies when multiple placental sites were examined (Artan *et al*., 1995). While low levels of trisomic cells in the placenta probably have little impact (and may even be normal), high levels may impair placental function and thus impede fetal growth.

Pregnancy outcome in the case of mosaicism is often associated with how the mosaicism arose. Mosaicism may originate through gain/loss of a chromosome in a normal diploid conceptus (somatic origin) or the abnormal cell line may be present at conception (meiotic origin), but the error "corrects" itself by loss of the supernumerary chromosome during development (Figure 1). Not surprisingly, when the conceptus originates from an abnormal zygote, there tends to be higher levels of trisomy in the placenta, higher risk for fetal mosaicism, and the pregnancy is at higher risk of fetal growth restriction, malformation, and intrauterine or neonatal death (Robinson *et al*., 1997). Uniparental disomy (*see* Article 19, **Uniparental disomy**, Volume 1) (both copies of a chromosome pair originating from the same parent) of the normal cell line can also occur in such cases, which may have a significant phenotypic effect when imprinted genes are located along the involved chromosome. Nonetheless, in many cases, there is a selective advantage of the normal cell line, thus making it possible for a nonmosaic diploid fetus to result from an abnormal conception. For example, trisomy 16 mosaics virtually always derive from a trisomic zygote, but it has been inferred that it is possible to form a baby with entirely (or predominantly) normal cells from only a single diploid cell from the inner cell mass of the blastocyst (Lau *et al*., 1997; Robinson *et al*., 2002). While pregnancies with prenatally diagnosed trisomy 16 mosaicism are at increased risk of complications (poor growth, maternal hypertension, and fetal malformations) due to placental trisomy or low-level trisomy in the fetus, it is truly remarkable how often they proceed successfully. In fact, most cases of mosaicism, even when detected in AF cultures, will have a good prognosis (see e.g., Hsu *et al*., 1997; Wallerstein *et al*., 2000).

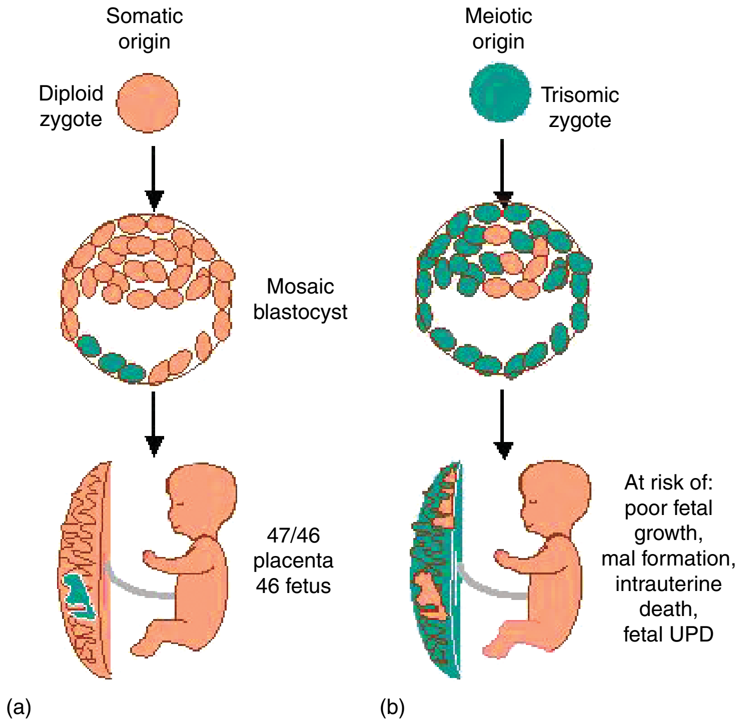


Figure 1 Origin of trisomy mosaicism may be somatic (a) or meiotic (b)

On the flip side, however, is the concern that there may be long-term consequences of mosaicism not apparent at birth. A number of malignancies are associated with chromosomally abnormal cells (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1). Some examples include hepatoblastoma with trisomy 20, various hematological malignancies with trisomy 8 (Maserati *et al*., 2002), certain leukemias with trisomy 21, and gonadoblastoma with 45,X/46,XY mosaicism. While some of these trisomies arise by somatic missegregation in the affected tissue, some may be associated with undiagnosed low-level mosaicism already present at birth. For example, a case of erythroleukemia was diagnosed in a 16-month-old girl with normal development, in which the trisomy 21 cell line found in blood exhibited two different maternal alleles and was thus determined to be of meiotic origin (Minelli *et al*., 2001). Trisomy 8 in cases of myelodysplasia and acute leukemia can be found in other tissues in 15–20% of cases, suggesting an early embryonic origin (Maserati *et al*., 2002).

Malignancies are not the only impact of mosaicism. Skin pigmentation anomalies, such as hypomelanosis of Ito, and asymmetric growth are the clinical features that immediately raise the question of possible chromosomal mosaicism. Rheumatoid arthritis, osteoarthritis, and other inflammatory joint diseases have been associated with trisomy 7 mosaicism in the affected joints (Kinne *et al*., 2001), and trisomy 21 has been found at increased frequency in the peripheral lymphocytes of individuals with Alzheimer disease (Geller and Potter, 1999). The increase of

chromosomally abnormal cells with age may also contribute to the "decay" of the body in ways we do not yet understand.

Although, distinct from mosaicism, chromosomally distinct cells may also be found in an individual because of chimerism or microchimerism. In this case, the cells arise not from a single conceptus but from two separate individuals. True chimeras may result from the fusion of two distinct embryos early in development, but proven examples of this are very few. Some cases of 46,XX/46,XY mosaicism have probably been falsely assumed to be chimeras, when in fact they may result from two chromosome loss events from a 47,XXY conceptus (Niu *et al*., 2002). Microchimerism can occur in twin pregnancies, whereby cells from one twin can populate the haematopoietic stem cells of another twin through "twin to twin transfusion." Furthermore, there appears to be a normal exchange of cells between fetus and mother and vice versa during pregnancy that can persist throughout the life span of the recipient individual (Bianchi, 2000). The pregnancy need not go to term, and as many as 500 000 fetal nucleated cells are transfused following elective first-trimester termination of pregnancy. Some investigators have hypothesized that the presence of foreign white blood cells might help to explain certain autoimmune diseases that tend to be more common in women after the age of 40 (Bianchi, 2000; Nelson, 2003). There is also evidence for the involvement of maternal cells in the etiology of neonatal or juvenile autoimmune disorders (Stevens *et al*., 2003).

Clearly, chromosomal mosaicism and microchimerism play important roles in human disease, which are likely to be appreciated more as clinicians and researchers become more aware of their possible impact. However, tracking these rare cells throughout the body can be a real challenge for researchers. The first step in considering the role of low levels of abnormal cells is to rationally consider both the possibility that they may play a role in disease and also that they may not. Unbiased ascertainment of mosaic cases and long-term follow-up will be key to accurately evaluate these possibilities.

References

Artan S, Basaran N, Hassa H, Özalp S, Sener T, Sayli BS, Cengiz C, Ozdemir M, Durak T and Dolen I (1995) Confined placental mosaicism in term placenta: analysis of 125 cases. *Prenatal Diagnosis*, **15**, 1135–1142.

Bianchi DW (2000) Fetomaternal cell trafficking: a new cause of disease? *American Journal of Medical Genetics*, **91**, 22–28.

Bruyere H, Barrett IJ, Kalousek DK and Robinson WP (1999) Tissue specific involvement in fetal trisomy 16. *American Journal of Human Genetics*, **65**, A173.

Bukvic N, Gentile M, Susca F, Fanelli M, Serio G, Buonadonna L Capurso A and Guanti G (2001) Sex chromosome loss, micronuclei, sister chromatid exchange and aging: a study including 16 centenarians. *Mutation Research*, **498**, 159–167.

Désilets VA, Yong SL, Langlois S, Wilson RD, Kalousek DK and Pantzar TJ (1996) Trisomy 22 mosaicism and maternal uniparental disomy. *American Journal of Human Genetics (Suppl)*, **59**, A319.

Fitzgerald PH and McEwan CM (1977) Total aneuploidy and age-related sex chromosome aneuploidy in cultured lymphocytes of normal men and women. *Human Genetics*, **39**(3), 329–337.

Geller LN and Potter H (1999) Chromosome missegregation and trisomy 21 mosaicism in Alzheimer's disease. *Neurobiology of Disease*, **6**, 167–179.

Gianaroli L, Magli MC and Ferraretti AP (2001) The in vivo and in vitro efficiency and efficacy of PGD for aneuploidy. *Molecular and Cellular Endocrinology*, **183**(Suppl.1), S13–S18.

Hammer P, Holzgreve W, Karabacak Z, Horst J and Miny P (1991) 'False-negative' and 'false-positive prenatal cytogenetic results due to 'true' mosaicism. *Prenatal Diagnosis*, **11**, 133–136.

Hoffman LH and Wooding FB (1993) Giant and binucleate trophoblast cells of mammals. *The Journal of Experimental Zoology*, **266**, 559–577.

Horsman DE, Dill FJ, McGillivray BC and Kalousek DK (1987) X chromosome aneuploidy in lymphocyte cultures from women with recurrent spontaneous abortions. *American Journal of Medical Genetics*, **28**, 981–987.

Hsu LYF, Yu M-T, Neu RL, Van Dyke DL, Benn PA, Bradshaw CL, Shaffer LG, Higgins RR, Khodr GS, Morton CC, *et al*. (1997) Rare trisomy mosaicism diagnosed in amniocytes, involving an autosome other than chromosomes 13, 18, 20 and 21: Karyotype/Phenotype correlations. *Prenatal Diagnosis*, **17**, 201–242.

Kaushal D, Contos JJA, Treuner K, Yang AH, Kingsbury MA, Rehen SK, McConnell MJ, Okabe M, Barlow C and Chun J (2003) Alteration of gene expression by chromosome loss in the postnatal mouse brain. *Journal of Neuroscience*, **23**, 5599–5606.

Kinne RW, Liehr T, Beensen V, Kunisch E, Zimmermann T, Holland H, Pfeiffer R, Stahl HD, Lungershausen W, Hein G, *et al*. (2001) Mosaic chromosomal aberrations in synovial fibroblasts of patients with rheumatoid arthritis, osteoarthritis, and other inflammatory joint diseases. *Arthritis Research*, **3**, 319–330.

Lau AW, Brown CJ, Langlois S, Kalousek DK and Robinson WP (1997) Skewed X-chromosome inactivation is common in fetuses or newborns associated with confined placental mosaicism. *American Journal of Human Genetics*, **61**, 1353–1361.

Magli MC, Gianaroli L and Ferraretti AP (2001) Chromosomal abnormalities in embryos. *Molecular and Cellular Endocrinology*, **183**(Suppl 1), S29–S34.

Maserati E, Aprili F, Vinante F, Locatelli F, Amendola G, Zatterale A, Milone G, Minelli A, Bernardi F, Lo Curto F, *et al*. (2002) Trisomy 8 in myelodysplasia and acute leukemia is constitutional in 15–20% of cases. *Genes Chromosomes & Cancer*, **33**, 93–97.

Minelli A, Morerio C, Maserati E, Olivieri C, Panarello C, Bonvini L, Leszl A, Rosanda C, Lanino E, Danesino C, *et al*. (2001) Meiotic origin of trisomy in neoplasms: evidence in a case of erythroleukaemia. *Leukemia*, **15**, 971–975.

Nelson JL (2003) Microchimerism in human health and disease. *Autoimmunity*, **36**, 5–9.

Niu DM, Pan CC, Lin CY, Hwang B and Chung MY (2002) Mosaic or chimera? Revisiting an old hypothesis about the cause of the 46,XX/46,XY hermaphrodite. *The Journal of Pediatrics*, **140**, 732–735.

Nowinski GP, Van Dyke DL, Tilley BC, Jacobsen G, Babu VR, Worsham MJ, Wilson GN and Weiss L (1990) The frequency of aneuploidy in cultured lymphocytes is correlated with age and gender but not with reproductive history. *American Journal of Human Genetics*, **46**, 1101–1111.

Van Opstal D, Van den Berg C, Deelen WH, Brandenburg H, Cohen-Overbeek TE, Halley DJ, Van den Ouweland AM, In 't Veld PA and Los FJ (1998) Prospective prenatal investigations on potential uniparental disomy in cases of confined placental trisomy. *Prenatal Diagnosis*, **18**, 35–44.

Robinson WP, Barrett IJ, Bernard L, Bernasconi F, Wilson RD, Best R, Howard-Peebles PN, Langlois S and Kalousek DK (1997) A meiotic origin of trisomy in confined placental mosaicism is correlated with presence of fetal uniparental disomy, high levels of trisomy in trophoblast and increased risk of fetal IUGR. *American Journal of Human Genetics*, **60**, 917–927.

Robinson WP, Barrett IJ, Kuchinka B, Penaherrera MS, Bruyere H, Best R, Pediera D, McFadden DE, Langlois S and Kalousek DK (2002) Origin of amnion and implications for evaluation of the fetal genotype in cases of mosaicism. *Prenatal Diagnosis*, **22**, 1078–1087.

Ruangvutilert P, Delhanty JDA, Serhal P, Simopoulou M, Rodeck CH and Harper JC (2000) FISH analysis on day 5 post-insemination of human arrested and blastocyst stage embryos. *Prenatal Diagnosis*, **20**, 552–560.

Schinzel A (1991) Tetrasomy 12p (Pallister-Killian syndrome). *Journal of Medical Genetics*, **28**, 122–125.

Stevens AM, Hermes HM, Rutledge JC, Buyon JP and Nelson JL (2003) Myocardial-tissue-specific phenotype of maternal microchimerism in neonatal lupus congenital heart block. *Lancet*, **362**, 1617–1623.

Wallerstein R, Yu MT, Neu RL, Benn P, Lee Bowen C, Crandall B, Disteche C, Donahue R, Harrison B, Hershey D, *et al*. (2000) Common trisomy mosaicism diagnosed in amniocytes involving chromosomes 13, 18, 20 and 21: karyotype-phenotype correlations. *Prenatal Diagnosis*, **20**, 103–122.

Wells D and Delhanty JD (2000) Comprehensive chromosomal analysis of human preimplantation embryos using whole genome amplification and single cell comparative genomic hybridization. *Molecular Human Reproduction*, **6**, 1055–1061.

Short Specialist Review

Uniparental disomy

Aaron P. Theisen
Health Research and Education Center, Washington State University, Spokane, WA, USA

Lisa G. Shaffer
Sacred Heart Medical Center, Washington State University, Spokane, WA, USA

In the mid-nineteenth century, Gregor Mendel discovered that when round garden peas were crossbred with wrinkled garden peas, all of the offspring were round, regardless of the parental origin of each trait in the cross. The resultant principle of equivalence – that genes are expressed equally, no matter what the parental origin is – and its corollary, the biparental inheritance of autosomal genes, have become a central dogma of genetics. However, in recent decades, researchers have discovered several phenomena that challenge the conventional Mendelian notions of equal, biparental inheritance. Genomic imprinting, the unequal expression of alleles depending on the parent of origin, is perhaps the most clinically significant exception to Mendel's laws of inheritance. The clinical consequences of genomic imprinting may be unmasked when a pair of homologous chromosomes are abnormally inherited from a single parent. This situation is termed *uniparental disomy* (*UPD*).

Eric Engel first suggested UPD as a mechanism for human genetic disease in 1980 on the basis of observations made by Searle and others (Searle *et al*., 1971; Lyon *et al*., 1975) that mice with translocations were susceptible to nondisjunction (*see* Article 16, **Nondisjunction**, Volume 1), in which the homologous chromosomes comprising the translocation would malsegregate during gametogenesis (*see* Article 13, **Meiosis and meiotic errors**, Volume 1). Engel hypothesized that by mating these translocation carriers, a subset of offspring would receive a nullisomic gamete from one parent and a disomic gamete from the other parent, resulting in a chromosomally balanced individual with both chromosome homologs coming from one parent (Engel, 1980). Further research in the mouse by Cattanach and others (1985) provided the first evidence of the clinical effects of mammalian uniparental disomy; however, the existence of UPD in humans did not come until several years later, when the development of DNA-based polymorphic markers allowed the parental origin of chromosome homologs to be determined (Spence *et al*., 1988).

The first documented case of human UPD arose from the investigation of a child with cystic fibrosis and short stature (Spence *et al*., 1988). Marker analysis revealed

that the child inherited both chromosomes 7 from her mother. The cystic fibrosis was likely the result of the presence of a recessive disease allele on both copies of identical homologs of chromosome 7 inherited by the child. Spence *et al*. (1988) proposed three mechanisms in addition to the gametic complementation theory hypothesized by Engel (1980), including postfertilization error and the "rescue" of a conception either by the loss of an extra chromosome in a trisomy or the duplication of a single chromosome in a monosomy. The UPD can contain one copy of each of the contributing parent's homologs (heterodisomy) or two copies of one of the parent's chromosomes (isodisomy). Recessive disease alleles, such as that for cystic fibrosis, are exposed through isodisomies, which often result from a monosomy rescue or postfertilization recombination (Spence *et al*., 1988). UPD for nearly every chromosome has been documented as a result of the investigation of the abnormal inheritance of recessive disorders, including spinal muscular atrophy type III, osteogenesis imperfecta, and Bloom syndrome (reviewed in Cassidy, 1995; Shaffer, 2003).

Because of the high lethality of monosomic embryos and the requirement of a duplication event early in development, cases of UPD arising from a monosomy rescue are far rarer than those UPDs that arise after resolution of a trisomy. Mosaicism (*see* Article 18, **Mosaicism**, Volume 1) may result because of the presence of a mixture of trisomic cells and cells that have lost the extra chromosome copy. The trisomic cells in some conceptuses may be restricted only to the placenta, whereas the fetal cells contain normal chromosomal complements (termed *confined placental mosaicism*). Several UPD cases have been reported following a discrepancy between karyotyping performed on chorionic villus samples (CVS) from the placenta and amniotic fluid samples in which cells are derived from the fetus (Kalousek *et al*., 1993; Jones *et al*., 1995; Ledbetter and Engel, 1995).

Several structural chromosome abnormalities may prompt molecular investigations that may lead to the identification of UPDs, including Robertsonian translocations, nonacrocentric isochromosomes, marker chromosomes, derivative chromosomes, and reciprocal translocations (Shaffer *et al*., 2001a). Of these, Robertsonian translocations are the most likely to be involved in cases of UPD, probably due to their relatively high incidence in the human population and the increased risk of malsegregation during gametogenesis, which would result in aneuploid gametes (Berend *et al*., 2000). Resolution of trisomies may result in UPD. Malsegregation of structural abnormalities may result in gametic complementation and UPD; however, because gametic complementation requires two independent nondisjunction events, one in each parent, UPDs resulting from this mechanism are relatively rare.

Of the four mechanisms that may result in UPD, all but postfertilization error (somatic recombination) result in whole-chromosome uniparental disomies. Mitotic crossing-over may result in partial UPDs; sporadic cases of Beckwith–Wiedemann syndrome (BWS) (*see* Article 30, **Beckwith–Wiedemann syndrome**, Volume 1) are often caused by a partial paternal disomy of the distal short arm of chromosome 11 (Henry *et al*., 1991; Bischoff *et al*., 1995).

Studies of UPD in mouse and human have highlighted the differential maternal and paternal genetic contribution to development. Androgenotes, mouse embryos that contain only paternally derived chromosomes, exhibit poor growth while the extra-embryonic tissues develop relatively normally; in contrast, the

embryonic tissues of gynogenotes, which contain two copies of the maternal chromosomes, develop well, but the extra-embryonic structures fail to develop, resulting in embryonic death shortly after implantation (Surani *et al*., 1984; *see also* Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1). This phenomenon is the result of genomic imprinting and the developmental genetic need for contributions from both parents.

In humans, abnormal pregnancies in which the placenta appears as a cyst known as a hydatidiform mole demonstrate the same abnormal development: the moles, which contain all paternal chromosomes, consist entirely of placental tissue. The maternal counterpart, teratomas, which contain two copies of the maternal genome, contain only embryonic tissues (Jacobs *et al*., 1982). These examples suggest that, rather than the maternal and paternal genomes being equivalent, certain regions of the genome are not expressed equally from the maternal and paternal contributions. Because the differential contributions are not marked by changes to the DNA sequence, epigenetic modifications (chemical changes to the DNA or the surrounding proteins) must result in differential gene expression, depending on whether the genes are inherited from the mother or father.

UPDs may unmask additional phenotypic effects beyond those attributed to recessive disease through the duplication and subsequent overexpression, or deletion and lack of expression, of imprinted genes (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1). For example, 25–30% of patients with Prader–Willi syndrome (PWS) (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1) have maternal UPD for maternal chromosome 15. The other 70% have deletions of a small part of paternal chromosome 15, indicating that the loss of a paternally expressed gene results in the syndrome. Conversely, maternal deletion of 15q in ~70% of cases or paternal disomy for chromosome 15 in ~5% of cases results in a clinically distinct disorder, Angelman syndrome (AS; Nicholls *et al*., 1989). The disparity between frequencies of UPD in the two syndromes is likely a result of higher rates of female nondisjunction than male nondisjunction (Abruzzo and Hassold, 1995). However, most cases of AS caused by UPD are isodisomy, likely due to a rescue of a monosomic conceptus, whereas maternally derived UPD cases in PWS are usually heterodisomic, due to a rescue of a trisomic conceptus. Because monosomies are less viable than trisomies and tend to result in miscarriage sooner, there is a smaller "window of opportunity" in which the monosomic conceptus can be rescued, likely resulting in fewer cases of isodisomy (Shaffer, 2003). In addition to maternal and paternal disomies 15, several other chromosomes show clinically distinct phenotypes due to imprinting, including paternal disomy 6, maternal disomy 7, partial paternal disomy 11, and maternal and paternal disomies 14 (Ledbetter and Engel, 1995; Shaffer *et al*., 2001b; Kotzot, 1999). Phenotypes suggestive of imprinting effects are also found in maternal disomy 2, maternal disomy 6, paternal disomy 9, maternal disomy 16, paternal disomy 16, and maternal disomy 20; however, either because of insufficient cases reported, conflicting reports in the literature, or the confounding effects of mosaicism, the imprinting effects for these chromosomes remain uncertain (Shaffer *et al*., 2001b; Shaffer, 2003).

Imprinting disorders often involve cognitive and growth problems. Patients with PWS, for example, have short stature and are frequently obese (Cassidy, 1984). In contrast, the major clinical features of BWS include overgrowth (Henry *et al.*, 1991). The differential growth effects of some imprinted syndromes – particularly PWS, AS, BWS, and Silver–Russell syndrome (SRS; maternal UPD7 and small stature) – may indicate a conflict between the paternal and maternal genomes over the amount of resources demanded of the mother by her offspring; paternally expressed genes would encourage embryonic growth at the expense of the mother, whereas maternally expressed genes would encourage decreased offspring size (Moore and Haig, 1991). However, contradictory evidence exists for a number of chromosomes, although this may simply indicate that the differential maternal and paternal contributions to development are more complex than what the original models predict (Hurst and McVean, 1997).

Although it is a relatively newly described phenomenon, genomic imprinting has been shown to play a key role in several developmental processes. Disruption of these normal dosage imbalances (Shaffer *et al.*, 2001b) by uniparental disomy has permitted key insights into the possible mechanisms and functions of these epigenetic modifications, the ramifications of which continue to challenge the seemingly unshakable concepts that Mendel observed in peas over 100 years ago.

References

Abruzzo MA and Hassold TJ (1995) Etiology of nondisjunction in humans. *Environmental and Molecular Mutagenesis*, **26**, 38–47.

Berend SA, Horwitz J, McCaskill C and Shaffer LG (2000) Identification of uniparental disomy following prenatal detection of Robertsonian translocations and isochromosomes. *American Journal of Human Genetics*, **66**, 1787–1793.

Bischoff FZ, Feldman GL, McCaskill C, Subramanian S, Hughes MR and Shaffer LG (1995) Single cell analysis demonstrating somatic mosaicism involving 11p in a patient with paternal isodisomy and Beckwith-Wiedemann syndrome. *Human Molecular Genetics*, **4**, 395–399.

Cassidy SB (1984) Prader-Willi syndrome. *Current Problems in Pediatrics*, **14**, 1–55.

Cassidy SB (1995) Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25**, 13–20.

Engel E (1980) A new genetic concept: uniparental disomy and its potential effect, isodisomy. *American Journal of Medical Genetics*, **6**, 137–143.

Henry I, Bonaiti-Pellie C, Chehensse V, Beldjord C, Schwartz C, Utermann G and Junien C (1991) Uniparental paternal disomy in a genetic cancer-predisposing syndrome. *Nature*, **351**, 665–667.

Hurst LD and McVean GT (1997) Growth effects of uniparental disomies and the conflict theory of genomic imprinting. *Trends in Genetics*, **13**, 436–443.

Jacobs PA, Szulman AE, Funkhouser J, Matsuura JS and Wilson CC (1982) Human triploidy: relationship between parental origin of the additional haploid complement and development of partial hydatidiform mole. *Annals of Human Genetics*, **46**, 223–231.

Jones C, Booth C, Rita D, Jazmines L, Spiro R, McCulloch B, McCaskill C and Shaffer LG (1995) Identification of a case of maternal uniparental disomy of chromosome 10 associated with confined placental mosaicism. *Prenatal Diagnosis*, **15**, 843–848.

Kalousek DK, Langlois S, Barrett I, Yam I, Wilson DR, Howard-Peebles PN, Johnson MP and Giorgiutti E (1993) Uniparental disomy for chromosome 16 in humans. *American Journal of Human Genetics*, **52**, 8–16.

Kotzot D (1999) Abnormal phenotypes in uniparental disomy (UPD): fundamental aspects and a critical review with bibliography of UPD other than 15. *American Journal of Medical Genetics*, **82**, 265–274.

Ledbetter DH and Engel E (1995) Uniparental disomy in humans: development of an imprinting map and its implications for prenatal diagnosis. *Human Molecular Genetics*, **4**, 1757–1764.

Lyon MF, Ward HC and Simpson GM (1975) A genetic method for measuring non-disjunction in mice with Robertsonian translocations. *Genetical Research*, **26**, 283–295.

Moore T and Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics*, **7**, 45–49.

Nicholls RD, Knoll JH, Butler MG, Karam S and Lalande M (1989) Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. *Nature*, **342**, 281–285.

Searle AG, Ford CE and Beechey CV (1971) Meiotic disjunction in mouse translocations and the determination of centromere position. *Genetical Research*, **18**, 215–235.

Shaffer LG (2003) Uniparental disomy: mechanisms and clinical consequences. *Fetal and Maternal Medicine Review*, **14**, 155–175.

Shaffer LG, Agan N, Goldberg JD, Ledbetter DH, Longshore JW and Cassidy SB (2001a) American college of medical genetics statement of diagnostic testing for uniparental disomy. *Genetics in Medicine*, **3**, 206–211.

Shaffer LG, Ledbetter DH and Lupski JR (2001b) Molecular cytogenetics of contiguous gene syndromes: mechanisms and consequences of gene dosage imbalance. In *The Metabolic and Molecular Bases of Inherited Disease*, Scriver CR, Beaudet AL, Sly WS and Valle D (Eds.), McGraw-Hill: New York, pp. 1291–1326.

Spence JE, Perciaccante RG, Greig GM, Willard HF, Ledbetter DH, Hejtmancik JF, Pollack MS, O'Brien WE and Beaudet AL (1988) Uniparental disomy as a mechanism for human genetic disease. *American Journal of Human Genetics*, **42**, 217–226.

Surani MA, Barton SC and Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, **308**, 548–550.

Short Specialist Review

Cytogenetics of infertility

Maria Oliver-Bonet and Renée H. Martin
University of Calgary, Calgary, AB, Canada
Alberta Children's Hospital, Calgary, AB, Canada

1. Introduction

Infertility is a major problem in human society. Up to 15% of couples at reproductive age have either problems in conceiving or repetitive reproductive failures. Human fertility can be altered by many different factors, some of which are related to defined genetic syndromes or chromosome abnormalities, both in males and females (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1). Chromosomal disorders occur more frequently among the infertile population. Studies in infertile males have shown that the incidence of constitutional chromosomal abnormalities is higher in these patients than in the fertile population. Depending on the definition of infertility, the average ranges between 2% in males with combined indications of infertility and 14% in azoospermic men, whereas the incidence of chromosome abnormalities in newborns is 0.7% (Shi and Martin, 2001). Aberrant karyotypes have also been detected in 4.8% of women who are partners of infertile men scheduled for intracytoplasmic sperm injections (ICSI) (Gekas *et al*., 2001). The presence of a karyotypic abnormality can reduce fertility in two different ways. First, interruption of gametogenesis can lead to a decrease in the number of germ cells expected. Second, the production, through the meiotic process, of genetically unbalanced gametes leads to either embryonic or fetal death or to the birth of children with mental and physical disabilities.

2. Chromosomal aneuploidies

Somatic chromosome aneuploidies (loss or gain of a chromosome) are not only the most frequent type of anomaly but also those with the main clinical consequences (*see* Article 16, **Nondisjunction**, Volume 1). Klinefelter syndrome (XXY males) is the most common anomaly found among infertile males attending reproductive clinics. Apparent nonmosaic 47,XXY males are usually azoospermic, while mosaic 47,XXY/46,XY individuals are able to produce different amounts of spermatozoa that range from severe oligozoospermia to normozoospermiaxy (*see* Article 18, **Mosaicism**, Volume 1). Fluorescence in situ hybridization (FISH) analysis of the spermatozoa of these patients has demonstrated that the frequencies of sperm

chromosome aneuploidies in nonmosaic and mosaic Klinefelter patients range from 2 to 25% and from 1.5 to 7.0%, respectively, which is significantly higher than in controls (Shi and Martin, 2001). Specifically, the mean frequencies of XY and XX spermatozoa in nonmosaic and in mosaic patients are 8.1 and 3.6%, and 1.8 and 0.8%, respectively (Egozcue *et al*., 2000; Shi and Martin, 2001).

Nowadays, there is a debate in the literature on the meiotic behavior of chromosomes in Klinefelter syndrome patients. Some authors support the idea that small numbers of XXY spermatogonial stem cells are able to undergo meiosis (Yamamoto *et al*., 2002) and that these few cells are responsible for the higher XY and XX disomy frequency found in the sperm of these patients. Alternatively, other authors suggest that XXY cells are unable to enter meiosis and that the XY line is actually the only one able to produce viable spermatozoa (Blanco *et al*., 2001). According to these authors, the unusual frequency of sex disomy found in the sperm of these patients may be due to an abnormal testicular environment, which may itself predispose the sex chromosomes to nondisjunction events (Mroz *et al*., 1999).

Male carriers of an extra Y chromosome (47,XYY) exhibit a variable degree of fertility. Three color-FISH analyses on sperm of 47,XYY men have shown that XY and YY disomy represent only 1% or less of the sperm population (Shi and Martin, 2000). These results, together with meiotic analysis performed on testicular samples of 47,XYY patients, support the hypothesis that the extra Y chromosome is lost from most germ cells even before meiosis begins. However, some authors have reported the presence of an X univalent plus a YY bivalent at pachytene stage, thus suggesting that XYY cells are able to undergo and even survive meiotic processes (Solari and Rey Valzacchi, 1997). The frequency of autosomal aneuploidies has also been investigated in 47,XYY males, in order to assess the possibility of interchromosomal effects (ICE), but no convincing evidence for such effects has been found in these men (Shi and Martin, 2000).

Women with 47,XXX karyotype are able to conceive, without a high risk for aneuploid offspring. Apparently, the extra X chromosome is lost in premeiotic stages. Thus, these women produce only oocytes with a single X. However, a recent study suggests that the triple X syndrome may be associated with premature ovarian failure and also with poor pregnancy outcome (Goswami *et al*., 2003). On the other hand, women affected by Turner's syndrome (45,X) are infertile, because their oocytes are lost during early gestation stages. A recent study reported that approximately 70% of germ cells were apoptotic in the ovary of a 20-week old 45,X fetus (Modi *et al*., 2003). Some oocytes may be capable of survival and maturation; indeed, some follicles have been found in adolescent Turner's patients and spontaneous pregnancies are seen in 2–5% of these women (Hreinsson *et al*., 2002). To date, it is still unclear whether these oocytes are chromosomally balanced.

3. Structural chromosomal anomalies

Somatic structural chromosome anomalies include translocations (both reciprocal and Robertsonian), pericentric and paracentric inversions and insertions (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**,

Volume 1). The percentage of unbalanced gametes found in male and female Robertsonian translocation carriers ranges from 3 to 27% and from 7 to 68% respectively (Guttenbach *et al*., 1997). Reciprocal translocation carriers display an unbalanced average fraction of 53% for males (Shi and Martin, 2001) and 70% for females (Durban *et al*., 2001). In addition, evidence from studies of preimplantation genetic diagnosis suggest that ICE seem to play a role only in the case of Robertsonian translocations (Gianaroli *et al*., 2002; *see also* Article 21, **Preimplantation genetic diagnosis for chromosome abnormalities**, Volume 1). Inversions, either paracentric or pericentric, have a risk of generating unbalanced gametes owing to meiotic recombination processes within the inverted segment. Generally, long inversion segments have a larger risk than short segments. Insertions usually have a high risk of abnormal offspring, the average risk being 32% for the male carrier and 36% for the female (Van Hemel and Eussen, 2000). The incidence of this risk is closely related to the size of the inserted segment – small segments usually have greater risk than large segments.

4. Infertile individuals with a normal somatic karyotype

Reproductive difficulties have also been associated with cytogenetic abnormalities in the germ cells of infertile individuals with a normal constitutional karyotype both in males (Egozcue *et al*., 2000) and females (Plachot, 2001). These abnormalities include meiotic pairing defects and a low incidence of recombination sites. The occurrence of such anomalies can affect the normal course of meiosis through a total or partial meiotic arrest (*see* Article 13, **Meiosis and meiotic errors**, Volume 1). The control upon these aberrant cells is especially noticeable at pachytene, indicating the existence of a universal pachytene checkpoint. Data obtained in different studies strongly suggest that this control mechanism is more permissive in females than in males, which means that whereas spermatogenesis tends to halt when an anomaly is encountered, oogenesis continues, with the consequence of an increased frequency of aneuploidy in the resultant gametes. Indeed, an association between the fertility status of women and the incidence of anomalies in the oocytes was found after analyzing human oocytes that remained unfertilized after in vitro fertilization (IVF) protocols: 13.3% of them were hypohaploid, 8.1% were hyperhaploid, 3.5% were diploid, and 1.6% carried structural abnormalities (Plachot, 2001). On the other hand, the presence of apoptotic processes in males does not preclude a variable number of germ cells from becoming viable aneuploid spermatozoa. Thus, despite having a normal somatic karyotype, sperm collected from infertile men exhibit a significantly increased frequency of chromosomal abnormalities that can also result in both fertility reduction and aneuploid progeny. Different sperm FISH analyses have reported a higher incidence of disomy in infertile patients. Particularly, the frequency of sex chromosome aneuploidy is markedly increased, up to 2–3 times the aneuploidy rate observed in control donors (Shi and Martin, 2001). This frequency is in agreement with the frequency of 1% sex chromosomal abnormalities reported from prenatal diagnoses after ICSI (Liebaers *et al*., 1995).

5. Techniques used in cytogenetic analysis of infertility

The main source of human oocytes for analysis is oocytes that failed to fertilize after IVF. In addition to the egg cells, the first polar body (PB1), which carries a haploid set of chromosomes complementary to the set carried by the oocyte, is also useful in order to obtain information about the incidence of chromosomal anomalies derived from the first meiotic division. After spreading and fixation, chromosome contents can be analyzed by different cytogenetic approaches (Pujol *et al*., 2003).

The content of the spermatozoa can be analyzed using two different methods. The human-hamster system allows analysis of the whole chromosome complement of individual sperm cells (Martin *et al*., 1984). However, the number of chromosome complements that can be analyzed for each patient is, from the statistical point of view, rather low. A quantitative improvement in sperm analysis was the application of FISH techniques on chemically decondensed sperm nuclei (Wyrobek *et al*., 1994). This is a quick and comparatively simple method, which allows the analysis of thousands of spermatozoa in a relatively short time. This method also has its own limitations, though. It is based on an indirect method of detection of certain chromosomes, and thus more prone to error than direct karyotyping of human sperm.

Human meiotic processes have been analyzed using the microspreading technique (Evans *et al*., 1964). This measures the number and position of crossover events within each bivalent at diakinesis-metaphase I stage, and permits the study of all meiotic stages. Nowadays, the application of FISH to meiotic chromosomes enables the analysis of the meiotic behavior of specific chromosomes. However, the special morphology and highly contracted state of the bivalents during the diakinesis-metaphase I stage make the exact estimation of physical distances among the chiasmata of a bivalent troublesome. Immunocytogenetic approaches have also been used to analyze meiotic processes in humans (Barlow and Hultén, 1996). Studies using this method combined with FISH and multi-FISH techniques have provided information about the pairing process of homologous chromosomes during meiosis I. Moreover, the application of both analyses in parallel allows not only the identification of specific pairs of bivalents but also the study of meiotic exchange rates of the target chromosomes (Oliver-Bonet *et al*., 2003).

Cytogenetic research has proved to be one of the most successful tools in the study of fertility. By different approaches, cytogenetic analysis has provided us with significant information about many of the processes that take place during human gametogenesis. We must expect that with the introduction of new and increasingly powerful techniques, especially those resulting from the fusion of conventional cytogenetic techniques and molecular genetic approaches, we will be soon able to obtain a deeper knowledge about both the factors affecting and the basic mechanisms regulating human fertility.

References

Barlow AL and Hultén MA (1996) Combined immunocytogenetic and molecular cytogenetic analysis of meiosis I human spermatocytes. *Chromosome Research*, **4**, 562–573.

Blanco J, Egozcue J and Vidal F (2001) Meiotic behaviour of the sex chromosomes in three patients with sex chromosome anomalies (47,XXY, mosaic 46,XY/47,XXY and 47,XYY) assessed by fluorescence in-situ hybridization. *Human Reproduction*, **16**, 887–892.

Durban M, Benet J, Boada M, Fernandez E, Calafell JM, Lailla JM, Sanchez-Garcia JF, Pujol A, Egozcue J and Navarro J (2001) PGD in female carriers of balanced Robertsonian and reciprocal translocations by first polar body analysis. *Human Reproduction Update*, **7**, 591–602.

Egozcue S, Blanco J, Vendrell JM, Garcia F, Veiga A, Aran B, Barri PN, Vidal F and Egozcue J (2000) Human male infertility: chromosome anomalies, meiotic disorders, abnormal spermatozoa and recurrent abortion. *Human Reproduction Update*, **6**, 93–105.

Evans EP, Breckon G and Ford C (1964) An air-drying method for meiotic preparations from mammalian testes. *Cytogenetics*, **3**, 289–294.

Gekas J, Thepot F, Turleau C, Siffroi JP, Dadoune JP, Briault S, Rio M, Bourouillou G, Carre-Pigeon F, Wasels R, *et al*. (2001) Association des cytogeneticiens de langue francaise (chromosomal factors of infertility in candidate couples for ICSI: an equal risk of constitutional aberrations in women and men. *Human Reproduction*, **16**, 82–90.

Gianaroli L, Magli MC, Ferraretti AP, Munne S, Balicchia B, Escudero T and Crippa A (2002) Possible interchromosomal effect in embryos generated by gametes from translocation carriers. *Human Reproduction*, **17**, 3201–3207.

Goswami R, Goswami D, Kabra M, Gupta N, Dubey S and Dadhwal V (2003) Prevalence of the triple X syndrome in phenotypically normal women with premature ovarian failure and its association with autoimmune thyroid disorders. *Fertility and Sterility*, **80**, 1052–1054.

Guttenbach M, Engel W and Schmid M (1997) Analysis of structural and numerical chromosome abnormalities in sperm of normal men and carriers of constitutional chromosome aberrations. A review. *Human Genetics*, **100**, 1–21.

Hreinsson JG, Otala M, Fridstrom M, Borgstrom B, Rasmussen C, Lundqvist M, Tuuri T, Simberg N, Mikkola M, Dunkel L, *et al*. (2002) Follicles are found in the ovaries of adolescent girls with Turner's syndrome. *Journal of Clinical Endocrinology and Metabolism*, **87**, 3618–3623.

Liebaers I, Bonduelle M, Van Assche E, Devroey P and Van Steirteghem A (1995) Sex chromosome abnormalities after intracytoplasmic sperm injection. *Lancet*, **346**, 1095.

Martin R, Balkan W, Burns K, Rademaker A, Lin C and Rudd N (1984) The constitution of 1000 human spermatozoa. *Human Genetics*, **63**, 304–309.

Modi DN, Sane S and Bhartiya D (2003) Accelerated germ cell apoptosis in sex chromosome aneuploid fetal human gonads. *Molecular Human Reproduction*, **9**, 219–225.

Mroz K, Hassold TJ and Hunt PA (1999) Meiotic aneuploidy in the XXY mouse: evidence that a compromised testicular environment increases the incidence of meiotic errors. *Human Reproduction*, **14**, 1151–1156.

Oliver-Bonet M, Liehr T, Nietzel A, Heller A, Starke H, Claussen U, Codina-Pascual M, Pujol A, Abad C, Egozcue J, *et al*. (2003) Karyotyping of human synaptonemal complexes by cenM-FISH. *European Journal of Human Genetics*, **11**, 879–883.

Plachot M (2001) Chromosomal abnormalities in oocytes. *Molecular Cell and Endocrinology*, **183**(Suppl 1), S59–S63.

Pujol A, Boiso I, Benet J, Veiga A, Durban M, Campillo M, Egozcue J and Navarro J (2003) Analysis of nine chromosome probes in first polar bodies and metaphase II oocytes for the detection of aneuploidies. *European Journal of Human Genetics*, **11**, 325–336.

Shi Q and Martin RH (2000) Multicolor fluorescence in situ hybridization analysis of meiotic chromosome segregation in a 47,XYY male and a review of the literature. *American Journal of Medical Genetics*, **93**, 40–46.

Shi Q and Martin RH (2001) Aneuploidy in human spermatozoa: FISH analysis in men with constitutional chromosomal abnormalities, and in infertile men. *Reproduction*, **121**, 655–666.

Solari AJ and Rey Valzacchi G (1997) The prevalence of a YY synaptonemal complex over XY synapsis in an XYY man with exclusive XYY spermatocytes. *Chromosome Research*, **5**, 467–474.

Van Hemel JO and Eussen HJ (2000) Interchromosomal insertions. Identification of five cases and a review. *Human Genetics*, **107**, 415–432.

Wyrobek AJ, Robbins W, Mehraein Y and Weier H (1994) Detection of sex chromosomal aneuploidies X-X, Y-Y in human sperm using two-chromosome fluorescence in situ hybridization. *American Journal of Human Genetics*, **53**, 1–7.

Yamamoto Y, Sofikitis N, Mio Y, Loutradis D, Kaponis A and Miyagawa I (2002) Morphometric and cytogenetic characteristics of testicular germ cells and sertoli cell secretory function in men with non-mosaic Klinefelter's syndrome. *Human Reproduction*, **17**, 886–896.

Short Specialist Review

Preimplantation genetic diagnosis for chromosome abnormalities

Santiago Munné
Yale University, New Haven, CT, USA

1. Introduction

Preimplantation genetic diagnosis (PGD) is the earliest form of prenatal diagnosis, and usually involves the analysis of a single cell after biopsy from a three-day old embryo created through assisted reproductive techniques (ART), or sometimes the first and second polar bodies of the eggs prior fertilization, also through ART. Because the embryo should be replaced in the future mother in no more than one or two days after biopsy and as there are only one or two cells for analysis, the diagnostic tests must be fast and highly sensitive. However, obtaining analyzable metaphases of karyotyping quality from a single cell is not possible even with cell-conversion methods (Willadsen *et al*., 1999; Verlinsky and Evsikov, 1999); so the analysis for cytogenetic purposes is performed using FISH, which allows chromosome enumeration of interphase cell nuclei, that is, without the need for culturing cells or preparing metaphase spreads. Since 1993, FISH has been used for PGD of common human aneuploidies with either blastomeres (cells from 2- to 16-cell stage embryos) or oocyte polar bodies (Munné *et al*., 1993; Munné *et al*., 1995a,b; Munné *et al*., 1998b; Munné *et al*., 1999; Munné *et al*., 2003; Verlinsky *et al*., 1995; Verlinsky *et al*., 1996; Verlinsky *et al*., 1998; Verlinsky *et al*., 2001; Verlinsky and Kuliev, 1996; Gianaroli *et al*., 1997; Gianaroli *et al*., 1999; Gianaroli *et al*., 2001b; Pehlivan *et al*., 2002; Kahraman *et al*., 2000; Rubio *et al*., 2003). Currently, probes for at least chromosomes X, Y, 13, 15, 16, 18, 21, and 22 are being used simultaneously (Munné *et al*., 2003), with the potential of detecting 83% of aneuploidies found in spontaneous abortions (Jobanputra *et al*., 2002).

2. PGD to improve pregnancy outcome in ART

PGD was first thought of as a tool for selecting against genetically abnormal embryos from couples carrying genetic diseases; but about 90% of the PGD cycles performed so far have been for aneuploidy to improve the pregnancy outcome of ART patients, with over 5000 cases performed to date for that purpose (Munné *et al*., 1999; Munné *et al*., 2003; Gianaroli *et al*., 1999; Gianaroli *et al*., 2001b;

ESHRE PGD Consortium Steering Committee, 2002; Verlinsky and Kuliev, 2003) and close to a thousand babies born thereafter (Verlinsky *et al*., 2004).

The rationale for using PGD to increase pregnancy rates and reduce miscarriage rates is as follows. Oocyte quality is the major cause of reduced implantation with advancing maternal age (Navot *et al*., 1994), and one of the clearest links so far between maternal age and embryo competence is aneuploidy. The increase in aneuploidy with maternal age in spontaneous abortuses and live offspring (Hassold *et al*., 1980; Warburton *et al*., 1980; Warburton *et al*., 1986; Simpson, 1990) has also been observed in embryos and oocytes (Munné *et al*., 1995a; Márquez *et al*., 2000; Dailey *et al*., 1996) but with much higher rates of chromosome abnormalities than in spontaneous abortions, which indicates that a sizable part of chromosomally abnormal embryos are eliminated before clinical recognition. This embryo loss, rather than endometrial factors, largely accounts for the decline in implantation with maternal age.

To compensate for the low implantation potential of human embryos created in vitro, fertility centers normally generate a larger cohort of embryos (average >10); those with the highest potential to implant are then selected on the basis of morphology and developmental characteristics. Unfortunately, trisomy is not correlated with embryo morphology or development (Munné *et al*., 1995a; Márquez *et al*., 2000), and only some monosomies can be selected against by culturing the embryos to blastocyst stage (Sandalinas *et al*., 2001).

Because of the correlation between aneuploidy and declining implantation with maternal age, we hypothesized that negative selection of chromosomally abnormal embryos could reverse this trend (Munné *et al*., 1993). While the probes currently used in PGD check only a limited number of chromosomes, the results so far indicate that PGD of aneuploidy actually does increase implantation while reducing trisomic offspring and spontaneous abortions (Munné *et al*., 1999; Munné *et al*., 2003; Gianaroli *et al*., 1999; Gianaroli *et al*., 2001a,b; Werlin *et al*., 2003).

As determined in a recent study, the chromosomes most involved in aneuploidy at the cleavage stage (first 3 days of embryo development) are different than those found in prenatal diagnosis (Munné *et al*., 2004). So when the most common chromosomes are analyzed by PGD with eight or more probes, including probes for chromosomes 13, 15, 16, 18, 21, and 22, the implantation rate (embryos implanting/embryos replaced) doubles.

In one study, we observed a significant twofold increase of implantation, from 10.2 to 22.5% ($p < 0.001$) (Gianaroli *et al*., 1999); in another more recent study with an older population (average age 40), and with two or less previously failed IVF attempts, we found a 20% implantation rate after PGD compared to 10% in the control group ($p = 0.002$) (Munné *et al*., 2003).

It is clear that not only does implantation reduce with advancing maternal age but also those embryos that do implant have a higher risk of chromosomal abnormality and miscarriage. Since the objective of ART is to ensure a healthy baby for couples seeking to conceive, both factors are extremely important.

FISH with probes for 13, 15, 16, 18, 21, 22, X, and Y can detect 83% of all chromosomally abnormal fetuses detected by karyotyping (Jobanputra *et al*., 2002). Since this combination of probes is the current standard (Munné *et al*., 1999; Gianaroli *et al*., 1999), PGD is able to eliminate close to 80% of all chromosomally

abnormal embryos at risk of causing a miscarriage. Among many examples, two studies reported abortion rates of only 9% after PGD in women >36 years (Munné *et al*., 1999; Gianaroli *et al*., 2001a,b) compared to the 24% spontaneous abortion expected for such populations of infertile patients (SART-ASRM, 2000).

Increased implantation and decreased spontaneous abortion result in a higher chance of patients achieving viable pregnancies (Munné *et al*., 1999). However, PGD works properly only when a larger group embryos is available for testing in a given procedure. If patients have less than five embryos available, then replacing all the embryos will in general give the same result as if PGD had been performed.

3. PGD to reduce the risk of aneuploid conceptions

The current PGD technique works by analysis of a single cell with an error rate around 10%. This error rate is mostly due to mosaicism, which is very common in human cleavage-stage embryos (see review Munné *et al*., 2002). Thus, when diagnosing trisomic offspring, PGD can significantly reduce the occurrence but not completely prevent it. Indeed, four misdiagnoses have already occurred after PGD (Munné *et al*., 1998b; Gianaroli *et al*., 2001a). Nevertheless, the rate of trisomic offspring detected after PGD is significantly lower than expected ($p < 0.001$). For instance, 2 of 666 (0.3%) fetuses were found with aneuploidies for chromosomes XY, 13, 15, 16, 18, 21, and 22 (Munné *et al*., 2003 and unpublished results) compared to a 2.6% rate expected in a population of the same age range (Eiben *et al*., 1994). Interestingly, the reduction from 2.6% to 0.3% is a 90% reduction, which is as expected if the error rate is indeed 10%. Similarly, Verlinsky *et al*. (2001) reported 140 healthy children born after PGD of aneuploidy using polar body analysis, with no misdiagnoses.

4. PGD for translocations

Balanced translocations occur in 0.2% of the neonatal population. However, they are identified in 0.6% of infertile couples, 2–3.2% of infertile males requiring intracytoplasmic sperm injection, and 9.2% of fertile couples experiencing three or more consecutive first-trimester abortions (Testart *et al*., 1996; Meschede *et al*., 1998; Van der Ven *et al*., 1998; Stern *et al*., 1996; Stern *et al*., 1999). PGD for translocations can reduce spontaneous abortion and minimize the risk of conceiving an unbalanced baby, thus being a realistic alternative to prenatal diagnosis and pregnancy termination of unbalanced fetuses. So far there have been close to 500 cycles of PGD of translocations performed worldwide (Munné *et al*., 2002; Verlinsky *et al*., 2002; Cieslak *et al*., 2003; Gianaroli *et al*., 2003).

4.1. Methods

Several approaches to PGD of translocations have been developed. The first involved the analysis of first polar bodies, after the observation that more than 90%

of first polar bodies fixed for 6 or fewer hours after retrieval are in metaphase stages (Munné *et al*., 1998a). The translocation can then be identified using chromosome-painting probes for the two chromosomes involved in the translocation (Munné *et al*., 1998c,d; Durban *et al*., 2001).

Cell-conversion methods have also been used to transform blastomere nuclei (usually in interphase to metaphase) by fusing them to oocytes or zygotes (Willadsen *et al*., 1999; Verlinsky and Evsikov, 1999; Evsikov *et al*., 2000; Verlinsky *et al*., 2002), achieving close to 80% rates of analyzable metaphases. Alternatively, Tanaka *et al*. (2004) observed 4–6-cell stage embryos every hour, and when the nuclear envelope of a blastomere disappeared, the blastomere was biopsied within an hour. Disappearance of nuclear envelope was observed in 89% of embryos, and all produced analyzable metaphases.

Two different interphase approaches have also been developed for PGD of translocations. The first developed specific probes spanning the breakpoints of each translocation (Munné *et al*., 1998c; Weier *et al*., 1999) or inversion (Cassel *et al*., 1997). The second used probes distal to the breakpoints or telomeric probes in combination with proximal or centromeric probes, either for translocations (Munné *et al*., 1998c; Munné *et al*., 2000; Pierce *et al*., 1998; Van Assche *et al*., 1999) or inversions (Iwarsson *et al*., 1998). The exception is a Robertsonian translocation (RT), for which chromosome enumerator probes are used to detect aneuploid embryos (Conn *et al*., 1998; Munné *et al*., 1998c; Munné *et al*., 2000). Only the first approach (spanning probes) can differentiate between balanced and normal embryos.

4.2. *Results*

For most translocation patients, the risk of consecutive pregnancy loss is their major incentive in enrolling in a PGD program. The unbalanced products of a translocation are usually lethal and therefore the true risk is that of pregnancy loss. We have demonstrated that PGD of translocations substantially increases a couple's chances of sustaining a pregnancy to full term (Munné *et al*., 1998e; Munné *et al*., 2000). So far, 115 patients undergoing PGD for translocations have lost 84% (233/278) of their prior conceptions, but after PGD only 5% (4/78) was lost (Munné *et al*., 1998e; Munné *et al*., 2000, and unpublished data). Data from Verlinsky's group also indicates a significant reduction in spontaneous abortions to 20% (7/34) (Verlinsky *et al*., 2002; Cieslak *et al*., 2003), when 88% of pregnancies in these patients prior to undertaking PGD procedure resulted in spontaneous abortions.

However, some translocation patients produce 80% or more unbalanced gametes, and with the 50% baseline of chromosome abnormalities in cleavage-stage human embryos, it is nearly impossible to find normal ones for replacement. Previous studies have used clinically recognized pregnancies to formulate rules to predict unbalanced offspring (Jalbert *et al*., 1988). However, these specimens were probably the most viable segregation types because selective processes had already occurred. Thus, when analyzing zygotes and preimplantation embryos, it is not surprising that different translocations involving the same chromosomes show very different meiotic behavior (Escudero *et al*., 2000; Van Assche *et al*.,

1999). Escudero *et al*. (2003) determined the level of chromosome abnormalities in spermatozoa that would preclude a chromosomally normal conception, and found that the percentages of abnormal gametes and of abnormal embryos were correlated, thereby establishing that patients with 65% or fewer chromosomally abnormal spermatozoa have a good chance of conceiving.

5. Molecular methods for PGD of chromosome abnormalities

Ultimately, speedy and efficient analysis of all 24 chromosomes is PGD's true goal, as some embryos diagnosed as normal are undoubtedly still abnormal for other aneuploidies not analyzed in current protocols. One approach still needing improvement is comparative genome hybridization (CGH) (Kallioniemi *et al*., 1992). For CGH of single cells, the whole genome of the cell must be amplified (Wells *et al*., 1999). Trials applied to human blastomeres from discarded embryos have promising results (Wells and Delhanty, 2000; Voullaire *et al*., 1999; Voullaire *et al*., 2000), but so far, the process takes too long. To gain enough time for analysis, Wilton *et al*., (2001, 2003) applied this method to blastomeres from embryos that were frozen after biopsy, and the first babies have recently been borne following this procedure. However, cryopreservation and thaw destroys some embryos and ultimately outweighs the benefits of CGH.

Recently, we have been able to obviate cryopreservation by applying CGH to polar bodies and get results prior to embryo replacement on day four of development (Wells *et al*., 2002). But as with any polar body analysis, postzygotic abnormalities, which account for more than half of all abnormalities, as well as paternally derived aneuploidies, are not detectable.

DNA microarrays are being developed for aneuploidy and translocation analysis (Weier *et al*., 2001), but as presented in the last International symposium of PGD, current methods still need improvement to differentiate in single cells ratio changes of 0.5 in order to be used for PGD (Leight *et al*., 2003).

Once optimized, microarrays will have the advantage over CGH of being more robust and probably faster, and possibly not requiring embryo freezing. Also, they will supply redundancy and more accurate diagnosis of subchromosomal regions useful for translocation analysis.

References

Cassel MJ, Munné S, Fung J and Weier HUG (1997) Carrier-specific breakpoint-spanning DNA probes: an approach to preimplantation genetic diagnosis in interphase cells. *Human Reproduction*, **12**, 2019–2027.

Cieslak J, Zlatopolsky Z, Galat V, *et al*. (2003) *Preimplantation diagnosis in 146 cases of chromosomal translocations*. *Fifth International Symposium on Preimplantation Genetics*, Antalya, 5-7 June, p. 23, (abstract).

Conn CM, Harper JC, Winston RML and Delhanty JDA (1998) Infertility couples with Robertsonian translocations: preimplantation genetic analysis of embryos reveals chaotic cleavage divisions. *Human Genetics*, **102**, 117–123.

Dailey T, Dale B, Cohen J and Munné S (1996) Association between non-disjunction and maternal age in meiosis-II human oocytes detected by FISH analysis. *American Journal of Human Genetics*, **59**, 176–184.

Durban M, Benet J, Boada M, Fernández E, Calafell JM, Lailla JM, Sánchez-García JF, Pujol A, Egozcue J and Navarro J (2001) PGD in female carriers of balanced Robertsonian translocations and reciprocal translocations by first polar body analysis. *Human Reproduction Update*, **7**, 591–602.

Eiben B, Goebel R, Hansen S and Hammans W (1994) Early amniocentesis. A cytogenetic evaluation of over 1500 cases. *Prenatal Diagnosis*, **14**, 497–501.

Escudero T, Abdelhadi I, Sandalinas M and Munné S (2003) Predictive value of sperm chromosome analysis on the outcome of PGD for translocations. *Fertility and Sterility*, **79**(Suppl 3), 1528–1534.

Escudero T, Lee M, Carrel D, Blanco J and Munné S (2000) Analysis of chromosome abnormalities in sperm and embryos from two 45,XY,t(13;14)(q10;q10) carriers. *Prenatal Diagnosis*, **20**, 599–602.

ESHRE PGD Consortium Steering Committee (2002) ESHRE preimplantation genetic diagnosis consortium: data collection III (May 2001). *Human Reproduction*, **17**, 233–246.

Evsikov S, Cieslak MLT and Verlinsky Y (2000) Effect of chromosomal translocations on the development of preimplantation human embryos in vitro. *Fertility and Sterility*, **74**, 672–677.

Gianaroli L, Magli MC and Ferraretti AP (2001a) The in vivo and in vitro efficiency and efficacy of PGD for aneuploidy. *Molecular and Cellular Endocrinology*, **183**, S13–S18.

Gianaroli L, Magli MC, Ferraretti AP, Tabanelli C, Trombetta C and Boudjema E (2001b) The role of preimplantation diagnosis for aneuploidy. *Reproductive Biomedicine Online*, **4**, 31–36.

Gianaroli L, Magli MC, Ferraretti AP, Fiorentino A, Garrisi J and Munné S (1997) Preimplantation genetic diagnosis increases the implantation rate in human in vitro fertilization by avoiding the transfer of chromosomally abnormal embryos. *Fertility and Sterility*, **68**, 1128–1131.

Gianaroli L, Magli C, Ferraretti AP and Munné S (1999) Preimplantation diagnosis for aneuploidies in patients undergoing in vitro fertilization with a poor prognosis: identification of the categories for which it should be proposed. *Fertility and Sterility*, **72**, 837–844.

Gianaroli L, Magli C, Fiorentino F, Baldi M and Ferraretti AP (2003) Clinical value of preimplantiation genetic diagnosis. *Placenta*, **24**, S77–S83.

Hassold T, Jacobs PA, Kline J, Stein Z and Warburton D (1980) Effect of maternal age on autosomal trisomies. *Annals of Human Genetics*, **44**, 29–36.

Iwarsson E, Ahrlund-Richter L, Inzunza J, Rosenlund B, Fridstrom M, Hillensjo T, Sjoblom P, Nordenskjold M and Blennow E (1998) Preimplantation genetic diagnosis of a large pericentric inversion of chromosome 5. *Molecular Human Reproduction*, **4**, 719–723.

Jalbert P, Jalbert H and Sele B (1988) Types of imbalances in human reciprocal translocations: risks at birth. In *The Cytogenetics of Mammalian Autosomal Rearrangements*, Daniel A (Ed.), Alan R Liss: New York.

Jobanputra V, Sobrino A, Kinney A, Kline J and Warburton D (2002) Multiplex interphase FISH as a screen for common aneuploidies in spontaneous abortions. *Human Reproduction*, **17**, 1166–1170.

Kahraman S, Bahce M, Samli H, Imirzahoglu N, Yakisn K, Cengiz G and Donmez E (2000) Healthy births and ongoing pregnancies obtained by preimplantation genetic diagnosis in patients with advanced maternal age and recurrent implantation failure. *Human Reproduction*, **15**, 2003–2007.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative Genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Leight D, Wright D and Stojanov T (2003) *Development of customized microarrays for testing of chromosomal aberrations in human preimplantation embryos*. *Fifth International Symposium on Preimplantation Genetics*, Antalya, 5–7 June, p. 21, (abstract).

Márquez C, Sandalinas M, Bahçe M, Alikani M and Munné S (2000) Chromosome abnormalities in 1255 cleavage-stage human embryos. *Reproductive Biomedicine Online*, **1**, 17–27.

Meschede D, Lemcke B, Exeler J, De Geyter C, Behre HM, Nieschlag E and Horst J (1998) Chromosome abnormalities in 447 couples undergoing intracytoplasmatic sperm injection – prevalence, types, sex distribution and reproductive relevance. *Human Reproduction*, **13**, 576–582.

Munné S, Alikani M, Tomkin G, Grifo J and Cohen J (1995a) Embryo morphology, developmental rates and maternal age are correlated with chromosome abnormalities. *Fertility and Sterility*, **64**, 382–391.

Munné S, Dailey T, Sultan KM, Grifo J and Cohen J (1995b) The use of first polar bodies for preimplantation diagnosis of aneuploidy. *Human Reproduction*, **10**, 1015–1021.

Munné S, Bahçe M, Sandalinas M, Escudero T, Márquez C, Velilla E, Colls P, Oter M, Alikani M and Cohen J (2004) Differences in chromosome susceptibility to aneuploidy and survival to first trimester. *Reproductive Biomedicine Online*, **8**, 81–90.

Munné S, Scott R, Sable D and Cohen J (1998a) First pregnancies after preconception diagnosis of translocations of maternal origin. *Fertility and Sterility*, **69**, 675–681.

Munné S, Magli C, Bahçe M, Fung J, Legator M, Morrison L, Cohen J and Gianaroli L (1998b) Preimplantation diagnosis of the aneuploidies most commonly found in spontaneous abortions and live births: XY, 13, 14, 15, 16, 18, 21, 22. *Prenatal Diagnosis*, **18**, 1459–1466.

Munné S, Fung J, Cassel MJ, Márquez C and Weier HUG (1998c) Preimplantation genetic analysis of translocations: case-specific probes for interphase cell analysis. *Human Genetics*, **102**, 663–674.

Munné S, Bahçe M, Schimmel T, Sadowy S and Cohen J (1998d) Case report: chromatid exchange and predivision of chromatids as other sources of abnormal oocytes detected by preimplantation genetic diagnosis of translocations. *Prenatal Diagnosis*, **18**, 1450–1458.

Munné S, Morrison L, Fung J, Márquez C, Weier U, Bahçe M, Sable D, Grundfelt L, Schoolcraft B, Scott R, *et al*. (1998e) Spontaneous abortions are reduced after pre-conception diagnosis of translocations. *Journal of Assisted Reproduction and Genetics*, **15**, 290–296.

Munné S, Lee A, Rosenwaks Z, Grifo J and Cohen J (1993) Diagnosis of major chromosome aneuploidies in human preimplantation embryos. *Human Reproduction*, **8**, 2185–2191.

Munné S, Magli C, Cohen J, Morton P, Sadowy S, Gianaroli L, Tucker M, Márquez C, Sable D, Ferraretti AP, *et al*. (1999) Positive outcome after preimplantation diagnosis of aneuploidy in human embryos. *Human Reproduction*, **14**, 2191–2199.

Munné S, Sandalinas M, Escudero T, Fung J, Gianaroli L and Cohen J (2000) Outcome of preimplantation genetic diagnosis of translocations. *Fertility and Sterility*, **73**, 1209–1218.

Munné S, Sandalinas M, Escudero T, Marquez C and Cohen J (2002) Chromosome mosaicism in cleavage stage human embryos: evidence of a maternal age effect. *Reproductive Biomedicine Online*, **4**, 223–232.

Munné S, Sandalinas M, Escudero T, Velilla E, Walmsley R, Sadowy S, Cohen J and Sable D (2003) Improved implantation after preimplantation genetic diagnosis of aneuploidy. *Reproductive Biomedicine Online*, **7**, 91–97.

Navot D, Drews MR, Bergh PA, Guzman I, Karstaedt A, Scott RT, Garrisi GJ and Hofmann GE (1994) Age related decline in female fertility is not due to diminished capacity of the uterus to sustain embryo implantation. *Fertility and Sterility*, **61**, 97–101.

Pehlivan T, Rubio C, Rodrigo L, Romero J, Remohi J, Simon C and Pellicer A (2002) Impact of preimplantation genetic diagnosis on IVF outcome in implantation failure patients. *Reproductive Biomedicine Online*, **6**, 232–237.

Pierce KE, Fitzgerald LM, Seibel MM and Zilberstein M (1998) Preimplantation genetic diagnosis of chromosome imbalance in embryos from patient with a balanced reciprocal translocation. *Molecular Human Reproduction*, **4**, 167–172.

Rubio C, Simon C, Vidal F, Rodrigo L, Pehlivan T, Remohi J and Pellicer A (2003) Chromosomal abnormalities and embryo development in recurrent miscarriage couples. *Human Reproduction*, **18**, 182–188.

Sandalinas M, Sadowy S, Alikani M, Calderon G, Cohen J and Munné S (2001) Developmental ability of chromosomally abnormal human embryos to develop to the blastocyst stage. *Human Reproduction*, **16**, 1954–1958.

Simpson JL (1990) Incidence and timing of pregnancy losses: relevance to evaluating safety of early prenatal diagnosis. *American Journal of Medical Genetics*, **35**, 165–173.

Society for Assisted Reproduction and Technology and American Society for Reproductive Medicine (2000) Assisted reproductive technology in the United States: 1997 results generated from the American Society for Reproductive Medicine/ Society for assisted Reproduction and technology. *Fertility and Sterility*, **74**, 641–654.

Stern JJ, Dorfman AD and Gutierrez-Najar MD (1996) Frequency of abnormal karyotype among abortuses from women with and without a history of recurrent spontaneous abortions. *Fertility and Sterility*, **65**, 250–253.

Stern C, Pertile M, Norris H, Hale L and Baker HWG (1999) Chromosome translocations in couples with in-vitro fertilization implantation failure. *Human Reproduction*, **14**, 2097–2101.

Tanaka A, Nagayoshi M, Awata S, Mawatari Y, Tanaka I and Kusunoki H (2004) Preimplantation diagnosis of repeated miscarriage due to chromosomal translocations using metaphase chromosomes of a blastomeres biopsied from 4- to 6-cell-stage embryos. *Fertility and Sterility*, **81**, 30–34.

Testart J, Gautier E, Brami C, Rolet F, Sedmon E and Thebault A (1996) Intracytoplasmic sperm injection in infertile patients with structural chromosome abnormalities. *Human Reproduction*, **11**, 2609–2612.

Van Assche E, Staessen C, Vegetti W, Bonduelle M, Vandervorst M, Van Steirteghem A and Liebaers I (1999) Preimplantation genetic diagnosis and sperm analysis by fluorescence in-situ hybridization for the most common reciprocal translocation t(11;22). *Molecular Human Reproduction*, **5**, 682–690.

Van der Ven K, Peschka B, Montag M, Lange R, Schwanitz G and van der Ven HH (1998) Increased frequency of congenital chromosomal aberrations in female partners of couples undergoing intracytoplasmic sperm injection. *Human Reproduction*, **13**, 48–54.

Verlinsky Y, Cieslak V, Evsikov G, Galat V and Kuliev A (2002) Nuclear transfer for full karyotyping and preimplantation diagnosis for translocations. *Reproductive Biomedicine Online*, **5**, 300–305.

Verlinsky Y, Cieslak J, Frieidine M, Ivakhnenko V, Wolf G, Kovalinskaya L, White M, Lifchez A, Kaplan B, Moise J, *et al*. (1995) Pregnancies following pre-conception diagnosis of common aneuploidies by fluorescence in-situ hybridization. *Human Reproduction*, **10**, 1923–1927.

Verlinsky Y, Cieslak J, Ivakhnenko V, Lifchez A, Strom C, Kuliev A and Preimplantation genetic group (1996) Birth of healthy children after preimplantation diagnosis of common aneuploidies by polar body fluorescent in situ hybridization analysis. *Fertility and Sterility*, **66**, 126–129.

Verlinsky Y, Cieslkak J, Ivanhnenko V, Evsikov S, Wolf G, White M, Lifchez A, Kaplan B, Moise J, Valle J, *et al*. (1998) Preimplantation diagnosis of common aneuploidies by the first- and second-polar body FISH analysis. *Journal of Assisted Reproduction and Genetics*, **15**, 285–289.

Verlinsky Y, Cieslak V, Ivakhnenko S, Evsikov G, Wolf M, White M, Lifchez A, Kaplan B, Moise J, Valle J, *et al*. (2001) Chromosomal abnormalities in the first and second polar body. *Molecular and Cellular Endocrinology*, **183**, S47–S49.

Verlinsky Y, Cohen J, Munné S, Gianaroli L, Simpson JL, Ferraretti AP and Kuliev A (2004) Over a decade of preimplantation genetic diagnosis experience – a multicenter report. *Fertility and Sterility*, **82**, 292–294.

Verlinsky Y and Evsikov S (1999) Karyotyping of human oocytes by chromosomal analysis of the second polar body. *Molecular Human Reproduction*, **5**, 89–95.

Verlinsky Y and Kuliev A (1996) Preimplantation diagnosis of common aneuploidies in fertile couples of advanced maternal age. *Human Reproduction*, **11**, 2076–2077.

Verlinsky Y and Kuliev A (2003) Thirteen years' experience of preimplantation diagnosis: report of the fifth international symposium on preimplantation genetics. *Reproductive Biomedicine Online*, **8**, 229–235.

Voullaire L, Slater H, Williamson R and Wilton L (2000) Chromosome analysis of blastomeres from human embryos by using comparative genomic hybridization. *Human Genetics*, **106**, 210–217.

Voullaire L, Wilton L, Slater H and Williamson R (1999) Detection of aneuploidy in single cells using comparative genome hybridization. *Prenatal Diagnosis*, **19**, 846–851.

Warburton D, Kline J, Stein Z and Strobino B (1986) Cytogenetic abnormalities in spontaneous abortions of recognized conceptions. In *Perinatal Genetics: Diagnosis and Treatment*, Porter IH and Willey A (Eds.), Academic Press: New York, pp. 133–148.

Warburton D, Stein Z, Kline J and Susser M (1980) Chromosome abnormalities in spontaneous abortion: data from the New York city study. In *Human Embryonic and Fetal Death*, Porter LH and Hook EB (Eds.), Academic press: New York, pp. 261–287.

Weier HUG, Munné S and Fung J (1999) Patient-specific probes for Preimplantation Genetic Diagnosis (PGD) of structural and numerical aberrations in interphase cells. *Journal of Assisted Reproduction and Genetics*, **16**, 182–191.

Weier HUG, Munné S, Lersch RA, Hsieh HB, Smida J, Chen XN, Korenberg JR, Pedersen RA and Fung J (2001) Towards a full karyotype screening of interphase cells: 'FISH and chip' technology. *Molecular and Cellular Endocrinology*, **183**, S41–S45.

Wells D and Delhanty JDA (2000) Comprehensive chromosomal analysis of human preimplantation embryos using whole genome amplification and single cell comparative genomic hybridization. *Molecular Human Reproduction*, **6**, 1055–1062.

Wells D, Escudero T, Levy B, Hirschhorn K, Delhanty JDA and Munné S (2002) First clinical application of comparative genome hybridization (CGH) and polar body testing for Preimplantation Genetic Diagnosis (PGD) of aneuploidy. *Fertility and Sterility*, **78**, 543–549.

Wells D, Sherlock JK, Handyside AH and Delhanty DA (1999) Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genome hybridization. *Nucleic Acids Research*, **27**, 1214–1218.

Werlin L, Rodi I, DeCherney A, Marello E, Hill D and Munné S (2003) Preimplantation Genetic Diagnosis (PGD) as both a therapeutic and diagnostic tool in assisted reproductive technology. *Fertility and Sterility*, **80**, 467–468.

Willadsen S, Levron J, Munné S, Schimmel T, Márquez C, Scott R and Cohen J (1999) Rapid visualization of metaphase chromosomes in single human blastomeres after fusion with in vitro matured bovine eggs. *Human Reproduction*, **2**, 470–475.

Wilton L, Voullaire L, Sargeant P, Williamson R and McBain J (2003) Preimplantation aneuploidy screening using comparative genomic hybridization or fluorescence in situ hybridization of embryos from patients with recurrent implantation failure. *Fertility and Sterility*, **80**, 860–868.

Wilton L, Williamson R, McBain J, Edgar D and Voullaire L (2001) Birth of a healthy infant after preimplantation confirmation of euploidy by comparative genomic hybridization. *The New England Journal of Medicine*, **345**, 1537–1541.

Basic Techniques and Approaches

FISH

P. Nagesh Rao
David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Mark J. Pettenati
Wake Forest University School of Medicine, Winston-Salem, NC, USA

Standard or conventional chromosomal banding analysis is limited to actively dividing cells and the resolution is limited to chromosomal aberrations greater than 3 Mb in size. On the other hand, fluorescence in situ hybridization (FISH) has developed into a meaningful and clinically accepted higher resolution method for analyzing the genetic characteristics of cells. Different FISH technologies provide increased resolution for the elucidation of structural chromosome abnormalities that cannot be resolved by more conventional cytogenetic analyses, including microdeletion syndromes, cryptic or subtle duplications and translocation, complex rearrangements involving many chromosomes, and marker chromosomes. There is a broad array of FISH techniques for both diagnostic and research applications. The FISH procedure has been developed for the tagging of DNA and RNA with labeled nucleic acid probes, and is a process whereby chromosomes or portions of chromosomes are vividly painted with fluorescent molecules that anneal to specific regions. This technique has been used widely for gene mapping and for the identification of chromosomal abnormalities. The method enables enumeration of multiple copies of chromosomes or detection of specific regions of DNA or RNA that represent associations with certain genetic characteristics and infectious diseases. Utilization of the FISH methods requires two essential components. The first is a nucleic acid probe that is complementary to the chromosome region of interest and the second is a buffer that solubilizes the probe and potentiates the denaturation of the cellular proteins and strands of DNA and RNA to be hybridized. The advent of fluorescent dyes for use as labels has made it possible to visualize multiple probes at the same time. This has opened up new possibilities for investigating and diagnosing diseases. There are now whole batteries of commercially available probes that can be used to identify each of the human chromosomes and can be used to identify cells in which chromosomal aberrations occur. Probe hybridization conditions generally depend on the probe's base-pair sequence and length of bases required to span the region of interest. Generally, probes less than 40 bases long are synthetically manufactured and probes that range from 80 Kb to 1 Mb long are manufactured with molecular biology methods using cosmids, YACs, PACs, or BACs. Probes can also be manufactured with PCR

techniques. The FISH probes that are developed in the labs have to be optimized for the FISH procedure. It is also important to compromise between maximizing the sensitivity of the test and minimizing the cost of the study. As hybridization efficiency increases proportionally with the increasing target size, the larger YAC and BAC probes usually give high-efficiency signals, and thus fewer cells need to be scored. In contrast, small single-copy signals hybridize less efficiently; thus, more cells need to be evaluated. The key criteria for development of probes for the FISH technique include bright signal intensity, compactness, and retention of cellular and nuclear morphology. These features provide a signal that is visible at lower magnification, thereby minimizing interference from cross hybridization to other genomic targets. The protocols developed for use with commercially available probes for labeling chromosomes in interphase nuclei and metaphase preparations, are also compatible with synthetically, biologically, or enzymatically lab-produced probes. The buffers used should also be compatible with either direct or indirect labeled fluorescent probes, optimize nucleic acid probe performance for rapid hybridizations to provide bright signals and low background, same-day hybridizations with results ranging from 30 min to overnight depending on probe characteristics.

For all interphase FISH analyses, it is important to establish diagnostic cutoff levels using the probe of choice and a series of normal controls for the type of cell under investigation. These must be established in the laboratory and cannot be adopted from other studies. The addition of different fluorescent labels and the selection of fluorescent dyes have also added variability to the technology method. Further, after labeling with FISH, it is often necessary to scan, detect, and enumerate the fluorescently labeled spots. In order to computerize the images for analysis, it is essential to have low background, specific and uniform labeling, and bright label intensities. After FISH, slides should be stored in the dark, either refrigerated or frozen, and are stable for at least 3 months.

Three different types of probes are commonly used, each with different ranges of applications.

Gene-specific probes target DNA sequences present in only one copy per chromosome. They are used to identify chromosomal translocations, inversions and deletions, contiguous gene syndromes, and chromosomal amplifications in interphase and metaphase chromosomes (Figure 1). Repetitive sequence probes bind to chromosomal regions that are represented by short repetitive base-pair sequences that are present in multiple copies (e.g., centromeric and telomeric probes). Centromeres are usually A-T rich, whereas telomeres are known to have repetitive TTAGGG sequences. Centromeric probes are extremely useful for identifying marker chromosomes and for detecting copy number chromosome abnormalities in interphase nuclei (Figure 2). On the other hand, telomeric probes are frequently used to identify subtle or submicroscopic chromosomal rearrangements. The relative ease of performance and high resolution (0.5 Mb) of repetitive sequence FISH has made it popular to screen for the common chromosomal aneuploidies, subtelomeric deletions, and marker chromosomes.

Whole-genome painting probes (WCP) are complex DNA probes that are generated by degenerate oligonucleotide polymerase chain reaction (DOP-PCR) or through flow sorting. WCP paints have high affinity for the whole chromosome

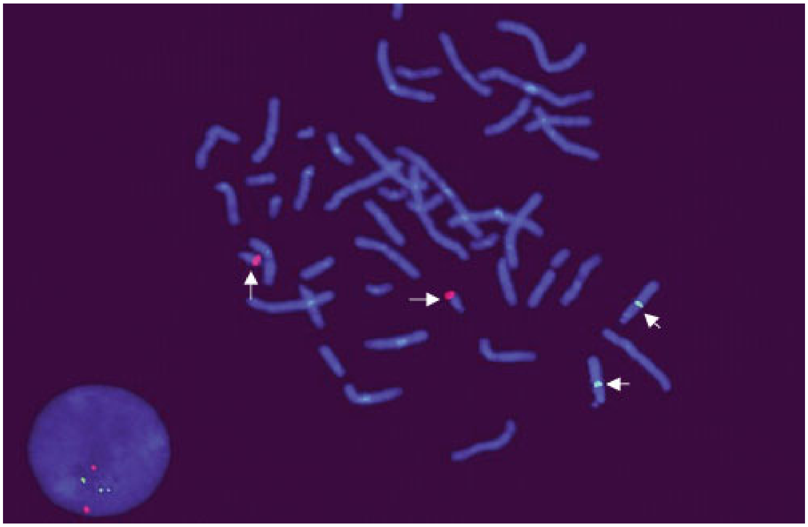


Figure 1 Locus-specific single-copy DNA probes mapping to the long arm of chromosome 21q22 (Red signals) and chromosome 13q14 (Green signals)

along its entire length, with the exception of the centromeric and telomeric regions (Figure 3). These probes are most suitable for identifying genomic imbalances in metaphase chromosomes, especially the complex chromosomal arrangements observed in many cancers. Two variants of the WCP have been developed, multicolor-FISH (M-FISH) and spectral karyotyping (SKY). WCP probes have been developed that paint all human chromosomes in different colors (48 paints). And WCP can also offer simultaneous detection of each arm of all human chromosomes in a single hybridization. This technique is usually used in conjunction with chromosomal banding techniques for a more precise identification of chromosome aberrations. The two greatest limitations of WCP are that there must be extensive knowledge of genetic abnormality to enable the correct selection of the probes – unless the full complement of paints is to be used – and there is limited resolution to detect chromosomal inversions and very small deletions/amplifications and translocations due to its limited resolution of greater than 2–3 Mb.

Comparative genomic hybridization (CGH) is a molecular cytogenetic technique that emerged from FISH and standard cytogenetics to look for genome-wide DNA copy number changes in patients with a possible chromosomal disorder. This technique is especially valuable because, unlike standard cytogenetics, cell culture is not necessary. CGH can be performed on DNA extracted from archival tissue as well as from fresh-frozen specimens. Chromosomal copy number changes can be detected without having to selectively perform FISH for a specific chromosomal sequence. Briefly, the CGH involves two-color FISH of tumor and reference DNA to normal metaphase chromosomes. In order to prevent nonspecific hybridization, the differentially labeled tumor and control DNA are mixed together with Cot-1 DNA (containing repetitive sequences of the genome). Images of metaphase spreads are obtained with a CCD camera and fluorochrome-specific optical filters to

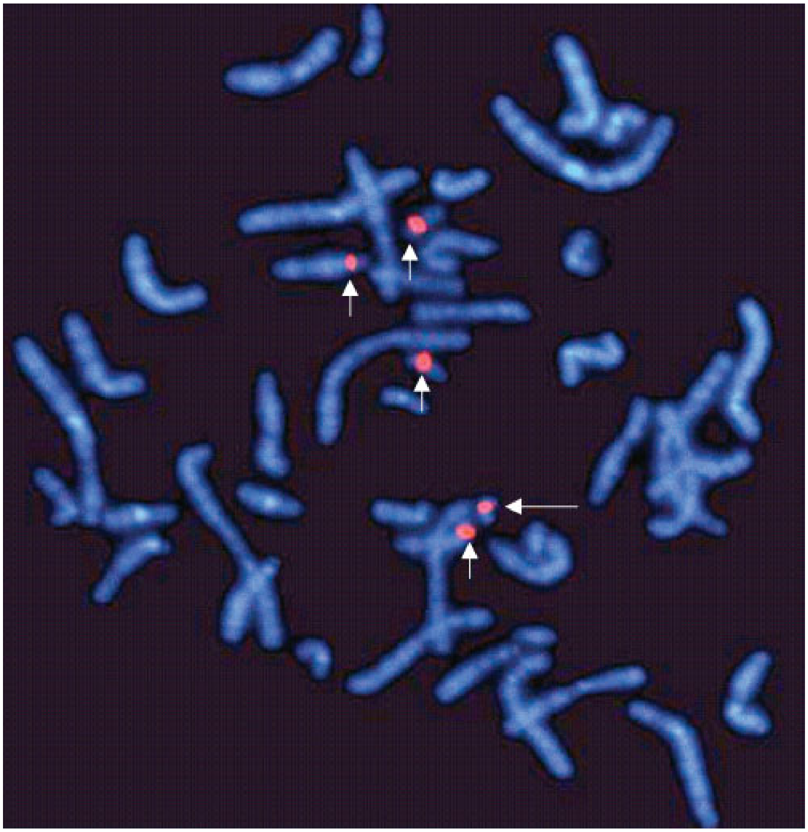


Figure 2 Repetitive sequence probes binding specifically to the centromeres of chromosome 14 and 22 (short arrows). The fifth signal (long arrow) shows a marker chromosome derived from either chromosome 14 or 22

capture the FITC (fluorescein-5-isothiocyanate) and TRITC (tetramethylrhodamine isothiocyanate) fluorescence. The copy number changes in the genome relative to the normal are assessed on the basis of differences in fluorescence intensities along the chromosome. The limitations of this technique include resolution of 3–10 Mb, manual hybridization reactions, and high cost. Recently, array-based formats of CGH (array CGH) have emerged into clinical practice as an alternative. In this procedure, the metaphase chromosomes are replaced with specific DNA sequences spotted in arrays on a glass slide. These DNA sequences may be either large insert genomic clones (e.g., BACs or oligonucleotides-synthesized short DNA sequences. The array readout is digital and quantitative. The greatest advantage of these techniques is the resolution.

The human BAC arrays for CGH consist of linker-adapter PCR representations of BAC clones spotted onto a variety of slide substrates. Hybridization to the human BAC arrays allows detection of single-copy gains and losses. The applications of BAC array technology include identification of diagnostic and prognostic chromosomal aberrations in birth defects, mental retardation, and cancers. The

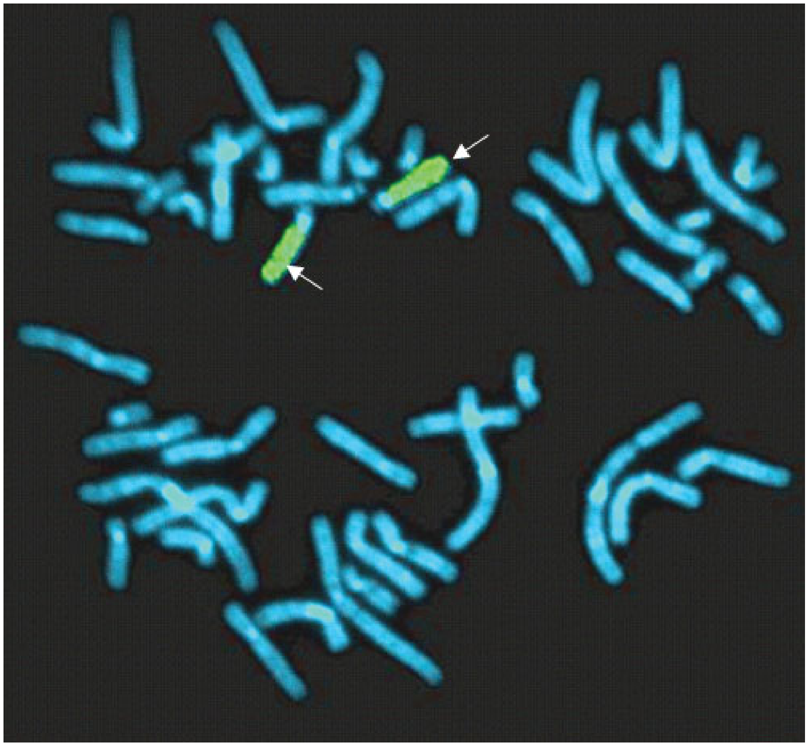


Figure 3 A metaphase cell showing two chromosome 15s with whole-chromosome paint probes (arrow, green signal) of chromosome 15

whole-genome analysis usually requires anywhere from several hundred nanograms to a couple of micrograms of genomic DNA, depending on the source of BAC arrays. One commercially available kit for human BAC arrays currently contains 2632 BAC clones distributed approximately uniformly at a spacing of approximately 1 Mb across the genome. Each clone contains at least one sequence tag that is mapped to the human genome sequence. The kit is composed of BACs that are spotted in duplicate allowing greater precision of the results. Clones containing unique sequences near the telomeres and clones containing genes known to be significant in cancer are included. Higher-density whole-genome- and chromosome-specific arrays have been developed in research laboratories with the goal of improving resolution. This slide format uses relatively large amounts of genomic DNA (approximately 6–10 ∝g at 400–500 ng uL^{-1}) from test specimens for analysis. The signal is highly quantitative and reproducible with low "background noise". However, the preparation and spotting of thousands of BACs onto slides is cumbersome and labor-intensive. The resolution is limited to the map distance between each clone. The signals produced by low copy, centromeric and telomeric repeats may be nonlinear, and thus nonindicative of the copy number. Array-based CGH technology offers 45 kb to 1 Mb resolution, depending on the density of the array.

Oligonucleotide-based arrays have gained popularity in the scientific community to monitor the quantitative expression of the arrayed genes in mRNA populations from cell/tissues, and more recently has been applied to detect DNA sequence and copy number polymorphisms in genomic DNA – array CGH. In brief, the arrays contain oligonucleotides matching all wild-type and single nucleotide substitution sequences in a gene, thus providing the basis for detecting single nucleotide polymorphisms (SNPs) as well as genomic copy number polymorphisms. Oligonucleotide array CGH can be applied to testing for mutations, as well as short- and long-range deletions and duplications in the genome, including aneuploidies and gene amplifications. When DNA derived from somatic (i.e., cancerous) cells and normal tissues is compared, loss of heterozygosity events can be detected. The resolution is limited by the density of the oligonucleotide arrays – currently 45–85 kb. Higher-density SNP arrays are currently undergoing β-testing in several commercial outfits, and these will greatly increase the ability to detect small regions of chromosomal changes and to identify the boundaries of the regions of loss or gain, thereby delineating candidate genes. Additional improvements will come from high-throughput automation of the process, including data analysis, and from development of protocols for handling fragmented DNA.

It is important to emphasize that the FISH methodologies reviewed here complement, but do not replace, standard chromosome banding analysis. Every technique has its strengths and limitations.

Further reading

Albertson DG (2003) Profiling breast cancer by array CGH. *Breast Cancer Research and Treatment*, **78**, 289–298.

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, **14**, 287–295.

Cai WW, Mao JH, Chow CW, Damani S, Balmain A and Bradley A (2002) Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nature Biotechnology*, **20**, 393–396.

Gray JW, Kallioniemi A, Kallioniemi O, Pallavicini M, Waldman F and Pinkel D (1992) Molecular cytogenetics: diagnosis and prognostic assessment. *Current Opinion in Biotechnology*, **3**, 623–631.

Harrison CJ, Kempski H, Hammond DW and Kearney L (2004) Molecular cytogenetics in childhood leukemia. *Methods in Molecular Medicine*, **91**, 123–137.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

Mantripragada KK, Buckley PG, de Stahl TD and Dumanski JP (2004) Genomic microarrays in the spotlight. *Trends in Genetics*, **20**, 87–94.

Mundle SD and Sokolova I (2004) Clinical implications of advanced molecular cytogenetics in cancer. *Expert Review of Molecular Diagnostics*, **4**, 71–81.

Saracoglu K, Brown J, Kearney L, Uhrig S, Azofeifa J, Fauth C, Speicher MR and Eils R (2001) New concepts to improve resolution and sensitivity of molecular cytogenetic diagnostics by multicolor fluorescence in situ hybridization. *Cytometry*, **44**, 7–15.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263–264.

Tönnies H (2002) Modern molecular cytogenetic techniques in genetic diagnostics. *Trends in Molecular Medicine*, **8**, 246–250.

Veltman IM, Veltman JA, Arkesteijn G, Janssen IM, Vissers LE, de Jong PJ, van Kessel AG and Schoenmakers EF (2003) Chromosomal breakpoint mapping by arrayCGH using flow-sorted chromosomes. *Biotechniques*, **35**, 1066–1070.

Comparative genomic hybridization

Brynn Levy
Mount Sinai School of Medicine, New York, NY, USA

1. Introduction

The history of human cytogenetics spans approximately 130 years from when chromosomes were first observed in plant material by Eduard Strasburger in 1875 and in animals by Walter Flemming in 1879–1889 (Lawce and Brown, 1997). However, the major advances in cytogenetics occurred in the 1950s when colchicine and hypotonic treatments were introduced (Hsu, 1952; Hughes, 1952; Makino and Nishimura, 1952) and 1970s with the development of various banding techniques that utilized either fluorescent dyes or Giemsa stain (Drets and Shaw, 1971; Patil *et al*., 1971; Seabright, 1971; Sumner *et al*., 1971) and enabled structural changes to be discerned and correlated with clinical phenotypes (Hirschhorn *et al*., 1973). The combination of molecular and cytogenetic techniques in the early 1970s led to the field of molecular cytogenetics, which now plays a significant role in both clinical diagnostics as well as clinical research. The in situ hybridization (ISH) procedures utilized in the 1970s and early 1980s primarily relied on radioactive probes and were quickly and preferably substituted for by fluorescence in situ hybridization (FISH) techniques that utilized nonradioactive probes labeled with haptens or fluorochromes (Manning *et al*., 1975; Pinkel *et al*., 1986). Clinical and cancer cytogeneticists readily embraced the new FISH techniques as it provided them with a way to detect complicated, cryptic, and submicroscopic rearrangements that remained undetected or undecipherable by conventional cytogenetic analysis (Kuwano *et al*., 1991; Leana-Cox *et al*., 1993; Franke *et al*., 1994; Blennow *et al*., 1994). FISH was also an attractive technique as it could be performed using both interphase nuclei and metaphase spreads (Kuo *et al*., 1991; Klever *et al*., 1992). The clinical utility of FISH in cancer cytogenetics quickly became apparent in the hematological cancers, but FISH on metaphases spreads from solid tumors was fraught with the same technical difficulties as conventional cytogenetic analysis because very few metaphase spreads are obtained after cell culture and their quality is often suboptimal. Comparative genomic hybridization (CGH), a DNA-based molecular cytogenetic technique, was developed in 1992 primarily as a way to overcome the obstacles associate with analyzing solid tissue tumors (Kallioniemi *et al*., 1992). CGH provided a very novel way to detect chromosomal

imbalances in a single-step global wide scan of the genome. Many factors made CGH an appealing technique for cancer research, the primary one being that CGH is a DNA-based technique, which therefore obviates the requirement of specimen culturing and does not have to consider the availability and quality of the specimen metaphase spreads. This property has also allowed for the analysis of formalin-fixed paraffin-embedded neoplastic tissue and even nonviable tissues, such as that derived from the products of conception. The ability to obtain a genome-wide search for imbalances without any prior information of the chromosomal aberration in question has been particularly useful in clinical cytogenetics, especially for samples for which a complete and detailed karyotype could not be obtained by conventional methods.

2. The basics of CGH

Comparative genomic hybridization (CGH) is a powerful DNA-based cytogenetic technique that allows the entire genome to be scanned for chromosomal imbalances without requiring the sample material to be mitotically active. CGH effectively reveals any DNA sequence copy number changes (i.e., gains, amplifications or losses) in a particular specimen and maps these changes on normal chromosomes (Kallioniemi *et al*., 1992; Kallioniemi *et al*., 1994). CGH can detect changes that are present in as little as 30–50% or more of the specimen cells (Kallioniemi *et al*., 1994). It does not reveal balanced translocations, inversions, and other aberrations that do not change copy number. CGH is accomplished by in situ hybridization of differentially labeled total genomic specimen DNA and normal reference DNA to normal human metaphase chromosome spreads (Kallioniemi *et al*., 1992; Kallioniemi *et al*., 1994). Hybridization of the specimen and reference DNA can be distinguished by their different fluorescent colors. The relative amounts of specimen and reference DNA hybridized at a particular chromosome position are contingent on the relative excess of those sequences in the two DNA samples and can be quantified by calculation of the ratio of their different fluorescent colors (Kallioniemi *et al*., 1992; Kallioniemi *et al*., 1994). Specimen DNA is traditionally labeled with a green fluorochrome such as fluorescein isothiocyanate (FITC) and the normal reference DNA with a red fluorochrome like Texas Red®. A gain of chromosomal material in a specimen would be detected by an elevated green:red ratio, while deletions or chromosomal losses would produce a reduced green:red ratio (Kallioniemi *et al*., 1992; Kallioniemi *et al*., 1994) (Figure 1). Computer-assisted measurement of the red:green ratio along each chromosome allows all over/underrepresented regions to be identified (Figure 1). The incorporation of standard reference intervals into the software program confers a considerably higher sensitivity to CGH analysis compared to fixed diagnostic thresholds and is often referred to as "high resolution CGH" (Kirchhoff *et al*., 1998). The sensitivity of conventional CGH is in the megabase range, with a theoretical detection limit of deletions estimated to be about 2 Mb (Piper *et al*., 1995). The introduction of high-resolution CGH has demonstrated that CGH can detect imbalances around

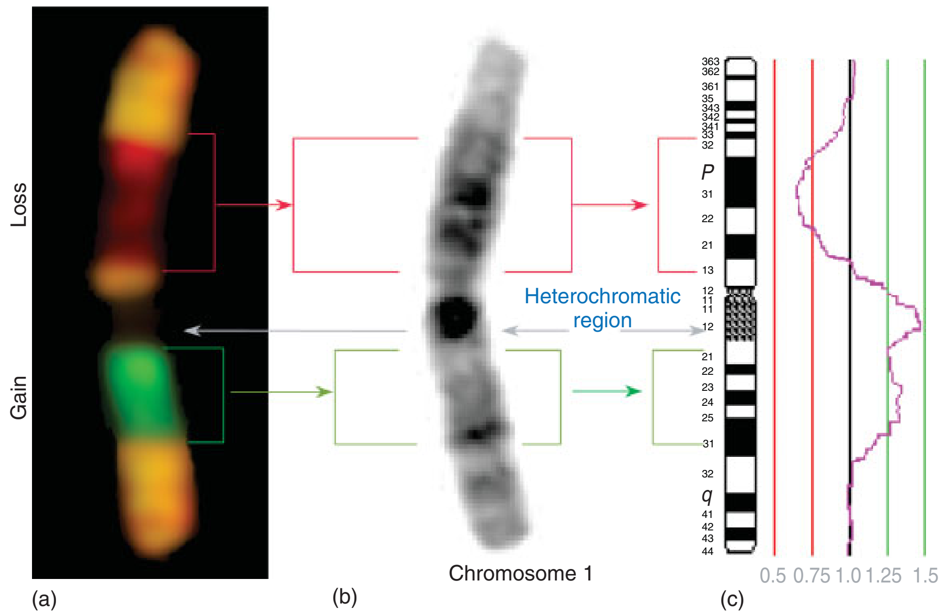


Figure 1 Mapping of chromosomal gains and losses detected by CGH. A loss in the short arm, dim (p21p31) and gain in the long arm, enh (q21q32.1) are illustrated. (a) Visual inspection of gains and losses of chromosomal material. Green regions represent a gain of specimen DNA at that region while red areas reflect greater hybridization of reference DNA at that location due to the depletion of specimen DNA. (b) Inverted DAPI image allows for identification of the chromosome. (c) Computer-generated CGH fluorescence ratio profile. The center line in the CGH profile represents the balanced state of the chromosomal copy number (ratio 1.0). Gains are viewed to the right (> 1.20) and losses (< 0.80) to the left of the center line

2–3 Mb in size (Kirchhoff *et al*., 2001), which appears comparable to that of a high-resolution karyotype (750–1000 band level).

3. Overview of CGH

The three main components of the CGH assay are the specimen DNA, the reference DNA, and the normal metaphase spreads. The specimen can be any patient sample from which DNA can be extracted. In cancer cytogenetics, this may include solid tumors and formalin-fixed preparations and in clinical cytogenetics this may include peripheral blood or prenatal samples such as chorionic villi or amniocytes. Reference DNA is extracted from karyotypically normal individuals, and the metaphase spreads are preferentially prepared from karyotypically normal 46,XY male individuals so that both the X and Y chromosomes are available as hybridization targets. Nick translation is used to label the specimen and reference (normal) DNA with different fluorochromes (usually green for the test specimen and red for the normal reference). In situ hybridization using equal amounts of labeled specimen and reference DNAs together with unlabeled Cot1-DNA is then performed using the normal 46,XY metaphase spreads. Following

this, any nonhybridized/unbound DNA is washed off and the metaphase spreads are counterstained with DAPI, which allows for chromosome identification. Fluorescence microscopy utilizing fluorochrome specific filters is required to visualize and digitally capture color ratio differences along the chromosomes. Chromosome identification is aided by the inverted DAPI banding pattern, which is similar to G-bands. Computer software quantitates the DNA copy number differences by generating a ratio profile of the specimen to reference DNA fluorescence intensities along each chromosome and is therefore critical for the actual analysis and interpretation. Combining the profiles from several metaphase spreads is necessary to improve the significance of the results.

4. CGH in cancer cytogenetics

Classical and molecular cytogenetic analysis of neoplastic specimens has produced a wealth of information regarding the hematological malignancies (Mitelman, 1994; Najfeld, 1997). In contrast, there is significantly less information known about the cytogenetics and molecular cytogenetics of solid tumors. This is mainly due to the technical difficulties encountered when preparing metaphase spreads from these tumor cells. Karyotype analysis requires viable, proliferating cells that can be arrested in the metaphase stage of the cell cycle. Since many solid tumor cells fail to proliferate in vitro, conventional cytogenetic analysis of these tumors is often not possible. For those tumors where metaphase spreads are produced, the quality is often inadequate and does not allow for recognition of banding patterns. There is also the possibility that the cytogenetic data derived from in vitro tumor cell culture is not accurate as small subclones in vivo may take advantage of the in vitro conditions, and thus, the nonproliferating cells that constitute the main clone in vivo may escape detection by conventional cytogenetic analysis (Becher *et al*., 1997). In addition, many aberrant chromosomal regions may not be identified because of the highly complex karyotypes of cultured cancer cells carrying both multiple numerical and structural chromosomal abnormalities.

Since CGH is a DNA-based analysis, initiation of a cell culture from a tumor specimen is not necessary as DNA copy number imbalances can be assessed directly from the analysis of genomic DNA extracted from these tumor specimens. The detection of chromosomal imbalances by CGH is dependent on the aberrations being present in 30–50% or more of cells from which the DNA is extracted (Kallioniemi *et al*., 1994). Therefore, the results derived using this technique reflect changes that are genuinely present in the majority of tumor cells. CGH analysis is, however, limited by its inability to provide information about balanced rearrangements, the origins of structural anomalies, or the ploidy status of the cells. Nevertheless, identification of specific regions of imbalance has been sufficient to provide a location for candidate genes (oncogenes and tumor suppressor genes) causative of the initiation and progression of these tumors.

The use of CGH in cancer genetics has revealed a number of novel recurring chromosomal gains, amplifications, losses, and deletion sites that escaped detection using traditional cytogenetic analysis in various tumors, including testicular germ

cell tumors, gliomas, sarcomas, breast cancer, prostate cancer, uterine leiomyomata, small-cell lung carcinoma, head, neck and pancreatic carcinomas, and uveal melanomas. The chromosomal aberrations detected by CGH has also provided prognostic information in a number of neoplasms including esophageal squamous cell carcinoma, renal cell carcinomas, breast carcinomas, cervical carcinomas, endometrioid carcinoma, malignant fibrous histiocytoma, neuroblastoma, bladder cancer, uveal melanoma, and prostate cancer. Various international CGH databases have been established that provide a wealth of information on the CGH studies that have been done since 1992 including the Progenetix cytogenetic on-line database (**http://www.progenetix.net**), the Tokyo Medical and Dental University CGH database (**http://www.cghtmd.jp/cghdatabase/index_e.html**), the database of Humboldt-University of Berlin (**http://amba.charite.de/~ksch/cghdatabase/index.htm**) and the National Cancer Institute and National Center for Biotechnology Information Spectral Karyotyping SKY, and Comparative Genomic Hybridization CGH Database (2001), (**http://www.ncbi.nlm.nih.gov/sky**). The use of CGH in cancer cytogenetics has now led to the discovery of new tumor suppressor genes and oncogenes that play a role in the initiation and/or progression of solid tumors and has therefore laid the groundwork for the development of appropriate and tissue-specific therapeutic agents.

5. CGH in clinical cytogenetics

Clinical cytogenetics in the late 1990s realized that the powerful genome-wide scanning attribute of CGH could be used to identify karyotypes with unbalanced or unrecognizable G-banded cytogenetic material (Bryndorf *et al*., 1995; Levy *et al*., 1998). This was an exciting application of the relatively new molecular cytogenetic technique as it could more precisely characterize the regions of imbalance in patients with chromosomal abnormalities, and thus pave the way for discovery of the gene(s) that are responsible for the clinical features observed in such patients. In the past, identification of unbalanced cytogenetic material of unknown origin required a comprehensive molecular cytogenetic work-up using various FISH probes. This approach is both expensive and labor intensive as numerous probes and/or whole-chromosome paints may be required until the origin of the chromosomal material is identified. In addition, the number of commercially available region-specific probes is limited and covers only a fraction of the genome. CGH also has the advantage over conventional FISH with whole chromosome paints (wcps) and multicolor FISH by its ability to identify not only the chromosome from which the extra unknown material originated but also to map the region involved to specific G-bands on the source chromosome. In addition, the CGH analysis does not require any prior information of the chromosomal imbalance in question as it scans the entire genome in a single step. The DNA-based nature of CGH also allows for the analysis of specimens that have failed to grow in cell culture (Fritz *et al*., 2001). This has been especially effective in the analysis of spontaneous abortions, which are estimated to have a chromosome abnormality (mainly aneuploidy) approximately 50–70% of the time (Fritz *et al*., 2001; Gardner and Sutherland, 2004).

To date, more than 1800 articles have been published on CGH with the most reporting the utility of CGH to identify novel chromosomal imbalances in neoplastic specimens (**http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?**). Less that 10% of CGH papers have dealt with technical aspects and only a limited number have described the application of CGH in a clinical cytogenetics setting. CGH has been shown to be a valuable tool in clinical cytogenetics by allowing for the characterization, in a single hybridization, of deletions, intrachromosomal duplications, unbalanced translocations, and marker chromosomes, including neocentric marker chromosomes (Kirchhoff *et al*., 2001; Levy *et al*., 1998; Levy *et al*., 2000). Various CGH studies have provided clinical correlation between specific chromosomal regions of imbalance and various clinical entities such as the duplication 5q phenotype (Levy *et al*., 2002) and partial tetrasomy 10p which resulted from an analphoid marker chromosome with a neocentromere (Levy *et al*., 2000). Such studies have helped to further define critical chromosomal regions, which are associated with normal and adverse phenotypic outcomes (Levy *et al*., 2000; Levy *et al*., 2002) thus providing prognostic information for genetic counseling. This type of information benefits prenatally ascertained cases of marker chromosomes as it may provide couples with a means to make rational and informed decisions concerning the pregnancy. In pediatric cases, such information may provide the parents with a realistic prognosis and will be important for the clinical management of the infant.

6. CGH in preimplantation genetic diagnosis

Molecular cytogenetic analyses of human preimplantation embryo have revealed extremely high levels of chromosomal abnormality (Munne *et al*., 1993; Delhanty *et al*., 1997). Upwards of 70% of abnormally developing embryos show chromosome abnormalities and mosaicism (Munne *et al*., 1993; Delhanty *et al*., 1997). Many women referred for in vitro fertilization (IVF) are of advanced maternal age, with an elevated risk for conceiving and delivering a child with chromosome aneuploidies. The rate of aneuploidy (approximately 40%) in normally developing embryos from older women (>40 years) is about 10 times higher than the incidence of aneuploidies (4%) in embryos from younger women (20–34 years) (Munne *et al*., 1995). The presence of aneuploidy impacts directly on the success of implantation, which is also one reason why embryos from older women are subject to higher implantation failure rates (Munne *et al*., 1995; Gianaroli *et al*., 1997; Munne *et al*., 1999). Current preimplantation genetic diagnosis (PGD) aneuploidy is carried out via multicolor FISH on interphase blastomeres and typically includes screening for chromosomes 13, 16, 18, 21, 22, X and Y (Munne *et al*., 1998). The ability to screen for additional aneuploidies is hampered by the number of chromosome-specific probes that can be used per hybridization. The power of CGH to detect total aneuploidy makes it an appealing diagnostic tool for use in preimplantation genetic diagnosis (PGD) as the selective transfer of diploid embryos would presumably result in higher implantation rates as well as lower miscarriage rates (Wilton *et al*., 2001; Wells *et al*., 2002; Wells and Levy,

2003). Various investigators have demonstrated that CGH can detect aneuploidy and partial aneuploidy as small as 3 MB in a clinical setting (Levy *et al*., 1998; Kirchhoff *et al*., 1999). Since the smallest autosome is in excess of 50 MB, CGH analysis provides a very powerful and sensitive method for PGD detection of whole chromosome aneuploidy and a few studies have now effectively demonstrated this in polar bodies and preimplantation embryos (Wilton *et al*., 2001; Wells *et al*., 2002; Wells and Delhanty, 2000). The major stumbling block in this application is the technical difficulties associated with reliably amplifying the DNA from a single cell so that sufficient DNA is available for the CGH analysis. Newer whole genome amplification methodologies may provide a robust way for a CGH-type analysis to become an integral part of PGD for aneuploidy.

7. Conclusion

Comparative genomic hybridization permits detailed cytogenetic data to be obtained from a wide range of tissues not usually amenable to analysis by conventional cytogenetic methods. These include formalin-fixed neoplastic specimens, freshly obtained solid tumors, mitotically inactive cells derived from products of conception, and single cells biopsied from polar bodies and preimplantation embryos. As a cancer research tool, CGH has proven invaluable for the ascertainment of chromosomal duplications, amplifications, and deletions that contain critical genes, which contribute to the neoplastic transformation of a whole battery of tissues. The application of CGH to prenatal and pediatric samples has also proven extremely beneficial, allowing the delineation of complex or cryptic chromosomal rearrangements that could not be defined using classical cytogenetic techniques. The use of CGH as a tool for PGD may allow for the detection and preferential transfer of the embryos most likely to form a viable pregnancy and thus lead to improvements in the outcome of assisted reproductive procedures.

CGH methodology is currently evolving as chromosomal targets are being replaced by arrays consisting of genomic clones, which are spotted onto glass microscope slides (Pinkel *et al*., 1998). These technical advances allow for the detection of much smaller imbalances (100–200 kb), such as those observed in the Digeorge and Prader–Willi syndromes (Pinkel *et al*., 1998). Current arrays have genomic coverage spaced about 1 Mb apart, and ongoing efforts are being made by researchers to create tiled arrays that span the entire genome. This technology remains expensive and will likely not be a common finding in all cytogenetics laboratories for a good few years. The use of regular CGH for the time being still remains an attractive and powerful accessory to routine clinical cytogenetic analysis.

Further reading

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes, Chromosomes and Cancer*, **20**, 399–407.

References

Becher R, Korn WM and Prescher G (1997) Use of fluorescence in situ hybridization and comparative genomic hybridization in the cytogenetic analysis of testicular germ cell tumors and uveal melanomas. *Cancer Genetics and Cytogenetics*, **93**, 22–28.

Blennow E, Bui TH, Kristoffersson U, Vujic M, Annerén G, Holmberg E and Nordenskjold M (1994) Swedish survey on extra structurally abnormal chromosomes in 39 105 consecutive prenatal diagnoses: Prevalence and characterization by fluorescence in situ hybridization. *Prenatal Diagnosis*, **14**, 1019–1028.

Bryndorf T, Kirchhoff M, Rose H, Maahr J, Gerdes T, Karhu R, Kallioniemi A, Christensen B, Lundsteen C and Philip J (1995) Comparative genomic hybridization in clinical cytogenetics. *American Journal of Human Genetics*, **57**, 1211–1220.

Delhanty JD, Harper JC, Ao A, Handyside AH and Winston RM (1997) Multicolour FISH detects frequent chromosomal mosaicism and chaotic division in normal preimplantation embryos from fertile patients. *Human Genetics*, **99**, 755–760.

Drets ME and Shaw MW (1971) Specific banding patterns of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 2073–2077.

Franke UC, Scambler PJ, Loffler C, Lons P, Hanefeld F, Zoll B and Hansmann I (1994) Interstitial deletion of 22q11 in DiGeorge syndrome detected by high resolution and molecular analysis. *Clinical Genetics*, **46**, 187–192.

Fritz B, Hallermann C, Olert J, Fuchs B, Bruns M, Aslan M, Schmidt S, Coerdt W, Muntefering H and Rehder H (2001) Cytogenetic analyses of culture failures by comparative genomic hybridisation (CGH)-Re-evaluation of chromosome aberration rates in early spontaneous abortions. *European Journal of Human Genetics*, **9**, 539–547.

Gardner RJM and Sutherland GR (2004) *Chromosome Abnormalities and Genetic Counseling*, Third Edition, Oxford University Press: New York.

Gianaroli L, Magli MC, Ferraretti AP, Fiorentino A, Garrisi J and Munne S (1997) Preimplantation genetic diagnosis increases the implantation rate in human in vitro fertilization by avoiding the transfer of chromosomally abnormal embryos. *Fertility and Sterility*, **68**, 1128–1131.

Hirschhorn K, Lucas M and Wallace I (1973) Precise identification of various chromosomal abnormalities. *Annals of Human Genetics*, **36**, 375–379.

Hsu TC (1952) Mammalian chromosomes *in vitro*: The karyotype of man. *Journal of Heredity*, **43**, 167–172.

Hughes A (1952) *The Mitotic Cycle*, Academic Press: New York.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kallioniemi OP, Kallioniemi A, Piper J, Isola J, Waldman FM, Gray JW and Pinkel D (1994) Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. *Genes, Chromosomes and Cancer*, **10**, 231–243.

Kirchhoff M, Gerdes T, Maahr J, Rose H, Bentz M, Dohner H and Lundsteen C (1999) Deletions below 10 megabasepairs are detected in comparative genomic hybridization by standard reference intervals. *Genes, Chromosomes and Cancer*, **25**, 410–413.

Kirchhoff M, Gerdes T, Rose H, Maahr J, Ottesen AM and Lundsteen C (1998) Detection of chromosomal gains and losses in comparative genomic hybridization analysis based on standard reference intervals. *Cytometry*, **31**, 163–173.

Kirchhoff M, Rose H and Lundsteen C (2001) High resolution comparative genomic hybridisation in clinical cytogenetics. *Journal of Medical Genetics*, **38**, 740–744.

Klever M, Grond-Ginsbach CJ, Hager HD and Schroeder-Kurth TM (1992) Chorionic villus metaphase chromosomes and interphase nuclei analysed by chromosomal in situ suppression (CISS) hybridization. *Prenatal Diagnosis*, **12**, 53–59.

Kuo WL, Tenjin H, Segraves R, Pinkel D, Golbus MS and Gray J (1991) Detection of aneuploidy involving chromosomes 13, 18, or 21, by fluorescence in situ hybridization (FISH) to interphase and metaphase amniocytes. *American Journal of Human Genetics*, **49**, 112–119.

Kuwano A, Ledbetter SA, Dobyns WB, Emanuel BS and Ledbetter DH (1991) Detection of deletions and cryptic translocations in Miller-Dieker syndrome by in situ hybridization. *American Journal of Human Genetics*, **49**, 707–714.

Lawce HJ and Brown MG (1997) Cytogenetics: An overview. In *The AGT Cytogenetics Laboratory Manual*, Barch MJ, Knutsen T and Spurbeck JL (Eds.), Lippincott-Raven: New York, pp. 19–50.

Leana-Cox J, Levin S, Surana R, Wulfsberg E, Keene CL, Raffel LJ, Sullivan B and Schwartz S (1993) Characterization of de novo duplications in eight patients by using fluorescence in situ hybridization with chromosome-specific DNA libraries. *American Journal of Human Genetics*, **52**, 1067–1073.

Levy B, Dunn TM, Kaffe S, Kardon N and Hirschhorn K (1998) Clinical applications of comparative genomic hybridization. *Genetics in Medicine*, **1**, 4–12.

Levy B, Dunn TM, Kern JH, Hirschhorn K and Kardon NB (2002) Delineation of the dup5q phenotype by molecular cytogenetic analysis in a patient with dup5q/del 5p (cri du chat). *American Journal of Medical Genetics*, **108**, 192–197.

Levy B, Papenhausen PR, Tepperberg JH, Dunn TM, Fallet S, Magid MS, Kardon NB, Hirschhorn K and Warburton PE (2000) Prenatal molecular cytogenetic diagnosis of partial tetrasomy 10p due to neocentromere formation in an inversion duplication analphoid marker chromosome. *Cytogenetics and Cell Genetics*, **91**, 165–170.

Makino S and Nishimura I (1952) Water pretreatment squash technic. A new and simple practical method for the chromosome study of animals. *Stain Technology*, **27**, 1–7.

Manning JE, Hershey ND, Broker TR, Pellegrini M, Mitchell HK and Davidson N (1975) A new method of in situ hybridization. *Chromosoma*, **53**, 107–117.

Mitelman F (1994) *Catalog of Chromosome Abberations in Cancer*, Fifth Edition, Wiley-Liss: New York.

Munne S, Alikani M, Tomkin G, Grifo J and Cohen J (1995) Embryo morphology, developmental rates, and maternal age are correlated with chromosome abnormalities. *Fertility and Sterility*, **64**, 382–391.

Munne S, Lee A, Rosenwaks Z, Grifo J and Cohen J (1993) Diagnosis of major chromosome aneuploidies in human preimplantation embryos. *Human Reproduction*, **8**, 2185–2191.

Munne S, Magli C, Bahce M, Fung J, Legator M, Morrison L, Cohert J and Gianaroli L (1998) Preimplantation diagnosis of the aneuploidies most commonly found in spontaneous abortions and live births: XY, 13, 14, 15, 16, 18, 21, 22. *Prenatal Diagnosis*, **18**, 1459–1466.

Munne S, Magli C, Cohen J, Morton P, Sadowy S, Gianaroli L, Tucker M, Marquez C, Sable D, Ferraretti AP, *et al*. (1999) Positive outcome after preimplantation diagnosis of aneuploidy in human embryos. *Human Reproduction*, **14**, 2191–2199.

Najfeld V (1997) FISHing among myeloproliferative disorders. *Seminars in Hematology*, **34**, 55–63.

Patil SR, Merrick S and Lubs HA (1971) Identification of each human chromosome with a modified Giemsa stain. *Science*, **173**, 821–822.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Pinkel D, Straume T and Gray JW (1986) Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 2934–2938.

Piper J, Rutovitz D, Sudar D, Kallioniemi A, Kallioniemi OP, Waldman FM, Gray JW and Pinkel D (1995) Computer image analysis of comparative genomic hybridization. *Cytometry*, **19**, 10–26.

Seabright M (1971) A rapid banding technique for human chromosomes. *Lancet*, **2**, 971–972.

Sumner AT, Evans HJ and Buckland RA (1971) New technique for distinguishing between human chromosomes. *Nature: New Biology*, **232**, 31–32.

Wells D and Delhanty JD (2000) Comprehensive chromosomal analysis of human preimplantation embryos using whole genome amplification and single cell comparative genomic hybridization. *Molecular Human Reproduction*, **6**, 1055–1062.

Wells D and Levy B (2003) Cytogenetics in reproductive medicine: The contribution of Comparative Genomic Hybridization (CGH). *Bioessays*, **25**, 289–300.

Wells D, Escudero T, Levy B, Hirschhorn K, Delhanty JDA and Munne S (2002) First clinical application of comparative genomic hybridization (CGH) and polar body testing for preimplantation genetic diagnosis (PGD) of aneuploidy. *Fertility and Sterility*, **78**, 543–549.

Wilton L, Williamson R, McBain J, Edgar D and Voullaire L (2001) Birth of a healthy infant after preimplantation confirmation of euploidy by comparative genomic hybridization. *The New England Journal of Medicine*, **345**, 1537–1541.

Cytogenetic analysis of lymphomas

Douglas E. Horsman
British Columbia Cancer Agency, Vancouver, BC, Canada

1. Introduction

Historically, the application of cytogenetic analysis to the investigation of malignant lymphoma has provided the entry point for the identification of critical gene deregulations associated with specific subtypes of B- and T-cell malignant lymphomas (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1). These include the IGH-MYC gene fusion created by the t(8;14)(q24;q32) of Burkitt leukemia/lymphoma, the IGH-CCND1 fusion resulting from the t(11;14)(q13;q32) of mantle cell lymphoma, and the IGH-BCL2 fusion associated with the t(14;18)(q32;q21) of follicular lymphoma and the BCL6 translocations found in diffuse large cell lymphomas, as well as other less common examples. This important cancer genetic information has been accessible to laboratory investigation due to the relative ease with which tissue specimens can be obtained from patients affected by these diseases and the ability to successfully propagate these cell samples in vitro, allowing examination of the chromosomal makeup of the malignant cells. The investigation of malignant lymphoma karyotypes continues to have important research implications, with a shifting of emphasis away from the well-defined primary alterations mentioned above to a scrutiny of the secondary chromosomal changes that characterize clonal expansion and disease evolution. These secondary changes may have an important role in determining the pace of disease progression or the transformation of low-grade disease to a more aggressive type of disease. At the clinical level, the detection of disease-specific changes may be of great help to the pathologist and oncologist to determine the proper diagnosis and to detect important prognostic or predictive factors that will aid in treatment planning and disease follow-up. The importance of genetic analysis in the assessment of malignant haematopoietic disorders has been emphasized in the most recent version of the World Health Organization Classification of Tumours: Pathology and Genetics of Haematopoietic and Lymphoid Tissues (2001).

The successful application of cytogenetic analysis to malignant lymphoma specimens is predicated on appropriate sampling of representative diseased tissue from the patient. The specimen must be transported to the lab in a viable condition, followed by optimal culturing, harvesting, slide making, chromosome

banding, and microscopic analysis to allow the identification of cell divisions or metaphases that are representative of the clonal population within the tissue sample. Numerous technical resources are available that describe the theoretical and technical basis of cytogenetic analysis of constitutional and cancer cell specimens as well as the standard nomenclature that is used in clinical practice for the description of normal and altered chromosome morphology. For these types of information, the interested reader is referred to the ACT Cytogenetic Laboratory Manual (1991) and the International System of Cytogenetic Nomenclature (ISCN, 1995) and other pertinent references (Therman, 1986; Verma and Babu, 1995). An in-depth description of the application of cytogenetic analysis to cancer specimens and the types of alterations found in various types of cancer, including malignant lymphoma, has been written by Heim and Mitelman (1995). In addition, a number of useful websites are available that can be accessed to search for chromosomal alterations associated with various types of cancer. These include the Atlas of Genetics and Cytogenetics in Oncology and Haematology (**http://babbage.infobiogen.fr/services/chromcancer/index.html**) and the Mitelman Database of Chromosome Aberrations in Cancer (**http://cgap.nci.nih.gov/Chromosomes/Mitelman**). Information on the description and classification of malignant lymphomas is provided in the World Health Organization Classification of Tumours: Pathology and Genetics of Haematopoietic and Lymphoid Tissues (2001).

2. Specimen collection and transport

The source of clinical specimens for cytogenetic analysis includes any tissue or fluid that may contain infiltrates of malignant lymphoma cells, such as lymph node biopsies, marrow aspirates, peripheral blood samples, fine needle aspirates, and fluids obtained from body cavities. Appropriate aseptic technique should be followed during specimen acquisition. The specimen may be sent directly to the laboratory, or placed in a minimal essential media (i.e., RPMI 1640) to ensure the preservation of viability during transit to the laboratory. The specimens should be kept cool, possibly in a 4°C refrigerator or on ice, with precautions taken to prevent freezing. Upon receipt in the laboratory, triage of the specimen to other investigations such as immunophenotyping and morphology examination may be appropriate, that is, for lymph nodes and marrow aspirates, or the entire specimen may be utilized for cytogenetic analysis if other investigations are not required. Histologic, cytologic, or immunophenotypic evaluation of a portion of the specimen may be helpful in confirming that the malignant cells are represented in the sample that has been submitted for cytogenetic analysis.

3. Culture and harvest, slide making, and chromosome banding

Solid specimens such as lymph node biopsies should be subjected to disaggregation by gentle mincing and vortexing in culture media or saline solution to obtain a

single cell suspension. Such cell suspensions, including aspirate or fluid specimens, are most suitable for the culture procedure. The standard methods used for leukemia specimens have proven successful for malignant lymphoma. A 24-h culture in a minimal essential media such as RPMI 1640 with the addition of fetal calf serum and glutamine supplement is normally used. Other growth stimulants or mitogenic agents are generally not required or recommended. In certain disease subtypes where the proliferation rate is extremely low, or when the clonal cells are sensitive to apoptosis in vitro, the use of growth stimulants may be helpful to obtain metaphases; however, such conditions or agents may induce selective growth of normal cells or subpopulations of the malignant cell population, thus rendering bias into the subsequent karyotype interpretation.

Harvesting of the cultured cells, the making of slides to obtain metaphase spreads and the banding of the chromosomes should be undertaken using standard approaches that have been adapted to local laboratory conditions and the experience of the practitioners (*see* Article 12, **The visualization of chromosomes**, Volume 1). A variety of banding techniques, including Q-, G-, and R-banding have been utilized, with equal success for the interpretation of numerical and structural chromosomal alterations associated with lymphoma. When possible, the documentation of both normal and abnormal metaphases from the specimen should be sought, the former to ensure that possible inherited chromosomal alterations in normal cells are not misinterpreted to represent acquired, lymphoma-related alterations.

4. Metaphase analysis

The great majority of malignant lymphomas will show clonal karyotypic alterations, with some notable exceptions, such as chronic lymphocytic leukemia where the clonal cells are sensitive to apoptosis in vitro, rendering the identification of clonal metaphases difficult, and in Hodgkin disease, where the proportion of malignant cells in the specimen represents a small minority of the total cell population and obtaining metaphases from these malignant cells is notoriously difficult, requiring the use of molecular methods to detect possible chromosome rearrangements and dosage alterations. In most lymphoma subtypes, however, a clonal karyotype will be evident with the analysis of 10 to 20 metaphases. The complexity of the karyotypes may vary from simple whole-chromosome gains or losses such as +3 or −13, single balanced translocations such as t(8;14)(q24;q32), to more complex karyotypes with a combination of chromosomal numerical and structural alterations including balanced and unbalanced translocations, inversions, insertions, deletions, duplications, and the presence of complex derivative chromosomes called marker or ring chromosomes or chromosomal additions where the origin of the extra chromosomal material cannot be ascertained from the evident banding pattern (see Figure 1) (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1). The description of the karyotypic changes should conform to the guidelines outlined in the International System of Cytogenetic Nomenclature (ISCN, 1995). In certain situations, the short-form description for observed chromosome alterations may not be sufficient to accurately describe



Figure 1 A G-banded karyotype of a representative case of malignant lymphoma with a primary chromosomal translocation t(14;18)(q32;q21) (arrows point to the derivative 14 and 18 chromosomes) and a single secondary change consisting of an extra copy of chromosome 12 (arrow)

the evident changes, and the more detailed, long-form description of individual derivative chromosomes may be required.

Most lymphoma karyotypes are in the near-diploid to hyperdiploid range, with polyploidy karyotypes being found in a minority of cases. These polyploid karyotypes may be in the near-triploid to near-tetraploid chromosome complement range, and are seldom, if ever, balanced. Single unique stemlines may be identified, but closely related sidelines are often evident if sufficient metaphases are examined. Such sidelines most commonly represent the result of sequential or divergent evolution of the clonal karyotype, with preservation of preexisting stemline changes in the evolved sidelines. Occasionally, independent clones may be identified with no shared alteration that links them to a common precursor stemline. These may harbor subcytogenetic alterations that link them to a common stemline or they may truly represent multiclonal proliferations. In some cases, there is a marked variation between sidelines and even between individual metaphases, and such heterogeneity being indicative of an extreme level of genetic instability.

5. Clinical indications for cytogenetic studies

The clinical utility of a full karyotype analysis of a malignant lymphoma should be determined by the need for such information to assist in patient management or follow-up. A partial list of disease-specific chromosomal alterations that have been detected in malignant lymphoma is provided in Table 1. Interpretation of chromosomal alterations in lymphoma specimens should always be undertaken in conjunction with information obtained from the morphologic and phenotypic evaluation of the specimen and with knowledge of the clinical situation that initiated the investigation. The necessity for chromosomal data in malignant lymphoma has not yet reached the level of importance that is assumed for the leukemias, and

Table 1 Chromosomal translocations associated with specific subtypes of malignant lymphoma

Chromosome alteration	Gene alteration	Disease subtype
t(8;14)(q24;q32)	IGH-MYC	Burkitt lymphoma
t(11;14)(q13;q21)	IGH-CCND1	Mantle cell lymphoma
t(14;18)(q32;q21)	IGH-BCL2	Follicular lymphoma
t(3;14)(q27;q32)	IGH-BCL6	Diffuse large cell lymphoma
t(11;18)(q21;q32)	API1-MALT1	Extranodal marginal zone lymphoma
del(7q) or dup(3q)	Not known	Splenic marginal zone lymphoma
t(2;5)(p23;q35)	ALK-NPM	Anaplastic large cell lymphoma
+3, +15, +X	Not known	Angioimmunoblastic lymphadenopathy
del(9q34-34)	Not known	Enteropathy-associated T-cell lymphoma
dup(7q)	Not known	Hepatosplenic gamma/delta T-cell lymphoma

in many situations surrogate information obtained by immunophenotyping may be adequate to infer the presence of an underlying chromosomal alteration. Currently, a number of clinically useful indications for chromosome analysis can be cited, including the differential diagnosis of Burkitt leukemia/lymphoma from other high-grade lymphomas, the identification of ALK translocations in anaplastic large cell lymphomas, and the differential diagnosis of mantle cell lymphoma from atypical forms of chronic lymphocytic leukemia or other low-grade lymphoproliferative disorders. The diagnosis and management of the common forms of follicular lymphoma and diffuse large cell lymphoma does not routinely require chromosomal information. However, as more information on the prognostic influence of BCL6 rearrangements and other gene deregulations becomes established, methods to obtain objective information on these genetic factors may be required to assist in the clinical management of a larger proportion of lymphoma patients. For many indications, the availability of fluorescence in situ hybridization (FISH) probes to detect specific chromosomal changes has supplanted the need for full karyotypic analysis (Siebert and Weber-Matthiesen, 1997). This is exemplified in chronic lymphocytic leukemia, where it has been demonstrated that prognostic subgroups can be identified on the basis of the presence or absence of trisomy 12 and deletion of 11q22, 17p13, and 13q14. This type of multigene assessment is best undertaken by FISH analysis using specific DNA probes for the chromosomal region or genes involved, owing to the difficulty in obtaining metaphases from chronic lymphocytic leukemia samples and the often cryptic nature of the deletions affecting these sites.

6. Supplemental molecular cytogenetics

In situations in which the exact chromosomal composition of certain abnormal chromosomes cannot be confidently determined by the chromosomal banding pattern, verification can be obtained through the application of appropriate FISH techniques, using chromosome centromere-specific probes, locus-specific probes, chromosome painting probes and multicolor karyotyping reagents (*see* Article 22, **FISH**, Volume 1). Such FISH techniques can be applied to a variety of cell sources, including imprints made from biopsy specimens, air-dried cell suspensions

or smears (i.e., from blood, marrow or fluid samples), from the methanol-acetic acid cells pellets remaining from the specimen used for karyotype analysis, and also from frozen or fixed tissue. Of particular importance is the ability to use tissue samples that have been preserved in formalin and embedded in paraffin for histological interpretation. These samples can be used either as paraffin sections mounted on glass slides, or as disaggregated nuclei obtained by dissection from the paraffin block, targeting specific areas in the block to ensure appropriate representation of the disease in the sample. A particularly useful approach in this regard is a tissue array coring device, which can be used to sample 0.5- or 1-mm cores of paraffin-embedded tissue, selected on the basis of an H&E slide sectioned from the same block. The core or section of tissue must be subjected to appropriate deparaffinization and pretreatment prior to hybridization with the DNA probe.

7. Types of FISH probes for malignant lymphoma investigation

Locus-specific FISH: a variety of probes are available from commercial sources, such as Vysis Inc. that have been developed and optimized for specific lymphoma translocation or rearrangement analysis, such as MYC, BCL6, t(8;14), t(11;14), t(14;18), and ALK. The dual-color split-apart probes are particularly useful for the

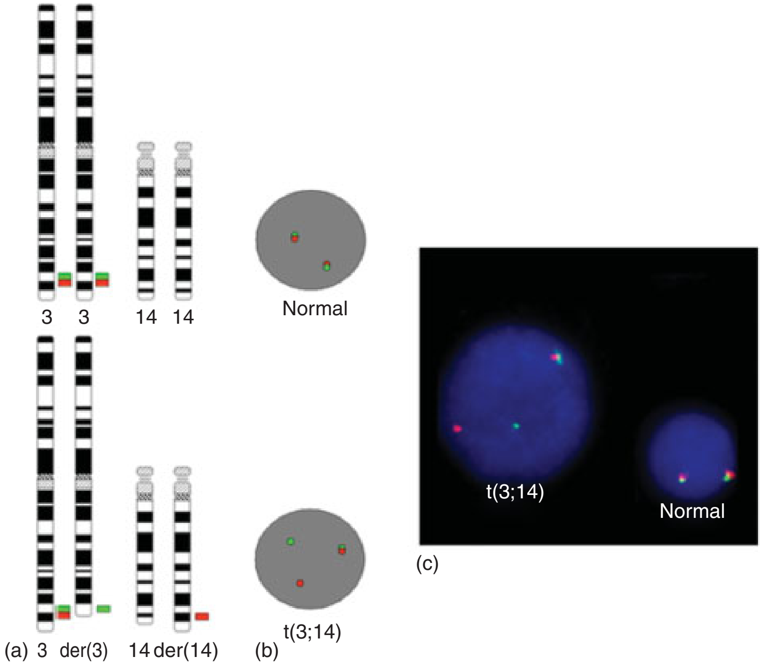


Figure 2 The hybridization patterns produced by a dual-color split-apart FISH probe used to detect BCL6 translocations in malignant lymphoma. The normal gene configuration shows two red-green fusion signals in each normal nucleus. A malignant cell nucleus with a BCL6 translocation such as t(3;14) shows one red, one green, and one fusion signal

interrogation of the configuration of an individual gene when multiple translocation partners have been identified, such as for BCL6 and ALK (see Figure 2). The MYC probe may be helpful in determining the possible presence of variant translocations not detected by the t(8;14) probe. Similarly, probing with a BCL2 probe may be necessary to detect variant translocations associated with follicular lymphoma, since commercial sources of the kappa and lambda immunoglobulin gene probes are not currently available to detect IGL-BCL2 translocations. Dual color, dual fusion probes are particularly useful for the detection of chromosomal translocations associated with Burkitt leukemia/lymphoma, follicular lymphoma, or mantle cell lymphoma (see Figure 3). Variant translocation signal patterns may be obtained in some cases, because of associated deletions and duplications, rendering interpretation difficult. In these situations, the availability of metaphases in the preparation may help resolve the interpretation of the signal pattern, where the signal localizations on individual chromosomes can be determined from the reverse

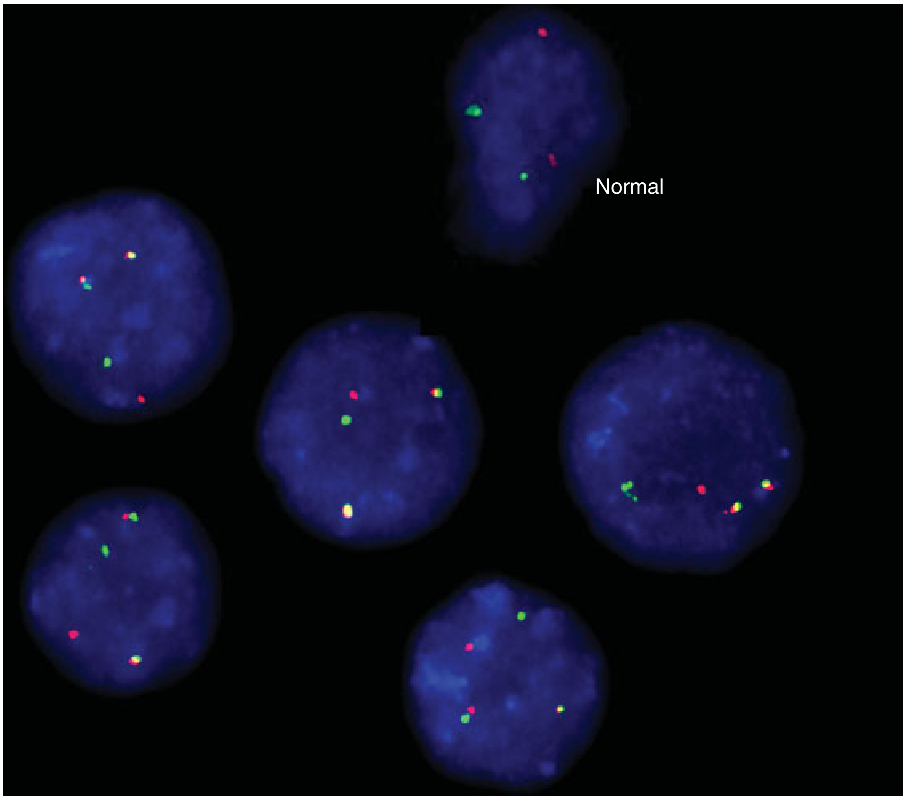


Figure 3 The hybridization patterns found with a dual color, dual fusion translocation FISH probe used to detect translocations such as the t(11;14)(q13;q32) in malignant lymphoma. The normal gene configuration shows two red and two green signals in each nucleus, indicating the two copies each of the IGH and CCND1 genes in normal nuclei. A nucleus with a t(11;14) shows one red, one green, and two fusion signals, indicating the presence of IGH-CCND1 fusions

banding pattern on the chromosomes resulting from the DAPI (4′-6′-diamidino-2′-phenylindole) staining, or by comparing to G-banded metaphases if these are also available. If only interphase nuclei are available, such as from paraffin-embedded tissue, an accurate interpretation of atypical signal patterns may not be possible.

Currently, there are a limited number of probes available for malignant lymphoma investigation, although additional commercial probes are continually being developed and marketed. Alternatively, readily available bacterial artificial chromosomes (BACs) or other vector constructs containing known fragments of human DNA can be obtained for preparation of "in house" or "home brew" FISH probes. Appropriate propagation, purification, labeling, and validation of the probe with local positive and negative controls is required prior to the utilization of such probes for clinical purposes. Protocols are available through organizations such as the National Committee for Clinical Laboratory Standards (NCCLS) (**http://www.nccls.org/**) to guide the preparation and validation of such probes.

Multicolor Karyotyping: The development of multicolor chromosome painting techniques, including the so-called SKY and MFISH methods (Schrock *et al*., 1996; Speicher *et al*., 1996), have provided the capability to more accurately

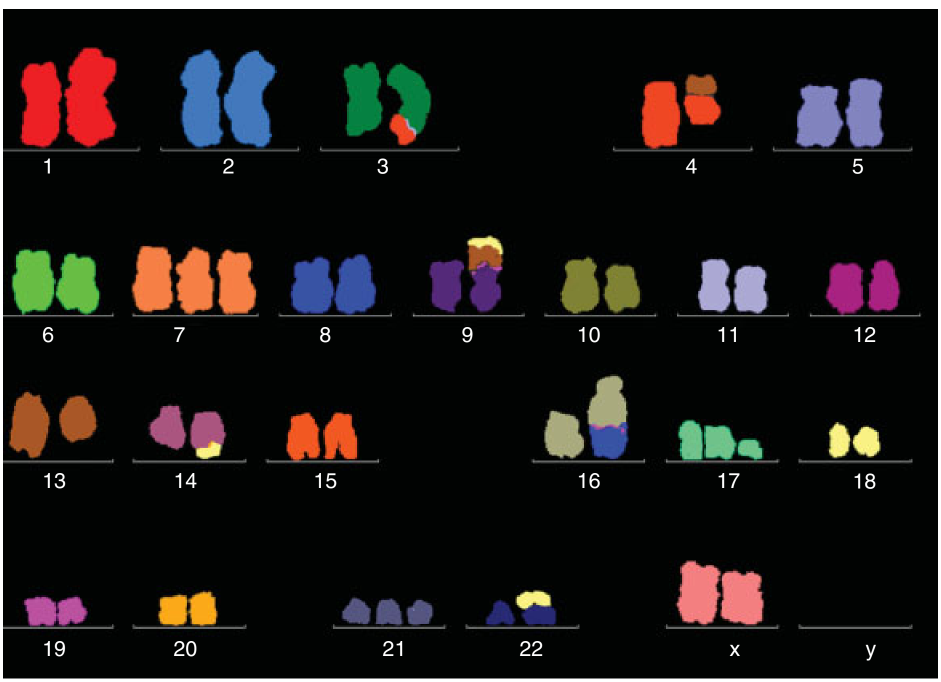


Figure 4 MFISH image of a complex lymphoma karyotype using the Metasystems MFISH reagent. Each chromosome is identified by a unique assigned color based on a unique fluorescence signal. The karyotype contains a t(14;18) with only the derivative chromosome 14 apparent, as the small portion of chromosome 14 that is translocated to chromosome 18 cannot be visualized with this reagent. Additional secondary chromosome changes are evident, including an extra whole chromosome 7 and 21, an additional partial chromosome 17, unbalanced translocations between 3 and 4, 4 and 13, 16 and 8, and 22 and 18, and a more complex derivative 9 containing material from chromosomes 9, 13, and 18

define complex chromosomal changes that have defied interpretation by standard chromosome banding methods. In particular, it has allowed the deciphering of the chromosomal makeup of marker chromosomes and ring chromosomes, and chromosomal additions were the small pieces of extra chromatin that cannot be identified by chromosomes banding methods. The application of these reagents has shown that many marker and derivative chromosomes are made up of multiple segments of material from different chromosomal sources (see Figure 4). In some situations, the regional source of this material can be deduced from the chromosome banding pattern, but if the segment is small, it may not be possible to determine whether it comes from the p or q arm of the donating chromosome. In these situations, additional verification using locus specific probes or multicolor banding reagents (see Figure 5) (Chudoba *et al*., 1999) may be required to fully ascertain the identity of the chromosomal segments within these complex rearranged chromosomes.

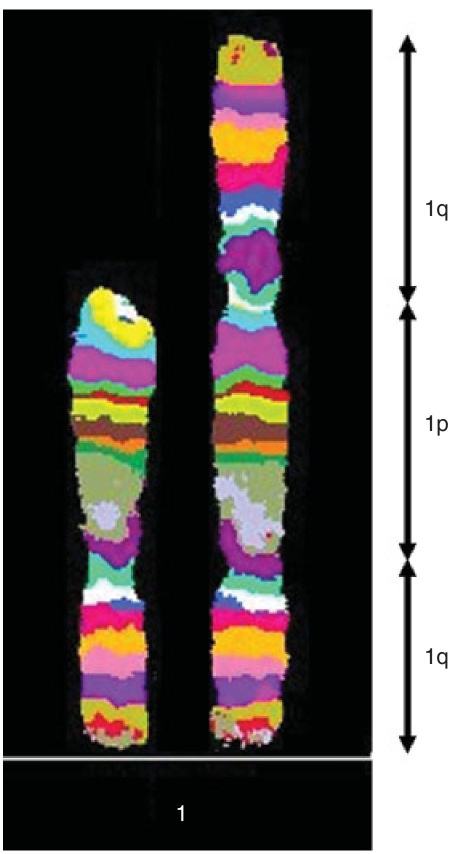


Figure 5 Multicolor banding of chromosome 1 using the Metasystems MBAND1 probe reagent. The two chromosomes on the left depict normal copies of chromosome 1. The two chromosomes on the right display a large regional duplication of the q arm of chromosome 1

Acknowledgments

I would like to thank the cytogenetic technologists of the BC Cancer Agency for the provision of the images for Figures 1, 2, and 3, Dr. Valia Lestou for the images for Figures 4 and 5, and Ms. Chris Salski for proofreading of the manuscript.

References

Barch MJ (1991) *The ACT Cytogenetics Laboratory Manual*, Raven Press: New York.

Chudoba I, Plesch A, Lorch T, Lemke J, Claussen U and Senger G (1999) High resolution multicolor-banding: A new technique for defined FISH analysis of human chromosomes. *Cytogenetics and Cell Genetics*, **84**, 156–160.

Heim S and Mitelman F (1995) *Cancer Cytogenetics: Chromosomal and Molecular Genetic Aberrations of Tumor Cells*, Wiley-Liss: New York.

ISCN (1995) *International System for Cytogenetic Nomenclature*, Karger: Basil.

Jaffe ES, Harris NL, Stein H and Vardiman JW (2001) *World Health Organization Classification of Tumours: Pathology and Genetics of Haematopoietic and Lymphoid Tissues*, IARC Press: Lyon.

Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, *et al*. (1996) Multicolor spectral karyotyping of human chromosomes. *Science*, **273**, 494–497.

Siebert R and Weber-Matthiesen K (1997) Fluorescence in situ hybridization as a diagnostic tool in malignant lymphomas. *Histochemistry and Cell Biology*, **108**, 391–402.

Speicher MR, Gwyn Ballard S and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, **12**, 368–375.

Therman E (1986) *Human Chromosomes: Structure, Behavior, Effects*, Springer-Verlag: New York.

Verma RS and Babu A (1995) *Human Chromosomes: Principles and Techniques*, McGraw-Hill: New York.

Human sperm–FISH for identifying potential paternal risk factors for chromosomally abnormal reproductive outcomes

Andrew J. Wyrobek, Thomas E. Schmid and Francesco Marchetti
University of California, Livermore, CA, USA

1. Introduction

A substantial challenge in the identification of human germ cell mutagens is to classify offspring carrying inherited genetic defects, to identify the responsible parent who transmitted the defect, and to determine whether specific genetic defects were caused by prefertilization exposure to environmental, occupational, medical, or other agents. The relatively large baseline frequencies of abnormal reproductive outcomes among human beings contribute to this problem. Every year in the United States, more than 20 million conceptions are lost before the 20th week of gestation and among these, about 50% carry numerical or structural chromosomal defects (McFadden and Friedman, 1997). Chromosomally defective offspring who survive to birth are at higher risks of malformations and other negative health effects (Jacobs, 1992). There is growing evidence that a substantial fraction of chromosomal defects, especially structural aberrations, are transmitted by the sperm (Robbins *et al*., 1997), pointing to the need for effective sperm assays for transmissible genetic damage. However, the sperm nucleus presents an unusual biophysical challenge. It is very dense with highly cross-linked chromatin, making it difficult to assess its chromosomal content.

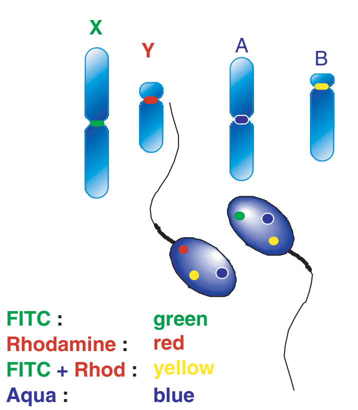
Rudak *et al*. (1978), working in the Yanagamachi laboratory in Hawaii, developed one of the first direct methods for analyzing human-sperm chromosomes (human-sperm/hamster-egg cytogenetic method, or hamster-egg method for short) (review by Brandriff *et al*., 1994; *see also* Article 12, **The visualization of chromosomes**, Volume 1). Using this method, human-sperm chromosomes can be examined by conventional cytogenetic staining at the first metaphase after fusing capacitated human sperm with enzymatically denuded hamster oocytes. This highly efficient matchup has not been equaled using gametes from other species. The hamster-egg system showed that sperm of normal men are about 2% aneuploid

with 2–7% having chromosomal structural aberrations, and that certain cancer therapies induced higher frequencies of chromosomally abnormal sperm (review by Brandriff *et al*., 1994). However, this technique is very labor-intensive; only a handful of laboratories ever mastered it, and no recent publication has utilized it.

2. Sperm-FISH assay

In the late 1980s, fluorescence in situ hybridization (FISH) (*see* Article 22, **FISH**, Volume 1) technology was adapted for the detection of chromosomally defective sperm (e.g., Wyrobek *et al*., 1990). It is an extremely simple procedure, compared to the hamster technique. Briefly, semen is smeared onto glass slides and sperm chromatin is decondensed (e.g., by dithiothreitol) so that fluorescently labeled chromosome-region-specific DNA-probes can penetrate the chromatin to hybridize to the target region in the sperm nucleus (Figure 1).

A variety of human sperm-FISH assays have been developed for detecting aneuploidy (Figure 1). Combinations of up to four chromosome-specific DNA-probes are each labeled with a different fluorescent color. Individual sperm are also marked with a DNA dye (such as DAPI) to identify the nucleus, and the number of specifically colored fluorescent domains within the nucleus are counted under the fluorescence microscope. Sperm with two domains of the same color are assumed to be disomic, while sperm lacking the same domain are assumed to be nullisomic for that chromosome. Controlling of technical factors is critical for the reliability of the sperm-FISH assay, especially when small changes are expected between exposed and control groups (Schmid *et al*., 2001). Comparisons between different laboratories have demonstrated the importance of training, harmonizing scoring criteria, and rigorously blinding scorers. More recently, a sperm-FISH technique was developed (ACM assay) to detect sperm carrying structural as well as numerical chromosomal abnormalities, as illustrated in Figure 2 (Sloter *et al*., 2000). The ACM FISH assay uses DNA-probes specific for three regions of chromosome 1 to detect human sperm that carry numerical chromosomal abnormalities plus two



Sperm genotype →	Syndrome/loss
Sex chromosomes	
X - Y	**X-X-Y,** Klinefelter
X - X	**X-X-X**
Y - Y	**X-Y-Y**
0	**X-0,** Turner
Autosomal disomy	
8 - 8	Early loss
16 - 16	Spont. abortion
13 - 13	Patau
18 - 18	Edwards
21 - 21	Down

Figure 1 Four-chromosome FISH for detecting aneuploid human sperm

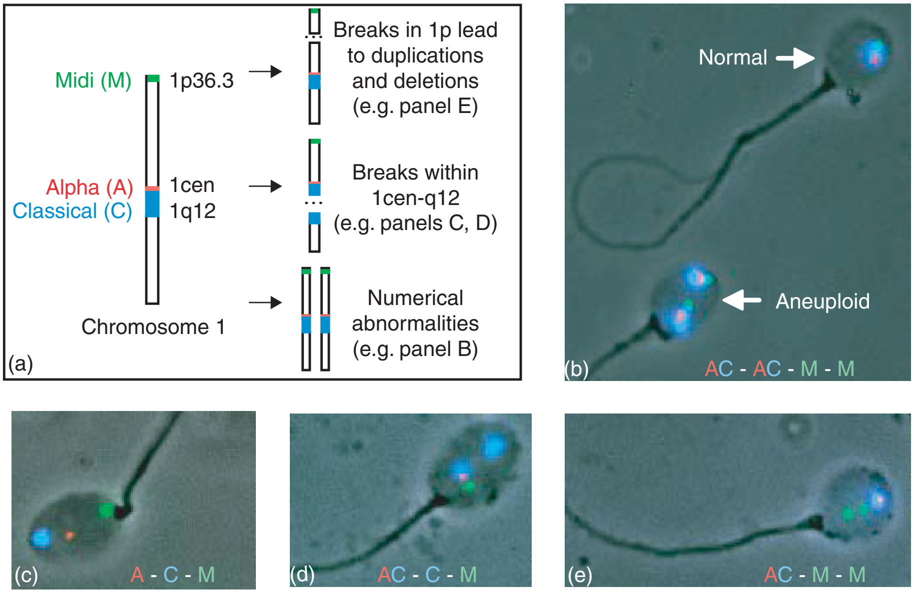


Figure 2 ACM human sperm–FISH for detecting sperm carrying either numerical or structural chromosomal abnormalities

categories of structural aberrations: duplications and deletions of 1pter and 1cen, and chromosomal breaks within the 1cen-1q12 region.

3. Advantages of the sperm-FISH assay

Sperm-FISH methods are gaining in popularity for assessing the effects of exposures because of their relative ease in collecting data when compared with epidemiological studies of human offspring, animal breeding studies, or the hamster technique (Brandriff *et al*., 1994). Compared to the hamster-egg method, which required fresh samples, sperm-FISH can utilize frozen semen specimens (for example, sperm-FISH has been effective for sperm stored for over 20 years). In addition, sperm-FISH requires much less scoring time allowing the analyses of 10 000 sperm and more per sample compared to 10–100 sperm typically scored by the hamster-egg method. The advantage of analyzing large numbers of sperm in a relatively short amount of time by sperm-FISH confers a relatively high level of sensitivity and statistical power to these assays, so that small increases can be detected by analyzing sperm from a small number of donors per experimental group. Table 1 shows the baseline variations for sperm aneuploidy and aberrations using data from the ACM assay in a group of young healthy men. Structural chromosomal abnormalities occur at higher frequencies than numerical abnormalities, while breaks are more prevalent than partial duplications and deletions. Men can vary significantly in their baseline frequencies for specific classes of chromosomally abnormal sperm, and these differences can persist over years affecting aneuploidy in both sperm and blood (Rubes *et al*., 1998).

Table 1 Chromosomally abnormal sperm detected by ACM human sperm–FISH: baseline frequencies among healthy individuals and the statistical sample size requirements for detecting increases in frequencies after toxicant exposure and varying lifestyle factors

Categories of sperm chromosomal defects detected by ACM sperm-FISH	Baseline frequencies[a]	Sample size to detect percentage increase or decrease[b]	
		50%	100%
Segmental aneuploidies			
1 pter duplication	6.4 ± 3.1	20	5
1 pter deletion	4.4 ± 4.7	96	24
Total	10.8 ± 5.5	22	6
1 cen-1q12 duplication	1.0 ± 1.3	143	36
1 cen-1q12 deletion	0.8 ± 1.4	258	65
Total	1.8 ± 1.6	67	17
Chromosomal breaks			
Between 1cen and 1q12	1.5 ± 1.7	108	27
Within 1q12	3.1 ± 2.3	47	12
Total	4.6 ± 3.2	41	10
Disomy and diploidy			
Disomy 1 or diploidy	22.7 ± 11.2	21	6

[a]Per 10 000 sperm: Mean + SD of 10 men (20–30 years of age, nonsmokers).
[b]Sample size = number of men needed per group; equal number of exposed and controls so that the total number of men for a comparison study would be twice that shown (10 000 sperm per sample).

Sperm-FISH provides a promising approach to identifying potentially damaging host factors and exposures that may increase the production of chromosomally defective sperm. Understanding these risk factors would help us design better epidemiological investigations of paternally transmitted abnormal reproductive outcomes. Table 1 provides statistical estimates of the sample size requirements to detect induced increases of 50 or 100% over baseline frequencies in the normal population. These sample sizes are dependent on the means and variances in the normal population and on the expected increase and variances among the exposed. The rarer the abnormality among normal men, the higher is the number of donors needed per group to detect an induced effect. In general, a doubling of the frequencies of chromosomally defective sperm can be detected with a sample size of about 10 men per group (range 6 to 65 for various subtypes of damage, Table 1). Sperm-FISH has already identified several risk factors and exposures that may increase the frequencies of sperm with chromosomal abnormalities (Table 2), including lifestyle, medical drugs, and occupational exposures.

4. Interspecies applications and challenges of sperm-FISH

Sperm-FISH has the intrinsic advantage of being equally applicable to any laboratory and domestic species (Lowe *et al*., 1996; Lowe *et al*., 1998; Hill *et al*., 2003). Rodent sperm–FISH may provide a platform for systematic tests of the genetic damage to germ cells of the myriads of chemicals present in the environment, and

Table 2 Examples of applications of human sperm–FISH to identify exposure and lifestyle factors that induce sperm chromosomal abnormalities

Lifestyle factors	Caffeine	Robbins *et al*. (1997)
	Smoking	Robbins *et al*. (1997); Rubes *et al*. (1998); Shi *et al*. (2001)
	Alcohol	Robbins *et al*. (1997)
Occupational exposures	Pesticides	Padungtod *et al*. (1999); Recio *et al*. (2001); Smith *et al*. (2004)
	Benzene	Li *et al*. (2001); Liu *et al*. (2003); Zhao *et al*. (2004)
	Styrene	Naccarati *et al*. (2003)
	Acrylonitrile	Xu *et al*. (2003)
Medical drugs	Diazepam	Baumgartner *et al*. (2001)
	Chemotherapy	Martin *et al*. (1997); Robbins *et al*. (1997); Martin *et al*. (1999); De Mas *et al*. (2001); Frias *et al*. (2003)

for prioritizing human epidemiological studies of paternally mediated abnormal reproductive outcomes. The main limitation for developing a sperm-FISH assay in any new species is the availability of reliable chromosome-region-specific DNA-probes for that species. For humans, probes are now commercially available for all chromosomes. Also, as the interest in sperm-FISH across species continues it will be possible, for the first time, to compare the sperm response among species and to select the best animal model for screening for human male germ cell mutagens.

The new sperm-FISH assay for structural aberrations also provides a direct approach to identify host factors and environmental exposures that increase chromosomal damage in stem cells. Damage induced in these cells may persist throughout the reproductive life of the individual. Thus, combining the human-sperm ACM assay (Figure 2) with one of the sperm-FISH assays for aneuploidy (Figure 1) promises a robust approach for detecting paternally transmissible chromosomal damage of both the numerical and the structural type.

Sperm-FISH has several remaining challenges that, while unresolved, will continue to limit its acceptance and utility (Shi and Martin, 2000). These include the following: (1) only a few chromosomes can be investigated within any one assay using visual assessments, (2) the microscope scoring criteria require intensive training and remain subjective, and (3) microscopic scoring is time-consuming limiting the throughput of sperm-FISH analyses. These challenges await the development of automated scoring methods to reduce subjectivity and improve throughput in human as well as rodent sperm-FISH assays (e.g., flow-cytometric analysis or computer-controlled microscopy); and this research is in progress.

Acknowledgments

The authors thank Jack Bishop of NIEHS for his long-standing support and encouragement of rodent sperm-FISH methods. This work was conducted under the auspices of the US DOE by the University of California, LLNL under contract

W-7405-ENG-48 with support from grants NIH/NIEHS ES09117-02 and NIH/NIEHS P42ES04705.

References

Baumgartner A and Schmid TE, Schuetz CG and Adler ID (2001) Detection of aneuploidy in rodent and human sperm by multicolor FISH after chronic exposure to diazepam. *Mutation Research*, **490**(1), 11–19.

Brandriff BF, Meistrich ML, Gordon LA, Carrano AV and Liang JC (1994) Chromosomal damage in sperm of patients surviving Hodgkin's disease following MOPP (nitrogen mustard, vincristine, procarbazine, and prednisone) therapy with and without radiotherapy. *Human Genetics*, **93**(3), 295–299.

De Mas P, Daudin M, Vincent MC, Bourrouillou G, Calvas P, Mieusset R and Bujan L (2001) Increased aneuploidy in spermatozoa from testicular tumour patients after chemotherapy with cisplatin, etoposide and bleomycin. *Human Reproduction*, **16**(6), 1204–1208.

Frias S, Van Hummelen P, Meistrich ML, Lowe XR, Hagemeister FB, Shelby MD, Bishop JB and Wyrobek AJ (2003) NOVP chemotherapy for Hodgkin's disease transiently induces sperm aneuploidies associated with the major clinical aneuploidy syndromes involving chromosomes X, Y, 18, and 21. *Cancer Research*, **63**(1), 44–51.

Hill FS, Marchetti F, Liechty M, Bishop J, Hozier J and Wyrobek AJ (2003) A new FISH assay to simultaneously detect structural and numerical chromosomal abnormalities in mouse sperm. *Molecular Reproduction and Development*, **66**(2), 172–180.

Jacobs PA (1992) The chromosome complement of human gametes. *Oxford Reviews of Reproductive Biology*, **14**, 47–72.

Li X, Zheng LK, Deng LX and Zhang Q (2001) Detection of numerical chromosome aberrations in sperm of workers exposed to benzene series by two-color fluorescence in situ hybridization. *Yi Chuan Xue Bao = Acta Genetica Sinica*, **28**(7), 589–594.

Liu XX, Tang GH, Yuan YX, Deng LX, Zhang Q and Zheng LK (2003) Detection of the frequencies of numerical and structural chromosome aberrations in sperm of benzene series-exposed workers by multi-color fluorescence in situ hybridization. *Yi Chuan Xue Bao = Acta Genetica Sinica*, **30**(12), 1177–1182.

Lowe X, O'Hogan S, Moore D, Bishop J and Wyrobek A (1996) Aneuploid epididymal sperm detected in chromosomally normal and robertsonian translocation-bearing mice using a new three-chromosome FISH method. *Chromosoma*, **105**, 204–210.

Lowe XR, de Stoppelaar JM, Bishop J, Cassel M, Hoebee B, Moore D II and Wyrobek AJ (1998) Epididymal sperm aneuploidies in three strains of rats detected by multicolor fluorescence in situ hybridization. *Environmental and Molecular Mutagenesis*, **31**(2), 125–132.

Martin RH, Ernst S, Rademaker A, Barclay L, Ko E and Summers N (1997) Analysis of human sperm karyotypes in testicular cancer patients before and after chemotherapy. *Cytogenetics and Cell Genetics*, **78**(2), 120–123.

Martin R, Ernst S, Rademaker A, Barclay L, Ko E and Summers N (1999) Analysis of sperm chromosome complements before, during, and after chemotherapy. *Cancer Genetics and Cytogenetics*, **108**(2), 133–136.

McFadden DE and Friedman JM (1997) Chromosome abnormalities in human beings. *Mutation Research*, **396**(1–2), 129–140.

Naccarati A, Zanello A, Landi S, Consigli R and Migliore L (2003) Sperm-FISH analysis and human monitoring: a study on workers occupationally exposed to styrene. *Mutation Research*, **537**(2), 131–140.

Padungtod C, Hassold TJ, Millie E, Ryan LM, Savitz DA, Christiani DC and Xu X (1999) Sperm aneuploidy among Chinese pesticide factory workers: scoring by the FISH method. *American Journal of Industrial Medicine*, **36**(2), 230–238.

Recio R, Robbins WA, Borja-Aburto V, Moran-Martinez J, Froines JR, Hernandez RM and Cebrian ME (2001) Organophosphorous pesticide exposure increases the frequency of sperm sex null aneuploidy. *Environmental Health Perspectives*, **109**(12), 1237–1240.

Robbins WA, Meistrich ML, Moore D, Hagemeister FB, Weier HU, Cassel MJ, Wilson G, Eskenazi B and Wyrobek AJ (1997) Chemotherapy induces transient sex chromosomal and autosomal aneuploidy in human sperm. *Nature Genetics*, **16**, 74–78.

Rubes J, Lowe X, Moore D II, Perreault S, Slott V, Evenson D, Selevan SG and Wyrobek AJ (1998) Smoking cigarettes is associated with increased sperm disomy in teenage men. *Fertility and Sterility*, **70**(4), 715–723.

Rudak E, Jacobs PA and Yanagimachi R (1978) Direct analysis of the chromosome constitution of human spermatozoa. *Nature*, **274**(5674), 911–913.

Schmid TE, Lowe X, Marchetti F, Bishop J, Haseman J and Wyrobek1 AJ *et al*. (2001) Evaluation of inter-scorer and inter-laboratory reliability of the mouse epididymal sperm aneuploidy (m-ESA) assay. *Mutagenesis*, **16**(3), 189–195.

Shi Q, Ko E, Barclay L, Hoang T, Rademaker A and Martin R (2001) Cigarette smoking and aneuploidy in human sperm. *Molecular Reproduction and Development*, **59**(4), 417–421.

Shi Q and Martin RH (2000) Aneuploidy in human sperm: a review of the frequency and distribution of aneuploidy, effects of donor age and lifestyle factors. *Cytogenetics and Cell Genetics*, **90**(3–4), 219–226.

Sloter ED, Lowe X, Moore ID, Nath J and Wyrobek AJ (2000) Multicolor FISH analysis of chromosomal breaks, duplications, deletions, and numerical abnormalities in the sperm of healthy men. *American Journal of Human Genetics*, **67**(4), 862–872.

Smith JL, Garry VF, Rademaker AW and Martin RH (2004) Human sperm aneuploidy after exposure to pesticides. *Molecular Reproduction and Development*, **67**(3), 353–359.

Wyrobek AJ, Ahlborn T, Balhorn R, Stanker L and Pinkel D (1990) Fluorescence in situ hybridization to Y chromosomes in decondensed human sperm nuclei. *Molecular Reproduction and Development*, **27**(3), 200–208.

Xu DX, Zhu QX, Zheng LK, Wang QN, Shen HM, Deng LX and Ong CN (2003) Exposure to acrylonitrile induced DNA strand breakage and sex chromosome aneuploidy in human spermatozoa. *Mutation Research*, **537**(1), 93–100.

Zhao T, Liu XX, He Y, Deng LX and Zheng LK (2004) Detection of numerical aberrations of chromosomes 7 and 8 in sperms of workers exposed to benzene series by two-color fluorescence in situ hybridization. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*, **21**(4), 360–364.

Introductory Review

Imprinting and epigenetic inheritance in human disease

Constantin Polychronakos
McGill University Health Center, Montreal, QC, Canada

1. Introduction

According to the rules of Mendelian inheritance, genetic diseases due to faulty genes located on autosomes (chromosomes other than X or Y) are equally transmitted from the parent of either sex. This is indeed the case for the vast majority of genetic diseases. Why should it matter to the offspring whether a given genetic material comes from the father or the mother? Indeed, once paternal and maternal chromosomes mix upon fertilization, how can the offspring even distinguish which autosomal genes came from which parent?

Actually, in diseases involving the small number of genes subject to the strange phenomenon of parental imprinting, the sex of the transmitting parent does matter. Passing through either the testis or the ovary (the gonads), these genes are "stamped" with an imprint that identifies them as paternal or maternal, respectively. This imprint persists through replication of the DNA as cells derived from the fertilized egg divide to form the entire body and will only be erased and reapplied in the gonads for transmission to the next generation. It usually involves DNA methylation and results in silencing of either the maternal or the paternal copy of the gene involved. Therefore, the level of the protein product of these genes normally corresponds to what comes out of a single gene copy, unlike the case with most other genes, where normal gene dosage corresponds to the product of two copies. Thus, imprinting can be seen as regulating gene dosage, a concept important in understanding its effects on genetic disease.

2. How can imprinting affect disease?

The several ways in which imprinting can play a role in disease can be divided into two broad categories. First, the normal process of imprinting may be disrupted, in which case the offspring has two copies that both behave as if they were maternal, or both as if paternal. This can result in either total lack of expression of the gene (both gene copies silenced) or a double gene dosage (neither copy is silenced). We will see that either of these two situations can result in disease.

Imprinting may also be responsible for disease even when not disrupted. Most genetic diseases are recessive, which means that both copies of the gene involved must be inactivated through mutation for the disease to manifest. If the gene is imprinted, only the single expressed copy need carry the mutation, thus converting what would have been recessive inheritance into dominant. Inheritance does not quite follow the usual autosomal dominant mode, in that (1) the transmitting parent is always the father or always the mother, as the case may be, and (2) the transmitting parent need not be affected, depending on which parent he or she inherited the faulty gene from.

3. The Prader–Willi and Angelman syndromes

The archetypal example of disease due to imprinting is this pair of distinct syndromes each of which can be due to exactly the same mutation, depending on whether it was inherited from, respectively, the father or the mother. In most cases, the mutation is a large DNA deletion on the long arm of chromosome 15 (15q11-13) (Knoll *et al*., 1989). When this deletion is inherited from the father, it results in the Prader–Willi syndrome (PWS). Affected individuals are short, and have symptoms most of which can be attributed to some fault in the development of the central nervous system, specifically the hypothalamus (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1). The most characteristic is an inability to feel satiety after eating, which results in excessive calorie intake and obesity. Their hypothalamus is also unable to stimulate the pituitary gland to make sufficient amounts of gonadotropins, the hormones necessary for pubertal development and sexual function (hypogonadotropic hypogonadism). As newborns, they have markedly decreased muscular tone (floppy babies) that can even be detected by experienced mothers as decreased fetal movements during pregnancy. Some intellectual impairment is common.

Curiously, when inherited from the mother, the exact same chromosome 15 deletion, derived from the same ancestor within an extended family, causes Angelman syndrome (AS) a totally different condition (Knoll *et al*., 1989; Clayton-Smith and Pembrey, 1992). Intellectual impairment in AS is much more severe. Affected individuals have increased muscular tone, jerky, puppet-like movements, a characteristic facial appearance with protruding tongue and bursts of laughter. There is no problem with excessive eating or sexual development (Clayton-Smith and Pembrey, 1992).

About 70% of cases of each syndrome are due to a chromosomal deletion (Knoll *et al*., 1989). The most straightforward explanation is that the deleted region contains (a) gene(s) with exclusive paternal expression, and also gene(s) with exclusive maternal expression, an explanation consistent with the known aggregation of imprinted genes of mixed maternal and paternal expression in the same chromosomal region. Indeed, the *SNRPN*, *IPW*, *Necdin*, and *ZNF127* genes, located in the common PWS/AS region on chr. 15 are paternally expressed, while *UBE3A* shows exclusive maternal expression in the brain. *UBE3A* is almost certainly the gene responsible for AS (Kishino *et al*., 1997), while PWS is more

complex, probably requiring loss of function in more than one of the paternally expressed genes (contiguous-gene syndrome).

The cause of PWS and AS in the 30% of cases with no deletion is a good illustration of the diverse mechanisms by which imprinting can cause disease (Figure 1). Most of such cases of PWS (Nicholls *et al*., 1989) and some of AS (Malcolm *et al*., 1991) have uniparental disomy (UPD) of chr. 15. The term refers to a rare reproductive accident that results in both copies of chr. 15 being derived from the same parent. PWS cases have maternal UPD, designated as $UPD_{mat}[15]$. AS cases, as might be expected, have $UPD_{pat}[15]$. Having two copies of an inactivated gene has the same effect here as a deletion of the sole active copy.

Even more interesting are PWS or AS cases with no deletion and with normal biparental derivation of chr. 15. Methylation analysis of the DNA in such cases shows incorrect imprinting, resulting in both chromosomes behaving as paternal (AS cases) or as maternal (PWS cases) at that locus (Ohta *et al*., 1999). In most cases, the reason is a small DNA deletion that does not affect any of the genes involved but disrupts the imprinting center (IC), a DNA sequence that marks the locus so that the imprinting machinery of the egg or sperm can properly identify it and impart the physical imprint (DNA methylation and, possibly, other less well understood modifications).

Finally, in AS individuals with no evidence of any of the above mechanisms, point mutations in *UBE3A* can cause AS but only if inherited from the mother, thus pinpointing this gene as the one responsible for AS (Kishino *et al*., 1997). No such findings single out one of the paternally expressed genes in PWS, suggesting a contiguous-gene syndrome.

4. Too much of a good thing: transient neonatal diabetes

There is one more imprinting-related, disease-causing mechanism not illustrated in the multiple facets of PWS/AS. Disease can also result from a double dose of a gene. For most genes, dosage need not be regulated precisely. One may have twice as much or half as much of the gene product without consequences (the reason that carriers of recessive disease are healthy). Some genes, however, may cause problems if expressed at higher than normal levels. Such a gene is responsible for transient neonatal diabetes mellitus (TNDM). The pancreas of affected newborns is unable to produce insulin, and death is certain without insulin injections (Temple *et al*., 2000). Characteristically, after a number of weeks or months, the pancreas recovers and normal blood sugars are maintained without treatment. Later in life, a milder form of diabetes may recur that usually does not require insulin treatment (Temple *et al*., 2000).

TNDM may be sporadic or familial. Almost all sporadic cases have $UPD_{pat}[6]$, suggesting either deficiency of a maternally expressed gene on chr. 6 or excess of a paternally expressed one (Temple *et al*., 2000). Not being heritable, UPD cannot explain the familial cases, in which the father is, without exception, the transmitting parent. The mutation inherited from these fathers is a duplication: chr. 6 contains two copies of a small region on its long arm (6q24). This clearly indicates that paternal excess, not maternal deficiency of (an) imprinted gene(s)

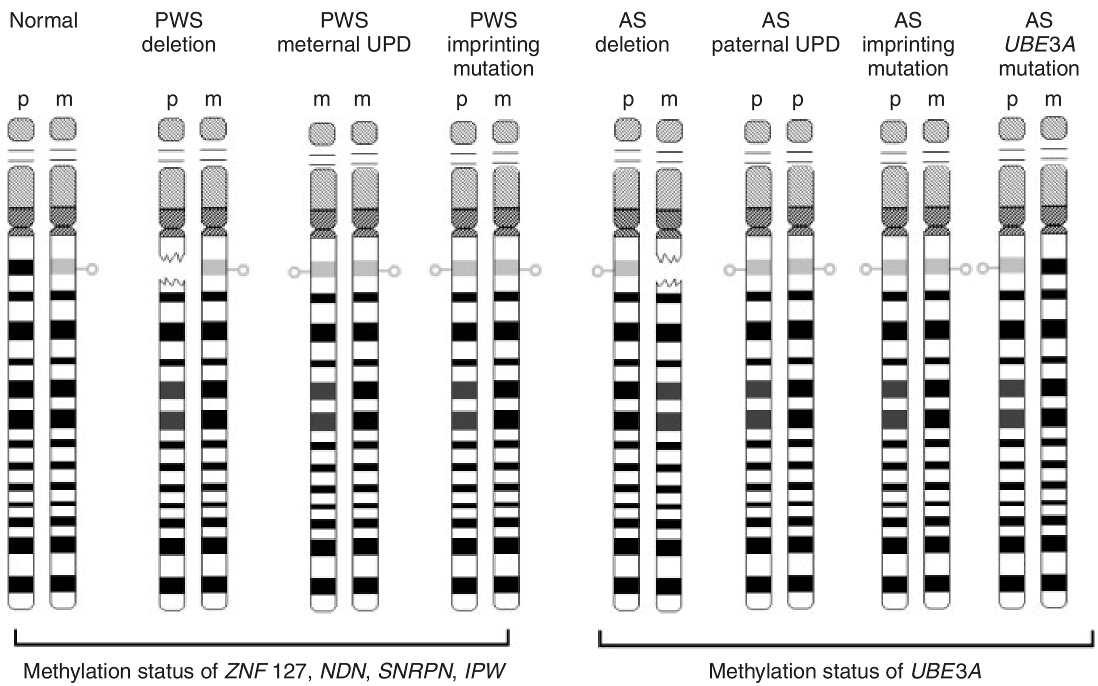


Figure 1 A simple schematic diagram of how parental imprinting is involved in causing Prader–Willi or Angelman syndrome. The various mechanisms, explained in the text, all result in either mutational loss of the only active copy, or the silencing of both, otherwise normal, copies. The light gray area represents the inactivated genes and the attached circle symbolizes DNA methylation

in 6q24 is responsible for TNDM (Temple *et al.*, 2000). Unlike UPD, which usually involves the entire chromosome, duplications are usually quite small. This was used to narrow down the TNDM region to a relatively short interval shared by all the overlapping duplications in different TNDM families (Gardner *et al.*, 2000). Two paternally expressed genes were located in that interval of which the more interesting and better studied is *ZAC*, a DNA-binding protein. *ZAC* is a tumor suppressor (TS), whose role is to prevent uncontrolled cell proliferation. In excess (twice physiologic), it appears to hinder normal proliferation of the insulin-producing β-cells in the developing fetal pancreas, which, for some reason, have a specific sensitivity to this effect.

5. Two other diseases where imprinting plays a role

The Beckwith–Wiedemann syndrome (BWS) is characterized by fetal overgrowth and propensity to embryonal tumors. It is associated with UPD_{pat}[11] and several different chromosomal rearrangements or imprinting defects that essentially make both copies of a region near the tip of the short arm of chr. 11 (11p15.5) behave as if they were paternal. In one subset of BWS cases, the fetal overgrowth is caused by a double dose of the insulin-like growth factor II, a peptide that stimulates fetal growth whose gene (*IGF2*) is normally expressed only from the paternal copy (Giannoukakis *et al.*, 1993). These cases, which also show a gain of DNA methylation in DNA insulator sequences upstream of the nearby and oppositely imprinted *H19* gene, show a high propensity to develop pediatric Wilms tumors. A second, somewhat larger, subset of BWS cases are due to the loss of DNA methylation at another position on chromosome band 11p15.5 (the KvDMR1 imprinting control region), and as a result have reduced expression of a third imprinted gene, *CDKN1C*. These individuals are not highly prone to tumors, but do show severe developmental anomalies, notably abdominal wall defects. This interesting "epigenotype–phenotype correlation" is reviewed in an accompanying article (*see* Article 30, **Beckwith–Wiedemann syndrome**, Volume 1).

Pseudohypoparathyroidism (PHPT Albright's hereditary osteodystrophy) is characterized by resistance to the calcium-regulating parathyroid hormone (PTH) that results in hypocalcemia, despite very high levels of PTH. The defect is on a subunit of a G-protein required for PTH receptor function. Isoforms of the subunit result from transcription of the *GNAS1* gene from alternative promoters, which can be imprinted paternally, maternally, or not at all (*see* Article 31, **Imprinting at the *GNAS* locus and endocrine disease**, Volume 1). Transmission is maternal although the biologically active isoform is not imprinted. We have proposed an explanation for this complex picture in Polychronakos and Kukuvitis (2002), based on the hypothesis that the imprinted isoform is actually a dominant negative inhibitor.

6. Imprinting and cancer

Imprinting may play a role in cancer through silencing of TS genes. Two mutations inactivating both copies of a TS gene in one of the billions of cells in a tissue

gives this one cell a growth advantage that results in a tumor (Knudson'e two-hit hypothesis, *see* Article 65, **Complexity of cancer as a genetic disease**, Volume 2). Often, the second mutation is a large deletion of a chromosomal segment that can be detected as loss of heterozygosity (LOH) at adjacent polymorphic markers. LOH of 11p15.5 (the segment involved in BWS) often occurs in Wilms tumor, a carcinoma of the kidney seen in infants and toddlers, as well as in some other, less common early life cancers that all occur with high frequency in BWS (*see* Article 30, **Beckwith–Wiedemann syndrome**, Volume 1). The deleted segment is always maternal (Schroeder *et al*., 1987), which indicates loss of a maternally expressed TS gene, with imprinting serving as the first "hit". Imprinting of *IGF2* is not universal, as certain tissues of some individuals may express both copies. We first noticed this in white blood cells (Vafiadis *et al*., 1996), and it was subsequently found that this relaxation of imprinting in blood cells correlates with a similar relaxation in the mucosa of the colon (Cui *et al*., 2003). The resulting double dose of the growth stimulator IGF2 may also predispose to colon cancer (Cui *et al*., 2003).

7. Conclusion

Although rare, the diseases caused by imprinting or its disruptions constitute an interesting experiment of nature that has given researchers the opportunity to discover imprinted genes and elucidate some of the mechanisms of this fascinating phenomenon.

References

Clayton-Smith J and Pembrey ME (1992) Angelman syndrome. *Journal of Medical Genetics*, **29**, 412–415.

Cui H, Cruz-Correa M, Giardiello FM, Hutcheon DF, Kafonek DR, Brandenburg S, Wu Y, He X, Powe NR and Feinberg AP (2003) Loss of IGF2 imprinting: A potential marker of colorectal cancer risk. *Science*, **299**(5613), 1753–1755.

Gardner RJ, Mackay DJ, Mungall AJ, Polychronakos C, Siebert R, Shield JP, Temple IK and Robinson DO (2000) An imprinted locus associated with transient neonatal diabetes mellitus. *Human Molecular Genetics*, **9**(4), 589–596.

Giannoukakis N, Deal C, Paquette J, Goodyer CG and Polychronakos C (1993) Parental genomic imprinting of the human IGF2 gene. *Nature Genetics*, **4**(1), 98–101.

Kishino T, Lalande M and Wagstaff J (1997) UBE3A/E6-AP mutations cause Angelman syndrome. *Nature Genetics*, **15**, 70–73.

Knoll JHM, Nicholls RD, Magenis RE, Graham JM Jr, Lalande M and Latt SA (1989) Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *American Journal of Medical Genetics*, **32**, 285–290.

Malcolm S, Clayton-Smith J, Nichols M, Robb S, Webb T, Armour JAL, Jeffreys AJ and Pembrey ME (1991) Uniparental paternal disomy in Angelman's syndrome. *Lancet*, **337**, 694–697.

Nicholls RD, Knoll JHM, Butler MG, Karam S and Lalande M (1989) Genetic imprinting suggested by maternal heterodisomy in non-deletion Prader-Willi syndrome. *Nature*, **342**, 281–285.

Ohta T, Buiting K, Kokkonen H, McCandless S, Heeger S, Leisti H, Driscoll DJ, Cassidy SB, Horsthemke B and Nicholls RD (1999) Molecular mechanism of Angelman syndrome in two

large families involves an imprinting mutation. *American Journal of Human Genetics*, **64**, 385–396.

Polychronakos C and Kukuvitis A (2002) Parental genomic imprinting in endocrinopathies. *European Journal of Endocrinology*, **147**(5), 561–569.

Schroeder WT, Chao LY, Dao DD, Strong LC, Pathak S, Riccardi V, Lewis WH and Saunders GF (1987) Nonrandom loss of maternal chromosome 11 alleles in Wilms tumors. *American Journal of Human Genetics*, **40**(5), 413–420.

Temple IK, Gardner RJ, Mackay DJ, Barber JC, Robinson DO and Shield JP (2000) Transient neonatal diabetes: Widening the understanding of the etiopathogenesis of diabetes. *Diabetes*, **49**(8), 1359–1366.

Vafiadis P, Bennett ST, Colle E, Grabs R, Goodyer CG and Polychronakos C (1996) Imprinted and genotype-specific expression of genes at the IDDM2 locus in pancreas and leucocytes. *Journal of Autoimmunity*, **9**(3), 397–403.

Introductory Review

Regulation of DNA methylation by Dnmt3L

Déborah Bourc'his

U741 INSERM/Paris 7 University, Institut Jacques Monod, Paris, France

Timothy H. Bestor

College of Physicians and Surgeons of Columbia University, New York, NY, US

The mammalian genome contains roughly 3×10^7 CpG dinucleotides, and about 60% of these are methylated at the 5-position of the cytosine. Most 5-methycytosine (m^5C) is in transposable elements and their remnants, and removal of methylation by means of mutations in DNA-methyltransferase genes causes the transcriptional activation of transposons in germ and somatic cells. The small fraction of methylation that is not in transposons is involved in the transcriptional repression of certain imprinted genes and in X chromosome inactivation in females. The promoters of tissue-specific genes are not methylated in a pattern that prevents transcription, and the CpG-rich 5′ domains that contain the promoters of 75% of mammalian genes are normally unmethylated at all developmental stages. The common perception of a role for dynamic methylation changes in the regulation of development has not been confirmed, and no gene has been proven to be activated or repressed by reversible DNA methylation. The major biological functions of DNA methylation are transposon repression, monoallelic expression at certain imprinted loci, and X chromosome inactivation in females. Even modest disruption of genomic methylation patterns is lethal to mammals.

Genomic methylation patterns are established and maintained by DNA (cytosine-5) methyltransferases (DNMTs). As shown in Figure 1, mammals have three enzymatically active DNMTs (DNMT1, DNMT3A, and DNMT3B), a tRNA methyltransferase (RNMT2, formerly DNMT2) that is closely related to DNA methyltransferases in sequence and structure, and DNMT3L, a protein that is related to DNMT3A and DNMT3B in framework sequences but which lacks the catalytic motifs that carry out the transmethylation reaction.

Dnmt3A and Dnmt3B are closely related and have low but approximately equivalent enzymatic activities on unmethylated and hemimethylated substrates (Okano *et al*., 1998). Deletion of Dnmt3A does not cause detectable alteration of genomic methylation patterns in somatic cells of homozygous mice, although adult mice lack germ cells and die of a condition similar to aganglionic megacolon (Okano *et al*., 1999). Mice that lack Dnmt3B die as embryos with demethylation

of major satellite DNA but normally methylated euchromatic DNA; the Dnmt3A-Dnmt3B double mutant dies very early with demethylation of all genomic sequences in a manner similar to that of Dnmt1 null mutants (Okano *et al*., 1999). The rare human genetic disorder ICF syndrome (immunodeficiency, centromere instability, and facial anomalies) is due to recessive loss-of-function mutations in the *DNMT3B* gene (Xu *et al*., 1999). Patients with ICF syndrome fail to methylate classical satellite (also known as satellite 2 and 3) sequences on the juxtacentromeric regions of chromosomes 1, 9, and 16; these demethylated chromosomes gain and lose long arms at a very high rate to produce the multiradiate pinwheel chromosomes unique to this disorder (Jeanpierre *et al*., 1993). Dnmt3B is the only mammalian DNA methyltransferase that has been reported to be affected by histone methylation; there is partial demethylation of major satellite DNA (but no reported chromosome destabilization) in mice that lack the heterochromatic histone H3 K9 methyltransferases Suv39h1 and Suv39h2 (Lehnertz *et al*., 2003). DNA methylation abnormalities have not been reported in mouse embryos that lack the euchromatic histone methyltransferase G9a, nor have there been convincing reports of abnormalities of genomic imprinting, X chromosome inactivation, or transposon silencing in mouse embryos that lack specific histone modifying enzymes (reviewed by Goll and Bestor, 2002, 2005). In mammals, all three processes are dependent on DNA methylation.

Dnmt3A and Dnmt3B are clearly required for the establishment of genomic methylation patterns, but neither enzyme has inherent sequence specificity (reviewed by Goll and Bestor, 2005). The outstanding problem in the mammalian DNA methylation field is undoubtedly the source of the sequence specificity for de novo methylation. Recent studies have shown that unidentified regulatory inputs act through Dnmt3L to guide Dnmt3A and Dnmt3B to target sequences, although the initiating signal or signals remain elusive.

1. Dnmt3L: Expression in male and female germ cells

As shown in Figure 2, expression of full-length Dnmt3L is confined to germ cells. The timing of Dnmt3L expression shows striking sexual dimorphism, it is expressed in males only in perinatal prospermatogonia, which will differentiate into spermatogonia and undergo many mitotic divisions before entering meiosis, but in females expression is limited to growing oocytes, which have completed the pachytene stage of Meiosis I.

1.1. *Functions of Dnmt3L in oogenesis*

As shown in Figure 1, Dnmt3L lacks the conserved motifs that mediate transmethylation but is related to Dnmt3A and Dnmt3B in framework regions (Aapola *et al*., 2000). Dnmt3L also fails to methylate DNA in biochemical tests (data not shown). However, Dnmt3L was of special interest because it is the only DNA-methyltransferase homolog whose expression is confined to germ cells at stages at which de novo methylation occurs (Bourc'his *et al*., 2001).

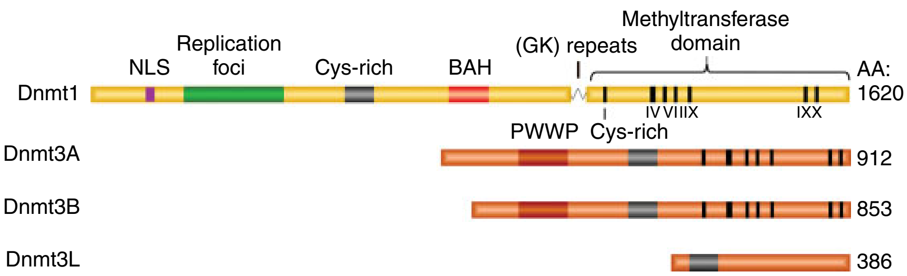


Figure 1 Structure and motif organization of mammalian DNA cytosine methyltransferases. Dnmt3L is expressed specifically in germ cells and is responsible for guiding Dnmt3A and Dnmt3B to target sequences. See Goll and Bestor (2005) for more information

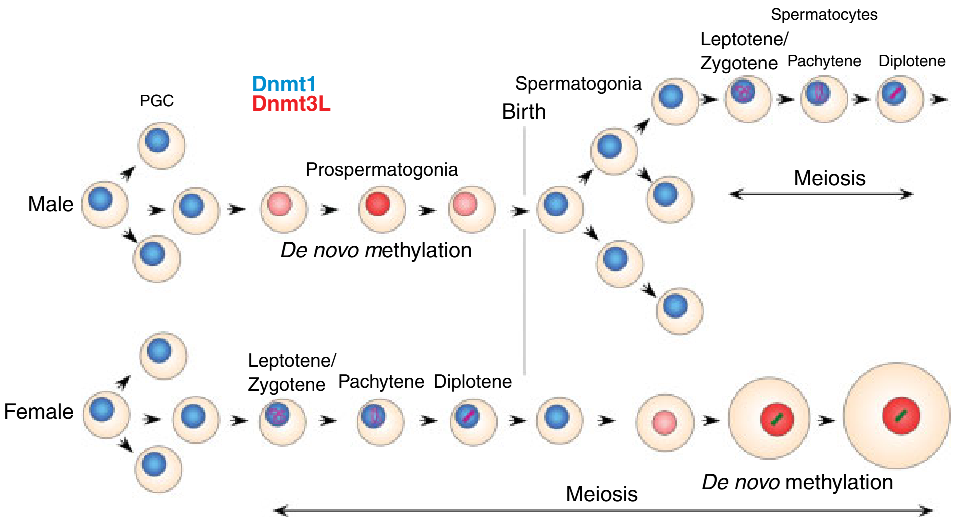


Figure 2 Expression of Dnmt3L and Dnmt1 in male and female germ cells. Intensity of red coloration indicates levels of Dnmt3L, as evaluated by intensity of β-galactosidase expression in animals heterozygous for a β-geo Dnmt3L knock-in allele (Bourc'his *et al*., 2001). Dnmt3L is expressed in premeiotic male germ cells but only in postpachytene (midmeiotic) oocytes in females. Dnmt3L is present only at the stages where genomic imprints are established and, in male germ cells, where transposons undergo de novo DNA methylation

Disruption of the *Dnmt3L* gene by gene targeting in ES cells and insertion of a promoterless β-geo marker into the locus showed that Dnmt3L is expressed in growing oocytes (Bourc'his *et al*., 2001), the stage at which maternal genomic imprints are established (Kono *et al*., 1996). We found that mice homozygous for the disrupted *Dnmt3L* gene were viable and without overt phenotype, although both sexes were sterile. Mutant males were azoospermic, but oogenesis and early development of heterozygous embryos derived from homozygous mutant oocytes was normal; the lethal phenotype was only manifested at e9.5. Such embryos showed signs of nutritional deprivation, and further analysis revealed a failure of chorioallantoic fusion and other dysmorphia of extraembryonic structures (Bourc'his *et al*., 2001). Analysis of expression of imprinted genes showed a

complete loss of imprinting at maternally imprinted loci and a lack of methylation of maternally methylated differentially methylated regions (DMRs). Bisulfite genomic sequencing showed that the imprinting defect was due to a failure to establish genomic imprints in the oocyte, and the normal imprinting of paternally silenced genes in heterozygous offspring of homozygous Dnmt3L-deficient females showed that imprint maintenance in the embryo was normal (Bourc'his *et al*., 2001). This contrasted with the situation in mice that lack Dnmt1o (an oocyte-specific isoform of Dnmt1), when we found that imprint establishment was normal but imprint maintenance in preimplantation embryos was defective (Howell *et al*., 2001). Methylation of sequences other than imprinted regions was normal in heterozygous embryos derived from homozygous *Dnmt3L* mutant oocytes (Bourc'his *et al*., 2001).

1.2. Functions of Dnmt3L in spermatogenesis

In male mice Dnmt3L is expressed at significant levels only in perinatal prospermatogonia, the stage at which paternal genomic imprints are established (Davis *et al*., 1999) and transposons undergo de novo methylation (Walsh *et al*., 1998). Male mice that lack Dnmt3L are outwardly normal except for hypogonadism as adults (Bourc'his *et al*., 2001). The germ cell population is normal at birth, but only the first cohort of germ cells begins meiosis, and none reach the pachytene stage. All mutant meiotic cells show extreme abnormalities of synapsis; grossly abnormal concentrations of synaptonemal complex proteins and nonhomologous synapsis are obvious in nearly all leptotene and zygotene spermatocytes. Adult males are devoid of all germ cells (Bourc'his and Bestor, 2004). This is in striking contrast to Dnmt3L-deficient females, where meiosis and oogenesis are normal and the phenotype is an imprinting defect apparent in heterozygous offspring of homozygous females (Bourc'his *et al*., 2001).

The fact that Dnmt3L-deficient male germ cells show a phenotype only after the stage at which Dnmt3L protein is no longer expressed suggests an epigenetic or gene silencing defect. Homozygous mutant male germ cells were purified and found to suffer global demethylation of the euchromatic genome (Bourc'his and Bestor, 2004).

Transposons contain the large majority of m^5C present in the mammalian genome (Yoder *et al*., 1997), and demethylation of the major transposon classes (IAP elements and LINE-1 elements) was observed in Dnmt3L-deficient male germ cells. However, there was little or no demethylation of major or minor satellite DNA when compared with controls. This indicates that the methylation of heterochromatic satellite DNA is controlled by mechanisms distinct from those that control the methylation of euchromatic sequences. Other data support this conclusion; mutations in the *DNMT3B* gene in humans cause demethylation only of classical satellite (which is analogous to mouse major satellite) in ICF syndrome patients (Jeanpierre *et al*., 1993; Xu *et al*., 1999), and the methylation status of major satellite (but not of other sequences) is affected in mice by loss of the histone methyltransferases Suv39h1 and Suv39h2 (Lehnertz *et al*., 2003), although the magnitude of the effect is much smaller.

The host defense hypothesis predicts that demethylation of transposons in germ cells will cause their transcriptional activation (Bestor, 1990; Yoder *et al*., 1997; Bestor, 2003). Deprivation of Dnmt3L causes mass reanimation of LINE-1 and IAP transcripts observed by blot hybridization and in situ hybridization (Bourc'his and Bestor, 2004). *Dnmt3L* is therefore the first gene shown to be required for the silencing of transposons in germ cells of any organism. It is notable that homozygous loss-of-function mutations in Dnmt1 cause reactivation of IAP transcription in somatic cells (Walsh *et al*., 1998), but LINE-1 elements are not reactivated (Bourc'his and Bestor, 2004); this is likely to reflect the germ cell-specific nature of the promoter in LINE-1 elements (Ostertag *et al*., 2002). LINE-1 elements are thought to be the source of reverse transcriptase for all retroposons, and the coexpression of IAP elements and LINE-1 elements suggests that active transposition of multiple retroposon classes will occur in Dnmt3L-deficient prospermatogonia and spermatogonia.

Dnmt3L is evolving rapidly by comparison to other mammalian DNA-methyltransferase orthologs (Bestor and Bourc'his, 2004). Rapid evolution often reflects an evolutionary chase in which a parasite evolves at a high rate to escape host defense systems, which are then brought under selective pressures for rapid evolution to counter the innovation of the parasite. Transposons represent the most rapidly diverging sequences within host genomes as a result of incessant selective pressures to evade host defense mechanisms. The rate of evolution of DNA methyltransferases is constrained by the requirement to preserve enzymatic activity. This constraint will limit the diversification of these enzymes and favor the evolution of adapter proteins free of this constraint and therefore capable of evolution at a much greater rate. It is suggested that Dnmt3L arose from an enzymatically active Dnmt3 family member in this way, and that the protein now has a regulatory function and cooperates with other proteins in the recognition and de novo methylation of transposons (in male germ cells) and imprinted genes in both germlines.

2. Conclusion

While recent progress has been significant, a number of outstanding questions remain with respect to the biological function of Dnmt3L and the regulation of de novo methylation in germ cells. Why is Dnmt3L required for the methylation of dispersed repeats and largely dispensable for methylation at DMRs in male germ cells, and why is it dispensable for methylation of dispersed repeats but essential for the methylation of single-copy sequences associated with imprinted genes in female germ cells? What signal does Dnmt3L interpret before recruiting Dnmt3A and Dnmt3B; is it recognition of atypical DNA structures such as cruciforms or homology–heterology boundaries in strand-exchange intermediates? Is it particular patterns of histone modifications, or combinations of proteins of the Polycomb and trithorax groups? Are there pathways that regulate DNA methylation in a manner that is independent of Dnmt3L? The identification of Dnmt3L as a regulator of de novo methylation provides an opportunity to discover interacting factors and to

finally identify the cues that designate specific regions of the genome for de novo methylation in germ cells.

References

Aapola U, Kawasaki K, Scott HS, Ollila J, Vihinen M, Heino M, Shintani A, Kawasaki K, Minoshima S, Krohn K *et al*. (2000) Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics*, **65**, 293–298.

Bestor TH (1990) DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **326**, 179–187.

Bestor TH (2003) Cytosine methylation mediates sexual conflict. *Trends in Genetics*, **19**, 185–190.

Bestor TH and Bourc'his D (2004) Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harbor Symposia in Quantitative Biology*, **69**, 381–387.

Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, **431**, 96–99.

Bourc'his D, Xu GL, Lin CS, Bollman B and Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science*, **294**, 2536–2539.

Davis TL, Trasler JM, Moss SB, Yang GJ and Bartolomei MS (1999) Acquisition of the H19 methylation imprint occurs differentially on the parental alleles during spermatogenesis. *Genomics*, **58**, 18–28.

Goll MG and Bestor TH (2002) Histone modification and replacement in chromatin activation. *Genes and Development*, **16**, 1739–1742.

Goll MG and Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annual Reviews of Biochemistry*, **74**, 481–514.

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM and Chaillet JR (2001) Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell*, **104**, 829–838.

Jeanpierre M, Turleau C, Aurias A, Prieur M, Ledeist F, Fischer A and Viegas-Pequignot E (1993) An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome. *Human Molecular Genetics*, **2**, 731–735.

Kono T, Obata Y, Yoshimzu T, Nakahara T and Carroll J (1996) Epigenetic modifications during oocyte growth correlates with extended parthenogenetic development in the mouse. *Nature Genetics*, **13**, 91–94.

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T and Peters AH (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current Biology*, **13**, 1192–1200.

Okano M, Bell DW, Haber D and Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.

Okano M, Xie S and Li E (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics*, **19**, 219–220.

Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL and Kazazian HH Jr (2002) A mouse model of human L1 retrotransposition. *Nature Genetics*, **32**, 655–660.

Walsh CP, Chaillet JR and Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, **20**, 116–117.

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ and Viegas-Pequignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, **402**, 187–191.

Yoder JA, Walsh CP and Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, **13**, 335–340.

Specialist Review

The histone code and epigenetic inheritance

Andrew R. Hoffman and Thanh H. Vu
Stanford University School of Medicine, Palo Alto, CA, USA

1. Histones and chromatin

Histones constitute a family of remarkably conserved proteins that are assembled into nucleosomes that are composed of eight core proteins arrayed as four heterodimers of canonical histones (H2A:H2B, H3:H4, H3:H4, H2A:H2B) around which 147 bp of DNA is wrapped 1.7 times (Luger *et al*., 1997). The nucleosomes, which are connected to each other by 20–60 bp of linker DNA, can then be compacted into a 30-nm fiber to form a more closed and silenced structure (Horn and Peterson, 2002). In conjunction with other proteins involved with DNA replication and packaging, the nucleosomes and DNA are referred to as chromatin. The DNA that is wound about the nucleosome is partially blocked, and only the portion of the DNA that faces away from the histones or is a part of the linker DNA has ready access to DNA polymerases, and regulatory proteins and complexes. Thus, while the histones serve to compact the DNA into the small nuclear space, biochemical machinery must exist to allow DNA to dissociate from the nucleosome on a temporary basis to allow for contact with DNA-modifying protein or RNA complexes and to permit DNA replication and gene transcription. Much of this work may be achieved through nucleosome remodeling complexes, which expose DNA to potential protein binders and partners. By hydrolyzing ATP to provide the energy needed to weaken the DNA:nucleosome interaction and to provide torsion on the DNA helix, DNA loops are opened, and bulges appear, thereby allowing the nucleosomes to slide over the DNA strand. Covalent modifications of the histone octamer represent another mechanism that can alter nucleosome:DNA interactions.

2. The histone code

Histones are substrates for an enormous number of posttranslational enzymatic modifications, including acetylation, methylation, phosphorylation, ubiquitination, and ADP-ribosylation on specific amino acid residues. These modifications occur

in a regulated and nonrandom manner, and the discovery that certain modifications are associated with changes in DNA transcription led to the hypothesis that these modified amino acid residues constituted a distinct "histone code" that regulates gene expression (Turner, 2000; Jenuwein and Allis, 2001). The presence of a code stipulates that each histone molecule possesses a linear array of modifiable amino acid residues and that these modifications enhance biologically productive interactions with chromatin-associated and other nuclear proteins. Higher-order chromatin structure is dependent on the operational sum of these modifications, and the code could be combinatorial. The code is read by nonhistone proteins whose binding to chromatin is determined by specific histone modifications, and the code is deciphered by viewing it as a series of independent modifications arrayed in a linear sequence, or as explicit combinations of modified amino acids. Moreover, the presence of one modification may stimulate further histone modifications on the same (in cis) or on adjoining histones (in trans).

The abundant evidence that acetylated histones were associated with open chromatin, and increased gene transcription was the first suggestion that histone modifications are a molecular signal (Grunstein, 1997). Acetylation of histone 3, lysine 9 (H3K9) was associated with gene expression. The demonstration that histone methylation could lead to gene silencing when it was on H3K9 and to gene expression when histone 3, lysine 4 (H3K4) was methylated demonstrated that the specificity of the code was determined by the particular amino acid residue in conjunction with the precise chemical modification (Noma *et al*., 2001). This detailed specificity was further delineated when it was shown that trimethylation, but not mono- or dimethylation, of H3K4 coded for gene activation (Santos-Rosa *et al*., 2002).

Most histone modifications have been described on the histone tail, where the enzymatic machinery has relatively free access to the relevant amino acid substrates. These modified histones then become adhesive, receptor-like molecules that can attract, recruit, and bind extremely specific protein complexes that can modulate transcription. Acetylated lysines on histone H3 bind a series of proteins containing a bromodomain motif, leading to an increase in gene expression, while H3K9 methylation attracts chromodomain-containing proteins and HP1 (Jenuwein and Allis, 2001), leading to the formation of heterochromatin and gene silencing.

In addition to modifications on the histone tail, the core portion of histones can also undergo posttranslational additions. These core modifications, which regulate nucleosome mobility, can also be governed by the histone code (Cosgrove *et al*., 2004). Since much of the core is associated with tightly wound DNA, it is likely that remodeling complexes must first weaken DNA:nucleosome interactions in order to expose the amino acid substrates to the acetylating, phosphorylating, or methylating enzymes. While the functions of core modifications are less clearly understood, it is clear that their histone code may play a vital transcriptional role, as a stable remodeled state can be generated even with tailless nucleosomes. Covalent histone modifications can also lead to an altered electrochemical "charge patch" that could theoretically modulate histone structure or its binding to DNA independent of any histone signal decoding.

3. Combinatorial code

It is likely that the consequence of any specific histone modification depends upon the modifications of its neighbors. The histone code is combinatorial in that multiple modifications may reside on a single histone molecule, and the set of modifications may be read together as an intelligible message. For example, one modification may lead to the recruitment of enzymatic complexes that promote other nearby histone modifications. In some systems, the combination of acetylated histone 3, lysine 9 (H3K9-Ac) and phosphorylated histone 3, serine 10 (H3 S10-P) synergize to stimulate gene expression, and, conversely, methylated histone 3, lysine 9 (H3K9-Me) may inhibit the establishment of H3 S10-P. In an elegant explication of the control IFN-beta, it was shown that three specific lysine acetylation reactions (H4K8, H3K9, and H3K14) at the promoter region were necessary to recruit two specific bromodomain-containing transcription factors and stimulate gene expression (Agalioti *et al*., 2002). Schreiber and Bernstein have proposed a creative dynamic model of chromatin, comparing the histone code to cellular signaling networks. As in the case with signal transducing receptors, posttranslational modifications of histones leads to the creation of docking sites that can recruit regulatory proteins in high concentrations. In both systems, multiple modifications, occurring in varying regions and in varying chronological sequences, lead to feedback loops, which can provide stability, adaptability, and robustness, thereby promoting an array of consequences under exquisitely ordered control (Schreiber and Bernstein, 2002).

The task of interpreting the histone code is still in its early stages, as it is not yet clear how we have identified the full gamut of covalent modifications and attachments, or how the various combinations of such changes (e.g., H4 acetylation and H3 K9 methylation) on a single nucleosome or on adjacent nucleosomes interact with one another to modify gene expression (Spotswood and Turner, 2002). While histone acetylation and phosphorylation are rapidly reversible modifications, histone methylation at first appeared to be a permanent mark, as no direct demethylases had been discovered. This situation changed very recently, however, with the discovery of an enzyme that can specifically demethylate the stimulatory modification, dimethyl-lysine 4 in histone H3 (Shi *et al*., 2004). Direct enzymatic demethylation of the inhibitory H3 K9 methyl modification has not yet been demonstrated. Moreover, indirect demethylation of CpG can occur through methylcytosine deamination, which leads to a C-T transitional mutation, in conjunction with T:G mismatch repair (Morgan *et al*., 2004), and methylarginine in histone proteins can be converted to citrulline by deimination (Cuthbert *et al*., 2004).

4. Epigenetic code

Epigenetics refers to a mechanism whereby genetic information is heritably passed to the next generation of daughter cells without altering the DNA sequence. Methylation of CpG nucleotides in CpG islands in DNA is generally associated with silencing of gene transcription, and this methylation pattern is passed down to daughter cells after mitosis. Thus, this form of epigenetic gene silencing may play

an important role in cellular differentiation, allowing cells to express a restricted repertoire of genes. Histone modifications as well as DNA methylation, moreover, have the unique ability to be encoded and maintained through cell division as epigenetic memory. The interactions between histone modifications and DNA methylation in the regulation of genomic imprinting and the related phenomenon of X-inactivation have been intensively investigated. For example, methylation of H3K4 is seen on the active X chromosome, while H3K9 methylation is associated with the inactive X chromosome (Boggs *et al*., 2002). In *Saccharomyces*, H3 methylation is dependent upon prior ubiquitination of histone 2B (Dover *et al*., 2002). DNA methylation of *Snrpn* is related to histone H3 and not H4 deacetylation (Gregory *et al*., 2001), while simultaneous H3 and H4 acetylation is associated with transcription of *Igf2/H19* (Grandjean *et al*., 2001). Using compounds such as butyrate and trichostatin A, which act as histone deacetylase inhibitors, we have shown that changing the state of histone acetylation can result in de novo expression of the imprinted allele. Moreover, the enhanced histone acetylation induced by these drugs also leads to decreases in DNA methylation, but drugs that inhibit histone deacetylases (HDACs) often do not restore full expression of previously silenced genes, as some DNA methylation will not be affected by changes in histone acetylation (Hu *et al*., 1998).

DNA methyltransferases (DNMTs) and HDACs can interact in complex and intricate ways to alter chromatin structure (Li, 2002). The primacy of the histone code versus DNA methylation in the establishment of the epigenetic state has been extensively studied. Thus, DNMT1 binds HDACs (Fuks *et al*., 2000), and the process of DNA methylation can lead directly to histone deacetylation, or vice versa. Moreover, a number of methyl-CpG-binding proteins become associated with HDAC-protein complexes, further integrating DNA methylation with changes in histone structure (Jones *et al*., 1998). De novo CpG methylation normally occurs *after* chromatin changes have rendered a gene transcriptionally silenced, suggesting that the major function of DNA methylation is to affect the stable silencing of a gene. Using an ingenious transgenic approach, Cedar's lab demonstrated that DNA methylation induces the deacetylation of H4 and the methylation of H3K9, and inhibits the methylation of H3K4. They conclude that DNA methylation is sufficient to induce a closed chromatin structure (Hashimshony *et al*., 2003). In *Neurospora*, DNA methylation occurs only after histone H3K9 methylation has been achieved (Tamaru and Selker, 2001). It is our view that the minute-by-minute transcriptional regulation of a gene is determined by the rapid and reversible on/off modifications to histone molecules (Spotswood and Turner, 2002), primarily acetylation and phosphorylation reactions. Long-term silencing, on the other hand, is initiated by histone methylation (e.g., H3K9 or H3K27) and is firmly cemented by the relatively irreversible methylation of CpG dinucleotides. In this paradigm, epigenetic regulation stems from dynamic histone modifications, leading to changes in DNA methylation. In the absence of DNA methylation (DNA methyltransferase-I [*Dnmt1*] deficient mice), imprinting is severely disrupted, indicating that histone modifications alone are not sufficient for persistent monoallelic expression (Howell *et al*., 2001).

5. How is the epigenetic code transmitted to daughter cells?

The mechanism for the transmission of DNA methylation after mitosis appears to be straightforward. After the DNA is replicated, one strand of the double helix will contain the original DNA (which is methylated), while the other strand of newly synthesized DNA will contain unmethylated DNA. This hemi-methylated DNA is the preferred substrate for DNMT1, which then transfers methyl groups to all of the hemi-methylated symmetric CpGs in the DNA daughter strand (Bestor, 1992). Replicating the histone code after mitosis is more complicated, however, because the myriad of modifications are regulated by a large array of enzymes and remodeling complexes. After DNA replication, the various sets of modified histones are distributed in a random semiconservative pattern to the newly formed nucleosomes where they are matched with newly synthesized (and presumably unmodified) histones, so that, on average, each set of the histone modifications are present in a density 50% of that seen in the original nucleosome. It has been suggested that chromatin assembly factor (CAF)-I is recruited by proliferating cell nuclear antigen as the DNA is being replicated. CAF1 binds to HP1, which then recruits a set of enzymes, including DNMT1, histone deacetylases, and histone methyltransferases, which can theoretically replicate the histone code (Maison and Almouzni, 2004). However, it is not at all clear how the very specific combination of modifications can be duplicated, since it is not apparent what template the enzymes could utilize. It has been argued that modifications such as H3K9-Me can recruit a H3K9-methyltransferase that can duplicate the modification (or lead to spreading of the modification in cis) (Maison and Almouzni, 2004), and acetylated lysines can recruit bromodomain proteins in complexes that may contain histone acetyltransferase (HAT) activity, but self-duplicating machinery has not been demonstrated for all of the modifications. Therefore, it has been suggested that in some cases, duplication of the code does not occur during replication, but during gene transcription when the nucleosome:DNA interaction is weakened. This would result in two complementary processes for heritable transmission of the code, one that is replication-dependent (RD) and one that is transcription-dependent (TD) (Henikoff *et al*., 2004). The RD method would replicate the histone code via chromodomain-containing proteins, and the TD would operate through proteins with bromodomains. The transcription-coupled nucleosome replacement theory suggests that active nucleosomes are replaced by ATP-dependent remodeling complexes that deposit newly synthesized histones containing the variant histone H3.3, which may alter the interactions of the nucleosome with DNA. H3.3 is deposited during all phases of the cell cycle except S-phase, and it has been shown to replace H3 containing H3K9Me after the induction of transcription. This theory suggests that the *variant* histones, more than a histone code dependent upon posttranslational modifications, determine the transcriptional valence of the chromatin, and the theory also suggests that the histones of all actively transcribed genes will be replaced during one-cell generation. Data to confirm or reject this theory awaits the availability of specific antibodies to variant histones, but it is our view that only H3 histones containing silencing modifications (K9-Me and K27-Me) may be replaced by H3.3 and the activating modifications remain in place. Thus, it is likely that some but not all of the nucleosomes

are replaced. Transcriptionally active chromatin normally harbors no or very low levels of H3K9-Me and H3K27-Me. The procession of transcription may help to "clean out" some of the sporadic silent code (H3K9-Me, H3K27-Me) along the transcriptional passage. It has been shown that read-through transcription can activate repressed chromatin domains marked by H3K27-Me and the Polycomb complex in Drosophila bithorax (Bender and Fitzgerald, 2002). However, as we have noticed in a number of imprinted genes, such as *Gnas*, which contain overlapping sense and antisense transcripts, on each parental chromosome, read -through transcription in a repressed promoter region fails to switch the promoter from a repressed state to an active one (Li *et al*., 2004). On each parental chromosome, the repressed promoters are strictly maintained by histone H3K9 methylation and by DNA methylation. DNA methylation is generally absent in Drosophila, a fact that might partly explain the observed switching from a silenced to an active state by read-through transcription in the fly. It is likely that epigenetic inheritance and maintenance in mammals, at least in imprinted genes, is more strictly governed by both histone methylation and DNA methylation, and that repressed chromatin domains cannot be reversed to an active state by read-through transcription.

6. Epigenetic inheritance and genomic imprinting

While most autosomal genes are expressed in an equivalent manner from both parental alleles, a small number of important genes are expressed only from one specific parental allele. These genes, which are often important in growth, development, and behavior, are said to be imprinted, and enormous interest has been aroused in solving the puzzle of how a cell can discern the parental origin of an allele and how the transcriptional machinery is able to operate on only one of the two alleles. A number of imprinted genes contain reciprocally imprinted antisense transcripts that originate within the gene and are transcribed from the parental allele opposite of that of the sense allele.

7. Epigenetic domains and the fine print of the histone code of *Gnas*

In the formation of heterochromatin, gene silencing enforced by H3K9 methylation leads to the formation of binding sites for proteins like Swi6 (in yeast), which then recruits histone methyltransferases, leading to the spreading of H3K9 Me and further silencing of adjacent nucleosomes. This spreading can be diffused, or it may be interrupted. Ultimately, this progression is blocked by insulators that prevent this repressive spread and allows for the subsequent development of histone modifications that promote gene expression. In some circumstances, a single nucleosome with a specific modification can control gene expression: one nucleosome containing H3K9 inhibits *RB* expression (Nielsen *et al*., 2001). In yet other systems, stimulatory and inhibitory regions of the genome lie side by side,

and the histone code should be able to provide the epigenetic message to regulate gene expression.

In the *Gnas* gene, there are various imprinted transcripts that arise within several thousand bases of one another. On the maternal chromosome, the *Nesp* transcript is transcribed through three silenced domains corresponding to the Nesp antisense (*Nespas*), *Xlαs*, and Exon 1A; conversely, on the paternal allele, the active *Nespas* is transcribed through the silenced *Nesp* region. Since the region between *Nesp* and *Xlαs* is relatively small (∼10 kb) and since the change from active to suppressed regions must involve epigenetically determined boundaries, we studied the sequential changes in the histone code in nucleosomes spanning this gene (Li *et al*., 2004). Surprisingly, the various activating and silencing signals segregated independently on nucleosomes over this area. We found that there were two allelic switch regions (ASR) that marked separate boundaries of epigenetic signaling in the region between the reciprocally imprinted *Nesp* and *Nespas* promoters. ASR1 has a terminal boundary just downstream of *Nesp*, and it is characterized by an activating region that is rich in histone acetylation and H3K4 methylation. Further downstream, the nucleosomes assume another configuration (ASR2), which is characterized by silencing epigenetic changes (H3K9Me and DNA methylation). A relative boundary element may exist at this interface between the two ASRs, although it is possible that there exist nucleosomes that harbor both activating as well as silencing signals. It seems likely that there are two independent regulatory mechanisms at play, one governing the activating modifications and the one silencing changes. The variety of modifications seen throughout the *Gnas* chromatin raises the question of where the histone code is read. Does the epigenetic transcriptional machinery operate only at promoter regions or is there an important histone code that resides over most of the gene or even the entire imprinted domain?

8. Histone code overrides the DNA imprint

The relative roles of DNA methylation and the histone code in the regulation of genomic imprinting has been a topic of great interest. While many studies have suggested that DNA methylation is the primordial imprint that determines allelic expression throughout development, several recent studies have shown that the histone modifications can override DNA methylation in the expression of imprinted genes. The imprinting of the mouse *Igf2r* gene in peripheral issues is regulated through a differentially methylated DNA region in the first intron and by the presence of a nontranslated RNA antisense, *Air* (Sleutels *et al*., 2002). In the central nervous system, however, *Igf2r* is biallelically expressed, even though *Air* is expressed (Hu *et al*., 1999). In human *IGF2R*, expression is always biallelic; even though the intronic DMR (differentially methylated region) is present in the human gene, no antisense RNA is transcribed (Vu *et al*., 2004). This discrepancy was difficult to reconcile with a general antisense model of imprinting until it was shown that the tissue-specific and species-specific imprinting status of *Igf2r* could be predicted by examining the histone code, rather than the DNA methylation or RNA antisense (Vu *et al*., 2004). In all cases, allele-specific expression is

determined by the presence of H3K4-Me (stimulatory) or H3K9-Me (inhibitory) at the promoter region.

Another example of the dominance of the histone code over DNA methylation in determining gene expression is found in the regulation of a cluster of genes on mouse chromosome 7. The imprinting center 2 (IC2) on mouse chromosome 7 controls the regulation of several imprinted genes, where one germ-line DMR and two somatically acquired DMRs have been recognized in this cluster. Imprinting in embryonic stem (ES) cells and in extraembryonic tissues (placenta) extends beyond the imprinting domain in embryos to ∼700-kb region flanking the IC2. The imprinting of these distal imprinted genes, in contrast to the proximal genes, does not depend upon DNA methylation, as the distal genes remain imprinted in the *Dnmt1* -/- placenta. The imprinting of both distal and proximal genes, however, is consistent with the association of H3K9Me and H3K27Me methylation in the silenced allele, and H3KAc and H3K4Me in the expressed allele. Allele-specific histone modifications and imprinting are disrupted by IC2 deletion (Lewis *et al*., 2004). Feil, Reik, and colleagues proposed that the histone code is the primordial regulator of imprinting, and remains the major factor allowing specific allelic expression in ES cells and in the placenta. In later development, in differentiated ES cells and in embryos, stable imprinting requires the recruitment of DNA methylation to make the imprinting permanent.

9. Conclusion

The multitude of posttranslational histone modifications constitute a code that can be read by the cell's transcriptional machinery to alter gene expression. Numerous questions concerning the histone code remain to be answered however. In particular, the precise mechanisms underlying the heritability of the various amino acid modifications are yet to be convincingly elucidated. Are there developmental-specific or tissue-specific aspects to the code? Is the code read the same way for all genes? Where do we look for the code: should we concentrate on nucleosomes associated with the promoter regions of the gene only, or must we examine all of the nucleosomes associated with a gene to understand how the histone code regulates gene expression?

Related articles

Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1, Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1, Article 31, **Imprinting at the *GNAS* locus and endocrine disease**, Volume 1, Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1, Article 37, **Evolution of genomic imprinting in mammals**, Volume 1 and Article 38, **Rapidly evolving imprinted loci**, Volume 1

References

Agalioti T, Chen G and Thanos D (2002) Deciphering the transcriptional histone acetylation code for a human gene. *Cell*, **111**(3), 381–392.

Bender W and Fitzgerald DP (2002) Transcription activates repressed domains in the Drosophila bithorax complex. *Development*, **129**(21), 4923–4930.

Bestor TH (1992) Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *The EMBO Journal*, **11**(7), 2611–2617.

Boggs BA, Cheung P, Heard E, Spector DL, Chinault AC and Allis CD (2002) Differentially methylated forms of histone H3 show unique association patterns with inactive human X chromosomes. *Nature Genetics*, **30**(1), 73–76.

Cosgrove MS, Boeke JD and Wolberger C (2004) Regulated nucleosome mobility and the histone code. *Nature Structural and Molecular Biology*, **11**(11), 1037–1043.

Cuthbert GL, Daujat S, Snowden AW, Erdjument-Bromage H, Hagiwara T, Yamada M, Schneider R, Gregory PD, Tempst P, Bannister AJ, *et al*. (2004) Histone deimination antagonizes arginine methylation. *Cell*, **118**(5), 545–553.

Dover J, Schneider J, Tawiah-Boateng MA, Wood A, Dean K, Johnston M and Shilatifard A (2002) Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *The Journal of Biological Chemistry*, **277**(32), 28368–28371.

Fuks F, Burgers WA, Brehm A, Hughes-Davies L and Kouzarides T (2000) DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nature Genetics*, **24**(1), 88–91.

Grandjean V, O'Neill L, Sado T, Turner B and Ferguson-Smith A (2001) Relationship between DNA methylation, histone H4 acetylation and gene expression in the mouse imprinted Igf2-H19 domain. *FEBS Letters*, **488**(3), 165–169.

Gregory RI, Randall TE, Johnson CA, Khosla S, Hatada I, O'Neill LP, Turner BM and Feil R (2001) DNA methylation is linked to deacetylation of histone H3, but not H4, on the imprinted genes Snrpn and U2af1-rs1. *Molecular and Cellular Biology*, **21**(16), 5426–5436.

Grunstein M (1997) Histone acetylation in chromatin structure and transcription. *Nature*, **389**(6649), 349–352.

Hashimshony T, Zhang J, Keshet I, Bustin M and Cedar H (2003) The role of DNA methylation in setting up chromatin structure during development. *Nature Genetics*, **34**(2), 187–192.

Henikoff S, Furuyama T and Ahmad K (2004) Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics*, **20**(7), 320–326.

Horn PJ and Peterson CL (2002) Molecular biology. Chromatin higher order folding–wrapping up transcription. *Science*, **297**(5588), 1824–1827.

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM and Chaillet JR (2001) Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell*, **104**(6), 829–838.

Hu JF, Balaguru KA, Ivaturi RD, Oruganti H, Li T, Nguyen BT, Vu TH and Hoffman AR (1999) Lack of reciprocal genomic imprinting of sense and antisense RNA of mouse insulin-like growth factor II receptor in the central nervous system. *Biochemical and Biophysical Research Communications*, **257**(2), 604–608.

Hu JF, Oruganti H, Vu TH and Hoffman AR (1998) The role of histone acetylation in the allelic expression of the imprinted human insulin-like growth factor II gene. *Biochemical and Biophysical Research Communications*, **251**(2), 403–408.

Jenuwein T and Allis CD (2001) Translating the histone code. *Science*, **293**(5532), 1074–1080.

Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J and Wolffe AP (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, **19**(2), 187–191.

Lewis A, Mitsuya K, Umlauf D, Smith P, Dean W, Walter J, Higgins M, Feil R and Reik W (2004) Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. *Nature Genetics*, **36**(12), 1291–1295.

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, **3**(9), 662–673.

Li T, Vu TH, Ulaner GA, Yang Y, Hu JF and Hoffman AR (2004) Activating and silencing histone modifications form independent allelic switch regions in the imprinted Gnas gene. *Human Molecular Genetics*, **13**(7), 741–750.

Luger K, Mader AW, Richmond RK, Sargent DF and Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**(6648), 251–260.

Maison C and Almouzni G (2004) HP1 and the dynamics of heterochromatin maintenance. *Nature Reviews Molecular Cell Biology*, **5**(4), 296–304.

Morgan HD, Dean W, Coker HA, Reik W and Petersen-Mahrt SK (2004) Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: IMPLICATIONS FOR EPIGENETIC REPROGRAMMING. *The Journal of Biological Chemistry*, **279**(50), 52353–52360.

Nielsen SJ, Schneider R, Bauer UM, Bannister AJ, Morrison A, O'Carroll D, Firestein R, Cleary M, Jenuwein T, Herrera RE, *et al*. (2001) Rb targets histone H3 methylation and HP1 to promoters. *Nature*, **412**(6846), 561–565.

Noma K, Allis CD and Grewal SI (2001) Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. *Science*, **293**(5532), 1150–1155.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J and Kouzarides T (2002) Active genes are tri-methylated at K4 of histone H3. *Nature*, **419**(6905), 407–411.

Schreiber SL and Bernstein BE (2002) Signaling network model of chromatin. *Cell*, **111**(6), 771–778.

Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA and Casero RA (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, **119**(7), 941–953.

Sleutels F, Zwart R and Barlow DP (2002) The noncoding air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**(6873), 810–813.

Spotswood HT and Turner BM (2002) An increasingly complex code. *The Journal of Clinical Investigation*, **110**(5), 577–582.

Tamaru H and Selker EU (2001) A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature*, **414**(6861), 277–283.

Turner BM (2000) Histone acetylation and an epigenetic code. *Bioessays*, **22**(9), 836–845.

Vu TH, Li T and Hoffman AR (2004) Promoter-restricted histone code, not the differentially methylated DNA regions or antisense transcripts, marks the imprinting status of IGF2R in human and mouse. *Human Molecular Genetics*, **13**(19), 2233–2245.

Specialist Review

Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes

Mellissa R. W. Mann
The University of Western Ontario, London, ON, Canada

1. Introduction

Twenty years ago, the developmental fate of uniparental embryos was described in mammals. Elegant nuclear transplantation studies in the mouse determined that both a maternal and a paternal genome are required to complete normal development (McGrath and Solter, 1984; Barton *et al*., 1984). Parthenogenetic and gynogenetic embryos possess only maternally derived chromosomes. The most advanced of these embryos fail to proceed beyond the early limb-bud stage. Extraembryonic tissues are poorly developed. Conversely, androgenetic embryos possess only paternally derived genomes. These embryos rarely advance to the unturned, head-fold stage, and the embryo proper shows poor development. These phenotypes suggest that genes required for embryonic development are expressed from the maternal genome, while genes required for extraembryonic development are transcribed from the paternal genome.

Further analysis indicates that this is a somewhat simplified view. Examination of uniparental embryos and aggregation chimeras revealed phenotypic abnormalities in both embryonic and extraembryonic lineages (Tables 1 and 2). Developmental differences possibly lie in the ability of uniparental cells to proliferate and differentiate. Parthenotes are characterized by a failure to maintain undifferentiated stem cell populations; terminally differentiated cells are overabundant (Sturm *et al*., 1994; Newman-Smith and Werb, 1995). Androgenetic chimeras display an increased rate of cell proliferation (Fundele *et al*., 1995). Thus, the original observation of an antipodal effect of androgenetic and parthenogenetic cells on growth and development is supported.

Noncomplementation of parental genomes indicates that transcriptional regulation of specific genes is dependent on germline origin. Two epigenetic phenomena direct monoallelic expression based on parental origin in the mouse, imprinted X chromosome inactivation (XCI), and genomic imprinting.

Table 1 Phenotypic characteristics of uniparental embryos

Parthenogenetic embryos	Androgenetic embryos
Growth retarded	Early embryonic growth retardation
Few proliferating cells	Trophoblast hyperplasia
Overabundant terminally differentiated cells	Hyperproliferation trophoblast giant cells
Degeneration of polar trophectoderm	Hyperproliferation chorion
Reduced endoreduplication in giant cells	Poor infiltration of chorion by allantois
Lack ectoplacental cone and extraembryonic ectoderm	Poor infiltration chorionic ectoderm into giant cell layer
Thickened extracellular matrix	Distension of pericardial cavity
Chorio-allantoic failure	Weight increase
Enlarged or bulbous allantois	Increased somite number
Poor yolk sac vasculature	Increase in anterior–posterior length
Mesoderm absent/disorganized	
Neural tube defects	
Abnormal, small somites	
Reduced somite number	
Thinner myocardial layer in heart	

Sturm *et al*. (1994) and Obata *et al*. (2000).

Table 2 Phenotypic characteristics of uniparental cells in aggregation chimeras

Parthenogenetic chimeras	Androgenetic chimeras
Growth retarded	Weight/size increase
Poor yolk sac vasculature	Increase in anterior–posterior length
Chorio-allantoic failure	Craniofacial abnormalities
Swollen or bulbous allantois	Sternum shortened/compressed
Defective labyrinthine development	Vertebrae shortened/thickened
Stunted/reduced somite number	Scoliosis
Thinner myocardial layer in heart	Rib cartilage hyperplasia
	Hypo-ossification ribs and skull
	Enlarged, distorted, and fused ribs
	Shortened limbs
	Enlarged and/or fused digits
	Postaxial polydactyly
	Enlarged and disorganized heart
	Umbilical hernia
	Abdominal swelling/liver enlargement
	Abdominal muscle wall defects
	Lack mature brown fat
	Eyelid fusion failure

Spindle *et al*. (1996), Barton *et al*. (1991), Fundele *et al*. (1995) and McLaughlin *et al*. (1997).

2. Imprinted XCI in extraembryonic development

Dosage compensation of X-linked genes is accomplished by inactivation of one X chromosome in female mammals (*see* Article 40, **Spreading of X-chromosome inactivation**, Volume 1 and Article 41, **Initiation of X-chromosome inactivation**, Volume 1). In rodents, extraembryonic tissues undergo preferential inactivation of

the paternal X chromosome (X^P). X-inactivation is regulated in cis by the *Xist* noncoding RNA and its antisense transcript, *Tsix* (Verona *et al*., 2003; Takagi, 2003). *Xist* is transcribed from the paternal X chromosome during preimplantation development and associates with the X chromosome that is to be inactivated, while *Tsix* is transcribed from the maternal X (X^M).

As parthenogenetic embryos possess only maternally derived chromosomes, developmental abnormalities could result from aberrant X-linked gene dosage. In fact, parthenotes fail to undergo XCI; both maternal X chromosomes repress *Xist* and remain active (Figure 1) (Takagi, 2003). If the phenotype associated with parthenogenesis is due only to the inability to inactivate a second maternal X chromosome, then X^MO parthenotes should not suffer the same fate, as X chromosome dosage should be normal. X^MO parthenogenetic embryos display the same compromised development as X^MX^M parthenotes (Mann and Lovell-Badge, 1987), indicating that developmental defects cannot be attributed solely to excessive X-linked gene expression; lack of paternally transcribed, imprinted genes must also contribute to parthenote degeneration.

Conversely, if parthenogenetic failure is due solely to a supernumerary maternal X, embryos with maternal X disomy should be equivalent to parthenotes. $X^MX^MX^P$ and X^MX^MY embryos repress *Xist* and possess two active X chromosomes (Figure 1) (Takagi, 2003). Phenotypically, they bear a striking resemblance to parthenotes. Aggregation of maternal X disomic embryos with tetraploid embryos, which contribute functional trophectoderm, rescues this phenotype. Thus, the

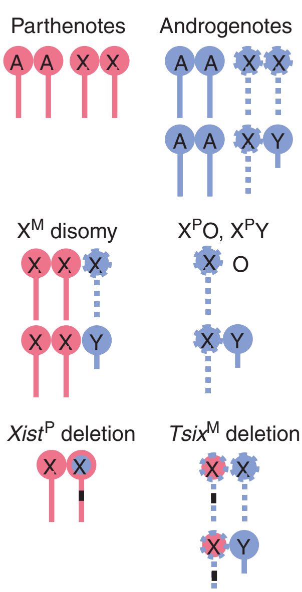


Figure 1 Imprinted XCI in various mouse embryos. Red and blue are maternal (M) and paternal (P) chromosomes, respectively (autosomes: A, X chromosome: X). Solid lines are active and dashed lines are inactive Xs. Black bar indicates deletion

inability to inactivate an extra maternal X chromosome contributes to early lethality of parthenogenetic embryos by impairing extraembryonic development.

Fatality induced by two active X chromosomes was unequivocally demonstrated by engineering mice with a paternally inherited *Xist* deletion. Loss of *Xist* expression results in the paternal X chromosome adopting a maternal epigenotype (Marahrens *et al*., 1997). Thus, neither the maternal nor the paternal X chromosome is inactivated (Figure 1). Mutant embryos display a remarkably similar phenotype to parthenotes. Mice that are $X^{P}O$ and carry the *Xist* deletion are normal, demonstrating that it is a second active X chromosome that is responsible for defective trophoblast development.

Similar to parthenotes, aberrant X-linked gene expression may factor into the developmental fate of androgenetic embryos. During preimplantation, androgenotes display ectopic *Xist* localization from all X^{P} chromosomes, indicative of imprinted XCI (Figure 1) (Takagi, 2003). If the phenotype of androgenetic embryos is due solely to functional nullisomy for the X chromosome, then $X^{P}O$ and $X^{P}Y$ embryos should display similar developmental sequelae to androgenotes. These embryos exhibit decreased trophoblast differentiation, developmental delay, and a small ectoplacental cone (Jamieson *et al*., 1998); their less severe phenotype indicates that lack of expression from maternally transcribed, imprinted genes contributes to androgenetic demise. Reactivation of genes distally located from the X-inactivation center or the X chromosome itself may also ameliorate the phenotype.

One possible explanation for ectopic *Xist* expression is loss of *Tsix* expression. Maternal deletion of *Tsix* leads to derepression of the silent, maternal *Xist* allele (Lee, 2000; Sado *et al*., 2001), and the maternal X chromosome takes on a paternal epigenotype (Figure 1). *Tsix* mutants are characterized by severe embryonic losses and display features common to androgenetic embryos. Early postimplantation, *Tsix* mutants are developmentally retarded. The ectoplacental cone fails to expand into decidua, and allantois and chorion are disorganized (Sado *et al*., 2001). One significant difference between androgenotes and *Tsix* deficient mice is that a small percentage of the latter survive, albeit with growth retardation (Lee, 2000; Sado *et al*., 2001), indicating that aberrant XCI is not the only determinant in defective extraembryonic development and androgenetic lethality. In summary, extraembryonic tissues are susceptible to perturbation in X chromosome dosage. However, imprinted XCI alone cannot account for the noncomplementarity of the parental genomes in placental development; genomic imprinting must also play a role.

3. Genomic imprinting during embryogenesis

Genomic imprinting is an epigenetic transcriptional regulatory mechanism that is established in the parental gametes and is manifested as parental-restricted expression in developing offspring (Verona *et al*., 2003). Biological functions for imprinted genes come from studies involving mice with uniparental disomy/duplications, and with specific gene deficiencies. Uniparental duplication is a condition in which an individual inherits two copies of a chromosomal region from one parent and no copy from the other parent. An imprinted phenotype may result from overexpression and/or loss of expression of genes that exhibit

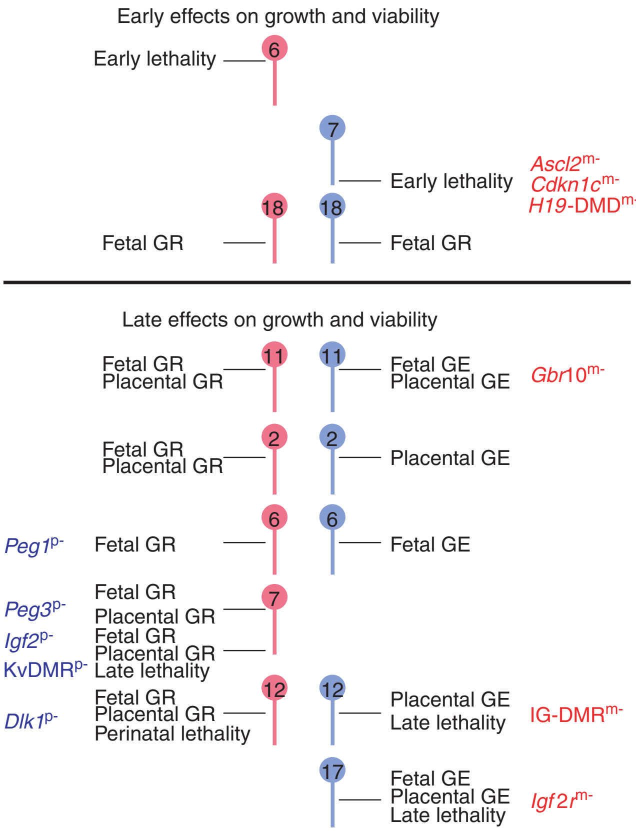


Figure 2 Effects of uniparental duplications on fetal and placental development. Only chromosomal region with imprinting effects on embryogenesis are shown (Beechey *et al*., 2003). Maternal chromosomes/genes are in red, while paternal chromosomes/genes are in blue. GR: growth retardation, GE: growth enhancement. Imprinted genes and imprinting control regions contributing to growth and viability, as determined by targeted mutation, are indicated (m- maternal deletion, p- paternal deletion) (Reproduced by permission of MRC Mammalian Genetics Unit, Harwell, Oxfordshire, Imprinting. Website **http://www.mgu.har.ac.uk/research/imprinting**)

parental-origin-dependent transcription. Several chromosomal regions affect embryonic development and viability when present from only one parent (Figure 2) (Beechey *et al*., 2003). Note that fetal and placental growth effects may occur independently and that not every chromosomal region produces a reciprocal effect when the opposite duplication is considered.

If a specific chromosomal region is responsible for the defective development of uniparental embryos, early lethality should be observed upon uniparental duplication. Only two chromosomal regions are associated with early lethality. Imprinted genes residing on proximal 6 may contribute to parthenogenetic failure (Figure 2). Little is known about maternal duplication for proximal 6 (MatDp(prox6)) except that it is embryonic-lethal prior to day 11.0 of gestation. Genes located on distal 7 may affect the survival of androgenotes as paternal

duplication of distal 7 (PatDp(dist7)) is embryonic-lethal around day 9.5 (Beechey *et al*., 2003). These embryos lack spongiotrophoblast and possess a thickened giant cell layer (McLaughlin *et al*., 1996). This phenotype is very similar to that of *achaete-scute homologue 2* (*Ascl2*) maternally deficient embryos (Guillemot *et al*., 1995). Mutant placentas display an increase in giant cells and a deficiency of proliferating spongiotrophoblast that leads to placental failure and lethality.

Three other imprinted genes in the region may also play a role in androgenetic placental development, *pleckstrin-homology domain A2* (*Phlda2/Ipl/Tssc3*), *cyclin-dependent kinase inhibitor 1C* (*Cdkn1c*), and *insulin-like growth factor 2* (*Igf2*). *Phlda2* and *Cdkn1c* are maternally transcribed genes, while *Igf2* is expressed from the paternal allele (Verona *et al*., 2003). Maternal deletion of *Phlda2* leads to placental overgrowth due to expansion of spongiotrophoblast and an increase in glycogen cells (Frank *et al*., 2002). Maternal deletion of *Cdkn1c* results in an increase in proliferating cells and placentomegaly (Takahashi *et al*., 2000). Lastly, deletion of the *H19* gene including the differentially methylated domain (DMD) results in biallelic expression of the adjacent *Igf2* gene that in turn produces somatic and placental overgrowth (Eggenschwiler *et al*., 1997).

Cdkn1c and *Igf2* have been suggested to have opposing effects on cell proliferation (Caspary *et al*., 1999). Double mutants of *Cdkn1c* and the *H19* DMD display exacerbation of the placental phenotype with overproliferation of trophoblast cells and disruptions in placental architecture, due to excess *Igf2* in the absence of *Cdkn1c* (Caspary *et al*., 1999). Thus, at least three genes on distal chromosome 7, *Acsl2*, *Cdkn1c*, and *Igf2*, are involved in cell proliferation and differentiation. These genes, as well as *Phlda2*, may play a pivotal role in aberrant proliferation of androgenetic cells and placental overgrowth. The lethality associated with PatDp(dist 7), however, occurs at a more advanced stage than the majority of androgenetic embryos, indicating the involvement of imprinted genes elsewhere in the genome.

In addition to lethality and placental structure defects, parthenogenetic and androgenetic cells exert parental-origin effects on growth and development of the embryo proper. Several chromosomal regions are associated with intrauterine growth retardation when maternally duplicated or with intrauterine growth enhancement when paternally duplicated (Figure 2). Targeted deletion studies have identified specific genes that contribute to these growth effects. Loss of maternal-specific expression of *growth factor receptor bound protein 10*, and of *insulin-like growth factor 2 receptor* (*Igf2r*) may account for the fetal and placental growth enhancement of PatDp(prox 11) and PatDp(prox 17) phenotypes, respectively (Beechey *et al*., 2003; Cattanach *et al*., 2004). Deficiency for *paternally expressed gene 3* (*Peg3*) and *Peg1* may be a factor in MatDp(prox 7) fetal and placental growth retardation and MatDp(subprox 6) fetal growth retardation, respectively (Beechey *et al*., 2003; Cattanach *et al*., 2004). In humans, disomy for maternal chromosome 7, which contains the orthologous *Peg1* imprinting domain, results in the development of Silver–Russell Syndrome (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1), characterized by pre- and postnatal growth retardation, and a small, triangular face (Hitchins *et al*., 2001).

With respect to specific developmental abnormalities, genes within the *gene-trap locus 2* imprinted domain on distal chromosome 12 are likely contributors

to skeletal defects and liver enlargement in androgenotes. Paternal disomy for chromosome 12 causes placental growth enhancement, skeletal muscle enhancement, protruding thorax, abdominal extension, enlarged liver, costal cartilage defects, and hypo-ossification (Georgiades *et al*., 2000). In humans, patients with paternal disomy for chromosome 14 display similar phenotypic characteristics (Kurosawa *et al*., 2002). Uniparental duplications in mice represent suitable models for investigating human imprinting disorders originating from uniparental disomy (*see* Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1).

Mouse embryos chimeric for PatDp(dist 7) share a large number of developmental anomalies produced by androgenesis (McLaughlin *et al*., 1997). These same pathogenic features are present in Beckwith–Weidemann Syndrome (BWS), an overgrowth disorder in humans (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1 and Article 30, **Beckwith–Wiedemann syndrome**, Volume 1) involving the orthologous imprinting region (Eggenschwiler *et al*., 1997). Double mutants for *H19* and *Igf2r* and for *H19* and *Cdkn1c* in mice can recapitulate the majority of symptoms seen in BWS patients (Eggenschwiler *et al*., 1997; Caspary *et al*., 1999).

Collectively, these data signify that genes within multiple imprinted domains must act synergistically during embryonic development. There is experimental precedence for this. Individually, MatDp(prox 2) and PatDp(prox 11) affect growth but not viability. However, when inherited together, the combination is fatal (Cattanach *et al*., 2004). Additionally, mutations in the maternal *H19* DMD on distal 7 and the maternal *Igf2r* allele on proximal 17 produce an increased frequency and earlier onset of lethality (Eggenschwiler *et al*., 1997), demonstrating the combinatorial effect of imprinting lesions.

Further evidence for synergism comes from mutations of epigenetic regulators that result in loss of imprinting at multiple loci. *DNA methyltransferase 1* (*Dnmt1*) mutants are growth retarded, and die around embryonic day 9.0 (Lei *et al*., 1996). Their phenotype is reminiscent of parthenogenetic embryos and may arise from the requirement for *Dnmt1* in cell proliferation.

Conversely, *Dnmt3a*-, *Dnmt3a3b*-, and *Dnmt3L*-deficient embryos (*see* Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1) display a hyperproliferation phenotype similar to androgenetic fetuses, characterized by pericardial distension, neural tube defects, chorionic-allantoic fusion failure, thickened chorion, and hyperproliferation of secondary trophoblastic giant cells and yolk sac endoderm (Bourc'his *et al*., 2001; Hata *et al*., 2002; Kaneda *et al*., 2004). Defects in these embryos are due to loss of maternal imprints in oocytes. Interestingly, a recessive human disorder, familial biparental complete hydatidiform mole (phenotypically identical to androgenetic moles) occurs from a failure to acquire maternal-specific imprints (El-Maarri *et al*., 2003).

Bypassing in toto, a double dose of maternal imprinting marks greatly improves development. Gynogenotes generated by nuclear transfer of one nucleus that had and one nucleus that had not been maternally reprogrammed exhibited expression of genes that are normally transcribed from the paternal genome (Kono *et al*., 2004). However, development was highly variable, a factor likely dependent on the epigenetic state of each immature donor nucleus. Thus, the low rates of

viability further substantiate the requirement for a maternal and paternal genome in mammalian development and point to the complex etiology of parthenogenesis and androgenesis.

In conclusion, poor embryonic and extraembryonic development of uniparental embryos is attributed to misregulation of multiple genes governed by genomic imprinting and imprinted XCI. The most important of these genes likely have roles in cell proliferation and differentiation. Parthenogenetic and androgenetic embryos represent good models for investigating genomic imprinting and imprinted XCI. Further studies could provide insight into critical epigenetic events that occur during mammalian embryogenesis, including the stability of imprinted XCI and genomic imprinting in advanced stage uniparental embryos. Furthermore, as androgenetic embryos possess only paternally derived chromosomes, these embryos may be advantageous for determining the hotly contested question of whether X^P is reactivated in the zygote. Finally, given the similarity of developmental defects, investigation of uniparental embryos will be complementary to studies involving interspecific hybrids, somatic cell nuclear transfer embryos (*see* Article 34, **Epigenetics and imprint resetting in cloned animals**, Volume 1), and large offspring syndrome (*see* Article 35, **Imprinted QTL in farm animals: a fortuity or a common phenomenon?**, Volume 1) that together will further our understanding of epigenetics and mammalian embryogenesis.

Acknowledgments

I thank Tamara Davis and Raluca Verona for critical reading of this manuscript.

References

Barton SC, Ferguson-Smith AC, Fundele R and Surani MA (1991) Influence of paternally imprinted genes on development. *Development*, **113**, 679–687.

Barton SC, Surani MAH and Norris ML (1984) Role of paternal and maternal genomes in mouse development. *Nature*, **311**, 374–376.

Beechey CV, Cattanach BM, Blake A and Peters J (2003) *Mouse imprinting data and references. MRC Mammalian Genetics Unit, Harwell*, Oxfordshire, **http://www.mgu.har.mrc.ac.uk/research/imprinting/**.

Bourc'his D, Xu GL, Lin CS, Bollman B and Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science*, **294**, 2536–2539.

Caspary T, Cleary MA, Perlman EJ, Zhang P, Elledge SJ and Tilghman SM (1999) Oppositely imprinted genes p57(Kip2) and igf2 interact in a mouse model for Beckwith-Wiedemann syndrome. *Genes and Development*, **13**, 3115–3124.

Cattanach BM, Beechey CV and Peters J (2004) Interactions between imprinting effects in the mouse. *Genetics*, **168**, 397–413.

Eggenschwiler J, Ludwig T, Fisher P, Leighton PA, Tilghman SM and Efstratiadis A (1997) Mouse mutant embryos overexpressing IGF-II exhibit phenotypic features of the Beckwith-Wiedemann and Simpson-Golabi-Behmel syndromes. *Genes and Development*, **11**, 3128–3142.

El-Maarri O, Seoud M, Coullin P, Herbiniaux U, Oldenburg J, Rouleau G and Slim R (2003) Maternal alleles acquiring paternal methylation patterns in biparental complete hydatidiform moles. *Human Molecular Genetics*, **12**, 1405–1413.

Frank D, Fortino W, Clark L, Musalo R, Wang W, Saxena A, Li CM, Reik W, Ludwig T and Tycko B (2002) Placental overgrowth in mice lacking the imprinted gene Ipl. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7490–7495.

Fundele R, Li-Lan L, Herzfeld A, Barton SC and Surani MA (1995) Proliferation and differentiation of androgenetic cells in fetal mouse chimeras. *Roux's Archives of Developmental Biology*, **204**, 494–501.

Georgiades P, Watkins M, Surani MA and Ferguson-Smith AC (2000) Parental origin-specific developmental defects in mice with uniparental disomy for chromosome 12. *Development*, **127**, 4719–4728.

Guillemot F, Caspary T, Tilghman SM, Copeland NG, Gilbery DJ, Jenkins NA, Anderson DJ, Joyner AL, Rossant J and Nagy A (1995) Genomic imprinting of Mash-2, a mouse gene required for trophoblast development. *Nature Genetics*, **9**, 235–241.

Hata K, Okano M, Lei H and Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development*, **129**, 1983–1993.

Hitchins MP, Stanier P, Preece MA and Moore GE (2001) Silver-Russell syndrome: a dissection of the genetic aetiology and candidate chromosomal regions. *Journal of Medical Genetics*, **38**, 810–819.

Jamieson RV, Tan SS and Tam PP (1998) Retarded postimplantation development of X0 mouse embryos: impact of the parental origin of the monosomic X chromosome. *Developmental Biology*, **201**, 13–25.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E and Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, **429**, 900–903.

Kono T, Obata Y, Wu Q, Niwa K, Ono Y, Yamamoto Y, Park ES, Seo JS and Ogawa H (2004) Birth of parthenogenetic mice that can develop to adulthood. *Nature*, **428**, 860–864.

Kurosawa K, Sasaki H, Sato Y, Yamanaka M, Shimizu M, Ito Y, Okuyama T, Matsuo M, Imaizumi K, Kuroki Y, *et al*. (2002) Paternal UPD14 is responsible for a distinctive malformation complex. *American Journal of Medical Genetics*, **110**, 268–272.

Lee JT (2000) Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell*, **103**, 17–27.

Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R and Li E (1996) De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development*, **122**, 3195–3205.

Mann JR and Lovell-Badge RH (1987) The development of XO gynogenetic mouse embryos. *Development*, **99**, 411–416.

Marahrens Y, Panning B, Dausman J, Strauss W and Jaenisch R (1997) *Xist*-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes and Development*, **11**, 156–166.

McGrath J and Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, **37**, 179–183.

McLaughlin KJ, Kochanowski H, Solter D, Schwarzkopf G, Szabo PE and Mann JR (1997) Roles of the imprinted gene Igf2 and paternal duplication of distal chromosome 7 in the perinatal abnormalities of androgenetic mouse chimeras. *Development*, **124**, 4897–4904.

McLaughlin KJ, Szabo P, Haegel H and Mann JR (1996) Mouse embryos with paternal duplication of an imprinted chromosome 7 region die at midgestation and lack placental spongiotrophoblast. *Development*, **122**, 265–270.

Newman-Smith ED and Werb Z (1995) Stem cell defects in parthenogenetic peri-implantation embryos. *Development*, **121**, 2069–2077.

Obata Y, Ono Y, Akuzawa H, Kwon OY, Yoshizawa M and Kono T (2000) Post-implantation development of mouse androgenetic embryos produced by in-vitro fertilization of enucleated oocytes. *Human Reproduction*, **15**, 874–880.

Sado T, Wang Z, Sasaki H and Li E (2001) Regulation of imprinted X-chromosome inactivation in mice by Tsix. *Development*, **128**, 1275–1286.

Spindle A, Sturm KS, Flannery M, Meneses JJ, Wu K and Pedersen RA (1996) Defective chorioallantoic fusion in mid-gestation lethality of parthenogenone<–>tetraploid chimeras. *Developmental Biology*, **173**, 447–458.

Sturm KS, Flannery ML and Pedersen RA (1994) Abnormal development of embryonic and extraembryonic cell lineages in parthenogenetic mouse embryos. *Developmental Dynamics*, **201**, 11–28.

Takagi N (2003) Imprinted X-chromosome inactivation: enlightenment from embryos in vivo. *Seminars in Cell and Developmental Biology*, **14**, 319–329.

Takahashi K, Kobayashi T and Kanayama N (2000) p57(Kip2) regulates the proper development of labyrinthine and spongiotrophoblasts. *Molecular Human Reproduction*, **6**, 1019–1025.

Verona RI, Mann MRW and Bartolomei MS (2003) Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annual Review of Cell and Developmental Biology*, **19**, 237–259.

Specialist Review

Imprinting in Prader–Willi and Angelman syndromes

Bernhard Horsthemke and Karin Buiting

Institut für Humangenetik, Universitätsklinikum Essen, Essen, Germany

1. Clinical, cytogenetic, and molecular findings in PWS and AS

The Prader–Willi syndrome (PWS) and the Angelman syndrome (AS) are well known examples of human diseases involving imprinted genes. PWS is characterized by neonatal muscular hypotonia and failure to thrive, hyperphagia and obesity starting in early childhood, hypogonadism, short stature, small hands and feet, sleep apnea, behavioral problems, and mild to moderate mental retardation. As shown in Table 1, patients with PWS have a deletion of the paternal chromosome 15 [del(15)(q11-q13)pat], maternal uniparental disomy 15 [upd(15)mat], or a maternal imprint on the paternal chromosome (imprinting defect). All three lesions lead to the lack of expression of imprinted genes that are active on the paternal chromosome only. At present, several such genes are known: *MKRN3*, *MAGEL2*, *NDN, SNURF-SNRPN,* and more than 70 genes encoding small nucleolar RNAs (snoRNAs) (see Figure 1 and below). The contribution to PWS of any of these genes is unknown, but it is generally believed that PWS involves the loss of function of two or more genes in 15q11-q13.

AS is characterized by microcephalus, ataxia, absence of speech, abnormal EEG pattern, severe mental retardation, and frequent laughing. Similar to PWS, the most frequent lesions in AS are a deletion, uniparental disomy, and an imprinting defect, but these lesions affect the maternal chromosome (Table 1). Therefore, it was concluded that AS involves a maternally expressed gene. On the basis of the finding of point mutations in patients with AS who do not have one of the above mentioned lesions, *UBE3A* has been identified as the gene affected in AS. Interestingly, *UBE3A* is imprinted only in the brain. There is one other gene in 15q11-q13 that is expressed from the maternal chromosome only: *ATP10C*. Although this gene is not expressed in AS patients with a deletion, uniparental disomy, or an imprinting defect, its role in AS is unclear (see also below).

2. Imprinted genes in 15q11-q13

MAKORIN3 (*MKRN3*, formerly *ZNF127*) is a ubiquitously expressed intronless gene, which encodes a polypeptide with a RING zinc-finger and multiple C_3H

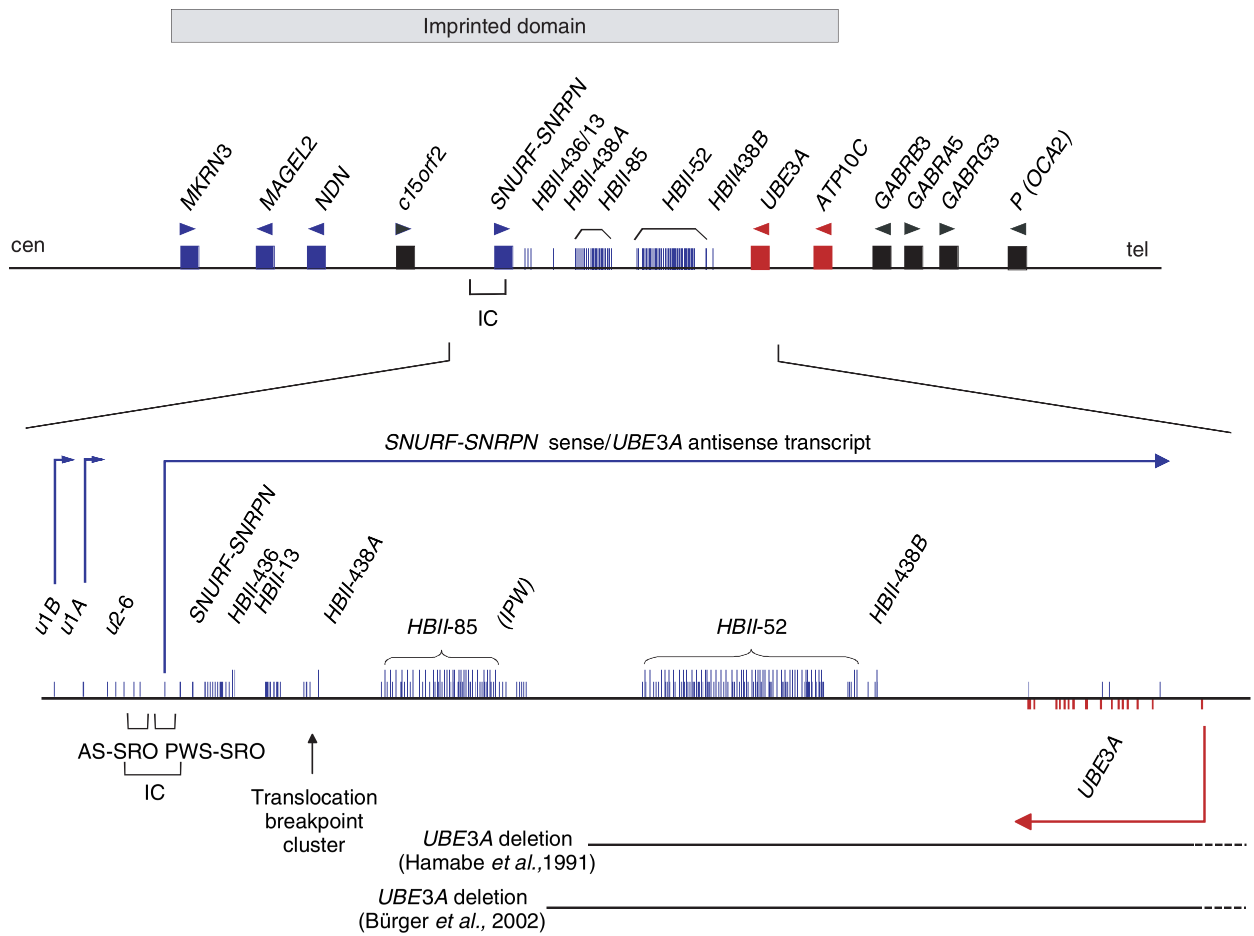


Figure 1 Physical map of the imprinted domain in 15q11-q13. Top: Gene order. Blue, paternally expressed genes; red, maternally expressed genes; black, biparentally expressed genes; arrowheads, orientation of transcription; IC, imprinting center; cen, centromere; tel, telomere. Bottom: structure of the *SNURF-SNRPN* transcription unit and the IC. U1B and u1A are alternative start sites. Short vertical blue lines, *SNURF-SNRPN* exons; long vertical blue lines, snoRNA genes; red vertical lines, *UBE3A* exons

Table 1 Genetic lesions in PWS and AS

Genetic lesion	PWS	AS
Deletion 15q11-q13	70% (pat)	70% (mat)
Uniparental disomy	28% (mat)	1% (pat)
Imprinting defect	1% (pat)	4% (mat)
Single gene mutation	–	5% (*UBE3A* mat)
Balanced translocation	Rare cases	–
Unknown	–	20%

zinc-finger motifs. Its function is unknown. *Mkrn3* knockout mice appear to be normal. Just telomeric to *MKRN3* are two other intronless genes, *MAGEL2* and *NECDIN* (*NDN*). Both genes encode proteins that are part of the melanoma-associated antigen (MAGE) protein family. *MAGEL2* is expressed only in the brain and placenta. Mouse models for *Magel2* are not yet available.

Whereas the human *NDN* gene is expressed in all tissues studied, with highest levels in brain and placenta, the murine *Ndn* gene is expressed predominantly in postmitotic neurons. The highest expression levels were found in the hypothalamus and other brain regions at late embryonic and early postnatal stages. It is upregulated during neuronal differentiation, and in vitro experiments have shown that overexpression of this gene leads to suppression of cell proliferation. Data obtained from different mouse models suggest that the loss of function of *Ndn* may contribute to respiratory distress and behavioral changes seen in patients with PWS (Ren *et al*., 2003).

The most complex locus in 15q11-q13 is *SNURF-SNRPN*. The original gene was found to consist of 10 exons, which encode two different proteins (Gray *et al*., 1999). Exons 1 to 3 encode SNURF (SNRPN upstream reading frame), a small polypeptide of unknown function, while exons 4 to 10 encode the small nuclear ribonucleioprotein N, a spliceosomal protein involved in RNA splicing in the brain. Mice lacking Snrpn appear to be normal (Yang *et al*., 1998). Knockouts of the promoter/exon 1 region of *Snurf-Snrpn* impair imprinting in the whole domain, because this region overlaps with the imprinting center (IC, see below).

In the past few years, many more 5′ and 3′ exons of *SNURF-SNRPN* have been identified. These exons have two peculiar features: they do not have any protein coding potential and they occur in many different splice forms of the primary transcript. Alternative transcripts containing novel 5′ exons were described by Dittrich *et al*. (1996) and characterized in detail by Färber *et al*. (1999). These transcripts start at two sites that share a high degree of sequence similarity. The 3′ exons are described in Runte *et al*. (2001). This analysis also showed that the *IPW* exons, which were previously thought to represent an independent gene ("imprinted in Prader–Willi syndrome"; Wevrick *et al*., 1994), are part of the *SNURF-SNRPN* transcription unit. Some of these splice variants are found predominately in brain and span the *UBE3A* gene in an antisense orientation.

Interestingly, the *SNURF-SNRPN* transcript also serves as a host for several snoRNAs, which are encoded within introns of this complex transcription unit. The genes are present as single copy genes (*HBII-13*, *HBII-436*, *HBII438A,* and *HBII-438B*) or as multigene clusters (*HBII-85* with 27 gene copies and *HBII-52*

with 47 gene copies) (Cavaillé *et al*., 2000; Runte *et al*., 2001). In contrast to other snoRNAs, which are usually involved in the modification of ribosomal RNAs, these snoRNAs do not have a region complementary to ribosomal RNA and might be involved in the modification of mRNAs. In *HBII-52*, 18 nucleotides are complementary to the serotonin receptor 2C mRNA (Cavaillé *et al*., 2000). It is possible that these snoRNAs are involved in the editing and/or alternative splicing of this mRNA.

In two unrelated families, a small deletion spanning *UBE3A* and the *HBII-52* gene cluster has been identified (Hamabe *et al*., 1991; Bürger *et al*., 2002). Whereas maternal transmission of the deletion leads to AS, paternal transmission is not associated with an obvious clinical phenotype. This excludes the *HBII-52* snoRNAs from a major role in PWS. The *HBII-85* gene cluster is distal to three balanced translocation breakpoints in patients with some features of PWS. As *HBII-85* is not expressed in one of the patients studied, these snoRNA may play a role in PWS. Knockout mice for the murine orthologs are currently being generated.

UBE3A encodes a ubiquitin-protein ligase that transfers ubiquitin to substrate proteins. Ubiquitin is a highly conserved 76 amino acid peptide that targets proteins for degradation. Thus, AS appears to result from a defect in proteasome-mediated protein degradation. *Ube3a* knockout mice (Jiang *et al*., 1998) serve as a model for AS.

ATP10C encodes a putative aminophospholipid translocase. Dhar *et al*. (2000) reported that maternal inheritance of deletions of the mouse *Atp10c* gene resulted in increased body fat. The obese phenotype was consistently observed in the mouse model for AS with paternal uniparental disomy as well as in a subset of sporadic AS ID patients (Gillessen-Kaesbach *et al*., 1999). However, loss of expression of *ATP10C*, which occurs in many patients with AS, is unlikely to play a role in the neurological features of this disease.

The murine orthologs of these genes map to mouse chromosome 7 (for a more detailed review on the human and murine genes see Nicholls and Knepper, 2001). Imprinted expression of the murine genes appears to be regulated in a similar way as in humans (see below).

3. Imprints

Paternal-only expression of *MKRN3*, *NDN,* and *SNURF-SNRPN* is associated with differential DNA methylation. Whereas the promoter/exon 1 regions of these genes are unmethylated on the expressing paternal chromosome, the silent maternal alleles are methylated (Glenn *et al*., 1993; Zeschnigk *et al*., 1997a). At the *SNURF-SNRPN* locus, a second differentially methylated region is found in intron 7. This region is preferentially methylated on the expressing paternal chromosome. The methylation status of the paternally expressed *MAGEL2* gene has not been thoroughly examined so far.

Differential DNA methylation can easily be detected with the help of methylation-sensitive restriction enzymes or by methylation-specific PCR (e.g., Sutcliffe *et al*., 1994; Zeschnigk *et al*., 1997b). AS and PWS patients with a deletion

15q11-q13, uniparental disomy 15, or an imprinting defect lack a methylated or an unmethylated band, respectively. Methylation analysis of *SNURF-SNRPN* is now the gold standard for testing patients suspected of having AS or PWS. Additional tests are necessary to distinguish between the different genetic lesions or, in AS patients with a normal methylation pattern, to search for a *UBE3A* mutation.

The promoter/exon 1 region of *SNURF-SNRPN*, which overlaps with the IC (see below), is the best studied region in 15q11-q13. In mice, the methylation imprint at this locus is clearly inherited from the gametes, but there is some controversy about the timing of methylation in human oogenesis. El-Maarri *et al*. (2001) have reported that in human oocytes *SNURF-SNRPN* is unmethylated and acquires a methylation imprint around or after fertilization. In contrast, Geuns *et al*. (2003) have found that oocytes are fully methylated at *SNURF-SNRPN*. This apparent discrepancy may be due to experimental problems or due to different protocols used for superovulation.

The methylation imprints are erased in primordial germ cells and newly established during later stages of gametogenesis. Interestingly, the acquisition of the maternal methylation imprint of the two alleles during oogenesis is asynchronous, at least in mice. Lucifero *et al*. (2004) found that the methylation imprint at the *Snurf-Snrpn* locus was initially established in preantral early growing oocytes on the maternally inherited allele and that the paternally inherited allele became methylated in more mature oocytes derived from antral follicles. These data indicate that the two alleles retain their parental identity after erasure of the parental imprints.

The paternally expressed snoRNA genes located between *SNURF-SNRPN* and *UBE3A* lack a direct methylation imprint. The snoRNAs are expressed from the paternal allele only, because they are processed from the paternally expressed *SNURF-SNRPN* sense/*UBE3A* antisense transcript. Thus, imprinted expression of the snoRNAs is indirectly regulated through *SNURF-SNRPN* methylation.

In vivo nuclease hypersensitive studies revealed that the two *SNURF-SNRPN* alleles do not only differ in DNA methylation but also in chromatin conformation (Schweizer *et al*., 1999). The paternal *SNURF-SNRPN* allele is associated with acetylated histones H4 and H3, whereas the maternal allele is hypoacetylated (Saitoh and Wada, 2000), in keeping with the general roles of DNA methylation and histone deacetylation to cooperate in transcriptional silencing. Xin *et al*. (2001) have demonstrated that the *SNURF-SNRPN* promoter/exon 1 region shows parent-specific complementary patterns of histone H3 lysine 9 (H3K9) and lysine 4 (H3K4) methylation. H3K9 is methylated on the maternal copy, and H3K4 is methylated on the paternal copy. They suggested that H3K9 methylation is a candidate maternal gametic imprint for this region. The authors also obtained evidence for a role of histone methylation in the establishment and maintenance of parent-specific DNA methylation (Xin *et al*., 2003).

In contrast to the paternally active *MKRN3*, *NDN,* and *SNURF-SNRPN* genes, the maternally active *UBE3A* gene lacks differential DNA methylation. Another striking difference is that imprinted *UBE3A* expression is tissue-specific. It has been proposed that paternal-only expression of a brain-specific antisense gene might prevent transcription of the paternal *UBE3A* gene (Rougeulle *et al*., 1998). A paternal deletion of the murine IC results in the loss of the antisense transcript

(Chamberlain and Brannan, 2001). As shown by Runte *et al*. (2001), the *UBE3A* antisense RNA is the 3′ end of the *SNURF-SNRPN* transcript, thus explaining its loss in mice and humans carrying a paternally inherited IC deletion. As the ultimate 3′ end of the *SNURF-SNRPN* sense/*UBE3A* antisense transcript has not yet been defined, it is possible that it extends into the *ATP10C* locus and plays a role in paternal silencing of this gene also.

In addition to DNA methylation and histone modification, the parental copies of 15q11-q13 differ in replication timing. Although there is some heterogeneity across this region, most of the paternal copy replicates before the maternal copy (Kitsberg *et al*., 1993). Asynchronous replication is erased during gametogenesis and maintained in the early embryo. Imprinting and replication timing are different processes, but appear to be regulated in a coordinate manner.

Another feature of 15q11-q13 is homologous association (LaSalle *et al*., 1996). A temporal and spatial association between maternal and paternal chromosomes 15 was observed in human T lymphocytes by three-dimensional fluorescence in situ hybridization. This association occurred only during the late S phase of the cell cycle. Cells from PWS and AS patients were deficient in association. This observation has not been confirmed by other groups.

4. Imprinting defects and the control of genomic imprinting in 15q11-q13

As shown in Table 1, approximately 1% of patients with PWS have a maternal imprint on the paternal chromosome and approximately 4% of patients with AS have a paternal imprint on the maternal chromosome. These imprinting defects are associated with silencing of the paternally expressed genes and the maternally expressed genes in 15q11-q13, respectively. Imprinting defects offer a unique opportunity to identify some of the factors and mechanisms involved in imprint erasure, resetting and maintenance. In 10–15% of cases, the imprinting defects are caused by a microdeletion affecting the 5′ end of the *SNURF-SNRPN* locus. These deletions define the 15q IC, which regulates imprinting in the whole domain (Sutcliffe *et al*., 1994; Buiting *et al*., 1995).

4.1. Imprinting defects caused by an IC deletion

To date, 32 deletions and one inversion affecting the IC have been reported. All the microdeletions found in patients with PWS affect the *SNURF-SNRPN* promoter/exon 1 region. Some of the deletions have occurred de novo on the paternal chromosome, but in most cases they have been inherited from the father. The deletions are without any phenotypic effect when transmitted through the female germline. The shortest region of deletion overlap (PWS-SRO) is 4.3 kb (Ohta *et al*., 1999). Segregation analysis suggested that this region is required for establishing or maintaining the paternal imprint in 15q11-q13. Studies of a PWS family in which the father is mosaic for an IC deletion on his paternal chromosome provided evidence that the deletion chromosome acquired a maternal methylation imprint in his

somatic cells. Identical findings were made in chimaeric mice generated from two independent embryonic stem (ES) cell lines harboring a similar deletion (Bielinska *et al*., 2000). Additional studies have shown that the sperm DNA from males with a maternally inherited PWS IC deletion has a normal paternal methylation imprint throughout human 15q11-q13, indicating that the maternal methylation imprint on the mutant maternal chromosome was successfully erased in the paternal germline and a paternal imprint established (El-Maarri *et al*., 2001). Thus, the incorrect maternal methylation pattern on the mutant paternal chromosome of patients with PWS and an IC deletion must have occurred after fertilization. These findings indicate that the PWS-SRO is not necessary for the establishment of the paternal imprint, but for maintaining the paternal methylation imprint during embryonic development, as proposed by Tilghman *et al*. (1998).

Targeted deletions in the mouse have revealed that the murine gene cluster on chromosome 7 is regulated in a similar way. The paternal transmission of a 42-kb deletion including the *Snurf-Snrpn* promoter/exon 1 region resulted in maternal imprint on the paternal chromosome (Yang *et al*., 1998). The mutant mice were smaller than their wild type littermates, exhibited muscular hypotonia and died within the first days of life. In contrast, paternal transmission of smaller deletion had little or no effect (Bressler *et al*., 2001).

In contrast to the PWS IC deletions, none of the IC deletions found in patients with AS affects the *SNURF-SNRPN* promoter/exon 1 region. Nevertheless, a shortest region of overlap (AS-SRO) could be defined (Buiting *et al*., 1999). The AS-SRO is 880 bp and maps 35 kb proximal to *SNURF-SNRPN* exon 1. Some of the deletions have occurred de novo on the maternal chromosome, but in most cases they have been inherited from the mother. The deletions are without any phenotypic effect when transmitted through the male germline and appear to have a role in establishing the maternal imprint in the female germline. The AS-SRO includes upstream exons u5 and u6 of the 5′ alternative *SNURF-SNRPN* transcripts (Dittrich *et al*., 1996; Färber *et al*., 1999). Dittrich *et al*. (1996) have suggested a model in which transcripts containing these exons are a major factor for setting up the maternal imprint. In this model, the AS-SRO is an imprintor that acts on the imprint switch initiation site (the PWS-SRO) to establish the maternal imprint.

The basic idea that the AS-SRO interacts with the PWS-SRO was confirmed by Perk *et al*. (2002), who showed that the maternal AS-SRO is essential for setting up the DNA methylation state and closed chromatin structure of the neighboring PWS-SRO. In contrast, the PWS-SRO had no influence on the epigenetic features of the AS-SRO. These results suggested a stepwise, unidirectional program in which structural imprinting at the AS-SRO brings about allele-specific repression of the maternal PWS-SRO. Transgenic mouse studies by Shemer *et al*. (2000) further supported this view, but challenged a role of alternative *SNURF-SNRPN* transcript in this process. The authors showed that a minitransgene with 1.0 kb of the human AS-SRO sequence fused to 0.2 kb of the mouse *Snurf-Snrpn* minimal promoter (homologous to the PWS-SRO) was appropriately imprinted after maternal and paternal transmission. These data suggest that the AS-SRO contains or overlaps one or more binding site for *trans*-acting factors. Although the AS-SRO is not differentially methylated in blood (Buiting *et al*., 2003), the demonstration of increased nuclease hypersensitivity on the maternal chromosome in the AS-SRO

region is consistent with the binding of *trans*-factors (Schweizer *et al*., 1999; Perk *et al*., 2002). Gel shift experiments as well as targeted deletions of different parts of the 0.2 kb *Snurf-Snrpn* minimal promoter of the transgene revealed the presence of five cis elements and putative trans-acting factors important for various steps in the imprinting process (Kantor *et al*., 2004).

It is unclear how the imprint spreads from the IC throughout the whole imprinted domain, but factors involved in chromatin conformation, such as heterochromatin-forming areas (matrix attachment regions, MARs) and transcription factor binding sites may play a role in this process. Genetic studies have shown that the IC is composed of an unusually high density of MARs with maternal-specific condensation, located in close proximity to the AS-SRO and PWS-SRO (Greally *et al*., 2000) and contains six “phylogenetic footprints”, DNA sequences of 7–10 bp that are conserved in the human and mouse *SNURF-SNRPN* promoter region (Ohta *et al*., 1999). Some of the latter sites are probably transcription factor binding sites for regulation of somatic expression of the gene, but one or more may also have IC regulatory function as sites for binding certain transcription factors during imprint erasure and resetting.

4.2. *Imprinting defects not caused by an IC deletion*

By analyzing more than one hundred PWS and AS patients with an imprinting defect, Buiting *et al*. (2003) have determined that in 85–90% of these cases the defect is not caused by an IC deletion, but that it is an epimutation (aberrant epigenetic state) that occurred spontaneously in the absence of DNA sequence changes. The apparent absence of point mutations, at least in the known critical elements of the IC (the AS-SRO and PWS-SRO) may indicate that the IC can tolerate small sequence changes or that it contains multiple, redundant elements. The latter notion is supported by the findings of multiple small IC sequence elements in the mouse (Kantor *et al*., 2004) (Figure 2).

Buiting *et al*. (2003) have also found that in PWS the imprinting defect was always on the chromosome 15 that was inherited from the paternal grandmother. This indicates that the paternal germline had failed to erase the maternal imprint so that the grandmaternal imprint was transmitted to the grandchild. This is the first demonstration of epigenetic inheritance in man.

In contrast, the imprinting defect in AS patients was on the chromosome 15 inherited from either the maternal grandmother or the maternal grandfather. This suggests that the maternal germline failed to establish a maternal imprint or that the maternal imprint was not maintained after fertilization. A postzygotic error in imprint maintenance is most likely in patients who are somatic mosaics. These patients, who comprise at least one-third of cases, have cells with an imprinting defect as well as normal cells. Methylation mosaicism is rarely observed in PWS. Using a novel quantitative MS-PCR assay, Nazlican *et al*. (2004) have found that the percentage of normal cells in peripheral blood of mosaic AS patients can vary considerably and that patients with a higher percentage of normal cells tend to have milder symptoms. Interestingly, some of these patients have an atypical

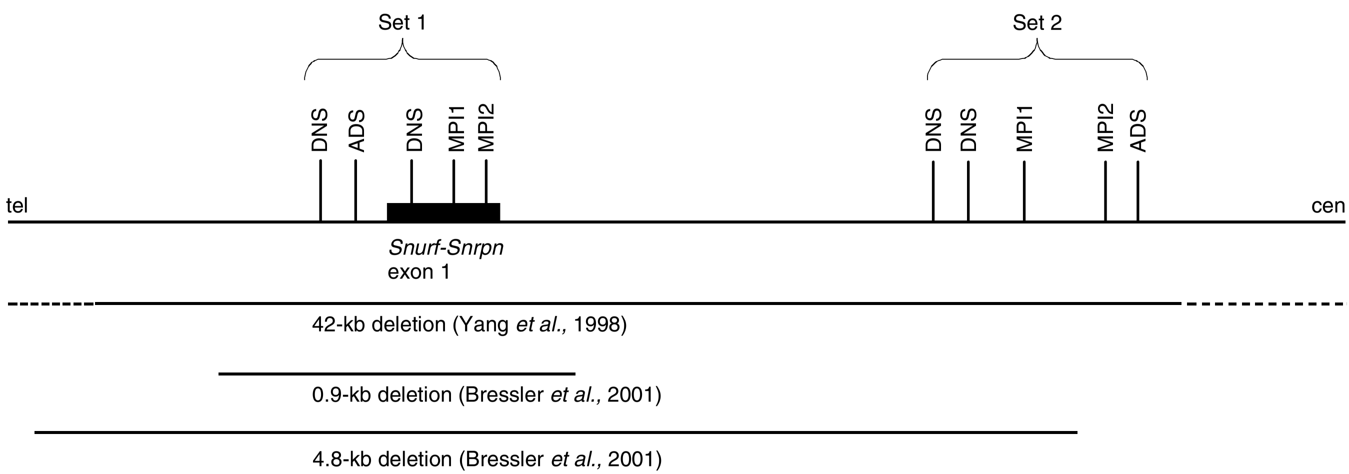


Figure 2 Sequence elements of the murine PWS-SRO. Set 1 overlaps *Snurf-Snrpn* exon 1 (black box). Set 2 is in intron 1. DNS, de novo methylation signal; ADS, allele discrimination signal; MPI, elements required for maintaining the paternal imprint (Reproduced from Kantor *et al*., 2004 by permission of Oxford University Press)

phenotype characterized by muscular hypotonia, early onset obesity and ability to speak (Gillessen-Kaesbach *et al*., 1999).

It is unclear, why, despite of an intact IC, the maternal imprint may not be established or be lost. Possible explanations include stochastic errors of the enzymatic machinery, genetic predisposition, and exogenous factors. A very critical period for imprint maintenance are the first few days after fertilization. Although methylation imprints normally survive the wave of global demethylation occurring during preimplantation development, it may occasionally happen that this protection fails in one of the cells of the preembryo. The methylation imprint may also be lost during later stages of development, if the methylation imprint is not copied onto the newly synthesized DNA strand after DNA replication.

In addition to stochastic errors, it is also possible that common sequence variants of the IC are associated with an increased risk of imprinting defects. These sequence variants may not impair the binding of trans-acting factors, but have a somewhat lower affinity to these factors, thus leading to a somewhat increased error rate. A study addressing this possibility has been initiated by the authors.

Last, but not least, exogenous factors may play a role. Reports on three AS patients conceived by intracytoplasmic sperm injection (ICSI) have suggested that assisted reproductive technology (ART) may increase the risk of imprinting defects (Cox *et al*., 2002; Ørstavik *et al*., 2003). The same concern was raised with regard to Beckwith–Wiedemann syndrome, another disease involving imprinted genes. Further studies are necessary to confirm these findings and to investigate whether hormonal stimulation, gamete manipulation, or culture conditions are the responsible factors. Ludwig *et al*. (2005) found that infertility per se and hormonal stimulation are associated with an increased risk of imprinting defects leading to Angelman syndrome.

Further reading

Boccaccio I, Glatt-Deeley H, Watrin F, Roeckel N, Lalande M and Muscatelli F (1999) The human *MAGEL2* gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. *Human Molecular Genetics*, **8**, 2497–2505.

Buiting K, Dittrich B, Endele S and Horsthemke B (1997) Identification of novel exons 3′ of the human *SNRPN* gene. *Genomics*, **40**, 132–137.

Cattanach BM, Barr JA, Evans EP, Burtenshaw M, Beechey CV, Leff SE, Brannan CI, Copeland NG, Jenkins NA and Jones J (1992) A candidate mouse model for Prader-Willi syndrome which shows an absence of SNRPN expression. *Nature Genetics*, **2**, 270–274.

DeBaun MR, Niemitz EL and Feinberg AP (2003) Association of in vitro fertilization with Beckwith-Wiedemann syndrome and epigenetic alterations of LIT1 and H19. *American Journal of Human Genetics*, **72**, 156–160.

Dittrich B, Robinson W, Knoblauch H, Buiting K, Schmidt K, Gillessen-Kaesbach G and Horsthemke B (1992) Molecular diagnosis of the Prader-Willi and Angelman syndromes by detection of parent-of-origin specific DNA methylation in 15q11-13. *Human Genetics*, **90**, 313–315.

Driscoll DJ, Waters MF, Williams CA, Zori RT, Glenn CC, Avidano KM and Nicholls RD (1992) A DNA methylation imprint, determined by the sex of the parent, distinguishes the Angelman and Prader-Willi syndromes. *Genomics*, **13**, 917–924.

Gicquel C, Gaston V, Mandelbaum J, Flahault A and Le Bouc Y (2003) In vitro fertilization may increase the risk of Beckwith-Wiedemann syndrome related to the abnormal imprinting of the *KCN1OT* gene. *American Journal of Human Genetics*, **72**, 1338–1341.

Fulmer-Smentek SB and Francke U (2000) Association of acetylated histones with paternally expressed genes in the Prader-Willi deletion region. *Human Molecular Genetics*, **10**, 645–652.

Herzing LB, Kim SJ, Cook EH and Ledbetter DH (2001) The human aminophospholipid-transporting ATPase gene *ATP10C* maps adjacent to *UBE3A* and exhibits similar imprinted expression. *American Journal of Human Genetics*, **68**, 1501–1505.

Jay P, Rougeulle C, Massacrier A, Moncla A, Mattei MG, Malzac P, Roeckel N, Taviaux S, Lefranc JL, Cau P, *et al*. (1997) The human necdin gene, *NDN*, is maternally imprinted and located in the Prader-Willi syndrome chromosomal region. *Nature Genetics*, **17**, 357–361.

Jong MTC, Gray TA, Ji Y, Glenn CC, Saitoh S, Driscoll DJ and Nicholls RD (1999) A novel imprinted gene, encoding a RING zinc-finger protein, and overlapping antisense transcript in the Prader-Willi syndrome critical region. *Human Molecular Genetics*, **8**, 783–793.

Kishino T, Lalande M and Wagstaff J (1997) *UBE3A/E6-AP* mutations cause Angelman syndrome. *Nature Genetics*, **15**, 70–73.

Knoll JH, Cheng SD and Lalande M (1994) Allele specificity of DNA replication timing in the Angelman/Prader-Willi syndrome imprinted chromosomal region. *Nature Genetics*, **6**, 41–46.

Kubota T, Das S, Christian SL, Baylin SB, Herman JG and Ledbetter DH (1996) Methylation-specific PCR simplifies imprinting analysis. *Nature Genetics*, **16**, 16–17.

Lee S, Kozlov S, Hernandez L, Chamberlain SJ, Brannan CI, Stewart CL and Wevrick R (2000) Expression and imprinting of *MAGEL2* suggest a role in Prader-Willi syndrome and the homologous murine imprinting phenotype. *Human Molecular Genetics*, **9**, 1813–1819.

MacDonald H and Wevrick R (1997) The *necdin* gene is deleted in Prader-Willi syndrome and is imprinted in human and mouse. *Human Molecular Genetics*, **11**, 1873–1878.

Maher ER, Brueton LA, Bowdin SC, Luharia A, Cooper W, Cole TR, Macdonald F, Sampson JR, Barratt CL, Reik W, *et al*. (2003) Beckwith-Wiedemann syndrome and assisted reproduction technology (ART). *Journal of Medical Genetics*, **40**, 62–64; Erratum in *Journal of Medical Genetics*, **40**, 304.

Matsuura T, Sutcliffe JS, Fang P, Nakao M, Kondo I, Saitoh S and Oshimura M (1997) De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (*UBE3A*) in Angelman syndrome. *Nature Genetics*, **5**, 74–77.

Meguro M, Kashiwagi A, Mitsuya K, Nakao M, Kondo I, Saitoh S and Oshimura M (2001) A novel maternally expressed gene, *ATP10C* encodes a putative aminophospholipid translocase associated with Angelman syndrome. *Nature Genetics*, **28**, 19–20.

Özcelik T, Leff S, Robinson W, Donlon T, Lalande M, Sanjines E, Schinzel A and Francke U (1992) Small nuclear ribonucleoprotein polypeptide N (*SNRPN*), an expressed gene in the Prader-Willi syndrome critical region. *Nature Genetics*, **2**, 265–269.

Rougeulle C, Glatt H and Lalande M (1997) The Angelman syndrome candidate gene, *UBE3A/E6-AP*, is imprinted in brain. *Nature Genetics*, **17**, 14–15.

Runte M, Kroisel PM, Gillessen-Kaesbach G, Varon R, Horn D, Cohen MY, Wagstaff J, Bernhard Horsthemke B and Buiting K (2004) *SNURF-SNRPN* and *UBE3A* transcript levels in patients with Angelman syndrome. *Human Genetics*, **114**, 553–561.

Schumacher A, Buiting K, Zeschnigk M, Doerfler W and Horsthemke B (1998) Methylation analysis of the PWS/AS region does not support an enhancer-competition model. *Nature Genetics*, **19**, 324–325.

Simon I, Tenzen T, Reubinoff BE, Hillman D, McCarrey JR and Cedar H (1999) Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature*, **401**, 929–932.

Vu TH and Hoffman AR (1997) Imprinting of the Angelman syndrome gene, *UBE3A*, is restricted to brain. *Nature Genetics*, **17**, 12–13.

Wirth J, Back E, Hüttenhofer A, Nothwang H-G, Lich C, Groß S, Menzel C, Schinzel A, Kioschis P, Tommerup N, *et al*. (2001) A translocation breakpoint cluster disrupts the newly defined 3′ end of the *SNURF-SNRPN* transcription unit on chromosome 15. *Human Molecular Genetics*, **10**, 201–210.

Yamasaki K, Joh K, Ohta T, Masuzaki H, Ishimaru T, Mukai T, Niikawa N, Ogawa M, Wagstaff J and Kishino T (2003) Neurons but not glial cells show reciprocal imprinting of sense and antisense transcripts of Ube3a. *Human Molecular Genetics*, **12**, 837–847.

References

Bielinska B, Blaydes SM, Buiting K, Yang T, Krajewska-Walasek M, Horsthemke B and Brannan CI (2000) De novo deletions of the *SNRPN* exon 1 region in early human and mouse embryos result in a paternal to maternal imprint switch. *Nature Genetics*, **25**, 74–78.

Bressler J, Tsai TF, Wu MY, Tsai SF, Ramirez MA, Armstrong D and Beaudet AL (2001) The *SNRPN* promoter is not required for genomic imprinting of the Prader-Willi/Angelman domain in mice. *Nature Genetics*, **28**, 232–240.

Buiting K, Groß S, Lich C, Gillessen-Kaesbach G, El-Maarri O and Horsthemke B (2003) Epimutations in Prader-Willi and Angelman syndrome: a molecular study of 136 patients with an imprinting defect. *American Journal of Human Genetics*, **72**, 571–577.

Buiting K, Lich C, Cottrell S, Barnicoat A and Horsthemke B (1999) A 5-kb imprinting center deletion in a family with Angelman syndrome reduces the shortest region of deletion overlap to 880 bp. *Human Genetics*, **105**, 665–666.

Buiting K, Saitoh S, Groß S, Dittrich B, Schwartz S, Nicholls R and Horsthemke B (1995) Inherited microdeletions in the Angelman and Prader-Willi syndromes define an imprinting center on human chromosome 15. *Nature Genetics*, **9**, 395–400.

Bürger J, Horn D, Tonnies H, Neitzel H and Reis A (2002) Familial interstitial 570 kbp deletion of the *UBE3A* gene region causing Angelman syndrome but not Prader-Willi syndrome. *American Journal of Medical Genetics*, **111**, 233–237.

Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie J-P, Brosius J and Hüttenhofer A (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proceedings of the National Academy of Sciences of the United States of America*, **7**, 14311–14316.

Chamberlain SJ and Brannan CI (2001) The Prader-Willi syndrome imprinting-center activates the paternally expressed murine *Ube3a* antisense transcript, but represses paternal *Ube3a*. *Genomics*, **73**, 316–322.

Cox GF, Bürger J, Lip V, Mau UA, Sperling K, Wu BL and Horsthemke B (2002) Intracytoplasmic sperm injection (ICSI) may increase the risk for imprinting defects. *American Journal of Human Genetics*, **71**, 162–164.

Dhar M, Webb LS, Smith L, Hauser L, Johnson D and West DB (2000) A novel ATPase on mouse chromosome 7 is a candidate gene for increased body fat. *Physiological Genomics*, **4**, 93–100.

Dittrich B, Buiting K, Korn B, Rickard S, Buxton J, Saitoh S, Nicholls RD, Poustka A, Winterpacht A, Zabel B, *et al*. (1996) Imprint switching on human chromosome 15 may involve alternative transcripts of the *SNRPN* gene. *Nature Genetics*, **14**, 163–170.

El-Maarri O, Buiting K and Peery EG (2001) Maternal methylation imprints on human chromosome 15 are established during or after fertilization. *Nature Genetics*, **27**, 341–344.

Färber C, Dittrich B, Buiting K and Horsthemke B (1999) The chromosome 15 imprinting centre (IC) region has undergone multiple duplication events and contains an upstream exon of *SNRPN* that is deleted in all Angelman syndrome patients with an IC microdeletion. *Human Molecular Genetics*, **8**, 337–343.

Geuns E, De Rycke M, Van Steirteghem A and Liebaers I (2003) Methylation imprints of the imprint control region of the SNRPN-gene in human gametes and preimplantation embryos. *Human Molecular Genetics*, **12**, 2873–2879.

Gillessen-Kaesbach G, Demuth S, Thiele H, Theile U, Lich Ch and Horsthemke B (1999) A previously unrecognised phenotype characterised by obesity, muscular hypotonia, and ability to speak in patients with Angelman syndrome caused by an imprinting defect. *European Journal of Human Genetics*, **7**, 638–644.

Glenn CC, Nicholls RD, Robinson WP, Saitoh S, Niikawa N, Schinzel A, Horsthemke B and Driscoll DJ (1993) Modification of 15q11-q13 DNA methylation imprints in unique Angelman and Prader-Willi patients. *Human Molecular Genetics*, **2**, 1377–1382.

Gray TA, Saitoh S and Nicholls RD (1999) An imprinted, mammalian bicistronic transcript encodes two independent proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5616–5621.

Greally JM, Gray TA, Gabriel JM, Song L, Zemel S and Nicholls RD (1999) Conserved characteristics of heterochromatin-forming DNA at the 15q11-q13 imprinting center. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14430–14435; Erratum in: (2000) *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4410.

Hamabe J, Kuroki Y, Imaizumi K, Sugimoto T, Fukushima Y, Yamaguchi A, Izumikawa Y and Niikawa N (1991) DNA deletion and its parental origin in Angelman syndrome patients. *American Journal of Medical Genetics*, **41**, 64–68.

Jiang YH, Armstrong D, Albrecht U, Atkins CM, Noebels JL, Eichele G, Sweatt JD and Beaudet AL (1998) Mutation of the Angelman ubiquitin ligase in mice causes increased cytoplasmic p53 and deficits of contextual learning and long-term potentiation. *Neuron*, **21**, 799–811.

Kantor B, Makedonski K, Green-Finberg Y, Shemer R and Razin A (2004) Control elements within the PWS/AS imprinting box and their function in the imprinting process. *Human Molecular Genetics*, **13**, 751–762.

Kitsberg D, Selig S, Brandeis M, Simon I, Keshet I, Driscoll DJ, Nicholls RD and Cedar H (1993) Allele-specific replication timing of imprinted gene regions. *Nature*, **364**, 459–463.

LaSalle JM and Lalande M (1996) Homologous association of oppositely imprinted chromosomal domains. *Science*, **272**, 725–728.

Lucifero D, Mann MR, Bartolomei MS and Trasler JM (2004) Gene-specific timing and epigenetic memory in oocyte imprinting. *Human Molecular Genetics*, **13**, 839–849.

Ludwig M, Katalinic A, Groß S, Sutcliffe A, Varon R and Horsthemke B (2005) Increased prevalence of imprinting defects in patients with Angelman syndrome born to subfertile couples. *Journal of Medical Genetics*, **42**, 289–291.

Nazlican H, Zeschnigk M, Claussen U, Michel S, Boehringer S, Gillessen-Kaesbach G, Buiting K and Horsthemke B (2004) Somatic mosaicism in patients with Angelman syndrome and an imprinting defect. *Human Molecular Genetics*, **13**, 2547–2555.

Nicholls RD and Knepper JL (2001) Genome organization: function and Imprinting in Prader-Willi and Angelman syndromes. *Annual Review of Genomics and Human Genetics*, **2**, 153–175.

Ohta T, Gray TA, Rogan PK, Buiting K, Gabriel JM, Saitoh S, Muralidhar B, Bilienska B, Krajewska-Walasek M, Driscoll DJ, *et al*. (1999) Imprinting-mutation mechanisms in Prader-Willi syndrome. *American Journal of Human Genetics*, **64**, 397–413.

Ørstavik KH, Eiklid K, van der Hagen CB, Spetalen S, Kierulf K, Skjeldal O and Buiting K (2003) Another case of imprinting defect in a girl with Angelman syndrome who was conceived by intracytoplasmic semen injection. *American Journal of Human Genetics*, **72**, 218–219.

Perk J, Makedonski K, Lande L, Cedar H, Razin A and Shemer R (2002) The imprinting mechanism of the Prader-Willi/ Angelman regional control center. *The EMBO Journal*, **21**, 5807–5814.

Ren J, Lee S, Pagliardini S, Gerard M, Stewart CL, Greer JJ and Wevrick R (2003) Absence of *Ndn*, encoding the Prader-Willi syndrome-deleted gene necdin, results in congenital deficiency of central respiratory drive in neonatal mice. *Journal of Neuroscience*, **23**, 1569–1573.

Rougeulle C, Cardoso C, Fontes M, Colleaux L and Lalande M (1998) An imprinted antisense RNA overlaps *UBE3A* and a second maternally expressed transcript. *Nature Genetics*, **19**, 15–16.

Runte M, Hüttenhofer A, Gross S, Kiefmann M, Horsthemke B and Buiting K (2001) The IC-*SNURF-SNRPN* transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for *UBE3A*. *Human Molecular Genetics*, **10**, 2687–2700.

Saitoh S and Wada T (2000) Parent-of-origin specific histone acetylation and reactivation of a key imprinted gene locus in Prader-Willi syndrome. *American Journal of Human Genetics*, **66**, 1958–1962.

Schweizer J, Zynger D and Francke U (1999) In vivo nuclease hypersensitivity studies reveal multiple sites of parental origin-dependent differential chromatin conformation in the 150 kb *SNRPN* transcription unit. *Human Molecular Genetics*, **8**, 555–566.

Shemer R, Hershko AY, Perk J, Mostoslavsky P, Tsuberi B, Buiting K and Razin A (2000) The imprinting box of the Prader-Willi/Angelman syndrome domain. *Nature Genetics*, **26**, 440–443.

Sutcliffe JS, Nakao M, Christian S, Orstavik KH, Tommerup N, Ledbetter DH and Beaudet AL (1994) Deletions of a differentially methylated CpG island at the *SNRPN* gene define a putative imprinting control region. *Nature Genetics*, **8**, 52–58.

Tilghman S, Caspary T and Ingram RS (1998) Competitive edge at the imprinted Prader-Willi/Angelman region? *Nature Genetics*, **18**, 206–208.

Wevrick R, Kerns JA and Francke U (1994) Identification of a novel paternally expressed gene in the Prader-Willi syndrome region. *Human Molecular Genetics*, **3**, 1877–1882.

Xin Z, Allis CD and Wagstaff J (2001) Parent-specific complementary patterns of histone H3 lysine 9 and H3 lysine 4 methylation at the Prader-Willi syndrome imprinting center. *American Journal of Human Genetics*, **69**, 1389–1394.

Xin Z, Tachibana M, Guggiari M, Heard E, Shinkai Y and Wagstaff J (2003) Role of histone methyltransferase G9a in CpG methylation of the Prader-Willi syndrome imprinting center. *Journal of Biological Chemistry*, **278**, 14996–15000.

Yang T, Adamson TE, Resnick JL, Leff S, Wevrick R, Francke U, Jenkins NA, Copeland NG and Brannan CI (1998) A mouse model for Prader-Willi syndrome imprinting-centre mutations. *Nature Genetics*, **19**, 25–31.

Zeschnigk M, Schmitz B, Dittrich B, Buiting K, Horsthemke B and Dörfler W (1997a) Imprinted segments in the human genome: different DNA methylation patterns in the Prader-Willi/Angelman syndrome region as determined by the genomic sequencing method. *Human Molecular Genetics*, **6**, 387–395.

Zeschnigk M, Lich C, Buiting K, Horsthemke B and Dörfler W (1997b) A single-tube PCR test for the diagnosis of Angelman and Prader-Willi syndrome based on allelic methylation differences at the *SNRPN* locus. *European Journal of Human Genetics*, **5**, 94–98.

Specialist Review

Beckwith–Wiedemann syndrome

Benjamin Tycko
Columbia University, New York, NY, USA

Marcel Mannens
University of Amsterdam, Amsterdam, The Netherlands

1. BWS: historical aspects and clinical features

In 1964, Beckwith and, independently, Wiedemann reported a syndrome of macroglossia and omphalocele, associated with adrenal cortical cytomegaly, fetal gigantism, and other abnormalities. The original abstracts and articles describing these cases have been reviewed recently, along with an interesting discussion of gigantism in folklore and early medicine (Beckwith, 1998). The Beckwith–Wiedemann syndrome (BWS, MIM130650), as this constellation of findings came to be named, is thus characterized by overgrowth of many organs during fetal development and, as documented by numerous subsequent reports, an increased susceptibility to childhood tumors. Abdominal wall defects, usually umbilical hernia or omphalocele, can be severe in some cases and require surgical repair after birth. These defects are probably a consequence of the organomegaly, aggravated by a primary defect in the development of the abdominal wall. Macroglossia is often prominent (Figure 1), and partial glossectomy to reduce the size of the tongue is another surgical procedure sometimes performed on children with BWS. In other cases in which glossectomy can be avoided, the jaws grow to accommodate the tongue, making the macroglossia less evident later in life. The kidneys are often increased in size, and they sometimes contain substantial collections of primitive metanephric cells, the so-called nephrogenic rests (Beckwith *et al*., 1990). No doubt, related to this histological abnormality, BWS is associated with a roughly 7% incidence of Wilms tumor, an embryonal kidney cancer arising from metanephric precursor cells that are defective in their ability to differentiate into mature epithelial structures (Li *et al*., 2002).

Generalized somatic overgrowth, producing a variable degree of gigantism in the neonate, often accompanied by placentomegaly, is another cardinal feature of BWS. In some cases, the overgrowth can be asymmetrical, producing so-called hemihypertrophy, which is probably more accurately termed *hemihyperplasia*. Lastly, ear pits and creases and facial nevus can be part of the syndrome (Figure 1), and neonatal hypoglycemia is found in about half of BWS cases for which this information is available. As described below, BWS is not a unitary disorder, but instead it has multiple genetic and epigenetic etiologies.



Figure 1 Macroglossia, nevus flameus, and ear creases in a child with BWS

2. Differential diagnosis of BWS: overgrowth syndromes

The clinical differential diagnosis of BWS includes other overgrowth disorders, notably the Simpson–Golabi–Behmel (SGBS), Perlman, and Sotos syndromes, and molecular tests are now a substantive aid in distinguishing these conditions. In Sotos syndrome (MIM117550), caused by deletions and point mutations in the *NSD1* gene encoding a nuclear regulatory protein, macrocephaly and a characteristic facial gestalt are major consistent features, while overgrowth and advanced bone age are sometimes but not always observed. The Weaver syndrome is a related disorder, and some Weaver syndrome cases have been reported to carry *NSD1* mutations, making this disorder allelic with Sotos syndrome (Rio *et al*., 2003). In contrast, SGBS (MIM312870), caused in most cases by mutations in the glypican gene *GPC3* (or in other cases the adjacent *GPC4* gene) encoding a cell surface proteoglycan, includes somatic overgrowth as a major consistent feature. Renal dysplasia, polydactyly, macrocephaly and coarse facial features, and placentomegaly are additional features of SGBS, which have been comprehensively discussed in the context of a mouse model, the *Gpc3* knockout, which largely mimics the human disorder (Chiao *et al*., 2002). Perlman syndrome (MIM267000), comprising renal hamartomas, nephroblastomatosis, and fetal gigantism, but not omphalocele, is also in the differential diagnosis of overgrowth. Facial dysmorphism and a high perinatal mortality are additional features of this syndrome, which may show autosomal recessive inheritance (Greenberg *et al*., 1986). Because of its rarity, Perlman syndrome has not yet been defined in molecular terms. In a very complete differential diagnosis, one could also include the Klippel–Trenaunay–Weber syndrome, Proteus syndrome, and even neurofibromatis, because of the hemihyperplasia in this condition, albeit due to vascular malformations. Isolated hemihyperplasia can be a BWS patient with minimal features of BWS. Such individuals have a high tumor risk (5.9%) and

should be offered tumor surveillance (Hoyme *et al*., 1998). Lastly, overgrowth of the fetus is common and well known in the setting of maternal diabetes, but this type of macrosomia does not include omphalocele, macroglossia, or disproportionate visceromegaly; that is, the organ weights, while increased, are appropriate for the overall body size.

3. Genomic imprinting in the BWS region of chromosome 11p15

In a broad sense, BWS maps to human chromosome band 11p15, a chromosomal region that contains a large number of imprinted genes. So, to understand the various BWS-associated molecular defects, it is helpful to review the structure and imprinting of this DNA region. The basic concept of genomic or parental imprinting is reviewed elsewhere in this volume (*see* Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1 and Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1); briefly, imprinting is an epigenetic process, occurring in gametogenesis, which marks certain genes for allele-specific, parent-of-origin-dependent, mRNA expression in the conceptus. A large body of evidence implicates DNA methylation at critical CpG-rich sequences as the fundamental epigenetic mark controlling imprinting in mammals. Such "imprinting centers", which are at most several kilobases in length, control the allele-specific expression of multiple flanking genes, which can be found dispersed in up to several megabases of flanking DNA. Imprinting centers are differentially methylated on the maternal versus paternal chromosome homolog, and are therefore alternatively referred to as "differentially methylated regions" or DMRs (not every differentially methylated DNA sequence is an imprinting center, but the examples discussed here do function in this capacity). As shown in Figure 2, there are two distinct imprinting centers in chromosome band 11p15, one immediately upstream of the *H19* gene (the *H19* DMR) and the other within an intron of the large *KCNQ1* gene (the KvDMR1element, also known as the *LIT1-associated DMR*). The *H19* DMR controls the allele-specific expression of *H19* and *IGF2*, while the KvDMRl element controls the allele-specific expression of a larger cluster of imprinted genes, including *CDKN1C*. Thus, chromosome band 11p15 contains two distinct "imprinted domains". A list of the imprinted genes in this chromosomal region, with their biochemical functions, is in Table 1, and the transcriptionally active and silent alleles, that is, the direction of imprinting, for several of the relevant genes are diagramed in Figure 2. Most of these genes are maternally expressed/paternally silenced (a pattern often referred to as "paternally imprinted"), but an important exception is the growth-promoting gene *IGF2*, which is imprinted in the opposite direction, paternal allele active/maternal allele repressed.

As proven by knockout experiments in mice, each of the two imprinting centers acts in cis to enforce imprinting of its flanking genes (Fitzpatrick *et al*., 2002; Leighton *et al*., 1995). Somatic cell genetics using human chromosomes with an engineered deletion of KvDMR1/LIT1 also supports this conclusion (Horike *et al*.,

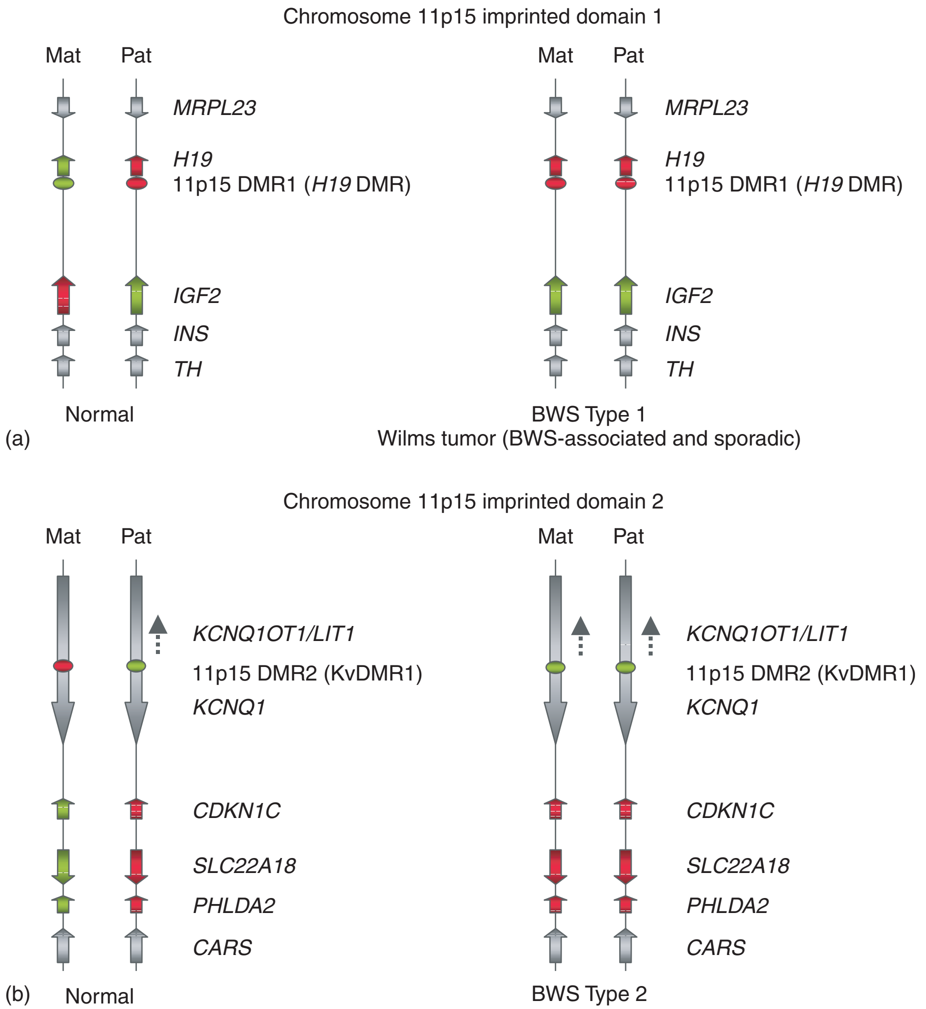


Figure 2 Two clusters of imprinted genes in the BWS-associated region of human chromosome 11p15. (a) The more distally located cluster of imprinted genes includes *H19* and *IGF2*. (b) The more proximally located cluster includes *CDKN1C* and several other imprinted genes. The red shading indicates lack of expression; green shading indicates active transcription. Gray shading indicates either lack of imprinting, weak tissue-specific imprinting or incompletely characterized imprinting. The cis-acting imprinting control elements (DMRs) are shown as ovals, with green indicating lack of CpG methylation and red indicating substantial CpG methylation. Directions of transcription are shown by the arrows. Mat: maternal chromosome; Pat: paternal chromosome. The abnormalities that lead to Type 1 and Type 2 BWS are indicated

2000). Several lines of data have led to a credible mechanistic model of how the *H19* DMR works to maintain the opposite allele-specific expression of *H19* and *IGF2* (Thorvaldsen and Bartolomei, 2000). As diagramed in Figure 3, this DMR acts as a chromatin insulator when it is unmethylated (maternal allele), and loses its insulator function when it is methylated (paternal allele). The unmethylated insulator, complexed to an insulator-binding protein called CTCF, blocks the

Table 1 Imprinted genes in the BWS-associated region of chromosome 11p15

Gene	Aliases	Expressed allele	Tissue-specific imprinting	Gene product
H19		Maternal	Many tissues	Nontranslated RNA
IGF2		Paternal	Many tissues	Trophic growth factor
INS		Paternal	Yolk sac	Insulin
ASCL2	*HASH2*	Maternal (stronger imprint in mice)	Extravillus trophoblast	Transcription factor for placental development
CD81	*TAPA1*	Maternal	Not known	Involved in signal transduction and lymphoma cell growth
TRPM5	*MTR1*	Paternal	Not known	Ion-channel
KCNQ1	*KvLQT1*	Maternal	Many tissues but not the heart	Ion-channel
KCNQ1OT1	*LIT1*	Paternal	Many tissues	Nontranslated RNA
CDKN1C	*p57*$^{\mathrm{KIP2}}$	Maternal	Many tissues	Cyclin-dependent kinase (cdk) inhibitor; regulates cell cycle and tissue growth
SLC22A18	*IMPT1* , *ITM*	Maternal	Many tissues	Membrane protein, putative multi drug resistance pump
PHLDA2	*TSSC3, IPL, BWR1C*	Maternal	Placenta and fetal liver	PH-domain protein regulating placental growth
OSBPL5	*ORP5*	Maternal	Placenta	Regulation of sterol metabolism
ZNF215		Maternal	Liver, lung, kidney, testis (not brain and heart)	Zinc-finger protein putative transcription factor

interaction of the *IGF2* promoter with downstream enhancer sequences while permitting the *H19* promoter to bind to these enhancers. In this situation (maternal allele), *H19* nontranslated RNA is produced, while *IGF2* mRNA is not. The opposite situation pertains for the paternal allele, in which CTCF binding is prevented by DNA methylation, and the *IGF2* promoter is therefore free to interact with the downstream enhancer elements. Precisely how the analogous KvDMRl element acts to maintain imprinting of its flanking genes is not yet clear. Like the *H19* DMR, this element is close to the initiation site for a nontranslated RNA (called *LIT1/KCNQ1OT1*). The function of this RNA is uncertain, but the KvDMR1 element is CG-rich, conserved in mammalian evolution, differentially methylated on the two alleles, and essential for imprinting of distant genes in cis, so it seems likely that some of the principles established for the *H19* DMR will also apply to KvDMR1.

4. Genetic and epigenetic etiologies of BWS

The phenotypic manifestations of BWS are variable, and this clinical heterogeneity can now be largely explained by an underlying molecular heterogeneity. As shown in Table 2, the majority of BWS cases can be assigned to one of several molecular categories. The observation that some individuals with BWS show mosaic uniparental paternal disomy of chromosome 11 on RFLP (restriction

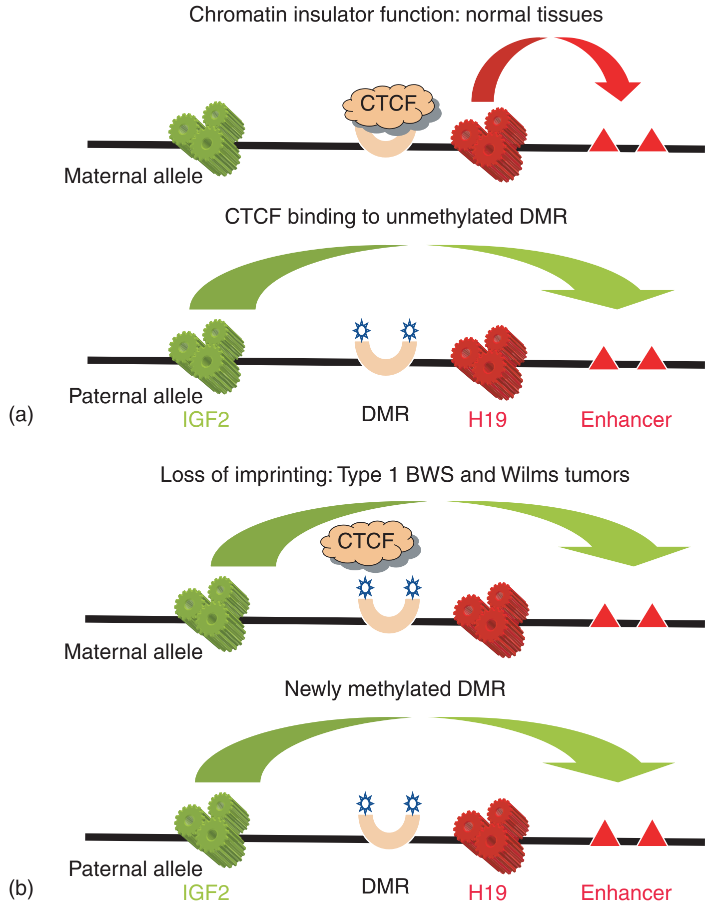


Figure 3 The chromatin insulator model for opposite imprinting of H19 and IGF2. (a) The insulator binding protein CTCF occupies the *H19* upstream DMR on the maternal allele, thereby preventing the *IGF2* promoter from interacting with the shared downstream enhancer elements, and enforcing monoallelic expression of *H19* RNA and *IGF2* mRNA from opposite alleles. (b) After de novo methylation of the DMR/insulator, *IGF2* becomes actively expressed from both alleles, and *H19* RNA is lost

fragment length polymorphism) analysis of blood or fibroblast DNA was an early indication of a role for imprinted genes in this disorder (Henry *et al*., 1991; Henry *et al*., 1993), and a substantial minority of cases, about 20%, can be attributed to this abnormality. Similarly, an important early finding was the existence of rare families in which BWS was transmitted by mothers but not fathers, and in which linkage could be established to markers on chromosome 11p15 (Koufos *et al*., 1989; Ping *et al*., 1989). In retrospect, most such cases are likely due to either mutations in *CDKN1C* (Hatada *et al*., 1996; Lam *et al*., 1999; O'Keefe *et al*., 1997) or recently described DNA microdeletions in the *CDKN1C* enhancer (Niemitz *et al*., 2004) or the *H19* DMR (Sparago *et al*., 2004). Another small subset of cases is accounted for by rare constitutional chromosomal translocations involving band 11p15, which

Table 2 Genetic and epigenetic defects causing BWS

Molecular defect	Dysregulated gene[a]	Proportion of cases	Clinical correlations
CDKN1C mutation	*CDKN1C*	3–5%	Omphalocele
KvDMR1[mat] loss of methylation	*CDKN1C*	50%	Omphalocele
			Rare tumors other than Wilms
H19[mat] gain of methylation	*IGF2*[b]	10%	Wilms tumors
Mosaic patUPD 11p15	*CDKN1C* and *IGF2*	20%	Wilms tumors
			Hemihyperplasia
Chromosomal translocations; chromosome band 11p15	*CDKN1C*	∼3%	Not known
	ZNF215 dysregulated		
Trisomy involving chromosome band 11p15	*IGF2*	5%	Overgrowth
			Wilms tumors
Unexplained (low level mosaicism?)	*CDKN1C?*	10–15%	Not known
	IGF2?		

[a]See the text for a discussion of additional chromosome 11p15 genes that may contribute to phenotypic variability in BWS.
[b]*H19* noncoding RNA also silenced in these cases.

also cause the syndrome only when transmitted maternally (Mannens *et al*., 1996). Some of these translocations have DNA breakage and rejoining close to KvDMR1, and presumably silence the expression of *CDKN1C* (11p15.5) (Lee *et al*., 1997b), while others influence the expression of a less well understood paternally imprinted gene, *ZNF215*, in chromosome band 11p15.4 (Alders *et al*., 2000). Currently, two patients define this more proximal BWS-associated chromosomal region; both demonstrated hemihyperplasia and minimal signs of BWS, and one developed a Wilms tumor.

In distinction to these genetic causes of BWS, two additional large groups of BWS cases are accounted for purely by epigenetic defects (Figures 2 and 3). In about 10% of cases, there is a pathological finding of gain of methylation (GOM) on the maternal *H19* DMR. This epigenetic lesion causes loss of imprinting (LOI) of *IGF2*, producing a double gene dosage of insulin-like growth factor II (Figures 2a and 3). "Nonsyndromic *IGF2* overgrowth disorder", also due to tissue mosaicism for GOM at the *H19* DMR, but with clinical hallmarks of gigantism and predisposition to Wilms tumor, but not macroglossia or abdominal wall defects, is a forme fruste of this category of BWS (Ogawa *et al*., 1993; Reeve, 1996). We refer this category of BWS as "Type 1" in Figure 2. The second epigenetic class of BWS, labeled "Type 2" in Figure 2, is attributed to loss of DNA methylation (LOM) on

the maternal allele of KvDMR1/LIT1 (Lee *et al*., 1999; Smilinich *et al*., 1999). As shown in Figure 2, this epigenetic defect leads to pathological biallelic expression of the *LIT1/KCNQ1OT1* untranslated RNA, and correlates with transcriptional downregulation (via gain-of-imprinting) of *CDKN1C* (Diaz-Meyer *et al*., 2003). Lastly, in about 10–15% of BWS no genetic or epigenetic defect is known. One possibility is that variable tissue mosaicism for LOM at KvDMR1 or GOM at the *H19* DMR may underlie some of these cases. Alternatively, some or all of these cases may actually be affected by Sotos syndrome, SGBS, or other BWS mimics.

5. Imprinted genes and growth regulation

From the data discussed above, and from mouse models (Caspary *et al*., 1999; Eggenschwiler *et al*., 1997), it is clear that the primary genes in the pathogenesis of BWS are *CDKN1C* and *IGF2*. These are good examples of imprinted genes that act to control growth – *IGF2* by encoding an antiapoptotic trophic factor with a positive net effect on tissue growth and *CDKN1C* by encoding a cyclin/cdk inhibitor with a negative net effect on cell proliferation and tissue growth. But the role of imprinted genes in mammalian growth regulation goes well beyond these examples. When this area was reviewed in 2002, there were nine examples of imprinted protein-coding genes proven by in vivo genetic data to control pre- or postnatal growth in mice and/or humans (Tycko and Morison, 2002), and an update from the recent literature reveals at least two additional examples (Charalambous *et al*., 2003; Moon *et al*., 2002). Strikingly, for each of these genes, the effect on growth correlates systematically with the direction of imprinting, with paternally expressed genes exerting a positive effect and maternally expressed genes a negative effect on net growth. This correlation is as predicted by the parental conflict theory of imprinting, which was first articulated in the early 1990s after the initial reports of opposite growth phenotypes in mice mutant for the oppositely imprinted *Igf2* and *Igf2r* genes (Moore and Haig, 1991).

Given the large number of imprinted growth-regulating genes in addition to *IGF2* and *CDKN1C*, it is interesting to consider whether aberrant expression of any of these genes might also contribute to overgrowth in humans. On chromosome 11p15 and the syntenic region of distal chromosome 7, in addition to *CDKN1C*, the *PHLDA2* gene (also known as *IPL/TSSC3*) is a bona fide growth suppressor, with placental overgrowth in *Phlda2* knockout mice and placental stunting in conceptuses engineered to overexpress this gene (Frank *et al*., 2002; Salas *et al*., 2004). Whether loss of expression of *PHLDA2* contributes to placentomegaly in BWS is an open question. While this also needs more study, the insulin gene (*INS*), closely upstream of *IGF2*, may well be overexpressed along with *IGF2* in the category of BWS with GOM at the *H19* DMR. Increased production of insulin may account at least in part for the finding of hypoglycemia in BWS (DeBaun *et al*., 2000), and since insulin can be mitogenic, such overexpression may also contribute to overgrowth. A related issue is whether any other imprinted chromosomal regions might contain genes that contribute to human overgrowth.

Uniparental maternal disomy for human chromosome 7 produces a growth-retardation syndrome, Silver–Russell syndrome, but we do not yet know of additional recurrent human chromosomal UPDs, other than those involving the chromosome 11p15 BWS region, that lead to overgrowth.

6. Cancer risk in BWS: epigenotype–phenotype correlation

Most cancers encountered in BWS are Wilms tumors (Bliek *et al*., 2004; Bliek *et al*., 2001; Gaston *et al*., 2001), but adrenal cortical carcinomas, hepatoblastomas, rhabdomyosarcomas, neuroblastomas, and other pediatric malignancies are also reported. Perhaps the most striking aspect of heterogeneity in BWS is the fact that cancer occurs only in a small subset, about 7%, of affected individuals. This percentage is lower than that seen in other well-studied human cancer syndromes, such as hereditary retinoblastoma, Li–Fraumeni syndrome, adenomatous polyposis coli and others, an observation that suggests either “incomplete penetrance” or molecular heterogeneity. The latter has proven to be correct, and a welcome recent advance has been the correlation of specific aspects of the BWS phenotype, including cancer risk, with the different molecular etiologies that underlie this disorder.

As early as 1999, a review of the literature suggested that predisposition to Wilms tumor in BWS is high in cases with Chr11p15 UPD or *H19* DMR GOM, and low or nonexistent in cases with *CDKN1C* mutations (Tycko, 1999). This conclusion was confirmed and strengthened shortly thereafter by a number of independent analyses of a large series, including a total of more than 250 molecularly characterized cases, which assessed these three classes of BWS, and also the large fourth category of affecteds with KvDMR1 LOM (Table 3). Frustratingly, the data are analyzed differently in each study, with some investigators choosing to describe the percentages of all tumor-bearing patients that have each molecular abnormality, and others describing the incidence of tumors in patients with a given molecular abnormality. The latter approach is easier to understand, so we have converted the data from each study to this uniform format in Table 3. The reader is also referred to a very recent combined European study of cancer risk in BWS (Bliek *et al*., 2004). These “meta-analyses” show that Wilms tumors, which in all of the series are the most common cancer, are increased in frequency only in those individuals that have BWS due to Chr11p15 UPD or *H19* DMR GOM. Since these two categories of BWS account for a minority of cases, this information nicely accounts for the limited Wilms tumor predisposition.

Larger series of cases will be needed to answer whether the rarer types of BWS-associated neoplasms, including adrenal cortical carcinoma, hepatoblastoma, and neuroblastoma are also specifically associated with a specific epigenotype in BWS. In fact, some of these non-Wilms neoplasms have been identified in BWS cases associated with the KvDMR1 imprinted domain: 2 neuroblastomas have been reported in children with *CDKN1C* mutations, and 2 hepatoblastomas, 2 rhabdomyosarcomas, and 1 gonadoblastoma were found in BWS-affecteds with KvDMR1 LOM (Lee *et al*., 1997a; Weksberg *et al*., 2001).

Table 3 Epigenotype–phenotype correlations in BWS

Study	Molecular category[a]	Tumor[b] (*n*) %	Exomphalos[c] (*n*) %	Conclusions
Bliek *et al*. (2001)[d]	*CDKN1C* mutation (*n* = 1[e])	Nd	Nd	Neoplasia associated with *H19* GOM and with UPD11p15, but not with KvDMR1 LOM.
	KvDMR1 LOM (*n* = 31)	(0) 0%	Nd	
	H19 GOM (*n* = 4)	(2) 50%	Nd	
	UPD 11p15 (*n* = 11)	(3) 27%	Nd	
	Undefined (*n* = 10)	(2) 20%	Nd	
Lam *et al*. (1999)	*CDKN1C* mutation (*n* = 13)[f]	(0) 0%	(11) 85%	Exomphalos, but not neoplasia, with *CDKN1C* mutation.
	H19 GOM (*n* = 5)	(0) 0%	(0) 0%	
DeBaun *et al*. (2002)[g]	*CDKN1C* mutation	Nd	Nd	Neoplasia with *H19* GOM and with UPD11p15, but not with KvDMR1 LOM. Exomphalos associated with KvDMR1 LOM.
	KvDMR1 LOM (*n* = 39)	(1) 3%	(33) 85%	
	H19 GOM (*n* = 10)	(4) 40%	(6) 60%	
	UPD 11p15 (*n* = 12)	(5) 42%	(8) 66%	
Engel *et al*. (2000)	*CDKN1C* mutation (*n* = 15)	(0) 0%	(13) 87%	Neoplasia with *H19* GOM or UPD11p15, but not with KvDMR1 LOM or *CDKN1C* mutation; exomphalos only in the latter two categories.
	KvDMR1 LOM (*n* = 29)	(0) 0%	(20) 69%	
	H19 GOM (*n* = 5)	(1) 20%	(0) 0%	
	UPD 11p15 (*n* = 22)	(2) 9%	(0) 0%	
Gaston *et al*. (2001)[h]	*CDKN1C* mutation (*n* = 2)	(1) 50%	(2) 100%	Neoplasia increased in frequency with *H19* GOM, UPD11p15 and trisomy; exomphalos predominantly in cases with KvDMR1 LOM or *CDKN1C* mutation.
	KvDMR1 LOM (*n* = 45)	(1) 2.2%	(18) 40%	
	H19 GOM (*n* = 11)	(3) 27.3%	(0) 0%	
	UPD 11p15 (*n* = 13)	(4) 30.8%	(2) 15.4%	
	Trisomy 11p15 (*n* = 2)	2	(0) 0%	
Weksberg *et al*. (2001)	*CDKN1C* mutation (*n* = 5)	(0) 0.0%	Nd	Wilms tumors in BWS cases with *H19* GOM or UPD 11p15; other tumor types (rhabdomyosarcoma, hepatoblastoma, gonadoblastoma) in cases with KvDMR1 LOM.
	KvDMR1 LOM (*n* = 35)	(5) 14.3%	Nd	

Table 3 (*continued*)

Study	Molecular category[a]	Tumor[b] (*n*) %	Exomphalos[c] (*n*) %	Conclusions
	H19 GOM ($n = 3$)	(1) 33.3%	Nd	
	UPD 11p15 ($n = 21$)	(6) 28.6%	Nd	

Nd: not described.

[a]*H19* GOM and KvDMR1 LOM refer to cases with these isolated molecular defects; cases with UPD are listed separately.

[b]Incidence of neoplasia within the indicated molecular category of BWS.

[c]Incidence of exomphalos or, generically, midline abdominal wall defects (exomphalos or umbilical hernia), within the indicated molecular category of BWS.

[d]This study reported 7 Wilms tumors and 1 hepatoblastoma in 113 individuals with BWS. Of these, 56 individuals were fully characterized by molecular criteria to assess tumor risk.

[e]This single *CDKN1C* mutant case, found in SSCP screening of 102 patients, had a maternally transmitted missense mutation, but was not further described. This case is not included in the percentages, which are based on 56 fully characterized cases.

[f]Six patients with Wilms tumor and somatic overgrowth, without classical BWS, were also analyzed. No *CDKN1C* mutations were found.

[g]Numbers are extracted from Table 2 of DeBaun *et al*. (2002) and percentages calculated. Please note that the table in this original publication lists percentages in a different format (% of all BWS-associated tumors accounted for by each category of BWS, rather than tumor incidence in a given category of BWS). Also, in extracting the data from that table, we have segregated the UPD cases, which have both *H19* GOM and KvDMR1 LOM from the non-UPD cases, which have one, but not both, of these molecular defects.

[h]Data are extracted from the text and Table 2 and Figure 3 of this publication. In extracting the data from Figure 3 of that publication, we have segregated the UPD cases, which have both *H19* GOM and KvDMR1 LOM from the non-UPD cases, which have one, but not both, of these molecular defects.

7. Developmental defects in BWS: epigenotype–phenotype correlation

The major category of BWS-affecteds with KvDMR1 LOM, or the rarer but seemingly equivalent genetic lesion of *CDKN1C* mutation, instead of developing Wilms tumors tend to show abdominal wall defects. As is true for the cancer correlations discussed above, the epigenotype–phenotype correlations for abdominal wall defects in all of the large case series analyzed to date are in complete agreement (Table 3).

8. Genetic and epigenetic mosaicism in BWS and Wilms tumor

Paternal UPD for chromosome 11p15 in BWS is found as tissue mosaicism (Slatter *et al*., 1994), consistent with the known lethal effect of complete paternal UPD for the orthologous chromosomal region in mice (Ferguson-Smith *et al*., 1991). But tissue mosaicism in BWS is not restricted to this molecular class and is also frequently observed in cases with GOM at the *H19* DMR or LOM at KvDMR1. In these cases, the epigenetic mosaicism is detected as incomplete

loss (KvDMR1) or gain (*H19* DMR and *H19* gene) of specific methylated bands on Southern blots after digesting the DNA with methylation-sensitive restriction enzymes. Whether asymmetrical tissue mosaicism can account for hemihyperplasia/hemihypertrophy in BWS is a difficult question. Mice with mosaic patUPD for the region corresponding to human chromosome 11p15 did not manifest asymmetrical growth and the hemihypertrophy seen in the two translocation cases of BWS with chromosome breakpoints in band 11p15.4 was not accounted for by mosaicism, since the translocations were constitutional (Alders *et al*., 2000).

As discussed above, tumor predisposition in BWS tracks with gain of DNA methylation at the imprinting center immediately upstream of the *H19* gene, and with paternal uniparental disomy. But this situation is not restricted to BWS: the same types of abnormalities are found as tissue mosaicism in the kidneys of a sizable group of sporadic Wilms tumor patients, who do not manifest the other features of BWS (Moulton *et al*., 1994; Moulton *et al*., 1996). In fact, this type of epigenetic mosaicism in the kidneys of children with Wilms tumor is a particularly clear example of a cancer-associated “field effect” determined by early developmental events (Tycko, 2003). What accounts for the de novo GOM in the H19 upstream DMR sequences early in development remains an important unsolved problem.

9. Causal role for epigenetic defects in BWS: discordance in twins

If epigenetic lesions are the primary cause of many cases of BWS, a strong prediction is that monozygotic twins might sometimes be found discordant for this disorder, that is, it should be possible to find rare but informative examples of individuals who share an identical DNA sequence throughout their genomes, but which are nonetheless discordant for the molecular and phenotypic features of BWS. In fact, the older literature contains several case reports of twins discordant for BWS, and a number of these examples were in monozygotic twins (“identical” being a misnomer in this situation) (Hall, 1996; Junien, 1992). The largest single study is a recent compilation of clinical and molecular data for 10 monozygotic twin pairs that were discordant for BWS (Weksberg *et al*., 2002). As predicted from the hypothesis that BWS is often purely epigenetic in etiology, each of the affected probands showed loss of DNA methylation at KvDMR1. These investigators postulated that there was a failure to maintain methylation of the paternal KvDMR1 allele early in postzygotic development, either coincident or shortly after the twinning event. In addition to this straightforward conclusion, there are two further aspects of BWS and twinning that remain to be explained: there is an excess of females over males among BWS-discordant twin pairs, and the frequency of monozygotic twins with BWS is greater than expected from the rate of twinning in the general population. A discussion of possible explanations for these observations, based on failure of proper subcellular localization and/or function of

the DNMT1 methyltransferase enzyme immediately prior to the twinning event, can be found in a recent review (Bestor, 2003).

10. BWS and assisted reproductive technology (ART)

Recent studies from BWS registries in the United States, England, and France have reported a total of 19 cases of this disorder occurring in children produced by in vitro fertilization or intracytoplasmic sperm injection (DeBaun *et al*., 2003; Gicquel *et al*., 2003; Maher *et al*., 2003). On the basis of the total number of cases surveyed, it has been estimated that slightly more than 4% of BWS in developed nations may be associated with assisted reproductive technology (ART). Since it is estimated that ART accounts for only about 1% of all births in these countries, there may be an increased risk of BWS after ART. The denominator in this calculation is uncertain, however, and larger surveys, in which the rate of BWS is studied in unselected pregnancies after ART, are needed. One prior study, concerning 61 pregnancies that were viable to term or late gestation following ART, found a single case of BWS (Olivennes *et al*., 2001).

While the statistical conclusions will need to be confirmed by larger studies, the available molecular data do provide some support for a causal connection between ART and epigenetic disorders. In particular, 18/19 of the reported ART-associated cases of BWS showed loss of methylation at KvDMR1, with only one case showing GOM at the *H19* DMR. This ratio deviates from the expected general distribution of molecular causes of BWS, and the data are suggestive of an ART-related failure of methylation of the maternal KvDMR1 allele, that is, an "oocyte problem" rather than a "sperm problem". Whether this mechanistic hypothesis can be validated in an experimental system, such as mice conceived by ART, remains to be seen, but further circumstantial support comes from reports of three cases of a second epigenetic disorder, Angelman syndrome, occurring after ART and all showing lack of appropriate DNA methylation of the maternal allele in the relevant chromosome 15 imprinting control region, the *SNRPN* DMR (Cox *et al*., 2002; Orstavik *et al*., 2003).

Acknowledgments

The authors thank Christine Gicquel for helpful information.

References

Alders M, Ryan A, Hodges M, Bliek J, Feinberg AP, Privitera O, Westerveld A, Little PF and Mannens M (2000) Disruption of a novel imprinted zinc-finger gene, ZNF215, in Beckwith-Wiedemann syndrome. *American Journal of Human Genetics*, **66**, 1473–1484.

Beckwith JB (1998) Vignettes from the history of overgrowth and related syndromes. *American Journal of Medical Genetics*, **79**, 238–248.

Beckwith JB, Kiviat NB and Bonadio JF (1990) Nephrogenic rests, nephroblastomatosis, and the pathogenesis of Wilms' tumor. *Pediatric Pathology*, **10**, 1–36.

Bestor TH (2003) Imprinting errors and developmental asymmetry. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **358**, 1411–1415.

Bliek J, Gicquel C, Maas SM, Gaston V, Le Bouc Y and Mannens M (2004) Epigenotyping as a tool for the prediction of tumor risk and tumor type in patients with Beckwith-Wiedemann syndrome (BWS). *The Journal of Pediatrics*, **145**, 796–799.

Bliek J, Maas SM, Ruijter JM, Hennekam RC, Alders M, Westerveld A and Mannens MM (2001) Increased tumour risk for BWS patients correlates with aberrant H19 and not KCNQ1OT1 methylation: occurrence of KCNQ1OT1 hypomethylation in familial cases of BWS. *Human Molecular Genetics*, **10**, 467–476.

Caspary T, Cleary MA, Perlman EJ, Zhang P, Elledge SJ and Tilghman SM (1999) Oppositely imprinted genes p57(Kip2) and igf2 interact in a mouse model for Beckwith-Wiedemann syndrome. *Genes & Development*, **13**, 3115–3124.

Charalambous M, Smith FM, Bennett WR, Crew TE, Mackenzie F and Ward A (2003) Disruption of the imprinted Grb10 gene leads to disproportionate overgrowth by an Igf2-independent mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8292–8297.

Chiao E, Fisher P, Crisponi L, Deiana M, Dragatsis I, Schlessinger D, Pilia G and Efstratiadis A (2002) Overgrowth of a mouse model of the Simpson-Golabi-Behmel syndrome is independent of IGF signaling. *Developmental Biology*, **243**, 185–206.

Cox GF, Burger J, Lip V, Mau UA, Sperling K, Wu BL and Horsthemke B (2002) Intracytoplasmic sperm injection may increase the risk of imprinting defects. *American Journal of Human Genetics*, **71**, 162–164.

DeBaun MR, King AA and White N (2000) Hypoglycemia in Beckwith-Wiedemann syndrome. *Seminars in Perinatology*, **24**, 164–171.

DeBaun MR, Niemitz EL and Feinberg AP (2003) Association of in vitro fertilization with Beckwith-Wiedemann syndrome and epigenetic alterations of LIT1 and H19. *American Journal of Human Genetics*, **72**, 156–160.

DeBaun MR, Niemitz EL, McNeil DE, Brandenburg SA, Lee MP and Feinberg AP (2002) Epigenetic alterations of H19 and LIT1 distinguish patients with Beckwith-Wiedemann syndrome with cancer and birth defects. *American Journal of Human Genetics*, **70**, 604–611.

Diaz-Meyer N, Day CD, Khatod K, Maher ER, Cooper W, Reik W, Junien C, Graham G, Algar E, Der Kaloustian VM, *et al*. (2003) Silencing of CDKN1C (p57KIP2) is associated with hypomethylation at KvDMR1 in Beckwith-Wiedemann syndrome. *Journal of Medical Genetics*, **40**, 797–801.

Eggenschwiler J, Ludwig T, Fisher P, Leighton PA, Tilghman SM and Efstratiadis A (1997) Mouse mutant embryos overexpressing IGF-II exhibit phenotypic features of the Beckwith-Wiedemann and Simpson-Golabi-Behmel syndromes. *Genes & Development*, **11**, 3128–3142.

Engel JR, Smallwood A, Harper A, Higgins MJ, Oshimura M, Reik W, Schofield PN and Maher ER (2000) Epigenotype-phenotype correlations in Beckwith-Wiedemann syndrome. *Journal of Medical Genetics*, **37**, 921–926.

Ferguson-Smith AC, Cattanach BM, Barton SC, Beechey CV and Surani MA (1991) Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature*, **351**, 667–670.

Fitzpatrick GV, Soloway PD and Higgins MJ (2002) Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nature Genetics*, **32**, 426–431.

Frank D, Fortino W, Clark L, Musalo R, Wang W, Saxena A, Li CM, Reik W, Ludwig T and Tycko B (2002) Placental overgrowth in mice lacking the imprinted gene Ipl. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7490–7495.

Gaston V, Le Bouc Y, Soupre V, Burglen L, Donadieu J, Oro H, Audry G, Vazquez MP and Gicquel C (2001) Analysis of the methylation status of the KCNQ1OT and H19 genes in leukocyte DNA for the diagnosis and prognosis of Beckwith-Wiedemann syndrome. *European Journal of Human Genetics: EJHG*, **9**, 409–418.

Gicquel C, Gaston V, Mandelbaum J, Siffroi JP, Flahault A and Le Bouc Y (2003) In vitro fertilization may increase the risk of Beckwith-Wiedemann syndrome related to the abnormal imprinting of the KCN1OT gene. *American Journal of Human Genetics*, **72**, 1338–1341.

Greenberg F, Stein F, Gresik MV, Finegold MJ, Carpenter RJ, Riccardi VM and Beaudet AL (1986) The Perlman familial nephroblastomatosis syndrome. *American Journal of Human Genetics*, **24**, 101–110.

Hall JG (1996) Twinning: mechanisms and genetic implications. *Current Opinion in Genetics & Development*, **6**, 343–347.

Hatada I, Ohashi H, Fukushima Y, Kaneko Y, Inoue M, Komoto Y, Okada A, Ohishi S, Nabetani A, Morisaki H, *et al*. (1996) An imprinted gene p57KIP2 is mutated in Beckwith-Wiedemann syndrome. *Nature Genetics*, **14**, 171–173.

Henry I, Bonaiti-Pellie C, Chehensse V, Beldjord C, Schwartz C, Utermann G and Junien C (1991) Uniparental paternal disomy in a genetic cancer-predisposing syndrome. *Nature*, **351**, 665–667.

Henry I, Puech A, Riesewijk A, Ahnine L, Mannens M, Beldjord C, Bitoun P, Tournade MF, Landrieu P and Junien C (1993) Somatic mosaicism for partial paternal isodisomy in Wiedemann-Beckwith syndrome: a post-fertilization event. *European Journal of Human Genetics: EJHG*, **1**, 19–29.

Horike S, Mitsuya K, Meguro M, Kotobuki N, Kashiwagi A, Notsu T, Schulz TC, Shirayoshi Y and Oshimura M (2000) Targeted disruption of the human LIT1 locus defines a putative imprinting control element playing an essential role in Beckwith-Wiedemann syndrome. *Human Molecular Genetics*, **9**, 2075–2083.

Hoyme HE, Seaver LH, Jones KL, Procopio F, Crooks W and Feingold M (1998) Isolated hemihyperplasia (hemihypertrophy): report of a prospective multicenter study of the incidence of neoplasia and review. *American Journal of Human Genetics*, **79**, 274–278.

Junien C (1992) Beckwith-Wiedemann syndrome, tumourigenesis and imprinting. *Current Opinion in Genetics & Development*, **2**, 431–438.

Koufos A, Grundy P, Morgan K, Aleck KA, Hadro T, Lampkin BC, Kalbakji A and Cavenee WK (1989) Familial Wiedemann-Beckwith syndrome and a second Wilms tumor locus both map to 11p15.5. *American Journal of Human Genetics*, **44**, 711–719.

Lam WW, Hatada I, Ohishi S, Mukai T, Joyce JA, Cole TR, Donnai D, Reik W, Schofield PN and Maher ER (1999) Analysis of germline CDKN1C (p57KIP2) mutations in familial and sporadic Beckwith-Wiedemann syndrome (BWS) provides a novel genotype-phenotype correlation. *Journal of Medical Genetics*, **36**, 518–523.

Lee MP, DeBaun M, Randhawa G, Reichard BA, Elledge SJ and Feinberg AP (1997a) Low frequency of p57KIP2 mutation in Beckwith-Wiedemann syndrome. *American Journal of Human Genetics*, **61**, 304–309.

Lee MP, Hu RJ, Johnson LA and Feinberg AP (1997b) Human KVLQT1 gene shows tissue-specific imprinting and encompasses Beckwith-Wiedemann syndrome chromosomal rearrangements. *Nature Genetics*, **15**, 181–185.

Lee MP, DeBaun MR, Mitsuya K, Galonek HL, Brandenburg S, Oshimura M and Feinberg AP (1999) Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5203–5208.

Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A and Tilghman SM (1995) Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature*, **375**, 34–39.

Li C-M, Guo M, Borczuk A, Powell CA, Wei M, Thaker HM, Friedman R, Klein U and Tycko B (2002) Gene expression in Wilms tumors mimics the earliest committed stage in the metanephric mesenchymal-epithelial transition. *American Journal of Pathology*, **160**, 2181–2190.

Maher ER, Brueton LA, Bowdin SC, Luharia A, Cooper W, Cole TR, Macdonald F, Sampson JR, Barratt CL, Reik W, *et al*. (2003) Beckwith-Wiedemann syndrome and assisted reproduction technology (ART). *Journal of Medical Genetics*, **40**, 62–64.

Mannens M, Alders M, Redeker B, Bliek J, Steenman M, Wiesmeyer C, de Meulemeester M, Ryan A, Kalikin L, Voute T, *et al*. (1996) Positional cloning of genes involved in the Beckwith-Wiedemann syndrome, hemihypertrophy, and associated childhood tumors. *Medical and Pediatric Oncology*, **27**, 490–494.

Moon YS, Smas CM, Lee K, Villena JA, Kim KH, Yun EJ and Sul HS (2002) Mice lacking paternally expressed Pref-1/Dlk1 display growth retardation and accelerated adiposity. *Molecular and Cellular Biology*, **22**, 5585–5592.

Moore T and Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics: TIG*, **7**, 45–49.

Moulton T, Chung WY, Yuan L, Hensle T, Waber P, Nisen P and Tycko B (1996) Genomic imprinting and Wilms' tumor. *Medical and Pediatric Oncology*, **27**, 476–483.

Moulton T, Crenshaw T, Hao Y, Moosikasuwan J, Lin N, Dembitzer F, Hensle T, Weiss L, McMorrow L, Loew T, *et al*. (1994) Epigenetic lesions at the H19 locus in Wilms' tumor patients. *Nature Genetics*, **7**, 440–447.

Niemitz EL, DeBaun MR, Fallon J, Murakami K, Kugoh H, Oshimura M and Feinberg AP (2004) Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. *American Journal of Human Genetics*, **75**, 844–849.

O'Keefe D, Dao D, Zhao L, Sanderson R, Warburton D, Weiss L, Anyane-Yeboa K and Tycko B (1997) Coding mutations in p57KIP2 are present in some cases of Beckwith-Wiedemann syndrome but are rare or absent in Wilms tumors. *American Journal of Human Genetics*, **61**, 295–303.

Ogawa O, Becroft DM, Morison IM, Eccles MR, Skeen JE, Mauger DC and Reeve AE (1993) Constitutional relaxation of insulin-like growth factor II gene imprinting associated with Wilms' tumour and gigantism. *Nature Genetics*, **5**, 408–412.

Olivennes F, Mannaerts B, Struijs M, Bonduelle M and Devroey P (2001) Perinatal outcome of pregnancy after GnRH antagonist (ganirelix) treatment during ovarian stimulation for conventional IVF or ICSI: a preliminary report. *Human Reproduction*, **16**, 1588–1591.

Orstavik KH, Eiklid K, van der Hagen CB, Spetalen S, Kierulf K, Skjeldal O and Buiting K (2003) Another case of imprinting defect in a girl with Angelman syndrome who was conceived by intracytoplasmic semen injection. *American Journal of Human Genetics*, **72**, 218–219.

Ping AJ, Reeve AE, Law DJ, Young MR, Boehnke M and Feinberg AP (1989) Genetic linkage of Beckwith-Wiedemann syndrome to 11p15. *American Journal of Human Genetics*, **44**, 720–723.

Reeve AE (1996) Role of genomic imprinting in Wilms' tumour and overgrowth disorders. *Medical and Pediatric Oncology*, **27**, 470–475.

Rio M, Clech L, Amiel J, Faivre L, Lyonnet S, Le Merrer M, Odent S, Lacombe D, Edery P, Brauner R, *et al*. (2003) Spectrum of NSD1 mutations in Sotos and Weaver syndromes. *Journal of Medical Genetics*, **40**, 436–440.

Salas M, John RM, Saxena A, Barton S, Frank D, Fitzpatrick GV, Higgins MJ and Tycko B (2004) Placental growth retardation due to loss of imprinting of Phlda2. *Mechanisms of Development*, **121**, 1199–1210.

Slatter RE, Elliott M, Welham K, Carrera M, Schofield PN, Barton DE and Maher ER (1994) Mosaic uniparental disomy in Beckwith-Wiedemann syndrome. *Journal of Medical Genetics*, **31**, 749–753.

Smilinich NJ, Day CD, Fitzpatrick GV, Caldwell GM, Lossie AC, Cooper PR, Smallwood AC, Joyce JA, Schofield PN, Reik W, *et al*. (1999) A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 8064–8069.

Sparago A, Cerrato F, Vernucci M, Ferrero GB, Silengo MC and Riccio A (2004) Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. *Nature Genetics*, **36**, 958–960.

Thorvaldsen JL and Bartolomei MS (2000) Molecular biology. Mothers setting boundaries. *Science*, **288**, 2145–2146.

Tycko B (1999) Genomic imprinting and cancer. *Results and Problems in Cell Differentiation*, **25**, 133–169.

Tycko B (2003) Genetic and epigenetic mosaicism in cancer precursor tissues. *Annals of the New York Academy of Sciences*, **983**, 43–54.

Tycko B and Morison IM (2002) Physiological functions of imprinted genes. *Journal of Cellular Physiology*, **192**, 245–258.

Weksberg R, Nishikawa J, Caluseriu O, Fei YL, Shuman C, Wei C, Steele L, Cameron J, Smith A, Ambus I, *et al*. (2001) Tumor development in the Beckwith-Wiedemann syndrome is associated with a variety of constitutional molecular 11p15 alterations including imprinting defects of KCNQ1OT1. *Human Molecular Genetics*, **10**, 2989–3000.

Weksberg R, Shuman C, Caluseriu O, Smith AC, Fei YL, Nishikawa J, Stockley TL, Best L, Chitayat D, Olney A, *et al*. (2002) Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Human Molecular Genetics*, **11**, 1317–1325.

Specialist Review

Imprinting at the *GNAS* locus and endocrine disease

Lee S. Weinstein, Min Chen, Akio Sakamoto and Jie Liu
National Institutes of Diabetes, Digestive, and Kidney Diseases, National Institute of Health, Bethesda, MD, USA

1. Introduction

$G_s\alpha$ is a ubiquitously expressed G protein α-subunit that couples numerous receptors to the enzyme adenylyl cyclase and is therefore required for the intracellular cyclic AMP response to many hormones and other signaling molecules (Weinstein *et al*., 2001; Weinstein, 2004). *GNAS*, the gene encoding $G_s\alpha$ that is located at 20q13 is a complex imprinted gene that generates multiple gene products through the use of alternative promoters and first exons that splice onto a common set of downstream exons (exons 2-13, see Figure 1) (*see* Article 30, **Alternative splicing: conservation and function**, Volume 3). The mouse ortholog *Gnas* is located within a syntenic region within chromosome 2 and has very similar overall structure and imprinting patterns (*see* Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1, Article 20, **Synteny mapping**, Volume 3, Article 47, **The mouse genome sequence**, Volume 3, and Article 48, **Comparative sequencing of vertebrate genomes**, Volume 3). The most upstream promoter generates transcripts for the chromogranin-like protein NESP55, which is structurally and functionally unrelated to $G_s\alpha$. The NESP55 coding region is fully encoded by its specific upstream exon, and $G_s\alpha$ exons 2-13 form part of the 3′ untranslated region of NESP55 transcripts (Ischia *et al*., 1997). The next promoter generates transcripts encoding the neuroendocrine-specific $G_s\alpha$ isoform XLαs, which is structurally identical to $G_s\alpha$ except for the presence of a long amino-terminal extension encoded by its specific first exon (Klemke *et al*., 2000; Pasolli *et al*., 2000). NESP55 and XLαs are oppositely imprinted (Hayward *et al*., 1998a,b; Peters *et al*., 1999). NESP55 is expressed only from the maternal allele and its promoter region is DNA methylated on the paternal allele, while XLαs is only expressed from the paternal allele and its promoter is methylated on the maternal allele (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1 and Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1). The XLαs promoter region appears to contain a primary imprint mark where methylation is established during gametogenesis (Coombes

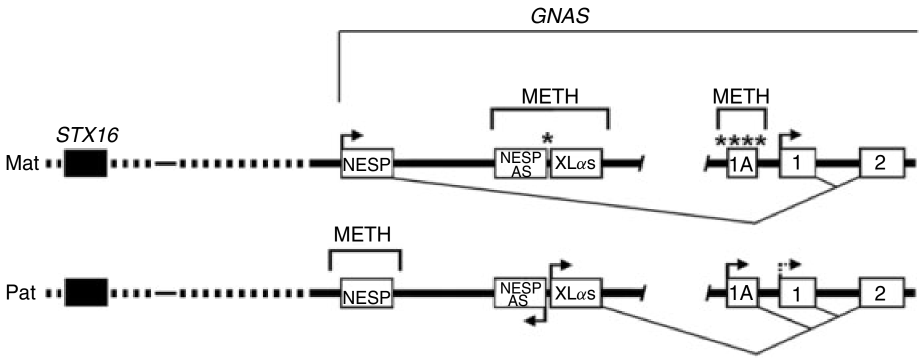


Figure 1 General organization and epigenotypes of the *GNAS* (and *Gnas*) locus. Maternal (Mat) and paternal (Pat) alleles of *GNAS* are depicted with alternative first exons for NESP55 (NESP), XLαs, exon 1A-specific untranslated mRNAs, and $G_s\alpha$ (exon 1) shown spliced to a common exon 2. Regions that are differentially methylated (METH) are noted above and splicing patterns are shown below in each panel. Transcriptionally active promoters are indicated by horizontal arrows in the direction of transcription. The dashed arrow for the paternal $G_s\alpha$ (exon 1) promoter indicates that $G_s\alpha$ expression from the paternal allele is suppressed in some tissues. The first exon for paternally expressed antisense transcripts is labeled as NESPAS. The location of two primary methylation imprint marks in *Gnas* is noted with asterisks. Common downstream exons 3-13 in the sense direction and downstream exons of the antisense transcript are not shown. A 3-kb deletion mutation within the *STX16* gene (black box), located 220 kb upstream of *GNAS* exon 1A, is associated with loss of exon 1A imprinting and familial PHP1B. The diagram is not drawn to scale

et al., 2003). This region also generates paternal-specific antisense transcripts that may be required for NESP55 imprinting, which has been shown in mice to not be established until after implantation (Hayward and Bonthron, 2000; Liu *et al*., 2000b; Wroe *et al*., 2000; *see also* Article 27, **Noncoding RNAs in mammals**, Volume 3).

The XLαs promoter is located ~35 kb upstream of the $G_s\alpha$ promoter (Figure 1). Studies in both mice and humans have shown $G_s\alpha$ to be imprinted in a tissue-specific manner. $G_s\alpha$ is equally expressed from both alleles in most tissues, but is expressed primarily from the maternal allele in certain hormone-target tissues (renal proximal tubules, a site of parathyroid hormone (PTH) action, thyroid, pituitary, ovary) (Germain-Lee *et al*., 2002; Hayward *et al*., 2001; Liu *et al*., 2003; Mantovani *et al*., 2002; *see also* Article 36, **Variable expressivity and epigenetics**, Volume 1). In some tissues (e.g., thyroid), expression from the paternal allele is only partially suppressed. The $G_s\alpha$ promoter is located within a CpG island that is euchromatic and unmethylated on both parental alleles despite the allele-specific differences in expression (Liu *et al*., 2000b; Sakamoto *et al*., 2004; *see also* Article 28, **The distribution of genes in human genome**, Volume 3 and Article 33, **Transcriptional promoters**, Volume 3). It has been shown in mice that allele-specific differences in $G_s\alpha$ gene expression are associated with differences in the extent of methylation of the lysine 4 residue of histone H3 within its first exon, a parameter that has been shown to be correlated with transcriptional activity in nonmammalian species (Sakamoto *et al*., 2004; *see also* Article 27, **The histone code and epigenetic inheritance**, Volume 1). Just upstream of the $G_s\alpha$ promoter is a differentially methylated region (DMR), which appears to be a primary imprint mark, as methylation of the maternal allele in this region is

established during oogenesis and maintained throughout development (Liu *et al*., 2000a,b). Within this region is another alternative promoter and first exon (exon 1A) that generates nontranslated mRNA transcripts only from the paternal allele (*see* Article 27, **Noncoding RNAs in mammals**, Volume 3). Both clinical and mouse studies suggest that this region (the exon 1A DMR) is necessary for tissue-specific imprinting of $G_s\alpha$ (see below).

2. Albright hereditary osteodystrophy

Albright hereditary osteodystrophy (AHO) is a congenital syndrome that is characterized by obesity, short stature, brachydactyly, subcutaneous ossifications, and, in some cases, neurobehavioral deficits (Weinstein *et al*., 2001). In most cases, AHO is associated with heterozygous loss-of-function mutations within or surrounding the $G_s\alpha$ coding exons, which may disrupt $G_s\alpha$ mRNA expression or protein expression or function (Aldred and Trembath, 2000). Consistent with the presence of a heterozygous null mutation, $G_s\alpha$ protein levels or bioactivity are reduced by $\sim$50% in membranes isolated from easily accessible tissues, such as erythrocytes and fibroblasts (Levine *et al*., 1983). Although these mutations could disrupt expression of other *GNAS* gene products, several lines of evidence point to $G_s\alpha$ haploinsufficiency as the likely molecular defect that is the cause of AHO. First, while all other known *GNAS* gene products are only expressed from one parental allele, the AHO phenotype is present in all patients with $G_s\alpha$ mutations, regardless of whether the mutation is inherited maternally or paternally. Second, some missense mutations are not predicted to affect NESP55 expression (Rickard and Wilson, 2003) and NESP55 deficiency in some patients with pseudohypoparathyroidism type 1B (PHPIB) does not lead to AHO (Liu *et al*., 2000a). Finally, mutations within $G_s\alpha$ exon 1, which would be predicted to only affect $G_s\alpha$ expression, also lead to the AHO phenotype. Some patients with $G_s\alpha$ null mutations develop a more severe ectopic ossification syndrome known as progressive osseous heteroplasia (Kaplan and Shore, 2000).

AHO patients who inherit the disease maternally (or have a de novo mutation on the maternal allele) also develop resistance to several hormones whose receptors activate $G_s\alpha$, including parathyroid hormone (PTH), thyrotropin, and gonadotropins, a condition also referred to as pseudohypoparathyroidism type 1A (PHP1A) (Davies and Hughes, 1993; Weinstein *et al*., 2001; Weinstein, 2004). In contrast, patients with paternal mutations develop AHO but do not develop hormone resistance, a condition also referred to as pseudopseudohypoparathyroidism. Studies in *Gnas* knockout mice showed a similar imprinted pattern of inheritance of renal PTH resistance and confirmed that this was due to tissue-specific $G_s\alpha$ imprinting (Yu *et al*., 1998; *see also* Article 38, **Mouse models**, Volume 3 and Article 41, **Mouse mutagenesis and gene function**, Volume 3). In renal proximal tubules (a major site of PTH action in the kidney), mutation of the active maternal allele leads to $G_s\alpha$ deficiency and PTH resistance (Figure 2). In contrast, mutation of the inactive paternal allele has little effect on $G_s\alpha$ expression or PTH sensitivity. The highly tissue-specific nature of the imprinting explains why the parent-of-origin effects are limited to a few hormone-signaling defects. This may also explain why

these patients fail to develop hormone resistance in other hormone-target tissues where $G_s\alpha$ is not imprinted (Weinstein *et al*., 2000).

3. Pseudohypoparathyroidism type 1B

PHPIB patients develop renal PTH resistance but do not develop the AHO phenotype (Silve *et al*., 1986). This disorder usually occurs sporadically but occasionally is familial. Like in PHPIA, the urinary cyclic AMP response to administered PTH is markedly reduced in PHPIB (Levine *et al*., 1983), implicating *GNAS* as a possible candidate disease gene for PHPIB as well. Despite this, erythrocyte $G_s\alpha$ levels and activity are normal in PHPIB (Levine *et al*., 1983; Silve *et al*., 1986), ruling out typical $G_s\alpha$ null mutations as the cause of PHPIB. Despite this, the familial PHPIB gene was mapped to 20q13 in the vicinity of *GNAS* (Juppner *et al*., 1998). Moreover, subjects within these kindreds only developed PTH resistance when the trait was inherited maternally, an inheritance pattern for PTH resistance similar to that observed in AHO patients with $G_s\alpha$ null mutations.

One mechanism that could explain the clinical features of PHPIB is the presence of a *GNAS* imprinting defect that results in both alleles having a paternal imprinting pattern or epigenotype. As $G_s\alpha$ is normally expressed primarily from the maternal allele in renal proximal tubules, the presence of a paternal epigenotype on both parental alleles would be expected to result in $G_s\alpha$ deficiency in proximal tubules and PTH resistance. In contrast, there should be little effect on $G_s\alpha$ expression in most other tissues, where $G_s\alpha$ is normally equally expressed from both parental alleles.

In most cases of familial PHPIB, a 3-kb deletion mutation is present within the closely linked *STX16* gene located upstream of *GNAS* in association with loss of maternal imprinting (methylation) of the exon 1A DMR (Bastepe *et al*., 2003; Liu *et al*., 2005b). Presumably, the deletion includes an imprinting control region that is critical for the establishment of exon 1A imprinting during oogenesis. In this scenario, maternal inheritance of the deletion leads to a paternal epigenotype within the exon 1A DMR on both parental alleles, which alters $G_s\alpha$ expression and hormone signaling (see below), while paternal inheritance of the deletion has no effect on exon 1A imprinting or $G_s\alpha$ expression, resulting in a clinically silent carrier state. The *STX16* deletion has no effect on imprinting of the intervening NESP55 and XLαs promoter regions, which suggests that exon 1A and NESP55/XLαs reside within independently regulated imprinted domains. In one family, PHPIB resulted from maternal inheritance of a $G_s\alpha$ missense mutation (deletion of isoleucine 382) that leads to selective uncoupling of $G_s\alpha$ from the PTH receptor (Wu *et al*., 2001).

Sporadic PHPIB is associated with *GNAS* imprinting defects in the absence of the *STX16* or any other known mutation (Bastepe *et al*., 2001b; Bastepe *et al*., 2003; Jan de Beur *et al*., 2003; Liu *et al*., 2000a; Liu *et al*., 2005b). In some sporadic cases, patients only have loss of exon 1A imprinting without the *STX16* mutation. Others have an imprinting defect where there is a paternal epigenotype in both parental alleles throughout the whole *GNAS* locus, and some have an imprinting defect involving exon 1A, NESP55, and the XLαs promoter, but not the XLαs

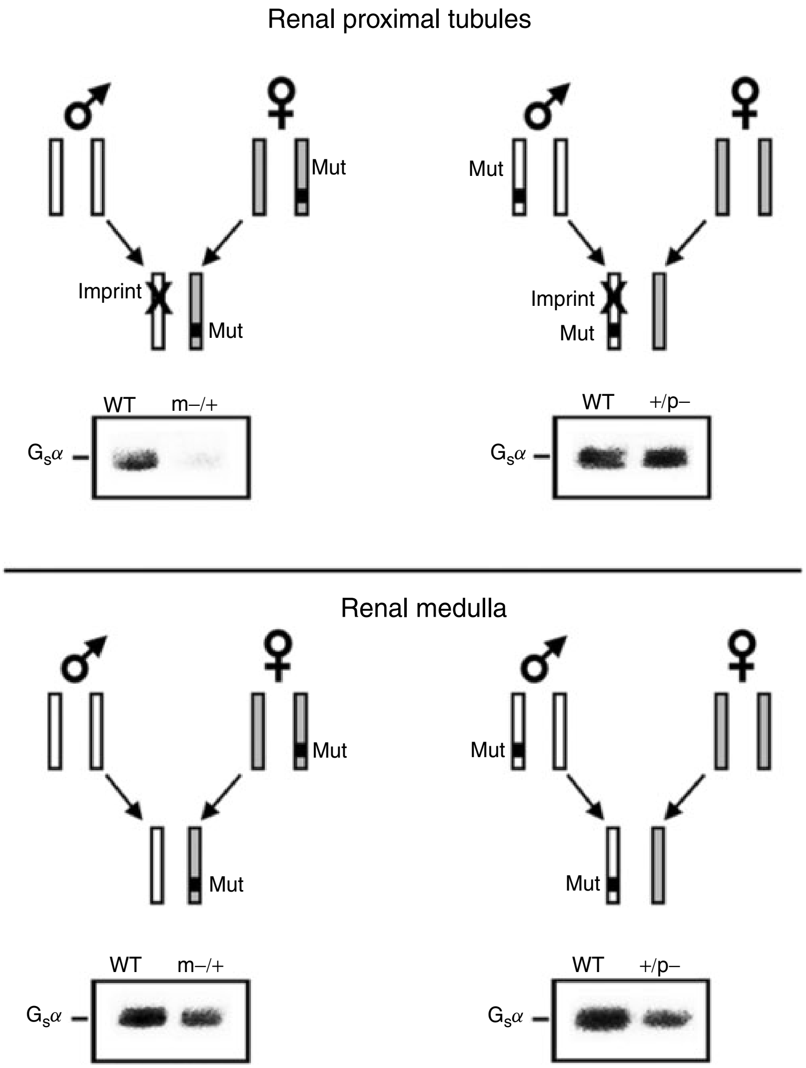


Figure 2 The effect of tissue-specific imprinting and heterozygous null mutations on $G_s\alpha$ expression in different tissues. In renal proximal tubules (upper panels), $G_s\alpha$ is paternally imprinted (denoted with X). Mutation (Mut) on the active maternal allele (left panel) leads to $G_s\alpha$ deficiency and PTH resistance while mutation on the inactive paternal allele (right panel) has little effect on $G_s\alpha$ expression or PTH sensitivity. Immunoblots of renal cortical membranes isolated from wild-type mice (WT) and mice with disruption of the *Gnas* maternal (m−/+) and paternal (+/p−) alleles respectively (Yu *et al*., 1998) confirms this imprinted pattern of $G_s\alpha$ expression. In most other tissues (lower panels), there is no $G_s\alpha$ imprinting and therefore both maternal and paternal mutations lead to ∼50% loss of $G_s\alpha$ expression (haploinsufficiency), as shown in immunoblots of renal inner medulla membranes from the same mice. This $G_s\alpha$ haploinsufficiency probably is the underlying molecular defect that leads to AHO in all patients with heterozygous $G_s\alpha$ mutations (both maternal and paternal) (Adapted from Weinstein *et al*. (2001) Endocrine manifestations of stimulatory G protein ∝-subunit mutations and the role of genomic imprinting. *Endocrine Reviews*, **22**, 675–705. © The Endocrine Society)

first exon. Patients with abnormal imprinting of the NESP55 promoter have loss of NESP55 expression due to methylation of the NESP55 promoter on both parental alleles without any obvious further clinical manifestations (Liu *et al*., 2000a). It is unclear whether the *GNAS* epigenetic defects in these sporadic mutations are the result of other underlying (and possibly de novo) mutations or rather result from the rare occurrence of the imprinting process going awry. One case of paternal uniparental disomy of 20q was associated with the expected *GNAS* imprinting abnormality along with PTH resistance and other clinical manifestations (Bastepe *et al*., 2001a; *see also* Article 19, **Uniparental disomy**, Volume 1).

The fact that PTH resistance in PHPIB is virtually always associated with loss of exon 1A imprinting provides strong support that the exon 1A DMR is a critical element in the regulation of tissue-specific $G_s\alpha$ imprinting. The fact that exon 1A and $G_s\alpha$ mRNA transcripts have a very similar pattern of tissue distribution (Liu, 2000b) makes it very unlikely that there is "competition" between these two promoters or that exon 1A transcripts themselves are involved in the $G_s\alpha$ imprinting mechanism. We have proposed that tissue-specific $G_s\alpha$ imprinting results from the presence of a cis-acting negative regulatory element (e.g., silencer or insulator) within the exon 1A DMR that is both tissue-specific and methylation-sensitive (Liu *et al*., 2000a; Weinstein *et al*., 2001). In the example shown in Figure 3, the exon 1A DMR contains a silencer. In renal proximal tubules, a tissue-specific repressor protein binds to the silencer and suppresses $G_s\alpha$ expression on the paternal allele, but is unable to bind to the silencer on the maternal allele because the site is methylated, allowing the maternal $G_s\alpha$ promoter to remain transcriptionally active. In most other tissues, exon 1A is still differentially methylated but there is no $G_s\alpha$ imprinting because the repressor is not expressed. In PHPIB, the methylation of exon 1A on the maternal allele is absent. In renal proximal tubules, the repressor can now bind to and inhibit $G_s\alpha$ expression from both parental alleles, resulting in $G_s\alpha$ deficiency and PTH resistance. In most other tissues, the loss of methylation has no effect on $G_s\alpha$ expression, because the repressor is absent. Support for this model is provided by studies in mice with deletion of the exon 1A DMR. Paternal, but not maternal, deletion of the exon 1A DMR results in $G_s\alpha$ overexpression in proximal tubules and lower circulating levels of PTH, indicative of increased PTH sensitivity, with little effect on $G_s\alpha$ expression in other tissues where $G_s\alpha$ is normally not imprinted (Williamson *et al*., 2004; Liu *et al*., 2005a; *see also* Article 38, **Mouse models**, Volume 3 and Article 41, **Mouse mutagenesis and gene function**, Volume 3).

$G_s\alpha$ is also partially imprinted in human thyroid, with ∼70–75% of the $G_s\alpha$ transcripts derived from the maternal allele (Germain-Lee *et al*., 2002; Liu *et al*., 2003; Mantovani *et al*., 2002). This explains why PHPIA patients (in whom the maternal allele is mutated) typically have mild to moderate hypothyroidism due to TSH resistance. One might predict that PHPIB patients would also have thyroid-specific $G_s\alpha$ deficiency, although to a lesser extent than in PHPIA, due to the fact that both alleles have a paternal epigenotype and therefore functionally behave as paternal alleles. Although TSH resistance has been considered to not be a feature of PHPIB, recent studies show that a significant number of patients with PHPIB and

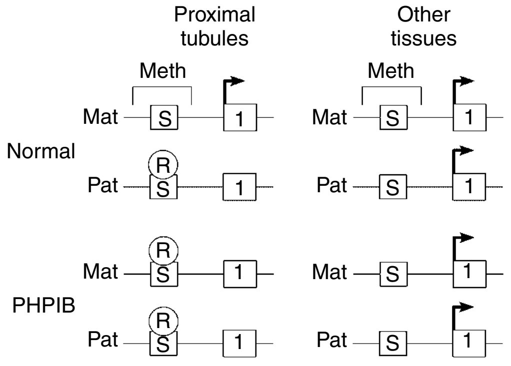


Figure 3 Proposed model for tissue-specific $G_s\alpha$ imprinting and pathogenesis of PHP1B. Maternal (Mat) and paternal (Pat) alleles of the exon 1A DMR-$G_s\alpha$ promoter region are depicted with a cis-acting silencer (S) within the exon 1A DMR and the first $G_s\alpha$ exon labeled as "1". Normally (upper panels), the silencer is methylated (Meth) on the maternal allele. In proximal tubules (left hand panel), a tissue-specific trans-acting repressor (R) binds to the silencer and suppresses $G_s\alpha$ expression on the paternal allele. The repressor fails to bind to the maternal allele due to methylation of the silencer, allowing the maternal $G_s\alpha$ promoter to remain active. In most other tissues (right hand panel), the repressor is not expressed and therefore the $G_s\alpha$ promoter remains transcriptionally active on both parental alleles. In PHP1B (lower panels), methylation is absent, allowing the repressor to bind to both alleles in proximal tubules, resulting in $G_s\alpha$ deficiency in this tissue. In most other tissues, $G_s\alpha$ expression is not affected owing to absence of the repressor

abnormal *GNAS* imprinting have evidence for borderline or mild TSH resistance (Bastepe *et al*., 2001a; Liu *et al*., 2003).

4. Potential role of $G_s\alpha$ imprinting on the clinical effects of activating $G_s\alpha$ mutations

Like all G proteins, $G_s\alpha$ is activated by receptors through release of bound GDP and binding of ambient GTP and is deactivated by an intrinsic GTPase that hydrolyzes bound GTP to GDP. Missense mutations in $G_s\alpha$ residues known to be catalytically important for the GTPase activity (arginine 201, glutamine 227) have been shown to be constitutively activating, and such dominant-acting somatic mutations are present in ~40% of growth hormone-secreting pituitary tumors (leading to acromegaly) and a small number of thyroid and other endocrine tumors (Lyons *et al*., 1990). More widespread distribution of arginine 201 mutations that presumably result from somatic mutation occurring during early embryonic development leads to the McCune–Albright syndrome, a condition characterized by the presence of precocious puberty, fibrous dysplasia of bone, café-au-lait skin lesions, and other endocrine and nonendocrine manifestations (Weinstein *et al*., 1991; *see also* Article 18, **Mosaicism**, Volume 1). Somatic arginine 201 mutations are also found in fibrous dysplasia of bone in patients who do not have the other manifestations of McCune–Albright syndrome. Recent studies show that $G_s\alpha$ is imprinted in the pituitary and that in almost all growth hormone-secreting pituitary

tumors harboring a $G_s\alpha$ activating mutation (both sporadic and in the setting of McCune–Albright syndrome), the mutation was present on the maternal allele (Hayward *et al*., 2001; Mantovani *et al*., 2004). These results suggest that the clinical manifestations of $G_s\alpha$-activating mutations may be affected by which parental allele harbors the mutation in tissues where there are allele-specific differences in $G_s\alpha$ expression.

References

Aldred MA and Trembath RC (2000) Activating and inactivating mutations in the human *GNAS1* gene. *Human Mutation*, **16**, 183–189.

Bastepe M, Lane AH and Jüppner H (2001a) Paternal uniparental disomy of chromosome 20q- and the resulting changes in *GNAS1* methylation- as a plausible cause of pseudohypoparathyroidism. *American Journal of Human Genetics*, **68**, 1283–1289.

Bastepe M, Pincus JE, Sugimoto T, Tojo K, Kanatani M, Azuma Y, Kruse K, Rosenbloom AL, Koshiyama H and Jüppner H (2001b) Positional dissociation between the genetic mutation responsible for pseudohypoparathyroidism type Ib and the associated methylation defect at exon A/B: Evidence for a long-range regulatory element within the imprinted *GNAS1* locus. *Human Molecular Genetics*, **10**, 1231–1241.

Bastepe M, Frohlich LF, Hendy GN, Indridason OS, Josse RG, Koshiyama H, Korkko J, Nakamoto JM, Rosenbloom AL, Slyper AH, *et al*. (2003) Autosomal dominant pseudohypoparathyroidism type Ib is associated with a heterozygous microdeletion that likely disrupts a putative imprinting control element of *GNAS*. *Journal of Clinical Investigation*, **112**, 1255–1263.

Coombes C, Arnaud P, Gordon E, Dean W, Coar EA, Williamson CM, Feil R, Peters J and Kelsey G (2003) Epigenetic properties and identification of an imprint mark in the *Nesp-Gnasxl* domain of the mouse *Gnas* imprinted locus. *Molecular and Cellular Biology*, **23**, 5475–5488.

Davies SJ and Hughes HE (1993) Imprinting in Albright's hereditary osteodystrophy. *Journal of Medical Genetics*, **30**, 101–103.

Germain-Lee EL, Ding C-L, Deng Z, Crane JL, Saji M, Ringel MD and Levine MA (2002) Paternal imprinting of Gαs in the human thyroid as the basis of TSH resistance in pseudohypoparathyroidism type 1a. *Biochemical and Biophysical Research Communications*, **296**, 67–72.

Hayward BE, Kamiya M, Strain L, Moran V, Campbell R, Hayashizaki Y and Bonthron DT (1998a) The human *GNAS1* gene is imprinted and encodes distinct paternally and biallelically expressed G proteins. *Proceedings of the National Academy of Sciences United States of America*, **95**, 10038–10043.

Hayward BE, Moran V, Strain L and Bonthron DT (1998b) Bidirectional imprinting of a single gene: *GNAS1* encodes maternally, paternally, and biallelically derived proteins. *Proceedings of the National Academy of Sciences United States of America*, **95**, 15475–15480.

Hayward BE and Bonthron DT (2000) An imprinted antisense transcript at the human *GNAS1* locus. *Human Molecular Genetics*, **9**, 835–841.

Hayward BE, Barlier A, Korbonits M, Grossman AB, Jacquet P, Enjalbert A and Bonthron DT (2001) Imprinting of the $G_s\alpha$ gene *GNAS1* in the pathogenesis of acromegaly. *Journal of Clinical Investigation*, **107**, R31–R36.

Ischia R, Lovisetti-Scamihorn P, Hogue-Angeletti R, Wolkersdorfer M, Winkler H and Fischer-Colbrie R (1997) Molecular cloning and characterization of NESP55, a novel chromogranin-like precursor of a peptide with 5-HT_{1B} receptor antagonist activity. *Journal of Biological Chemistry*, **272**, 11657–11662.

Jan de Beur S, Ding C, Germain-Lee E, Cho J, Maret A and Levine MA (2003) Discordance between genetic and epigenetic defects in pseudohypoparathyroidism type 1b revealed by inconsistent loss of maternal imprinting at *GNAS1*. *American Journal of Human Genetics*, **73**, 314–322.

Juppner H, Schipani E, Bastepe M, Cole DE, Lawson ML, Mannstadt M, Hendy GN, Plotkin H, Koshiyama H, Koh T, *et al*. (1998) The gene responsible for pseudohypoparathyroidism type Ib is paternally imprinted and maps in four unrelated kindreds to chromosome 20q13.3. *Proceedings of the National Academy of Sciences United States of America*, **95**, 11798–11803.

Kaplan FS and Shore EI (2000) Progressive osseous heteroplasia. *Journal of Bone and Mineral Research*, **15**, 2084–2094.

Klemke M, Pasolli HA, Kehlenbach RH, Offermanns S, Schultz G and Huttner WB (2000) Characterization of the extra-large G protein α-subunit XLαs. II. Signal transduction properties. *Journal of Biological Chemistry*, **275**, 33633–33640.

Levine MA, Downs RW Jr, Moses AM, Breslau NA, Marx SJ, Lasker RD, Rizzoli RE, Aurbach GD and Spiegel AM (1983) Resistance to multiple hormones in patients with pseudohypoparathyroidism. Association with deficient activity of guanine nucleotide regulatory protein. *American Journal of Medicine*, **74**, 545–556.

Liu J, Chen M, Deng C, Bourch'his D, Nealon JG, Erlichman B, Bestor TH and Weinstein LS (2005a) Identification of the control region for tissue-specific imprinting of the stimulatory G protein α-subunit. *Proceedings of the National Academy of Sciences United States of America*, **102**, 5513–5518.

Liu J, Erlichman B and Weinstein LS (2003) The stimulatory G protein α-subunit $G_s\alpha$ is imprinted in human thyroid glands: Implications for thyroid function in pseudohypoparathyroidism types 1A and 1B. *Journal of Clinical Endocrinology and Metabolism*, **88**, 4336–4341.

Liu J, Litman D, Rosenberg MJ, Yu S, Biesecker LG and Weinstein LS (2000a) A *GNAS1* imprinting defect in pseudohypoparathyroidism type IB. *Journal of Clinical Investigation*, **106**, 1167–1174.

Liu J, Nealon JG and Weinstein LS (2005b) Distinct patterns of abnormal GNAS imprinting in familial and sporadic pseudohypoparathyroidism type IB. *Human Molecular Genetics*, **14**, 95–102.

Liu J, Yu S, Litman D, Chen W and Weinstein LS (2000b) Identification of a methylation imprint mark within the mouse *Gnas* locus. *Molecular and Cellular Biology*, **20**, 5808–5817.

Lyons J, Landis CA, Harsh G, Vallar L, Grünewald K, Feichtinger H, Duh Q-Y, Clark OH, Kawasaki E, Bourne HR, *et al*. (1990) Two G protein oncogenes in human endocrine tumors. *Science*, **249**, 655–659.

Mantovani G, Ballare E, Giammona E, Beck-Peccoz P and Spada A (2002) The Gsα gene: Predominant maternal origin of transcription in human thyroid gland and gonads. *Journal of Clinical Endocrinology and Metabolism*, **87**, 4736–4740.

Mantovani G, Bandioni S, Lania AG, Corbetta S, de Sanctis L, Cappa M, DiBattista E, Chanson P, Beck-Peccoz P and Spada A (2004) Parental origin of $G_s\alpha$ mutations in McCune-Albright syndrome and in isolated endocrine tumors. *Journal of Clinical Endocrinology and Metabolism*, **89**, 3007–3009.

Pasolli HA, Klemke M, Kehlenbach RH, Wang Y and Huttner WB (2000) Characterization of the extra-large G protein α-subunit XLαs. I. Tissue distribution and subcellular localization. *Journal of Biological Chemistry*, **275**, 33622–33632.

Peters J, Wroe SF, Wells CA, Miller HJ, Bodle D, Beechey CV, Williamson CM and Kelsey G (1999) A cluster of oppositely imprinted transcripts at the *Gnas* locus in the distal imprinting region of mouse chromosome 2. *Proceedings of the National Academy of Sciences United States of America*, **96**, 3830–3835.

Rickard SJ and Wilson LC (2003) Analysis of *GNAS1* and overlapping transcripts identifies the parental origin of mutations in patients with sporadic Albright hereditary osteodystrophy and

reveals a model system in which to observe the effects of splicing mutations on translated messenger RNA. *American Journal of Human Genetics*, **72**, 961–974.

Sakamoto A, Liu J, Greene A, Chen M and Weinstein LS (2004) Tissue-specific imprinting of the G protein α-subunit $G_s\alpha$ is associated with tissue-specific differences in histone methylation. *Human Molecular Genetics*, **15**, 819–828.

Silve C, Santora A, Breslau N, Moses A and Spiegel A (1986) Selective resistance to parathyroid hormone in cultured skin fibroblasts from patients with pseudohypoparathyroidism type Ib. *Journal of Clinical Endocrinology and Metabolism*, **62**, 640–644.

Weinstein LS (2004) *GNAS* and McCune-Albright syndrome/fibrous dysplasia, Albright hereditary osteodystrophy/pseudohypoparathyroidism type 1A, progressive osseous heteroplasia, and pseudohypoparathyroidism type 1B. In *Molecular Basis of Inborn Errors of Development*, Epstein CJ, Erickson RP and Wynshaw-Boris A (Eds.), Oxford University Press: San Francisco, pp. 849–866.

Weinstein LS, Shenker A, Gejman PV, Merino MJ, Friedman E and Spiegel AM (1991) Activating mutations of the stimulatory G protein in the McCune-Albright syndrome. *New England Journal of Medicine*, **325**, 1688–1695.

Weinstein LS, Yu S and Ecelbarger CA (2000) Variable imprinting of the heterotrimeric G protein G_s α-subunit within different segments of the nephron. *American Journal of Physiology*, **278**, F507–F514.

Weinstein LS, Yu S, Warner DR and Liu J (2001) Endocrine manifestations of stimulatory G protein α-subunit mutations and the role of genomic imprinting. *Endocrine Reviews*, **22**, 675–705.

Williamson CM, Ball ST, Nottingham WT, Skinner JA, Plagge A, Turner MD, Powles N, Hough T, Papworth D, Fraser WD, *et al*. (2004) A *cis*-acting control region is required exclusively for the tissue-specific imprinting of *Gnas*. *Nature Genetics*, **36**, 894–899.

Wroe SF, Kelsey G, Skinner JA, Bodle D, Ball ST, Beechey CV, Peters J and Williamson CM (2000) An imprinted transcript, antisense to *Nesp*, adds complexity to the cluster of imprinted genes at the mouse *Gnas* locus. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3342–3346.

Wu WI, Schwindinger WF, Aparicio LF and Levine MA (2001) Selective resistance to parathyroid hormone caused by a novel uncoupling mutation in the carboxyl terminus of $G_{\alpha s}$. A cause of pseudohypoparathyroidism type Ib. *The Journal of Biological Chemistry*, **276**, 165–171.

Yu S, Yu D, Lee E, Eckhaus M, Lee R, Corria Z, Accili D, Westphal H and Weinstein LS (1998) Variable and tissue-specific hormone resistance in heterotrimeric G_s protein α-subunit ($G_{s\alpha}$) knockout mice is due to tissue-specific imprinting of the $G_{s\alpha}$ gene. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8715–8720.

Specialist Review

Developmental regulation of DNA methyltransferases

Diane J. Lees-Murdock and Colum P. Walsh
School of Biomedical Sciences, University of Ulster, Centre for Molecular Biosciences, Coleraine, UK

1. Introduction

It is well established that DNA methylation is important for maintaining genes in a transcriptionally repressed state (*see* Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1 and other chapters in this section). Due to the ability of methylation marks to be passed on to daughter cells after division, repression, once established, is very stable. Methylation patterns appear highly conserved in mice and humans; so to understand how and where methylation occurs in humans, extensive studies have been done in mouse, a more amenable experimental system. We will predominantly talk about results gleaned there, pointing out significant differences in humans where they may occur. Methylation in mice is used to repress a variety of different target sequences, some of which are parasitic in nature, such as endogenous retroviruses and retrotransposons (Bourc'his and Bestor, 2004; Walsh *et al*., 1998), and others that are endogenous genes required for normal development, such as the imprinted genes (Li *et al*., 1993, Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1, Article 31, **Imprinting at the *GNAS* locus and endocrine disease**, Volume 1, Article 30, **Beckwith–Wiedemann syndrome**, Volume 1, Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1) and those on the inactive X chromosome (Csankovszki *et al*., 2001, Article 15, **Human X chromosome inactivation**, Volume 1, Article 41, **Initiation of X-chromosome inactivation**, Volume 1 and Article 40, **Spreading of X-chromosome inactivation**, Volume 1). Once methylation is established on the inactive X and the silent copies of imprinted genes, they remain stably shut down in each cell lineage during the lifetime of the organism due to the faithful copying of methylation patterns.

Methylation now appears to be established on unmethylated DNA de novo by the enzymes Dnmt3a and Dnmt3b (Kaneda *et al*., 2004; Okano *et al*., 1999); to do this, they require the presence of the cofactor Dnmt3L (Bourc'his and Bestor, 2004; Bourc'his *et al*., 2001; Hata *et al*., 2002; Webster *et al*., 2005). After methylation has been laid down, the pattern is faithfully copied to daughter cells by the activity of the maintenance enzyme Dnmt1 (Li *et al*., 1992), though Dnmt3a and Dnmt3b may also be required for maintenance of methylation at some sites in mouse

(Chen *et al*., 2003) and humans (Ting *et al*., 2004). Earlier knockout experiments suggested that Dnmt2, a protein containing conserved methyltransferase motifs, was unlikely to be involved in development (Okano *et al*., 1998) and the protein has recently been shown to encode a tRNA methyltransferase (Grace Goll *et al*., 2006) and so will not be discussed further here.

During gamete production in the mature animal, inactivation of the X chromosome and of the silent imprinted allele must be reversed, otherwise dosage compensation would fail and the offspring might inherit two inactive or two active alleles (*see* Article 15, **Human X chromosome inactivation**, Volume 1, Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1). As expected then, during embryonic germ cell development, DNA methylation undergoes radical reprogramming on single copy sequences (Hajkova *et al*., 2002; Li *et al*., 2004, *see also* Article 33, **Epigenetic reprogramming in germ cells and preimplantation embryos**, Volume 1). On the parasitic sequences, however, methylation does not change as much and most of these sequences appear to retain a high level of methylation throughout the embryonic life (Lane *et al*., 2003; Lees-Murdock *et al*., 2003).

2. Surfing the methylation wave

Removal of methylation on imprinted genes and the X chromosome to reset epigenetic memory appears to occur in the germ cells in the primordial gonads at around embryonic day 11.5 in mice, which is about a day after the germ cells arrive at the presumptive gonad following their migration from the base of the allantois (Ginsburg *et al*., 1990) (see Figure 1). When the cells first start arriving at e10.5, methylation is present on the silent copy of each imprinted gene (Hajkova *et al*., 2002; Li *et al*., 2004) and one X chromosome is inactive (Tam *et al*., 1994). By e12.5, methylation marks have been erased from imprinted genes in both male and female germ cells (Davis *et al*., 2000; Hajkova *et al*., 2002; Li *et al*., 2004) and the inactive X chromosome is also reactivated (Tam *et al*., 1994). It has been suggested that the erasure of imprints may be an active process due to the very short period in which it occurs (Hajkova *et al*., 2002). Some methylation is also lost from repetitive DNA elements at this stage, but the bulk of *IAP*, *LINE1* and minor satellite sequences remain methylated (Hajkova *et al*., 2002; Lees-Murdock *et al*., 2003). Continued methylation of selfish elements may help prevent them from spreading in the germ line (Bourc'his and Bestor, 2004; Walsh *et al*., 1998), while methylation of nontranscribed repeats may also be important for genome stability (Hansen *et al*., 1999; Okano *et al*., 1999; Xu *et al*., 1999).

The repeat elements are rapidly de novo methylated in the male germ line between e15.5 and e17.5, whereas methylation of these sequences in the female germ line remains lower at this stage (Lees-Murdock *et al*., 2003). The timing of this event coincides with the onset of methylation of the paternally imprinted genes, *H19* (Davis *et al*., 2000) and *Rasgrf1* (Li *et al*., 2004) that is initiated at e17.5, but unlike the repeat sequences, which have become completely methylated by

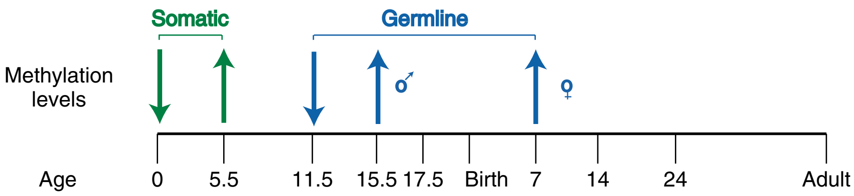


Figure 1 Timeline of development in the mouse with major methylation changes indicated. Embryonic or postnatal age in days is indicated below the line, with zero being the day of conception; embryonic ages are by convention given in half-day intervals. Mice are generally born at embryonic day 19 or 20, and then take three (female) to five (male) weeks to become sexually mature adults. Above the line is indicated the timing of *onset* of major methylation changes, decreases in methylation are a downward-pointing arrow, increases an upward-pointing one. Somatic (green) and germ cell (blue) lineages undergo separate methylation events, and de novo methylation begins at different times in the male and female germ lines. In male germ cells, methylation begins at approximately e15.5, but in females de novo methylation first occurs *postpartum* in the growing oocyte

e17.5, the paternally methylated imprinted genes do not become fully methylated until after birth (Bourc'his and Bestor, 2004; Davis *et al*., 2000; Kaneda *et al*., 2004; Li *et al*., 2004). In the female germ line, de novo methylation occurs in the growing oocyte for both repeats and imprinted genes (Lucifero *et al*., 2004).

In the preimplantation embryo as well, there is some evidence for ongoing changes in DNA methylation. In mice, the male pronucleus appears to become actively demethylated in the one-cell embryo, as shown by staining with antibodies to methylated cytosine (Santos *et al*., 2002) and by bisulfite sequencing of the *H19* gene (Olek and Walter, 1997), though this is not the case in all mammals (Beaujean *et al*., 2004). The female pronucleus is resistant to this active demethylation but becomes demethylated passively during cleavage divisions (Olek and Walter, 1997; Santos *et al*., 2002). *IAP*s, but not *LINE1* elements, are maintained at high methylation levels throughout this period (Lane *et al*., 2003; Walsh *et al*., 1998). After implantation, methylation levels overall appear to increase again (reviewed in (Yoder *et al*., 1997)). For single copy genes, the timing of methylation and its extent depends completely on the individual gene, since some do not become methylated at all, while some only become methylated when cell lineages become more firmly established, or during tumorigenesis (Jones and Baylin, 2002; Walsh and Bestor, 1999).

It is clear therefore that methylation must be tightly regulated to ensure that it is established and maintained correctly at the appropriate times and also to prevent maintenance or de novo methylation from occurring when cells are being reprogrammed in the germ line and in early development. In theory, this could involve proteins which bind to transcriptional control regions and either attract or prevent DNA methylation. As yet only a few candidates for such factors have been identified (Fedoriw *et al*. 2004; Suzuki *et al*., 2005), but this number would be expected to grow. Alternatively, control of when and where the cytosine methyltransferases themselves are expressed should also allow restriction of methylation to specific periods. There is good

evidence that this latter control mechanism is also used and here we review known mechanisms used to limit DNA methyltransferase expression during development.

3. Use of alternative promoters

One method used for restricting methyltransferase expression is by utilizing alternative promoters specific for particular cell types. This phenomenon was first seen at the *Dnmt1* locus in mouse, where it was discovered that three types of transcripts could be produced (Mertineit *et al*., 1998), originating at separate promoters that were active in different cell types (Figure 2). One promoter drives transcription of a full-length protein, Dnmt1^{s}, in all somatic cells. However, in the mouse testis, a separate promoter is used to drive expression specifically in the pachytene spermatocytes. This transcript (*Dnmt1*^{p}) is longer than the somatic form, but contains many upstream Open Reading Frames (ORFs) before the main protein-coding ORF, which are thought to inhibit protein expression (Mertineit *et al*., 1998). Supporting this view, the pachytene transcripts are not associated with polysomes and the protein cannot be detected in these cells (Jue *et al*., 1995). A recent report suggests that this transcript may also be active in myoblasts and showed it could produce some protein in an in vitro transcription-translation (IVTT) system (Aguirre-Arteta *et al*., 2000); however, the ability of the transcript to be

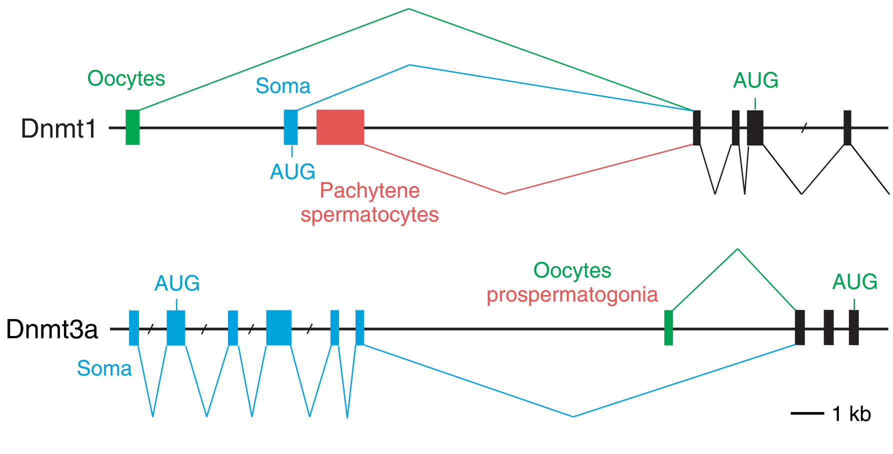


Figure 2 Alternative promoter and initiation codon usage for DNA methyltransferase genes in mice. The structures of the *Dnmt1* and *Dnmt3a* loci are shown schematically. Alternative promoters for *Dnmt1* direct the production of unique transcripts in oocytes (*Dnmt1*^{o}–green), pachytene spermatocytes (*Dnmt1*^{p}–red) and somatic cells (*Dnmt1*^{s}– blue), which differ only at the 5′ ends. The initiation codon used by *Dnmt1*^{s} is missing in *Dnmt1*^{o}, which instead initiates at an internal AUG codon to give a shorter protein with altered stability. Most evidence suggests that *Dnmt1*^{p} does not produce a protein. For *Dnmt3a*, one promoter (blue) produces a long transcript present at low levels in most tissues. A second promoter active in both germ lines produces a shorter mRNA (green). This also uses an internal initiation codon (indicated) to produce the smaller *Dnmt3a2* protein isoform. *Dnmt3a2* is also detected in somatic cells undergoing de novo methylation

translated may differ substantially in vivo from the IVTT system, so the balance of evidence is against a functional role for this form.

What then is it doing? The testis produces many alternative transcripts not seen in somatic cells (reviewed in (Kleene, 2001)). Some of these transcripts may reflect a need for a modified mRNA or an alternative protein in the highly specialized sperm cells. For other transcripts, intensive investigation by homologous deletion and other means have failed to find any function so far. It is possible that, like some alternative transcripts, the mRNA has no function and is a means of downregulating protein production while maintaining the locus in an active conformation. This might facilitate transcription of a functional isoform early in development or promote homologous recombination at this site. The gold standard test would be to carry out targeted mutation of the pachytene promoter or first exon and check for effects on male fertility. Existing *Dnmt1*-null mutant mice that target downstream exons generally do not survive to adulthood due to removal of the functional protein in somatic cells as well, so effects on male fertility are not known.

The third promoter for *Dnmt1* is active only in the oocyte and produces a longer transcript, but with a slightly shorter ORF (Mertineit *et al*., 1998). This shorter protein, Dnmt1^o, appears to be fully functional, however, and is the only isoform produced in the egg. This was demonstrated in knockout mice with homozygous deletion of the Dnmt1^o first exon and proximal promoter, which showed loss of Dnmt1^o in the egg, but normal expression of Dnmt1^s from implantation onward (Howell *et al*., 2001). Female mice with homozygous Dnmt1^o deletions were normal and produced eggs which also showed normal imprint establishment by meiosis II, showing that Dnmt1^o was not required for de novo methylation establishment on imprinted genes in the egg. However, these eggs did not go on to produce normal offspring, embryo loss postimplantation occurred and when examined, all imprinted genes had lost 50% of their methylation, due to the passive demethylation occurring during cleavage division (Figure 3) – normally, Dnmt1^o acts to prevent loss of methylation on imprinted genes at this stage. In mutants, the result is aberrant expression or silencing of the genes and since many are crucial for development, this is the likely cause of the embryonic lethality observed. The reason for the 50% loss is due to a nuclear trafficking event, discussed below. Although, Howell and colleagues did not observe any marked global depletion of methylation on *IAP* elements in Dnmt1^o knockout mice, later studies using Dnmt1^o knockouts coupled to a specific *IAP* insertion at the *agouti* locus suggest that Dnmt1^o may be required for maintenance of methylation in the preimplantation embryo on some *IAP*s as well (Gaudet *et al*., 2004). Confirming the importance of promoter switching for Dnmt, oocyte-specific promoters and mRNAs equivalent to the mouse Dnmt1^o also exist in human (Hayward *et al*., 2003) and opossum (Ding *et al*., 2003).

For Dnmt3a as well, an alternative promoter has been identified in both mouse and human that appears to be more active in germ cells and in postimplantation embryos, stages where de novo methylation is occurring (Chen *et al*., 2002). This promoter produces a shorter isoform called Dnmt3a2 that has been shown to have similar ability to rescue methylation in knockout mice as the longer Dnmt3a (Chen *et al*., 2003). The presence of a large number of pseudogenes in the rat

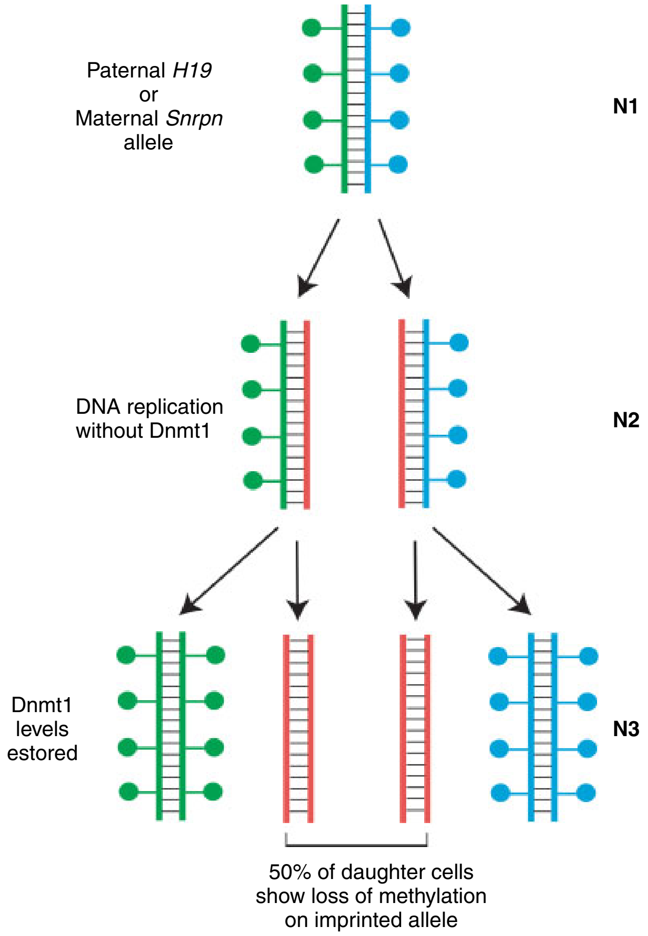


Figure 3 Consequences of a single cell duplication without efficient methylation maintenance activity. The DNA duplex encoding an imprinted gene is shown in the first generation (N1) at top. On the transcriptionally silent allele of imprinted genes, each strand of the DNA has methyl groups (circles). For *H19*, this is on the paternal chromosome (the maternal chromosome which is unmethylated is not shown, because its status will be unaffected by the loss of maintenance activity), whereas for *Snrpn*, it is the maternal copy of the gene that is methylated. On replication, the two strands of the helix are separated and act as templates for two new daughter strands (red). Methylation is added postreplication: if this does not occur, the daughter duplexes in N2 are hemimethylated. At the next round of replication, the unmethylated strands will act as templates for new duplexes. *Dnmt1* uses the parental strand as a template to add methylation to the newly synthesized strand and is not able to add methyl groups de novo. Even if maintenance activity is restored in N3, 50% of the cells will have lost methylation and will show dysregulation of imprinted gene transcription. This is what appears to occur in embryos, which develop from eggs lacking Dnmt1°, where both paternally and maternally methylated imprinted alleles show a 50% loss of methylation and aberrant transcription is seen. This gives an overall ratio of 75% unmethylated to 25% methylated strands, counting the unmethylated active alleles, which are unaffected in the mutant (not shown)

genome matching the shorter isoform, but not the longer, suggests that Dnmt3a2 is the ancestral transcript and that the additional exons found upstream in Dnmt3a may have been acquired later in evolution (Lees-Murdock *et al*., 2004). Cre-lox mediated mutation of an exon common to Dnmt3a and Dnmt3a2 specifically in germ cells demonstrated that the gene has a vital role in de novo methylation of imprinted genes in this lineage (Kaneda *et al*., 2004) though again, isoform-specific targeting would be needed to confirm that Dnmt3a2, rather than Dnmt3a, is more important in germ cells.

Despite identification of the putative promoter regions of the major players and some work on cloning and characterization (Aapola *et al*., 2004; Ishida *et al*., 2003; Ko *et al*., 2005; McCabe *et al*., 2005), on the whole little has been done to identify what signals they respond to and what transcription factors may play a role in regulating them.

4. Alternative splicing

One mechanism known to operate at the *Dnmt3b* locus to regulate protein production is the use of alternative splicing. This is a posttranscriptional level of control and thus simple genetic approaches are less useful to determine which factors regulate the process, and other experimental tools will be needed to work out the relevant details. A large number of mature transcripts of *Dnmt3b* exist in mouse and human (Chen *et al*., 2002; Lees-Murdock *et al*., 2005; Xu *et al*., 1999), differing in the splicing of a number of exons. Many of these are also capable of being translated into proteins in vivo (Chen *et al*., 2002), giving rise to different isoforms of the protein. While it has not yet been determined if all of these isoforms are functionally different, it is clear that in a number of these, vital motifs necessary for catalytic activity have been spliced out. Some of these variants have been tested for their ability to remethylate DNA and do indeed show little or no catalytic activity (Chen *et al*., 2003). Transcription of such mRNAs, which do not appear to produce functional proteins, is reminiscent of the situation for *Dnmt1* transcripts in the pachytene spermatocytes, and probably represents a way of downregulating the protein. Examination of the types of isoforms present at various stages of development can help to identify stages at which no functional isoforms are produced (Lees-Murdock *et al*., 2005; Sakai *et al*., 2004) and suggest that Dnmt3b2, a transcript lacking exon 10, may be most important for de novo methylation.

5. Nuclear import/export

Methylation relies on the proximity of the enzyme to its target in the nucleus of the cell and this can also act as a point of control for regulation of methylation. This was first demonstrated for Dnmt1, where studies showed that the protein undergoes complex cycles of entry and exit to the nucleus during development (Leonhardt *et al*., 1992). This form of control may be most important in the very earliest stages of development prior to implantation. As discussed above, in mouse oocytes Dnmt1

is produced from an alternative sex-specific $5'$ exon resulting in a truncated protein, Dnmt1^{o}. This short protein is the predominant form of Dnmt1 in oogenesis and early preimplantation development (Howell *et al*., 2001). Growing oocytes exhibit high levels of Dnmt1^{o} in the nucleus. As oogenesis proceeds, Dnmt1^{o} is either actively exported or simply blocked from entering the nucleus by factors retaining it in the cytoplasm and becomes sequestered subcortically in mature oocytes (Howell *et al*., 2001; Mertineit *et al*., 1998). Following fertilization, Dnmt1^{o} remains in the cytoplasm for the first two divisions, but moves back into the nucleus at the eight cell stage for a single cell division, before being again cytoplasmically localized at the 16-cell and morula stages (Carlson *et al*., 1992). At implantation, Dnmt1^{o} is downregulated and transcription of Dnmt1^{s} is activated in the blastocyst/egg cylinder and becomes localized to the nuclei (Trasler *et al*., 1996). Subcellular localization in the early embryo is most likely achieved by retention of Dnmt1^{o} by a factor in the egg cytoplasm rather than active nuclear export, as a tagged version of Dnmt1^{o}, when expressed in several different somatic cell lines, shows clear nuclear localization (Cardoso and Leonhardt, 1999) and embryos treated with a nuclear exporter inhibitor do not show accumulation of the protein in the nucleus (Doherty *et al*., 2002).

This complex choreography can perhaps best be understood in terms of what we know Dnmt1 is required for, i.e. maintaining existing methylation patterns. Restricting the protein to the cortex allows passive demethylation to occur during the cleavage divisions. The enzyme moves back in at the eight cell stage specifically to prevent loss of methylation on all imprinted genes (Howell *et al*., 2001). The fact that methylated alleles lose 50% of their methylation suggests that a single cell division is affected (see Figure 3); here then is an explanation for why Dnmt1^{o} must move back into the nucleus for a single cell division at the eight cell stage.

The mechanism causing the translocation is unknown but has been shown to be independent of DNA replication, transcription, protein synthesis, and compaction (Doherty *et al*., 2002). This unusual peripheral cytoplasmic localization of Dnmt1^{o} could be mediated by an Annexin V binding site present in the $5'$ end of the protein. Indeed, Dnmt1^{o} has been shown to bind to this protein (Ohsawa *et al*., 1996) and colocalize with it in preimplantation embryos (Doherty *et al*., 2002). Other binding sites have been described in both the amino terminus of the protein (Rountree *et al*., 2000) and the $3'$ untranslated region (UTR) of the mRNA (Lees-Murdock *et al*., 2004) that may be involved in subcellular localization, although their importance during development has not yet been fully elucidated.

Protein localization also appears to play a role in regulating the de novo methyltransferases' access to the DNA. During germ cell development, when imprints need to be removed after the germ cells colonize the gonad, Dnmt3b is excluded from germ cell nuclei while Dnmt3a is undetectable (Hajkova *et al*., 2002). At later stages, Dnmt3a is found in the male germ cells beginning at e15.5 (Lees-Murdock *et al*., 2005; Sakai *et al*., 2004), when methylation is occurring, while it is absent from the female germ cells at the same stage. Localization studies on Dnmt3a and Dnmt3b in the developing egg show similar patterns to that of Dnmt1, being localized to the nucleus at the earliest stages, but then moving out and becoming confined to the cortex in the MII oocyte (Lees-Murdock *et al*., 2004). In line with these observations, homozygous deletion of the Dnmt3a gene

results in the failure of de novo methylation in both the prospermatogonia and the growing oocyte (Kaneda *et al*., 2004). This knockout study did not determine which isoform, Dnmt3a or Dnmt3a2, is more important in germ cells, but a recent study indicates that Dnmt3a2 is the only form detected in the male germ cells during the period of de novo methylation (Sakai *et al*., 2004). A Dnmt3b deletion was reported not to effect methylation in germ cells (Kaneda *et al*., 2004), but no primary data was presented in the paper to support this, and additional work on this question will be of interest.

6. Protein stability

Finally, stabilization of Dnmt1 protein has recently been shown to be important for regulating its activity. The shorter oocyte form of the protein, when expressed in parallel with the somatic form in embryonic stem (ES) cells, showed greater stability than the latter (Ding and Chaillet, 2002). This increased stability would be an advantage during long-term storage in the egg. There is also a conserved KEN box in the middle of the protein which is a target for ubiquitin addition and subsequent proteasomal degradation of the protein: degradation occurs in the nucleus, and is prevented by cytoplasmic sequestration of the protein (Ghoshal *et al*., 2005). This may contribute to the increased stability of the oocyte form of the protein.

7. Importance of coordination

We now know that Dnmt1 appears to be almost exclusively involved in maintaining methylation patterns, after these patterns have been put on the DNA by the de novo enzymes Dnmt3a2 and Dnmt3b2. In the germ line (but not the soma), these latter appear dependent on the presence of Dnmt3L (see the article by Bestor and Bourc'his in this volume): in the absence of this protein, methylation fails to be properly established in either the female or male sex cells in mice (Bourc'his and Bestor, 2004; Bourc'his *et al*., 2001; Hata *et al*., 2002; Hata *et al*., 2006; Kaneda *et al*., 2004; Webster *et al*., 2005). It would therefore appear vital to coordinate expression of Dnmt3L and the de novo enzymes so that both an active enzyme and Dnmt3L are present in the nucleus at the same time. The mechanisms outlined above converge to provide only limited periods during development when Dnmt3L is actively transcribed and the active isoforms of the de novo enzymes are present in the nucleus (La Salle *et al*., 2004; Lees-Murdock *et al*., 2005; Sakai *et al*., 2004). Once established, these patterns of methylation are faithfully copied by Dnmt1 at every cell division: failure to do so for even a single cell cycle can lead to loss of imprints and embryo death, as evidenced by the Dnmt1^{o} knockout (Howell *et al*., 2001).

It is interesting to note that studies on human (Huntriss *et al*., 2004) and rhesus monkey (Vassena *et al*., 2005) found that *DNMT3L* was not transcriptionally active in the egg, but was turned on at the blastocyst stage. There was also one report indicating that maternal methylation imprints in human are established after

fertilization at the *SNRPN* gene (El-Maarri *et al*., 2001), though this was later disputed (Geuns *et al*., 2003). Further studies are needed to clarify the situation with regard to the timing of imprint establishment and the role of *DNMT3L* for maternal imprints in humans.

8. Concluding remarks

It is perhaps because of the powerful nature of gene repression through CpG methylation that such manifold layers of control have arisen to restrict methyltransferase expression during development. Although it can take as little as few hours for methylation to be established in germ cells, repression of the target genes can be efficiently maintained through decades of life for the organism. Evidently, with such power came the intricate series of checks and balances reviewed here.

References

Aapola U, Maenpaa K, Kaipia A and Peterson P (2004) Epigenetic modifications affect Dnmt3L expression. *The Biochemical Journal*, **380**, 705–713.

Aguirre-Arteta AM, Grunewald I, Cardoso MC and Leonhardt H (2000) Expression of an alternative Dnmt1 isoform during muscle differentiation. *Cell Growth & Differentiation*, **11**, 551–559.

Beaujean N, Taylor JE, McGarry M, Gardner JO, Wilmut I, Loi P, Ptak G, Galli C, Lazzari G, Bird A *et al*. (2004) The effect of interspecific oocytes on demethylation of sperm DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7636–7640.

Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, **431**, 96–99.

Bourc'his D, Xu GL, Lin CS, Bollman B and Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science*, **294**, 2536–2539.

Cardoso MC and Leonhardt H (1999) DNA methyltransferase is actively retained in the cytoplasm during early development. *The Journal of Cell Biology*, **147**, 25–32.

Carlson LL, Page AW and Bestor TH (1992) Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting. *Genes & Development*, **6**, 2536–2541.

Chen T, Ueda Y, Dodge JE, Wang Z and Li E (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and Cellular Biology*, **23**, 5594–5605.

Chen T, Ueda Y, Xie S and Li E (2002) A novel dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *The Journal of Biological Chemistry*, **277**, 38746–38754.

Csankovszki G, Nagy A and Jaenisch R (2001) Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *The Journal of Cell Biology*, **153**, 773–784.

Davis TL, Yang GJ, McCarrey JR and Bartolomei MS (2000) The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Human Molecular Genetics*, **9**, 2885–2894.

Ding F and Chaillet JR (2002) In vivo stabilization of the Dnmt1 (cytosine-5)- methyltransferase protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14861–14866.

Ding F, Patel C, Ratnam S, McCarrey JR and Chaillet JR (2003) Conservation of Dnmt1o cytosine methyltransferase in the marsupial Monodelphis domestica. *Genesis*, **36**, 209–213.

Doherty AS, Bartolomei MS and Schultz RM (2002) Regulation of stage-specific nuclear translocation of Dnmt1o during preimplantation mouse development. *Developmental Biology*, **242**, 255–266.

El-Maarri O, Buiting K, Peery EG, Kroisel PM, Balaban B, Wagner K, Urman B, Heyd J, Lich C, Brannan CI *et al*. (2001) Maternal methylation imprints on human chromosome 15 are established during or after fertilization. *Nature Genetics*, **27**, 341–344.

Fedoriw AM, Stein P, Svoboda P, Schultz RM and Bartolomei MS (2004) Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science*, **303**, 238–240.

Gaudet F, Rideout WM, Meissner A, Dausman J, Leonhardt H and Jaenisch R 3rd (2004) Dnmt1 expression in pre- and postimplantation embryogenesis and the maintenance of IAP silencing. *Molecular and Cellular Biology*, **24**, 1640–1648.

Geuns E, De Rycke M, Van Steirteghem A and Liebaers I (2003) Methylation imprints of the imprint control region of the SNRPN-gene in human gametes and preimplantation embryos. *Human Molecular Genetics*, **12**, 2873–2879.

Ghoshal K, Datta J, Majumder S, Bai S, Kutay H, Motiwala T and Jacob ST (2005) 5-Aza-deoxycytidine induces selective degradation of DNA methyltransferase 1 by a proteasomal pathway that requires the KEN box, bromo-adjacent homology domain, and nuclear localization signal. *Molecular and Cellular Biology*, **25**, 4727–4741.

Ginsburg M, Snow MH and McLaren A (1990) Primordial germ cells in the mouse embryo during gastrulation. *Development*, **110**, 521–528.

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE and Bestor TH (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, **311(5759)**, 395–398.

Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J and Surani M (2002) Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of Development*, **117**, 15.

Hansen RS, Wijmenga C, Luo P, Stanek AM, Canfield TK, Weemaes CM and Gartler SM (1999) The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14412–14417.

Hata K, Kusumi M, Yokomine T, Li E and Sasaki H (2006) Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. *Molecular Reproduction and Development*, **73**, 116–122.

Hata K, Okano M, Lei H and Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development*, **129**, 1983–1993.

Hayward BE, De Vos M, Judson H, Hodge D, Huntriss J, Picton HM, Sheridan E and Bonthron DT (2003) Lack of involvement of known DNA methyltransferases in familial hydatidiform mole implies the involvement of other factors in establishment of imprinting in the human female germline. *BMC Genetics*, **4**, 2.

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM and Chaillet JR (2001) Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell*, **104**, 829–838.

Huntriss J, Hinkins M, Oliver B, Harris SE, Beazley JC, Rutherford AJ, Gosden RG, Lanzendorf SE and Picton HM (2004) Expression of mRNAs for DNA methyltransferases and methyl-CpG-binding proteins in the human female germ line, preimplantation embryos, and embryonic stem cells. *Molecular Reproduction and Development*, **67**, 323–336.

Ishida C, Ura K, Hirao A, Sasaki H, Toyoda A, Sakaki Y, Niwa H, Li E and Kaneda Y (2003) Genomic organization and promoter analysis of the Dnmt3b gene. *Gene*, **310**, 151–159.

Jones PA and Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature Reviews. Genetics*, **3**, 415–428.

Jue K, Bestor TH and Trasler JM (1995) Regulated synthesis and localization of DNA methyltransferase during spermatogenesis. *Biology of Reproduction*, **53**, 561–569.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E and Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, **429**, 900–903.

Kleene KC (2001) A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mechanisms of Development*, **106**, 3–23.

Ko YG, Nishino K, Hattori N, Arai Y, Tanaka S and Shiota K (2005) Stage-by-stage change in DNA methylation status of Dnmt1 locus during mouse early development. *The Journal of Biological Chemistry*, **280**, 9627–9634.

Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J and Reik W (2003) Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*, **35**, 88–93.

La Salle S, Mertineit C, Taketo T, Moens PB, Bestor TH and Trasler JM (2004) Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells. *Developmental Biology*, **268**, 403–415.

Lees-Murdock DJ, De Felici M and Walsh C (2003) Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics*, **82**, 230–237.

Lees-Murdock DJ, McLoughlin GA, McDaid JR, Quinn LM, O'Doherty A, Hiripi L, Hack CJ and Walsh CP (2004) Identification of 11 pseudogenes in the DNA methyltransferase gene family in rodents and humans and implications for the functional loci. *Genomics*, **84**, 193–204.

Lees-Murdock DJ, Shovlin TC, Gardiner T, De Felici M and Walsh CP (2005) DNA methyltransferase expression in the mouse germ line during periods of de novo methylation. *Developmental Dynamics*, **232**, 992–1002.

Leonhardt H, Page AW, Weier HU and Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, **71**, 865–873.

Li E, Beard C and Jaenisch R (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Li E, Bestor TH and Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

Li JY, Lees-Murdock DJ, Xu GL and Walsh CP (2004) Timing of establishment of paternal methylation imprints in the mouse. *Genomics*, **84**, 952–960.

Lucifero D, Mann MR, Bartolomei MS and Trasler JM (2004) Gene-specific timing and epigenetic memory in oocyte imprinting. *Human Molecular Genetics*, **13**, 839–849.

McCabe MT, Davis JN and Day ML (2005) Regulation of DNA methyltransferase 1 by the pRb/E2F1 pathway. *Cancer Research*, **65**, 3624–3632.

Mertineit C, Yoder JA, Taketo T, Laird DW, Trasler JM and Bestor TH (1998) Sex-specific exons control DNA methyltransferase in mammalian germ cells. *Development*, **125**, 889–897.

Ohsawa K, Imai Y, Ito D and Kohsaka S (1996) Molecular cloning and characterization of annexin V-binding proteins with highly hydrophilic peptide structure. *Journal of Neurochemistry*, **67**, 89–97.

Okano M, Bell DW, Haber DA and Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.

Okano M, Xie S and Li E (1998) Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic Acids Research*, **26**, 2536–2540.

Olek A and Walter J (1997) The pre-implantation ontogeny of the H19 methylation imprint [letter]. *Nature Genetics*, **17**, 275–276.

Rountree MR, Bachman KE and Baylin SB (2000) DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature Genetics*, **25**, 269–277.

Sakai Y, Suetake I, Shinozaki F, Yamashina S and Tajima S (2004) Co-expression of de novo DNA methyltransferases Dnmt3a2 and Dnmt3L in gonocytes of mouse embryos. *Gene Expression Patterns*, **5**, 231–237.

Santos F, Hendrich B, Reik W and Dean W (2002) Dynamic reprogramming of DNA methylation in the early mouse embryo. *Developmental Biology*, **241**, 172–182.

Suzuki M, Yamada T, Kihara-Negishi F, Sakurai T, Hara E, Tenen DG, Hozumi N and Oikawa T (2005) Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. *Oncogene*, epub doi:10.1038.

Tam PP, Zhou SX and Tan SS (1994) X-chromosome activity of the mouse primordial germ cells revealed by the expression of an X-linked lacZ transgene. *Development*, **120**, 2925–2932.

Ting AH, Jair KW, Suzuki H, Yen RW, Baylin SB and Schuebel KE (2004) CpG island hypermethylation is maintained in human colorectal cancer cells after RNAi-mediated depletion of DNMT1. *Nature Genetics*, **36**, 582–584.

Trasler JM, Trasler DG, Bestor TH, Li E and Ghibu F (1996) DNA methyltransferase in normal and Dnmtn/Dnmtn mouse embryos. *Developmental Dynamics*, **206**, 239–247.

Vassena R, Dee Schramm R and Latham KE (2005) Species-dependent expression patterns of DNA methyltransferase genes in mammalian oocytes and preimplantation embryos. *Molecular Reproduction and Development*, **72**, 430–436.

Walsh CP and Bestor TH (1999) Cytosine methylation and mammalian development. *Genes & Development*, **13**, 26–34.

Walsh CP, Chaillet JR and Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, **20**, 116–117.

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R *et al*. (2005) Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4068–4073.

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ and Viegas-Pequignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, **402**, 187–191.

Yoder JA, Walsh CW and Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, **13**, 335–340.

Specialist Review

Epigenetic variation: amount, causes, and consequences

Elena de la Casa-Esperón
University of Texas at Arlington, Arlington, TX, US

Carmen Sapienza
Temple University, Philadelphia, PA, US

1. Introduction

The diversity of human phenotypes that we observe is the result of genetic and epigenetic variation and the interaction of these "biological" variables with environmental factors. Both large-scale and small-scale genome sequencing projects, as well as more recent efforts to define structural variation (copy number variation and subkaryotypic insertions, deletions and rearrangements), have resulted in an important initial description of the amount and type of genetic variation in the human genome. On the other hand, the scale of epigenetic variation in the human population is only beginning to be investigated.

Epigenetic variation may arise by diverse mechanisms but, at the molecular level, it reflects differences in the spatial configuration of chromatin and its interactions and function. Multiple biochemical processes (DNA methylation, histone methylation, acetylation, phosphorylation, sumoylation, etc.) are associated with these differences. One important consequence of this variability is the resultant variation in gene expression, although many other effects have also been described (see the following text).

In the same way that somatic mutations can be transmitted through successive cell divisions, epigenetic marks can change during the lifespan of an organism and also be transmitted somatically through subsequent cell divisions. In fact, the normal phenotypic diversity found between the different cell types of an organism is, with a few notable exceptions in the immune system, epigenetically controlled.

Interestingly, traits that result from particular patterns of epigenetic modification can also be transmitted between generations in some circumstances. The term "epialleles" has been coined to describe such different epigenetic states (Article 36, **Variable expressivity and epigenetics**, Volume 1). However, unlike DNA sequence changes, epigenetic modifications are often reversible at much higher frequencies than the mutation rate. This is an important characteristic, because epigenetic marks can be reset between generations and they can change in response

to the environment. Because epigenetic variation can also be genetically controlled, it constitutes a potentially important link between environmental and genetic factors (Cui *et al*., 1998; Nakagawa *et al*., 2001; Sandovici *et al*., 2003). Such a response to the environment could be mediated by metabolic changes that result in epigenetic modifications (Paldi, 2003; Waterland and Jirtle, 2004; Wolff *et al*., 1998). Consequently, epigenetic variability is not only a source of phenotypic plasticity in response to the environment, but these epigenetic alterations can also, potentially, be transmitted between generations, with very important implications in evolution (Rutherford and Henikoff, 2003; Sollars *et al*., 2003).

To better understand the relevance of epigenetic variation, we will discuss the extent (how much?), the origin (what are the causes?), and the implications (what are the consequences?) of this important source of phenotypic variation.

2. Epigenetic variation: how much?

2.1. *Epigenetic variation arises from multiple mechanisms*

It is difficult to estimate the precise extent of epigenetic variation because it occurs at multiple levels and as a result of multiple processes. The epigenetic variation resulting from inactivation of X chromosome provides a classic example of how multiple and distinct processes can give rise to very large fluctuation in phenotype among genetically similar or identical (Fraga *et al*., 2005) individuals. In human females (as in other female mammals), one of the two X chromosomes is inactivated by epigenetic means. Once one of the two X chromosomes is chosen for inactivation early in development, and the same X chromosome remains inactive in all descendants of that cell (Article 41, **Initiation of X-chromosome inactivation**, Volume 1). The inactive X chromosome becomes a cytologically visible heterochromatic body. This cytological manifestation of femaleness (the Barr body) is to a large extent (but not completely (Disteche, 1995)) transcriptionally inert. This means that each single cell expresses only one allele of most (approximately 85%) (Carrel and Willard, 2005) X-linked genes. If both X chromosomes have the same probability of being inactivated, the "average" women will have the paternal X chromosome inactive in 50% of her cells and the maternal X inactive in the remaining 50% of her cells. However, because the process of choosing the X chromosome for inactivation has a large stochastic component (Article 41, **Initiation of X-chromosome inactivation**, Volume 1), individual women will have different patterns of X-inactivation (Figure 1a). In fact, there is a minor fraction of females in whom >90% of the cells have the same X chromosome inactivated (Figure 1a). These females will have highly preferential expression of either maternal or paternal alleles of all X-linked genes affected by the inactivation process.

In addition to this partly stochastic, partly genetic variability in the fraction of cells in which a particular X chromosome remains active (see next section), there is also population-level and intraindividual variability in the extent of X-inactivation. It has been documented that a fraction of X-linked genes have escaped inactivation' (reviewed in Disteche, 1995). Interestingly, some genes are

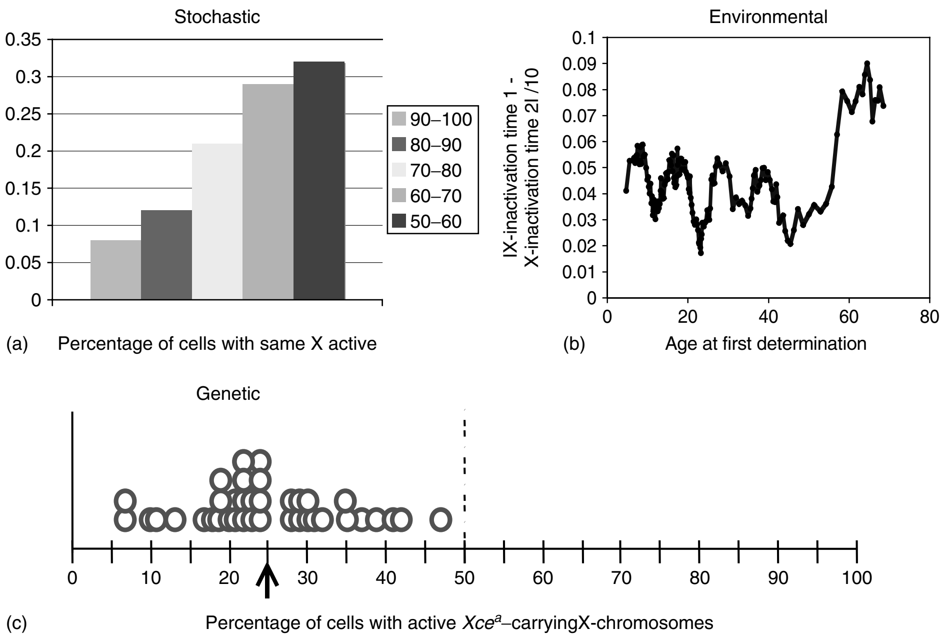


Figure 1 Origin of X-inactivation variation. (a) Much of the variation in the human results from the stochastic component of the X-inactivation choice process. Y-axis represents the fraction of women with the indicated percent of cells with the same X chromosome active (Naumova *et al*., 1996 and our unpublished data). Approximately one-third of women have either X chromosomes inactivated in one-half of their of cells (purple bar) and approximately 60% of women (purple bar plus green bar) have X-inactivation ratios between 50:50 and 70:30. However, approximately 7% of women have highly skewed patterns of X-inactivation, that is, greater than 90:10 (blue bar) in favor of the inactivation of a particular X chromosome. (b) Moving average of change in X-inactivation score in individual females (over nearly two decades; see Sandovici *et al*., 2004) as a function of age. Females who were greater than 60 years of age when the first sample was taken show significantly more variation over time than younger females. (c) Heritable effects on X-chromosome inactivation variation. Distribution of X-inactivation ratios in heterozygous Xce^a /Xce^c mouse females – each circle represents an individual female mouse (de la Casa-Esperon *et al*., 2002). The X-controlling element locus affects the probability that an X chromosome will become inactive, so that X chromosomes carrying the Xce^a allele have a higher probability of being inactivated than X chromosomes carrying the Xce^c allele. The observed mean X-inactivation ratio of this population of females is 25% of cells with an active Xce^a -carrying X chromosome

inactivated in some human samples, but escape inactivation in others (Carrel and Willard, 2005). In addition, the level of expression of such "escapees" also differs between samples (Carrel and Willard, 2005). Therefore, even genetically identical women (monozygotic twins) can differ in their mosaic pattern of X-inactivation, the number of genes that escape X-inactivation, and the levels of expression of some X-linked genes. In addition, some genes that have been inactivated may become reactivated as a function of age (Wareham *et al*., 1987) or other environmental factors, although not all X-linked genes appear equally susceptible to reactivation (Migeon *et al*., 1988; Pagani *et al*., 1990).

A similar variability in a long-term inactivation phenomenon has been observed for another class of monoallelically expressed genes located in the autosomes, the imprinted genes. Imprinted genes are expected to be expressed exclusively, or

nearly exclusively, from the paternal or the maternal copy (Article 37, **Evolution of genomic imprinting in mammals**, Volume 1). Several studies have shown that some imprinted genes (e.g., *IGF2*, *HTR2A* genes) are expressed from both alleles in a small fraction of normal individuals (Bunzel *et al*., 1998; Sakatani *et al*., 2001), while others (*IGF2R*) exhibit the reciprocal characteristic of being imprinted in only a small fraction of individuals (Xu *et al*., 1993).

Expression levels between alleles have been found to be variable for several imprinted genes in human tissues (Dao *et al*., 1998; McMinn *et al*., 2006), and have also been observed at nonimprinted autosomal genes. In fact, large-scale transcription profiling studies in humans have shown differential expression of alleles at a large proportion of loci (up to 54%, depending on the cutoff level of differential expression selected (Lo *et al*., 2003)) and, interestingly, the degree of difference in expression between particular alleles varies between individuals (Lin *et al*., 2005; Lo *et al*., 2003; Pant *et al*., 2006; Pastinen *et al*., 2004). Moreover, skewing of allelic expression is not necessarily in the same direction: in some individuals who are heterozygous for the same alleles, the allele that is preferentially expressed differs (Lo *et al*., 2003; Pastinen *et al*., 2004). This observation suggests that trans-modifiers and epigenetic variation are involved in the control of allelic differences in expression, in addition to polymorphisms in cis-regulatory sequences. Such extensive variation in allelic expression must have a large impact in generating phenotypic diversity.

2.2. *Variability in the biochemical "marks" associated with epigenetic variation*

The types of epigenetic marks that result in allelic variation in gene expression can be of diverse nature. The best known and most extensively investigated are covalent modifications of DNA and core histones. DNA methylation at CpG sites shows a degree of variability between different individuals at multiple loci. This is the case for imprinted genes like *IGF2/H19* and *IGF2R*, for which interindividual variation in methylation patterns has been observed in the differentially methylated regions associated with their expression (Sandovici *et al*., 2003). Interestingly, alterations in normal methylation patterns of these regions have been associated with loss of imprinting (LOI), a common observation in several types of cancer (Cui *et al*., 1998; Nakagawa *et al*., 2001). Another interesting example of interindividual variation at an imprinted gene is *PEG1*: this gene codes two isoform, one imprinted (isoform 1) and one expressed biallelically in multiple tissues (isoform 2). However, in a large subset of human placentae, isoform 2 allelic expression differences are observed, as well as interindividual variation in methylation of an associated CpG island (McMinn *et al*., 2006).

Interindividual variability in methylation patterns has been also described outside of imprinted genes or even protein coding regions: this is the case of methylation differences between humans that is observed in specific Alu repeated sequences (Sandovici *et al*., 2005). These observations reflect the fact that DNA methylation may have roles in addition to transcriptional control (de la Casa-Esperon and Sapienza, 2003; Pardo-Manuel de Villena *et al*., 2000; Sandovici *et al*., 2005).

Studies in other organisms also support the idea that variation in DNA methylation could be a widespread phenomenon. For instance, variation in cytosine methylation has been described in rRNA genes of natural accessions of the flowering plant *Arabidopsis thaliana* (Riddle and Richards, 2002), as well as in retrotransposons (Rangwala *et al*., 2006). Also, differentially methylated P1 pigment gene alleles have been observed in maize (Das and Messing, 1994). Importantly, studies in Arabidopsis have also shown that both natural and induced methylation changes can be transmitted to the offspring and result in developmental abnormalities in some instances (Kakutani *et al*., 1999; Rangwala *et al*., 2006; Riddle and Richards, 2005).

3. Epigenetic variation: what are the causes?

Epigenetic variation is the result of three types of processes: stochastic, environmental, and heritable. Variation in X-inactivation illustrates all three of these processes: during embryogenesis, one of the two X chromosomes is inactivated in each cell and clonally transmitted through successive mitotic divisions. Because this choice has a stochastic component (although some deterministic models are also capable of explaining the observations (Williams and Wu, 2004)), the X-inactivation patterns of a population of females approximates a normal distribution. The average female has about half of her cells with the maternal X chromosome inactive and half with the paternal X chromosome inactive. However, a small proportion of females show skewed patterns with a particular X chromosome being inactive in most cells (Figure 1a). Therefore, females are a mosaic for the expression of X-linked genes, and not even genetically identical females need show the same mosaic pattern.

The so-called skewing of X-inactivation is not always the rare consequence of the stochastic nature of the choice process. In some instances, skewing is the result of selection against X chromosomes carrying deleterious mutations, and the cell type-specificity of this skewing, as in X-linked agammaglobulinemia (skewing for inactivation of the mutant *XLA/BTK* allele in B-lymphocytes but not in T-lymphocytes), highlights the role of functional cellular selection (Fearon *et al*., 1987; reviewed in Belmont, 1996). In addition, skewing appears more common in older women, which suggests the contribution of environmental factors throughout their lifespan (Busque *et al*., 1996; Gale *et al*., 1997; Sharp *et al*., 2000). In this regard, X-inactivation seems to remain quite stable over many years during earlier ages (Sandovici *et al*., 2004) (Figure 1b). Many older females, however, exhibit substantial changes over the timescales at which younger females do not exhibit changes (Sandovici *et al*., 2004). In this regard, we have speculated (Sandovici *et al*., 2004) that acquired skewing of X-inactivation in older females may result from discontinuous or catastrophic processes that result in decreased numbers of stem cells or an age-related tendency toward bone marrow clonality or myelodysplasia.

Additionally, preference for the inactivation of a particular X chromosome can have a completely different origin compared to the selection for particular clonal cell populations or against disadvantageous mutations. Several studies in human and mice have shown that preference for X-inactivation can be heritable and genetically

controlled (Cattanach and Isaacson, 1967; Naumova *et al*., 1996, 1998; Plenge *et al*., 1997) In the mouse, the X-controlling element (*Xce*) is well known for its participation in the X-inactivation choice, so chromosomes carrying different alleles of *Xce* have different probabilities of being inactivated (Cattanach and Isaacson, 1967) (Figure 1c). Additional autosomal loci also participate in the genetic control of the choice of the X chromosome to be inactivated in mice (Chadwick and Willard, 2005; Percec *et al*., 2002, 2003). Moreover, parent-of-origin effects have also been observed in both mice (Takagi and Sasaki, 1975) and humans (Chadwick and Willard, 2005).

Stochastic, environmental, and genetic factors result in variability in X-chromosome inactivation and, consequently, generate a gamut of phenotypes for each of the X-linked genes, with multiple implications. The relative abundance of transcripts of each allele of any gene subject to X-inactivation reflects the fraction of cells with each of the two chromosomes active, as well as any allelic differences in expression that are intrinsic to specific alleles. Variations in such relative expression result in the spectrum of phenotypes observed in the population. For instance, a correlation between X-inactivation patterns and meiotic recombination levels (genomewide) has been described in female mice (de la Casa-Esperon *et al*., 2002). The biological importance of this trait (recombination levels) in the human population cannot be overestimated as it is a major determinant of female fecundity and reproductive lifespan. If recombination levels are controlled by gene/s in the X chromosome, then levels of recombination can change accordingly with the relative expression of different alleles of such gene/s. Because this is only one of the numerous genes in the X chromosome, the phenotypic diversity generated by similar phenomena related to X-inactivation processes is expected to be large in female mammals.

Similarly, epigenetic variability between individuals at multiple autosomal loci can be the result of multiple processes. Since erasure and establishment of epigenetic marks is a dynamic process that occurs during the lifespan of organisms, especially during gametogenesis and embryogenesis (reviewed in Latham, 1999; Mann and Bartolomei, 2002; Article 33, **Epigenetic reprogramming in germ cells and preimplantation embryos**, Volume 1), there is ample room for stochastic factors to contribute to the diversity of patterns observed. Environmental effects have also been described. Nutritional factors can induce epigenetic modifications such as changes in the expression of imprinted genes; moreover, maternal diet can affect the methylation status of transposable elements and the expression of nearby genes in mice (reviewed in Waterland and Jirtle, 2004). Examples of environmental effects have also been reported in rats, in which variations in maternal care behavior result in epigenetic changes in the offspring at the level of histone acetylation and DNA methylation of the consensus sequence for the NGFI-A transcription factor of the glucocorticoid receptor gene. Consequently, expression of this gene in the hippocampus can be modified by maternal care, which might be the basis for the changes in stress response observed in this gene in the offspring (reviewed in Fish *et al*., 2004). Environmental effects could be also the basis for the changes observed in epigenetic marks over time. DNA methylation patterns change with aging in a complex fashion, although overall hypomethylation has been observed in most vertebrate tissues (Mays-Hoopes *et al*., 1986; Richardson, 2003). For

instance, changes in the methylation profile of the c-*myc* proto-oncogene have been described during the aging process of mice. Because this is a gene involved in many tumor processes, similar temporal alterations of epigenetic marks might be part of the basis of the increasing incidence of cancer with age (Ono *et al*., 1986, 1989).

Finally, epigenetic diversity can be the result of heritable variants that affect the formation or stability of epigenetic marks. It has been observed that allelic differences in the expression of several genes are transmitted in families, although the patterns of transmission are variable (Pastinen *et al*., 2004; Yan *et al*., 2002). In some instances, the transmission of allelic imbalance is compatible with Mendelian inheritance, and even associated with transmission of particular polymorphisms (haplotypes), suggesting the participation of cis-acting elements in the regulation of allelic expression (Yan *et al*., 2002), whether they are of genetic or epigenetic origin. In fact, studies showing transmission of de novo induced methylation changes indicate that chromatin modifications, per se, are heritable (Kakutani *et al*., 1999; Stokes *et al*., 2002). Moreover, abnormal methylation patterns at the differentially methylated regions of the *IGF2/H19* and *IGF2R* imprinted genes have been found to cluster in families (Sandovici *et al*., 2003). Also, methylation levels at particular Alu repeated sequences show interindividual differences when the insertions were paternally versus maternally transmitted (Sandovici *et al*., 2005). In the case of imprinting defects, epimutations in an imprinting control region of human chromosome 15 have been associated with a substantial percentage of cases of the neurodevelopmental disorders Angelman and Prader–Willi syndromes (see Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1). Recent studies have shown that both cis- and trans-acting factors seem to increase the risk of conceiving a child with Angelman syndrome (AS) (Zogel *et al*., 2006). Trans-acting genetic elements have also been involved in changes in the imprinting status of the *Dlk1* gene in mouse brain (Croteau *et al*., 2005). In this case, reactivation of the normally silent maternal allele correlates with the methylation status of a differentially methylated region. Therefore, epigenetic information constitutes a code superimposed on the genetic information, thereby increasing phenotypic diversity. Much future research will no doubt focus on determining whether epigenetic variation makes a significant contribution to common "complex genetic disorders", such as diabetes, hypertension, schizophrenia, Alzheimer's disease and the like, in humans.

4. Epigenetic variation: what are the consequences?

Phenotypic diversity is the direct consequence of much epigenetic variation. As we mentioned before, epigenetic modifications can result in allelic expression imbalance within (differential expression levels) or between cells (monoallelic and mosaic expression). This, in turn, can result in phenotypic differences between cells, tissues, and/or individuals. The most obvious example is that of monozygotic twins: although genetically identical, numerous phenotypic differences appear during their life span. The same is true at the epigenetic level: recent studies have shown that differences in DNA methylation and histone acetylation between twins are present

throughout the genome (Fraga *et al*., 2005). Therefore, epigenetic differences could be the basis of many phenotypic discordances observed between twins, including their susceptibility to complex diseases (Wong *et al*., 2005).

4.1. Epigenetic variation and disease

Epigenetic variation is particularly important for genes involved in diseases. For instance, the fragile-X syndrome of mental retardation is associated with an expansion in the number of CGG repeats in the promoter and 5′ untranslated region of the *FMR1* gene on chromosome X. This expansion results in hypermethylation of the region and silencing of the *FMR1* gene (Hansen *et al*., 1992). Short expansions (premutations) do not have apparent phenotypic effects, while long expansions are observed in affected individuals. Notably, the severity of the disease ranges from severe mental retardation to only mild learning disabilities. It is possible that the observed gamut of symptoms depends, at least in part, on epigenetic differences, because variability in methylation in this region has been observed between and within individuals (Genc *et al*., 2000; Stoger *et al*., 1997) and changes in the CGG repeat length might also result in additional chromatin and transcriptional modifications.

Another interesting example of mosaicism has been observed in a small group of AS patients, in whom an imprinting defect silences the maternal copy of the *UBE3A* gene. However, some of these patients show mosaic maternal expression and methylation of this gene, which, again, suggests the possibility of an epigenetic effect on the observed variability in the severity of clinical symptoms (Nazlican *et al*., 2004).

Cancer has also been associated with epigenetic alterations, such as losses and gains of methylation and LOI (Feinberg *et al*., 1988, 2002; Cui *et al*., 2003; Jones and Baylin, 2002; Nakagawa *et al*., 2001). Interestingly, some of these alterations are also observed in normal tissues of the same individuals, as highlighted by the gain of DNA methylation in the imprinting control region upstream of *H19* in human Wilms tumors and in the non-neoplastic kidney parenchyma adjacent to these tumors (Cui *et al*., 1998, 2003; Moulton *et al*., 1994). Hence, epigenetic variation between individuals is probably involved in susceptibility to develop cancer as well as other genetic diseases. Moreover, since heritable epigenetic variation has been observed in many instances, it can actually play an important role in quantitative trait variation, and selection acting on such epialleles might result in rapid phenotypic changes, making it a formidable force in evolution (Rutherford and Henikoff, 2003; Sollars *et al*., 2003).

4.2. Epigenetic variation and development

Epigenetic variation also has important consequences in development and differentiation. A potentially important example of epigenetic changes as a result of environmental effects is the effects of culture conditions on the expression of imprinted genes in mouse embryos. It has been shown that some culture media

perturbs gene expression and results in aberrant methylation and expression of imprinted genes (Doherty *et al*., 2000; Mann *et al*., 2004; Rinaudo and Schultz, 2004). Although some of these abnormalities can be restored in the embryo proper (Mann *et al*., 2004), many persist in the extraembryonic tissues and can potentially affect the development of the embryo. In fact, several epidemiological studies suggest that assisted reproductive technologies (ART) might result in an increased frequency of diseases caused by imprinting defects, such as AS and Beckwith-Wiedemann syndrome (BWS) (Article 30, **Beckwith–Wiedemann syndrome**, Volume 1).

Despite the many reassuring reports on the safety of ART, there have been a small number of recent reports suggesting that ART children may be at increased risk for rare congenital malformation syndromes that are related to defects in genome imprinting (Cox *et al*., 2002; DeBaun *et al*., 2003; Halliday *et al*., 2004; Horsthemke *et al*., 2003; Niemitz *et al*., 2004; Olivennes *et al*., 2001; Orstavik *et al*., 2003). At least three children conceived by intracytoplasmic sperm injection (ICSI) have been diagnosed with AS (Horsthemke *et al*., 2003; Orstavik *et al*., 2003) and at least 28 ART children (both in vitro fertilization (IVF) and ICSI cases) have been diagnosed with BWS (Boerrigter *et al*., 2002; Bonduelle *et al*., 2002; DeBaun *et al*., 2003; Gicquel *et al*., 2003; Halliday *et al*., 2004; Koudstaal *et al*., 2000; Maher *et al*., 2003; Olivennes *et al*., 2001; Sutcliffe *et al*., 1995). Because both AS and BWS are rare disorders (each affects approximately 1 in 15 000 children (Nicholls *et al*., 1998)), the appearance of even small numbers of cases is unexpected except among a large sample of births. Therefore, the current data strongly suggests that there is an association between increased risk for AS and BWS and ART. With respect to BWS, the number of affected individuals observed is estimated to be up to nine times the expected incidence (Halliday *et al*., 2004).

The epidemiological assessment that ART may lead to an increase in the frequency of defective genome imprints is also supported by biochemical characterization of alleles at the relevant disease loci. All three cases of AS show allelic DNA methylation patterns characteristic of a sporadic imprinting defect at the AS locus (i.e., complete or mosaic absence of methylation on both maternal and paternal alleles (Horsthemke *et al*., 2003; Orstavik *et al*., 2003)). None of the patients has a cytogenetically visible alteration of chromosome 15 (which occurs in 70% of all AS cases (Nicholls *et al*., 1998)) and none has a detectable microdeletion at the imprinting center, suggesting that all three cases are due to sporadic, primary, epigenetic defects rather than genetic changes. Given that such imprinting defects account for less than 5% of all AS cases (Buiting *et al*., 2001, 2003; Nicholls *et al*., 1998), there is at least a suspicion that all three cases occurring in patients following ICSI are of this type.

The case for the presence of primary epigenetic defects in the majority of the BWS patients found among ART children is also supported by molecular analyses of alleles at the BWS locus on chromosome 11. Nineteen of the 24 patients have been analyzed for "loss of imprinting" ("LOI"; defined, in this context, as transcription of both maternal and paternal alleles; or the specific changes in DNA methylation that track with this phenomenon and provide a more robust marker in clinical samples) at one or more imprinted genes within the BWS locus and 13 of the 19 cases showed LOI at either *KCNQ1OT1* (DeBaun *et al*., 2003; Gicquel

et al., 2003; Maher *et al*., 2003) or *H19/IGF2* (DeBaun *et al*., 2003). With the addition of the BWS patients described by Halliday *et al*., 2004, 16 out of a total of 22 cases examined showed LOI. Although imprinting defects are more common in BWS than in AS, LOI still appears to be overrepresented among BWS cases in ART children and ART is, in turn, overrepresented among BWS cases.

4.3. Epigenetic variation diversity

During the last several years, there has been a dramatic increase in the number of studies attempting to elucidate the patterns and interrelationship between DNA methylation, histone modifications, noncoding RNAs, binding of nonhistone chromatin proteins, nuclear positioning and interactions, and so on, which are part of the "epigenetic code" (Article 27, **The histone code and epigenetic inheritance**, Volume 1). Alterations of the chromatin configuration can affect interactions between DNA regions, between chromosomes, and with other molecules. Most of the studies in epigenetic variation have been focused on the different mechanisms and effects on gene expression and its phenotypic consequences, including allelic differences and disease, enhancers and insulators, trans-sensing and paramutation, long-range interactions and nuclear colocation, and so on. However, epigenetic changes have also been found to affect many other chromosomal functions (see the following text). A classical example is the centromere, in which multiple chromatin modifications and proteins play a major role in binding to the poles of the spindle and promoting chromosome segregation. Interestingly, epigenetic changes can generate new domains with similar properties (neocentromeres) that affect the segregation of chromosomes during mitosis and meiosis (Pardo-Manuel de Villena and Sapienza, 2001; Rhoades and Dempsey, 1966; Warburton, 2004). Consequently, changes in the segregation of chromosomes or chromatids can favor the transmission of particular alleles to the next generations, with important consequences in evolution and disease (Pardo-Manuel de Villena and Sapienza, 2001).

Another example of a biochemical process for which there is a strong epigenetic effect is asynchronous DNA replication. Asynchronous replication is characteristic of regions containing monoallelically expressed genes (Mostoslavsky *et al*., 2001; Simon *et al*., 1999) and, therefore, epigenetic differences seem to be the basis for the differential replication between homologs at such regions. Consequently, these chromosomal regions are interesting examples of how epigenetic modifications of chromosomal regions have not one but multiple effects (on replication and expression). In addition, a recent survey of asynchronously replicated regions have found that they are located in close proximity to areas of tandem gene duplication (Gimelbrant and Chess, 2006) – although whether such epigenetic marks play a role in chromosome stability in regions of duplications remains to be determined.

Meiotic pairing and recombination constitute another example of a cellular process in which epigenetic marking appears to play an important role. Functional and epigenetic differences between paternal and maternal chromosomes are a common observation in sexually reproducing organisms (reviewed in de la Casa-Esperon and Sapienza, 2003; Pardo-Manuel de Villena *et al*., 2000). However, only

a few of such differences have been associated with imprinted gene expression. Consequently, it has been postulated that parent-of-origin epigenetic differences share a common origin and function in all sexually reproducing organisms: to allow the recognition (and distinction) between homologous chromosomes during the processes of recombination and repair (de la Casa-Esperon and Sapienza, 2003; Pardo-Manuel de Villena *et al*., 2000). Indeed, a recent study has shown that DNA methylation has a role in early meiotic stages: mice deficient in the DNA methyltransferase 3-like (*Dnmt3L*) gene are sterile and display abnormal chromosome synapsis during meiosis (Bourc'his and Bestor, 2004). Curiously, normal expression of *Dnmt3L* occurs not in the meiotic cells, but in their precursors. Hence, the epigenetic signals must be inherited through multiple cell divisions. Such epigenetic signals are observed as DNA methylation of retrotransposons, which appear demethylated in *Dnmt3L* knockout male germ cells. While methylation participates in the normal silencing of mobile elements, retrotransposons are transcribed in the mutant mice. Therefore, *Dnmt3L* mutant mice represent an example of how epigenetic changes can not only affect transcription but can also reshape the genome by affecting synapsis and allowing the mobilization of retrotransposons into new locations, with multiple consequences. Consequently, studies of epigenetic variation cannot be restricted to effects on gene expression, because it can also modulate many other chromosome functions (de la Casa-Esperon and Sapienza, 2003; Pardo-Manuel de Villena *et al*., 2000; Sandovici *et al*., 2005).

5. Conclusions

When discussing epigenetic variation, it is important to remember that we know little about either the underlying mechanisms or the consequences. To mention a few recent examples, studies on the viable yellow allele of the mouse agouti locus (A^{vy}) have shown that the expression of the agouti gene is correlated with the methylation status of upstream sequences (Article 36, **Variable expressivity and epigenetics**, Volume 1). Interestingly, epigenetic inheritance at this locus is not due to such methylation marks, because they are erased during embryonic development (Blewitt *et al*., 2006). Therefore, other epigenetic marks are responsible for the transmission of this epiallele to the offspring. In this review, although we have mostly mentioned examples of variability in methylation (because it has been the most frequently studied epigenetic mark in mammals and is the first subject of the Human Epigenome Project (Eckhardt *et al*., 2004)), we hope that current and future studies will bring to light epigenetic variation at many other levels. For instance, studies of the effects of histone tail modifications at multiple amino acid residues are an expanding field, because the spectrum of modifications and residues affected continues to grow (Article 27, **The histone code and epigenetic inheritance**, Volume 1). In addition, new epigenetic marks and inheritance modes are likely to be discovered. For instance, the role of small RNAs on epigenetic changes has become prominent since the discovery of RNA interference in *Caenorhabditis elegans* (Fire *et al*., 1998). Recent studies have revealed striking new roles for RNA in non-Mendelian epigenetic inheritance, similar to paramutation in plants.

The homozygous wild-type progeny of mice that are heterozygous for a mutation in the *Kit* gene (Rassoulzadegan *et al*., 2006) are found to exhibit the white spotting phenotype that is characteristic of mice that carry a *Kit* mutation. Elaboration of this phenotype is related to the zygotic inheritance of abnormally processed RNAs of the normal allele.

A realistic description of the scale of epigenetic variation is hampered by the diversity of causes and consequences and because the mechanism by which many epigenetic marks are heritable remains obscure. An increasing number of studies are aiming to integrate profiles from different epigenetic marks and gene expression patterns of particular chromosomal regions, in order to better understand the possibilities of variations on the epigenetic code. The complexity and diversity of the epigenetic marks and their implications poses a tremendous challenge, but understanding the nature of the immense phenotypic diversity that surrounds us makes it worth the effort.

Further Readings

Kochanek S, Renz D and Doerfler W (1994) Variability in allelic DNA methylation in spermatozoa. *Human Genetics*, **94**, 203–206.

Maher ER (2005) Imprinting and assisted reproductive technology. *Human Molecular Genetics*, **14**(Spec No. 1), R133–R138.

References

Belmont JW (1996) Genetic control of X inactivation and processes leading to X-inactivation skewing. *American Journal of Human Genetics*, **58**, 1101–1108.

Blewitt ME, Vickaryous NK, Paldi A, Koseki H and Whitelaw E (2006) Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. *PLoS Genetics*, **2**, e49.

Boerrigter PJ, de Bie JJ, Mannaerts BM, van Leeuwen BP and Passier-Timmermans DP (2002) Obstetrical and neonatal outcome after controlled ovarian stimulation for IVF using the GnRH antagonist ganirelix. *Human Reproduction*, **17**, 2027–2034.

Bonduelle M, Liebaers I, Deketelaere V, Derde MP, Camus M, Devroey P and Van Steirteghem A (2002) Neonatal data on a cohort of 2889 infants born after ICSI (1991–1999) and of 2995 infants born after IVF (1983–1999). *Human Reproduction*, **17**, 671–694.

Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, **431**, 96–99.

Buiting K, Barnicoat A, Lich C, Pembrey M, Malcolm S and Horsthemke B (2001) Disruption of the bipartite imprinting center in a family with Angelman syndrome. *American Journal of Human Genetics*, **68**, 1290–1294.

Buiting K, Gross S, Lich C, Gillessen-Kaesbach G, el-Maarri O and Horsthemke B (2003) Epimutations in Prader-Willi and Angelman syndromes: a molecular study of 136 patients with an imprinting defect. *American Journal of Human Genetics*, **72**, 571–577.

Bunzel R, Blumcke I, Cichon S, Normann S, Schramm J, Propping P and Nothen MM (1998) Polymorphic imprinting of the serotonin-2A (5-HT2A) receptor gene in human adult brain. *Molecular Brain Research*, **59**, 90–92.

Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y, Maragh M and Gilliland DG (1996) Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood*, **88**, 59–56.

Carrel L and Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, 400–404.

de la Casa-Esperon E, Loredo-Osti JC, Pardo-Manuel de Villena F, Briscoe TL, Malette JM, Vaughan JE, Morgan K and Sapienza C (2002) X chromosome effect on maternal recombination and meiotic drive in the mouse. *Genetics*, **161**, 1651–1659.

de la Casa-Esperon E and Sapienza C (2003) Natural selection and the evolution of genome imprinting. *Annual Review of Genetics*, **37**, 349–370.

Cattanach BM and Isaacson JH (1967) Controlling elements in the mouse X chromosome. *Genetics*, **57**, 331–346.

Chadwick LH and Willard HF (2005) Genetic and parent-of-origin influences on X chromosome choice in Xce heterozygous mice. *Mammalian Genome*, **16**, 691–699.

Cox GF, Burger J, Lip V, Mau UA, Sperling K, Wu BL and Horsthemke B (2002) Intracytoplasmic sperm injection may increase the risk of imprinting defects. *American Journal of Human Genetics*, **71**, 162–164.

Croteau S, Roquis D, Charron MC, Frappier D, Yavin D, Loredo-Osti JC, Hudson TJ and Naumova AK (2005) Increased plasticity of genomic imprinting of Dlk1 in brain is due to genetic and epigenetic factors. *Mammalian Genome*, **16**, 127–135.

Cui H, Cruz-Correa M, Giardiello FM, Hutcheon DF, Kafonek DR, Brandenburg S, Wu Y, He X, Powe NR and Feinberg AP (2003) Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science*, **299**, 1753–1755.

Cui H, Horon IL, Ohlsson R, Hamilton SR and Feinberg AP (1998) Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. *Nature Medicine*, **4**, 1276–1280.

Dao D, Frank D, Qian N, O'Keefe D, Vosatka RJ, Walsh CP and Tycko B (1998) IMPT1, an imprinted gene similar to polyspecific transporter and multi-drug resistance genes. *Human Molecular Genetics*, **7**, 597–608.

Das OP and Messing J (1994) Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. *Genetics*, **136**, 1121–1141.

DeBaun MR, Niemitz EL and Feinberg AP (2003) Association of in vitro fertilization with Beckwith-Wiedemann syndrome and epigenetic alterations of LIT1 and H19. *American Journal of Human Genetics*, **72**, 156–160.

Disteche CM (1995) Escape from X inactivation in human and mouse. *Trends in Genetics*, **11**, 17–22.

Doherty AS, Mann MR, Tremblay KD, Bartolomei MS and Schultz RM (2000) Differential effects of culture on imprinted H19 expression in the preimplantation mouse embryo. *Biology of Reproduction*, **62**, 1526–1535.

Eckhardt F, Beck S, Gut IG and Berlin K (2004) Future potential of the Human Epigenome Project. *Expert Review of Molecular Diagnostics*, **4**, 609–618.

Fearon ER, Winkelstein JA, Civin CI, Pardoll DM and Vogelstein B (1987) Carrier detection in X-linked agammaglobulinemia by analysis of X-chromosome inactivation. *New England Journal of Medicine*, **316**, 427–431.

Feinberg AP, Cui H and Ohlsson R (2002) DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. *Seminars in Cancer Biology*, **12**, 389–398.

Feinberg AP, Gehrke CW, Kuo KC and Ehrlich M (1988) Reduced genomic 5-methylcytosine content in human colonic neoplasia. *Cancer Research*, **48**, 1159–1161.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

Fish EW, Shahrokh D, Bagot R, Caldji C, Bredy T, Szyf M and Meaney MJ (2004) Epigenetic programming of stress responses through variations in maternal care. *Annual New York Academy of Sciences*, **1036**, 167–180.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, *et al* (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10604–10609.

Gale RE, Fielding AK, Harrison CN and Linch DC (1997) Acquired skewing of X-chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age. *British Journal of Haematology*, **98**, 512–519.

Genc B, Muller-Hartmann H, Zeschnigk M, Deissler H, Schmitz B, Majewski F, von Gontard A and Doerfler W (2000) Methylation mosaicism of 5′-(CGG)(n)-3′ repeats in fragile X, premutation and normal individuals. *Nucleic Acids Research*, **28**, 2141–2152.

Gicquel C, Gaston V, Mandelbaum J, Siffroi JP, Flahault A and Le Bouc Y (2003) In vitro fertilization may increase the risk of Beckwith-Wiedemann syndrome related to the abnormal imprinting of the KCN1OT gene. *American Journal of Human Genetics*, **72**, 1338–1341.

Gimelbrant AA and Chess A (2006) An epigenetic state associated with areas of gene duplication. *Genome Research*, **16**, 723–729.

Halliday J, Oke K, Breheny S, Algar E and Amor DJ (2004) Beckwith-Wiedemann syndrome and IVF: a case-control study. *American Journal of Human Genetics*, **75**, 526–528.

Hansen RS, Gartler SM, Scott CR, Chen SH and Laird CD (1992) Methylation analysis of CGG sites in the CpG island of the human FMR1 gene. *Human Molecular Genetics*, **1**, 571–578.

Horsthemke B, Nazlican H, Husing J, Klein-Hitpass L, Claussen U, Michel S, Lich C, Gillessen-Kaesbach G and Buiting K (2003) Somatic mosaicism for maternal uniparental disomy 15 in a girl with Prader-Willi syndrome: confirmation by cell cloning and identification of candidate downstream genes. *Human Molecular Genetics*, **12**, 2723–2732.

Jones PA and Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, **3**, 415–428.

Kakutani T, Munakata K, Richards EJ and Hirochika H (1999) Meiotically and mitotically stable inheritance of DNA hypomethylation induced by ddm1 mutation of Arabidopsis thaliana. *Genetics*, **151**, 831–838.

Koudstaal J, Braat DD, Bruinse HW, Naaktgeboren N, Vermeiden JP and Visser GH (2000) Obstetric outcome of singleton pregnancies after IVF: a matched control study in four Dutch university hospitals. *Human Reproduction*, **15**, 1819–1825.

Latham KE (1999) Epigenetic modification and imprinting of the mammalian genome during development. *Current Topics in Developmental Biology*, **43**, 1–49.

Lin W, Yang HH and Lee MP (2005) Allelic variation in gene expression identified through computational analysis of the dbEST database. *Genomics*, **86**, 518–527.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH and Lee MP (2003) Allelic variation in gene expression is common in the human genome. *Genome Research*, **13**, 1855–1862.

Maher ER, Brueton LA, Bowdin SC, Luharia A, Cooper W, Cole TR, Macdonald F, Sampson JR, Barratt CL, Reik W, *et al* (2003) Beckwith-Wiedemann syndrome and assisted reproduction technology (ART). *Journal of Medical Genetics*, **40**, 62–64.

Mann MR and Bartolomei MS (2002) Epigenetic reprogramming in the mammalian embryo: struggle of the clones. *Genome Biology*, **3**, Reviews 1003.

Mann MR, Lee SS, Doherty AS, Verona RI, Nolen LD, Schultz RM and Bartolomei MS (2004) Selective loss of imprinting in the placenta following preimplantation development in culture. *Development*, **131**, 3727–3735.

Mays-Hoopes L, Chao W, Butcher HC and Huang RC (1986) Decreased methylation of the major mouse long interspersed repeated DNA during aging and in myeloma cells. *Developmental Genetics*, **7**, 65–73.

McMinn J, Wei M, Sadovsky Y, Thaker HM and Tycko B (2006) Imprinting of PEG1/MEST isoform 2 in human placenta. *Placenta*, **27**, 119–126.

Migeon BR, Axelman J and Beggs AH (1988) Effect of ageing on reactivation of the human X-linked HPRT locus. *Nature*, **335**, 93–96.

Mostoslavsky R, Singh N, Tenzen T, Goldmit M, Gabay C, Elizur S, Qi P, Reubinoff BE, Chess A, Cedar H, *et al* (2001) Asynchronous replication and allelic exclusion in the immune system. *Nature*, **414**, 221–225.

Moulton T, Crenshaw T, Hao Y, Moosikasuwan J, Lin N, Dembitzer F, Hensle T, Weiss L, McMorrow L, Loew T, *et al* (1994) Epigenetic lesions at the H19 locus in Wilms' tumour patients. *Nature Genetics*, **7**, 440–447.

Nakagawa H, Chadwick RB, Peltomaki P, Plass C, Nakamura Y and de La Chapelle A (2001) Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of H19-associated CTCF-binding sites in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 591–596.

Naumova AK, Olien L, Bird LM, Smith M, Verner AE, Leppert M, Morgan K and Sapienza C (1998) Genetic mapping of X-linked loci involved in skewing of X chromosome inactivation in the human. *European Journal of Human Genetics*, **6**, 552–562.

Naumova AK, Plenge RM, Bird LM, Leppert M, Morgan K, Willard HF and Sapienza C (1996) Heritability of X chromosome-inactivation phenotype in a large family. *American Journal of Human Genetics*, **58**, 1111–1119.

Nazlican H, Zeschnigk M, Claussen U, Michel S, Boehringer S, Gillessen-Kaesbach G, Buiting K and Horsthemke B (2004) Somatic mosaicism in patients with Angelman syndrome and an imprinting defect. *Human Molecular Genetics*, **13**, 2547–2555.

Nicholls RD, Saitoh S and Horsthemke B (1998) Imprinting in Prader-Willi and Angelman syndromes. *Trends in Genetics*, **14**, 194–200.

Niemitz EL, DeBaun MR, Fallon J, Murakami K, Kugoh H, Oshimura M and Feinberg AP (2004) Microdeletion of LIT1 in familial Beckwith-Wiedemann syndrome. *American Journal of Human Genetics*, **75**, 844–849.

Olivennes F, Mannaerts B, Struijs M, Bonduelle M and Devroey P (2001) Perinatal outcome of pregnancy after GnRH antagonist (ganirelix) treatment during ovarian stimulation for conventional IVF or ICSI: a preliminary report. *Human Reproduction*, **16**, 1588–1591.

Ono T, Takahashi N and Okada S (1989) Age -associated changes in DNA methylation and mRNA level of the c-myc gene in spleen and liver of mice. *Mutation Research*, **219**, 39–50.

Ono T, Tawa R, Shinya K, Hirose S and Okada S (1986) Methylation of the c-myc gene changes during aging process of mice. *Biochemical and Biophysical Research Communication*, **139**, 1299–1304.

Orstavik KH, Eiklid K, van der Hagen CB, Spetalen S, Kierulf K, Skjeldal O and Buiting K (2003) Another case of imprinting defect in a girl with Angelman syndrome who was conceived by intracytoplasmic semen injection. *American Journal of Human Genetics*, **72**, 218–219.

Pagani F, Toniolo D and Vergani C (1990) Stability of DNA methylation of X-chromosome genes during aging. *Somatic Cell and Molecular Genetics*, **16**, 79–84.

Paldi A (2003) Stochastic gene expression during cell differentiation: order from disorder? *Cell and Molecular Life Sciences*, **60**, 1775–1778.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR and Frazer KA (2006) Analysis of allelic differential expression in human white blood cells. *Genome Research*, **16**, 331–339.

Pardo-Manuel de Villena F, de la Casa-Esperon E and Sapienza C (2000) Natural selection and the function of genome imprinting: beyond the silenced minority. *Trends in Genetics*, **16**, 573–579.

Pardo-Manuel de Villena F and Sapienza C (2001) Nonrandom segregation during meiosis: the unfairness of females. *Mammalian Genome*, **12**, 331–339.

Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, *et al* (2004) A survey of genetic and epigenetic variation affecting human gene expression. *Physiological Genomics*, **16**, 184–193.

Percec I, Plenge RM, Nadeau JH, Bartolomei MS and Willard HF (2002) Autosomal dominant mutations affecting X inactivation choice in the mouse. *Science*, **296**, 1136–1139.

Percec I, Thorvaldsen JL, Plenge RM, Krapp CJ, Nadeau JH, Willard HF and Bartolomei MS (2003) An N-ethyl-N-nitrosourea mutagenesis screen for epigenetic mutations in the mouse. *Genetics*, **164**, 1481–1494.

Plenge RM, Hendrich BD, Schwartz C, Arena JF, Naumova A, Sapienza C, Winter RM and Willard HF (1997) A promoter mutation in the *XIST* gene in two unrelated families with skewed X-chromosome inactivation. *Nature Genetics*, **17**, 353–356.

Rangwala SH, Elumalai R, Vanier C, Ozkan H, Galbraith DW and Richards EJ (2006) Meiotically stable natural epialleles of sadhu, a novel arabidopsis retroposon. *PLoS Genetics*, **2**, e36.

Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I and Cuzin F (2006) RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, **441**, 469–474.

Rhoades MM and Dempsey E (1966) The effect of abnormal chromosome 10 on preferential segregation and crossing over in maize. *Genetics*, **53**, 989–1020.

Richardson B (2003) Impact of aging on DNA methylation. *Ageing Research Reviews*, **2**, 245–261.

Riddle NC and Richards EJ (2002) The control of natural variation in cytosine methylation in Arabidopsis. *Genetics*, **162**, 355–363.

Riddle NC and Richards EJ (2005) Genetic variation in epigenetic inheritance of ribosomal RNA gene methylation in Arabidopsis. *Plant Journal*, **41**, 524–532.

Rinaudo P and Schultz RM (2004) Effects of embryo culture on global pattern of gene expression in preimplantation mouse embryos. *Reproduction*, **128**, 301–311.

Rutherford SL and Henikoff S (2003) Quantitative epigenetics. *Nature Genetics*, **33**, 6–8.

Sakatani T, Wei M, Katoh M, Okita C, Wada D, Mitsuya K, Meguro M, Ikeguchi M, Ito H, Tycko B, *et al* (2001) Epigenetic heterogeneity at imprinted loci in normal populations. *Biochemical and Biophysical Research Communication*, **283**, 1124–1130.

Sandovici I, Kassovska-Bratinova S, Loredo-Osti JC, Leppert M, Suarez A, Stewart R, Bautista FD, Schiraldi M and Sapienza C (2005) Interindividual variability and parent of origin DNA methylation differences at specific human Alu elements. *Human Molecular Genetics*, **14**, 2135–2143.

Sandovici I, Leppert M, Hawk PR, Suarez A, Linares Y and Sapienza C (2003) Familial aggregation of abnormal methylation of parental alleles at the IGF2/H19 and IGF2 R differentially methylated regions. *Human Molecular Genetics*, **12**, 1569–1578.

Sandovici I, Naumova AK, Leppert M, Linares Y and Sapienza C (2004) A longitudinal study of X-inactivation ratio in human females. *Human Genetics*, **115**, 387–392.

Sharp A, Robinson D and Jacobs P (2000) Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. *Human Genetics*, **107**, 343–349.

Simon I, Tenzen T, Reubinoff BE, Hillman D, McCarrey JR and Cedar H (1999) Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature*, **401**, 929–932.

Sollars V, Lu X, Xiao L, Wang X, Garfinkel MD and Ruden DM (2003) Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nature Genetics*, **33**, 70–74.

Stoger R, Kajimura TM, Brown WT and Laird CD (1997) Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene FMR1. *Human Molecular Genetics*, **6**, 1791–1801.

Stokes TL, Kunkel BN and Richards EJ (2002) Epigenetic variation in Arabidopsis disease resistance. *Genes and Development*, **16**, 171–182.

Sutcliffe AG, D'Souza SW, Cadman J, Richards B, McKinlay IA and Lieberman B (1995) Minor congenital anomalies, major congenital malformations and development in children conceived from cryopreserved embryos. *Human Reproduction*, **10**, 3332–3337.

Takagi N and Sasaki M (1975) Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, **256**, 640–642.

Warburton PE (2004) Chromosomal dynamics of human neocentromere formation. *Chromosome Research*, **12**, 617–626.

Wareham KA, Lyon MF, Glenister PH and Williams ED (1987) Age related reactivation of an X-linked gene. *Nature*, **327**, 725–727.

Waterland RA and Jirtle RL (2004) Early nutrition, epigenetic changes at transposons and imprinted genes, and enhanced susceptibility to adult chronic diseases. *Nutrition*, **20**, 63–68.

Williams BR and Wu CT (2004) Does random X-inactivation in mammals reflect a random choice between two X chromosomes? *Genetics*, **167**, 1525–1528.

Wolff GL, Kodell RL, Moore SR and Cooney CA (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB Journal*, **12**, 949–957.

Wong AH, Gottesman II and Petronis A (2005) Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics*, **14**(Spec No. 1), R11–R18.

Xu Y, Goodyer CG, Deal C and Polychronakos C (1993) Functional polymorphism in the parental imprinting of the human IGF2 R gene. *Biochemical and Biophysical Research Communication*, **197**, 747–754.

Yan H, Yuan W, Velculescu VE, Vogelstein B and Kinzler KW (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.

Zogel C, Bohringer S, Gross S, Varon R, Buiting K and Horsthemke B (2006) Identification of cis- and trans-acting factors possibly modifying the risk of epimutations on chromosome 15. *European Journal of Human Genetics*, **14**, 752–758.

Specialist Review

How to get extra performance from a chromosome: recognition and modification of the X chromosome in male *Drosophila melanogaster*

Ying Kong and **Victoria H. Meller**
Wayne State University, Detroit, US

1. Differentiated sex chromosomes cause genetic imbalance

Many organisms have a single X chromosome in males and two in females. The X chromosome is gene-rich and carries genes required in both sexes. By contrast, the Y is often gene-poor, and may carry genes only required in males. The resulting imbalance in the dosage of X-linked genes is fatal if not addressed early in life. Equalization of expression between the sexes is an essential feature of differentiation in flies, mammals, and the worm *Caenorhabditis elegans*. Although the problem is common, the strategy used to solve it in each of these organisms is distinct (Figure 1). To equalize expression between *C. elegans* males (XO; one X chromosome but no Y chromosome) and XX hermaphrodites, hermaphrodites reduce transcription from both X chromosomes by 50%. Mammalian females silence most genes on a single X chromosome. The remaining X chromosome is transcriptionally equivalent to the single X chromosome of males. By contrast, *Drosophila* males increase expression from their single X chromosome about twofold. Although these methods for equalizing expression are overtly very different, each organism regulates the X chromosome through modulation of chromatin architecture.

These animals must accurately and selectively modulate a single chromosome or a pair of chromosomes in one sex. Interestingly, both mammals and flies use large noncoding RNAs to direct chromatin-modifying proteins that regulate expression. The large *Xist* (*X inactive specific transcript*) is transcribed from and directs silencing to the inactive X chromosome of mammalian females (reviewed by Plath *et al*., 2002). As silencing of both X chromosomes would be lethal, this process is restricted to chromatin *in cis* to a single *Xist* allele. *Drosophila* males have two large, noncoding transcripts, *roX1* (*RNA on the X1*), and *roX2* (*RNA on the X2*), that are necessary for localization of a complex of proteins and *roX* RNA to the male X chromosome (Meller and Rattner, 2002). In both mammals and flies, this process likely involves two steps: recruitment of a protein complex,

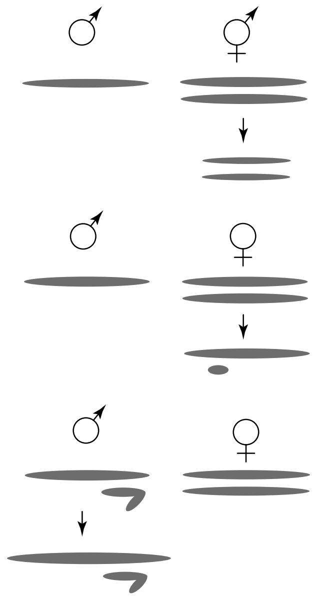


Figure 1 Organisms use divergent strategies to compensate sex chromosome gene dosage. (a) *C. elegans* hermaphrodites (right) have 2 X chromosomes, whereas males have a single X chromosome and no Y chromosome (left). Association of the repressive DCC complex reduces expression of both hermaphrodite X chromosomes by about 50%. (b) Mammalian females (right) randomly silence a single X chromosome. The remaining active X chromosome is transcriptionally equivalent to the single X chromosome of males. (c) *Drosophila* males increase expression from their X chromosome by modulation of chromatin structure (left). Female X chromosomes remain unchanged

followed by modulation of gene expression. This review will focus on advances in understanding the process of recognition and modulation in flies. It will center on the role of the *roX* transcripts in recognition and modification of the X chromosome.

2. RNA and protein coat the X chromosome of *Drosophila* males

Many of the genes necessary for dosage compensation in flies were identified through male-specific lethal (*msl*) mutations. These genes, *maleless* (*mle*), the *male-specific lethals 1*, *-2*, and *-3* (*msl1*, *-2*, and *-3*), and *males absent on first* (*mof*), are collectively known as the *male-specific lethals* (recently reviewed by Mendjan and Akhtar, 2007). Mutations in these genes cause developmental delay and lethality in males, but none is essential in females. The genes that encode the *roX* RNAs are X linked and functionally redundant for dosage compensation. Both properties make them unlikely to be identified by conventional mutagenesis and phenotypic analysis. Accordingly, the *roX* genes were discovered serendipitously (Amrein and Axel, 1997; Meller *et al*., 1997). Immunolocalization of MSL proteins or *in situ* hybridization to *roX* on polytene preparations reveals finely banded

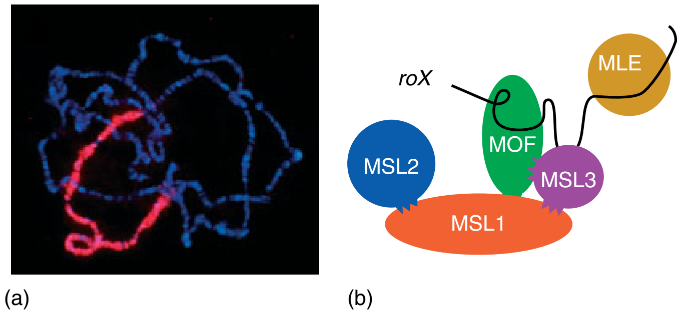


Figure 2 MSL proteins and *roX* RNA form a complex that binds to the *Drosophila* X chromosome. (a) Immunodetection of MSL2 on a polytene chromosome preparation from a male larva. The X chromosome binds MSL2, MSL3 detected with Texas Red. DNA appears blue. (b) Molecular interactions between MSL proteins and RNA. Interactions between proteins are denoted by teeth. Potential interactions are modeled between a single *roX* transcript (black line) and proteins reported to have RNA-binding activity. Protein/protein and protein/RNA interactions are reported by Akhtar *et al*. (2000); Buscaino *et al*. (2003); Copps *et al*. (1998); Li *et al*. (2005); Morales *et al*. (2004); Scott *et al*. (2000)

enrichment along the X chromosome (Figure 2). The MSL proteins and *roX* RNA coimmunoprecipitate, demonstrating that they form a complex (Meller *et al*., 2000; Smith *et al*., 2000). Removal of individual members of the complex disrupts its localization and can reduce the stability of remaining molecules. This is particularly dramatic for the *roX* RNAs, which are unstable upon elimination of any MSL protein (Meller *et al*., 1997; Amrein and Axel, 1997). Mutation of *mle*, *msl3*, or *mof* reduces X-chromosome binding by remaining members of the complex, but a subset of sites able to bind the remaining proteins is retained on the X chromosome. The most prominent of these sites are the *roX* genes themselves (reviewed by Kelley, 2004). MSL1 and MSL2 have a more central role in regulation and assembly of the MSL complex as elimination of either of these proteins prevents all chromatin binding by remaining complex members (Lyman *et al*., 1997). In contrast, simultaneous elimination of both *roX* RNAs shifts the MSL proteins from the X chromosome to ectopic autosomal sites and results in reduced X-linked gene expression (Deng and Meller, 2006; Deng *et al*., 2005; Meller and Rattner, 2002). Recognition of the X chromosome is thus a property of the intact MSL complex, and is not attributable solely to a single participating molecule.

3. Proteins associated with the MSL complex modify chromatin

Increased expression of the male X chromosome is believed to result from changes in chromatin architecture induced by MSL complex. One member of the MSL complex, MOF, is an acetyltransferase specific for lysine 16 of H4 (Akhtar and Becker, 2000; Hilfiker *et al*., 1997; Smith *et al*., 2000). This modification is generally associated with active chromatin, and is highly enriched on the male X chromosome of flies (Turner *et al*., 1992). Acetylation of H4K16 by MOF

increases transcription *in vitro* and *in vivo* (Akhtar and Becker, 2000). Effector proteins that mediate transcriptional change bind to some modified histones, but none specific for H4K16Ac has been found. A recent study demonstrated that acetylation of H4K16 inhibits the formation of highly compact chromatin by disrupting charge-based internucleosomal interactions (Shogren-Knaak *et al*., 2006). This structural effect partially decondenses chromatin, thereby increasing the accessibility of the DNA template. In humans, H4acK16 is found uiquitously on all chromosomes except for the inactive X chromosome (Jeppesen and Turner, 1993). In flies, H4AcK16 colocalizes with the MSL complex (Bone *et al*., 1994; Turner *et al*., 1992).

A second modification linked to increased expression, phosphorylation of H3 on serine 10 (H3pS10), is also enriched on the male X chromosome (Jin *et al*., 1999; Mahadevan *et al*., 2004). H3pS10 in interphase cells is directed by the JIL-1 kinase (Wang *et al*., 2001). Proper dosage compensation of the X-linked *white* gene requires JIL-1 function (Lerach *et al*., 2005). In addition to compensation of the male X chromosome, JIL-1 has a general role in maintenance of chromatin structure and limits the spread of heterochromatin into euchromatic regions (Zhang *et al*., 2005). Accordingly, JIL-1 is an essential gene required in both sexes (Wang *et al*., 2001). The male X chromosome is therefore marked with at least two histone modifications that are associated with elevated transcription and decreased chromatin condensation. It is likely that the primary function of the MSL complex is to direct and control these modifications.

4. The MSL complex increases transcription by a general method

X chromosome compensation affects hundreds of genes with different expression levels and profiles. It must therefore be superimposed on genes with distinct regulatory strategies. Interestingly, chromatin immuno and affinity precipitation of DNA bound by the MSL complex detects modest levels of these proteins in promoter regions, but higher levels within the body of most actively transcribed genes (Alekseyenko *et al*., 2006; Gilfillan *et al*., 2006; Legube *et al*., 2006). H4Ac16 has also been found to be high in the body of X-linked genes (Smith *et al*., 2001). Reduced chromatin compaction may increase the speed or processivity of RNA polymerase. Enhanced expression is thus likely to result from facilitation of transcriptional elongation, rather than increased initiation (Henikoff and Meneely, 1993). Alternatively, modifications at the 3′ end of transcription units may enhance reinitiation by recycled RNA polymerase (Dieci and Sentenac, 2003).

A second theory of how the MSL complex enhances expression stems from a recent study that identified nuclear pore components copurifying with tagged MOF and MSL3 (Mendjan *et al*., 2006). This study found no classical transcriptional factors but identified exosome components and interband binding proteins in association with MSL proteins. Knock down of the nuclear pore proteins Mtor and Nup153 disrupted the location of MSL proteins and compensation of some X-linked genes, suggesting that interaction with the nuclear pore is important for

localization of the MSL complex (Mendjan *et al*., 2006). An association with the nuclear pore might facilitate transcriptional elongation by affecting RNA processing and transportation. The nuclear pore might alternatively establish a transcriptionally active compartment or a region of facilitated chromatin remodeling within the nucleus (Casolari *et al*., 2005; Feuerbach *et al*., 2002). Tethering transcription units to nuclear pores facilitates expression in yeast, supporting the involvement of this structure in activation (Cabal *et al*., 2006; Taddei *et al*., 2006). Recruitment of silenced genes into a repressive nuclear compartment has been proposed as the mechanism of X-chromosome inactivation in mammals. The inactive X chromosome (X_i) occupies a region adjacent to the nucleolus during replication (Zhang *et al*., 2007). This may ensure epigenetic maintenance of the silent state through replication. X-linked gene inactivation is accompanied by movement of individual genes from the outer fringe of the domain occupied by the X_i to a more interior position from which RNA polymerase is excluded (Chaumeil *et al*., 2006). Thus X inactivation and its perpetuation may rely on recruitment of genes and the X chromosome into specific nuclear compartments. It is unknown if dosage compensation in flies involves repositioning of genes into transcriptionally active domains. However, this method is appealing as it is well adapted to control closely linked genes and could do so by a general method that is superimposed on genes with different regulatory strategies.

5. Large RNAs that control X chromosomes: powerful but mysterious molecules

Regulatory RNAs that coat the X chromosome play a key role in dosage compensation in mammals and *Drosophila*. In spite of the central role of *roX* transcripts in fly dosage compensation, how they interact with the MSL proteins, and how this changes the properties of the MSL complex, remains speculative. A comparison with *Xist* may prove valuable. *Xist* is well studied and shares unusual properties with the *roX* RNAs. Both RNAs coat the dosage-compensated X chromosomes, direct protein complexes to chromatin, and are able to recruit chromatin-modifying activities *in cis* to the site of RNA synthesis.

Xist is transcribed from the *Xic* (*X inactivation center*) and is essential for initiation and propagation of X-chromosome inactivation (reviewed by Plath *et al*., 2002). *Xist* is selectively expressed from one X chromosome and spreads *in cis* from the site of synthesis to coat most of this chromosome. *Xist* recruits polycomb proteins that introduce repressive histone modifications (Plath *et al*., 2004; Silva *et al*., 2003). Several days after the initiation of X inactivation, the inactivation becomes largely independent of *Xist*. Additional changes in chromatin, such as enrichment for variant histones and methylation of CpG islands, characterize the differentiated X_i (reviewed by Lucchesi *et al*., 2005). Distinct sequences within *Xist* are responsible for localization to the X chromosome and for silencing (Figure 3a; Wutz *et al*., 2002). Several widely separated *Xist* sequences act cooperatively to direct X localization, and a repeated element that folds into shortstem loops mediates silencing as well as localization (Wutz *et al*., 2002).

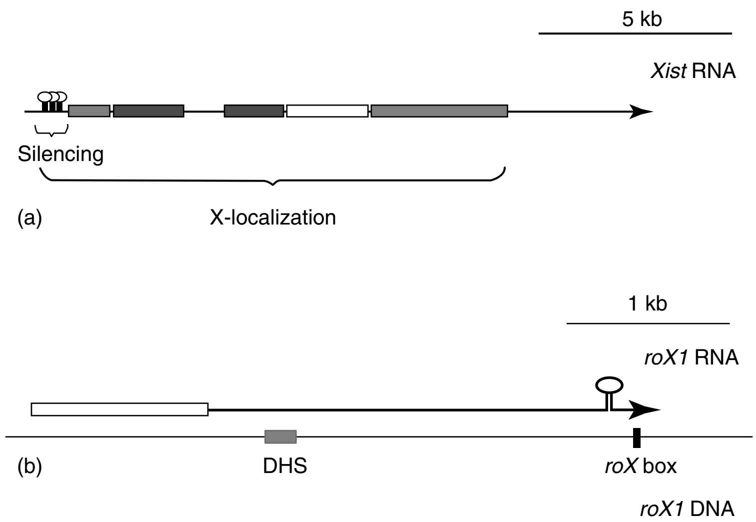


Figure 3 *roX1* and *Xist* have distinct regions necessary for gene function. (a) The *Xist* transcript has a series of 15 shortstem loops near the 5′ end that are necessary for silencing. Distributed elements that contribute to X localization are shown as gray and white boxes. The strongest of these are darkest. Figure is based on (Wutz, 2003; Wutz *et al*., 2002). (b) Functional and conserved regions of *roX1* RNA (top) and DNA (bottom) are represented. One kb at the 5′ end of *roX1* (open box on left) is necessary for wild-type localization of the MSL complex. Between this and the 3′ stem loop (right) there is no identified element necessary for RNA function. The 200 bp *roX1* DNase hypersensitive site (DHS) is shown as a gray box on the *roX1* DNA. This sequence attracts the MSL complex. The 30 bp "*roX* box" (black) is at the right. This is based on (Kageyama *et al*., 2001; Park *et al*., 2003; Stuckenholz *et al*., 2003)

This repeat is also necessary for relocalization of silenced genes inside the domain occupied by the X_i (Chaumeil *et al*., 2006).

roX1 also has multiple regions necessary for function (Figure 3b). In spite of being redundant, the *roX* transcripts share little similarity that can be used to identify potentially important sequences (Amrein and Axel, 1997; Park *et al*., 2003). A highly conserved 30 bp "*roX* box" is present at the 3′ end of both *roX* RNAs (Franke and Baker, 1999). This region is dispensable for function (Stuckenholz *et al*., 2003). A weakly conserved 200 bp sequence within each *roX* gene strongly attracts the MSL proteins and forms a male-specific DNaseI hypersensitive site (Kageyama *et al*., 2001; Park *et al*., 2003). The *roX1* DNase hypersensitive site (DHS) acts as an enhancer of *roX1* transcription in males and a repressor in females (Bai *et al*., 2004). However, internal deletions of *roX1* lacking the DHS are still regulated in a sex-specific manner, and *roX1* alleles and transgenes lacking this sequence retain full activity (Deng *et al*., 2005; Rattner and Meller, 2004; Stuckenholz *et al*., 2003). Although the role of the DHS remains speculative, all evidence points to its function as DNA, rather than RNA. Two *roX1* regions necessary for transcript activity have been identified. A stem loop close to the 3′ end is the only structural feature linked to *roX1* function (Stuckenholz *et al*., 2003). Transgenes deleted for the stem loop or with disrupted pairing of the stem have low rescue of *roX1*$^-$ *roX2*$^-$ males in spite of substantial recruitment of MSL

proteins to the X chromosome. It is possible that this sequence influences chromatin modification or gene activation by the MSL complex. Deletion of 1 kb at the 5′ end of *roX1* also destroys activity. When large portions of this region are removed by internal deletion, *roX1* activity is reduced commensurate with the amount deleted (Deng *et al*., 2005). Small (∼300 bp) deletions scanning the 5′ end have failed to identify discrete elements, suggesting redundancy (Stuckenholz *et al*., 2003). Males that carry a *roX1* allele with a large part of the 5′ end missing display ectopic MSL binding and reduced coverage of the X chromosome, suggesting that this region is necessary for recognition of the X chromosome (Deng *et al*., 2005). An internal deletion of 2.4 kb that retains 0.8 kb of 5′ end and 0.6 kb of the 3′ end, including the stem loop, supports full male survival (Deng *et al*., 2005). However, simultaneous expression of separate 3′ and 5′ fragments of *roX1* does not rescue either MSL localization or male survival (Meller and Rattner, 2002; Stuckenholz *et al*., 2003). Taken together, these observations suggest that *roX* activity requires simultaneous interaction with different molecules. An attractive model is that *roX1*, like *Xist*, has distinct domains necessary for X chromosome localization and gene activation.

The major *roX2* splice form is 600 bp, and functional domains within this molecule remain to be identified. A multitude of alternative *roX2* splice forms with decreased activity has been found (Park *et al*., 2005). *roX2* molecules with different levels of activity may modulate the activity of the MSL complex, thus fine-tuning the level of X chromosome activation.

6. MSL proteins have RNA-binding activity

MLE is an RNA/DNA helicase with higher activity on RNA substrates (Lee *et al*., 1997). The helicase activity of MLE is essential for normal localization of the MSL complex on the X chromosome and for movement of the *roX* RNAs from their sites of synthesis, suggesting a role in integration of *roX* into the mature MSL complex (Gu *et al*., 2000; Meller *et al*., 2000). MLE itself does not interact with other MSL proteins and can only be coimmunoprecipitated under nonstringent conditions using antibodies that pull down other MSL proteins (Smith *et al*., 2000). MLE can be released from polytene chromosomes by RNase A digestion, suggesting that it associates with the MSL complex through an RNA (Figure 2b; Richter *et al*., 1996). The stability of *roX1* is particularly dependent on MLE, supporting the idea of a direct interaction between these molecules (Meller, 2003).

Both MSL3 and MOF have RNA-binding activity *in vitro* and their localization on the X chromosome is destabilized by RNase digestion (Akhtar *et al*., 2000; Buscaino *et al*., 2003). Both proteins have variant chromo domains that have been implicated in RNA binding. Whereas the canonical chromo domains of Heterochromatin protein 1 (HP1) and Polycomb (Pc) bind methylated histones by aromatic residues, the variant structures found in MOF and MSL3, named *chromo barrel domains*, may have different functions (Lachner *et al*., 2001; Bannister *et al*., 2001). The MOF chromo barrel lacks aromatic residues that recognize methylated peptides (Nielsen *et al*., 2005). This region contributes to MOF’s ability to bind RNA *in vitro* (Akhtar *et al*., 2000). The chromo barrel of MSL3 has also been implicated in RNA binding, but retains the aromatic residues necessary for

methyl group binding (Nielsen *et al*., 2005). Mutation of the MSL3 chromo barrel prevents increased transcription of X-linked genes, but does not affect targeting of the complex to the X chromosome (Buscaino *et al*., 2006). Deletion of a different domain of MSL3 blocks X chromosome localization, reinforcing the idea that MSL complex localization and gene activation are separable.

These studies suggest that multiple RNA/protein contacts within the MSL complex might fine-tune the activity of the complex. The H4 acetyltransferase activity of MOF is greatly increased by association with MSL1 and MSL3, suggesting a mechanism for limiting MOF activity until it assembles with a regulatory complex (Morales *et al*., 2004). MSL1, -2, -3, and MOF continue to associate in the absence of *roX*, but only low levels of H4Ac16 are detected at sites bound by these proteins (Deng and Meller, 2006). This may reflect a reduced MOF activity in the absence of *roX* RNA.

7. Do *roX* and MLE recruit a preexisting chromatin-binding complex?

The discovery that the yeast NuA4 transcriptional regulator contains subunits similar to MSL3 and MOF Esa/p-associated factor (Eaf3p and Esa1p) and characterization of a mammalian complex containing MSL homologs suggests that the association of these proteins is ancient (Eisen *et al*., 2001; Marin and Baker, 2000; Smith *et al*., 2005; Taipale *et al*., 2005). Human MOF (hMOF) participates in multiple protein assemblies and is required for normal function of human ATM (ataxia-telangiectasia-mutated) protein in DNA repair (Gupta *et al*., 2005; Taipale *et al*., 2005).

roX RNAs have only been identified in closely related Drosophilids, but helicases with similarity to MLE have been identified from yeast to mammals (Park *et al*., 2003; Sanjuan and Marin, 2001). MLE homologs have not yet been isolated in complexes of MSL-like proteins outside of flies. MLE has a peripheral association with the fly MSL complex, and is presumably tethered by RNA. It thus seems plausible that the addition of MLE to the MSL complex depends on the presence of *roX*. The importance of *roX* in correct targeting of the MSL complex suggests that the addition of noncoding RNA was a major step in recruitment for the purpose of X chromosome compensation.

8. Recognition of X chromosomes

A mechanism that targets changes in expression to a single chromosome is a fundamental requirement of dosage compensation. Two distinct strategies for accomplishing this have been described. The chromosome may be controlled by spread of regulation from *cis*-acting elements. The *Xic* is a strong *cis*-acting element capable of directing silencing to an entire X chromosome (Figure 4a). It can also silence autosomal chromatin if *Xist* is inserted on the autosome. *Xist* RNA produced from the *Xic* does not work *in trans*, thus protecting one X chromosome from inactivation. An alternative mechanism for distinguishing a chromosome is finely dispersed

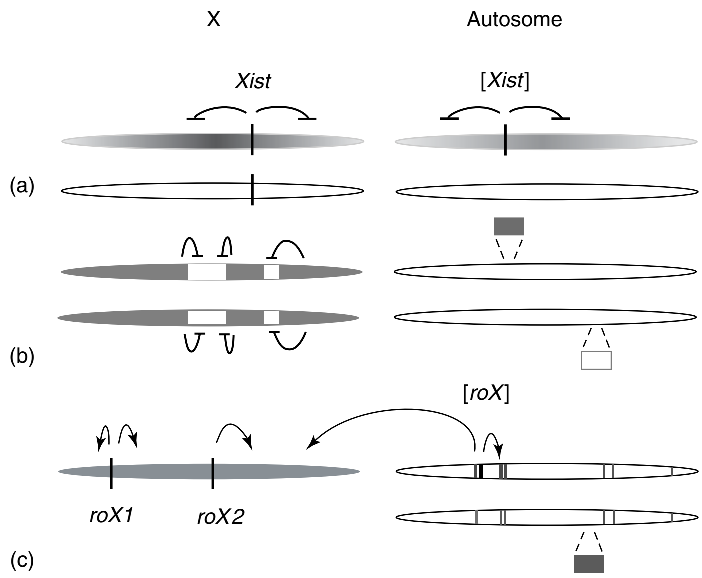


Figure 4 Strategies for X chromosome recognition in mammals, flies and worms. (a) One of two *X inactivation centers* present in females produces *Xist* RNA (top left). The chromosome carrying this allele becomes the inactive X (shaded). A transgene carrying *Xist* can silence autosomal chromatin *in cis* (shaded, right). (b) The *C. elegans* X chromosome is distinguished by sequence elements (gray shading). The distribution of these elements is uneven, leaving large gaps (white). The repressive DCC spreads into these gaps from flanking regions. Segments separated from the rest of the X chromosome attract the DCC if they have X-recognition sequences (autosomal insertion, top) but remain uncompensated if they lack these elements (autosomal insertion, bottom). (c) The *Drosophila* X chromosome is finely marked by sequences that attract the MSL complex (gray). Translocated X chromosome fragments are recognized accurately (autosomal insertion, bottom). Weak and scattered MSL-binding sites on the autosomes do not attract the MSL complex in normal males (gray lines, right). *roX1* and *roX2* (vertical black lines) produce *roX* RNA and are *cis*-acting elements that enhance recognition of the X chromosome. A *roX* transgene (top right) enables MSL binding to closely linked autosomal sites. The *roX* transgene also produces transcript that acts *in trans* to compensate an X chromosome

sequence elements. Two short sequences that participate in recognition of the *C. elegans* X chromosome have been identified (McDonel *et al*., 2006). Interestingly, these are not exclusive to the X but cooccur near one another on the X chromosome. This suggests that cooperativity between multiple DNA-binding molecules underlies recognition X chromatin in worms. However, large regions of the *C. elegans* X chromosome fail to bind the repressive dosage compensation complex (DCC) when separated from the X chromosome, but are coated by it when on the X chromosome (Figure 4b) (Csankovszki *et al*., 2004). This indicates that the ability of the DCC to spread *in cis* is necessary for a complete coverage of the *C. elegans* X chromosome.

Translocated segments of the *Drosophila* X chromosome are faithfully recognized by the MSL complex, indicating the presence of finely distributed sequences marking this chromosome (Figure 4c; Fagegaltier and Baker, 2004; Oh *et al*., 2004). Autosomal *roX* transgenes can fully rescue male viability, indicating that *roX* RNA can act *in trans* to its site of synthesis. But under

some conditions, autosomal *roX* insertions also direct MSL binding to chromatin flanking the insertion site (Kelley *et al*., 1999; Park *et al*., 2002). Regional spreading of chromatin modification from the *roX* genes can also be observed on the X chromosome (Bai *et al*., 2007; Oh *et al*., 2003). It therefore appears that recognition of the *Drosophila* X chromosome involves strong, *cis*-acting elements as well as sequences identifying the X chromosome.

Subdivision of DNA clones that recruit the MSL complex and a functional assay for MSL recruitment have identified short sequences that contribute to MSL binding (Dahlsveen *et al*., 2006; Gilfillan *et al*., 2007; Oh *et al*., 2004). These sequences are divergent and display a wide range in affinity for the MSL proteins. Overexpression of MSL proteins also identifies autosomal sites that can recruit the MSL complex, indicating that potential binding sites are not limited to the X chromosome (Demakova *et al*., 2003). An attractive hypothesis is that a dense distribution of strong and weak recruitment sites act cooperatively to mark the X chromosome (Dahlsveen *et al*., 2006; Demakova *et al*., 2003; Fagegaltier and Baker, 2004). Local elevation of the MSL complex by strong sites will enable weaker ones to be bound. The DHS within the *roX1* and *roX2* genes are extraordinarily strong MSL recruitment sites (Kageyama *et al*., 2001). Their ability to induce binding of the MSL complex in autosomal chromatin flanking transgene insertions likely results from enhancement of weak autosomal binding sites (Dahlsveen *et al*., 2006; Kelley *et al*., 1999). Although not absolutely essential for compensation, the situation of the *roX* DHS on the X chromosome will enhance recognition of the X.

X-chromosome binding of the MSL proteins is disrupted in *roX1*$^{-}$*roX2*$^{-}$ males, but these proteins continue to colocalize at ectopic autosomal sites. The *roX* transcripts are therefore not essential for chromatin binding, but ensure high selectivity of the intact MSL complex for the X chromosome. Assembly with *roX* might enhance the ability of the MSL complex to recognize *cis*-acting elements on the X chromosome. Alternatively, the resulting change in the complex could promote cooperative binding to closely situated sites. This would favor the X chromosome, proposed to have dense mix of strong and weak sites, over the autosomes, which have more scattered sites capable of recruiting MSL proteins (Demakova *et al*., 2003).

The distribution pattern of the MSL complex in the body of genes suggests that localization is largely established by transcriptional activity (Alekseyenko *et al*., 2006; Gilfillan *et al*., 2006; Legube *et al*., 2006). This could occur by association with the transcription machinery or interaction with nascent transcripts. Alternatively, the MSL complex could be targeted to modified histones in the wake of a transcribing polymerase. These methods would identify transcribed regions, but are unable to distinguish between X-linked and autosomal genes. Closely linked genes will be transcribed in proximity to one another, and thus may be influenced by their neighbors. Linked mammalian genes often associate at "transcription factories" (Osborne *et al*., 2004). Although there is no evidence for analogous transcription factories in *Drosophila*, identification of proteins from the nuclear pore in association with the MSL complex suggests recruitment to a transcriptionally active region. It is possible that some elements marking the X chromosome direct transcribed genes to regions where MSL loading can occur, rather than interacting directly with the MSL complex itself. This would explain

why sequences necessary for compensation of the *white* gene are found in the promoter as well as within the body of the gene (Qian and Pirrotta, 1995).

9. Concluding remarks

Dosage compensation in *Drosophila* is a remarkable model system for the study of epigenetic regulation. It is also rich in common principles of chromatin-based transcription control. Regulation of the male X chromosome involves histone modifications that are proposed to act by increasing the speed, processivity, or reinitiation rate of RNA polymerase. Similar mechanisms will be relevant for the regulation of all eukaryotic genes. Modification of the *Drosophila* X chromosome is targeted by cues including a favorable density of *cis*-acting DNA elements, transcriptional activity and possibly recruitment to regions where cotranscriptional MSL loading is promoted or transcription is facilitated. Together these produce highly selective recognition and modulation of a single chromosome. Understanding how noncoding RNAs such as *roX* coordinate this process will enhance our understanding of long-range epigenetic processes in all eukaryotes.

References

Akhtar A and Becker PB (2000) Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in *Drosophila*. *Molecular Cell*, **5**, 367–375.

Akhtar A, Zink D and Becker PB (2000) Chromodomains are protein-RNA interaction modules. *Nature*, **407**, 405–409.

Alekseyenko AA, Larschan E, Lai WR, Park PJ and Kuroda MI (2006) High resolution ChIP-chip analysis reveals that the *Drosophila* MSL complex selectively indentifies active genes on the male X chromosome. *Genes & Development*, **20**, 848–857.

Amrein H and Axel R (1997) Genes expressed in neurons of adult male *Drosophila*. *Cell*, **88**, 459–469.

Bai X, Alekseyenko AA and Kuroda MI (2004) Sequence-specific targeting of MSL complex regulates transcription of the *roX* RNA genes. *The EMBO Journal*, **23**, 2853–2861.

Bai X, Larschan E, Kwon SY, Badenhorst P and Kuroda MI (2007) Regional control of chromatin organization by noncoding *roX* RNAs and the NURF remodeling complex in *Drosophila melanogaster*. *Genetics*, **176**, 1491–1499.

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC and Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, **410**, 120–124.

Bone JR, Lavender J, Richman R, Palmer MJ, Turner BM and Kuroda MI (1994) Acetylated histone H4 on the male X chromosome is associated with dosage compensation in *Drosophila*. *Genes & Development*, **8**, 96–104.

Buscaino A, Kocher T, Kind JH, Holz H, Taipale M, Wagner K, Wilm M and Akhtar A (2003) MOF-regulated acetylation of MSL3 in the *Drosophila* dosage compensation complex. *Molecular Cell*, **11**, 1265–1277.

Buscaino A, Legube G and Akhtar A (2006) X-chromosome targeting and dosage compensation are mediated by distinct domains in MSL3. *EMBO Reports*, **7**, 531–538.

Cabal GG, Genvesio A, Rodriguez-Navarro S, Zimmer C, Gadal O, Lesne A, Buc H, Feuerbach-Fournier F, Olivio-Marin JC, Hurt EC, *et al* (2006) SAGA interacting factors confine subdiffusion of transcribed genes to the nuclear envelope. *Nature*, **441**, 770–773.

Casolari JM, Brown CR, Drubin DA, Rando OJ and Silver PA (2005) Developmentally induced changes in transcriptional program alter spatial organization across chromosomes. *Genes & Development*, **19**, 1188–1198.

Chaumeil J, Le Baccon P, Wutz A and Heard E (2006) A nevel role for *Xist* RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & Development*, **20**, 2223–2237.

Copps K, Richman R, Lyman LM, Chang KA, Rampersad-Ammons J and Kuroda MI (1998) Complex formation by the *Drosophila* MSL proteins: role of the MSL2 RING finger in protein complex assembly. *Embo Journal*, **17**, 5409–5417.

Csankovszki G, McDonel P and Meyer BJ (2004) Recruitment and spreading of the *C. elegans* dosage compensation complex along X chromosomes. *Science*, **303**, 1182–1185.

Dahlsveen IK, Gilfillan GD, Shelest VI, Lamm R and Becker PB (2006) Targeting determinants of dosage compensation in *Drosophila*. *PLoS Genetics*, **2**, e5.

Demakova OV, Kotlikova IV, Gordadze PR, Alekseyenko AA, Kuroda MI and Zhimulev IF (2003) The MSL complex levels are critical for its correct targeting to the chromosomes in *Drosophila melanogaster*. *Chromosoma*, **112**, 103–115.

Deng X and Meller VH (2006) *roX* RNAs are required for increased expression of X-linked genes in *Drosophila melanogaster* males. *Genetics*, **174**, 1859–1866.

Deng X, Rattner BP, Souter S and Meller VH (2005) The severity of *roX1* mutations are predicted by MSL localization on the X chromosome. *Mechanisms of Development*, **122**, 1094–1105.

Dieci G and Sentenac A (2003) Detours and shortcuts to transcription reinitiation. *Trends in Biochemical Sciences*, **28**, 202–209.

Eisen A, Utley RT, Nourani A, Allard S, Schmidt P, Lane WS, Lucchesi JC and Cote J (2001) The yeast NuA4 and *Drosophila* MSL complexes contain homologous subunits important for transcription regulation. *Journal of Biological Chemistry*, **276**, 3484–3491.

Fagegaltier D and Baker BS (2004) X chromosome sites autonomously recruit the dosage compensation complex in *Drosophila* males. *PLoS Biology*, **2**, e341.

Feuerbach F, Galy V, Trelles-Sticken E, Fromont-Racine M, Jacquier A, Gilson E, Olivo-Marin JC, Scherthan H and Nehrbass U (2002) Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast. *Nature Cell Biology*, **4**, 214–221.

Franke A and Baker BS (1999) The *roX1* and *roX2* RNAs are essential components of the compensasome, which mediates dosage compensation in *Drosophila*. *Molecular Cell*, **4**, 117–122.

Gilfillan GD, Konig C, Dahlsveen IK, Prakoura N, Straub T, Lamm R, Fauth T and Becker PB (2007) Cumulative contribution of weak DNA determinants to targeting the *Drosophila* dosage compensation complex. *Nucleic Acids Research*, **35**, 3561–3572.

Gilfillan GD, Straub T, de Wit E, Greil F, Lamm R, van Steensel B and Becker PB (2006) Chromosome-wide gene-specific targeting of the *Drosophila* dosage compensation complex. *Genes & Development*, **20**, 858–870.

Gu W, Wei X, Pannuti A and Lucchesi JC (2000) Targeting the chromatin-remodeling MSL complex of *Drosophila* to its sites of action on the X chromosome requires both acetyl transferase and ATPase activities. *EMBO Journal*, **19**, 5202–5211.

Gupta A, Sharma GG, Young CS, Agarwal M, Smith ER, Paull TT, Lucchesi JC, Khanna KK, Ludwig T and Pandita TK (2005) Involvement of human MOF in ATM function. *Molecular and Cellular Biology*, **25**, 5292–5305.

Henikoff S and Meneely PM (1993) Unwinding dosage compensation. *Cell*, **72**, 1–2.

Hilfiker A, Hilfiker-Kleiner D, Pannuti A and Lucchesi JC (1997) *mof*, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in *Drosophila*. *EMBO Journal*, **16**, 2054–2060.

Jeppesen P and Turner BM (1993) The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell*, **74**, 281–289.

Jin Y, Wang Y, Walker DL, Dong H, Conley C, Johansen J and Johansen KM (1999) JIL-1: a novel chromosomal tandem kinase implicated in transcriptional regulation in *Drosophila*. *Molecular Cell*, **4**, 129–135.

Kageyama Y, Mengus G, Gilfillan G, Kennedy HG, Stuckenholz C, Kelley RL, Becker PB and Kuroda MI (2001) Association and spreading of the *Drosophila* dosage compensation complex from a discrete *roX1* chromatin entry site. *EMBO Journal*, **20**, 2236–2245.

Kelley RL (2004) Path to equality strewn with *roX*. *Developmental Biology*, **269**, 18–25.

Kelley RL, Meller VH, Gordadze PR, Roman G, Davis RL and Kuroda MI (1999) Epigenetic spreading of the *Drosophila* dosage compensation complex from *roX* RNA genes into flanking chromatin. *Cell*, **98**, 513–522.

Lachner M, O'Carroll D, Rea S, Mechtler K and Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, **410**, 116–120.

Lee CG, Chang KA, Kuroda MI and Hurwitz J (1997) The NTPase/helicase activities of *Drosophila maleless*, an essential factor in dosage compensation. *EMBO Journal*, **16**, 2671–2681.

Legube G, McWeeney SK, Lercher MJ and Akhtar A (2006) X-chromosome-wide profiling of MSL-1 distribution and dosage compensation in *Drosophila*. *Genes & Development*, **20**, 871–883.

Lerach S, Zhang W, Deng H, Bao X, Girton J, Johansen J and Johansen KM (2005) JIL-1 kinase, a member of the Male-specific lethal (MSL) complex, is necessary for proper dosage compensation of eye pigmentation in *Drosophila*. *Genesis*, **43**, 213–215.

Li F, Parry DA and Scott MJ (2005) The amino-terminal region of *Drosphila* MSL1 contains basic, glycine-rich and leucine zipper-like motifs that promote X chromosome binding, self-association and MSL2 binding, respectively. *Molecular and Cellular Biology*, **25**, 8913–8924.

Lucchesi JC, Kelly WG and Panning B (2005) Chromatin remodeling in dosage compensation. *Annual Review of Genetics*, **39**, 615–651.

Lyman LM, Copps K, Rastelli L, Kelley RL and Kuroda MI (1997) *Drosophila male-specific lethal-2* protein: structure/function analysis and dependence on MSL-1 for chromosome association. *Genetics*, **147**, 1743–1753.

Mahadevan LC, Clayton AL, Hazzalin CA and Thomson S (2004) Phosphorylation and acetylation of histone H3 at inducable genes: two controversies revisited. *Novartis Foundation Symposium*, **259**, 102–111.

Marin I and Baker BS (2000) Origin and evolution of the regulatory gene male-specific lethal-3. *Molecular Biology and Evolution*, **17**, 1240–1250.

McDonel P, Jans J, Peterson BK and Meyer BJ (2006) Clustered DNA motifs mark X chromosomes for repression by a dosage compensation complex. *Nature*, **444**, 614–618.

Meller VH (2003) Initiation of dosage compensation in *Drosophila* embryos depends on expression of the *roX* RNAs. *Mechanisms of Development*, **120**, 759–767.

Meller VH, Gordadze PR, Park Y, Chu X, Stuckenholz C, Kelley RL and Kuroda MI (2000) Ordered assembly of *roX* RNAs into MSL complexes on the dosage- compensated X chromosome in *Drosophila*. *Current Biology*, **10**, 136–143.

Meller VH and Rattner BP (2002) The *roX* genes encode redundant *male-specific lethal* transcripts required for targeting of the MSL complex. *EMBO Journal*, **21**, 1084–1091.

Meller VH, Wu KH, Roman G, Kuroda MI and Davis RL (1997) *roX1* RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system. *Cell*, **88**, 445–457.

Mendjan S and Akhtar A (2007) The right dose for every sex. *Chromosoma*, **116**, 95–106.

Mendjan S, Taipale M, Kind J, Holz H, Gebhardt P, Schelder M, Vermeulen M, Buscaino A, Duncan K, Mueller J, *et al* (2006) Nuclear pore components are involved in the transcriptional regulation of dosage compensation in *Drosophila*. *Molecular Cell*, **21**, 811–823.

Morales V, Straub T, Neumann MF, Mengus G, Akhtar A and Becker PB (2004) Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *The EMBO Journal*, **23**, 2258–2268.

Nielsen PR, Nietlispach D, Buscaino A, Warner RJ, Akhtar A, Murzin AG, Murzina NV and Laue ED (2005) Structure of the chromo barrel domain from the MOF acetyltransferase. *Journal of Biological Chemistry*, **37**, 32326–32331.

Oh H, Bone JR and Kuroda MI (2004) Multiple classes of MSL binding sites target dosage compensation to the X chromosome of *Drosophila*. *Current Biology*, **14**, 481–487.

Oh H, Park Y and Kuroda MI (2003) Local spreading of MSL complexes from *roX* genes on the *Drosophila* X chromosome. *Genes & Development*, **17**, 1334–1339.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, et al (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, **36**, 1065–1071.

Park Y, Kelley RL, Oh H, Kuroda MI and Meller VH (2002) Extent of chromatin spreading determined by *roX* RNA recruitment of MSL proteins. *Science*, **298**, 1620–1623.

Park Y, Mengus G, Bai X, Kageyama Y, Meller VH, Becker PB and Kuroda MI (2003) Sequence-specific targeting of *Drosophila roX* genes by the MSL dosage compensation complex. *Molecular Cell*, **11**, 977–986.

Park Y, Oh H, Meller VH and Kuroda MI (2005) Variable splicing of non-coding *roX2* RNAs influences targeting of MSL dosage compensation complexes in *Drosophila*. *RNA Biology*, **2**, 157–164.

Plath K, Mlynarczyk-Evans S, Nusinow DA and Panning B (2002) *Xist* RNA and the mechanism of X chromosome inactivation. *Annual Review of Genetics*, **36**, 233–278.

Plath K, Talbot D, Hameer KM, Otte AP, Yang TP, Jaenisch R and Panning B (2004) Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *Journal of Cell Biology*, **167**, 1025–1035.

Qian S and Pirrotta V (1995) Dosage compensation of the *Drosophila* white gene requires both the X chromosome environment and multiple intragenic elements. *Genetics*, **139**, 733–744.

Rattner BP and Meller VH (2004) *Drosophila* Male Specific Lethal 2 protein controls male-specific expression of the *roX* genes. *Genetics*, **166**, 1825–1832.

Richter L, Bone JR and Kuroda MI (1996) RNA-dependent association of the *Drosophila maleless* protein with the male X chromosome. *Genes to Cells*, **1**, 325–336.

Sanjuan R and Marin I (2001) Tracing the origin of the compensasome. *Molecular biology and evolution*, **18**, 330–343.

Scott MJ, Pan LL, Cleland SB, Knox AL and Heinrich J (2000) MSL1 plays a central role in assembly of the MSL complex, essential for dosage compensation in *Drosophila*. *EMBO Journal*, **19**, 144–155.

Shogren-Knaak M, Ishii H, Sun J-M, Pazin MJ, Davie JR and Peterson CL (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, **311**, 844–847.

Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, Peters AH, Jenuwein T, Otte AP and Brockdorff N (2003) Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Developmental Cell*, **4**, 481–495.

Smith ER, Allis CD and Lucchesi JC (2001) Linking global histone acetylation to the transcription enhancement of X-chromosomal genes in *Drosophila* males. *Journal of Biological Chemistry*, **276**, 31483–31486.

Smith ER, Pannuti A, Gu W, Steurnagel A, Cook RG, Allis CD and Lucchesi JC (2000) The *Drosophila* MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. *Molecular and Cellular Biology*, **20**, 312–318.

Smith RR, Cayrou C, Huang R, Lane WS, Cote J and Lucchesi JC (2005) A human protein complex homologous to the *Drosophila* MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Molecular and Cellular Biology*, **25**, 9175–9188.

Stuckenholz C, Meller VH and Kuroda MI (2003) Functional redundancy within *roX1*, a noncoding RNA involved in dosage compensation in *Drosophila melanogaster*. *Genetics*, **164**, 1003–1014.

Taddei A, Van Houwe G, Hediger F, Kalck V, Cubizolles F, Schober H and Gasser SM (2006) Nuclear pore association confers optimal expression levels for an inducible yeast gene. *Nature*, **441**, 774–778.

Taipale M, Rea S, Richter K, Vilar A, Lichter P, Imhof A and Akhtar A (2005) hMOF histone acetyltransferase is required for histone H4 lysine 16 acetylation in mammalian cells. *Molecular and Cellular Biology*, **25**, 6798–6810.

Turner BM, Birley AJ and Lavender J (1992) Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. *Cell*, **69**, 375–384.

Wang Y, Zhang W, Jin Y, Johansen J and Johansen KM (2001) The JIL-1 tandem kinase mediates histone H3 phosphorylation and is required for maintenance of chromatin structure in *Drosophila*. *Cell*, **105**, 433–443.

Wutz A (2003) RNAs templating chromatin structure for dosage compensation in animals. *Bioessays*, **25**, 434–442.

Wutz A, Rasmussen TP and Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics*, **30**, 167–174.

Zhang L-F, Huynh KD and Lee JT (2007) Perinucleolar targeting of the inactive X during S phase. *Cell*, **129**, 693–706.

Zhang W, Deng H, Bao X, Lerach S, Girton J, Johansen J and Johansen K (2005) The JIL-1 histone H3S10 kinase regulates dimethyl H3K9 modifications and heterochromatic spreading in *Drosophila*. *Development*, **133**, 229–235.

Short Specialist Review

DNA methylation in epigenetics, development, and imprinting

Hiroyuki Sasaki
National Institute of Genetics, Mishima, Japan

1. DNA methylation in epigenetics

In the vertebrate genome, over 60% of CpG dinucleotides are methylated at the 5 position of the cytosine residue. The product of this methylation is 5-methylcytosine (5mC), which is the only physiologically modified base seen in the vertebrate genome. 5mC provides additional information to DNA sequences and serves as an excellent epigenetic mark. The idea that 5mC might influence gene expression was first postulated by Holiday and Pugh (1975) and Riggs (1975). As will be discussed below, vertebrates, including mammals, indeed use methylation for developmental and tissue-specific control of gene expression.

The eukaryotic DNA methylation system has its roots in the restriction/modification system of prokaryotes (a defense mechanism against bacteriophages). The catalytic domain of mammalian DNA methyltransferases (DNMTs) shares conserved motifs with prokaryotic cytosine modification enzymes (Bestor *et al*., 1988; Okano *et al*., 1998). Although methylation is present almost universally in eukaryotes, there are exceptions such as yeast (*Saccharomyces cerevisiae*) and nematode (*Caenorhabditis elegans*), which live happily without it. Because these organisms have relatively small genomes, it is proposed that the primary function of methylation is to differentiate active and inactive regions in higher eukaryotes with complex genomes. Silencing of transposons may be another but related primary function of methylation.

A key feature of methylation in vertebrates is that it occurs at CpG dinucleotides. This means that DNA can be methylated symmetrically on both strands (Figure 1). Unmodified CpGs are methylated by de novo methyltransferases DNMT3A and/or DNMT3B (Okano *et al*., 1998), which transfer a methyl group from the methyl donor *S*-adenosylmethionine to the target cytosine. Upon DNA replication, fully methylated CpGs become hemimethylated, a configuration that is the favored targets for maintenance methyltransferase DNMT1 (Bestor *et al*., 1988). DNMT1 faithfully restores the methylation patterns (Figure 1). If DNMT1 is not present, the sites will eventually lose 5mC after cycles of replication (passive demethylation). Although the presence of an active demethylation system has been suggested, its biochemical nature remains elusive.

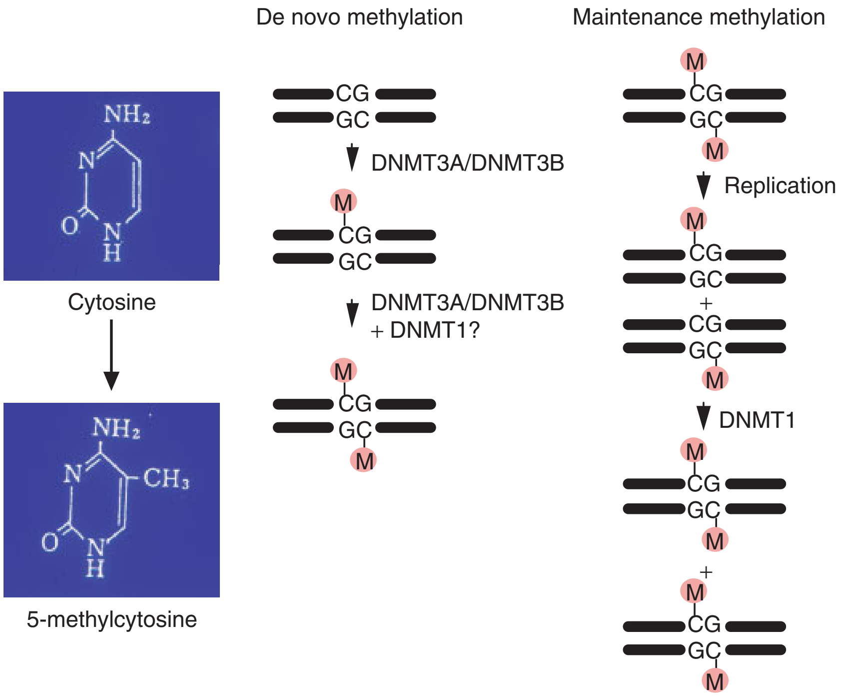


Figure 1 Creation and maintenance of DNA methylation patterns

Because spontaneous deamination of 5mC causes a C to T transition, CpG dinucleotide is underrepresented in the vertebrate genome. However, gene promoters and regulatory regions are often more rich in CpG. These CpGs are the targets for methylation in gene regulation. In particular, approximately half of the genes are associated with small regions (typically 1–2 kb) with a high density of CpG (CpG islands) (Bird, 1986). The CpG islands are normally methylation-free but can be methylated in genomic imprinting (see later) and X-chromosome inactivation (*see* Article 15, **Human X chromosome inactivation**, Volume 1, Article 40, **Spreading of X-chromosome inactivation**, Volume 1, and Article 41, **Initiation of X-chromosome inactivation**, Volume 1).

How does DNA methylation silence genes? Some transcription factors are known to be sensitive to methylation: if 5mC is present in the target sequence, it prevents the factor from binding. However, the number of such factors seems small. Rather, vertebrates evolved a family of methyl-CpG-binding proteins, which share a common methyl-CpG-binding domain (MBD proteins) (Hendrich and Bird, 1998). Binding of MBD proteins to methylated DNA may physically interfere with transcription factor binding. Furthermore, some MBD proteins such as MeCP2 and MBD2 form a multisubunit complex containing histone deacetylases (HDACs) to repress transcription (Nan *et al*., 1998; Jones *et al*., 1998). The complex formed with MBD2 also contains an ATP-dependent chromatin remodeling protein (Wade *et al*., 1999; Zhang *et al*., 1999). Thus, there is interplay between DNA methylation

and chromatin modification/remodeling. Both DNMT1 and DNMT3B also interact with HDACs and can repress transcription (Burgers *et al*., 2002).

2. DNA methylation in development

That DNA methylation plays an essential role in cell differentiation is best illustrated in experiments done with a nucleotide analogue 5-azacytidine (Taylor and Jones, 1979). When a fibroblast cell line was treated with this demethylating agent, various cell types, including fat cells and muscle cells, appeared in the culture dish. This suggested a massive reactivation of genes that are normally silent in fibroblast cells. Consistent with this, it is well established that different cell types have their respective methylation patterns.

How and when are such tissue-specific methylation patterns established? The global methylation level changes dynamically during mammalian development (Figure 2). The sperm- and oocyte-derived genomes first experience massive demethylation in cleavage stage embryos (*see* Article 33, **Epigenetic reprogramming in germ cells and preimplantation embryos**, Volume 1). (However, the paternal and maternal imprints are maintained even in this stage. See later.) This is mainly due to exclusion of DNMT1 from the nucleus, but there is evidence that the sperm-derived genome is actively demethylated soon after fertilization (*see* Article 33, **Epigenetic reprogramming in germ cells and preimplantation**

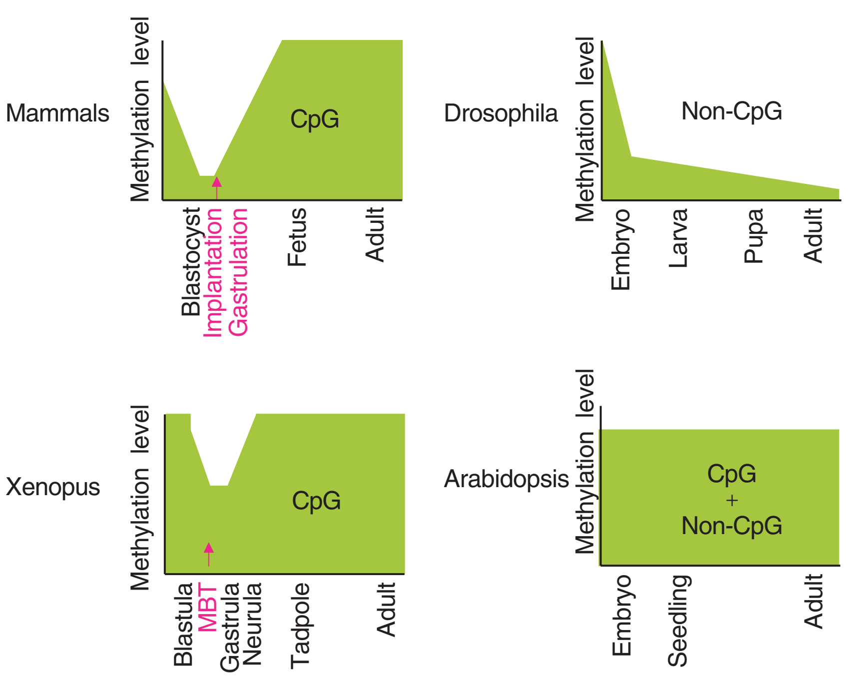


Figure 2 Global changes in DNA methylation during development

embryos, Volume 1). After implantation, DNMT3A and DNMT3B establish methylation patterns specific to individual cell lineages (Okano *et al*., 1999). DNMT1 then maintains the tissue-specific methylation patterns (Li *et al*., 1992). However, extraembryonic tissues and germ cells maintain relatively low methylation levels (*see* Article 33, **Epigenetic reprogramming in germ cells and preimplantation embryos**, Volume 1).

Differentiation-dependent changes in methylation are recapitulated in embryonic stem (ES) cells, which are derived from the inner cell mass of blastocysts. Undifferentiated ES cells have a methylation level lower than that of the somatic tissues but, when they are induced to differentiate, de novo methylation by DNMT3A and DNMT3B occurs. This provides a model to study epigenetic control of developmental gene expression.

Developmental changes in methylation are less well characterized in other organisms, but frogs (*Xenopus laevis*) show demethylation in early embryos toward the mid-blastula transition, just like preimplantation mammalian embryos. In fruit fly (*Drosophila melanogaster*), methylation is seen only during early development. All these dynamic changes suggest that methylation probably has a role in the development of these animals. However, in plants such as *Arabidopsis thaliana*, no global methylation change is seen.

Direct evidence that DNA methylation is crucial for normal development comes from gene knockout studies in mice. A targeted disruption of DNMT1, which removes almost all genomic methylation due to lack of maintenance activity, causes early embryonic death (Li *et al*., 1992). Similarly, mice deficient for DNMT3A die at an early postnatal stage and those deficient for DNMT3B die at a late embryonic stage, respectively (Okano *et al*., 1999). DNMT3A/DNMT3B double mutants die even earlier, similar to the stage at which the DNMT1 mutants die. Interestingly, DNMT3B mutations, but not DNMT3A mutations, cause loss of methylation of pericentromeric minor satellite DNA. Consistent with this observation, DNMT3B mutations cause ICF (immunodeficiency, centromeric instability, facial anomalies) syndrome in humans (Okano *et al*., 1999; Xu *et al*., 1999).

3. DNA methylation in imprinting

Normal mammalian development requires both a paternal and a maternal genome, and thus parthenogenesis is not possible. The functional differences between the paternal and maternal genomes are due to the differential expression of the paternal and maternal alleles of a small subset (up to a few hundred) of genes (*see* Article 45, **Bioinformatics and the identification of imprinted genes in mammals**, Volume 1 and Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1). For example, *Igf2* is only expressed from the paternal allele, and *H19* is only expressed from the maternal allele in both humans and mice. The differences between the parental alleles originate from the differential epigenetic modification of the genome in the male and female gametes, a phenomenon called *genomic imprinting*. Imprinting is crucial for normal mammalian development and is relevant to congenital malformation syndromes and cancers in humans (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1,

Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1, Article 30, **Beckwith–Wiedemann syndrome**, Volume 1, Article 31, **Imprinting at the *GNAS* locus and endocrine disease**, Volume 1, and Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1).

Imprinted genes tend to form clusters in the genome and are often associated with regions methylated differently on the paternal and maternal alleles. Such regions are called differentially methylated regions (DMRs) and many of them are CpG islands. There are at least two classes of DMRs, one methylated during gametogenesis and the other methylated after fertilization. Germ-line deletion experiments in mice have shown that some DMRs belonging to the former class control multiple imprinted genes of the cluster (imprint control region or imprinting center). The gamete-derived methylation patterns of imprinted genes are maintained in the somatic tissues throughout embryonic development, but are erased in primordial germ cells, consistent with the reversibility of imprinting (Figure 3). These findings suggest that DNA methylation is the epigenetic mark (imprint) that differentiates the paternal and maternal alleles.

A recent study demonstrated that de novo methylation is required for the establishment of imprints in the gametes. A disruption of DNMT3A, but not DNMT3B, in the female germ cells causes loss of methylation at maternally methylated DMRs and loss of monoallelic expression of their associated genes (Kaneda *et al*., 2004a). Also, disruption of DNMT3A in the male germ cells causes loss of methylation at paternally methylated DMRs (Kaneda *et al*., 2004a). Together with the data from other knockout studies, it is thought that DNMT3A cooperates with DNMT3L, a DNMT3-like protein with no methylation activity, to establish the gametic methylation patterns (Figure 3) (Bourc'his *et al*., 2001; Hata *et al*., 2002).

The gamete-derived methylation patterns of DMRs are maintained in the zygote by maintenance methyltransferase DNMT1 (Figure 3). Thus, a deletion of DNMT1

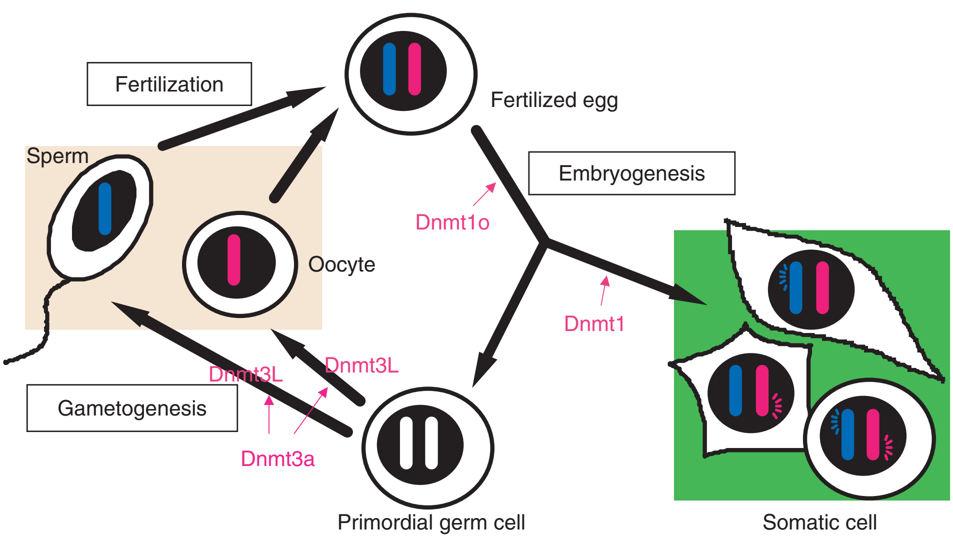


Figure 3 DNA methylation and the cycle of genomic imprinting

causes loss of methylation at both paternally and maternally methylated DMRs and loss of monoallelic expression of all imprinted genes (Li *et al*., 1993; Howell *et al*., 2001). Interestingly, some imprinted genes such as *H19* show reactivation of the normally silent allele but other imprinted genes such as *Igf2* and *Igf2r* show inactivation of the normally active allele (Li *et al*., 1993). This is consistent with the methylation-sensitive insulator model for *Igf2* imprinting and with the antisense regulation model for *Igf2r* imprinting.

Although DNA methylation is the primary mechanism of genomic imprinting, other epigenetic mechanisms are also involved. For example, DMRs show allelic differences in histone modifications (acetylation and methylation) and in accessibility by nucleases. Furthermore, imprinting of a placenta-specific gene *Mash2* has been shown to be tolerant to loss of maintenance methylation. However, even in such cases, the gamete-derived primary imprint is probably DNA methylation (Kaneda *et al*., 2004b).

Genomic imprinting is not unique to mammals, and it also occurs in plants. The *FWA* gene of *Arabidopsis* is only expressed from the maternal allele in the endosperm (a nutritive tissue that can be considered the placental equivalent in plants) (Kinoshita *et al*., 2004). Unlike the mammalian imprinted genes, *FWA* has no gamete-derived imprint. It is methylated in both male and female gametes. In plants, two identical male gametes are delivered by the pollen to two distinct female gametes. A double fertilization process leads to the development of the embryo surrounded by the nurturing endosperm. DNA methylation of *FWA* is maintained throughout embryonic development and plant vegetative phase, but in endosperm, the maternal *FWA* allele become demethylated and expressed (Kinoshita *et al*., 2004). Thus, imprinting of *FWA* does not involve de novo methylation and there is no imprint resetting in germ cells.

Further reading

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews. Genetics*, **3**, 662–672.

References

Bestor T, Laudano A, Mattaliano R and Ingram V (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of Molecular Biology*, **203**, 971–983.

Bird A (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.

Bourc'his D, Xu GL, Lin CS, Bollman B and Bestor TH (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science*, **294**, 2536–2539.

Burgers WA, Fuks F and Kouzarides T (2002) DNA methyltransferases get connected to chromatin. *Trends in Genetics*, **18**, 275–277.

Hata K, Okano M, Lei H and Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development*, **129**, 1983–1993.

Hendrich B and Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and Cellular Biology*, **18**, 6538–6547.

Holiday R and Pugh JE (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM and Chaillet JR (2001) Genomic imprinting disrupted by a maternal effect mutation in the *Dnmt1* gene. *Cell*, **104**, 829–838.

Jones PL, Veenstra GJC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J and Wolffe AP (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, **19**, 187–190.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E and Sasaki H (2004a) Essential role for *de novo* DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, **429**, 900–903.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E and Sasaki H (2004b) Role of *de novo* DNA methyltransferases in initiation of genomic imprinting and X-chromosome inactivation. *Cold Spring Harbor Symposia on Quantitative Biology*, in press.

Kinoshita T, Miura A, Choi Y, Kinoshita Y, Cao X, Jacobsen SE, Fischer RL and Kakutani T (2004) One-way control of *FWA* imprinting in Arabidopsis endosperm by DNA methylation. *Science*, **303**, 521–523.

Li E, Beard C and Jeanisch R (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Li E, Bestor T and Jeanisch R (1992) Targeted Mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

Nan X, Ng H-H, Jojnson CA, Laherty CD, Turner BM, Eisenman RN and Bird A (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**, 386–389.

Okano M, Bell DW, Haber DA and Li E (1999) DNA methyltransferases *Dnmt3a* and *Dnmt3b* are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.

Okano M, Xie S and Li E (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics*, **19**, 219–220.

Riggs AD (1975) X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics*, **14**, 9–25.

Taylor SM and Jones PA (1979) Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine. *Cell*, **17**, 771–779.

Wade PA, Gegonne A, Jones PL, Ballestar E, Aubry F and Wolffe AP (1999) Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature Genetics*, **23**, 62–66.

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ and Viegas-Pequignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, **402**, 187–191.

Zhang Y, Ng H-H, Erdjument-Bromage H, Tempst P, Bird A and Reinberg D (1999) Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development*, **13**, 1924–1935.

Short Specialist Review

Epigenetic reprogramming in germ cells and preimplantation embryos

Abraham L. Kierszenbaum
The Sophie Davis School of Biomedical Education, and The City University of New York Medical School, New York, NY, USA

1. Introduction

Gametogenesis and early embryogenesis in mammals are under the control of genetic and epigenetic mechanisms. A remarkable aspect of epigenetics is the reprogramming of allele-specific gene expression by DNA methylation and histone modifications (acetylation, phosphorylation, methylation, and ubiquitylation). A disruption of these two biochemical events leads to abnormal developmental processes, including Prader–Willi (PWS) and Angelman (AS) syndromes, and the Beckwith–Wiedemann syndrome (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1 and Article 30, **Beckwith–Wiedemann syndrome**, Volume 1) (see Nicholls and Knepper, 2001 for a comprehensive review).

Two clinically related issues are relevant to the epigenetic reprogramming and genomic imprinting in germ cells and preimplantation embryos. First, advances in assisted reproductive technology as an approach to treating infertility have attracted attention to the potential risk of birth defects when major epigenetic events can be disrupted when round spermatids and preimplantation embryo are developed or maintained in culture, respectively (Lucifero *et al*., 2002; Gosden *et al*., 2003). Second, prospects of epigenetic therapy based on inhibitors of enzymes controlling epigenetic modifications, in particular, DNA methyltransferases and histone deacetylases, have opened the possibility that genes that have undergone abnormal epigenetic silencing may be reactivated (Egger *et al*., 2004). This brief review summarizes current knowledge on the developmental occurrence of genomic imprinting during gametogenesis and in the preimplantation embryo. Knowledge of these highly timed events can contribute to implementing safe and efficient assisted reproductive technologies.

2. Components of the epigenetic reprogramming machinery

DNA methylation and histone modifications are epigenetic heritable changes functioning as efficient modulators of transcription. ATP-dependent chromatin

modifications contribute to DNA methylation and histone modification events (see Reik *et al*., 2001; Li, 2002 for comprehensive reviews). DNA methylation occurs at CpG dinucleotides and is catalyzed by DNA methyltransferase 1 (Dnmt1), a maintenance enzyme operating after DNA replication, and Dnmt3a and Dnmt3b, both required for de novo DNA methylation patterns during development. Both Dnmt1 and Dnmt3a interact with histone deacetylases (HDACs) to repress transcription. CpG-binding proteins, with a methyl-CpG-binding domain (MBD), recruit different chromatin-remodeling proteins and transcription regulatory complexes to DNA-methylated regions in the genome. Histone modifications occur at the lysine, arginine, and serine residues located at the amino-terminal tail of histones. Histone methyltransferases include H3-K4 methyltransferase (methylation of lysine 4 of Histone 3; associated with active gene expression), H3-K9 (methylation of lysine 9 of Histone 3; associated with transcriptional silencing), and five H3-K9 methyltransferases (Suv39 h1 and Suv39 h2, G9a, ESET/SetDB1 and Eu-MHTase1). Several HDACs have been identified, including transcription coactivators with intrinsic histone acetyltransferase activity. ATP-dependent chromatin-remodeling protein complexes (SWI/SNF/Brm, ISWI, and Mi-2/NuRD) use ATP hydrolysis to make nucleosomal DNA and the histone core accessible to DNA methylation and histone modifications.

3. Epigenetic reprogramming in germ cells

The most differentially methylated regions in imprinted genes of primordial germ cells (PGC) located in the genital ridge become demethylated or "erased" by embryonic day 13 to 14 in both females and males. Prior to this (embryonic days 11.5 or 12.5), PGC are highly methylated and *H19* (a paternally imprinted gene) and *Igf2r* (insulin-like growth factor 2 receptor gene, a regulator of fetal growth and embryonic development) display normal imprinting patterns. Both genes are more methylated in cells with an XY chromosome complement than those with an XX chromosome complement (Durcova-Hills *et al*., 2004). Following demethylation, male PGC in the testicular cords enter mitotic arrest and primary oocytes, surrounded by follicular cells in the fetal ovary, become arrested at the end of meiotic prophase. During spermatogenesis, epigenetic changes of the spermatocyte lineage and the derived postmeiotic haploid spermatids have enabled the use of nuclei of in vitro developed spermatids from spermatocytes precursors to generate normal offspring when injected into mature oocytes (Marh *et al*., 2003).

Genomic imprinting of the sperm and egg genomes is regulated by differential methylation, an activity dependent on DNA methyltransferases. Recent studies of the Dnmt3-Like (*Dnmt3L*) gene have shown that the encoded protein shares homology with Dnmt3a and Dnmt3b in the PHD-zinc-finger domain, but lacks both the highly conserved methyltransferase motifs and enzymatic activity. *Dnmt3L*-deficient females generate mature and functional oocytes, but derived embryos have neural tube and placental abnormalities and are nonviable by mid-gestation. Analysis of DNA methylation patterns of *Dnmt3L*-deficient oocytes shows that genes on different chromosomes (*Igf2r, Mest* (mesoderm-specific transcript) and *Peg3* (paternally expressed gene 3) and several genes in the *Snrpn* (a maternally

imprinted gene) locus) fail to display epigenetic maternalization and the monoallelic expression of all maternally imprinted genes is thought to be lost in the offspring.

The Dnmt3L protein interacts with the Dnmt3 family of DNA methyltransferases and might cooperate with Dnmt3a or Dnmt3b to regulate gamete-specific methylation of imprinted genes in oocytes. In fact, Dnmt3a/Dnmt3b-deficient oocytes also fail to establish epigenetic maternalization. *Dnmt1*-deficient oocytes can establish methylation imprints but cannot maintain imprinting in preimplantation embryos. Therefore, it appears that Dnmt1 is essentially a housekeeping methyltransferase required for maintaining tissue-specific methylation patterns. High levels of the ovarian Dnmt1o accumulate in the oocyte nuclei during the follicular growth phase, when genomic imprinting has been established. Like Dnmt1, Dnmt1o is a housekeeping DNA methyltransferase.

In the male, *Dnmt3L* gene deficiency results in spermatogenesis arrest and sterility due to spermatogonia entering meiosis and being killed by an asynapsis checkpoint just prior to pachytene (Bourc'his and Bestor, 2004). In the female, in contrast, mature and functional oocytes are produced in the *Dnmt3L* deficient mutant. *Dnmt3a* knockout male mice display testes with abnormal meiotic prophase spermatocytes and few mature sperm. In addition to DNA methylation, histone modifications are critical for spermatogenesis. Histone methyltransferase *Suv39h1* and *Suv39h2*-deficient mice are sterile because of meiotic arrest at the pachytene stage of meiotic prophase. These examples emphasize the impact of epigenetic modifications in male and female fertility.

DNA methylation and histone H3-K9 methylation during spermatogenesis correlate with histone deacetylation. Somatic histones are hyperacetylated in spermatogonia and in pre-leptotene spermatocytes but acetylated histones are not detected from leptotene on and in round spermatids. It appears that DNA methylation and histone modifications play a role in modulating meiotic chromosome architecture and in ribosomal and nonribosomal transcription activity. In summary, the timing and extent of remethylation following methylation erasure in PGC is slightly different during oogenesis and spermatogenesis (Figure 1). In the male, remethylation begins in prospermatogonia (gonocytes) by embryonic day 15 to 16 and continues throughout spermatogenesis. Therefore, remethylation takes place before mitotic amplification of the spermatogonial stem cell lineage at the time of puberty. Two representative genes, *H19* and *Mest*, display developmental epigenetic paternalization: both genes are unmethylated in fetal prospermatogonia, *Mest* remains unmethylated throughout spermatogenesis, and *H19* methylation first appears in spermatogonia and is maintained throughout spermatogenesis (Kerjean *et al*., 2000). In the female, remethylation is observed during the growth of oocytes, a prolonged developmental process enabling sequence methylation at different time points. Although germ cell–specific DNA methyltransferases have not been identified, Dnmt3a and Dnmt3b are good candidates (Kaneda *et al*., 2004).

4. Epigenetic reprogramming in preimplantation embryos

After fertilization, the chromatin of the paternal genome undergoes changes consisting in the replacement of protamines by acetylated histones and, from some

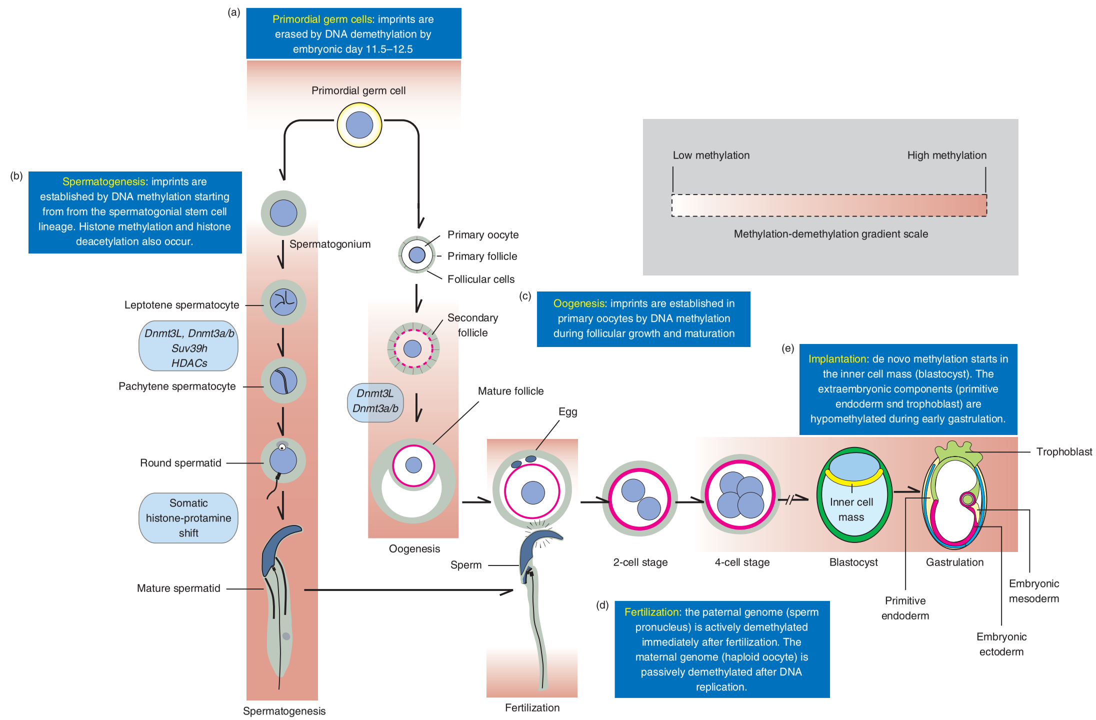


Figure 1 Epigenetic reprogramming during gametogenesis and early embryo implantation (mouse). (a) Demethylation at imprinted loci erase parental imprinting marks in primordial germ cells by embryonic day 11.5–12.5. (b) During spermatogenesis, methyltransferases Dnmt3a and Dnmt3b, in association with Dnmt3L, start to reestablish paternal methylation from spermatogonia on. In addition, histone hypoacetylation-deacetylation – controlled by histone deacetylases (HDACs) – and histone methyltransferases, including Suv39 h, regulate chromatin organization and transcription activities. During spermiogenesis, somatic histones are gradually replaced by transient basic proteins and finally by protamines. Consequently, the nucleosomal beaded chromatin pattern is replaced by smooth chromatin fibers and the genome becomes transcriptionally silent. (c) During oogenesis, maternal-specific genomic imprints are reestablished in the DNA of the oocyte starting during follicular growth and continuing throughout follicle maturation by the de novo activity of methyltransferases Dnmt3a, Dnmt3b, and Dnmt3 L. (d) Immediately after fertilization, the paternal genome (sperm pronucleus) in the zygote is demethylated by an active mechanism. Demethylation of the maternal genome (egg pronucleus) takes place by way of a passive mechanism after DNA replication has occurred. Chromatin decondenses and transcription activities of the zygote, essential for early development, take place. Most of the methylation marks inherited from the male and female gametes are erased by the blastocyst embryonic stage (embryonic day 3.5). (e) During early implantation, the embryonic DNA methylation patterns are established in a lineage-specific manner by de novo methylation starting in the inner cell mass of the blastocyst. DNA methylation levels increase in the primitive ectoderm. The DNA of extraembryonic cells (primitive endoderm and trophoblast) remains hypomethylated. Specific parental imprinted genes protected from demethylation are not indicated for clarity. Diagram not to scale. Data compiled from Reik *et al*. (2001), Kierszenbaum (2002), Lucifero *et al*. (2002), and Li (2002)

evidence, DNA demethylation by an active mechanism that is completed before DNA replication. Some sequences in the paternal chromosomes are protected from demethylation, in particular, the imprinted genes *H19* and *RasGrf1* . The maternal genome is demethylated by a passive mechanism dependent on DNA replication. Dnmt1 is not present in the nucleus and, therefore, passive demethylation occurs. In the eight-cell embryo (mouse), Dnmt1 reappears in the nucleus. At the time of implantation, both the paternal and maternal genomes are remethylated with the participation of Dnmt3a and Dnmt3. Variations in the methylation of imprinted genes in embryonic and extraembryonic cell lineages are characteristic. The postzygotic demethylation and remethylation sequence (mimicking to some extent the reprogramming saga of the germ cell lineage) presumably removes epigenetic modifications acquired during gametogenesis. An important issue is the consequence of using somatic nuclei from adult and embryonic donors during animal cloning. Somatic donor nuclei contain a highly methylated genome, a departure of the precise postzygotic demethylation-remethylation timely sequence. Although somatic nuclei are reprogrammed in clones, the timing and efficiency of demethylation-remethylation of genes critical for cellular differentiation may differ, thus leading to developmental abnormalities and lethality.

References

Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, **431**, 96–99.

Durcova-Hills G, Burgoyne P and McLaren A (2004) Analysis of sex differences in EGC imprinting. *Developmental Biology*, **268**, 105–110.

Egger G, Liang G, Aparicio A and Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, **429**, 457–463.

Gosden R, Trasler J, Lucidero D and Faddy M (2003) Rare congenital disorders, imprinted genes, and assisted reproductive technology. *Lancet*, **361**, 1975–1977.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E and Sasaki H (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, **429**, 900–903.

Kerjean A, Dupont J-M, Vasseur C, Le Tessier D, Cuisset L, Paldi A, Jouannet P and Jeanpierre M (2000) Establishment of the paternal methylation imprint of human *H19* and *MEST/PEG1* genes during spermatogenesis. *Human Molecular Genetics*, **9**, 2183–2187.

Kierszenbaum AL (2002) Genomic imprinting and epigenetic reprogramming: unearthing the garden of forking paths. *Molecular Reproduction and Development*, **63**, 269–272.

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, **3**, 662–673.

Lucifero D, Mertineit C, Clarke HJ, Bestor TH and Trasler JM (2002) Methylation dynamics of imprinted genes in mouse germ cells. *Genomics*, **79**, 530–538.

Marh J, Tres LL, Yamazaki Y, Yanagimachi R and Kierszenbaum AL (2003) Mouse round spermatids developed in vitro from preexisting spermatocytes can produce normal offspring by nuclear injection into in vivo-developed mature oocytes. *Biology of Reproduction*, **69**, 169–176.

Nicholls RD and Knepper JL (2001) Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes. *Annual Review of Genomics and Human Genetics*, **2**, 153–175.

Reik W, Dean W and Walter J (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.

Short Specialist Review

Epigenetics and imprint resetting in cloned animals

Sigrid Eckardt, Satoshi Kurosaka and K. John McLaughlin
University of Pennsylvania, Kennett Square, PA, USA

1. Introduction

Since virtually all cells of an individual animal contain the same genetic information, the diversity in differential gene activity required for the development and formation of specialized cells must be ensured by mechanisms that do not involve changes in genomic sequence while efficiently activating or silencing certain genes. These "mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Russo *et al*., 1996) are referred to as epigenetic, and include the modification of the genome by DNA methylation, the modification of histones, and the adoption of specific chromatin structures (Jaenisch and Bird, 2003; Reik *et al*., 2003). Although epigenetic modifications normally establish a stable cell identity, this cellular memory can be erased or altered. The development of clones from differentiated nuclei establishes that the ooplasm can create a substitute for the zygotic epigenotype from a somatic cell nucleus. Normally, the genome is "reset" during germline development, such that the gametic genomes are prepared to execute a developmental program once the fertilization and remodeling into a zygotic genome has occurred (Renard, 1998).

Germline reprogramming also involves the erasure and establishment of genomic imprints that regulate parent-of-origin-dependent gene expression (Mann, 2001). Apparently, exposure of a somatic nucleus to the ooplasm is sufficient to remodel its epigenetically distinct phenotype to one that is similar enough to that of the zygote, permitting development while bypassing reprogramming events that occur in the germline (Figure 1). Poor development, survival, frequent defects, and gene expression errors prevalent in mammalian clones (Ogura *et al*., 2002; Rideout *et al*., 2001; Wakayama and Yanagimachi, 1999b; Wells, 2003), and possibly the diminishing efficiency when cloning from clones for multiple generations (Wakayama *et al*., 2000), suggest that epigenetic reprogramming of somatic cell nuclei in clones is usually flawed or incomplete.

2. Genomic imprinting

The failure of clones, and the abnormalities observed in those that develop, are presumably due to incomplete reprogramming that affects gene expression. Clones

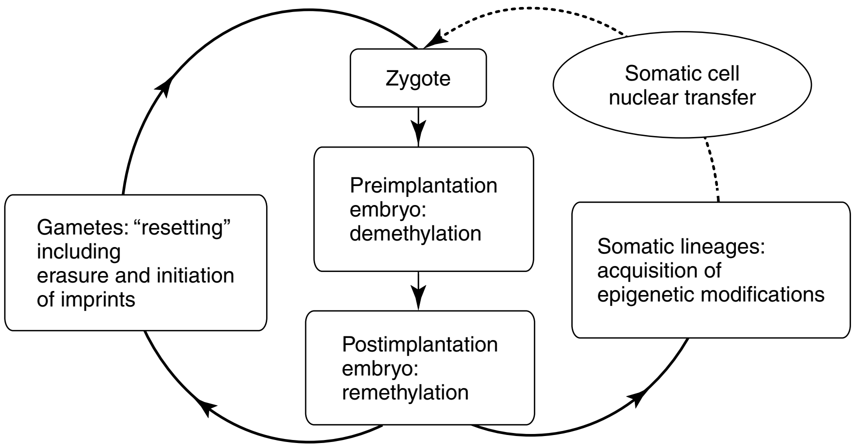


Figure 1 Somatic cell nuclear transfer bypasses germline reprogramming

exhibit gene expression abnormalities at several developmental stages (Daniels *et al*., 2000; Daniels *et al*., 2001; Humpherys *et al*., 2002; Inoue *et al*., 2002; Wrenzycki *et al*., 2001) including essential embryonic genes (Boiani *et al*., 2002; Bortvin *et al*., 2003). Genes subject to parental allele-specific imprinting have been considered key to the problems in clone development, as loss of the parental-specific imprint would require the germline passage to be reestablished (Jaenisch, 1997). Allelic expression of imprinted genes is regulated by parental-specific imprinting marks that are set in the germline, some of which involve differential methylation of regulatory regions. Dysregulation of imprinted genes has severe consequences on development, apparent in the early death of uniparental embryos (Barton *et al*., 1984; Mann and Lovell-Badge, 1984; McGrath and Solter, 1984; Surani *et al*., 1984), the abnormalities observed in those with uniparental duplications of chromosomal regions (Cattanach, 1986) and targeted disruption of imprinted gene expression (DeChiara *et al*., 1990; Eggenschwiler *et al*., 1997; Lau *et al*., 1994; Leighton *et al*., 1995). Abnormalities associated with disruption in the expression of imprinted genes regulating fetal growth, including insulin-like growth factor 2 (Igf2) and Igf2 receptor (Li *et al*., 1998; Reik and Maher, 1997), are similar to some of the most common phenotypes of clones: respiratory failure (Hill *et al*., 1999; Ogura *et al*., 2002; Wells, 2003), fetal overgrowth (Young *et al*., 1998), increased birth weight (Eggan *et al*., 2001), and placental hypertrophy (Hill *et al*., 2000; Humpherys *et al*., 2001; Ono *et al*., 2001; Tanaka *et al*., 2001; Wakayama and Yanagimachi, 1999a; Wakayama and Yanagimachi, 2001). Several imprinted genes are indeed expressed at abnormal levels in clones both at fetal or perinatal stages, but particularly so in the placentae (Humpherys *et al*., 2002; Inoue *et al*., 2002). The imprinted genes commonly implicated in the dysregulation of fetal growth, Igf2, Igf2r, and H19, are, however, expressed at normal levels and presumably not implicated (Humpherys *et al*., 2002; Inoue *et al*., 2002; Wells, 2003). Additionally, the abnormal expression levels observed in several imprinted genes analyzed were not associated with an imbalance in the allele specificity (Humpherys *et al*., 2002; Inoue *et al*., 2002). Therefore, the aberrant expression of imprinted genes seems to

be stochastic with no apparent correlation to phenotypes, at least in those clones that develop to midgestation or later. As this proportion of clones represents a minority of those generated, it does not preclude that dysregulation of imprinted genes contributes to the death of the large high proportion of clones that occurs early in gestation.

Analysis of clones at early developmental stages, such as the blastocyst, allows a more representative analysis of epigenetic and gene expression changes in the majority of clones, presumably indicative of reprogramming. In the mouse, most clones at the blastocyst stage retain or emulate the allele specificity of expression of imprinted genes with monoallelic expression patterns in both somatic cells and the early embryo, such as H19, Meg3, or Snrpn (Mann *et al*., 2003). This may reflect either continuation or a correct reestablishment of the allele-specific imprint. However, the differential methylation patterns of control regions involved in the regulation of allelic expression appear not to be established correctly for H19 and Snrpn. This may not influence preimplantation stage expression but could predict the loss of allele-specific expression in postimplanation development. For two autosomal imprinted genes that differ in expression between preimplantation stage and the soma, Ascl2 and Igfr2, being expressed biallelically in the early embryo but monoallelically in somatic cells, expression patterns are random with both mono- and biallelic expression patterns in clone mouse blastocysts. This suggests that reprogramming required for regulation of allele-specific expression often fails after somatic cell nuclear transfer. The proportion of clones at the blastocyst stage with correct reprogramming/expression of autosomal imprinted genes is very low (4%; Mann *et al*., 2003). Methylation and expression of imprinted genes are, however, altered in embryos due to culture in vitro and possibly have long-term effects on development (Doherty *et al*., 2000; Khosla *et al*., 2001; Young *et al*., 2001). Clones require considerable periods of in vitro culture, and are apparently not an exception to this phenomenon (Mann *et al*., 2003).

3. X inactivation

In contrast to autosomal imprinted genes, faithful recapitulation from mono- to biallelic activity has been observed in mouse clones with respect to X inactivation. The inactive X from a female somatic donor nucleus becomes activated in clones during preimplantation stages, followed by random X inactivation in the epiblast (Eggan *et al*., 2000). Since nonrandom inactivation is observed in the trophectoderm lineage of midgestation clones, with preferential inactivation of the previously inactive X chromosome of the somatic donor nucleus, it appears that an X chromosome silenced randomly in the epiblast carries an imprinting mark that is functionally equivalent to that of the paternal X chromosome of the zygotic genome. Placental tissue of perinatal bovine clones exhibited preferential inactivation of one X chromosome in viable clones, contrasting with biallelic activity in clones that died at birth, suggesting a similar recognition process in bovine clones but also a possible correlation of biased X inactivation in extraembryonic tissues with viability (Xue *et al*., 2002). However, as cloning of mammals from female versus male somatic

cells is similar in efficiency, X inactivation is apparently not a limiting factor in clone development (Heyman *et al*., 2002; Kato *et al*., 2000).

4. Global methylation changes in clones

The developmental stages subsequent to the stage at which nuclear transfer is performed, normally include a sequence of major changes in the methylation status of the zygotic genome. In murine and bovine embryos, the paternal genome appears to be rapidly demethylated within hours of fertilization ("active" demethylation), while demethylation of the maternal genome occurs in a replication-dependent manner during cleavage stages ("passive" demethylation) (Dean *et al*., 2001; Mayer *et al*., 2000; Oswald *et al*., 2000; Rougier *et al*., 1998). Global de novo methylation of the hypomethylated embryonic genome occurs at late preimplantation stages and after implantation. The recapitulation of stage-specific methylation patterns in somatic cell clones of several species is inconsistent, which can be interpreted as errors in reprogramming. Analysis of global methylation of genomic repetitive elements in the early cleavage (Bourc'his *et al*., 2001), morula (Dean *et al*., 2001), and blastocyst (Kang *et al*., 2001) stage, bovine clones revealed abnormally high levels of methylation, similar to those of the somatic donor nuclei. While observations differ in respect to conservation (Dean *et al*., 2001) or absence (Bourc'his *et al*., 2001) of an initial wave of active demethylation, there is a consensus that passive demethylation during early cleavage stages is lacking (Kang *et al*., 2001). Faithful recapitulation of preimplantation-stage-specific methylation changes occurred on one single-copy sequence analyzed (Kang *et al*., 2003) but not on satellite sequences (Bourc'his *et al*., 2001; Kang *et al*., 2001; Kang *et al*., 2002). These contrasts in methylation for different sequences have been ascribed to the structural differences between permanent (satellite sequences) and facultative (unique sequences) heterochromatin (Kang *et al*., 2003). Interpretation of methylation status changes at the early developmental stages cannot predict developmental outcomes, however, the widespread dysregulation of global methylation, particularly in the trophectoderm of bovine clone blastocysts (Kang *et al*., 2002), is consistent with the overall and, in particular, the placental gene expression abnormalities in clones at later developmental stages. Some variation from the normal state of DNA methylation is compatible with postnatal development (Humpherys *et al*., 2001; Ohgane *et al*., 2001). However, major level differences in global methylation coincide with loss of viability, as evident from null mutants for DNA methyltransferase in the mouse (Jaenisch and Bird, 2003) and the observation of genome-wide hypomethylation in the majority of aborted bovine clone fetuses at late gestational stages (Cezar *et al*., 2003).

5. Conclusions

The abnormal or failing development of mammalian somatic cell clones presumably reflects inappropriate gene expression, precipitated by failure to put in place the necessary epigenetic information. It should also be considered that the cloning

efficiency using donor nuclei from less-differentiated cells such as embryonic blastomeres and cells is low, and possibly limited by factors other than reprogramming of gene expression (Cheong *et al*., 1993; Eggan *et al*., 2001; Otaegui *et al*., 1994; Prather *et al*., 1987; Rideout *et al*., 2000; Tsunoda and Kato, 1997; Willadsen, 1986).

Reprogramming of the somatic cell nucleus is typically imperfect, evident at both the level of expression and in the epigenetic modifications of imprinted genes but is possibly due to the retention of characteristics of the somatic cell, at the exclusion of, or conflicting with normal developmental epigenetics. Additional evidence for the maintenance of the epigenome of the donor nucleus stems from studies indicating retention of somatic cell–like metabolism (Chung *et al*., 2002) and gene expression (Gao *et al*., 2003). As the abnormal phenotypes observed in clones coincide with abnormal gene expression but do not necessarily correlate with imprinted gene expression, reprogramming errors are apparently genome-wide (Humpherys *et al*., 2002). This is consistent with the finding that many genes are downregulated in clones. The threshold at which these differences prevent development are unknown, and gene expression and methylation variation in viable clones at term exemplifies that development is somewhat tolerant to variation in epigenetics. To assess the quality of reprogramming, it will be essential to define those epigenetic criteria that not only need to be reset after nuclear transfer but also correlate with viability.

Acknowledgments

We thank Jeffrey R. Mann for critical comments on the manuscript.

References

Barton SC, Surani MA and Norris ML (1984) Role of paternal and maternal genomes in mouse development. *Nature*, **311**, 374–376.

Boiani M, Eckardt S, Scholer HR and McLaughlin KJ (2002) Oct4 distribution and level in mouse clones: consequences for pluripotency. *Genes & Development*, **16**, 1209–1219.

Bortvin A, Eggan K, Skaletsky H, Akutsu H, Berry DL, Yanagimachi R, Page DC and Jaenisch R (2003) Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. *Development*, **130**, 1673–1680.

Bourc'his D, Le Bourhis D, Patin D, Niveleau A, Comizzoli P, Renard JP and Viegas-Pequignot E (2001) Delayed and incomplete reprogramming of chromosome methylation patterns in bovine cloned embryos. *Current Biology*, **11**, 1542–1546.

Cattanach BM (1986) Parental origin effects in mice. *Journal of Embryology and Experimental Morphology*, (97 Suppl), 137–150.

Cezar GG, Bartolomei MS, Forsberg EJ, First NL, Bishop MD and Eilertsen KJ (2003) Genome-wide epigenetic alterations in cloned bovine fetuses. *Biology of Reproduction*, **68**, 1009–1014.

Cheong HT, Takahashi Y and Kanagawa H (1993) Birth of mice after transplantation of early cell-cycle-stage embryonic nuclei into enucleated oocytes. *Biology of Reproduction*, **48**, 958–963.

Chung YG, Mann MR, Bartolomei MS and Latham KE (2002) Nuclear-cytoplasmic "tug of war" during cloning: effects of somatic cell nuclei on culture medium preferences of preimplantation cloned mouse embryos. *Biology of Reproduction*, **66**, 1178–1184.

Daniels R, Hall V and Trounson AO (2000) Analysis of gene transcription in bovine nuclear transfer embryos reconstructed with granulosa cell nuclei. *Biology of Reproduction*, **63**, 1034–1040.

Daniels R, Hall VJ, French AJ, Korfiatis NA and Trounson AO (2001) Comparison of gene transcription in cloned bovine embryos produced by different nuclear transfer techniques. *Molecular Reproduction and Development*, **60**, 281–288.

Dean W, Santos F, Stojkovic M, Zakhartchenko V, Walter J, Wolf E and Reik W (2001) Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13734–13738.

DeChiara TM, Efstratiadis A and Robertson EJ (1990) A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. *Nature*, **345**, 78–80.

Doherty AS, Mann MR, Tremblay KD, Bartolomei MS and Schultz RM (2000) Differential effects of culture on imprinted H19 expression in the preimplantation mouse embryo. *Biology of Reproduction*, **62**, 1526–1535.

Eggan K, Akutsu H, Hochedlinger K, Rideout W 3rd, Yanagimachi R and Jaenisch R (2000) X-Chromosome inactivation in cloned mouse embryos. *Science*, **290**, 1578–1581.

Eggan K, Akutsu H, Loring J, Jackson-Grusby L, Klemm M, Rideout WM 3rd, Yanagimachi R and Jaenisch R (2001) Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6209–6214.

Eggenschwiler J, Ludwig T, Fisher P, Leighton PA, Tilghman SM and Efstratiadis A (1997) Mouse mutant embryos overexpressing IGF-II exhibit phenotypic features of the Beckwith-Wiedemann and Simpson-Golabi-Behmel syndromes. *Genes & Development*, **11**, 3128–3142.

Gao S, Chung YG, Williams JW, Riley J, Moley K and Latham KE (2003) Somatic cell-like features of cloned mouse embryos prepared with cultured myoblast nuclei. *Biology of Reproduction*, **69**, 48–56.

Heyman Y, Zhou Q, Lebourhis D, Chavatte-Palmer P, Renard JP and Vignon X (2002) Novel approaches and hurdles to somatic cloning in cattle. *Cloning and Stem Cells*, **4**, 47–55.

Hill JR, Burghardt RC, Jones K, Long CR, Looney CR, Shin T, Spencer TE, Thompson JA, Winger QA and Westhusin ME (2000) Evidence for placental abnormality as the major cause of mortality in first-trimester somatic cell cloned bovine fetuses. *Biology of Reproduction*, **63**, 1787–1794.

Hill JR, Roussel AJ, Cibelli JB, Edwards JF, Hooper NL, Miller MW, Thompson JA, Looney CR, Westhusin ME, Robl JM, *et al*. (1999) Clinical and pathologic features of cloned transgenic calves and fetuses (13 case studies). *Theriogenology*, **51**, 1451–1465.

Humpherys D, Eggan K, Akutsu H, Friedman A, Hochedlinger K, Yanagimachi R, Lander ES, Golub TR and Jaenisch R (2002) Abnormal gene expression in cloned mice derived from embryonic stem cell and cumulus cell nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12889–12894.

Humpherys D, Eggan K, Akutsu H, Hochedlinger K, Rideout WM 3rd, Biniszkiewicz D, Yanagimachi R and Jaenisch R (2001) Epigenetic instability in ES cells and cloned mice. *Science*, **293**, 95–97.

Inoue K, Kohda T, Lee J, Ogonuki N, Mochida K, Noguchi Y, Tanemura K, Kaneko-Ishino T, Ishino F and Ogura A (2002) Faithful expression of imprinted genes in cloned mice. *Science*, **295**, 297.

Jaenisch R (1997) DNA methylation and imprinting: Why bother? *Trends in Genetics*, **13**, 323–329.

Jaenisch R, and Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, (33 Suppl), 245–254.

Kang YK, Koo DB, Park JS, Choi YH, Chung AS, Lee KK and Han YM (2001) Aberrant methylation of donor genome in cloned bovine embryos. *Nature Genetics*, **28**, 173–177.

Kang YK, Park JS, Koo DB, Choi YH, Kim SU, Lee KK and Han YM (2002) Limited demethylation leaves mosaic-type methylation states in cloned bovine pre-implantation embryos. *The EMBO Journal*, **21**, 1092–1100.

Kang YK, Yeo S, Kim SH, Koo DB, Park JS, Wee G, Han JS, Oh KB, Lee KK and Han YM (2003) Precise recapitulation of methylation change in early cloned embryos. *Molecular Reproduction and Development*, **66**, 32–37.

Kato Y, Tani T and Tsunoda Y (2000) Cloning of calves from various somatic cell types of male and female adult, newborn and fetal cows. *Journal of Reproduction and Fertility*, **120**, 231–237.

Khosla S, Dean W, Brown D, Reik W and Feil R (2001) Culture of preimplantation mouse embryos affects fetal development and the expression of imprinted genes. *Biology of Reproduction*, **64**, 918–926.

Lau MM, Stewart CE, Liu Z, Bhatt H, Rotwein P and Stewart CL (1994) Loss of the imprinted IGF2/cation-independent mannose 6-phosphate receptor results in fetal overgrowth and perinatal lethality. *Genes & Development*, **8**, 2953–2963.

Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A and Tilghman SM (1995) Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature*, **375**, 34–39.

Li M, Squire JA and Weksberg R (1998) Overgrowth syndromes and genomic imprinting: from mouse to man. *Clinical Genetics*, **53**, 165–170.

Mann JR (2001) Imprinting in the germ line. *Stem Cells*, **19**, 287–294.

Mann JR and Lovell-Badge RH (1984) Inviability of parthenogenones is determined by pronuclei, not egg cytoplasm. *Nature*, **310**, 66–67.

Mann MR, Chung YG, Nolen LD, Verona RI, Latham KE and Bartolomei MS (2003) Disruption of imprinted gene methylation and expression in cloned preimplantation stage mouse embryos. *Biology of Reproduction*, **69**, 902–914.

Mayer W, Niveleau A, Walter J, Fundele R and Haaf T (2000) Demethylation of the zygotic paternal genome. *Nature*, **403**, 501–502.

McGrath J and Solter D (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, **37**, 179–183.

Ogura A, Inoue K, Ogonuki N, Lee J, Kohda T and Ishino F (2002) Phenotypic effects of somatic cell cloning in the mouse. *Cloning and Stem Cells*, **4**, 397–405.

Ohgane J, Wakayama T, Kogo Y, Senda S, Hattori N, Tanaka S, Yanagimachi R and Shiota K (2001) DNA methylation variation in cloned mice. *Genesis*, **30**, 45–50.

Ono Y, Shimozawa N, Ito M and Kono T (2001) Cloned mice from fetal fibroblast cells arrested at metaphase by a serial nuclear transfer. *Biology of Reproduction*, **64**, 44–50.

Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R, Dean W, Reik W and Walter J (2000) Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, **10**, 475–478.

Otaegui PJ, O'Neill GT, Campbell KH and Wilmut I (1994) Transfer of nuclei from 8-cell stage mouse embryos following use of nocodazole to control the cell cycle. *Molecular Reproduction and Development*, **39**, 147–152.

Prather RS, Barnes FL, Sims MM, Robl JM, Eyestone WH and First NL (1987) Nuclear transplantation in the bovine embryo: assessment of donor nuclei and recipient oocyte. *Biology of Reproduction*, **37**, 859–866.

Reik W and Maher ER (1997) Imprinting in clusters: lessons from Beckwith-Wiedemann syndrome. *Trends in Genetics*, **13**, 330–334.

Reik W, Santos F and Dean W (2003) Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology*, **59**, 21–32.

Renard JP (1998) Chromatin remodelling and nuclear reprogramming at the onset of embryonic development in mammals. *Reproduction, Fertility, and Development*, **10**, 573–580.

Rideout WM 3rd, Eggan K and Jaenisch R (2001) Nuclear cloning and epigenetic reprogramming of the genome. *Science*, **293**, 1093–1098.

Rideout WM 3rd, Wakayama T, Wutz A, Eggan K, Jackson-Grusby L, Dausman J, Yanagimachi R and Jaenisch R (2000) Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nature Genetics*, **24**, 109–110.

Rougier N, Bourc'his D, Gomes DM, Niveleau A, Plachot M, Paldi A and Viegas-Pequignot E (1998) Chromosome methylation patterns during mammalian preimplantation development. *Genes & Development*, **12**, 2108–2113.

Russo VEA, Martienssen RA and Riggs AD (1996) *Epigenetic Mechanisms of Gene Regulation*, Cold Spring Harbor Laboratory Press: Cold Spring Harbor.

Surani MA, Barton SC and Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, **308**, 548–550.

Tanaka S, Oda M, Toyoshima Y, Wakayama T, Tanaka M, Yoshida N, Hattori N, Ohgane J, Yanagimachi R and Shiota K (2001) Placentomegaly in cloned mouse concepti caused by expansion of the spongiotrophoblast layer. *Biology of Reproduction*, **65**, 1813–1821.

Tsunoda Y and Kato Y (1997) Full-term development after transfer of nuclei from 4-cell and compacted morula stage embryos to enucleated oocytes in the mouse. *The Journal of Experimental Zoology*, **278**, 250–254.

Wakayama T, Shinkai Y, Tamashiro KL, Niida H, Blanchard DC, Blanchard RJ, Ogura A, Tanemura K, Tachibana M, Perry AC, *et al*. (2000) Cloning of mice to six generations. *Nature*, **407**, 318–319.

Wakayama T and Yanagimachi R (1999a) Cloning of male mice from adult tail-tip cells. *Nature Genetics*, **22**, 127–128.

Wakayama T and Yanagimachi R (1999b) Cloning the laboratory mouse. *Seminars in Cell & Developmental Biology*, **10**, 253–258.

Wakayama T and Yanagimachi R (2001) Mouse cloning with nucleus donor cells of different age and type. *Molecular Reproduction and Development*, **58**, 376–383.

Wells DN (2003) Cloning in livestock agriculture. *Reproduction Supplement*, **61**, 131–150.

Willadsen SM (1986) Nuclear transplantation in sheep embryos. *Nature*, **320**, 63–65.

Wrenzycki C, Wells D, Herrmann D, Miller A, Oliver J, Tervit R and Niemann H (2001) Nuclear transfer protocol affects messenger RNA expression patterns in cloned bovine blastocysts. *Biology of Reproduction*, **65**, 309–317.

Xue F, Tian XC, Du F, Kubota C, Taneja M, Dinnyes A, Dai Y, Levine H, Pereira LV and Yang X (2002) Aberrant patterns of X chromosome inactivation in bovine clones. *Nature Genetics*, **31**, 216–220.

Young LE, Fernandes K, McEvoy TG, Butterwith SC, Gutierrez CG, Carolan C, Broadbent PJ, Robinson JJ, Wilmut I and Sinclair KD (2001) Epigenetic change in IGF2R is associated with fetal overgrowth after sheep embryo culture. *Nature Genetics*, **27**, 153–154.

Young LE, Sinclair KD and Wilmut I (1998) Large offspring syndrome in cattle and sheep. *Reviews of Reproduction*, **3**, 155–163.

Short Specialist Review

Imprinted QTL in farm animals: a fortuity or a common phenomenon?

Martien A. M. Groenen
Wageningen University, Wageningen, The Netherlands

The analysis of complex multifactorial traits has always been at the forefront in livestock species. Contrary to human and mouse, monogenic traits and disorders are generally easily recognized and removed by culling potential carrier animals. Initially, breeders applied these aspects in the algorithms and programs used to calculate breeding values for these animals, but with the development of detailed linkage maps and high-throughput genotyping systems in the last two decades, it formed the basis for the localization of many loci underlying such complex traits (for a recent overview of identified QTL in pigs and chicken, see **http://www.animalgenome.org/QTLdb/** and **http://acedb.asg.wur.nl/**). Because these traits are quantitative in nature and are being influenced by a large number of genes as well as by environmental factors, they are referred to as quantitative trait loci or QTL. The initial algorithms used to analyze the data from a genome-wide scan, were restricted in considering only Mendelian inheritance and generally a single QTL per individual chromosome. More recently, several groups have started to extend these programs to include epistatic and epigenetic effects as well (Carlborg and Andersson, 2002; Carlborg *et al*., 2003; Knott *et al*., 1998; Jeon *et al*., 1999; Nezer *et al*., 1999; de Koning *et al*., 2000). Population structure and the outbred nature of most livestock breeds make them particularly well suited to address these genetic effects, something that is generally not possible in crosses between inbred laboratory rats and mice.

Although the potential implications of genetic imprinting with regard to quantitative traits in farm animals had already been described by de Vries *et al*. (1994), the interest for epigenetic effects in livestock species was particularly triggered by the identification of the callipyge locus in sheep. The callipyge locus was observed to be segregating in his herd by a sheep breeder in Oklahoma in 1983. The mutation results in an exceptional muscularity, and a subsequent genetic analysis revealed an exceptional behavior of this locus. Although the segregation of the callipyge phenotype clearly indicated an underlying mechanism of imprinting, the fact that homozygous carriers of the callipyge mutation did not express the callipyge phenotype was an intriguing observation and the effect was referred to by the authors as polar overdominance (Cockett *et al*., 1996). In subsequent studies, it was shown

that the callipyge mutation is located in a potential longe-range control element (LRCE) located between a group of paternally expressed (*DLK1* and *PEG11*) and maternally expressed (*GTL2* and *MEG8*) genes (Freking *et al*., 2002). The current working hypothesis for the observed polar overdominance at the callipyge locus is that (one of) the paternally expressed genes has a strong positive effect on muscle development and that (one of) the maternally expressed genes is a negative regulator of the paternally expressed gene(s). In this model, the LRCE is a gain-of-function mutation that results in an increase of both the maternally and paternally expressed genes (Georges *et al*., 2003). Although the callipyge mutation clearly affects a quantitative trait, its large effect on the phenotype enabled the analysis of the locus as a monogenic trait rather than a QTL.

The interest with regard to imprinting in relation to QTL was further boosted by the simultaneous independent findings of two research groups that a QTL on the tip of the p-arm of pig chromosome 2 (SSC2), affecting muscle mass and fat deposition, showed clear evidence of being maternally imprinted (Nezer *et al*., 1999; Jeon *et al*., 1999). Subsequent research recently led to the identification of a mutation in a regulatory site of the maternally imprinted gene *IGF2* as being the causative mutation underlying the imprinted QTL (Van Laere *et al*., 2003). The mutation is located in a CpG island located within intron 3 of the *IGF2* gene, which results in the abrogation of the binding of a repressor, resulting in an increased expression of the *IGF2* gene in pig muscle.

Triggered by the findings of the callipyge and *IGF2* gene, de Koning *et al*. (2000) decided to include a parent-of-origin effect in their statistical model for the genome-wide identification of QTL in a divergent cross between Chinese Meishan with European white pigs. Their results indicated a surprisingly large number of QTL that seemed to be affected by either maternal or paternal imprinting. The paternally expressed QTL on SSC2 in that study was recently (Jungerius *et al*., 2004) shown to be caused by the same mutation in *IGF2* that was identified by Van Laere *et al*. (2003). This mutation, however, was not responsible for the imprinted QTL on SSC2 affecting teat number (Hirooka *et al*., 2001; Jungerius *et al*., 2004). Additional analysis by these researchers and by others using another similar cross resulted in the further identification of several QTL for different growth-related traits on a number of different chromosomes (Rattink *et al*., 2000; De Koning *et al*., 2001; Milan *et al*., 2002; Desautes *et al*., 2002; Quintanilla *et al*., 2002).

Why do so many QTL seem to be affected by imprinting? There are several aspects to consider in dealing with this question (*see* Article 37, **Evolution of genomic imprinting in mammals**, Volume 1). One of the reasons that might cause the unexpectedly large number of imprinted QTLs is due to the nature of the traits being studied in these studies, which mainly were related to body composition and growth. The parental conflict hypotheses for imprinting (Sleutels and Barlow, 2002) suggests that there is a battle between male and female based on a trade-off for the survival of offspring versus that of the survival and fecundity of the mother. Consequently, genes affecting growth might be expected to be affected by imprinting at a higher frequency then the average gene. The fact that the paternally expressed genes identified for callipyge and on SSC2 have a strong positive effect on growth would support this hypothesis. However, this is not the case for all the

observed QTL effects. Furthermore, an imprinted QTL controlling susceptibility to trypanosomiasis with no clear relevance to growth was recently identified in mice (Clapcott *et al*., 2000). The experimental design of the QTL study and, in particular, the family structure of the population being studied has been shown to strongly affect the chance of identifying spurious imprinted QTL (De Koning *et al*., 2002). In particular, when the number of F1 animals is small or for smaller QTL effects when the founder lines are not fixed for different QTL alleles, spurious detection of imprinted QTL is a serious problem. A skewed distribution of uninformative markers between the male and female parents potentially could also result in the spurious identification of imprinted QTL in these crosses, although this pitfall could be excluded in the pig studies described above. Finally, other phenomena such as epistatic interactions between different genes might further complicate the correct identification of imprinting effects. Recently, Carlborg *et al*. (2003) have shown that epistatic effects also seem to play a major role in QTL identified for a large number of growth related traits.

Although several of the identified imprinted QTL effects are likely to be spurious effects, the studies described above provide accumulating evidence that imprinting plays a more important role in multifactorial traits than previously anticipated. Furthermore, the identification of the callipyge and *IGF2* mutations provide further insight into the importance of genetic variation within regulatory regions on quantitative traits. For animal breeding practice, the identification of major imprinted loci affecting body composition has several implications. It calls for a revision of the breeding value evaluation methods and breeding strategies that are currently solely based on the assumption of a large number of genes showing Mendelian expression. Detecting the QTL and confirming the mode of inheritance in commercial populations would open important new opportunities for pig-breeding companies. Imprinted QTL for fatness, for example, offers the opportunity to produce crossbred sows that have higher levels of fat reserves (beneficial for their health and reproduction), while their offspring have lower amounts of fat (requested by the consumer).

The results from the QTL studies in farm animals clearly have emphasized the importance of the inclusion of statistical testing for imprinting in the analysis of complex traits not only in animal genetics but in human medical genetics as well. Subsequent detailed molecular analysis and definite proof of the imprinting effects have to await the identification of the underlying genes responsible for the observed QTL effects.

References

Carlborg Ö and Andersson L (2002) The use of randomization testing for detection of multiple epistatic QTL. *Genetical Research*, **79**, 175–184.

Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P and Andersson L (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Research*, **13**, 413–421.

Clapcott SJ, Teale AJ and Kemp SJ (2000) Evidence for genomic imprinting of the major QTL controlling susceptibility to trypanosomiasis in mice. *Parasite Immunology*, **22**, 259–263.

Cockett NE, Jackson SP, Shay TL, Farnir F, Berghmans S, Snowder GD, Nielsen DM and Georges M (1996) Polar overdominance at the ovine callipyge locus. *Science*, **273**, 236–238.

De Koning DJ, Bovenhuis H and van Arendonk JAM (2002) On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics*, **161**, 931–938.

De Koning DJ, Rattink AP, Harlizius B, Groenen MAM, Brascamp EW and van Arendonk JAM (2001) Detection and characterization of quantitative trait loci for growth and reproduction traits in pigs. *Livestock Production Science*, **72**, 185–198.

de Koning DJ, Rattink AP, Harlizius B, van Arendonk JAM, Brascamp EW and Groenen MAM (2000) Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 7947–7950.

Desautes C, Bidanel JP, Milan D, Iannuccelli N, Amigues Y, Bourgeois F, Caritez JC, Renard C, Chevalet C and Mormede P (2002) Genetic linkage mapping of quantitative trait loci for behavioural and neuroendocrine stress response traits in pigs. *Journal of Animal Science*, **80**, 2276–2285.

de Vries AG, Kerr R, Tier B and Long T (1994) Gametic imprinting effects on rate and composition of pig growth. *Theoretical and Applied Genetics*, **88**, 1037–1042.

Freking BA, Murphy SK, Wylie AA, Rhodes SJ, Keele JW, Leymaster KA, Jirtle RL and Smith TP (2002) Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Research*, **12**, 1496–1506.

Georges M, Charlier C and Cockett N (2003) The callipyge locus: evidence for the trans interaction of reciprocally imprinted genes. *Trends in Genetics*, **19**, 248–252.

Hirooka H, de Koning D-J, Harlizius B, van Arendonk JAM, Rattink AP, Groenen MAM, Brascamp EW and Bovenhuis H (2001) A whole-genome scan for quantitative trait loci affecting teat number in pigs. *Journal of Animal Science*, **79**, 2320–2326.

Jeon JT, Carlborg O, Tornsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundstrom K, *et al*. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nature Genetics*, **21**, 157–158.

Jungerius BJ, Van Laere A-S, te Pas MFW, van Oost BA, Andersson L and Groenen MAM (2004) The IGF2-intron3-G3072A substitution explains a major imprinted QTL effect on backfat thickness in a Meishan X European white pig intercross. *Genetical Research*, **84**, 95–101.

Knott SA, Marklund L, Haley CS, Andersson K and Davies W (1998) Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics*, **149**, 1069–1080.

Milan D, Bidanel JP, Iannuccelli N, Riquet J, Amigues Y, Gruand J, Le Roy P, Renard C and Chevalet C (2002) Detection of quantitative trait loci for carcass composition traits in pigs. *Genetics, Selection, Evolution*, **34**, 705–728.

Nezer C, Moreau L, Brouwers B, Coppieters W, Detilleux J, Hanset R, Karim L, Kvasz A, Leroy P and Georges M (1999) An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics*, **21**, 155–156.

Quintanilla R, Milan D and Bidanel JP (2002) A further look at quantitative trait loci affecting growth and fatness in a cross between Meishan and Large White pig populations. *Genetics, Selection, Evolution*, **34**, 193–210.

Rattink AP, de Koning DJ, Faivre M, Harlizius B, van Arendonk JAM and Groenen MAM (2000) Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mammalian Genome*, **11**, 656–661.

Sleutels F and Barlow DP (2002) The origins of genomic imprinting in mammals. In *Homology Effects, Advances in Genetics*, Dunlap JC and Wu C (Eds.), Academic Press, pp. 119–163.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, *et al*. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*, **425**, 832–836.

Short Specialist Review

Variable expressivity and epigenetics

Marnie E. Blewitt and Emma Whitelaw

University of Sydney, Sydney, Australia

1. Introduction

It is clear that the phenotype of an organism cannot be completely described by its genotype alone. The phenotypic differences between humans, or between individuals of other species, appear to be greater than can be explained by the frequency of single nucleotide polymorphisms (SNPs) even when the environment is "controlled". In some cases, we know that the variation in phenotype can be ascribed to epigenetic rather than genetic differences. Epigenetic changes involve modifications to the DNA, such as methylation of cytosine residues, and modifications to the chromatin, such as acetylation and methylation of histone proteins. One of the underlying aspects of epigenetic control of gene expression is that it is stochastic, that is, there is a certain probability of a particular epigenetic state being established at a particular locus, but this is not necessarily 100%, and has profound implications on phenotype.

2. Variegation

For centuries, variegated appearances have been observed in plants and mammals. Variegated leaves and brindle coat colors in dogs are good examples. In humans, the iris often has a flecked appearance. Variegation can be defined as the differential expression of a particular gene among cells of the same cell type. The more recent finding of variegation in the expression of transgenes has provided a tractable experimental system for studying the molecular basis of this phenomenon in mammals (Mintz and Bradl, 1991; for review, Martin and Whitelaw, 1996). For reasons that remain unclear, transgenes appear to be particularly sensitive to variegation, and, at least in these cases, differential expression of the foreign DNA element is the result of differential transcriptional activity. The differential transcriptional activity correlates with epigenetic modifications at the transgene locus. In general, expressing cells are found to display hypomethylation of cytosine residues and an open chromatin state at the transgene promoter, while nonexpressing or silent cells are found to be hypermethylated at CpG dinucleotides, with a closed chromatin state (Allen *et al*., 1990; Festenstein *et al*., 1996; Garrick *et al*., 1996; Weichman

and Chaillet, 1997; Sutherland *et al*., 2000; Kearns *et al*., 2000). This chromatin compaction is reminiscent of what had already been reported as the molecular basis of a phenomenon called position effect variegation (PEV) in *Drosophila* (Henikoff, 1990), although *Drosophila* lacks CpG methylation.

3. Variable expressivity

Variable expressivity, where a range of expression states is observed among littermates, is another interesting feature of some transgenes. Variable expressivity, also called incomplete penetrance, is a term frequently used by clinicians to describe the feature of diseases where despite several family members being carriers of the disease, only some actually go on to present with symptoms. In these cases, the variable expressivity of the disease is rationalized by differences in quantitative trait loci (QTLs) between family members. However, variable expressivity at transgenes made in genetically identical, inbred strains of mice cannot be so simply explained. Originally, this was reported for transgenes made in mixed genetic backgrounds (Allen *et al*., 1990; Dobie *et al*., 1996), or closed colonies (Sutherland *et al*., 2000; Kearns *et al*., 2000), but these results have also been confirmed in inbred strains, where the individuals are genetically identical (Weichman and Chaillet, 1997). In these cases, transgene expression can range from very high, to variegated, to completely silenced, an unexpected event in isogenic individuals.

4. Subtle parent-of-origin effects

Furthermore, the extent to which a transgene is expressed can be dependent on the parent-of-origin of the transgene. This is often a subtle parent-of-origin effect, which differs from classic parental imprinting (Maggert and Golic, 2002; Preis *et al*., 2003). Classic parental imprinting refers to a situation in which an allele is active when inherited from one parent and inactive when inherited from the other. Subtle parent-of-origin effects refer to situations in which the probability of expression shifts depending on whether the allele has been inherited from the mother or the father. For example, the amount of variegation at the transgene may vary by approximately 20%, depending on parental origin (Preis *et al*., 2003). If the transgene displays variable expressivity, that is, a range of expression states between individuals, then the parental origin can shift the proportion of individuals in each class (Allen *et al*., 1990; Kearns *et al*., 2000). This shift is a probabilistic event, rather than a complete switch in expression, as seen in parentally imprinted alleles.

Recent work on *Drosophila* P transposon insertions has shown that 22 of 23 insertions into the *Drosophila* Y chromosome display subtle parent-of-origin effects (Maggert and Golic, 2002). In this case, the phenotypic assay is eye color, where the amount of variegation between the red and white pigment varies according to the parental origin of the Y chromosome. For some insertion sites, the transgene is expressed more following maternal transmission, and at other sites the reverse is observed. This result implies that large regions of the Y chromosome in *Drosophila* are subject to parent-of-origin effects (Maggert and Golic, 2002). In mammals, it

is also possible that large portions of the genome are subtly imprinted since small effects could easily have been overlooked.

It is now clear that variegation, variable expressivity, and subtle parent-of-origin effects are not peculiar to transgenes but are seen in endogenous alleles in mice, plants, and *Drosophila*. To date, the reported alleles that display these effects produce visual phenotypes. It is still unclear how common such alleles are, and whether they exist in humans. In plants and mammals, these alleles are now termed metastable epialleles (Brink, 1960; Matzke *et al*., 1994; Hollick *et al*., 1995 and Rakyan *et al*., 2002).

Examples of murine metastable epialleles include *agouti viable yellow* (A^{vy}) (Perry *et al*., 1994), *agouti intracisternal A particle yellow* (A^{iapy}) (Michaud *et al*., 1994), and *agouti hypervariable yellow* (A^{hvy}) (Argeson, 1996), also *axin fused* ($axin^{Fu}$) (Reed, 1937; Ruvinsky and Agulnik, 1990; Zeng *et al*., 1997), *axial defects* (Essien *et al*., 1990), *disorganization* (Hummel, 1959), and *murine CDK5 activator binding protein IAP* ($mCABP^{IAP}$) (Druker, *et al*., 2004). It is interesting to note that A^{vy}, A^{iapy}, A^{hvy}, $axin^{Fu}$, and $mCABP^{IAP}$ all arose as a result of the stable integration of a retrotransposon. In the case of the *agouti* alleles and $axin^{Fu}$, the variable expressivity arises as a direct result of variable activity at a cryptic promoter in the retrotransposon long terminal repeat (LTR), which reads out into the adjacent DNA (Michaud *et al*., 1994; Perry *et al*., 1994 and Argeson, 1996). It is possible that all endogenous metastable epialleles will turn out to be under the control of nearby retrotransposons. Given that over 9% of the human genome has been classified as retrotransposon-like (International Human Genome Sequencing Consortium, 2001), metastable epialleles may also be present in reasonable numbers in humans.

The A^{vy} allele is perhaps the best-characterized metastable epiallele. In this case, the allele results from an intracisternal A particle (IAP) insertion, upstream of the *agouti* gene (Duhl *et al*., 1994). When the cryptic promoter in the IAP LTR is active, it causes constitutive expression of the *agouti* gene (Duhl *et al*., 1994). Agouti is a signaling molecule that causes a shift in pigment production from black to yellow. Normally, the *agouti* gene is under the control of hair-cycle-specific promoters, and is produced for only a short period in the hair growth cycle, to produce a subapical yellow band on an otherwise black hair. The resultant mouse appears brown (an agouti coat). However, when the *agouti* gene is constitutively expressed, a completely yellow coat results, along with other pleiotropic effects including obesity and diabetes. Sometimes, a genetically identical mouse will have an agouti-colored coat (termed pseudoagouti). In this case, the cryptic promoter is silent, and CpG-methylated, and the *agouti* gene undergoes normal spatiotemporal expression (Morgan *et al*., 1999). A spectrum of intermediate mottled phenotypes, where there is variegated A^{vy} allele expression, is also observed (see Figure 1).

The A^{vy} allele displays a subtle parent-of-origin effect (Morgan *et al*., 1999), as do many other endogenous murine metastable epialleles (Reed, 1937; Belyaev *et al*., 1981; Essien *et al*., 1990; Duhl *et al*., 1994; Argeson, 1996; Rakyan *et al*., 2003). In this case, a yellow female produces a greater proportion of yellow offspring than a yellow male. There is a 15% difference in the proportion of pseudoagouti mice produced from these reciprocal crosses (Morgan *et al*., 1999).

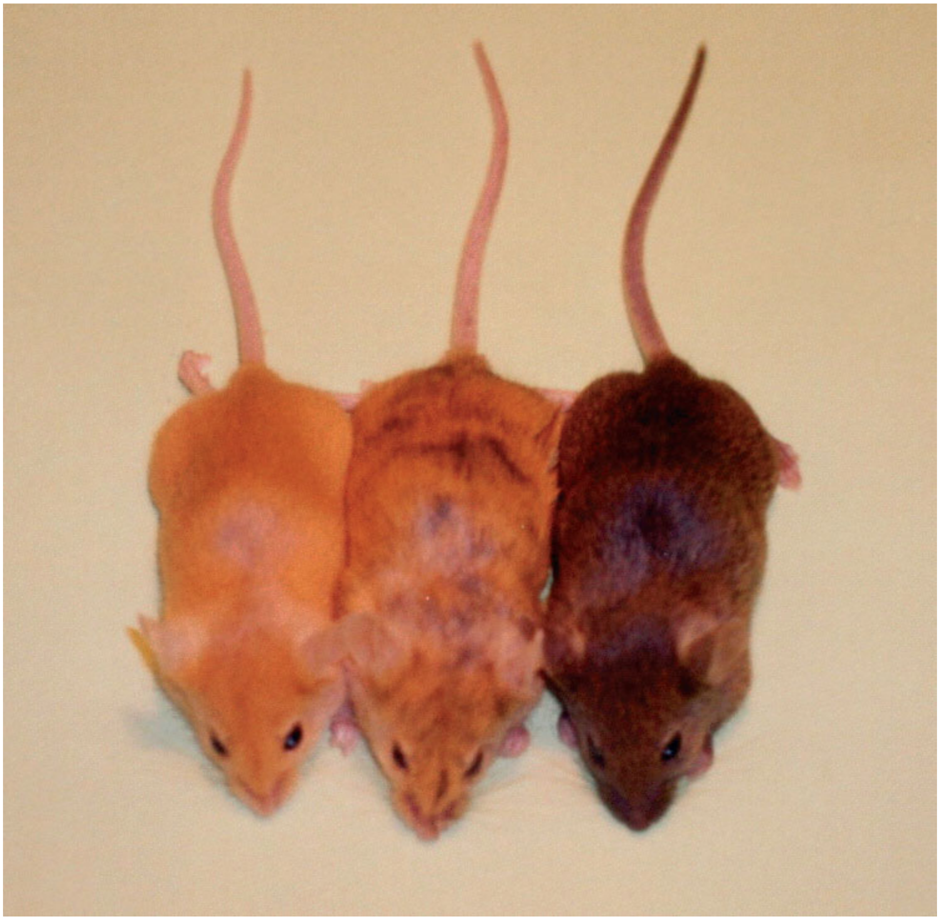


Figure 1 Genetically identical mice carrying the A^{vy} allele display variegation and variable expressivity. The mouse on the left is termed yellow, the middle mouse is termed mottled, and the mouse to the right is termed pseudoagouti

Again, this is a small, but significant shift in the spectrum of coat color phenotypes seen at the A^{vy} allele, which is dependent on parental origin (see Figure 2).

5. Transgenerational epigenetic inheritance

An additional effect seen at some murine transgenes (Hadchouel *et al*., 1987; Allen *et al*., 1990; Kearns *et al*., 2000; Sutherland *et al*., 2000) and some metastable epialleles, including the A^{vy} allele, is transgenerational epigenetic inheritance (Belyaev *et al*., 1981; Wolff, 1978; Morgan *et al*., 1999; Rakyan *et al*., 2003). A yellow-coated A^{vy} female produces more yellow offspring than a pseudoagouti female, despite being genetically identical (Morgan *et al*., 1999; see Figure 2). That is, there is some memory of the phenotype, and, therefore, the epigenotype of the mother. The mechanism of epigenetic inheritance is not known, but it seems likely

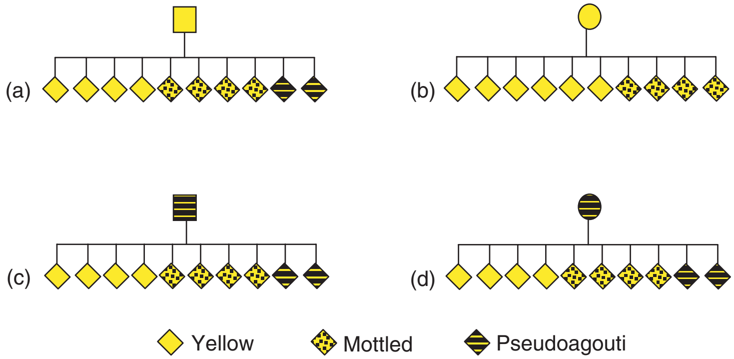


Figure 2 Schematic pedigrees of coat color inheritance at the A^{vy} allele. Mice heterozygous for the A^{vy} allele mated with congenic animals, carrying a null *agouti* allele. Only offspring carrying the allele are displayed. Comparing (a) and (b), there are more yellow mice observed following female transmission of the allele from a yellow parent. This is a subtle parent-of-origin effect, which shifts the proportion of each phenotype of offspring. Comparing (b) and (d), the yellow female produces more yellow offspring than the pseudoagouti female, so the range of phenotypes of the offspring is influenced by the phenotype of the dam. This is transgenerational epigenetic inheritance

that there is incomplete clearing of the epigenetic marks between generations. Again, it is unclear how often epigenetic inheritance may be occurring, but it certainly has profound implications on the inheritance of phenotype in general. Interestingly, although epigenetic inheritance can occur through the male germline (Rakyan *et al*., 2003), it is more common following female transmission. It is interesting to consider whether this mode of inheritance has some adaptive value. In higher organisms, the mother will often stay with the young after birth. It may be advantageous for the phenotype of the offspring to be more closely related to that of the mother, who is presumably well adapted to the surrounding physical environment.

A recent finding in *Drosophila* suggests that epigenetic inheritance of chromatin state could provide a mechanism for rapid evolution of morphological traits (Sollars *et al*., 2003). Sollars and coworkers found that epigenetic variants produced de novo, can cause phenotypic variations. The variants were produced by genetically altering levels of chromatin proteins. Unexpectedly, the phenotypic variation could be inherited through the female germline for successive generations, in the absence of the original genetic mutation. The inheritance resulted after only one meiotic generation, several generations faster than is required for genetic variation to drive to fixation (Rutherford and Henikoff, 2003).

The existence of variable expressivity and epigenetic inheritance, raises the possibility that phenotypic variation in humans may not be all QTL-based. Moreover, some of the evolution of quantitative traits may have been due to epigenetic, rather than genetic changes. Consider a situation in which an allele displays variable expressivity, due to differences in epigenetic state of the allele. One of the phenotypes may be selected owing to environmental change, and inherited through the

germline owing to transgenerational epigenetic inheritance. However, since epigenetic inheritance is never complete, variation is not lost, which may allow for a more plastic adaptation than the selection of genetic variation affords.

Further reading

Chong S and Whitelaw E (2004) Murine metastable epialleles and transgenerational epigenetic inheritance. *Cytogenetic and Genome Research*, **105**, 311–315.

Lloyd V (2000) Parental imprinting in *Drosophila*. *Genetica*, **109**, 35–44.

Vasicek TJ, Zeng LI, Zhang T, Costantini F and Tilghman SM (1997) Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics*, **147**, 777–786.

References

Allen ND, Norris ML and Surani MA (1990) Epigenetic control of transgene expression and imprinting by genotype-specific modifiers. *Cell*, **61**, 853–861.

Argeson AC (1996) The molecular basis of the pleiotropic phenotype of mice carrying the hypervariable yellow (A^{hvy}) mutation at the agouti locus. *Genetics*, **142**, 557–567.

Belyaev DK, Ruvinsky AO and Borodin PM (1981) Inheritance of alternative states of the fused gene in mice. *The Journal of Heredity*, **72**, 107–112.

Brink RA (1960) Paramutation and chromosome organization. *The Quarterly Review of Biology*, **35**, 120–137.

Dobie KW, Lee M, Fantes JA, Graham E and Clark AJ (1996) Variegated transgene expression in mouse mammary glans is determined by the transgene integration locus. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 6659–6664.

Druker R, Bruxner TJ, Lehrbach NJ and Whitelaw E (2004) Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Research*, **32**, 5800–5808.

Duhl DMJ, Vrieling H, Miller KA, Wolff GL and Barsh GS (1994) Neomorphic agouti mutations in obese yellow mice. *Nature Genetics*, **8**, 59–64.

Essien FB, Haviland MB and Naidoff AE (1990) Expression of a new mutation (Axd) causing axial defects in mice correlates with maternal phenotype and age. *Teratology*, **42**, 183–194.

Festenstein R, Tolaini M, Corbella P, Mamalaki C, Parrington J, Fox M, Miliou A, Jones M and Kioussis D (1996) Locus control region function and heterochromatin induced position effect variegation in transgenic mice. *Science*, **271**, 1123–1126.

Garrick D, Sutherland H, Robertson G and Whitelaw E (1996) Variegated expression of a globin transgene correlates with chromatin accessibility but not methylation status. *Nucleic Acids Research*, **24**, 4902–4909.

Hadchouel M, Farza H, Simon D, Tollias P and Pourcel C (1987) Maternal inhibition of hepatitis B surface antigen gene expression in transgenic mice correlates with de novo methylation. *Nature*, **329**, 454–456.

Henikoff S (1990) Position effect variegation after 60 years. *Trends in Genetics*, **6**, 422–426.

Hollick JB, Patterson GI, Coe EHJ, Cone KC and Chandler VL (1995) Allelic interactions heritably alter the activity of a metastable maize *pl* allele. *Genetics*, **141**, 709–719.

Hummel KP (1959) Developmental anomalies in mice resulting from the action of the gene disorganization, a semi-dominant lethal. *Pediatrics*, **23**, 212–221.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kearns M, Preis J, McDonald M, Morris C and Whitelaw E (2000) Complex patterns of inheritance of an imprinted murine transgene suggest incomplete germline erasure. *Nucleic Acids Research*, **28**, 3301–3309.

Maggert KA and Golic KG (2002) The Y chromosome of *Drosophila melanogaster* exhibits chromosome-wide imprinting. *Genetics*, **162**, 1245–1258.

Martin DI and Whitelaw E (1996) The vagaries of variegating transgenes. *Bioessays*, **18**, 919–923.

Matzke MA, Moscone EA, Parke YD, Papp I, Oberkofler H, Neuhuber F and Matzke AJ (1994) Inheritance and expression of a transgene insert in an aneuploidy tobacco line. *Molecular & General Genetics*, **245**(4), 471–485.

Michaud EJ, van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT and Woychik RP (1994) Differential expression of a new dominant agouti allele (A^{iapy}) is correlated with methylation state and is influenced by parental lineage. *Genes & Development*, **8**, 1463–1472.

Mintz B and Bradl M (1991) Mosaic expression of a tyrosinase fusion gene in albine mice yields a heritable striped coat colour in transgenic homozygotes. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(21), 9643–9647.

Morgan HD, Sutherland HGE, Martin DIK and Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics*, **23**, 314–318.

Perry WL, Copeland NG and Jenkins NA (1994) The molecular basis for dominant yellow agouti coat colour mutations. *Bioessays*, **16**, 705–707.

Preis JI, Downes M, Oates NA, Rasko JE and Whitelaw E (2003) Sensitive flow cytometric analysis reveals a novel type of subtle parent-of-origin effects in the mouse genome. *Current Biology*, **13**(11), 955–959.

Rakyan VR, Blewitt ME, Preis JI, Druker R and Whitelaw E (2002) Metastable epialleles in mammals. *Trends in Genetics*, **18**(7), 348–351.

Rakyan VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KVK and Whitelaw E (2003) Transgenerational inheritance of epigenetic states at the murine $Axin^{Fu}$ allele occurs following maternal and paternal transmission. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(5), 2538–2543.

Reed SC (1937) The inheritance and expression of fused, a new mutation in the house mouse. *Genetics*, **22**, 1–13.

Rutherford SL and Henikoff S (2003) Quantitative epigenetics. *Nature Genetics*, **33**, 6–8.

Ruvinsky AO and Agulnik A (1990) Genetic imprinting and the manifestation of the fused gene in the house mouse. *Developmental Genetics*, **11**, 263–269.

Sollars V, Lu X, Xiao L, Wang X, Garfinkel MD and Ruden DM (2003) Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nature Genetics*, **33**, 70–74.

Sutherland HGE, Kearns M, Morgan HD, Headley AP, Morris C, Martin DIK and Whitelaw E (2000) Reactivation of heritably silenced gene expression in mice. *Mamm Genome*, **11**, 347–355.

Weichman K and Chaillet JR (1997) Phenotypic variation in a genetically identical population of mice. *Molecular and Cellular Biology*, **17**, 5269–5274.

Wolff GL (1978) Influence of maternal phenotype on metabolic differentiation of agouti locus mutants in the mouse. *Genetics*, **88**, 529–539.

Zeng LI, Fagotto F, Zhang T, Hsu W, Vasicek TJ, Perry WL, Lee JJ, Tilghman SM, Gumbiner BM and Costantini F (1997) The mouse fused locus encodes axin, an inhibitor of the wnt signaling pathway that regulates embryonic axis formation. *Cell*, **90**, 191–192.

Short Specialist Review

Evolution of genomic imprinting in mammals

Hamish G. Spencer
University of Otago, Dunedin, New Zealand

1. Introduction

Genomic imprinting is the name of the form of non-Mendelian gene expression in which the two copies of a gene at a locus have different levels of expression. The archetypal case is that of insulin-like growth factor 2 (IGF-2) in which the maternal copy is silent in most fetal tissues, with only the paternally inherited copy being transcribed. Some 60 or so mammalian loci are currently known to be imprinted (Morison *et al*., 2001), but there is little consensus about the proportion of the genome subject to imprinting. By silencing (or at least downregulating) one copy of a gene, imprinting negates (or reduces) what is considered to be the major advantage of diploidy in mammals, namely, the ability to mask recessive deleterious mutations. Thus, the evolution of imprinting from an ancestral state of standard Mendelian (i.e., biallelic) expression appears paradoxical, apparently reducing an individual's fitness. Several hypotheses seeking to resolve this paradox have been proposed by a number of authors. Here we examine a number of the more plausible ideas, highlighting their various strengths and weaknesses. In brief, however, no single hypothesis appears able to explain all the observations.

2. Prevention of parthenogenesis and the ovarian time-bomb hypothesis

The oldest ideas note that if different loci are oppositely inactivated (i.e., the maternal copy is transcribed at one locus and the paternal copy at another), imprinting at essential loci would require both paternal and maternal contributions to the developing zygote, thereby preventing parthenogenesis. Indeed, recent experiments with mice show that if appropriate expression at the normally imprinted H19 and Igf2 loci occurs, at least some parthenogenetic embryos can survive to adulthood and reproduce (Kono *et al*., 2004). The observed absence of any parthenogenetic mammalian species consequently led some authors to argue that imprinting may have evolved for this purpose. Parthenogenesis is considered to be disadvantageous because it stops the genetic recombination that occurs as a

consequence of sexual reproduction. Parthenogenetic lineages are inclined to be evolutionary dead ends, lacking the ability to respond to novel selection pressures.

The main problem with this argument, however, is that it is "group-selectionist": the purported advantage of avoiding parthenogenesis accrues to the species, not to an individual. Indeed, an individual with parthenogenetic abilities might have a selective advantage, able to reproduce even in the absence of suitable mates. This selection for parthenogenesis at the level of the individual would subvert selection against it at the level of the species. In most cases of conflict between selection pressures in opposite directions at different levels, selection at the lower level – for instance, individuals rather than species – prevails. Thus, prevention of parthenogenesis for the good of the species is not considered a likely cause of the evolution of imprinting in mammals.

Nevertheless, an individual-level advantage for avoiding parthenogenesis has been suggested by Varmuza and Mann (1994). They argued that a haploid egg spontaneously developing in an ovary would amount to ovarian cancer, and imprinting may have evolved to prevent such a scenario. Inactivating the maternal copy of a growth-enhancing gene would defuse this "ovarian time bomb". Iwasa (1998) pointed out that the same protection is afforded by upregulating the maternal copy of a growth-inhibiting gene. Moreover, the level of expression in the diploid developing zygote can be maintained by concomitantly downregulating the paternal copy of this same gene, possibly to the point of silencing it. Thus, the ovarian time-bomb hypothesis predicts that growth-affecting genes active in the early stages of embryogenesis are likely candidates for imprinting and that growth enhancers should be maternally inactivated, whereas growth inhibitors should be paternally silenced. And indeed, this prediction is often met: the growth-enhancing Igf2 is maternally inactivated in fetal tissues in all mammalian species examined so far, and the growth-inhibiting Igf2r is paternally inactivated in mice and rats (but not in humans).

Mathematical modeling of the ovarian time-bomb hypothesis (Weisstein *et al*., 2002) implies that the verbal hypothesis is plausible: the selection pressures envisaged could lead to the evolution of imprinting. The modeling shows also that the selection pressure required to lead to the evolution of imprinting need not be very strong, contradicting the objection that ovarian cancer was too rare to be worth the cost of the loss of functional diploidy. Nevertheless, it is less clear why so many loci should be imprinted: surely, the imprinting of one or two critical loci would provide sufficient protection. The flip side of this problem is that the hypothesis at least offers a weak explanation for why not all growth-affecting genes important in early development (e.g., IGF1) are not imprinted. Finally, the hypothesis offers no explanation for genes that are not involved in early development.

3. The genetic-conflict hypothesis

Perhaps the best-known explanation for the evolution of imprinting is that invoking the different, conflicting genetic interests of mothers, fathers, and their offspring (Haig and Graham, 1991; Haig, 1992). A mammalian mother is equally related to all the offspring in a single (and subsequent) pregnancies, so her genetic contribution to the next generation is maximized by ensuring the survival of as many of these

offspring as possible. To at least the first level of approximation, these genetic interests are best served by equally dividing the nutrients and care she provides among these offspring. One way to accomplish this goal would be to turn off any growth-enhancing genes in her offspring, so she can control the transfer of her various resources. A father's genetic perspective is quite different, however, because most mammals have some degree of multiple paternity. A mammalian father has no assurance that all the offspring born to a female with which he has mated will be his. Consequently, his genetic success is greater if somehow his offspring obtain more maternal resources, maybe at the expense of any half-sibs or even the mother herself. Inactivating the paternal copy of a growth-inhibiting gene would achieve that end.

Thus, the genetic-conflict hypothesis makes the same predictions as the ovarian time-bomb hypothesis about the sorts of loci likely to be imprinted (i.e., growth-affecting genes important in fetal development) and the direction of imprinting (i.e., maternal inactivation of growth enhancers and paternal silencing of growth inhibitors). Mathematical modeling by various groups confirms the basic plausibility of the hypothesis (Spencer *et al*., 1999). Some modeling (Spencer *et al*., 1998) predicts that under certain circumstances, a locus can be polymorphic in imprinting status, with some individuals having two active copies of the genes and others just one. Indeed, two loci, the Wilm's tumor suppressor, WT1 (Jinno *et al*., 1994), and the serotonin-2A (5-HT_{2A}) receptor, HTR2A (Bunzel *et al*., 1998), appear to fulfill this prediction. Importantly, this prediction differs from that derived from modeling of the ovarian time-bomb hypothesis, so these observations lend important support to genetic conflict over the ovarian time bomb. The genetic-conflict hypothesis can also apply in arenas other than fetal development, for example in postnatal care, and so the range of loci likely to be imprinted by this mechanism is greater than under the ovarian time bomb. This prediction appears largely, but not completely, fulfilled (Tycko and Morison, 2002). Nevertheless, the genetic-conflict hypothesis is less able to explain why other growth-affecting genes such as Igf1 are not imprinted.

4. Differential selection on males and females

We have only a limited number of observations on imprinting at sex-linked loci. Indeed, the best known simply infer the presence of imprinting from observations of chromosomal abnormalities, especially X-chromosome monosomy in mice (*see* Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1). The parental origin of the single X in these XO mice has significant growth effects: if it is paternally derived, the mice are developmentally retarded (Jamieson *et al*., 1998), implying the presence of a paternally inactivated growth enhancer. Thus, the direction of imprinting at X-linked loci appears to be opposite to that predicted by both the genetic-conflict and ovarian time-bomb hypotheses. Observations like these led Iwasa and Pomiankowski (1999) to propose that selection for different phenotypes in males and females – especially different sizes – could lead to imprinting.

These authors noted that changing the expression level of a paternally derived gene on the X chromosome would affect only female offspring. In contrast, alteration of expression at a maternally derived locus subject to dosage compensation (*see* Article 15, **Human X chromosome inactivation**, Volume 1) would have greater effects on male offspring. Hence, greater male size, common in mammalian species, could be achieved by maternally inactivating an X-linked growth inhibitor and/or paternally silencing a growth enhancer. These predictions are the opposite of those made by the genetic-conflict hypothesis (Spencer *et al*., 2004). Moreover, the sorts of loci that might be subject to imprinting under Iwasa and Pomiankowski's hypothesis is considerably greater: any loci affecting characters for which optimum male and female phenotypes differ could be imprinted. The paucity of clear examples of imprinted X-linked genes, therefore, could be seen as evidence against this suggestion.

Spencer *et al*. (2004) argued that the above ideas could be extended to autosomal loci underlying characters for which being more similar to one parent than the other is advantageous. For example, given that male mammals usually disperse further than females, genes important in local adaptation in a heterogeneous habitat might be preferentially expressed from the better-adapted maternal copies and hence subject to paternal inactivation. There are, however, no current examples of imprinting that support these latest ideas.

In summary, several hypotheses have been proposed to explain the paradox of imprinting – the apparently disadvantageous functional haploidy at imprinted loci – but not one explains all the observations. Some suggestions not discussed here (e.g., better control of gene expression) have little support, either empirical or theoretical, but three hypotheses – ovarian time bomb, genetic conflict, and differential selection on males and females – appear to do far better in plausibly explaining many known observations.

Related articles

Article 15, **Human X chromosome inactivation**, Volume 1; Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1; Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1; Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1; Article 30, **Beckwith–Wiedemann syndrome**, Volume 1; Article 31, **Imprinting at the *GNAS* locus and endocrine disease**, Volume 1; Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1; Article 33, **Epigenetic reprogramming in germ cells and preimplantation embryos**, Volume 1; Article 36, **Variable expressivity and epigenetics**, Volume 1; Article 38, **Rapidly evolving imprinted loci**, Volume 1; Article 39, **Imprinting and behavior**, Volume 1; Article 41, **Initiation of X-chromosome inactivation**, Volume 1; Article 45, **Bioinformatics and the identification of imprinted genes in mammals**, Volume 1; Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1

Further reading

Haig D (2000) The kinship theory of genomic imprinting. *Annual Review of Ecology and Systematics*, **31**, 9–32.

Haig D and Trivers R (1995) The evolution of parental imprinting: a review of hypotheses. In *Genomic Imprinting: Causes and Consequences*, Ohlsson R, Hall K and Ritzen M (Eds.), Cambridge University Press: Cambridge, pp. 17–28.

Hurst LD (1997) Evolutionary theories of genomic imprinting. In *Genomic Imprinting: Frontiers in Molecular Biology*, Reik W and Surani A (Eds.), Oxford University Press: Oxford, pp. 211–237.

Hurst LD and McVean GT (1998) Do we really understand the evolution of genomic imprinting? *Current Opinion in Genetics & Development*, **8**, 701–708.

McDonald JF (1999) Genomic imprinting as a coopted evolutionary character. *Trends in Ecology & Evolution*, **14**, 359.

Moore T and Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics*, **7**, 45–49.

Ohlsson R, Paldi A and Graves JAM (2001) Did genomic imprinting and X chromosome inactivation arise from stochastic expression. *Trends in Genetics*, **17**, 136–141.

Spencer HG (2000) Population genetics and evolution of genomic imprinting. *Annual Review of Genetics*, **34**, 457–477.

References

Bunzel R, Blümcke I, Cichon S, Normann S, Schramm J, Propping P and Nöthen MN (1998) Polymorphic imprinting of the serotonin-2A (5-HT_{2A}) receptor gene in human adult brain. *Molecular Brain Research*, **59**, 90–92.

Haig D (1992) Genomic imprinting and the theory of parent-offspring conflict. *Seminars in Cell & Developmental Biology*, **3**, 153–160.

Haig D and Graham C (1991) Genomic imprinting and the strange case of insulin-like growth factor II receptor. *Cell*, **64**, 1045–1046.

Iwasa Y (1998) The conflict theory of genomic imprinting: how much can be explained? *Current Topics in Developmental Biology*, **40**, 255–293.

Iwasa Y and Pomiankowski A (1999) Sex specific X chromosome expression caused by genomic imprinting. *Journal of Theoretical Biology*, **197**, 487–495.

Jamieson RV, Tan S-S and Tam PPL (1998) Retarded postimplantation development of X0 mouse embryos: impact of the parental origin of the monosomic X chromosome. *Developmental Biology*, **201**, 13–25.

Jinno Y, Yun K, Nishiwaki K, Kubota T, Ogawa O, Reeve AE and Niikawa N (1994) Mosaic and polymorphic imprinting of the WT1 gene in humans. *Nature Genetics*, **6**, 305–309.

Kono T, Obata Y, Wu Q, Niwa K, Ono Y, Yamamoto Y, Park ES, Seo J-S and Ogawa H (2004) Birth of parthenogenetic mice that can develop to adulthood. *Nature*, **428**, 860–864.

Morison IM, Paton CJ and Cleverley SD (2001) The imprinted gene and parent-of-origin effect database. *Nucleic Acids Research*, **29**, 275–276, **http://www.otago.ac.nz/IGC**.

Spencer HG, Clark AG and Feldman MW (1999) Genetic conflicts and the evolutionary origin of genomic imprinting. *Trends in Ecology & Evolution*, **14**, 197–201.

Spencer HG, Feldman MW and Clark AG (1998) Genetic conflicts, multiple paternity and the evolution of genomic imprinting. *Genetics*, **148**, 893–904.

Spencer HG, Feldman MW, Clark AG and Weisstein AE (2004) The effect of genetic conflict on genomic imprinting and modification of expression at a sex-linked locus. *Genetics*, **166**, 565–579.

Tycko B and Morison IM (2002) Physiological functions of imprinted genes. *Journal of Cellular Physiology*, **192**, 245–258.

Varmuza S and Mann M (1994) Genomic imprinting – defusing the ovarian time bomb. *Trends in Genetics*, **10**, 118–123.

Weisstein AE, Feldman MW and Spencer HG (2002) An evolutionary genetic model of the ovarian time-bomb hypothesis for the evolution of imprinting. *Genetics*, **162**, 425–439.

Short Specialist Review

Rapidly evolving imprinted loci

Joomyeong Kim and Lisa Stubbs
Lawrence Livermore National Laboratory, Livermore, CA, USA

Most mammalian genes have evolved within the boundaries of functional constraints, which preserve their protein-coding capability and proper temporal and spatial expression patterns. Such constraints clearly have operated on most imprinted genes since most are highly conserved in structure and function (*see* Article 37, **Evolution of genomic imprinting in mammals**, Volume 1). However, some imprinted loci have evolved in very unusual ways, diverging significantly in protein-coding capabilities and gene expression patterns. One of the most dramatic cases includes a cluster of imprinted genes located in human chromosome 19q13.4 and the homologous region of proximal mouse chromosome 7 (Figure 1). Six imprinted genes are located within this domain, including paternally expressed genes, *Peg3, Usp29*, and *Zfp264*, and maternally expressed loci, *Zim1, Zim2*, and *Zim3* (Kim *et al*., 1997; Kim *et al*., 1999; Kim *et al*., 2000a,b; Kim *et al*., 2001; Kuroiwa *et al*., 1996). Three-way comparisons of sequence from human, mouse, and cow have revealed an unusual degree of evolutionary change in this domain (Kim *et al*., 2003; Kim *et al*., 2004). Even the genes that remain intact and are similarly imprinted in human and mouse appear to be evolving rapidly, showing significant differences in domain structure or in protein-coding sequence. The rapid pace of evolution that characterizes this gene cluster is unprecedented for imprinted domains, and may provide a new paradigm for mammalian gene evolution.

The most striking change that has occurred in this domain is the loss of protein-coding capacity in three of the six imprinted rodent loci, *Zim2*, *Zim3*, and *Zfp264*, which are intact genes in human and cow. All three genes have maintained transcriptional activity in rodents and are still clearly imprinted, suggesting adaptation to unknown functions. One of these genes, *Zim3*, produces an RNA species in mice and rats that serves as an antisense transcript gene for the neighboring gene, *Usp29*. Noncoding RNA genes and antisense transcripts are common features of imprinted domains (Sleutels *et al*., 2000; Ogawa and Lee, 2002), but rodent *Zim3* may be the first case where the origin of antisense transcript genes can be traced unequivocally. Another feature that is unique to the *Peg3*-containing domain is that several homologous human and mouse genes have evolved very different genomic structures. For example, in mouse and cow, *Peg3* and *Zim2* represent two separate genes with distinct sets of exons and independent promoters, but in human these two genes share seven small exons and are transcribed from a common promoter. *Zim2* and *Peg3* each possesses

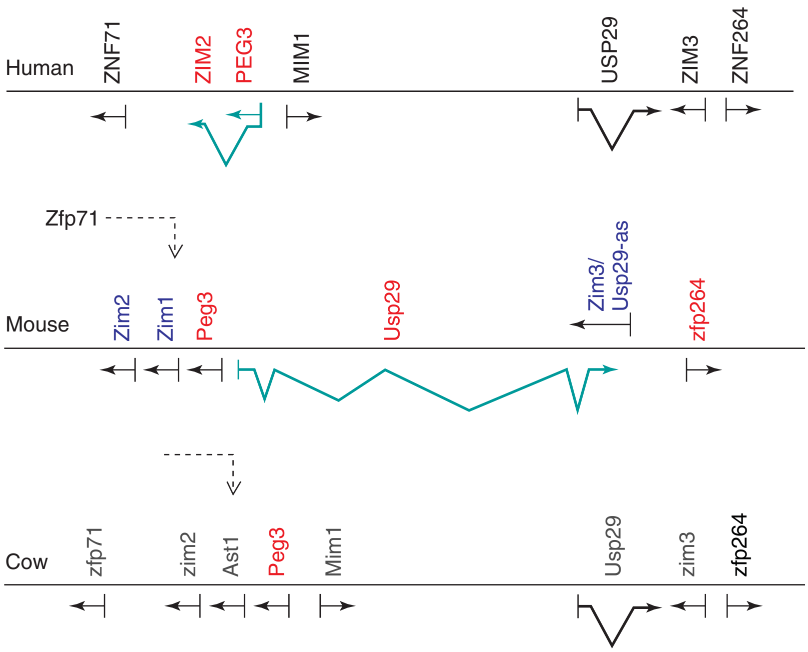


Figure 1 Comparative map of the Peg3-imprinted domain. The genomic organization of six imprinted genes in the Peg3 domain in three different mammals. Maternally expressed genes are marked in blue, whereas names of paternally expressed genes are shown in red. The genes with unknown imprinting status are marked in black. The imprinting status of cow *Zim2* and *Ast1* (Artiodactyls-specific transcript 1; Kim *et al*., 2004) are shown to be biallelic at adult testis, but the imprinting status at the other tissues is currently unknown. Lineage-specific changes have been observed in the downstream regions of *Peg3*. In human and cow, *ZNF71* is localized immediate downstream of *ZIM2*, whereas in mouse *Zim1*, a potential homolog or paralog of *ZNF71* has been moved to between *Peg3* and *Zim2*. A similar genomic insertion is also predicted to have happened in cow, the insertion of *Ast1* between *Peg3* and *Zim2*. These genomic rearrangements are marked as dotted arrows

the canonical structure of independent SCAN box–containing zinc-finger genes; together with the observation that the two loci are independently transcribed in mouse and cow, this fact has prompted the speculation that two originally separate genes merged to generate the transcription unit found in the primate lineage (Kim *et al*., 2004). Exon structure and promoter choice also differ significantly for human and mouse versions of the ubiquitin-specific protease gene, *Usp29*. In mouse, *Usp29* transcription starts from a shared bidirectional promoter at a site located ~150 bp from the start of *Peg3*, and the transcript is composed of seven exons distributed over more than 300 kb. In humans, however, transcripts arising from the same promoter region and containing sequences that correspond to the first two exons of mouse *Usp29* constitute parts of a novel untranslated human transcript, called *MIM1* (Mer-repeat containing imprinted transcript 1). The remaining five exons comprise the protein-coding transcript of the human gene, which is generated from an alternative downstream promoter. As a consequence, the expression patterns of human and mouse *Usp29* genes differ substantially; mouse

Usp29 is widely expressed with highest levels of transcription in the brain, whereas the human gene is testis-specific (Kim *et al*., 2000b). Preliminary data suggest that the exon structure of bovine *Usp29* is similar to that of the human gene. This suggests that the exon structure of mouse *Usp29* is a more neomorphic form than that of human and cow, and that *Mim1* and *Usp29* transcription units might have been fused to form a single large gene in the rodent lineage. Therefore, although in overall arrangement and content of genes, the *Peg3*-containing domain is well conserved among different mammals, the different genomic structures, promoter usage, and the shifting relationships between neighboring genes make this region unique among the known imprinted domains.

Why is this domain so unusual? The unique evolutionary patterns observed for genes within the *Peg3*-containing imprinted domain may be related to underlying features of the resident genes. The first feature is the relatively recent origin of most of the genes. Most of the known imprinted genes are ancient and highly conserved; for example, clear orthologs of *Igf2, Igf2r, and Peg1* are found in nonmammalian vertebrates such as fish and chickens (Nolan *et al*., 2001; Yang, S. K. and Chung, J. H., personal communication). By contrast, all genes in the *Peg3*-containing domain except *Usp29* are Krüppel-type zinc-finger (ZNF) genes, members of one of the largest and most evolutionarily dynamic mammalian gene families. In *Peg3* and *Zim2*, the DNA-binding ZNF region is attached to a SCAN effecter domain; *Zim1, Zim3,* and *Zfp264* belong to a second subclass containing a *Kruppel*-associated box, or KRAB effecter motif. SCAN- and KRAB-ZNFs are found only in amphibian, avian, and mammalian genomes, indicating the recent advent of these gene families. KRAB-ZNF genes, in particular, have expanded dramatically in mammalian lineages through repeated rounds of gene duplication (Hamilton *et al*., 2003; Huntley *et al*., 2004). Although the imprinted *Zim1/ZNF71*, *Zim2*, and *Zim3* genes encode relatively unique proteins, distantly related ZNF loci surround the imprinted domain in both human and mouse and these may provide some level of functional redundancy. Such redundancy may have reduced the otherwise potentially devastating effects of gene loss, resulting in the rapid loss of protein-coding capability of three mouse imprinted genes, *Zim2*, *Zim3*, and *Zfp264*.

Why have the imprinted genes maintained their transcriptional activity and imprinted regulation over the long evolutionary period since their coding sequence was inactivated? This may be related to the regulatory mechanisms that underlie genomic imprinting. A relatively large fraction of the transcripts produced within imprinted domains are expressed without the protein-coding capability, and many of these serve as antisense transcripts that are oppositely imprinted relative to the neighboring protein-coding genes (Sleutels *et al*., 2000; Ogawa and Lee, 2002). Some of these transcripts are unusually large, up to several hundred kilobases in length, as illustrated by the *Air* transcript (Lyle *et al*., 2000). In fact, large genes, such as *Usp29*, *Kcnq1*, *Snurp/Snurf*, and *Gnas,* are frequently found in imprinted domains. The prevalence of large transcription units, whether coding or not, may provide a hint that transcription per se may be required for establishing and/or maintaining specific types of chromatin structure in imprinted domains. This is consistent with the current view that many untranslated mammalian transcripts may have regulatory roles for chromatin structure and neighboring genes

(Kiyosawa *et al*., 2003; Tufarelli *et al*., 2003). The potential role of transcription itself for imprinted regulation may possibly help explain why the imprinted pseudogenes in the *Peg3*-containing domain have maintained transcriptional activity without protein-coding capability. The prevalence of long-distance transcription in imprinted domains might also have been a trigger for unusual gene fusion events that have been observed in the *Peg3*-containing region, *PEG3/ZIM2* in human and possibly *Usp29/Mim1* in mouse (Figure 1).

Evolutionary changes in sequence, imprinting status, and coding capacity have been demonstrated for a handful of other imprinted genes (Chai *et al*., 2003; Hitchins *et al*., 2001). However, the high level of evolutionary plasticity observed in the *Peg3* region has not been seen in the other known imprinted domains. In contrast to the genes in the human chromosome 19–imprinted domain, the other well-known imprinted loci were established, and embedded in well conserved pathways of development and behavior, long before the advent of imprinted regulation arose in mammals. In contrast, we predict that *Peg3* and its imprinted neighbors arose around the time that genomic imprinting began as a mechanism for gene regulation. This recent evolutionary history is likely to be one reason why the imprinted genes in the *Peg3*-containing domain appear to have evolved so rapidly and to have diverged in such a lineage-specific fashion. Dissecting the activities and developmental roles of the genes in this domain will be challenging due to the differences between humans and mice, since the mouse is the experimental system of choice for mammalian gene function and imprinting studies. However, the predicted roles of *Zim1, Zim2, Zim3, Znf264,* and possibly *Peg3* as transcriptional regulators suggest the impact of genomic imprinting on a wider network of mammalian genes. Studying the impact of their conservation and loss will provide a unique window to the evolution and functions of mammalian imprinted domains.

References

Chai JH, Locke DP, Greally JM, Knoll JH, Ohta T, Dunai J, Yavor A, Eichler EE and Nicholls RD (2003) Identification of four highly conserved genes between breakpoint hotspots BP1 and BP2 of the Prader-Willi/Angelman syndromes deletion region that have undergone evolutionary transposition mediated by flanking duplicons. *American Journal of Human Genetics*, **73**, 898–925.

Hamilton AT, Huntley S, Kim J, Branscomb E and Stubbs L (2003) Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. *Cold Spring Harbor Symposium on Quantitative Biology*, **68**, 131–140.

Hitchins MP, Monk D, Bell GM, Ali Z, Preece MA, Stanier P and Moore GE (2001) Maternal repression of the human GRB10 gene in the developing central nervous system; evaluation of the role for GRB10 in Silver-Russell syndrome. *European Journal of Human Genetics*, **9**, 82–90.

Huntley S, Hamiton AT, Kim J, Branscomb E and Stubbs L (2004) Tandem gene family expansion and genomic diversity. In *Comparative Genomics: A Guide to the Analysis of eukaryotic Genomes*, Adams M (Ed.), Humana Press.

Kim J, Ashworth L, Branscomb E and Stubbs L (1997) The human homologue of a mouse imprinted gene, *Peg3*, maps to a zinc finger gene-rich region of human chromosome 19q13.4. *Genome Research*, **7**, 532–540.

Kim J, Bergmann A and Stubbs L (2000a) Exon sharing of a novel human zinc-finger gene, *ZIM2*, and paternally expressed gene 3 (*PEG3*). *Genomics*, **64**, 114–118.

Kim J, Noskov V, Lu X, Bergmann A, Ren X, Warth T, Richardson P, Kouprina N and Stubbs L (2000b) Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13.4. *Genome Research*, **10**, 1138–1147.

Kim J, Bergmann A, Lucas S, Stone R and Stubbs L (2004) Lineage-specific imprinting and evolution of the zinc finger gene ZIM2. *Genomics*, in press.

Kim J, Bergmann A, Wehri E, Lu X and Stubbs L (2001) Imprinting and evolution of two Kruppel-type zinc-finger genes, Zim3 and ZNF264, located in the PEG3/USP29-imprinted domain. *Genomics*, **77**, 91–98.

Kim J, Kollhoff A, Bergmann A and Stubbs L (2003) Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, *Peg3*. *Human Molecular Genetics*, **12**, 233–245.

Kim J, Lu X and Stubbs L (1999) Zim1, a maternally expressed mouse Kruppel-type zinc-finger gene located in proximal chromosome 7. *Human Molecular Genetics*, **8**, 847–854.

Kiyosawa H, Yamanaka I, Osato N, Kondo S and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*, **13**, 1324–1334.

Kuroiwa Y, Kaneko-Ishino T, Kagitani F, Kohda T, Li L-L, Tada M, Suzuki R, Yokoyama M, Shiroishi T, Wakana S, *et al*. (1996) *Peg3* imprinted gene on proximal chromosome 7 encodes for a zinc finger protein. *Nature Genetics*, **12**, 186–190.

Lyle R, Watanabe D, te Vruchte D, Lerchner W, Smrzka OW, Wutz A, Schageman J, Hahner L, Davies C and Barlow DP (2000) The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature Genetics*, **25**, 19–21.

Nolan CM, Killian JK, Petitte JN and Jirtle RL (2001) Imprint status of M6P/IGF2 R and IGF2 in chickens. *Development Genes and Evolution*, **211**, 179–183.

Ogawa Y and Lee JT (2002) Antisense regulation in X inactivation and autosomal imprinting. *Cytogenetic and Genome Research*, **99**, 59–65.

Sleutels F, Barlow DP and Lyle R (2000) The uniqueness of the imprinting mechanism. *Current Opinion in Genetics & Development*, **10**, 229–233.

Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG and Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genetics*, **34**, 157–165.

Short Specialist Review

Imprinting and behavior

James P. Curley and Eric B. Keverne
University of Cambridge, Madingley, UK

1. Evidence for the role of imprinted genes in behavioral phenotypes

The crossbreeding of animal species with one another can often produce hybrids with parent-of-origin-specific changes in behavioral phenotypes. For instance, the offspring of a male zebra and a female donkey are noted for their stubbornness in comparison with the reciprocal cross (Gray 1972; Walton & Hammond 1938). For many years, it was unclear why these genetically equivalent individuals behaved in consistently different ways dependent upon parent of origin. Following the discovery of genomic imprinting, it is now acknowledged that these offspring differ in their expression of imprinted genes and that differences in the maternal and paternal alleles of reciprocal crosses at these loci could lead to the observed behavioral differences.

Hybrid F1 mice produced from mating a male from an inbred strain A and a female from inbred strain B can be compared with those produced from the reciprocal cross. If both sets of offspring are transferred as embryos into a different strain C, these offspring are genetically equivalent in every respect with the exception of their imprinted genes. Interestingly, it was found that reciprocal cross F1 offspring avoid female urine from their maternal strain and investigate more the urine from an unrelated strain D in a choice test (Isles *et al*., 2001). However, no preference was observed when these mice were given the opportunity to investigate female urine from either a neutral strain D or their genetic paternal strain. This finding has been replicated with at least two separate parental sets of inbred strains, demonstrating that the effect of imprinted genes on mate choice is likely to be universal and not an epiphenomenon of one particular inbred strain (Isles *et al*., 2002). Moreover, since the F1 offspring were embryo transferred to foster mothers of a separate strain, this eliminated the possibility that the avoidance of maternal strain odors was dependent upon any learning during pup development.

The importance of genomic imprinting in brain and behavioral development was further illustrated by identifying where in the brain cells expressing imprinted genes preferentially develop (Allen *et al*., 1995; Keverne *et al*., 1996a). Mice with two complete sets of maternal or paternal chromosomes were only viable until day 10 of gestation (*see* Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental**

genomes, Volume 1). However, this lethal phenotype could be rescued by producing chimeric mice that possess a mixture of genetically normal cells together with cells containing two copies of maternal chromosomes (parthenogenomes, Pg) or two copies of paternal chromosomes (androgenomes, Ag). Those mice that received a contribution of paternally disomic cells had smaller brains, but with localized distribution of Ag cells to the ventral limbic forebrain. Conversely, those mice with a higher contribution of maternally disomic cells had larger brains, with localized distribution of Pg cells in cortical structures and the striatum but with virtual absence of these cells in the hypothalamus. Interestingly, male Pg mice were more aggressive than control mice, being quicker to attack an opponent male (Allen *et al*., 1995). However, no direct link between specific imprinted genes and this behavior has yet been discovered.

More recent research has looked for specific imprinted genes and investigated their expression patterns in different tissues. Currently, nearly a hundred imprinted genes have been identified, many of which are expressed in the brain and have immediate significance for behavioral functioning. One obvious example is the maternally specific expression of the serotonin receptor 2A gene (*Htr2*) in both humans and mice (Kato *et al*., 1998; Kato *et al*., 1996). Three other genes, *Gnas* (a G protein involved in cell signaling), *Nesp* (a splice variant of *Gnas*), and *Nnat* (a brain specific transmembrane protein), that are all expressed in the brain have been put forward as candidate genes for the reciprocal activity phenotype observed in mice with uniparental disomies (UPDs, *see* Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1) of chromosome 2 (Cattanach & Kirk 1985; Isles & Wilkinson 2000). In this instance, offspring with a partial maternal or paternal UPD are hypokinetic or hyperkinetic respectively.

In humans, the importance of imprinted gene expression in the brain has been recognized from behavioral disorders (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1). Of major interest is a cluster of maternally and paternally imprinted loci on human chromosome 15q11-q13, which is homologous to a region on mouse chromosome 7 (Yang *et al*., 1998). Paternal and maternal deletions of this chromosomal region are associated with Prader–Willi Syndrome (symptoms include mild mental retardation and hyperphagia) and Angelman Syndrome (symptoms include mental retardation, absent speech and inappropriate tongue movements, and laughter) respectively (Nicholls & Knepper 2001). Several candidate-imprinted genes have been identified in this area including the paternally expressed *ZNF127*, *NECDIN*, and *IPW* and the maternally expressed *UBE3A*. These chromosome 15 imprinting disorders have been associated with malfunctioning chromosomal regions, but a mutation in the imprinting mechanism could also affect neighboring nonimprinted genes such as those coding for subunits of the $GABA_A$-receptor (DeLorey *et al*., 1998). Parent-of-origin effects on the inheritance of several other behavioral disorders including autism, schizophrenia, and bipolar affective disorder have also been reported (Isles & Wilkinson 2000), although candidate-imprinted genes have yet to be identified.

Most studies of imprinted genes are phenotype led and provide for the possibility of a role for imprinted genes in the regulation of behavior, but few link specific genes with specific phenotypes. However, recent studies using modern molecular genetic techniques to target specific imprinted genes have resulted in the

development of behavioral phenotypes. For instance, mice with a targeted mutation of the paternally expressed *Grf1* (involved in Ras signaling) have impaired long-term emotional memories (Brambilla *et al.*, 1997). Learning and memory deficits, in addition to motor dysfunctions and inducible seizures have been identified in mice lacking the putatively paternally expressed *Gabrb3* gene (codes for a $GABA_A$-receptor subunit) and the maternally expressed *Ube3a* gene (codes for E6-AP ubiquitin protein ligase).

In addition to the effects of these genes on cognitive behavior, the role of paternally expressed genes on maternal and infant behavior has been studied using mice carrying a targeted mutation of either the *Mest/Peg1* or the *Peg3* gene (Curley *et al.*, 2004; Lefebvre *et al.*, 1998; Li *et al.*, 1999). The unique inheritance pattern of paternally expressed genes enables the effect of the mutation on maternal and infant behavior to be studied independently. Females carrying the mutation give birth to wild-type offspring when mated with a wild-type male, whereas wild-type females give birth to offspring carrying the mutation when mated with a homozygous mutant male. When female mice carry a mutation in one of these two paternally expressed genes, they are impaired in a whole suite of maternally related functions and their offspring are growth retarded during both the prenatal and postnatal periods (Keverne 2001). They are slower to retrieve pups, build nests, and crouch over pups, and their resting body temperature is lower than that of the wild-type females, making it difficult to keep the pups warm. During pregnancy, mutant females consume less food than wild-type controls, and hence their wild-type embryos are weight retarded. After birth, mutant females are impaired in their letdown of milk from mammary glands, resulting in their wild-type offspring failing to suckle adequate milk and thereby suffering further growth retardation, leading to delayed puberty. In addition to these behavioral deficits, females carrying the *Mest/Peg1* mutation also do not show placentophagia. Mechanistically, the decreased maternal behavior associated with females carrying the *Peg3* mutation has been associated with a significant decrease in oxytocin positive neurons in the paraventricular nucleus of the hypothalamus immediately postpartum. Female mammals are reliant on a surge of oxytocin following birth to activate maternal behavior as well as milk letdown, and it appears that female mutant mice fail in this respect.

Genomic imprinting appears to have evolved approximately 150 million years ago in the ancestor of eutherian (placental) and marsupial mammals (John & Surani 2000). One favored hypothesis for its evolution is the conflict theory (Burt & Trivers 1998; Haig & Graham 1991), which argues that in promiscuous species paternally expressed genes promote offspring growth, even at the expense of the mother's future fitness, while maternally expressed genes counteract this by suppressing offspring growth. Most of the early empirical work supported this hypothesis; imprinted genes were found to be expressed in the placenta – the battleground for this parental tug-of-war – and maternal and paternal alleles were generally growth suppressing and promoting respectively. However, the expression of imprinted genes in the brain and their involvement in behavioral phenotypes cannot be intuitively explained in terms of parental conflict. Indeed, instead of being driven by conflict, the expression of paternally expressed genes in the maternal and

infant brain may have evolved through coadaptation for the same genes *Mest/Peg1* and *Peg3* that ensure that infant growth and survival are also essential for good nurturing behavior (Curley *et al*., 2004; Keverne 2001). Overall, imprinted genes expressed in the brain appear to be important in the regulation of brain development and, interestingly, may have been crucial to the remodeling of brain structures during mammalian evolution (Allen *et al*., 1995; Keverne 2001; Keverne *et al*., 1996b).

Related articles

Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1; Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1; Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1; Article 37, **Evolution of genomic imprinting in mammals**, Volume 1; Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1

References

Allen ND, Logan K, Lally G, Drage DJ, Norris ML and Keverne EB (1995) Distribution of parthenogenetic cells in the mouse brain and their influence on brain development and behavior. *Proceedings of the National Academy of Sciences USA*, **92**, 10782–10786.

Brambilla R, Gnesutta N, Minichiello L, White G, Roylance A, Herron C, Ramsey M, Wolfer D, Cestari V, Rossi-Arnaud C, *et al*. (1997) A role for the Ras signalling pathway in synaptic transmission and long-term memory. *Nature*, **390**, 281–286.

Burt A and Trivers R (1998) Genetic conflicts in genomic imprinting. *Proceedings of the Royal Society Series B*, **265**, 2393–2397.

Cattanach BM and Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. *Nature*, **315**, 496–498.

Curley JP, Barton SC, Surani MA and Keverne EB (2004) Co-adaptation in mother and infant regulated by a paternally expressed imprinted gene. *Proceedings of the Royal Society Series B*, **271**, 1303–1309.

DeLorey T, Handforth A, Anagnostaras S, Homanics G, Minassian B, Asatourian A, Fanselow M, Delgado-Escueta A, Ellison G and Olsen R (1998) Mice lacking the beta3 subunit of the GABAA receptor have epilepsy phenotype and many of the behavioral characteristics of Angelman syndrome. *Journal of Neuroscience*, **18**, 8505–8514.

Gray AP (1972) *Mammalian Hybrids*, Commonwealth Agricultural Bureaux: Slough.

Haig D and Graham C (1991) Genomic imprinting and the strange case of the insulin-like growth factor II receptor. *Cell*, **64**, 1045–1046.

Isles AR, Baum MJ, Ma D, Keverne EB and Allen ND (2001) Urinary odour preferences in mice. *Nature*, **409**, 783–784.

Isles AR, Baum MJ, Ma D, Szeto A, Keverne EB and Allen ND (2002) A possible role for imprinted genes in inbreeding avoidance and dispersal from the natal area in mice. *Proceedings of the Royal Society Series B*, **269**, 665–670.

Isles AR and Wilkinson LS (2000) Imprinted genes, cognition and behaviour. *Trends in Cognitive Sciences*, **4**, 309–318.

John RM and Surani MA (2000) Genomic imprinting, mammalian evolution, and the mystery of egg-laying mammals. *Cell*, **101**, 585–588.

Kato M, Ikawa Y, Hayashizaki Y and Shibata H (1998) Paternal imprinting of mouse serotonin receptor 2A gene *Htr2* in embryonic eye: a conserved imprinting regulation on the RB/Rb locus. *Genomics*, **47**, 146–148.

Kato M, Shimizu T, Nagayoshi M, Kaneko A, Sasaki M and Ikawa Y (1996) Genomic imprinting of the human serotonin-receptor (HTR2) gene involved in the development of retinoblastoma. *American Journal of Human Genetics*, **59**, 1084–1090.

Keverne EB (2001) Genomic imprinting, maternal care, and brain evolution. *Hormones and Behavior*, **40**, 146–155.

Keverne EB, Fundele R, Narasimha M, Barton SC and Surani MA (1996a) Genomic imprinting and the differential roles of parental genomes in brain development. *Developmental Brain Research*, **92**, 91–100.

Keverne EB, Martel FL and Nevison CM (1996b) Primate brain evolution: genetic and functional considerations. *Proceedings of the Royal Society Series B*, **262**, 689–696.

Lefebvre L, Viville S, Barton SC, Ishino F, Keverne EB and Surani MA (1998) Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene *Mest*. *Nature Genetics*, **20**, 163–169.

Li LL, Keverne EB, Aparicio SA, Ishino F, Barton SC and Surani MA (1999) Regulation of maternal behavior and offspring growth by paternally expressed *Peg3*. *Science*, **284**, 330–333.

Nicholls R and Knepper J (2001) Genome organisation, function, and imprinting in Prader-Willi and Angelman syndromes. *Annual Review of Genomics Human Genetics*, **2**, 153–175.

Walton A and Hammond J (1938) The maternal effects on growth and conformation in Shire horse-Shetland pony crosses. *Proceedings of the Royal Society Series B*, **125**, 311–335.

Yang T, Adamson T, Resnick J, Leff S, Wevrick R, Francke U, Jenkins N, Copeland N and Brannan C (1998) A mouse model for Prader-Willi syndrome imprinting-centre mutations. *Nature Genetics*, **19**, 25–31.

Short Specialist Review

Spreading of X-chromosome inactivation

Jason O. Brant and Thomas P. Yang
University of Florida College of Medicine, Gainesville, FL, USA

During early female mammalian embryogenesis, one of the two X chromosomes is randomly inactivated in each cell of the embryo (*see* Article 28, **Imprinting and epigenetics in mouse models and embryogenesis: understanding the requirement for both parental genomes**, Volume 1 and Article 41, **Initiation of X-chromosome inactivation**, Volume 1). The stable inactivation of genes on one of the two X chromosomes in females functionally equalizes the apparent dosage imbalance of X-linked genes between males and females (Lyon, 1961). This process of X-chromosome inactivation (XCI) is believed to occur in three steps: initiation (Lyon, 1961; Russell, 1963), spreading (Russell, 1963), and maintenance (Barr and Carr, 1962). Initiation of XCI is believed to occur in the X inactivation center (XIC) and involve the *XIST* (inactive X specific transcript) gene, which is located within the XIC (in Xq13 in humans). The *XIST* gene encodes a 17-kb noncoding transcript that is expressed exclusively from the inactive X chromosome (Xi) (Brown *et al*., 1992), localized to the nucleus, accumulates at the Xi (Brown *et al*., 1992), and is required for XCI (Penny *et al*., 1996). Once XCI is initiated, the inactivation process is believed to spread bidirectionally in cis from the XIC along the length of the X chromosome. In all subsequent somatic cell divisions, the inactive state of the same Xi is maintained in each daughter cell, ensuring that the pattern of XCI is heritably and stably maintained throughout the life of the organism. However, XCI does not occur uniformly along the entire length of the X chromosome because numerous domains and loci have been shown to partially or completely escape inactivation (Schneider-Gadicke *et al*., 1989; Carrel and Willard, 1999; Disteche, 1999; Tsuchiya *et al*., 2004).

The molecular mechanism by which inactivation spreads bidirectionally from the XIC along the length of the X chromosome, yet appears to "skip" over domains that escape inactivation, remains perhaps the least understood aspect of X-chromosome inactivation. The fact that the spreading of inactivation occurs during a narrow interval in early female development among a relatively small number of cells has hampered studies to examine the mechanisms involved in this phase of XCI.

One of the early observations of XCI in mice was that in X;autosome translocations, inactivation of genes on the Xi appeared capable of spreading into the adjacent autosomal chromosome (Russell, 1963). Using coat-color variegation as

a marker of autosomal gene silencing, it was shown that these translocations could result in variable and discontinuous spreading of inactivation into autosomal chromatin. The extent of spreading of inactivation into the adjacent autosomal regions was variable, depending upon the location of the translocation breakpoints, including the breakpoint on the X. This apparent spreading of gene silencing into autosomal regions from the Xi suggested that spreading of inactivation along the X chromosome might be an integral feature of the process of XCI.

Similar indications of spreading of silencing into adjacent autosomal regions have also been observed on the human X chromosome in human X;autosome translocations (Canun *et al*., 1998; White *et al*., 1998; Sharp *et al*., 2001; Sharp *et al*., 2002). Patients with X;autosome translocations can present with variable phenotypic severity, suggesting variable extent of spreading of inactivation into the translocated autosomal regions, which is dependent on the location of the translocation breakpoints. This has included patients with unbalanced X;autosome translocations who showed phenotypes less severe than the corresponding autosomal trisomy. For example, a high-resolution molecular analysis of the spread of inactivation into autosomal DNA was performed in a patient with an X;4q translocation (White *et al*., 1998). 4q duplications are usually associated with dysmorphic features and severe growth and mental retardation. However, this patient presented with a normal phenotype, suggesting spread of inactivation into the translocated 4q chromosome. RT-PCR was performed for 20 transcribed sequences spanning this 4q translocation; results showed that gene silencing had spread into the translocated autosomal chromatin, though 30% of the sequences examined escaped inactivation. These data, along with similar observations of the spreading of silencing in other X;autosome translocations (Canun *et al*., 1998; White *et al*., 1998; Sharp *et al*., 2002), further suggested that autosomal DNA might lack X chromosome–specific characteristics and/or sequences that allow efficient response to the signal for XCI and/or the maintenance of inactivation. The fact that X inactivation is capable of spreading into autosomal DNA in X;autosome translocations in a discontinuous and variable fashion suggested reduced efficiency of spreading into autosomal chromatin.

The apparent ability of XCI to spread along the length of the X chromosome suggested there must be something unique about the X chromosome that promotes an efficient response to inactivation and facilitates the efficient spreading of inactivation along the length of the chromosome. The Riggs "way station" model proposed the presence of DNA elements along the X chromosome that function as way stations or "boosters" to promote or enhance the spreading of inactivation (Riggs *et al*., 1985). This model was later expanded upon by suggesting that these way stations were somehow uniquely arranged on the X chromosome in a manner that facilitates the efficient spreading of inactivation along the length of the X chromosome (Riggs, 1990).

The identity of the proposed way station elements on the X chromosome has proved elusive. Lyon (1998) proposed long interspersed repetitive elements (LINE-1; L1) elements as potential candidates for the way station elements hypothesized in the original Riggs model. Evidence in support of a role for L1, and perhaps other repetitive elements, in spreading of XCI has remained circumstantial, but increasingly compelling. For L1 elements to function as way stations to

facilitate spreading of XCI, the X chromosome would be expected to be enriched in these elements as compared to autosomal DNA, and regions of autosomal DNA into which XCI more readily spreads in X;autosome translocations should also be enriched in L1 elements. Consistent with these expectations, Lyon showed a strong correlation between the spread of inactivation into autosomal regions and the density of L1 elements in these regions by analyzing previously characterized X;autosome translocations (Lyon, 1998). Additionally, when an *Xist* transgene was ectopically expressed on mouse chromosome 12 (Lee and Jaenisch, 1997), the spread of silencing was observed along the chromosome, but it was noted that the distal portion of the transgenic chromosome 12 lacked histone H4 hypoacetylation (histone H4 hypoacetylation is commonly associated with transcriptionally silent chromatin), suggesting that genes in this region may have escaped inactivation (Lee and Jaenisch, 1997). Fluorescence in situ hybridization (FISH) mapping of LINE elements on mouse chromosome 12 indicated a distinct lack of these repetitive elements in this distal region (Boyle *et al*., 1990), which Lyon proposed may explain the lack of the spread of silencing into this region. This again suggested that silencing spreads more efficiently in regions that are enriched in L1 elements rather than in regions where these elements are underrepresented (Lyon, 1998).

More recently, Bailey *et al*. (2000) investigated the distribution of L1 elements along the length of the entire X chromosome and compared this pattern to L1 distribution in human autosomes. Their findings showed that L1 elements are indeed enriched on the X chromosome by nearly twofold as compared to autosomal DNA, as would be expected according to Lyon's proposed role for L1 elements in spreading of XCI. Moreover, there was a nonrandom distribution of L1 elements on the X chromosome, with enrichment in the region of the XIC as Riggs had postulated earlier (Riggs, 1990). Furthermore, a significant decrease in the number of L1 elements was observed for regions of the Xi that escape inactivation, consistent with Lyon's proposed role for L1 elements in the spreading of XCI.

A detailed examination of a domain that escapes X inactivation in humans and the syntenic region in the mouse provided evidence, suggesting that other interspersed repetitive elements may also play a role in spreading of XCI (Tsuchiya *et al*., 2004). Sequence analysis of human Xp11.2 containing a domain of multiple genes that escape XCI, as well as the syntenic region of the mouse X chromosome, only showed a correlation between escape from XCI and the density of L1 repeats in the region of the *SMCX/Smcx* gene. No correlation between L1 repeat density and inactivation was noted for the remainder of the domain in humans where other genes also escape inactivation. However, a correlation was noted between long terminal repeat (LTR) density and the transcriptional status of genes in the domain that escape inactivation. The entire Xp11.2 domain that escapes inactivation in humans was noted to be reduced in LTR density, as compared to the whole X chromosome. Furthermore, in the syntenic region of the mouse, only the region containing the *Smcx* gene (which is the only gene in this region that escapes inactivation) was reduced in LTR density. These findings are consistent with a correlation between LTR density and the ability of genes to escape inactivation, and suggest that spreading of XCI may also be facilitated by LTR elements in addition to L1 elements (Tsuchiya *et al*., 2004).

Another element potentially involved in the mechanism for spreading of XCI is the *XIST* gene and its RNA product. The discovery and characterization of this unique transcript led to its emergence as a candidate to mediate the spread of X inactivation via association with the postulated way stations. Introduction of an inducible *Xist* transgene into an autosomal chromosome in male mouse embryonic stem (ES) cells demonstrated that expression of *Xist* RNA was required in cis for the spread of inactivation to occur in the flanking autosomal region (Wutz and Jaenisch, 2000). Although the expression of *XIST* RNA seems to be the first observable event leading to the spreading of inactivation, the mechanism by which the spread of inactivation occurs only in cis, how silencing is induced, and the identity of proteins or protein complexes that associate with *XIST* RNA are currently unknown. Although the precise role of *XIST* RNA in the spreading of XCI remains unclear, models of how *XIST* RNA may mediate spreading have been suggested on the basis of current knowledge of XCI and the Xi. Wutz *et al*. (2002) have suggested that there may be multiple low-affinity-binding motifs present on *Xist* RNA that could bind sites along the Xi in a cooperative manner, adding to stability of the chromatin-bound *Xist* RNA and leading to subsequent binding and accumulation of *Xist* RNA on adjacent chromatin sites, thereby facilitating the spreading of XCI. Previous studies have shown that *Xist* RNA that does not localize to chromatin is unstable and quickly degraded (Wutz *et al*., 2002), which may explain why the spreading of *Xist* RNA occurs only in cis. Lyon has proposed a role for *XIST* RNA in spreading of XCI that involves its interaction with L1 elements and subsequent inactivation via repeat-induced silencing (Lyon, 1998).

Repeat-induced silencing via an RNAi-mediated process has been postulated as a potential mechanism involved in XCI. Nascent L1, or other repetitive sequence transcripts, could induce silencing through the RNAi pathway (Hansen, 2003), similar to the heterochromatic silencing demonstrated in fission yeast (Verdel *et al*., 2004). In fission yeast, double-stranded (ds) RNA generated from transcription of centromeric repeats is processed by the RNAi machinery (Verdel *et al*., 2004). The processed siRNA is targeted to the site of transcription through sequence-specific interactions between the siRNA and chromosomal DNA, leading to recruitment of chromatin-modifying proteins such as SWI6 (Verdel *et al*., 2004). SWI6, the yeast ortholog of metazoan heterochromatin binding protein-1 (HP1), appears to facilitate the spread of heterochromatin formation in yeast (Schramke and Allshire, 2003). SWI6 binds to histone H3 methylated at lysine 9 (H3meK9), which leads to the formation of a silent chromatin nucleation center, which can then facilitate the spreading of silent chromatin (Schramke and Allshire, 2003). It is conceivable that similar mechanisms of HP1-mediated spreading of heterochromatin could also contribute to the spreading of silencing in XCI. Chadwick *et al*. have also drawn parallels between heterochromatin formation in yeast and mammalian X-chromosome silencing, in which HP1 recognizes methylated H3K9, and acts to promote the formation of heterochromatin and spreading of X inactivation (Chadwick and Willard, 2003).

Another possible mechanism that could facilitate spreading of XCI has recently been demonstrated for the autosomal mouse *Dntt* gene (Su *et al*., 2004). Analysis of histone modifications in the region surrounding the *Dntt* promoter in immature thymocytes showed that local changes in histone modifications at the promoter

could subsequently spread. Upon differentiation of immature thymocytes, where the *Dntt* gene undergoes functional silencing, deacetylation of H3K9 was observed near the *Dntt* promoter, which then spread bidirectionally throughout the locus. These results suggest that certain histone modification patterns associated with silent chromatin are capable of spreading bidirectionally from a nucleation site, a process that could conceivably be involved in the spreading of histone modification patterns associated with silent chromatin on the Xi following initiation of XCI at the inactivation center. Additionally, methylation of H3K27 and assembly of polycomb group (PcG) protein complexes on the Xi also appear to be involved in the process of initiation and establishment of inactivation (Plath *et al*., 2004).

The ability of XCI to spread along the length of the X chromosome, yet apparently skip over regions that escape inactivation, has proved problematic in regard to understanding the mechanism of spreading. As many as 15–20% of genes on the human Xi may escape inactivation to some degree (Carrel *et al*., 1999), and these genes that escape inactivation tend to localize in clusters and domains (Carrel *et al*., 1999). Thus, how does the signal that propagates spreading of XCI act discontinuously by skipping over domains as large as 235 kb, and then continue the spreading of inactivation downstream of these domains? This question may be explained by analysis of the X-linked mouse *Smcx* gene (Lingenfelter *et al*., 1998). Allele-specific expression patterns of the *Smcx* gene were examined at various time points throughout female development using RT-PCR. The results showed that *Smcx* is subject to complete inactivation during the initial silencing of the X chromosome in early female embryogenesis, but subsequently becomes reactivated, presumably though a progressive loss of maintenance of the inactive epigenetic state. This would suggest that XCI in fact spreads uniformly along the Xi in early embryogenesis, then certain loci and domains fail to maintain the transcriptionally silent state and become reactivated later in development.

XCI poses a variety of unusual mechanistic challenges to female mammalian cells, including the problem of how to heritably silence genes on one of two essentially identical X chromosomes within the same nucleus, and how to spread inactivation in cis from the XIC along the length of the X chromosome. Although some or all of the DNA sequences (e.g., L1 elements), factors (e.g., *XIST* RNA), and mechanisms (e.g., RNA-mediated gene silencing) described may be involved in the mechanisms responsible for spreading of XCI, a clear understanding of the process by which XCI inactivation spreads in cis along the X chromosome remains to be determined. The process by which this occurs, while becoming clearer, stills remains mechanistically elusive, as the potential binding partners of *XIST* RNA and other factors recruited to mediate the silencing of genes on the Xi have yet to be identified. As more protein complexes are identified that associate with the Xi, we may begin to gain a better understanding of the mechanisms involved in spreading of XCI.

References

Bailey JA, Carrel L, Chakravarti A and Eichler EE (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis.

Proceedings of the National Academy of Sciences of the United States of America, **97**(12), 6634–6639.

Barr ML and Carr DH (1962) Correlations between sex chromatin and sex chromosomes. *Acta Cytologica*, **6**, 34–45.

Boyle AL, Ballard SG and Ward DC (1990) Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **87**(19), 7757–7761.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J and Willard HF (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**(3), 527–542.

Canun S, Mutchinick O, Shaffer LG and Fernandez C (1998) Combined trisomy 9 and Ullrich-Turner syndrome in a girl with a 46,X,der(9)t(X;9)(q12;q32) karyotype. *American Journal of Medical Genetics*, **80**(3), 199–203.

Carrel L and Willard HF (1999) Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(13), 7364–7369.

Carrel L, Cottle AA, Goglin KC and Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(25), 14440–14444.

Chadwick BP and Willard HF (2003) Chromatin of the barr body: histone and non-histone proteins associated with or excluded from the inactive X chromosome. *Human Molecular Genetics*, **12**(17), 2167–2178.

Disteche CM (1999) Escapees on the X chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(25), 14180–14182.

Hansen RS (2003) X inactivation-specific methylation of LINE-1 elements by DNMT3B: implications for the Lyon repeat hypothesis. *Human Molecular Genetics*, **12**(19), 2559–2567.

Lee JT and Jaenisch R (1997) Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. *Nature*, **386**(6622), 275–279.

Lingenfelter PA, Adler DA, Poslinski D, Thomas S, Elliott RW, Chapman VM and Disteche CM (1998) Escape from X inactivation of Smcx is preceded by silencing during mouse development. *Nature Genetics*, **18**(3), 212–213.

Lyon MF (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Die Naturwissenschaften*, **190**, 372–373.

Lyon MF (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenetics and Cell Genetics*, **80**(1–4), 133–137.

Penny GD, Kay GF, Sheardown SA, Rastan S and Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**(6561), 131–137.

Plath K, Talbot D, Hamer KM, Otte AP, Yang TP, Jaenisch R and Panning B (2004) Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *Journal of Cellular Biology*, **167**, 1025–1035.

Riggs AD (1990) Marsupials and mechanisms of X chromosome inactivation. *Australian Journal of Zoology*, **37**, 419–441.

Riggs AD, Singer-Sam J and Keith DH (1985) Methylation of the PGK promoter region and an enhancer way-station model for X-chromosome inactivation. *Progress in Clinical and Biological Research*, **198**, 211–222.

Russell LB (1963) Mammalian X-chromosome action: inactivation limited in spread and region of origin. *Science*, **140**, 976–978.

Schneider-Gadicke A, Beer-Romero P, Brown LG, Nussbaum R and Page DC (1989) ZFX has a gene structure similar to ZFY, the putative human sex determinant, and escapes X inactivation. *Cell*, **57**(7), 1247–1258.

Schramke V and Allshire R (2003) Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science*, **301**(5636), 1069–1074.

Sharp A, Robinson DO and Jacobs P (2001) Absence of correlation between late-replication and spreading of X inactivation in an X;autosome translocation. *Human Genetics*, **109**(3), 295–302.

Sharp AJ, Spotswood HT, Robinson DO, Turner BM and Jacobs PA (2002) Molecular and cytogenetic analysis of the spreading of X inactivation in X;autosome translocations. *Human Molecular Genetics*, **11**(25), 3145–3156.

Su RC, Brown KE, Saaber S, Fisher AG, Merkenschlager M and Smale ST (2004) Dynamic assembly of silent chromatin during thymocyte maturation. *Nature Genetics*, **36**(5), 502–506.

Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP and Disteche CM (2004) Comparative sequence and x-inactivation analyses of a domain of escape in human xp11.2 and the conserved segment in mouse. *Genome Research*, **14**(7), 1275–1284.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, Grewal SI and Moazed D (2004) RNAi-mediated targeting of heterochromatin by the RITS complex. *Science*, **303**(5658), 672–676.

White WM, Willard HF, Van Dyke DL and Wolff DJ (1998) The spreading of X inactivation into autosomal material of an x;autosome translocation: evidence for a difference between autosomal and X-chromosomal DNA. *American Journal of Human Genetics*, **63**(1), 20–28.

Wutz A and Jaenisch R (2000) A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Molecules and Cells*, **5**(4), 695–705.

Wutz A, Rasmussen TP and Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics*, **30**(2), 167–174.

Short Specialist Review

Initiation of X-chromosome inactivation

Lygia V. Pereira and Raquel Stabellini
Universidade de São Paulo, São Paulo, Brazil

Dosage compensation is the process by which the amount of X-linked gene products between individuals with one and two X chromosomes is equalized (*see* Article 15, **Human X chromosome inactivation**, Volume 1). Currently, three different mechanisms of dosage compensation are known in nature (reviewed by Marin *et al*., 2000). In *Drosophila*, the single X chromosome in males has a twofold increased level of transcription when compared to each of the two X chromosomes in females. In contrast, in *Caenorhabditis elegans* (*C. elegans*), a twofold decrease in X-linked gene expression in hermaphrodites (XX) relative to males (XO) ensures equalization of X-linked gene expression.

In mammals, dosage compensation between XY males and XX females is achieved by the transcriptional silencing of one X chromosome in female somatic cells (Lyon, 1961), a process called X-chromosome inactivation (XCI). Therefore, while in *Drosophila* and *C. elegans*, the levels of gene expression from each X chromosome in the female or hermaphrodite, respectively, are indistinguishable, a formidable feature of the mechanism of dosage compensation in mammals is that female cells must differentiate the two X chromosomes, rendering only one transcriptionally active.

A snapshot of the inactive X (Xi) in somatic cells would show several epigenetic modifications of this chromosome when compared to the active X (Xa) (reviewed by Hall and Lawrence, 2003). The Xi is coated by RNA from the *Xist* gene expressed in cis. It presents higher degree of methylation of CpG islands, higher concentration of the histone variant macroH2A1, and association with the BRCA1 protein. In addition, the histones in the Xi have several posttranslational modifications associated with gene silencing, including hypoacethylation of histones H4, methylation of lysine 9 of histones H3 (H3K9), and dimethylation of lysine 4 of histones H3 (H3K4).

The issue is: how did the Xi get from its originally active state at the zygote to that state of inactivity during early embryonic development? This epigenetic transformation requires that the cell counts the number of X chromosome and chooses which X will be inactive and which will be active so that in a differentiated cell there will be only one Xa per diploid genome.

Counting X chromosomes requires that the cell somehow differentiates the X from the other chromosomes. The identity of the X chromosome is tightly linked

to the X-inactivation center (Xic), mapped at Xq13 in humans and in the syntenic region located in band D of the X in mice. Only when this minimal region of the X is present in an X:autosome translocation will the translocated chromosome be counted by the cell as an X to participate in XCI. To date, two genes within the Xic have been demonstrated to be involved in XCI: *Xist* (X inactive specific transcript), expressed exclusively from the Xi and required for initiation of XCI; and its antisense *Tsix*, which downregulates *Xist* in cis in undifferentiated cells.

The choice of which X to be inactive is made in two very different forms in the mouse embryo: in cells of the trophectoderm, the paternal X (Xp) is always chosen to be the Xi, whereas in the inner cell mass (ICM) the choice is random in each cell. Imprinted XCI in the trophectoderm has been associated with maternal imprinting of the *Tsix* gene. In that lineage, lack of *Tsix* expression from the Xp leads to stabilization of *Xist* RNA expressed in cis, triggering inactivation of that chromosome.

In the ICM, although generally either the Xp or Xm is randomly chosen to be inactive, the choice of the Xi may be influenced by the X-controlling element (Xce), located within the Xic. Stronger Xce alleles, which render the corresponding X more likely to remain active, have been described in mice. Mutation studies of the *Xist* and *Tsix* genes, also within the Xic, indicate a role for those also in choosing, for instance, a 65-kb deletion 3' of *Xist* including most of the *Tsix* gene leads to primary completely skewed inactivation of the mutant X. In addition, dominant mutations mapped to different mouse autosomes can disrupt normal random XCI. However, the mechanism of choice of the Xi in cells of the ICM is still unknown.

Once the future Xi is chosen, inactivation starts. Until recently, XCI was thought to initiate at cells of the trophectoderm, where the Xp is always inactivated. Subsequently, random XCI would take place in cells of the ICM, where in each cell either the maternal X (Xm) or the Xp is inactivated.

However, more recent studies have detected XCI as early as the cleavage states before cellular differentiation (Okamoto *et al*., 2004; Mak *et al*., 2004). At that stage, imprinted XCI takes place in all cells of the embryo, imposing sequential epigenetic modifications in the Xp (Figure 1). These modifications are triggered in cis by the expression of the *Xist* gene from the Xp. At the 4-cell stage, *Xist* RNA is first observed, coating the Xp; H3K4 hypomethylation and H3K9 hypoacetylation are detected at the 8-cell stage; at 16-cell stage, the Eed/Enx1 polycomb group complex and the histone variant macroH2A accumulate in the Xp; finally, at the early blastocyst stage, association of methylated H3K9 with the Xp takes place. Together, these epigenetic modifications of the Xp lead to its inactivation.

At the blastocyst stage, while cells of the trophectoderm and the primitive endoderm maintain the inactive state of the Xp, cells of the ICM reactivate this chromosome, erasing the epigenetic marks imposed during the imprinted XCI. Reversible XCI had been reported in embryonic stem (ES) cells carrying an inducible *Xist* transgene (Wutz and Jaenisch, 2000). In this important experimental model of initiation of XCI, expression of *Xist* before differentiation leads to *Xist*-dependent and reversible XCI. Similarly, in the ICM, loss of *Xist* expression from the Xp leads to dissociation of the Eed/Enx1 complex followed by loss of histone H3K9 and K27 methylation, so that at implantation, the Xp is reactivated.

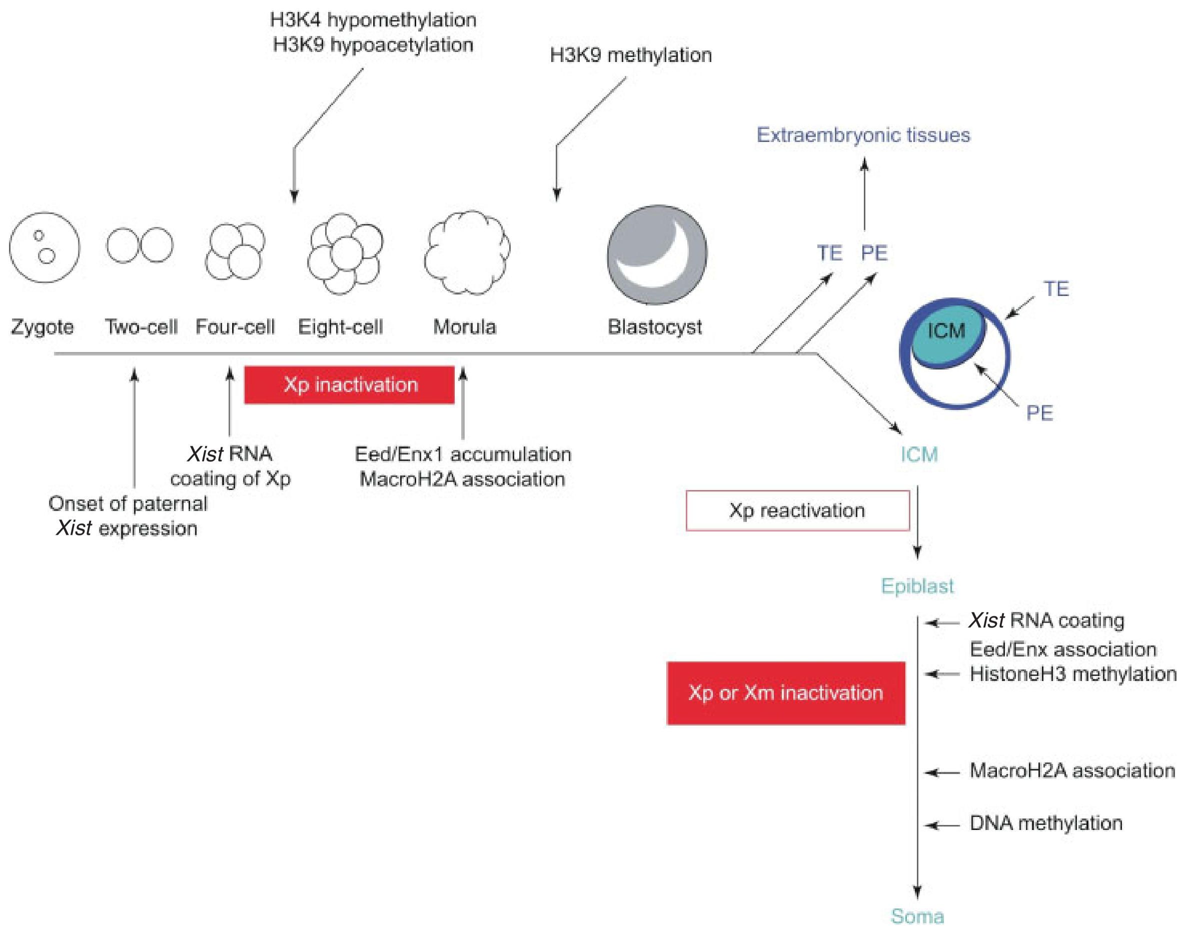


Figure 1 Kinetics of XCI during early mouse embryonic development. Imprinted inactivation of Xp established before cell differentiation is erased in the ICM. Random inactivation of either the Xp or Xm will then take place in cells of the epiblast (see text). PE, primitive ectoderm; TE, trophectoderm (Adapted from Heard E (2004) Recent advances in Z-chromosome inactivation. *Current opinion in Cell Biology*, **16**, 247–255, with permission from Elsevier)

At that point, a second round of XCI will take place in cells of the epiblast, starting with repression of *Tsix* in the future Xi, now chosen at random in each cell, and the consequential stabilization of *Xist* RNA in cis. The *Xist* RNA coating the future Xi recruits transiently the Eed/Enx1 complex required for methylation of histones H3, stabilizing the structure of the Xi chromatin. Further modifications of histones H4, recruitment of macroH2A, and methylation of CpG islands will lock that chromosome in an inactive state that is independent of *Xist* expression and heritable through mitosis.

XCI is traditionally thought of as a mechanism triggered by the presence of supernumerary Xs, involving counting the X chromosomes (or Xics) and randomly choosing the one to be inactivated. Alternatively, XCI can be looked at as a default mechanism in the mammalian cell: X chromosomes will be inactivated, unless they are somehow protected from inactivation. Therefore, the existence of a blocking or protective factor in the cell has been postulated, which must exist in very limited amounts, allowing protection from inactivation of only one X per diploid genome.

The affinity of the protective structure with an X, or with an Xic, may influence the probability of the corresponding X to remain active, that is, it may influence the choice of the Xa. Weaker Xce alleles may have a primary sequence with lower affinity with the protective structure, rendering the corresponding X unprotected from inactivation. Following that rationale, completely skewed inactivation of the X carrying the 65-kb *Tsix* deletion may be due to total inability of that chromosome to interact with the protective structure. One can also hypothesize that the autosomic factors influencing the choice of the Xi may be part of the protective structure, and therefore, identification of those factors may shed some light on the nature of that structure.

In the last few years, much has been learned about the nature and the dynamics of epigenetic modifications imposed in the Xi during XCI, particularly highlighting the role of posttranslational modifications of histones in defining the epigenetic state of the Xi (*see* Article 40, **Spreading of X-chromosome inactivation**, Volume 1). However, the mechanisms by which the female cell chooses and protects one X chromosome from inactivation and the players that impose those epigenetic modifications on the Xi remain obscure and a fascinating topic of research in modern biology.

Finally, it is important to notice that most, if not all, that is known about initiation of XCI comes from studies in the mouse, either in preimplantation embryos or in ES cells. Important differences between XCI in mouse and humans exist, including the apparent absence of a functional human *TSIX* gene, and of imprinted XCI in human extraembryonic tissues (reviewed by Vasques *et al*., 2002). Therefore, experimental systems for the study of XCI in humans must be developed. In that sense, the recent availability of human ES cell lines (Cowan *et al*., 2004) may allow the dissection of initiation of human XCI.

References

Cowan CA, Klimanskaya I, McMahon J, Atienza J, Witmyer J, Zucker JP, Wang S, Morton CC, McMahon AP, Powers D, *et al*. (2004) Derivation of embryonic stem-cell lines from human blastocysts. *New England Journal of Medicine*, **350**, 1353–1356.

Hall LL and Lawrence JB (2003) The cell biology of a novel chromosomal RNA: Chromosome painting by XIST/Xist RNA initiates a remodeling cascade. *Seminars in Cell and Developmental Biology*, **14**, 369–378.

Heard E (2004) Recent advances in Xchromosome inactivation. *Current Opinion in Cell Biology*, **16**, 247–255.

Lyon MF (1961) Gene action in the X chromosome of the mouse (*Mus musculus* L.). *Nature*, **190**, 372–373.

Mak W, Nesterova TB, de Napoles M, Appanah R, Yamanaka S, Otte AP and Brockdorff N (2004) Reactivation of the paternal X chromosome in early mouse embryos. *Science*, **303**, 666–669.

Marin I, Siegal ML and Baker BS (2000) The evolution of dosage-compensation mechanisms. *Bioessays*, **22**, 1106–1114.

Okamoto I, Otte AP, Allis CD, Reinberg D and Heard E (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science*, **303**, 644–649.

Vasques LR, Klockner MN and Pereira LV (2002) X chromosome inactivation: How human are mice? *Cytogenetic and Genome Research*, **99**, 30–35.

Wutz A and Jaenisch R (2000) A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Molecular and Cellular Proteomics*, **5**, 695–705.

Short Specialist Review

Mechanisms of epigenetic loss of chromosomes in insects

Clara Goday and Maria-Fernanda Ruiz
Consejo Superior de Investigaciones Científicas, Centro de Investigaciones Biologicas, Madrid, Spain

1. Introduction

The programmed exclusion of chromosomes from the genome is a remarkable developmental phenomenon that in insects is very diversified between different families and between individual species. The best-known examples are found in Diptera such as *Sciara* (Sciaridae), *Heteropeza*, *Miastor*, *Mayetiola*, *Wachtiella* (Cecidomyiidae), *Acritocopus* (Chironomidae), and also in Homoptera coccids (reviewed in White, 1973). A common feature is that specific chromosomes are eliminated from presomatic cells at early cleavage divisions by the time of somatic/germ-line separation. In some species, there is an additional chromosome loss from germ cells. In all cases, this process leads to a reduction in the number of chromosomes in the somatic tissues compared to the germline.

Somatic embryonic elimination involves chromosomes of the regular complement (mostly sex chromosomes) and/or chromosomes that are restricted to germline. In both cases, the timing of chromosome elimination in early embryos is species specific. A classic example of sex chromosomes elimination is that of *Sciara* where the loss of one or two X chromosomes at early cleavages determines the sex of the embryo (Metz and Moses, 1926). That this type of elimination is preceded by imprinting was discovered in *Sciara* when it was found that the discarded X chromosomes are invariably of paternal origin (Crouse, 1960; reviewed in Gerbi, 1986). Despite its importance, the parent-of-origin-specific mark(s) involved in this elimination system is still unknown. A substantial number of examples, including cecidomyiids species, show elimination of sex chromosomes linked to sex determination (Nicklas, 1959; reviewed in White, 1973). An extreme case is found in coccids diaspididae where the somatic elimination of the entire paternal chromosome set produces solely male embryos (reviewed in White, 1973). On the other hand, germ-line-limited chromosomes, present in certain sciarids and most cecidomyiids and chironomids ("L", "E", or "K" chromosomes, respectively), are all discarded from the presumptive somatic nuclei at early cleavages in both sexes. This is independent of their number, which can be extremely high and varies greatly between species. In this respect, the elimination of germ-line-limited chromosomes

in early cleavages reduces the chromosome number in the future somatic nuclei to a level characteristic of the Diptera in general. This may be particularly important, since the developmental pattern in Diptera has been adjusted to a low number of chromosomes in relatively small nuclei (Nicklas, 1959). Most of the germ-line-limited chromosomes are of heterochromatic nature and contain repetitive DNA sequences. The functional role of these chromosomes in the germline remains obscure. However, experimental elimination of the E chromosomes in pole cells of cecidomyiids species revealed that their presence is necessary for the female gonad development (Geyer-Duszynska, 1959).

In germ cells, chromosome elimination commonly involves the loss of the whole paternal regular set of chromosomes during male meiosis. In *Sciara* species, an exceptional and unique type of paternal X-chromosome nuclear exclusion also takes place in embryonic germ cells of both sexes (reviewed in Goday and Esteban, 2001).

Despite many years of study, chromosome loss in insects is far from being exhaustively analyzed. An intriguing question is whether the cellular and molecular mechanisms underlying these processes share common traits between different insects. With this in mind, we discuss here relevant data coming from examples in Diptera, Sciaridae, Cecidomyiidae, and Chironomidae.

2. Mechanisms of chromosome elimination in the soma

In all species, chromosome loss at early embryonic mitotic divisions is produced by an abnormal segregation of the chromosomes. A regular cytological feature of this event is the occurrence of “lagging chromosomes” at anaphase, so that these chromosomes fail to enter the daughter nuclei.

In sciarids, as classically described for L- and X-chromosome elimination processes in *S. coprophila* (Dubois, 1933; reviewed in Gerbi, 1986), the chromosomes begin their movement toward the poles at anaphase but are incapable of complete chromatid separation and remain at the equatorial plate (see Figures 1 and 2). These early observations led to the proposal that alterations in centromeric activity cause chromosome loss in sciarids (reviewed in Gerbi, 1986). In *S. coprophila*, moreover, it was found that a cis-acting locus or *controlling element* (CE) located in a heterochromatic block near the centromere of the X chromosome regulates the elimination process (Crouse, 1960). The molecular nature of the CE is still undetermined, but when the CE is translocated to an autosome it is able to direct its elimination in paternally inherited translocations (Crouse, 1979). Recent analysis of L- and X-chromosome elimination kinetics in *S. coprophila* through confocal microscopy showed that the centromeres remain attached to the spindle and stretched toward the poles at anaphase, while the chromatids remain joined at a region on the long arm of the X chromosome (de Saint-Phalle and Sullivan, 1996). It was proposed that anaphase lag (of X and L chromosomes) is caused primarily by a CE-controlled failure of chromatid separation rather than by a CE-controlled centromere dysfunction (de Saint-Phalle and Sullivan, 1996). To find the specific biochemical alterations in the chromatid separation processes at the anaphase transition would be extremely interesting to further develop this conclusion.

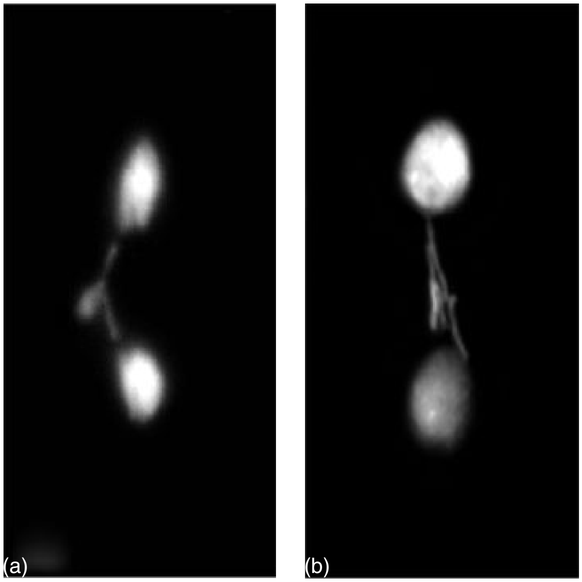


Figure 1 X-chromosome elimination from the soma in *S. ocellaris* embryos. DAPI-stained early syncitial somatic divisions. Lagging X chromosomes in somatic division undergoing elimination. Only one X chromosome is discarded from cells of female embryos (a), while two X chromosomes are discarded from cells of male embryos (b) (Reproduced by permission of John Wiley & Sons, Inc. from C. Goday and M. R. Esteban (2001) *BioEssays*, **23**, 242–250)

Another yet undiscovered issue is the nature of the cytoplasmic factor(s), produced by the mother and distributed within the egg, involved in regulating the number of X chromosomes eliminated in sciarids. As demonstrated in sciarid species, this factor(s) is produced in the oocyte during oogenesis (reviewed in Gerbi, 1986 and Goday and Esteban, 2001). Two main models have been put forward to explain the number of eliminated X chromosomes (de Saint-Phalle and Sullivan, 1996; Sánchez and Perondini, 1999). The one-factor model (de Saint-Phalle and Sullivan, 1996) is based on a maternal factor that regulates the differential X-chromosome elimination pattern through the quantity of this factor present in the egg. In the alternative, two-factor model (Sánchez and Perondini, 1999), a hypothetical chromosomal factor interacts with the X chromosome(s) causing its (their) elimination. In this model, the number of X chromosomes to be eliminated is controlled by a maternal factor that regulates the amount of free chromosomal factor interacting with the X chromosomes (reviewed in Goday and Esteban, 2001).

In contrast to Sciara, in the chironomid *Acritopus* during Ks-chromosomes elimination, the Ks sister chromatids are not stretched in the direction of the spindle poles, and their centromeres appear not to separate while somatic chromosomes move to the poles (Staiber, 2000). Since Ks chromosomes stay in the equatorial plate, it was concluded that their sister chromatids remain joined at their centromeric regions rather than at the chromosomes arms or the telomeres (see Figure 2). Moreover, it was proposed that proteins responsible for centromeric cohesion might be involved in this type of chromosome behavior (Staiber, 2000).

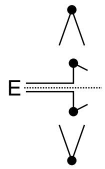
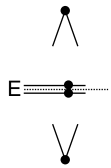
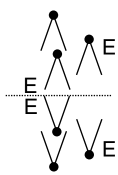
		Cytological events (anaphase)	Proposed elimination mechanisms
Sciaridae		Incomplete anaphasic movement of chromosomes	Failure of sister chromatid separation at chromosomes arms
Chironomidae		Lack of anaphasic movement of chromosomes	Failure of sister chromatid separation at centromeric regions
Cecidomyiidae		Aberrant anaphasic movement of chromosomes	Alterations in centromere kinetic activity

Figure 2 Chromosome elimination in embryonic somatic cells. Diagram summarizing the“lagging chromosomes” phenotypes, relevant cytological events, and proposed elimination mechanisms. E denotes the chromosomes undergoing elimination

Interestingly, a specific highly repetitive DNA family located in the paracentromeric heterochromatin of the Ks chromosomes was identified in *A. lucidus* (Staiber *et al*., 1977). Whether these repetitive DNA sequences are involved in identifying the eliminating chromosomes as proposed (Staiber *et al*., 1977) remains unknown.

The elimination of E chromosomes in cecidomyiids displays varied chromosome-lagging patterns (see Figure 2). In examples of *Miastor*, *Micophila,* and *Heteropeza*, E chromosomes remain at the equator as a result of the apparent absence or diminution of the usual mid-anaphase tension (Nicklas, 1959; Nicklas, 1960; White, 1973). In several of the observations, it is clear that among the chromosomes to be lost there are chromatid pairs that remain totally separate and yet fail to exhibit normal mid-anaphase movement. In *Mayetiola* and *Wachtliella* examples, all E chromatids separate completely but fail to continue anaphase along with S chromosomes (Geyer-Duszynska, 1959; reviewed in White, 1973). A remarkable case is that of *Heteropeza pygmaea* where until mid-anaphase both the E and S chromosomes segregate to the poles as in a normal cleavage division. Time-lapse cinémicrography revealed that the velocity of the E chromosomes, however, is less than that of the S chromosomes. After variable amounts of anaphase movements, the E chromosomes return toward the equator with their kinetochores being still oriented toward the poles (Camezind, 1974). From these, and other cytological observations, it was generally accepted that functional defects in the centromeres of E chromatids were responsible for causing elimination in cecidomyiids. So far,

nothing is known about the biochemical and molecular organization of the E-chromosomes' centromere with respect to those of S chromosomes; neither about specific DNA sequences candidates to constitute molecular chromosome landmarks for determining elimination in cecidomyiids. Both kinds of studies would no doubt shed light on the mechanisms and control of elimination.

Figure 2 shows a schematic diagram summarizing the main "lagging chromosomes" phenotypes and the emphasized elimination mechanisms.

3. Mechanisms of chromosome elimination in the male germline

In Sciaridae and Cecidomyiidae, there is an additional elimination of chromosomes in male germ cells during spermatogenesis. In this elimination event, the paternal chromosome set is discarded in Sciaridae. In Cecidomyiidae, elimination includes a haploid set of somatic chromosomes (presumably paternally derived) plus all, or nearly all, E chromosomes (reviewed in White, 1973 and Gerbi, 1986). In male meiosis I, the absence of homolog pairing, synapsis, and metaphase alignment is a common characteristic for both groups.

In sciarids, in anaphase I, only maternal (and L chromosomes when present) move toward the single pole of a monocentric first meiotic spindle, and become included into the daughter nucleus. The paternal set, in contrast, segregates away from the maternal set into a cytoplasmic bud that is later cast off from the spermatocyte (see Figure 3). Several observations support the conclusion that differential kinetic behavior of *Sciara* maternal and paternal chromosomes is accomplished by a monopolar spindle and by non-spindle cytoplasmic bud microtubules (Kubai, 1982; Fuge, 1994; Esteban *et al*., 1997). Furthermore, unorthodox microtubule-organizing centers (MTOCs) have been found to be responsible for the assembly and polarity of microtubules in the bud regions in *Sciara* spermatocytes (reviewed in Esteban *et al*., 1997). Such microtubules class are specifically involved in capturing and retaining paternal chromosomes in the spermatocytes bud. Most interesting, the presence of organized kinetochores in paternal chromosomes appears not to be necessary for their regular elimination among sciarids. Hence, an essential role of the centromeres in this kind of elimination seems to be discardable (reviewed in Goday and Esteban, 2001).

As in *Sciara*, in the cecidomyiids *Miastor*, *Heteropez,* and presumably in *Mayetiola*, a monopolar spindle directs anaphase I (Nicklas, 1959, 1960; White, 1973; Stuart and Hatchett, 1988). A haploid set of S chromosomes orientate themselves with their centromeres directed toward the single pole while the remainder (one haploid set of S chromosomes and the E chromosomes) remain, generally less condensed, in an unorientated group on the opposite nuclear side. Two kinds of secondary spermatocytes are formed that are of very different sizes, the larger being a residual cell that contains the discarded chromosomes. As in *Sciara*, the chromosomes that interact with, and migrate to, the single spindle pole are the ones that will be maintained and included in the sperm nucleus. So far, it is not known if in cecidomyiids unorthodox MTOCs generating nonspindle

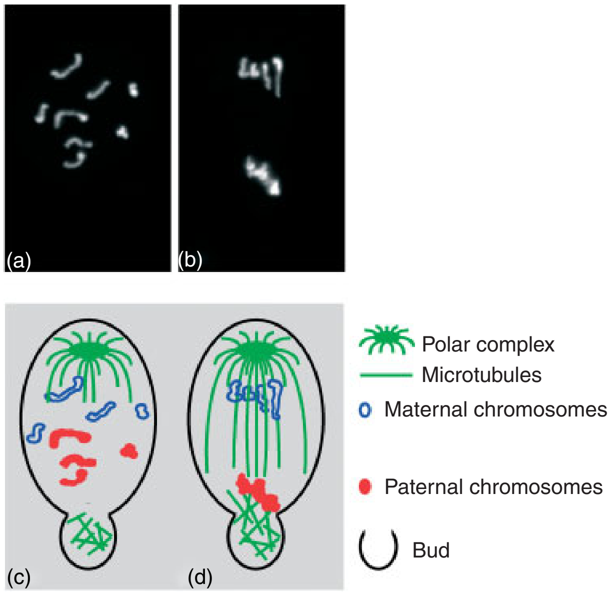


Figure 3 Chromosome elimination during first male meiosis in *S. ocellaris*. Upper row: DAPI-stained spermatocyte chromosomes. (a) Prophase I; (b) Anaphase I. Lower row: a diagrammatic representation of the same pictures illustrating the chromosome interactions with the microtubules of the first meiotic spindle and bud microtubules in the spermatocyte. (c) Prophasic chromosomes do not pair. Maternal and paternal chromosomes display a separate arrangement within the nucleus. A monopolar spindle is formed and nonspindle microtubules are generated in the cytoplasmic bud regions. (d) Maternal chromosomes move toward the single pole while paternal chromosomes segregate into the bud (Reproduced by permission of John Wiley & Sons, Inc. from C. Goday and M. R. Esteban (2001) *BioEssays*, **23**, 242–250)

microtubules are also involved, together with the spindle microtubules, in the elimination of chromosomes. If this is so, they may, as in *Sciara*, be part of the established cellular mechanism that assures the regular elimination of specific chromosomes.

A highly relevant feature, common to cecidomyiids and sciarids, is that there is a spatial compartmentalization of chromosomes within the meiotic prophase nuclei. Ultrastructural studies in the cecidomyiid *Monarthropalpus buxi* demonstrated that the non- and eliminated chromosomes are separated in the spermatogonium nucleus by a complex system of intranuclear lammellae (reviewed in Jazdowska-Zagrodzinska and Matuszewski, 1978). Similarly, in *S. coprophila* male germ cells, paternal and maternal chromosomal sets occupy distinct compartments in the meiotic prophase nuclei, and this accounts for their nonrandom chromosome segregation during anaphase I (Kubai, 1982). The territorial separation of the two chromosome sets most probably facilitates the proper interactions of each parental chromosome group with the microtubular system that further separates them at anaphase I (Kubai, 1982; Goday and Esteban, 2001). The existence of separate chromosomal territories is further supported by recent data indicating that histone acetylation between chromosomes of different parental origin highly differ, both in

	Cytological events (meiosis I)	Proposed elimination mechanisms
Sciaridae Cecydomyiidae	No pairing of homolog chromosomes No metaphase alignment Intranuclear chromosomes compartmentalization Monopolar spindle formation	Differential kinetic behavior of chromosomes at anaphase I Active role of monopolar spindle structure in capturing noneliminating chromosomes
Sciaridae	Generation of non-spindle citoplasmic microtubules	Active role of non-spindle microtubules in capturing eliminating chromosomes

Figure 4 Chromosome elimination at male first meiotic division. Summary of relevant cytological events and proposed elimination mechanisms

early germ nuclei as well as during first male meiotic division in *Sciara* (Goday and Ruiz, 2002).

The most relevant cytological events of chromosome elimination during first male meiosis and proposed mechanisms are summarized in Figure 4.

In conclusion, a common feature underlying chromosome elimination in insects is the occurrence of changes in the chromosome segregation modalities. Different tissue-specific mechanisms have evolved to achieve chromosome elimination in somatic cells and in germ cells. Modifications in centromeric functional activities and failure of sister chromatid separation are thought to account for somatic elimination events. On the other hand, chromosome elimination during gametogenesis can be conceived as the result of the combined role of a specific spindle structure (monopolar spindle) and cytoplasmic microtubules (generated by unorthodox MTOCs). The success of such a microtubule-based differential meiotic segregation seems to require a previous intranuclear specification of the chromosomal set domains.

Acknowledgments

We would like to thank Lucas Sánchez and James Haber for their valuable comments on the manuscript. We apologize to colleagues whose original research papers could not be cited because of space limitations.

References

Camezind R (1974) Chromosome elimination in *Heteropeza pygmea* I. *In vitro* observations. *Chromosoma*, **49**, 87–112.

Crouse HV (1960) The controlling element in sex chromosome behavior in *Sciara*. *Genetics*, **45**, 1429–1443.

Crouse HV (1979) X heterochromatin subdivision and cytogenetic analysis in *Sciara coprophila* (Diptera, Sciaridae) II. The controlling element. *Chromosoma*, **74**, 219–239.

de Saint-Phalle B and Sullivan W (1996) Incomplete sister chromatid separation is the mechanism of programmed chromosome elimination during early *Sciara coprophila* embryogenesis. *Development*, **122**, 3775–3784.

Dubois AM (1933) Chromosome behavior during cleavage n the eggs of *Sciara coprophila* (Diptera) in the relation to the problem of sex determination. *Z Wiss Biol Abt B-Z Zellforsch Mikrosk Anat*, **19**, 595–614.

Esteban MR, Campos MCC, Perondini ALP and Goday C (1997) Role of microtubules and microtubule organizing centers on meiotic chromosome elimination in *Sciara ocellaris*. *Journal of Cell Science*, **110**, 721–730.

Fuge H (1994) Unorthodox male meiosis in *Trichosia pubescens* (Sciaridae)- chromosome elimination involves polar organelle degeneration and monocentric spindles in first and second division. *Journal of Cell Science*, **107**, 99–312.

Gerbi SA (1986) Unusual chromosome movements in sciarid flies. In *Germ Line-Soma Differentiation. Results and Problems of Cell Differentiation*, Vol 13, Hennig W (Ed.), Springer-Verlag: New York, pp. 71–10.

Geyer-Duszynska I (1959) Experimental research on chromosome elimination in Cecidomyidae (Diptera). *Journal of Experimental Zoology*, **141**, 391–488.

Goday C and Esteban R (2001) Chromosome elimination in sciarid flies. *BioEssays*, **23**, 242–250.

Goday C and Ruiz MF (2002) Differential acetylation of histones H3 and H4 in paternal and maternal germline chromosomes during development of sciarid flies. *Journal of Cell Science*, **115**, 4765–4775.

Jazdowska-Zagrodzinska B and Matuszewski B (1978) Nuclear lamellae in the germ-line cells of gall midges (Cecidomyiidae, Diptera). *Experientia*, **34**, 777–778.

Kubai DF (1982) Meiosis in *Sciara coprophila*: Structure of the spindle and chromosome behavior during the first meiotic division. *Journal of Cell Biology*, **l93**, 655–669.

Metz CW and Moses MS (1926) Sex determination in *Sciara* (Diptera). *The Anatomical Record*, **34**, 170.

Nicklas RB (1959) An experimental and descriptive study of chromosome elimination in *Miastor* spec(Cecidomyidae, Diptera). *Chromosoma*, **10**, 301–336.

Nicklas RB (1960) The chromosome cycle of a primitive cecidomyiid-*Mycophila speryeri*. *Chromosoma*, **11**, 402–418.

Sánchez L and Perondini ALP (1999) Sex determination in sciarid flies: A model for the control of differential X-chromosome elimination. *Journal of Theoretical Biology*, **197**, 247–259.

Staiber W (2000) Immunocytological and FISH analysis of pole cell formation and soma elimination of germ line-limited chromosomes in the chironomid *Acritopus lucidus*. *Cell and Tissue Research*, **302**, 189–197.

Staiber W, Wech I and Preiss A (1977) Isolation and chromosomal characterization of a germ line-specific highly repetitive DNA family in *Acritopus lucidus* (Diptera, Chironomidae). *Chromosoma*, **106**, 267–275.

Stuart JJ and Hatchett JH (1988) Cytogenetics of the Hessian Fly: II. Inheritance and behavior of somatic and germ-line-limited chromosomes. *Journal of Heredity*, **79**, 190–199.

White MJD (1973) *Animal Cytology and Evolution*, *Third Edition*, Cambridge University Press.

Short Specialist Review

Epigenetic inheritance and RNAi at the centromere and heterochromatin

Kristin S. Caruana Scott
Duke University, Durham, NC, USA

Beth A. Sullivan
Boston University School of Medicine, Boston, MA, USA

1. Introduction

The discovery of RNA interference (RNAi) and regulatory RNAs has revolutionized views of gene regulation and chromosome structure. Originally implicated in posttranscriptional silencing, developmental regulation, and defense against transposition and viral invasion, RNAi involves the degradation of mRNA by small (21–23 nucleotide) inhibitory dsRNA duplexes (siRNAs). Argonaute, Dicer, and RNA-directed RNA polymerase (RdRP) mediate RNAi in many organisms. Dicer proteins, which are RNase III class enzymes, mediate cleavage of long dsRNAs into siRNAs. These siRNAs are incorporated into the multiprotein RNA-induced Silencing Complex (RISC) (Dykxhoorn *et al*., 2003). The antisense strand of the siRNA targets RISC to its homologous mRNA, triggering mRNA degradation or arresting the translation of mRNA into protein. Recent findings in the fission yeast *Schizosacchromyces pombe* (*S. pombe*) have uncovered a link between the RNAi pathway and assembly of pericentric heterochromatin.

2. RNAi directs heterochromatin assembly in fission yeast

Heterochromatin is a densely packaged chromatin that represses gene transcription, and is formed at repetitive DNA regions. It is required for cohesion between sister centromeres, ensuring proper chromosome segregation (Bernard *et al*., 2001; Lopez *et al*., 2000). Heterochromatin assembly depends on methylation of histone H3 at lysine 9 (H3 K9-Me) by the conserved histone methyltransferase, Clr4 [Su(var)3-9, SUVAR39H]. H3 K9-Me creates binding sites for proteins, such as Swi6 [Su(var)2-5, HP1] (Lachner *et al*., 2001), which are involved in both heterochromatin propagation and epigenetic inheritance of transcriptionally silent

chromatin (Grewal and Elgin, 2002; *see also* Article 27, **The histone code and epigenetic inheritance**, Volume 1).

Unlike higher eukaryotes, the *S. pombe* genome contains a single copy of each of the RNAi components: argonaute (*ago1*), dicer (*dcr1*), and RdRP (*rdp1*). Deleting any of these genes has no effect on cell viability. However, these mutants display chromosome segregation defects (Volpe *et al*., 2002), loss of H3 K9-Me at centromeres, and increased expression of pericentric reporter genes (Provost *et al*., 2002; Volpe *et al*., 2002). On the basis of the known RNAi pathway in nematodes and plants, RNAi machinery should process dsRNAs that originate from pericentric repeats. Indeed, ~22-nucleotide long Dicer cleavage products, complementary to both strands of centromeric repeats, called short heterochromatic RNAs (shRNAs), were detected in wild-type cells (Reinhart and Bartel, 2002).

By uncovering a novel role of RNAi components, these studies have defined a pathway for assembly of pericentric heterochromatin in fission yeast (Figure 1). First, tandemly repeated centromeric DNA is transcribed in the forward and reverse directions, producing dsRNAs. RdRP is involved in this process, perhaps to amplify RNA accumulation, since it is associated in vivo with centromeric repeats. Dicer-dependent processing of transcripts into shRNAs requires functional Ago1 and Rdp1, as well as Clr4/Su(var)3-9 (Schramke and Allshire, 2003). Once processed,

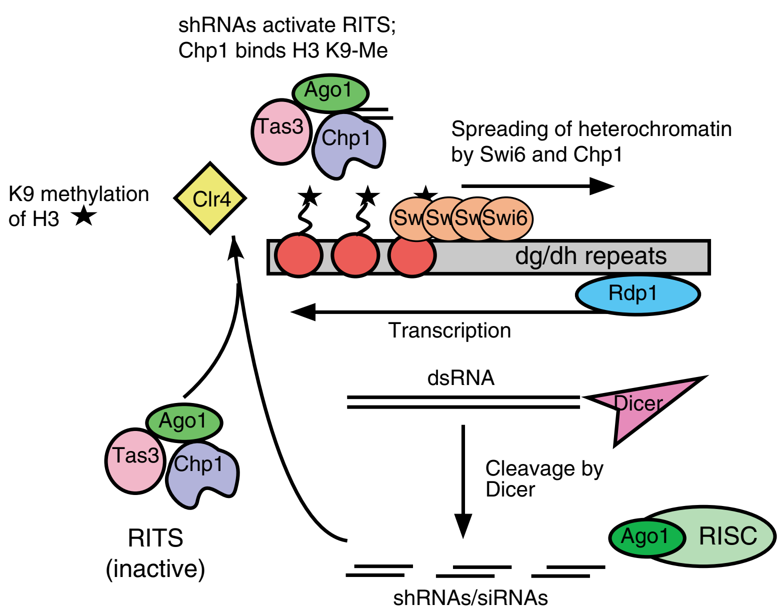


Figure 1 RNAi-dependent heterochromatin assembly at *S. pombe* centromeres. Transcripts homologous to centromeric repeats (dg/dh) (gray bar) are generated and amplified by Rdp1/RdRP and processed by Dicer into shRNAs/siRNAs. The RITS complex is activated upon interactions between Ago1 and shRNAs, and is directed to centromeric regions by the homologous small RNAs. Tas3 and Chp1 also act coordinately with the shRNAs to target the complex to nucleosomes. Clr4 methylation of H3 at lysine 9 (K9) (black stars) recruits both Chp1 and Swi6, which propagate or spread heterochromatin past the initiation site

shRNAs associate with the chromodomain protein Chp1, a component of the preassembled RITS (RNA-induced Initiation of Transcriptional Silencing) complex, which also contains Ago1 and Tas3, a novel protein that is necessary for localizing RITS to heterochromatin (Verdel *et al*., 2004). shRNA inclusion in RITS enables recognition of domains targeted for heterochromatin assembly, since, in *dcr1* mutants, the RITS complex forms, but histone modification and heterochromatin assembly is impaired. Thus, analogous to RISC, RITS directly links the RNAi components to heterochromatin through either RNA–RNA or RNA–DNA interactions (Verdel *et al*., 2004). The actions of histone deacetylases and histone methyltransferases must occur at this time, since only after lysine 9 of H3 is deacetylated can Clr4/Su(var)3-9 methylate this same residue. Swi6/HP1 and Chp1 then bind to H3 K9-Me and propagate the silent chromatin structure to surrounding sequences (see below). Remarkably, the RNAi machinery can trigger heterochromatin formation at noncentromeric sites in the presence of either a cis-linked centromere-related DNA (Hall *et al*., 2002) or by a hairpin RNA provided in trans and that has homology to a target locus (Schramke and Allshire, 2003). Moreover, Schramke and Allshire (2003) have shown that in *S. pombe*, LTRs maintain a number of meiosis-specific euchromatic genes transcriptionally silent by RNAi-dependent heterochromatin assembly.

3. How is heterochromatin maintained?

Similar to cell signaling pathways, a feedback loop may ensure heterochromatin integrity. The reverse strand of *S. pombe* centromeric repeats is always transcribed in wild-type cells, whereas the transcription of the forward strand is inhibited by heterochromatin assembly, specifically Swi6, and detected only in RNAi mutants (Volpe *et al*., 2002). Thus, stochastic loss of intact heterochromatin would catalyze the upregulation of forward-strand RNA, increasing the pool of dsRNAs available for processing and reassembly of heterochromatin over the forward-strand of centromeric repeats.

Several proteins implicated in the RNAi pathway act at multiple steps of heterochromatin assembly. For example, Chp1 associates with shRNAs in RITS, which is recruited to repetitive DNA as a complex (Verdel *et al*., 2004); it is necessary for Clr4-dependent H3 methylation at centromeres (Partridge *et al*., 2002), and finally, it binds methylated chromatin through its chromodomain (Partridge *et al*., 2000; Partridge *et al*., 2002). Thus, Chp1 may both establish and maintain pericentric heterochromatin. Clr4 is required for the processing of dsRNA transcripts by Dicer, suggesting that in addition to its downstream role in methylating histone H3, it acts coordinately, perhaps through its chromodomain, with the RNAi machinery early in heterochromatin assembly (Schramke and Allshire, 2003). Although both the chromo and SET domains of Clr4 are required in vivo for histone H3 methylation, mutations in *clr4* have varying effects on H3 K9-Me, Swi6 localization, and reporter gene silencing within pericentric heterochromatin (Nakayama *et al*., 2001). Swi6 is not required for the initial methylation event triggered by RNAi, yet it is clearly required for the spreading of H3 K9-Me beyond the sequence-specific initiation site (Hall *et al*., 2002; Schramke

and Allshire, 2003). Chp1 is also likely involved, since it binds to and directs histone H3 methylation. Although a direct interaction between Clr4 and Swi6 in *S. pombe* remains unproven, mammalian SUVAR39H has been shown to interact with HP1 to induce gene silencing, and presumably heterochromatin spreading (Schotta *et al*., 2002).

4. RNAi and heterochromatin in other eukaryotes

In Drosophila and mammals, the role of RNAi in heterochromatin assembly and function has only recently been explored. While the pathways have yet to be fully dissected, there is compelling evidence linking RNAi components to heterochromatic gene silencing in Drosophila. Single or multiple copies of a marker gene inserted within pericentric heterochromatin are subject to silencing by heterochromatin formation (Dorer and Henikoff, 1994). H3 K9-Me and heterochromatin proteins HP1 and HP2 are known to bind at the site of the silenced marker (Pal-Bhadra *et al*., 2004). Candidate genetic loci that participate in RNAi-mediated heterochromatin assembly can be identified by conserved protein motifs, such as PIWI and PAZ domains. A recent study showed that mutations in RNAi components *piwi* that encodes a PAZ domain/PIWI domain (PPD) protein, *aubergine* (*aub*), and homeless (*hls*), encoding a DEAD-motif RNA helicase, alleviated silencing of the marker gene or transgene array (Pal-Bhadra *et al*., 2004). *Piwi* and *aub* mutants partially alleviated silencing, but *hls* mutants showed dominant suppression of silencing. Accordingly, in *hls* mutants, HP1 and HP2, as well as H3 K9-Me, were markedly decreased within the pericentromeric regions and at the site of the normally silent marker genes. Even more striking, though, was the observation that these proteins were dramatically redistributed from the typically heterochromatic chromocenter to euchromatin. Similar effects on nuclear and chromosome architecture have been observed for the localization of Swi6 in RNAi mutants in *S. pombe* (Hall *et al*., 2003). Thus, RNAi has an additional, profound role in chromatin organization within distinct nuclear and chromosomal compartments.

Vertebrate centromeres are also located in repetitive DNA (Sullivan *et al*., 2001), and it is reasonable to assume a link between transcripts from centromeric repeats and the RNAi pathway. Studies in mouse suggest that RNA may be involved in pericentric heterochromatin assembly, as H3 K9-Me cannot be detected at centromeres after ribonuclease treatment of permeabilized mouse cells (Maison *et al*., 2002). Although transcripts homologous to centromeric satellite sequences in mice have been reported, a direct relationship between RNA–RNA or RNA–DNA interactions and heterochromatin assembly has not been demonstrated (Lehnertz *et al*., 2003; Rudert *et al*., 1995).

Centromeric DNAs, including satellite repeats and transposons, have historically been considered “junk DNA”, and it has been assumed that centric heterochromatin serves as a dumping ground for these elements and their relics. However, the demonstration that LTRs can regulate gene expression through the RNAi pathway (Schramke and Allshire, 2003) raises the possibility that these elements are actively involved in establishing pericentric heterochromatin. In addition to satellite DNAs, transposon- and LTR- like sequences are often found within or near centromeres

of higher eukaryotes. Intriguingly, de novo centromere function, as assayed by the presence of kinetochore proteins, cannot occur in the absence of a CENP-B box, a transposon-like motif present within human alpha satellite DNA (Ohzeki *et al*., 2002). Furthermore, some neocentromeres (centromeres formed on noncentromeric DNA) are flanked by or are within regions that are enriched for LTRs and/or tandem repeats. One intriguing possibility is that LTRs and the RNAi pathway create a heterochromatic environment flanking that promotes and regulates centromeric chromatin assembly.

Much remains to be learned about RNAi and epigenetic centromere assembly and regulation, but the importance of RNA–RNA or RNA–DNA interactions in chromosome structure and the complex inheritance of genome modification is an emerging theme. Future studies are needed to dissect the steps involved in targeting of RNAi machinery to specific sequences. Many of the genes required for RNAi are redundant in higher eukaryotes, indicating added levels of complexity in chromatin organization and regulation. A better understanding of the link between the RNAi pathway and heterochromatin assembly will also advance our knowledge of the formation of specialized nuclear and chromosomal domains, and the overall impact of these domains on chromosome biology and gene expression.

Related article

Article 27, **The histone code and epigenetic inheritance**, Volume 1

References

Bernard P, Maure JF, Partridge JF, Genier S, Javerzat JP and Allshire RC (2001) Requirement of Heterochromatin for Cohesion at Centromeres. *Science*, **294**, 2539–2542.

Dorer DR and Henikoff S (1994) Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. *Cell*, **77**, 993–1002.

Dykxhoorn DM, Novina CD and Sharp PA (2003) Killing the messenger: short RNAs that silence gene expression. *Nature Reviews. Molecular Cell Biology*, **4**, 457–467.

Grewal SI and Elgin SC (2002) Heterochromatin: new possibilities for the inheritance of structure. *Current Opinion in Genetics & Development*, **12**, 178–187.

Hall IM, Noma K and Grewal SI (2003) RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 193–198.

Hall IM, Shankaranarayana GD, Noma K, Ayoub N, Cohen A and Grewal SI (2002) Establishment and maintenance of a heterochromatin domain. *Science*, **297**, 2232–2237.

Lachner M, O'Carroll D, Rea S, Mechtler K and Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, **410**, 116–120.

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T and Peters AH (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current Biology*, **13**, 1192–1200.

Lopez JM, Karpen GH and Orr-Weaver TL (2000) Sister-chromatid cohesion via MEI-S332 and kinetochore assembly are separable functions of the Drosophila centromere. *Current Biology*, **10**, 997–1000.

Maison C, Bailly D, Peters AH, Quivy JP, Roche D, Taddei A, Lachner M, Jenuwein T and Almouzni G (2002) Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature Genetics*, **30**, 329–334.

Nakayama J, Rice JC, Strahl BD, Allis CD and Grewal SI (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, **292**, 110–113.

Ohzeki J, Nakano M, Okada T and Masumoto H (2002) CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *The Journal of Cell Biology*, **159**, 765–775.

Pal-Bhadra M, Leibovitch BA, Gandhi SG, Rao M, Bhadra U, Birchler JA and Elgin SC (2004) Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. *Science*, **303**, 669–672.

Partridge JF, Borgstrom B and Allshire RC (2000) Distinct protein interaction domains and protein spreading in a complex centromere. *Genes & Development*, **14**, 783–791.

Partridge JF, Scott KS, Bannister AJ, Kouzarides T and Allshire RC (2002) cis-acting DNA from fission yeast centromeres mediates histone H3 methylation and recruitment of silencing factors and cohesin to an ectopic site. *Current Biology*, **12**, 1652–1660.

Provost P, Dishart D, Doucet J, Frendewey D, Samuelsson B and Radmark O (2002) Ribonuclease activity and RNA binding of recombinant human Dicer. *The EMBO Journal*, **21**, 5864–5874.

Reinhart BJ and Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. *Science*, **297**, 1831.

Rudert F, Bronner S, Garnier JM and Dolle P (1995) Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, **6**, 76–83.

Schotta G, Ebert A, Krauss V, Fischer A, Hoffmann J, Rea S, Jenuwein T, Dorn R and Reuter G (2002) Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *The EMBO Journal*, **21**, 1121–1131.

Schramke V and Allshire R (2003) Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science*, **301**, 1069–1074.

Sullivan BA, Blower MD and Karpen GH (2001) Determining centromere identity: cyclical stories and forking paths. *Nature Reviews. Genetics*, **2**, 584–596.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, Grewal SI and Moazed D (2004) RNAi-mediated targeting of heterochromatin by the RITS complex. *Science*, **303**, 672–676.

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI and Martienssen RA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, **297**, 1833–1837.

Basic Techniques and Approaches

Techniques in genomic imprinting research

Todd A. Gray
Wadsworth Center, The Genomics Institute, Troy, NY, USA

1. RT-PCR of transcribed polymorphisms

Demonstration of parentally determined monoallelic transcription represents the gold standard in imprinted gene evaluation. One prerequisite for this analysis is that individuals being tested have a sequence difference, or polymorphism, in the candidate gene that can be used to distinguish the maternally and paternally derived alleles. For example, a "G" may be found in a specific position in a candidate gene from the mother, and a "T" in this same position in the gene from the father (Figure 1a). If the polymorphism is located in an exon of the gene, then the parental origin of the mRNA transcripts can be distinguished (Figure 1b). mRNAs prepared from such an informative individual are reverse transcribed (RT) into matching complementary DNA (cDNA), and are amplified by the polymerase chain reaction (PCR) to allow analysis of this gene's mRNAs out of the thousands of others in the original sample. Amplified cDNA products from the pool of mRNA transcripts subjected to this RT-PCR approach are inspected, usually by direct sequencing or by sequencing of >10 individual clones, to determine whether both, or only one, of the polymorphisms is present in the sample (Bonthron *et al*., 2000; Yin *et al*., 2004). Recovery of approximately equivalent levels of each polymorphism indicates biallelic expression, while the finding of only one of the two genomic polymorphisms in the cDNA pool suggests monoallelic expression consistent with imprinting (Figure 1b).

Since monoallelic expression can result from other epigenetic phenomena (e.g., X-inactivation (Lyon, 1999), mutual exclusion (Serizawa *et al*., 2000), aberrant gene silencing (Esteller, 2002)) or genetic alterations (e.g., promoter mutations, mRNA destabilizing changes such as certain premature protein truncation mutations (Frischmeyer and Dietz, 1999)), independent verification is necessary. For humans, verification usually means examination of additional parent/child subjects for other informative polymorphisms, or analysis of reciprocal transmission of the silenced allele from a parent of the other sex. In lab animals, such as mice, verification is systematically accomplished by reciprocal breeding in which the maternal strain in one mating is used as the paternal strain in another mating; if the monoallelic expression is due to imprinting, the gene from that strain should be actively transcribed in the pups from one cross, but silenced in the other.

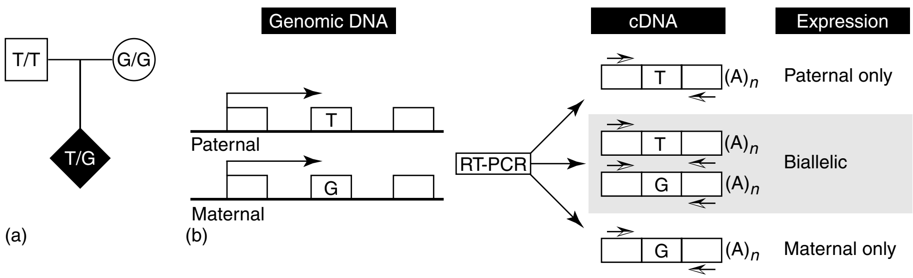


Figure 1 Assessment of monoallelic transcription by PCR amplification of expressed polymorphisms. (a) A simplified pedigree showing a homozygous T/T father, G/G mother, and heterozygous T/G offspring for a hypothetical candidate gene. (b) RT-PCR amplification of the candidate gene's mRNA products. Each parental genomic DNA structure is shown as three exons, with the embedded T or G polymorphism. Following reverse transcription of the mRNA, the cDNA is PCR-amplified and analyzed for the presence of the T or G polymorphism

Often, RNA samples are not available, thus precluding the relatively definitive RT-PCR analysis. DNA samples, in contrast, are much more widely available. Fortunately, DNA methylation of cytosine at cytosine/guanosine nucleotide tandems (CpG dinucleotides) near the start of a gene correlates with expression, such that actively transcribed genes are associated with unmethylated CpGs, whereas methylated CpG dinucleotides are associated with transcriptionally silent genes. Traditional methods of examining CpG methylation utilized restriction enzymes that could only cut DNA whose cognate sites had unmethylated CpGs (Kubota *et al*., 1996). This method is reproducible and unambiguous, but it requires a comparatively large amount of DNA, suitable restriction sites in the region of interest, and radioactive probes, and it has limited throughput. Current methods use the power of PCR amplification for similar analyses that bypass some of these drawbacks.

PCR amplification alone, however, cannot discriminate methylated from unmethylated template DNA. The epigenetic information contained in the cytosine methylation pattern of genomic DNA must first be converted into genetic, or sequence, information. To this end, bisulfite-conversion methodologies have been refined and implemented (Clark *et al*., 1994; Frommer *et al*., 1992; Herman *et al*., 1996; Sadri and Hornsby, 1996). The basic procedure is predicated on the conversion of cytosine, but not methyl-cytosine, to uracil upon exposure to high concentrations of sodium bisulfite, under certain conditions. All non-CpG cytosines will be converted to uracil, as will unmethylated CpG cytosines, but methylated CpG cytosines will remain unchanged (Figure 2). DNA that has been bisulfite-converted is used as a template for PCR amplification, with the newly created uracils recognized as thymidines.

The bisulfite-conversion process has harsh effects on the DNA (Grunau *et al*., 2001), and the loss of nearly all of the cytosines in the template DNA impairs the robustness and specificity of PCR amplification (Warnecke *et al*., 2002). Two rounds of amplification are therefore common, in which amplification is initially performed with primers that do not contain CpGs, and therefore do not discriminate between methylated and unmethylated samples. This round is intended to amplify the target sequence as template for a second round of amplification with primers

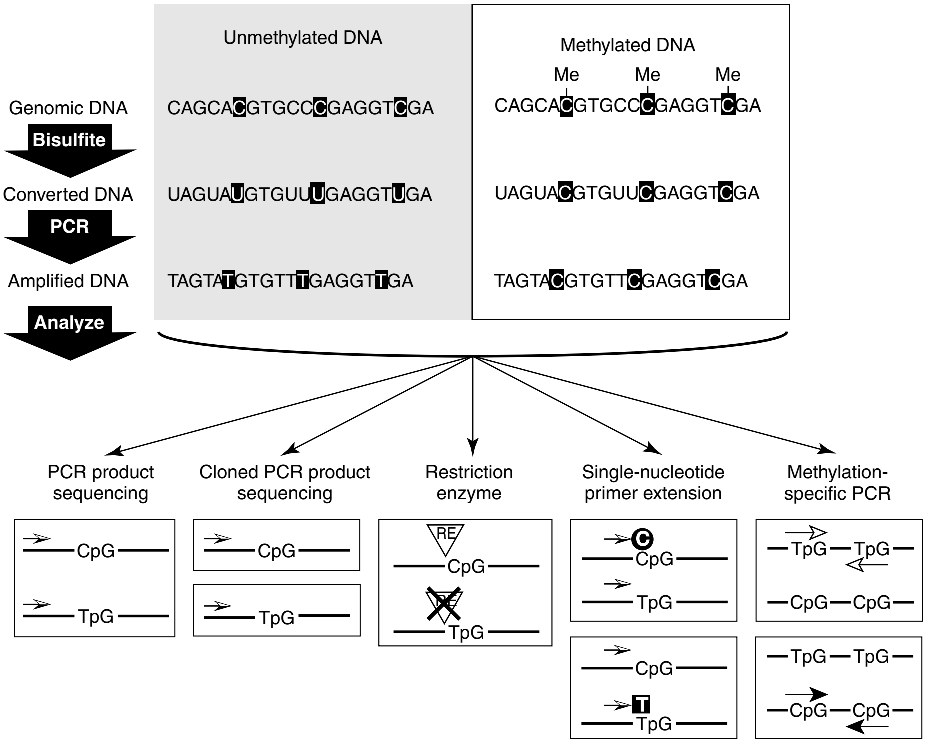


Figure 2 Bisulfite conversion and PCR evaluation of cytosine methylation in genomic DNA. Subjecting genomic DNA to a bisulfite-modification procedure converts cytosines to uracils, while methyl-cytosines are protected from conversion. Two successive rounds of amplification (nested PCR) are typically performed, to enhance sensitivity. Comparison of the amplified products for converted thymidines and protected cytosines reveals the methylation profile of the original template. Simplified schematic diagrams illustrating the basis of each assay are shown below, with each rectangular box representing individual reactions

that lie within the amplified interval. The resulting products can be analyzed for the presence of a thymidine or cytosine at CpGs in the candidate gene interval using many common genetic analysis tools (Figure 2). As with the first primer set, the second primer set may be nondiscriminatory, so that it will amplify all target fragments, regardless of original methylation status. These products are analyzed through: (1) sequencing of the PCR products either directly or as a subset of cloned PCR products (Frommer *et al*., 1992), (2) restriction enzyme analysis that will only cut converted DNA in which a specific cytosine has been either retained or transformed, but not both (Sadri and Hornsby, 1996; Xiong and Laird, 1997), and (3) single-nucleotide primer extension (SNuPE), in which an internal primer hybridizes adjacent to a specific CpG dinucleotide, and is incubated with a polymerase extension mix that contains only either labeled thymidine or cytosine dideoxynucleotides to generate single base extension products only from unmethylated or methylated templates, respectively (Gonzalgo and Jones, 1997). As an alternative, two sets of second-round primers may be used that discriminate between cytosine- and thymidine-containing templates. In this methylation-specific

PCR (Herman *et al*., 1996), an "unmethylated" primer set contains thymidines where all cytosines had originally been, and a "methylated" primer set contains cytosines at CpG dinucleotides. No one method is inherently the best, and which one is used will depend on technical resources and expertise, as well as features of the individual gene being assayed. Higher-throughput approaches are also being devised that facilitate the simultaneous analysis of many samples (Akey *et al*., 2002; Cottrell *et al*., 2004; Eads *et al*., 2000; Rand *et al*., 2002).

While the methods outlined above address the methylation status at specified genomic loci, screening procedures have been developed that first identify a difference in methylation, followed by locus identification. These procedures are remarkably diverse and employ methyl-sensitive restriction enzymes, bisulfite treatment, electrophoresis, or microarray platforms (Laird, 2003).

Definition of the profile of cytosines and thymidines in bisulfite-converted DNA allows the epigenetic methylation profile of the target sequence in the original sample to be inferred. Regardless of the assay used, a hallmark of imprinted genes is that they predictably exhibit a 50% (differential) methylation profile, with the copy from one parent being transcribed and generally unmethylated, and the copy from the other parent being transcriptionally silent and hypermethylated. In addition to their use in detection of imprinting, bisulfite-conversion methodologies are applicable to epigenetic analyses of cancer, aging, and mammalian cloning. As epigenetic profiling continues to gain prominence, new, higher-throughput, and more cost-effective technologies are likely to be developed, facilitating their routine application in both research and clinical settings.

References

Akey DT, Akey JM, Zhang K and Jin L (2002) Assaying DNA methylation based on high-throughput melting curve approaches. *Genomics*, **80**, 376–384.

Bonthron DT, Hayward BE, Moran V and Strain L (2000) Characterization of TH1 and CTSZ, two non-imprinted genes downstream of GNAS1 in chromosome 20q13. *Human Genetics*, **107**, 165–175.

Clark SJ, Harrison J, Paul CL and Frommer M (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, **22**, 2990–2997.

Cottrell SE, Distler J, Goodman NS, Mooney SH, Kluth A, Olek A, Schwope I, Tetzner R, Ziebarth H and Berlin K (2004) A real-time PCR assay for DNA-methylation using methylation-specific blockers. *Nucleic Acids Research*, **32**, e10.

Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV and Laird PW (2000) MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Research*, **28**, E32.

Esteller M (2002) CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, **21**, 5427–5440.

Frischmeyer PA and Dietz HC (1999) Nonsense-mediated mRNA decay in health and disease. *Human Molecular Genetics*, **8**, 1893–1900.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL and Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 1827–1831.

Gonzalgo ML and Jones PA (1997) Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Research*, **25**, 2529–2531.

Grunau C, Clark SJ and Rosenthal A (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Research*, **29**, E65–E65.

Herman JG, Graff JR, Myohanen S, Nelkin BD and Baylin SB (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 9821–9826.

Kubota T, Aradhya S, Macha M, Smith AC, Surh LC, Satish J, Verp MS, Nee HL, Johnson A, Christan SL, *et al*. (1996) Analysis of parent of origin specific DNA methylation at SNRPN and PW71 in tissues: implication for prenatal diagnosis. *Journal of Medical Genetics*, **33**, 1011–1014.

Laird PW (2003) The power and the promise of DNA methylation markers. *Nature Reviews Cancer*, **3**, 253–266.

Lyon MF (1999) X-chromosome inactivation. *Current Biology*, **9**, R235–R237.

Rand K, Qu W, Ho T, Clark SJ and Molloy P (2002) Conversion-specific detection of DNA methylation using real-time polymerase chain reaction (ConLight-MSP) to avoid false positives. *Methods*, **27**, 114–120.

Sadri R and Hornsby PJ (1996) Rapid analysis of DNA methylation using new restriction enzyme sites created by bisulfite modification. *Nucleic Acids Research*, **24**, 5058–5059.

Serizawa S, Ishii T, Nakatani H, Tsuboi A, Nagawa F, Asano M, Sudo K, Sakagami J, Sakano H, Ijiri T, *et al*. (2000) Mutually exclusive expression of odorant receptor transgenes. *Nature Neuroscience*, **3**, 687–693.

Warnecke PM, Stirzaker C, Song J, Grunau C, Melki JR and Clark SJ (2002) Identification and resolution of artifacts in bisulfite sequencing. *Methods*, **27**, 101–107.

Xiong Z and Laird PW (1997) COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Research*, **25**, 2532–2534.

Yin D, Xie D, De Vos S, Liu G, Miller CW, Black KL and Koeffler HP (2004) Imprinting status of DLK1 gene in brain tumors and lymphomas. *International Journal of Oncology*, **24**, 1011–1015.

Bioinformatics and the identification of imprinted genes in mammals

Melissa J. Fazzari and John M. Greally
Albert Einstein College of Medicine, Bronx, NY, USA

Currently, we know of the existence of tens of imprinted genes in human and mouse (Greally, 2002), representing only a fraction of the hundreds of imprinted genes that are predicted (Reik and Walter, 2001). It is technically difficult to prove that a gene is imprinted (Suda *et al*., 2003), as the assay has to be able to quantify the relative amount of expression from the paternal and maternal alleles of the gene, and may require isolating a specific cell type (Yamasaki *et al*., 2003) or developmental stage (Moore *et al*., 2001) for analysis (*see* Article 44, **Techniques in genomic imprinting research**, Volume 1 and Article 46, **UPD in human and mouse and role in identification of imprinted loci**, Volume 1). In the meantime, population geneticists have begun to recognize the increasing numbers of genetic diseases with parent-of-origin effects on their inheritance (Green *et al*., 2002; Karason *et al*., 2003; McInnis *et al*., 2003; Pezzolesi *et al*., 2004; Strauch *et al*., 2001; *see also* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1). When faced with a region of peak linkage, often millions of base pairs in size and containing dozens of genes, candidates for targeted analysis might currently be chosen on the basis of their functional properties. If parent-of-origin effects were noted, a further useful indicator of the gene responsible for the disease would be a characteristic predictive of its likelihood of imprinting.

Whole-genome approaches to predict imprinted gene status have been performed in two ways. Direct molecular analysis has been performed to try to identify loci at which gene expression (Nikaido *et al*., 2003) or cytosine methylation (Shibata *et al*., 1994) patterns differ on the paternal and maternal chromosomes. The focus of this review is the alternative, bioinformatic approach, based on our appreciation that imprinted genes have discriminatory sequence characteristics (Greally, 2002; Ke *et al*., 2002a; Ke *et al*., 2002b). However, as we stress again later, these are complementary, not competing approaches – concordance of predictions by multiple independent approaches would be expected to be more powerful than individual predictions alone.

The bioinformatic prediction of imprinted genes is predicated on the presence of sequence features at imprinted loci that discriminate them from nonimprinted loci. Our study of the 100-kb context of imprinted promoters revealed an unexpected

dearth of short interspersed nuclear elements (SINEs) (Greally, 2002). Nonimprinted loci can accumulate SINEs to similar low levels, so a comparison based on this sequence feature alone is very nonspecific. Other sequence features, in conjunction with our knowledge of SINE content, may further distinguish imprinted genes. Therefore, it becomes necessary to study imprinting in a multivariable setting, using statistical models that can accommodate more than one feature. By identifying discriminatory sequence features, these models provide insight (both mechanistic and evolutionary) into the underlying biology of the system, as well as a means to help predict whether imprinting is occurring at the remaining genes in the genome.

Statistical modeling allows each gene of unknown status to be assigned a predictive score, or relative likelihood of imprinting. We attempt to summarize sequence content information simultaneously through a multivariable model, which uses these characteristics as inputs and weights them on the basis of their relative effects on the probability of imprinting. Logistic regression modeling, a well-studied and powerful approach for binary outcome data (Hosmer and Lemeshow, 1989), models the log odds of imprinting probability using a linear combination of the predictors. Other methods may be used to generate predictive models (Bishop, 1995; Breiman *et al*., 1984; Fisher, 1936), but performance has been shown to vary with respect to different datasets examined, with no method consistently outperforming the others.

The statistical challenges in the prediction of imprinting are numerous. The small number of known imprinted genes (Greally, 2002) limits the size and complexity of the model that can be estimated. Adding features will always allow a better fit of the model to the data in hand, but this perfect fit will almost certainly mean poor predictions when applied to new samples of genes. This poor generalization is due to the model's fitting of each nuance in the data, rather than capturing only those effects that will be consistent from sample to sample. At the other extreme, low-complexity models that underfit the data may be quite biased. In this application, the number of candidate predictors that can be mined from sequence data annotations (such as those at the UCSC Genome Browser **http://genome.ucsc.edu/**) is in the hundreds, with varying levels of correlation. It is likely that the majority of these features have small to negligible effects. Model selection techniques must therefore be employed to find models that capture the bulk of the information available, without overfitting. Validation of the model is difficult since additional test samples are nonexistent; we must rely on computer-intensive methods (Hjorth, 1994) to gain a sense of predictive ability.

We show the principle of this approach in the following analysis. We used a logistic regression model to predict imprinting, mining DNA sequence annotations (repetitive element cumulative number and size, CpG island number) for the 10- and 100-kb regions flanking the transcription start sites for 28 human imprinted genes and 300 randomly chosen control loci. Model selection was performed using the "all subsets" approach and the Akaike's Information Criterion (AIC) (Akaike, 1973) as a measure of model information. This criterion is a useful measure of the model's predictive ability, while providing a penalty for model complexity. The model that we derived is representative of many competing imprinting models and is a function of five sequence characteristics:

Alu SINEs	(100 kb)
CR1 LINEs	(100 kb)
CR1 LINEs	(10 kb)
Low-complexity repeats	(10 kb)
Tip100 DNA elements	(100 kb)

The previous report of SINE exclusion from the 100-kb flanking promoters of imprinted genes (Greally, 2002) is confirmed with this analysis, while additional sequence features are also found to discriminate imprinted loci (the ancient chicken repeat 1 (CR1) LINEs, low-complexity repeats, and the Tip100 DNA transposon). To evaluate this model, general measures of predictive power may be obtained (Agresti, 1990). In the reported model, the area under the Receiver Operating Curve (ROC) is 0.91. The generalized R-square statistic (based on the log-likelihood) is 0.42. On the basis of these measures, we can reasonably conclude that sequence characteristics contain information with respect to imprinting status and that the model may be used as an analytic tool to highlight genes with the highest evidence of imprinting.

Predictive scores derived from the logistic model may be used to rank genes on the basis of imprinting evidence (Figure 1). This analytic approach is common in many areas of research, and has been successfully applied in clinical trials to

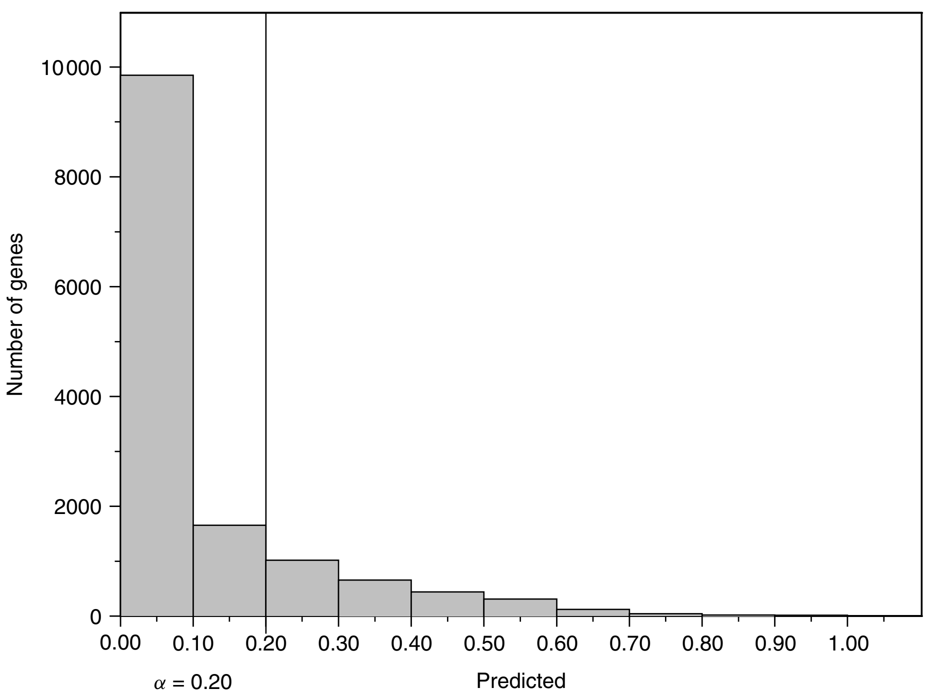


Figure 1 Genome-wide distribution of predicted scores based on logit model. Genes with scores below a threshold level (α) have the lowest probability of imprinting based on sequence features. Sensitivity and specificity of predictions are direct functions of the threshold specified

stratify patients on the basis of risk (Gail and Costantino, 2001). Consideration of the highest-ranked genes may offer information that, in conjunction with other discovery methods, quickly and systematically finds new imprinted genes.

On the basis of the sequence characteristics of the regions flanking gene promoters, we can assign a relative likelihood of imprinting to every gene in the genome. There are several ways by which we can improve our predictive capabilities, however. It is striking that certain promoters exist embedded within imprinted domains that escape imprinting (Hayward *et al*., 1998; Vu and Hoffman, 1994), indicating that flanking sequences are not the sole genomic determinant of imprinting. Identification of the sequences that discriminate imprinted from nonimprinted promoters may add a lot of power to bioinformatically based predictions of imprinted genes. As the human and mouse genomes become better annotated, especially with regard to promoter sites, that will allow much better sample assembly. The integration of bioinformatic with molecular predictions should also allow us to explore how much concordance exists between these datasets. In addition, novel statistical methods in model selection can be explored to further identify and characterize informative sequence features. As a final point, predictions are just predictions – how well they perform will be apparent only with testing, which is beyond the scope of any single laboratory, so these predictions need to be made publicly available for testing by investigators interested in different regions of the genome.

References

Agresti A (1990) *Categorical Data Analysis*, Wiley: New York.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, Petrov BN and Csaki F (Eds), Akademiai Kiado: Budapest, pp. 267–281.

Bishop CM (1995) *Neural Networks for Pattern Recognition*, Clarendon Press: Oxford.

Breiman L, Friedman JH, Olshen RA and Stone CJ (1984) *Classification and regression trees*, Wadsworth Statistical Press: Belmont.

Fisher RA (1936) The use of multiple measurements on taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Gail MH and Costantino JP (2001) Validating and improving models for projecting the absolute risk of breast cancer. *Journal of the National Cancer Institute*, **93**, 334–335.

Greally JM (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 327–332.

Green J, O'Driscoll M, Barnes A, Maher ER, Bridge P, Shields K and Parfrey PS (2002) Impact of gender and parent of origin on the phenotypic expression of hereditary nonpolyposis colorectal cancer in a large newfoundland kindred with a common MSH2 mutation. *Diseases of the Colon and Rectum*, **45**, 1223–1232.

Hayward BE, Moran V, Strain L and Bonthron DT (1998) Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 15475–15480.

Hjorth U (1994) *Computer Intensive Statistical Methods*, Chapman and Hall: London.

Hosmer DW and Lemeshow S (1989) *Applied Logistical Regression*, Wiley: New York.

Karason A, Gudjonsson JE, Upmanyu R, Antonsdottir AA, Hauksson VB, Runasdottir EH, Jonsson HH, Gudbjartsson DF, Frigge ML, Kong A, *et al*. (2003) A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting. *American Journal of Human Genetics*, **72**, 125–131.

Ke X, Thomas NS, Robinson DO and Collins A (2002a) A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes. *Human Genetics*, **111**, 511–520.

Ke X, Thomas SN, Robinson DO and Collins A (2002b) The distinguishing sequence characteristics of mouse imprinted genes. *Mammalian Genome*, **13**, 639–645.

McInnis MG, Lan TH, Willour VL, McMahon FJ, Simpson SG, Addington AM, MacKinnon DF, Potash JB, Mahoney AT, Chellis J, *et al*. (2003) Genome-wide scan of bipolar disorder in 65 pedigrees: supportive evidence for linkage at 8q24, 18q22, 4q32, 2p12, and 13q12. *Molecular Psychiatry*, **8**, 288–298.

Moore GE, Abu-Amero SN, Bell G, Wakeling EL, Kingsnorth A, Stanier P, Jauniaux E and Bennett ST (2001) Evidence that insulin is imprinted in the human yolk sac. *Diabetes*, **50**, 199–203.

Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M, *et al*. (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Research*, **13**, 1402–1409.

Pezzolesi MG, Nam M, Nagase T, Klupa T, Dunn JS, Mlynarski WM, Rich SS, Warram JH and Krolewski AS (2004) Examination of candidate chromosomal regions for type 2 diabetes reveals a susceptibility locus on human chromosome 8p23.1. *Diabetes*, **53**, 486–491.

Reik W and Walter J (2001) Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, **2**, 21–32.

Shibata H, Hirotsune S, Okazaki Y, Komatsubara H, Muramatsu M, Takagi N, Ueda T, Shiroishi T, Moriwaki K, Katsuki M, *et al*. (1994) Genetic mapping and systematic screening of mouse endogenously imprinted loci detected with restriction landmark genome scanning method (RLGS). *Mammalian Genome*, **5**, 797–800.

Strauch K, Bogdanow M, Fimmers R, Baur MP and Wienker TF (2001) Linkage analysis of asthma and atopy including models with genomic imprinting. *Genetic Epidemiology*, **21**(Suppl 1), S204–S209.

Suda T, Katoh M, Hiratsuka M, Fujiwara M, Irizawa Y and Oshimura M (2003) Use of real-time RT-PCR for the detection of allelic expression of an imprinted gene. *International Journal of Molecular Medicine*, **12**, 243–246.

Vu TH and Hoffman AR (1994) Promoter-specific imprinting of the human insulin-like growth factor-II gene. *Nature*, **371**, 714–717.

Yamasaki K, Joh K, Ohta T, Masuzaki H, Ishimaru T, Mukai T, Niikawa N, Ogawa M, Wagstaff J and Kishino T (2003) Neurons but not glial cells show reciprocal imprinting of sense and antisense transcripts of Ube3a. *Human Molecular Genetics*, **12**, 837–847.

UPD in human and mouse and role in identification of imprinted loci

Aaron P. Theisen
Health Research and Education Center, Washington State University, Spokane, WA, USA

Lisa G. Shaffer
Sacred Heart Medical Center, Washington State University, Spokane, WA, USA

Although most mammalian autosomal genes are biallelically expressed, a small number are functionally haploid with one of the parental contributions silenced while the homologous allele is active. "Genomic imprinting" refers to an epigenetic mark of the DNA, or associated proteins, that stably switch genes "on" or "off" depending on their parent of origin. Thus, genomic imprinting refers to a normal state of genomic imbalance. The search for imprinted genes began with the serendipitous detection of genomic alterations that unmasked candidate loci through their overexpression or lack of expression.

The pioneering nuclear transfer technique developed by Surani *et al*. (1984) confirmed that mammalian development required the contribution of both the maternal and paternal genomes. Exploiting the separation of the maternal and paternal nuclei shortly after fertilization, researchers removed the sperm-derived pronucleus and replaced it with a second egg-derived pronucleus. The gynogenetic embryos failed to develop normally, as did the alternative experiment with androgenetic embryos. The authors concluded that genomic imprinting occurred during gametogenesis such that the absence of contribution of active genes from one of the parental genomes resulted in the failure of proper embryonic development.

The first suggestion that specific chromosomal regions contained imprinted genes necessary for proper development came from earlier work with uniparental disomic (*see* Article 19, **Uniparental disomy**, Volume 1) mice, which received two copies of a chromosome or chromosomal region from one parent with no contribution from the other parent. The process of generating mice with whole-chromosome or partial uniparental disomies (UPDs), termed "translocation intercrossing", was developed by Snell in 1946. Because gametes carrying a translocation are susceptible to nondisjunction (*see* Article 16, **Nondisjunction**, Volume 1), if two mice carrying the same translocation are intercrossed, a subset of offspring should receive

two complementary aneuploid (either missing or carrying an extra chromosome involved in the translocation) gametes, forming a viable, chromosomally balanced zygote. If one of the mice involved in the intercross carries homozygous alleles for a suitable marker gene, the parental origin of the whole-chromosome or partial UPD can be identified (Snell, 1946).

Intercrosses of mice carrying reciprocal translocations (the exchange of chromosomal segments between nonhomologous chromosomes) are complicated by the types of chromosome segregation required to produce UPDs either proximal or distal to the translocation breakpoint. Balanced zygotes with duplications of regions distal to the translocation breakpoint require the relatively common adjacent-1 segregation of gametes; up to 16% of zygotes will mature into viable offspring. However, offspring with duplications proximal (centromeric) to the translocation breakpoint require the much rarer adjacent-2 segregation of gametes; recovery rates of these types of offspring are much lower than for distal-duplication mice, with 5% or less of zygotes resulting in live offspring (Searle *et al*., 1971). Based on the methods of calculating nondisjunction rates for mice with reciprocal translocations, Lyon *et al*. (1976) estimated the recovery rate for the offspring of mice with Robertsonian translocations (the centric translocations of the long arms of two chromosomes) to be 0–5%; the majority of zygotes will be unbalanced (monosomic or trisomic) and usually die preterm.

By comparing the percentage of viable offspring with that expected of adjacent-1 or -2 segregation, mouse reciprocal translocations could be used to identify the location of the centromere of one of the chromosomes involved relative to the translocation breakpoint (Searle *et al*., 1971). However, in some studies, the homozygous-marker offspring failed to appear in the litter or died shortly after birth. The first evidence that “complementation errors” might arise as a result of uniparental disomy for a chromosome region occurred in studies of the hairpin-tail T^{hp} allele on mouse chromosome 17, whereby inheritance of the allele from the mother resulted in death at birth or in utero, whereas paternal inheritance of the allele produced viable offspring (Johnson, 1974). Further studies by Searle and Beechey (1978) demonstrated the noncomplementation and resultant abnormal phenotypes of maternal duplications (and complementary paternal deficiencies) of distal mouse chromosomes 2, 7, and 8.

Following the successes of the early nuclear transfer experiments, researchers, using mice carrying translocations, discovered that maternal and paternal contributions of some mouse chromosomal regions could produce opposite growth effects: mice with Robertsonian translocations involving chromosomes 11 and 13 revealed that maternal UPDs of chromosome 11 resulted in offspring smaller than their littermates, whereas paternal UPDs of chromosome 11 produced mice significantly larger than their littermates. Reciprocal translocations involving chromosomes 2 and 11 confirmed that the growth effects were of proximal chromosome 11 origin (Cattanach and Kirk, 1985).

Although nuclear transfer experiments have confirmed that the parental genomes make differential contributions to development and narrowed the “imprinting window” in which these epigenetic marks are established, the search for imprinted loci has relied on imprinting map positions derived from studies of mouse whole-chromosome and partial UPDs. The imprinting maps, previously updated

annually and now available on-line (**http://www.mgu.har.mrc.ac.uk/research/imprinting**, 2005) indicate those areas of the mouse genome that have demonstrated differential growth or developmental defects in mice with UPDs.

Imprinted chromosomal regions may compose 10–25% of the mouse genome (Cattanach and Beechey, 1997), although estimates of the number of individual imprinted genes suggest that these constitute a lower percentage of all genes. Although a similar estimate has not been made for humans, the high homology between the mouse and human genomes has allowed researchers to identify those genes that are imprinted in the mouse genome that should display phenotypic effects in humans (**http://www.mgu.har.mrc.ac.uk**). Confirmation of the phenotypic effects, if any, of these imprinted loci depends on the discovery of rare individuals with UPDs of the corresponding chromosome.

Two individuals with maternal disomy for chromosome 15 and Prader–Willi syndrome (PWS) (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1), a disorder characterized by obesity and mental retardation, provided the first demonstration that UPD could be associated with human diseases arising from the imbalance of imprinted loci (Nicholls *et al*., 1989). Comparisons of the phenotypes of UPDs with those regions of the human and mouse genomes thought to contain imprinted loci have identified regions of genomic imprinting for chromosomes 6, 7, 11, 14, and 15 and phenotypes suggestive of imprinting effects on chromosomes 2, 9, 16, and 20 (Ledbetter and Engel, 1995; Kotzot, 1999). Although the same experiments that generate UPDs in mice cannot be performed in humans, the relatively high incidence of Robertsonian translocations in the general population (∼1:1000), coupled with the high proportion of structural chromosomal abnormalities resulting in UPD that are Robertsonian translocations (Shaffer, 2003), illustrates the importance of identifying imprinted loci in humans.

The search for imprinted loci using UPDs in mice and humans relies on several assumptions. First, although most imprinted genes are conserved between humans and mice, expression profiles have identified several that are not. For example, *U2afbp-rs* is imprinted in mice but not in humans, and conflicting reports suggest that *Igf2r*, which is imprinted in mice, may be biallelically expressed in humans (Kalscheuer *et al*., 1993). Furthermore, researchers have only been able to identify those regions in mice with easily recognizable phenotypes. Lethal imprinting effects that may be construed as nonheritable disease alleles due to a lack of observed offspring provide a confounding corollary for investigations into human imprinted loci.

Nonetheless, the search for imprinted loci in mice and humans using UPDs may continue to reveal imprinting regions in both species. Several chromosomes that display parent-of-origin effects have yet to have genes mapped to those regions (**http://www.mgu.har.mrc.ac.uk**), and UPDs for a few mouse chromosomes have not been generated or thoroughly investigated owing to a lack of suitable translocations or marker genes. Likewise, confirmation of regions of homologous imprinting loci in humans awaits the ascertainment of UPDs for those chromosomes, including chromosomes 18 and 19 (Shaffer, 2003). Just as the identification of disease-causing genes often awaits the ascertainment of the rare chromosomal rearrangement that disrupts their function (*see* Article 11, **Human**

cytogenetics and human chromosome abnormalities, Volume 1), the search for imprinted loci and the deleterious effects of their misexpression relies upon the discovery of the rare case of UPD that unmasks their existence.

References

Cattanach BM and Beechey CV (1997) Genomic imprinting in the mouse: possible final analysis. In *Genomic Imprinting: Frontiers in Molecular Biology*, Vol. 18, Reik W and Surani A (Eds.), IRC Press, Oxford University Press: Oxford, pp. 118–145.

Cattanach BM and Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. *Nature*, **315**, 496–498.

Johnson DR (1974) Hairpin-tail: a case of post-reductional gene action in the mouse egg? *Genetics*, **76**, 795–805.

Kalscheuer VM, Mariman EC, Schepens MT, Rehder H and Ropers HH (1993) The insulin-like growth factor type-2 receptor gene is imprinted in the mouse but not in humans. *Nature Genetics*, **5**, 74–78.

Kotzot D (1999) Abnormal phenotypes in uniparental disomy (UPD): fundamental aspects and a critical review with bibliography of UPD other than 15. *American Journal of Medical Genetics*, **82**, 265–274.

Ledbetter DH and Engel E (1995) Uniparental disomy in humans: development of an imprinting map and its implications for prenatal diagnosis. *Human Molecular Genetics*, **4**, 1757–1764.

Lyon MF, Ward HC and Simpson GM (1976) A genetic method for measuring non-disjunction in mice with Robertsonian translocations. *Genetical Research*, **26**, 283–295.

Mammalian Genetics Unit, Medical Research Council, Harwell, Oxfordshire OX11 0RD, UK, **http://www.mgu.har.mrc.ac.uk/research/imprinting**, 2005.

Nicholls RD, Knoll JH, Butler MG, Karam S and Lalande M (1989) Genetic imprinting suggested by maternal heterodisomy in non-deletion Prader-Willi syndrome. *Nature*, **342**, 281–285.

Searle AG and Beechey CV (1978) Complementation studies with mouse translocations. *Cytogenetics and Cell Genetics*, **20**, 282–303.

Searle AG, Ford CE and Beechey CV (1971) Meiotic nondisjunction in mouse translocations and the determination of centromere position. *Genetical Research*, **18**, 215–235.

Shaffer LG (2003) Uniparental disomy: mechanisms and clinical consequences. *Fetal and Maternal Medicine Review*, **14**, 155–175.

Snell GD (1946) An analysis of translocations in the mouse. *Genetics*, **31**, 157–180.

Surani MA, Barton SC and Norris ML (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, **308**, 548–550.

Introductory Review

Introduction to gene mapping: linkage at a crossroads

Nancy J. Cox
The University of Chicago, Chicago, IL, USA

1. Introduction

The success of gene mapping in Mendelian disorders has been remarkable; the general paradigm of linkage mapping followed by positional cloning has become near-foolproof for identifying the genetic variation responsible for disorders with a simple relationship between genotype and phenotype. Both the linkage mapping and the positional cloning studies of simple, Mendelian disorders have been accomplished with increasing speed and efficiency as human genomic resources have been developed and become more affordable. But the paradigm used so successfully for Mendelian disorders is widely perceived as being notably less successful in applications to complex disorders. Complex disorders are those that have a significantly elevated risk in relatives of an affected individual as compared with disease risk for the general population, but for which simple genetic models of transmission can be excluded. Such disorders, including asthma, cardiovascular disease, diabetes, familial cancers, and psychiatric disorders, are relatively common compared with disorders having simple, Mendelian patterns of transmission, and collectively contribute substantially to public health care expenditures. Complex disorders are generally believed to arise as a consequence of the actions and interactions of many risk factors, both genetic and nongenetic. Successful identification of genetic risk factors for these disorders should improve both our understanding of the primary pathophysiology of the disorders, and ultimately our ability to treat and/or prevent them. In particular, identification of genetic risk factors may allow us to improve the design of large-scale epidemiological studies, allowing the discovery of more specific nongenetic risk factors that might be cost-effective targets for disease prevention strategies.

We, as a society, have made substantial investments in genetic science and genomic technologies, and the expectations for returns on those investments are understandably high. Given these expectations, and our experience to date in applying genetic and genomic technologies to identify and characterize the genetic component to complex phenotypes, what is the right road forward? Can we modify the gene-mapping paradigm developed for Mendelian disorders to yield better success for complex phenotypes than we have enjoyed to date? Or is gene

mapping the wrong paradigm for understanding the genetic component to complex phenotypes? In the following sections, we will consider the recent methodological improvements to gene mapping, key questions related to the measurement of complex phenotypes, and novel approaches to gene mapping. The intent is not to provide a set of directions, sending readers to one road or another, but rather to illuminate the roadway.

2. Better methods for better results?

Among the first of the challenges for mapping complex phenotypes was the recognition that genetic models for these phenotypes were unknown, a considerable contrast to the largely transparent models used for Mendelian disorders. The evolution of approaches for linkage mapping that do not require specifying genetic models and the subsequent controversies on the merits of model-free approaches to gene mapping relative to the more traditional approaches requiring knowledge and specification of genetic models are summarized in Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1. The recognition of the value of the additional information obtained through the use of multiple markers simultaneously led to the development of algorithms for multipoint linkage mapping (Lathrop *et al*., 1984; Lathrop *et al*., 1985). The simultaneous stimulation provided by the development of multipoint algorithms for linkage mapping, and the development of different approaches (parametric and model-free) for assessing the evidence for linkage led to a series of advances in the computational algorithms for linkage mapping, summarized in Article 52, **Algorithmic improvements in gene mapping**, Volume 1. The most commonly used algorithms in linkage mapping involve a trade-off between the number of markers that can be simultaneously considered and the number of individuals in the pedigree that can be used in analysis. Thus, the conventional wisdom on the optimal size and structure of families for gene mapping has often been confounded by the most recent advances in algorithmic development (Cox, 2001), and the optimal choice of population as well as family size and structure remains a subject of considerable discussion (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1).

The initial difficulties in successfully mapping genes for complex disorders led to the development of approaches allowing for considerably more sophisticated genetic complexities, such as imprinting and epigenetic phenomena (*see* Article 49, **Gene mapping, imprinting, and epigenetics**, Volume 1). Similarly, it was realized that multipoint mapping was far from perfectly informative, and that measures of information content (*see* Article 53, **Information content in gene mapping**, Volume 1) might provide additional insight into how linkage-mapping results could be improved. Sex-averaged maps have generally been used in linkage mapping studies, despite the knowledge that male and female genetic maps differ considerably. Recent mapping approaches (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1) try to make better use of information on the differences in male and female genetic maps. While genotyping accuracy has almost certainly improved since the initial use of DNA

genotypes in linkage mapping, there are a variety of approaches for identifying and accommodating genotyping errors that should improve the quality of inferences from linkage analysis. We also have a growing appreciation of the variability of the human genome outside the site polymorphisms and simple length polymorphisms that we have commonly used as genetic markers. Polymorphic inversions, duplications, and deletions affect linkage mapping and make up at least part of the genetic component to a variety of complex disorders (*see* Article 55, **Polymorphic inversions, deletions, and duplications in gene mapping**, Volume 1). Our ability to detect and characterize these cytogenetic polymorphisms will increase as we utilize denser single-nucleotide polymorphism (SNP) maps in linkage studies, although there are real trade-offs in moving from the more highly polymorphic single tandem repeat polymorphisms (STRPs) to biallelic SNPs (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1).

Have all of these improvements in the methods we use to conduct linkage-mapping studies increased our ability to localize genes for complex phenotypes? It is hard to believe that we could have had much success in mapping complex disorders using two-point parametric analysis with the sparse genetic maps of the 1980s. But, as noted in the Introduction, we are not reliably successful at linkage mapping of complex disorders even with all of the advances in analysis methods and genotyping technologies that we currently enjoy.

3. The alternative route

Genome-wide association mapping has been suggested as an alternative to linkage mapping, particularly for complex disorders in which individual genetic risk factors may be sufficiently modest to preclude detection in linkage studies (Risch and Merikangas, 1996). The technology is now available to tackle these studies (Altshuler and Clark, 2005), and early successes (e.g., Klein *et al*., 2005) will likely lead to increased use of this approach. Linkage mapping, nevertheless, retains substantial appeal. We expect the magnitude of the signals for linkage mapping to reflect the magnitude of the contribution the locus makes to the genetic component to disease. That will not be the case for genetic risk factors identified through association mapping. However challenging and difficult the linkage mapping and positional cloning studies are, they are but a prelude to the functional and physiological studies that inevitably follow. It seems easier to justify those follow-up studies when they will be conducted on the genetic risk factors making the largest contributions to disease.

Moreover, there are reasons to be optimistic about the prospects for greater success in linkage mapping as we consider more informative phenotypes. A key challenge for all genetic studies on disease phenotypes is that the diagnostic criteria reflect clinical, rather than genetic, utility. There is increased emphasis on phenotypic characterization, and it is quite possible that increased phenotypic homogeneity will frequently be associated with increased genetic homogeneity, and attendant increases in power for linkage mapping. Similarly, with careful attention to study design, the gene mapping of quantitative phenotypes, including those related to disease, seems within reach (see Blangero, 2004, for a recent review).

Finally, there is a growing recognition that disorders with substantial nongenetic risk, as exemplified by rapid changes in prevalence (e.g., diabetes, asthma, obesity), may be more amenable to mapping by focusing on aspects of the phenotype that are relevant to the ultimate gene-by-environment interaction – natural selection (*see* Article 7, **Genetic signatures of natural selection**, Volume 1). For example, polymorphisms associated with salt-sensitive hypertension have allele frequencies showing a significant correlation with latitude (Thompson *et al*., 2004), consistent with the hypothesis that spatial variability in selection pressures have shaped the allele frequencies at these loci. This suggests an alternative approach for gene mapping that focuses on the correlation of allele frequencies in native human populations with measures of environment, such as latitude, that may reflect the effects of variation in natural selection.

There will, no doubt, be further advances in genomic technology and in computational biology that will enhance our ability to identify genetic variation for complex phenotypes. With equal certainty, geneticists will continue to debate the best road to take to future gene mapping studies. At the very least, we can expect an interesting ride.

References

Altshuler D and Clark AG (2005) Genetics. Harvesting medical information from the human family tree. *Science*, **307**(5712), 1052–1053.

Blangero J (2004) Localization and identification of human quantitative trait loci: king harvest has surely come. *Current Opinion in Genetics & Development*, **14**, 233–240.

Cox NJ (2001) Computational issues in mapping variation affecting susceptibility to complex disorders: the chicken and the egg. *Theoretical Population Biology*, **60**, 221–225.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Lathrop GM, Lalouel J, Julier C and Ott J (1984) Strategies for multilocus linkage in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3443–3446.

Lathrop GM, Lalouel JM, Julier C and Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics*, **37**, 482–498.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA and Di Rienzo A (2004) CYP3A variation and the evolution of salt-sensitivity variants. *American Journal of Human Genetics*, **75**(6), 1059–1069.

Introductory Review

Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping

Joan E. Bailey-Wilson
National Human Genome Research Institute, Baltimore, MD, USA

1. Introduction

In parametric or model-based linkage analysis (often called LOD score linkage analysis (*see* Article 47, **Introduction to gene mapping: linkage at a crossroads**, Volume 1 and Article 56, **Computation of LOD scores**, Volume 1)), one assumes that models describing both the trait and marker loci are known without error. This means that one assumes that the allele frequencies, dominance relationships among the alleles, and relationships between genotypes and phenotypes are known without error at both the trait and marker loci. Nonparametric (or model-free or weakly parametric) linkage methods make fewer assumptions about the trait model, although the assumption that the marker locus model(s) is known without error is still in force. These methods of analysis have different strengths and weaknesses that should be taken into account by the analyst when choosing how to analyze linkage data. Both parametric and nonparametric analyses may be performed between a trait locus and a single marker locus (two-point linkage) or between a trait locus and a map of multiple markers (multipoint linkage). Again, these methods have different strengths and weaknesses that should be understood before embarking upon their use.

2. Parametric versus nonparametric linkage

It is well known that when both the trait and marker locus models are specified without error, then parametric linkage analysis is more powerful than nonparametric linkage and there is no inflation of Type 1 error rates asymptotically (Amos and Williamson, 1993). However, if either the trait or marker locus models are misspecified in the analysis, then the power of parametric linkage is decreased and, in some situations, may be exceeded by the power of some nonparametric methods (Amos, 1988; Williamson and Amos, 1990). Furthermore, when both the trait and the marker locus models are misspecified, the Type 1 error rate of parametric linkage analysis is increased over the nominal level (Williamson and Amos, 1990).

In situations in which no marker data are available on parents, correct specification of the trait model but incorrect specification of the marker locus model can also lead to inflated Type 1 error rates (Mandal *et al*., 1998). The effect of misspecification of marker locus models on nonparametric linkage methods has been shown to vary depending on the type of test. Tests that rely on identity-by-state methods have been shown to have inflated Type 1 error rates if marker models are misspecified (Weeks and Lange, 1988) but at least some methods that use identity-by-descent methods (such as the Haseman–Elston sib-pair regression test) have been shown to be quite robust (Mandal *et al*., 1999, 2001). All further discussion in this section will assume that the marker locus model is correctly specified in the analyses.

Many authors have shown that various nonparametric methods may be weakly parametric in that they may have maximal power under certain genetic models or that they reduce the genetic model to simpler models with fewer parameters such as regression coefficients or variance components. It has been shown for affected sib-pair (ASP) data that the mean test is equivalent to a test based on a simple recessive model LOD score, and equivalences have also been shown between particular forms of LOD scores and some forms of the maximum-likelihood score statistic (MLS) (Risch, 1990; Knapp *et al*., 1994). It has also been shown that in the presence of locus heterogeneity, this MLS statistic under Holmans' possible triangle constraints (Holmans, 1993) is locally equivalent to the heterogeneity LOD score assuming a recessive trait model (HLOD/R) and that the one-parameter MLS assuming no dominance variance is locally equivalent to a recessive model LOD score assuming homogeneity (Huang and Vieland, 2001).

Misspecification of the trait model in parametric linkage analysis has been given as the motivating reason for use of nonparametric methods. The genetic model can be misspecified in various ways, with varying effects on the outcome of the tests. First, the penetrance functions (including dominance effects) can be specified incorrectly for a qualitative trait or the dominance effects on the genotypic means and variances can be incorrectly specified for a quantitative trait. Incorrectly specifying the dominance relationships among the trait genotypes can result in substantial decreases in power, as can ignoring the possibility of nonpenetrant susceptibility allele carriers or sporadic cases (for dichotomous traits). For qualitative traits, it has been shown that essentially equivalent power can be obtained when using a nonparametric affected-pair linkage test (NPL score) or by using two parametric models, one dominant and one recessive, both allowing 50% penetrance in the susceptible genotypes and adjusting the maximum LOD score by 0.3 LOD units for the multiple testing. This latter method, termed the MMLS-C, has been shown to be both powerful and robust under a variety of complex modes of inheritance including two-locus additive and heterogeneity models (Abreu *et al*., 1999). Variations of this approach include the Mod score, in which the LOD score is maximized over both the recombination fraction and the disease-model parameters (Clerget-Darpoux *et al*., 1986) and the MLOD (Sham *et al*., 2000), in which the LOD score is maximized over the recombination fraction and over the trait-model parameters with the constraint that the penetrances must result in a given population morbid risk and that they are fully dominant or fully recessive. The MOD score has the problem that its sampling distribution under the null hypothesis is uncertain. The MLOD has been shown to have similar power to the NPL score in affected relative

pairs (and not much lower power than that of a LOD score analysis performed under the assumption of the correct genetic parameters) in a variety of situations when there is genetic homogeneity (Sham *et al*., 2000).

The genetic model for the trait may also be misspecified by assuming a single trait locus when in fact multiple loci exist. Both parametric and nonparametric methods that do not make specific allowances for multiple loci can lose power in this situation. LOD scores assuming heterogeneity (HLODs) are more powerful than LOD scores assuming homogeneity in this situation. For quantitative traits, nonparametric tests have been shown to be powerful in the detection of linkage in the presence of heterogeneity, although this power is generally lower than that of appropriate HLODs. For qualitative traits, the MMLS test has also been shown to be more powerful than some nonparametric tests, such as the NPL, in the presence of locus heterogeneity (Abreu *et al*., 1999). However, the power of some nonparametric methods, such as the NPL score, is not greatly reduced from that of the HLOD, MMLS, and MLOD in some situations in which only affected pairs are used in the tests. Thus, the relative power of these parametric and nonparametric methods for qualitative traits often depends on whether unaffected as well as affected relatives are available for study and whether the diagnosis of "unaffected" provides good information on the likely genotype at the susceptibility locus. Although it is possible to devise nonparametric scoring functions that utilize information not only on the sharing of alleles among affected individuals but also on the discordance of sharing between affected and unaffected individuals and on the sharing of alleles among unaffected individuals, many of the commonly used software packages that allow rapid calculation of multipoint nonparametric linkage statistics do not provide such scoring functions as an option. Thus, at least some of the apparent discrepancies in the literature in the relative power of parametric and nonparametric linkage methods may reflect differences in the ability of the particular implementations of these methods to utilize information on unaffected individuals in families.

When considering the application of parametric versus nonparametric tests of linkage, the availability of estimates of the recombination fraction, θ, and the proportion of families linked to the marker locus under consideration, α, are often cited as advantages of parametric tests. For the analysis of simple Mendelian disorders with 100% penetrance, no phenocopies, and a known mode of inheritance, this is a reasonable interpretation of the biological meaning of these mathematical parameters of the models. However, for complex traits with incomplete penetrance, phenocopies, multiple trait loci, and possibly incorrectly specified dominance relationships, these mathematical parameters of the models cannot be so simply interpreted. If one attempts to interpret them thus, one finds that they give biased estimates. In fact, the estimate of θ is a measure of cosegregation of genotypes of a marker locus and inferred genotypes of a trait locus conditional on an assumed model. In any multifactorial trait, the model can never accurately reflect the truth and, therefore, the estimate of θ is not an accurate estimate of the recombination fraction (Goring and Terwilliger, 2000a,b). Similarly, estimates of α are generally biased when HLODs are calculated for complex traits because this parameter is actually the prior probability that the marker will show a given correlation structure in a given family, where that correlation structure is defined by the model and the estimate of the

recombination fraction (Goring and Terwilliger, 2000a; Vieland and Logue, 2002). Thus, we can use HLODs (and MMLS versions of HLODs) as valid tests of linkage for complex traits that may be as powerful or more powerful than nonparametric tests, but we should refrain from interpreting the estimates of θ and α obtained in these analyses as accurate estimates of the recombination fraction or as the estimate of proportion of linked families. These "advantages" of parametric methods over nonparametric methods are not truly useful in the case of complex traits.

3. Two-point versus multipoint linkage

In general, linkage between a trait and a single marker locus (two-point linkage) is less powerful than linkage between a trait locus and a map of multiple marker loci (multipoint linkage) (Fisher, 1954; Lathrop *et al.*, 1985). Multipoint linkage combines the information on the transmission of a haplotype of alleles at nearby loci to better track the transmission of alleles from parents to children in each mating at each location along a chromosome. This can result in large gains in power. However, multipoint linkage requires additional assumptions that are not made in two-point linkage; that is, one assumes that the order and intermarker distances are known without error (Ott, 1999). Multipoint linkage evaluates the probability that a trait locus is located at various places along this known marker map, and thus all multipoint linkage statistics test for linkage with a 0 recombination fraction to a certain location on the map. In two-point parametric linkage, as discussed above, misspecification of the trait model may result in the inflation of the estimate of the recombination fraction, with some loss of power to detect linkage (Ott, 1999). However, this inflation of the estimate of the recombination fraction is not possible in multipoint parametric linkage, so any misspecification of the trait model, including heterogeneity, tends to decrease power to detect linkage more strongly than in two-point linkage. It is very important to allow for the possibility of genetic heterogeneity in multipoint parametric linkage analysis. In addition, misspecification of the marker map order or (to a lesser extent) intermarker distances will also decrease power to detect linkage in both parametric and nonparametric multipoint linkage (Ott and Lathrop, 1987; Ott, 1999), but these problems are not relevant to two-point linkage. Therefore, most analysts perform both two-point and multipoint linkage analyses on their data, and compare the results carefully. It is also important to estimate marker allele frequencies from the data to be analyzed, analyze separate ethnic groups that may have different allele frequencies separately, obtain the best genetic maps possible for marker loci, check for evidence of excessive double recombinants between markers since this may indicate incorrect map specification, and if necessary drop marker loci that are not well mapped.

Further reading

Olson JM, Witte JS and Elston RC (1999) Tutorial in biostatistics. Genetic mapping of complex traits. *Statistics in Medicine*, **18**, 2961–2981.

Rao DC and Province MA (2001) *Genetic Dissection of Complex Traits*, Academic Press: San Francisco.

Thomas DC (2004) *Statistical Methods in Genetic Epidemiology*, Oxford University Press: Oxford.

Whittemore AS (1996) Genome scanning for linkage: an overview. *American Journal of Human Genetics*, **59**, 704–716.

References

Abreu PC, Greenberg DA and Hodge SE (1999) Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. *American Journal of Human Genetics*, **65**, 847–857.

Amos CI (1988) *Robust methods for detection of genetic linkage for data from extended families and pedigrees*. Ph.D. Dissertation, Louisiana State University Medical Center.

Amos CI and Williamson JA (1993) Robustness of the maximum-likelihood (LOD) method for detecting linkage. *American Journal of Human Genetics*, **52**, 213–214.

Clerget-Darpoux FC, Bonaiti-Pellie C and Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, **42**, 393–399.

Fisher RA (1954) The experimental study of multiple crossing-over. *Caryologia*, **6**(Suppl), 227–231.

Goring H and Terwilliger J (2000a) Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *American Journal of Human Genetics*, **66**, 1095–1106.

Goring H and Terwilliger J (2000b) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *American Journal of Human Genetics*, **66**, 1310–1327.

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, **52**, 362–374.

Huang J and Vieland V (2001) Comparison of ‘Model-free’ and ‘Model-based’ linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications. *Human Heredity*, **51**, 217–225.

Knapp M, Seuchter SA and Baur MP (1994) Linkage analysis in nuclear families. II. Relationship between affected sib-pair tests and lod score analysis. *Human Heredity*, **44**, 44–51.

Lathrop GM, Lalouel JM, Julier C and Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics*, **37**, 482–498.

Mandal DM, Wilson AF, Keats BJB and Bailey-Wilson JE (1998) Factors affecting inflation of Type I error of model-based linkage under random ascertainment. *American Journa of Human Genetics*, **63**, A298.

Mandal DM, Wilson AF, Elston RC, Weissbecker K, Keats BJ and Bailey-Wilson JE (1999) Effects of misspecification of allele frequencies on the Type I error rate of model-free linkage analysis. *Human Heredity*, **50**, 126–132.

Mandal DM, Wilson AF and Bailey-Wilson JE (2001) Effects of misspecification of allele frequencies on the power of Haseman-Elston sib-pair linkage method for quantitative traits. *American Journal of Medical Genetics*, **103**, 308–313.

Ott J (1999) *Analysis of Human Genetic Linkage*, *Third Edition*, The Johns Hopkins University Press: Baltimore.

Ott J and Lathrop GM (1987) Estimating the position of a locus on a known map of loci. *Cytogenetics and Cell Genetics*, **46**, 674.

Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics*, **46**, 242–253.

Sham PC, Lin M-W, Zhao JH and Curtis D (2000) Power comparison of parametric and nonparametric linkage tests in small pedigrees. *American Journal of Human Genetics*, **66**, 1661–1668.

Vieland V and Logue M (2002) HLODs, trait models, and ascertainment: implications of admixture for parameter estimation and linkage detection. *Human Heredity*, **53**, 23–35.

Weeks DE and Lange K (1988) The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, **42**, 315–326.

Williamson JA and Amos CI (1990) On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genetic Epidemiology*, **7**, 309–318.

Specialist Review

Gene mapping, imprinting, and epigenetics

Konstantin Strauch
University of Bonn, Bonn, Germany

1. Introduction

Genomic imprinting, which is also called parent-of-origin effect, is an important epigenetic factor that leads to complete or partial deactivation of either the paternally or maternally inherited copy of a gene (*see* Article 26, **Imprinting and epigenetic inheritance in human disease**, Volume 1). For example, if a disease gene is subject to complete maternal imprinting, individuals who are heterozygous at the disease locus express the trait if they have inherited the mutant allele from the father, but do not do so if they have received it from the mother. Maternal imprinting is therefore equivalent to paternal expression. With imprinting, the penetrance depends on the sex of the parent who transmits a certain allele, rather than the sex of the individual who receives the allele. It should be noted that imprinting needs to be distinguished from other parental effects such as maternal-fetal interactions. Imprinting is determined by chromosomal region (see e.g., Hall, 1990; Ainscough and Surani, 1996; Bartolomei and Tilghman, 1997; Reik and Walter, 2001). Two mechanisms known to be involved in the process of imprinting are DNA methylation (*see* Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1) and the differential packing density of DNA by histone proteins (*see* Article 27, **The histone code and epigenetic inheritance**, Volume 1). Evolutionary causes that have led to the establishment of imprinting were discussed by Pardo-Manuel de Villena *et al*. (2000), de la Casa-Esperón and Sapienza (2003), as well as by Wilkins and Haig (2003) (*see* Article 37, **Evolution of genomic imprinting in mammals**, Volume 1). Beckwith–Wiedemann, Prader–Willi, and Angelman syndromes are examples of rare disorders that show a parent-of-origin effect (Falls *et al*., 1999; *see also* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1 and Article 30, **Beckwith–Wiedemann syndrome**, Volume 1). Imprinting is also known or suspected to play a role in many complex traits as well, including type I diabetes (Bain *et al*., 1994; Paterson *et al*., 1999), polycystic ovarian syndrome (Bennett *et al*., 1997), atopy (Daniels *et al*., 1996; Moffatt and Cookson, 1998; *see also* Article 61, **Allergy and asthma**, Volume 2), celiac disease (Petronzelli *et al*., 1997), preeclampsia (Esplin *et al*., 2001), cancer (Falls *et al*., 1999; *see also* Article 65, **Complexity of cancer as a genetic disease**, Volume 2), epilepsy

(Greenberg *et al*., 2000), or bipolar disorder (Stine *et al*., 1995; McInnis *et al*., 2003).

2. How can imprinting be accounted for in the context of linkage analysis?

In order to model a parent-of-origin effect, it is necessary that parent-specific genotypes are available, or otherwise can be reconstructed, such as with half-sibs. This holds for parametric and nonparametric linkage analysis of dichotomous traits, as well as for an analysis of quantitative traits.

2.1. *Parametric analysis*

In the simplest case of a dichotomous trait that is governed by one locus, the trait model comprises the disease allele frequency and three penetrance parameters $\{f_{+/+}; f_{Het}; f_{m/m}\}$, that is, the probabilities to express the trait, given one of the three possible genotypes. Here, "+" and "*m*" refer to the wild-type and mutant allele, respectively, and "Het" stands for the heterozygous genotype. Usually, a linkage analysis represents the first step toward the mapping of a disease gene whose location is unknown. In the context of a parametric (model-based) linkage analysis, the disease allele frequency and penetrances have to be specified prior to the analysis. When performing linkage analysis of imprinted disease genes, however, the three-penetrance formulation is insufficient. In this case, the probability that a heterozygote expresses the trait is greatly influenced by the parent who transmitted the disease allele. Let us look at the example of a fully penetrant disease allele with no phenocopies and complete maternal imprinting, that is, exclusively paternal expression. If the specified heterozygote penetrance is 0, heterozygotes who have inherited the mutant allele from the mother will be treated correctly, whereas for those heterozygotes who have inherited the disease allele from the father, the parameter will be wrong. Likewise, if the heterozygote penetrance is specified to be 1, heterozygotes who have inherited the disease allele from the father will be treated correctly, but not those who have inherited the disease allele from the mother. Hence, no matter what heterozygote penetrance is specified for the analysis, it will not be optimal for the mapping of imprinted genes.

In a linkage study of bipolar affective disorder, Stine *et al*. (1995) classified the pedigrees according to "maternal" and "paternal" transmission of the disease, and performed separate linkage analyses for the two classes. Substantial differences between the LOD scores for the "maternal" and "paternal" pedigrees were interpreted as evidence for imprinting. However, such a classification does not make sense if the disease is transmitted through both males and females. As an alternative, it has been proposed to use sex-specific recombination fractions between marker and trait locus for the calculation of LOD scores (Smalley, 1993), or to fix the recombination fraction of the imprinting sex at 1/2. This allows for an explanation of unaffected heterozygotes by fictitious recombinations in the parent who has transmitted the mutant allele. It is also possible to define separate liability

classes for heterozygotes who have inherited the disease allele from the father and from the mother (Heutink *et al*., 1992). Still, in most cases, it is impossible to infer the parent-of-origin at first sight; it can only be inferred by likelihood calculation. Thus, the most natural approach to account for imprinting in the context of a parametric linkage analysis is to extend the trait model to contain two different heterozygote penetrances, one for paternal and one for maternal origin of the disease allele. This allows for flexibly taking the parent-of-origin into account in the likelihood calculation, by use of the appropriate penetrance. Altogether, the trait model with imprinting contains four penetrance parameters. This formulation was implemented into the program GENEHUNTER-IMPRINTING (Strauch *et al*., 2000a), which, in its current version, is based on the original GENEHUNTER v2.1 (Kruglyak *et al*., 1996; Kruglyak and Lander, 1998; Markianos *et al*., 2001). The program employs the Lander–Green algorithm (Lander and Green, 1987) and can therefore easily cope with a large number of markers (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1). On the other hand, pedigrees are restricted to be of moderate size. Version 2 of the program VITESSE (O'Connell and Weeks, 1995; O'Connell, 2001) can also take imprinting models with four penetrances into account. This program mainly uses the Elston–Stewart algorithm (Elston and Stewart, 1971) (although version 2 also incorporates features of the Lander–Green algorithm); therefore, it can handle larger pedigrees, but there is a limit on the number of markers that can be jointly analyzed. In terms of significance, LOD scores obtained under a four-penetrance model are directly comparable to LOD scores obtained under a three-penetrance nonimprinting model, provided that a single predefined model is used. However, this no longer holds if the LOD scores are maximized not only over the recombination fraction between marker and trait locus (or, in the context of multimarker analysis, over the genetic map position of the trait locus) but also with respect to the disease allele frequency and penetrances, as is done in a MOD-score analysis (Clerget-Darpoux *et al*., 1986).

It should be expected that the power to detect linkage for imprinted genes is higher under a four-penetrance imprinting model than it is under a standard trait model with three penetrances. Indeed, this is the finding obtained by Strauch *et al*. (1999). The result can be anticipated when looking at the diamond of inheritance (DOI), which is shown in Figure 1, for the special case of complete penetrance and no phenocopies. The DOI illustrates the parameter space formed by the two heterozygote penetrances of the four-penetrance imprinting model. The penetrances are given in the order $\{f_{+/+}; f_{m/+}; f_{+/m}; f_{m/m}\}$, with the paternally inherited allele listed first. In order to obtain a more instructive illustration of the relationships between different modes of inheritance (MOI), the two heterozygote penetrances $f_{m/+}$ and $f_{+/\mathrm{m}}$ are transformed into the dominance index D and the imprinting index I. These two values were previously defined in the context of Figure 1 in Strauch *et al*. (2000a). Here, in order to properly take the case of a nonzero phenocopy rate or reduced penetrance into account, the dominance and imprinting index are redefined as follows:

$$D = \frac{f_{m/+} + f_{+/m} - f_{+/+} - f_{m/m}}{f_{m/m} - f_{+/+}} \qquad I = \frac{f_{m/+} - f_{+/m}}{f_{m/m} - f_{+/+}} \tag{1}$$

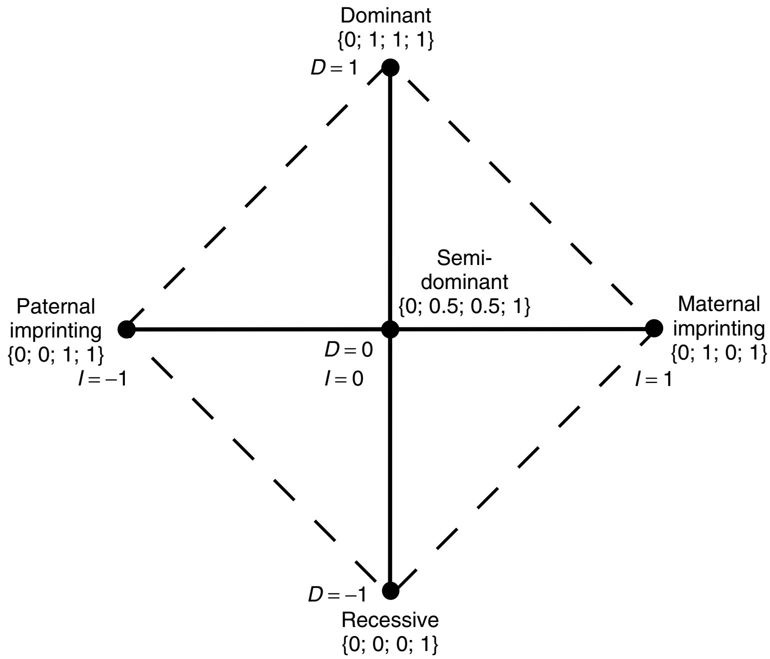


Figure 1 The diamond of inheritance visualizes the parameter space formed by the two heterozygote penetrances of the four-penetrance trait model that takes genomic imprinting into account, for the special case of no phenocopies and complete penetrance (i.e., $f_{+/+} = 0$ and $f_{m/m} = 1$). Dominant ($D = 1$), semidominant ($D = 0$), and recessive modes of inheritance ($D = -1$) are shown on the vertical axis, with the degree of imprinting I being 0. For these models, the heterozygote penetrances $f_{m/+}$ and $f_{+/m}$ are equal. A trait model on the left half of the diamond, with $f_{m/+} < f_{+/m}$ and $I < 0$, corresponds to paternal imprinting. Analogously, a model on the right half of the diamond, with $f_{m/+} > f_{+/m}$ and $I > 0$, corresponds to maternal imprinting. (Modified from Strauch K, Fimmers R, Kurz T, *et al*. Parametric and non-parametric multipoint linkage analysis with imprinting and two-locus-trait models in the *American Journal of Human Genetics*, 2000; **66**, 1945–1957, with permission)

The dominant and recessive MOI are the distal points on the vertical dominance axis, with the dominance index D ranging from -1 (recessive MOI, both heterozygote penetrances equal $f_{+/+}$) to 1 (dominant MOI, both heterozygote penetrances equal $f_{m/m}$). For a semidominant or additive MOI, the two heterozygote penetrances are halfway between the two homozygote penetrances $f_{+/+}$ and $f_{m/m}$, and D equals zero. All nonimprinting models, for which the two heterozygote penetrances are equal (i.e., $I = 0$), are represented by the central vertical line. The imprinting axis is perpendicular to the dominance axis; the imprinting index I ranges from -1 (complete paternal imprinting) to 1 (complete maternal imprinting). In these extreme cases of complete imprinting, one heterozygote penetrance equals $f_{+/+}$ and the other equals $f_{m/m}$. Figure 1 shows the diamond of inheritance with these new definitions of D and I. The graphical representation clearly illustrates that the paternal- and maternal-imprinting MOI lie far off the central axis of dominant–recessive inheritance. It explains the fact that the power to detect

linkage will drop if imprinting is not adequately accounted for in the analysis. This is equivalent to the statement for standard LOD-score analysis under three-penetrance trait models that the power to detect linkage is maximal if the analysis model roughly equals the true disease model (Clerget-Darpoux *et al*., 1986).

2.2. Nonparametric analysis

It is also possible to account for genomic imprinting in a nonparametric or model-free linkage analysis. Paterson *et al*. (1999) and McInnis *et al*. (2003) proposed to evaluate allele sharing separately for male and female meioses. Significant differences point to an imprinted gene. An option to perform such an analysis has been implemented into the program ASPEX (Hinds and Risch, 1996), in which two different χ^2 tests with 1 degree of freedom are performed, one each for sharing by male and female meioses. In a linkage study of psoriatic arthritis, Karason *et al*. (2003) used imprinting-based scoring functions, for nonparametric analysis of extended pedigrees, which have been implemented into the program ALLEGRO (Gudbjartsson *et al*., 2000). In particular, weights to allele sharing are assigned specific to parental origins. For example, the scoring function considers only the sharing of alleles transmitted to two affected relatives through their mothers, if paternal imprinting is to be modeled. Knapp and Strauch (2004) have developed a likelihood-based affected sib-pair test for imprinted disease genes. It is similar to Holmans' possible triangle test (Holmans, 1993), for the nonimprinting case, which is an extension of the likelihood-ratio test proposed by Risch (1990). In the approach taken by Knapp and Strauch, affected sib pairs who share one allele identical-by-descent (ibd) are distinguished by whether sharing is through the mother or father. Constraints on the sharing probabilities are derived for genetically possible models, which may include imprinting, similar to Holmans' triangular constraints for the nonimprinting case. The corresponding likelihood-ratio test proves to be substantially more powerful than Holmans' possible triangle test in the case that imprinting is present, at the cost of only a small reduction in power if there is no imprinting.

2.3. Quantitative trait locus analysis

Linkage analysis methods that adequately take imprinting into account also exist for quantitative trait loci (QTL). In the context of nuclear families, such methods were proposed by Hanson *et al*. (2001), for variance components analysis and Haseman–Elston regression, as well as by Shete and Amos (2002), for variance components analysis. Here, the proportion of alleles shared ibd is partitioned into a component derived from the father and a component derived from the mother. A further development of variance components analysis with imprinting for extended pedigrees has been described by Shete *et al*. (2003). In that context, allele-sharing ibd by two relatives is distinguished by the combination of sexes of the two transmitting parents.

3. Tests for linkage versus tests for imprinting

It is important to distinguish between tests for linkage that take imprinting into account and tests for imprinting per se. For example, a parametric analysis under a four-penetrance trait model represents a test for linkage that accounts for imprinting. However, a LOD-score analysis under such an extended trait model can also be used to infer the degree of imprinting, by calculation of MOD scores (i.e., maximization of the LOD score with respect to the disease-model parameters) (Strauch *et al*., 2000a). A difference between the two heterozygote penetrances $f_{m/+}$ and $f_{+/m}$ of the best-fitting model obtained by MOD-score analysis may indicate a parent-of-origin effect. To judge whether imprinting is indeed present, one can look at the difference between MOD scores obtained for analysis under models with and without imprinting (i.e., with four and three penetrances, respectively). The method employed by Paterson *et al*. (1999) and McInnis *et al*. (2003) allows for an investigation of linkage in the presence of imprinting, when looking at the allele sharing for male and female meioses. It can also be used to test for imprinting, when focusing on the difference between sharing for male and female meioses. The imprinting-based scoring functions for nonparametric analysis, used by Karason *et al*. (2003), can as well be used to test both for linkage and imprinting. For an assessment of linkage, one has to look at the results obtained for the paternal-imprinting and maternal-imprinting scoring functions, and for imprinting, one has to focus on their difference. The likelihood-based test proposed by Knapp and Strauch (2004) is a test for linkage that accounts for imprinting. In addition, a likelihood-based affected sib-pair test for imprinting is outlined in the discussion section of that paper. The latter is equivalent to a test for imprinting proposed by Olson and Elston (1998). The methods for QTL analysis developed by Hanson *et al*. (2001), Shete and Amos (2002), and Shete *et al*. (2003) allow for both the assessment of linkage under imprinting and the assessment of imprinting.

4. Confounding between imprinting and sex differences in recombination fractions

Researchers need to be aware of the fact that imprinting can be confounded with sex differences in recombination fractions. As described above, it is possible to use sex-specific recombination fractions to allow for imprinting in a linkage analysis (Smalley, 1993). In that approach, unaffected heterozygotes are accounted for by excess recombinations in the sex whose transmitted allele is imprinted. On the other hand, it is also possible that a true sex difference in recombination fractions is misinterpreted as evidence for imprinting, in cases in which actually no genomic imprinting is present. For example, if the recombination fraction in a certain genetic region is higher in females than in males, the additional recombinations in females can be explained by a maternal-imprinting model in the context of a parametric analysis. By this means, the additional female recombinations are regarded as nonexistent, and the appearingly transmitted disease alleles from the mother are interpreted as being nonpenetrant. This confounding also occurs with

nonparametric linkage analysis. In the case of sex differences in recombination fractions, the excess allele sharing, due to an existing linkage, will be smaller for the alleles inherited from the parent who has a higher recombination fraction between marker and trait locus. The smaller amount of allele sharing through this parental sex will be interpreted as imprinting. In order for an imprinting analysis to be robust with respect to sex differences in recombination fractions, Paterson (2000) proposed to perform multimarker analysis with sex-specific maps (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1). However, it can be argued that confounding does not represent a major problem in the context of multimarker analysis, even if sex-averaged recombination fractions are used (Strauch *et al*., 2000b). Here, each genetic position within the marker group is located between two flanking markers. In this situation, it is hardly possible to use fictitious recombinations to explain carriers of the disease allele who are unaffected because of imprinting. This is because such fictitious recombinations would have to be double recombinations, which are rather unlikely. By the same token, a truly existing sex difference in recombination fractions may lead to additional recombinations, in the meioses of one sex, between a putative disease-locus position and one marker, but not with both flanking markers at the same time. This will remove the advantage of an analysis method that accounts for imprinting, over one that does not, since it is no longer necessary to explain a recombination by a nonpenetrant case. However, these arguments no longer hold if the sex ratio of the recombination fractions takes extreme values in a particular genetic region, or if the marker density is low. In such a case, double recombinants are possible, and confounding becomes likely. However, the problem of double recombinants cannot be remedied by use of sex-specific marker maps; rather, additional markers should be typed in order to increase marker density. This is not only important to avoid confounding, but also to increase linkage information, which is otherwise going to be low for the meioses of the sex with the higher recombination fractions.

5. Imprinting and association analysis

The basic idea common to all imprinting association methods is to distinguish between maternally and paternally transmitted alleles when investigating an influence on the trait. For this reason, it is not possible to model a parent-of-origin effect in case–control association studies, and thus, family-based designs are a prerequisite. In the context of dichotomous traits, Weinberg *et al*. (1998) proposed a log-linear method, allowing researchers to perform a likelihood-ratio χ^2 test for association that takes imprinting into account. With this method, it is also possible to estimate association parameters, such as relative risks, and to test for imprinting itself. Furthermore, they devised a "transmission asymmetry test", in the spirit of the transmission/disequilibrium test (TDT), which also is a test for imprinting. Weinberg (1999) reviewed association methods for detection of imprinting, and showed that they can be invalid under certain scenarios. This holds, for example, for a stratification of the transmission/nontransmission counts by the parental sex if the sample includes case-parents triads with two heterozygous parents. In addition, problems can arise in the case of maternal effects. Therefore, in the same work,

Weinberg proposed two new methods, based on a logistic model that includes the parental mating type as well as the number of inherited copies of the allele under study. Methods of association that account for imprinting in the context of quantitative traits were described by van den Oord (2000), who proposed to use a finite mixture model, and by Whittaker *et al*. (2003), who employed simple linear models. For association analysis in general, confounding between sex differences in recombination fractions and imprinting should not represent a problem, since association will usually only be detected for markers tightly linked to a disease locus.

6. Conclusion

It is well known that a considerable portion of the human genome is subject to imprinting (Hall, 1990). However, not all imprinted genetic regions have been identified and localized yet. Since the power to detect imprinted genes will be highest if a parent-of-origin effect is adequately accounted for in the analysis, researchers should routinely employ methods that model imprinting, for both linkage and association analysis. This holds particularly for genetically complex diseases, for which the effect of a single gene on the trait is likely to be small. If genetic mapping studies in the past did not show positive results, it may, therefore, be reasonable to reanalyze the data with a method that accounts for imprinting. During the past years, a considerable number of methods and computer programs for analysis have been developed that take parent-of-origin effects into account. It is likely that they will contribute to the mapping of imprinted genes responsible for both Mendelian and complex traits.

Related articles

Article 45, **Bioinformatics and the identification of imprinted genes in mammals**, Volume 1; Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1; Article 51, **Choices in gene mapping: populations and family structures**, Volume 1

References

Ainscough JFX and Surani MA (1996) Organization and control of imprinted genes: the common features. In *Epigenetic Mechanisms of Gene Regulation*, Russo VEA, Martienssen RA and Riggs AD (Eds.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, pp. 173–194.

Bain SC, Rowe BR, Barnett AH and Todd JA (1994) Parental origin of diabetes-associated HLA types in sibling pairs with type I diabetes. *Diabetes*, **43**, 1462–1468.

Bartolomei MS and Tilghman SM (1997) Genomic imprinting in mammals. *Annual Review of Genetics*, **31**, 493–525.

Bennett ST, Todd JA, Waterworth DM, Franks S and McCarthy MI (1997) Association of insulin gene VNTR polymorphism with polycystic ovary syndrome. *Lancet*, **349**, 1771–1772.

Clerget-Darpoux F, Bonaïti-Pellié C and Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, **42**, 393–399.

Daniels SE, Bhattacharrya S, James A, Leaves NI, Young A, Hill MR, Faux JA, Ryan GF, le Souef PN, Lathrop GM, *et al*. (1996) A genome-wide search for quantitative trait loci underlying asthma. *Nature*, **383**, 247–250.

de la Casa-Esperón E and Sapienza C (2003) Natural selection and the evolution of genome imprinting. *Annual Review of Genetics*, **37**, 349–370.

Elston RC and Stewart J (1971) A general model for the genetic analysis of pedigree data. *Human Heredity*, **21**, 523–542.

Esplin MS, Fausett MB, Fraser A, Kerber R, Mineau G, Carrillo J and Varner MW (2001) Paternal and maternal components of the predisposition to preeclampsia. *The New England Journal of Medicine*, **344**, 867–872.

Falls JG, Pulford DJ, Wylie AA and Jirtle RL (1999) Genomic imprinting: implications for human disease. *American Journal of Pathology*, **154**, 635–647.

Greenberg DA, Durner M, Keddache M, Shinnar S, Resor SR, Moshe SL, Rosenbaum D, Cohen J, Harden C, Kang H, *et al*. (2000) Reproducibility and complications in gene searches: linkage on chromosome 6, heterogeneity, association, and maternal inheritance in juvenile myoclonic epilepsy. *American Journal of Human Genetics*, **66**, 508–516.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Hall JG (1990) Genomic imprinting: review and relevance to human diseases. *American Journal of Human Genetics*, **46**, 857–873.

Hanson RL, Kobes S, Lindsay RS and Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *American Journal of Human Genetics*, **68**, 951–962.

Heutink P, van der Mey AGL, Sandkuijl LA, van Gils APG, Bardoel A, Breedveld GJ, van Vliet M, van Ommen GJ, Cornelisse CJ, Oostra BA, *et al*. (1992) A gene subject to genomic imprinting and responsible for hereditary paragangliomas maps to chromosome 11q23-qter. *Human Molecular Genetics*, **1**, 7–10.

Hinds DA and Risch N (1996) *The ASPEX Package: Affected Sib-Pair Exclusion Mapping*. **http://sourceforge.net/projects/aspex/**.

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, **52**, 362–374.

Karason A, Gudjonsson JE, Upmanyu R, Antonsdottir AA, Hauksson VB, Runasdottir EH, Jonsson HH, Gudbjartsson DF, Frigge ML, Kong A, *et al*. (2003) A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting. *American Journal of Human Genetics*, **72**, 125–131.

Knapp M and Strauch K (2004) Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting. *Genetic Epidemiology*, **26**, 273–285.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Kruglyak L and Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, **5**, 1–7.

Lander ES and Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363–2367.

Markianos K, Daly MJ and Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, **68**, 963–977.

McInnis MG, Lan TH, Willour VL, McMahon FJ, Simpson SG, Addington AM, MacKinnon DF, Potash JB, Mahoney AT, Chellis J, *et al*. (2003) Genome-wide scan of bipolar disorder in 65 pedigrees: supportive evidence for linkage at 8q24, 18q22, 4q32, 2p12, and 13q12. *Molecular Psychiatry*, **8**, 288–298.

Moffatt MF and Cookson WO (1998) Maternal effects in atopic disease. *Clinical and Experimental Allergy*, **28**(Suppl 1), 56–61.

O'Connell JR (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. *Human Heredity*, **51**, 226–240.

O'Connell JR and Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

Olson JM and Elston RC (1998) Using family history information to distinguish true and false positive model-free linkage results. *Genetic Epidemiology*, **15**, 183–192.

Pardo-Manuel de Villena F, de la Casa-Esperón E and Sapienza C (2000) Natural selection and the function of genome imprinting: beyond the silenced minority. *Trends in Genetics*, **16**, 573–579.

Paterson AD (2000) Analysis of parental-origin effects in linkage data. *Molecular Psychiatry*, **5**, 125–126.

Paterson AD, Naimark DMJ and Petronis A (1999) The analysis of parental origin of alleles may detect susceptibility loci for complex disorders. *Human Heredity*, **49**, 197–204.

Petronzelli F, Bonamico M, Ferrante P, Grillo R, Mora B, Mariani P, Apollonio I, Gemme G and Mazzilli MC (1997) Genetic contribution of the HLA region to the familial clustering of coeliac disease. *Annals of Human Genetics*, **61**, 307–317.

Reik W and Walter J (2001) Genomic imprinting: parental influence on the genome. *Nature Reviews. Genetics*, **2**, 21–32.

Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics*, **46**, 242–253.

Shete S and Amos CI (2002) Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *American Journal of Human Genetics*, **70**, 751–757.

Shete S, Zhou X and Amos CI (2003) Genomic imprinting and linkage test for quantitative-trait loci in extended pedigrees. *American Journal of Human Genetics*, **73**, 933–938.

Smalley SL (1993) Sex-specific recombination frequencies: a consequence of imprinting? *American Journal of Human Genetics*, **52**, 210–212.

Stine OC, Xu J, Koskela R, McMahon FJ, Gschwend M, Friddle C, Clark CD, McInnis MG, Simpson SG, Breschel TS, *et al*. (1995) Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. *American Journal of Human Genetics*, **57**, 1384–1394.

Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF and Baur MP (2000a) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *American Journal of Human Genetics*, **66**, 1945–1957.

Strauch K, Fimmers R, Wienker TF, Baur MP, Cichon S, Propping P and Nöthen MM (2000b) Reply to Paterson. *Molecular Psychiatry*, **5**, 126–127.

Strauch K, Fimmers R, Windemuth C, Hahn A, Wienker TF and Baur MP (1999) Linkage analysis with adequate modeling of a parent-of-origin effect. *Genetic Epidemiology*, **17**(Suppl 1), S331–S336.

van den Oord EJ (2000) The use of mixture models to perform quantitative tests for linkage disequilibrium, maternal effects, and parent-of-origin effects with incomplete subject-parent triads. *Behavior Genetics*, **30**, 335–343.

Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *American Journal of Human Genetics*, **65**, 229–235.

Weinberg CR, Wilcox AJ and Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *American Journal of Human Genetics*, **62**, 969–978.

Whittaker JC, Gharani N, Hindmarsh P and McCarthy MI (2003) Estimation and testing of parent-of-origin effects for quantitative traits. *American Journal of Human Genetics*, **72**, 1035–1039.

Wilkins JF and Haig D (2003) What good is genomic imprinting: the function of parent-specific gene expression. *Nature Reviews. Genetics*, **4**, 359–368.

Specialist Review

Gene mapping and the transition from STRPs to SNPs

Ellen M. Wijsman
University of Washington, Seattle, WA, USA

1. Introduction

Human gene mapping is focused on two major goals. The first is linkage detection: to determine where, relative to a genetic map, gene(s) are located that contribute to the trait. The second goal is localization: to more accurately determine the location of the gene(s), in preparation for gene identification. Although these goals have not changed, the types of markers used to achieve these goals have changed over time.

The principles behind gene mapping in humans are identical to those for other diploid organisms. However, there are numerous practical complications. Meioses used in gene mapping are not observed directly, but are inferred through examination of genetic markers in pedigrees. In human pedigrees, the inability to control matings leads to the use of indirect statistical methods for inference regarding meiotic transmission. In addition, the need to use observational data has two important implications. First, the sample sizes and the cost of human studies are high, because the loss of information in uncontrolled crosses leads to the need for increased sample sizes, and identification of pedigrees with ideal characteristics for gene mapping can be difficult and labor-intensive (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2). Second, it is necessary to develop procedures that result in reliable and efficient extraction of mapping information on arbitrary pedigrees since pedigree structures vary widely (*see* Article 60, **Population selection in complex disease gene mapping**, Volume 2). These two issues create the need to search for approaches that maximize the use of available data.

The development of efficient human gene mapping studies involves interrelated areas, each of which will be discussed in more depth, below. An understanding of changes in mapping procedures in response to a move from STRPs (single tandem repeat polymorphisms) to SNPs (single-nucleotide polymorphism) requires an understanding of each of these areas, and how it interacts with the choice of the marker. The areas are as follows: (1) The choice of genetic marker has changed over time. The transition evolved from classical protein-based markers to the early DNA-based restriction fragment length polymorphisms (RFLPs), which were typically diallelic, to multiallelic variable number tandem repeat (VNTR) and finally to microsatellite STRPs. With each transition, automation and speed of

genotyping increased, as did the mapping information achievable. Costs dropped. More recently, the use of SNPs has been proposed, on the basis of high speed and low cost, which is potentially achievable. (2) Construction of the associated genetic maps involves the consideration of possible analysis approaches, and potential data. Both the density and the structure of such maps are relevant, as is the quality of the map estimates. (3) There is an ever-growing arsenal of approaches to analysis, with an increasing emphasis on the development of statistical methods suitable for complex traits. Depending on other aspects of the data and analysis, there are often constraints on either the number of markers that can simultaneously be used in the analysis or on the sizes of pedigrees that can be analyzed. (4) Finally, the choice of pedigrees selected for a study interacts with the analytical approaches and with the number of markers and map structure that is practical to use in analysis.

2. Genotyping technologies

Marker genotyping technologies have changed considerably over time (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4). The realization that DNA-based variation could provide an essentially limitless source of markers was a breakthrough (Botstein *et al*., 1980). Prior to this, identifying new markers involved serendipity, and large-scale genotyping was extremely expensive. DNA-based variation, on the other hand, can be assayed with generic methods, and the use of RFLPs, which were the first DNA-based markers, resulted in rapid growth in the number of mapped markers (Dib *et al*., 1996). Early DNA genotyping was still expensive, thus severely limiting the sizes of the studies that could be tackled and thus the complexity of the traits that could be analyzed with reasonable cost. In addition, the diallelic polymorphisms that were typical of RFLPs were relatively uninformative for gene mapping.

The success of RFLPs stimulated the development of better markers and identification and popularization of STRPs (Weber and May, 1989). STRPs have long remained the most popular type of markers, serving well for simple and complex traits, as well as for analysis of large and small pedigrees (*see* Article 67, **History of genetic mapping**, Volume 4). STRP genotyping has modest costs because of considerable automation, but rapid genotyping is achieved only in the largest facilities. Recently, SNPs have been proposed as a replacement (Hastbacka *et al*., 1992; Kruglyak, 1997), with the expectation that use of large numbers of SNPs can be used to offset the lack of mapping information intrinsic in diallelic markers, and that genotype scoring can be almost completely automated, thus increasing the speed and reducing the costs. However, in considering the transition to SNPs, it is useful to remember the four major reasons leading to the popularity of STRPs (*see* Article 67, **History of genetic mapping**, Volume 4). (1) STPRs provide high information for mapping (*see* Article 53, **Information content in gene mapping**, Volume 1) even in single-marker analysis. (2) STRPs are applicable to a wide variety of problems and data sets. (3) STRPs can be used for most analytical approaches (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1), with relative

insensitivity to simplifying assumptions in analysis. (4) With STRPs, the genotyping component of a study represents a relatively low fraction of the total cost of a study, including pedigrees and phenotype collection, genotyping, and analysis.

3. Information

A critical factor is the information obtained. Here, we describe two aspects of information (*see* Article 53, **Information content in gene mapping**, Volume 1) that are relevant in the comparison of STRP- and SNP-based mapping studies. First, there is the fraction of *mapping information*, or I_m, measured on a scale of 0 to 1, that can be extracted via a particular marker set. Second, there is the *total information*, I_t, potentially available in a pedigree data set. Such information increases with increasing sizes of data sets (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1), and is a function of aspects of the data such as number and sizes of pedigrees, as well as the phenotype. Finally, the combination gives the *realized information*, or I_r, where $I_m I_t = I_r$. The amount of data needed is a function of I_r. Clearly, the design of a study is a function of the information that can be extracted with a particular panel of markers, relative to the maximum that could be extracted with perfect markers, and the underlying absolute amount of information potentially available in a data set.

Information is most usefully defined in the context of a particular analytic method. The different measures of mapping information will not give identical results when applied to a particular set of markers in a particular data set. Examples of mapping information measures that are based on particular analytic frameworks are PIC (Botstein *et al*., 1980), which is a useful measure for methods that are based on scoring transmitted gametes, MPIC (Goddard and Wijsman, 2002), which is an extension to tightly clustered markers, and LIC (linkage information content) (Guo and Elston, 1999), which is useful for methods based on identity-by-descent (ibd) sharing in pairs of individuals. Another information measure is that of entropy, which has been proposed as a general measure of information for use in evaluating information in multilocus maps (Kruglyak, 1997). This measure, while strongly correlating with other measures, is not a direct measure of mapping information that is achieved with any particular method of linkage analysis. Each measure of information describes the proportion of meioses or other sampling units (e.g., sib pairs) that can be scored for the event of interest, assuming accuracy of all aspects of the map and marker model. For a particular measure, a value of $I_m = 0.5$ requires a doubling in sample size to achieve the same realized sample size as would $I_m = 1$. Investigators generally desire high values for I_m since for a particular choice of pedigree type and analytic approach (dictating I_t), this minimizes the necessary sample size collected.

High I_r can be achieved by using individual STRPs, by using multiple markers, or by a combination of strategies (*see* Article 53, **Information content in gene mapping**, Volume 1). Increased marker density increases I_m when used in multipoint analysis, as does an overall increase in the number of markers. Multipoint analysis is essential for the realization of high information in the context of SNP marker scans. In contrast, although it can improve the available information,

it is not always absolutely necessary for analysis with STRPs. Simulation studies demonstrate that even for STRPs at typical mapping densities, I_m increases considerably when multiple markers are used in multipoint analysis (Amos *et al*., 1997), and similar studies have shown that high values for I_m are obtainable for panels of SNP markers (International Multiple Sclerosis Genetics Consortium, 2004; John *et al*., 2004; Matise *et al*., 2003). There are also proposals for different ways to construct SNP multipoint mapping panels that end up with information that is at least as high as the information obtainable with STRPs. These include the use of SNP panels with a uniform marker density (Kruglyak, 1997) as well as the use of SNP mapping panels that are based on clusters of very closely linked SNPs (see Section 6) (Goddard and Wijsman, 2002).

It is I_r that is important in a mapping study. When discussing issues pertaining to the choice of STRP versus SNP markers for linkage analysis, there is a tendency to focus only on I_m. The pedigree and trait material used for analysis (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1 and Article 60, **Population selection in complex disease gene mapping**, Volume 2) are as important since this determines, in part, the pedigrees collected, which affects I_r. While I_m can increase by 30–40% for STRP markers with the use of multipoint instead of single-marker analysis, I_t can easily increase by 200–300% per sampled individual, when appropriate choice of pedigrees and/or phenotypes are considered (Wijsman and Amos, 1997). The part of a gene mapping study that is the slowest and most expensive is the collection of the pedigree and phenotype data. It may be overall most effective in many cases to collect pedigree and phenotype data that has maximum I_t per sampled individual, and then to use analysis methods that use these data most efficiently. This may have important ramifications in the choice of a marker panel.

4. Experience with real data

To date, there have been only a few reports describing the use of SNPs in linkage analysis of real data. Information about the performance of SNPs for use on real data will undoubtedly change soon, since multiple data collections and analyses are currently in progress. Real-life experience with SNPs will be needed to identify the problems that almost certainly will occur as investigators gain experience with the use and analysis of data derived from SNP marker panels.

Analyses on real data sets that have been typed for both STRP and SNP markers support the theoretical prediction that information estimated on these data sets increases with marker density (*see* Article 53, **Information content in gene mapping**, Volume 1). Computed information is greater for sufficiently dense panels of SNP markers than that for standard 10-cM density panels of STRP markers (International Multiple Sclerosis Genetics Consortium, 2004; John *et al*., 2004), and increases with the density of SNP marker panels (International Multiple Sclerosis Genetics Consortium, 2004). Note that these conclusions are based on the assumption that the model used to compute information in a multipoint setting is sufficiently close to correct, including the marker allele frequencies, linkage equilibrium between markers, genetic map distances, and mapping function

(*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4). We do not yet know how robust these assumptions are. However, there are suggestions that there may be inaccuracies in the model used to predict information in multipoint analysis. There are regions with substantial differences in the information computed for SNP versus STRP panels, yet the linkage analysis provided virtually identical results (Browning *et al*., 2004; John *et al*., 2004). In other regions, information in two marker panels was nearly identical, but the linkage signals were substantially different (Browning *et al*., 2004). Until these discrepancies are understood, caution is needed in interpreting the measures of information used to compare multipoint SNP versus STRP mapping panels.

For the one published comparison of genome scan results for a large nuclear-family data set with ∼10 000 SNPs and ∼400 STRPs, there was overall excellent agreement in the results obtained with both marker panels (John *et al*., 2004). In this study, the most convincing evidence for linkage was stronger for the analyses with STRPs than for that obtained with the SNPs, despite higher computed regional marker information for the SNPs. There were also several additional regions with virtually identical but modest linkage signals; with a few modestly higher for the SNPs and others for the STRPs. It is difficult to make definitive statements regarding the advantages or disadvantages of the use of either of the mapping panels, on the basis of these results.

5. Linkage analysis

5.1. *Analysis basics*

The choice of marker panel presents several issues. One can choose to use single-marker or multipoint analysis, and an analysis method based on marker ibd-sharing or one based on an explicit or estimated trait model (e.g., parametric LOD score or joint linkage and segregation analysis) (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1 and Article 56, **Computation of LOD scores**, Volume 1). The rationale for the choice of a strategy is best determined by the trait to be mapped, estimates of power to detect linkage under various analysis scenarios and marker choices, and the goal(s) of the analysis. For example, in cases of trait model misspecification, parametric linkage analysis with single markers may have higher power for linkage detection than would a multipoint analysis (Risch and Giuffra, 1992; Sullivan *et al*., 2003). A single-marker analysis with a typical STRP may also have higher power to detect linkage than a multipoint analysis with SNPs, if the use of SNPs involves significant pedigree pruning to make the computations feasible. On small pedigrees, a multipoint approach may be feasible and desirable, especially if the analysis is based on an ibd-scoring method. Accurate localization that follows linkage detection may be the primary goal in other studies, in which case a parametric multipoint analysis may be the approach of choice, since ibd-sharing methods, which are, strictly speaking, only linkage-detection methods, give poor information regarding localization regardless of marker density (Atwood and

Heard-Costa, 2003). Because there is no single analytical approach that is optimal for all data sets, there will also probably be no single-marker panel that is also optimal under all conditions.

While STRPs can be efficiently used for both single-marker and multipoint analysis, the use of SNPs *requires* multipoint analysis to maintain reasonable information (*see* Article 53, **Information content in gene mapping**, Volume 1). STRPs can be effectively used for ibd-sharing as well as trait-model-based analyses, with single-marker analysis sometimes preferred (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1). Also, for STRPs there are many statistical approaches available, with computational tools implementing such approaches (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1). In contrast, although some of these analytical methods can be used with SNPs, the effect of violation of assumptions underlying the analyses has not been evaluated. The long-term use of SNPs for the wide range of possible problems may need to include additional assumptions and methods. It also needs to include an understanding of assumptions behind the analysis, so that appropriate modifications and extensions to existing approaches can be developed.

5.2. Modifications and extensions needed

Current analysis methods were developed under the assumption of linkage equilibrium among markers (*see* Article 2, **Modeling human genetic history**, Volume 1). For STRP marker panels, which range from 400 to 2000 markers (~10 cM to ~2 cM density), this is a reasonable assumption since linkage disequilibrium (LD) (*see* Article 10, **Measuring variation in natural populations: a primer**, Volume 1 and Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3) at such spacing is rare (Dawson *et al.*, 2002; Stephens *et al.*, 2001). In contrast, some SNP panels are sufficiently dense that LD between neighboring markers may affect the analysis (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4, Article 71, **SNPs and human history**, Volume 4, and Article 73, **Creating LD maps of the genome**, Volume 4). Recent work suggests that the presence of LD among SNPs can lead to inflated evidence of linkage, when an analysis approach that assumes linkage equilibrium is used (Schaid *et al.*, 2004). While a few computer programs, such as FASTLINK (Cottingham *et al.*, 1993; *see also* Article 52, **Algorithmic improvements in gene mapping**, Volume 1), allow for LD among loci, these programs are unsuitable for use with large numbers of markers, as is needed for SNP panels. Alternatives that allow for local LD while handling large numbers of markers are necessary for the use of SNPs.

Multipoint analysis assumes accurate meiotic maps (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1). The pedigrees typically used to construct maps were assembled with relatively sparse maps in mind. Recent construction of denser maps and comparison of sequence-based and meiotic maps has identified discrepancies in map estimates for closely spaced markers (Hattori *et al.*, 2000), providing empirical support for the concern that small map

distances may be unreliably estimated. Statistical arguments predict that there may be somewhat inflated false-positive rates in the presence of map misspecification (Daw *et al*., 2000; *see also* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1). Analysis of real data under a variety of estimated maps has shown that results can be quite sensitive to map distance estimates (Gretarsdottir *et al*., 2002), suggesting that for reliable results from SNP maps, it may be necessary to both develop methods to estimate maps from the combined data across multiple studies and to develop systematic approaches to evaluate sensitivity of conclusions to map assumptions.

Genotyping error may influence linkage analysis results. For STRPs, it is safe to assume low error rates in linkage analysis, since most errors are readily identified (Epstein *et al*., 2000; Sieberts *et al*., 2002). For SNPs, error identification is more difficult, and unidentified errors may be influential (Cherny *et al*., 2001). More effort may be needed in data cleaning, and models that directly incorporate genotyping error models into the linkage analysis may be needed.

Computational algorithms will need modification and development (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1). Currently, there are practical constraints on the number of markers that can be used in a linkage analysis, even when the efficient Lander–Green algorithm serves as the basis for computation (Lander and Green, 1987). While there have been improvements (Abecasis *et al*., 2002; Gudbjartsson *et al*., 2000; Kruglyak, 1997; Markianos *et al*., 2001), computation with many hundreds of markers remains challenging, and remains limited to pedigrees of modest size. Other methods of analysis that allow multipoint computation on larger pedigrees, such as Markov chain Monte Carlo (MCMC) approaches (Heath, 1997; Sobel and Lange, 1996; Wijsman, 2003), are also likely to require substantial development to make them practical for use on SNP data. This could be critical since MCMC approaches are among the few that also allow for more complex trait models (Heath, 1997). One fundamental problem is that even the most efficient algorithms increase computation time linearly with the number of markers in the analysis. Although computers steadily increase in speed, it may take a decade for the hardware to catch up with the computational demands of current SNP panels.

6. Maps

There are multiple possible distributions of markers on a map. Uniform marker density provides the maximum information per typed marker (Goddard and Wijsman, 2002; *see also* Article 53, **Information content in gene mapping**, Volume 1). However, in order to realize this information, it is necessary to be able to use all the typed markers in a multipoint analysis. For some data sets, particularly consisting of large pedigrees (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1), multipoint analysis quickly becomes computationally infeasible (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1). One compromise that has been proposed is to use SNPs in clusters (Goddard and Wijsman, 2002). Each cluster is treated as a single multiallelic locus and can be used as an individual locus in

linkage analysis, or in a multipoint analysis with other such loci. Information about local linkage disequilibrium can be used to identify which markers create maximally informative clusters (*see* Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4). A map has been constructed with clusters of SNPs (Matise *et al*., 2003), with demonstration, as proof of principle, that it can extract as much or more information from real data than do STRPs (Browning *et al*., 2004). It may also be possible to use such an approach with groups of SNPs from other marker panels.

The density of markers is also important. Current SNP panels consist of 5000–11 000 SNPs. Traditional STRP panels consist of ∼400 markers, with a few panels with 800–2000 markers. While, in principle, the denser panels would be expected to provide more information, as well as a somewhat more accurate gene localization, it is not clear that in real data such gains will be realized. In addition, in most cases, it is unlikely that a true linkage signal would be completely missed by the use of a somewhat less dense panel of markers. The advantage of dense SNP panels may be more important for localization, if only because an investigation may be able to proceed directly to fine-scale localization without further genotyping.

7. Quality control

Linkage analysis can be exquisitely sensitive to data error. Methods for detecting error exist (Abecasis *et al*., 2002; Boehnke and Cox, 1997; Epstein *et al*., 2000; Sieberts *et al*., 2002) and take into account the map, allele frequencies, and observed genotype data. These methods are dependent on the same assumptions as are multipoint linkage analysis methods, and violation of these assumptions will also affect the detection of data error. There have been limited attempts to apply the same methods to SNP genotyping. Through simulation, it appears that a relatively high fraction of SNP genotype errors miss detection (Abecasis *et al*., 2002), especially when there are missing data in the pedigrees, as is typical of real data. The lower information for individual SNPs may mean less sensitivity of analyses to genotyping error. However, failure to account for such error will lead to the inefficient use of SNP genotyping. As for linkage analysis methods, extension and improvement of methods to detect or to model genotyping error will be needed to realize the potential of SNPs for linkage analysis.

Additionally, the density of SNP maps relative to STRP maps will draw increasing attention to the variability of human gene maps among individuals and populations. Dense SNP maps include markers that are located within regions that have polymorphic inversions, duplications, and deletions of DNA segments of varying length (*see* Article 55, **Polymorphic inversions, deletions, and duplications in gene mapping**, Volume 1). There is clearly a value in including such markers, as they permit some initial assessment of the possibility that these genomic regions are implicated in disease. This comes at the cost of further extension of methods for genetic analysis to accommodate whole new classes of polymorphisms that challenge the notion of a single, uniform genetic map.

8. Summary

SNPs are a new class of markers. They offer advantages in terms of cost and speed of genotyping. However, to realize the potential of SNPs, analysis must be done with multipoint methods. This has several consequences. The most important are that (1) assumptions that are adequate for sparser maps may not be appropriate for SNP-based analyses, (2) analytic methods and programs implementing such methods will need development and improvement, (3) choice of use of SNP versus STRP-based panels may be affected by factors such as the type of pedigrees and analytic methods to be used, and (4) experience with real data applications is needed to determine the real-life issues with the use of these markers.

Acknowledgments

Supported by NIH GM 46255, HD 33812, AG 14382, HD 35465, AG 05136, AG 11762, and AG 21544.

References

Abecasis G, Cherny S, Cookson W and Cardon L (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Amos CI, Krushkal TJ, Young A, Zhu DK, Boerwinkle E and de Andrade M (1997) Comparison of Model-free linkage mapping strategies for the study of a complex trait. *Genetic Epidemiology*, **14**, 743–748.

Atwood L and Heard-Costa N (2003) Limits of fine-mapping a quantitative trait. *Genetic Epidemiology*, **24**, 99–106.

Boehnke M and Cox N (1997) Accurate inference of relationships in sib-pair linkage studies. *American Journal of Human Genetics*, **61**, 423–429.

Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

Browning B, Brashear D, Butler A, Cyr D, Harris E, Nelsen A, Yarnall D, Ehm M and Wagner M (2004) Linkage analysis using single nucleotide polymorphisms. *Human Heredity*, **57**, 220–227.

Cherny S, Abecasis G, Cookson W, Sham P and Cardon L (2001) The effect of genotype and pedigree error on linkage analysis. *Genetic Epidemiology*, **21**(Suppl 1), S117–S122.

Cottingham RW, Idury RM and Schaffer AA (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics*, **53**, 252–263.

Daw E, Thompson E and Wijsman E (2000) Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology*, **19**, 336–380.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, *et al*. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**(6897), 544–548.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, *et al*. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.

Epstein MP, Duren WL and Boehnke M (2000) Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics*, **67**(5), 1219–1231.

Goddard K and Wijsman E (2002) Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genetic Epidemiology*, **22**, 205–220.

Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, Shkolny D, Einarsson G, Gudjonsdottir HM, Valdimarsson EM, *et al*. (2002) Localization

of a susceptibility gene for common forms of stroke to 5q12. *American Journal of Human Genetics*, **70**(3), 593–603.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**(1), 12–13.

Guo X and Elston RC (1999) Linkage information content of polymorphic genetic markers. *Human Heredity*, **49**, 112–118.

Hastbacka J, de la Chappelle A, Kaitila I, Sistonen P and Weaver A (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, **2**, 204–211.

Hattori M, Fujiyama A, Taylor T, Watanabe H, Yada T, Park H-S, Tyoda A, Ishii K, Totoki Y, Choi D, *et al*. (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

International Multiple Sclerosis Genetics Consortium (2004) Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Human Molecular Genetics*, **13**(17), 1943–1949.

John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, *et al*. (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *American Journal of Human Genetics*, **75**, 54–64.

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics*, **17**, 21–24.

Lander ES and Green P (1987) Construction of multilocus genetic maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363–2367.

Markianos K, Daly MJ and Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, **68**(4), 963–977.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, *et al*. (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *American Journal of Human Genetics*, **73**(2), 271–284.

Risch N and Giuffra L (1992) Model misspecification and multipoint linkage analysis. *Human Heredity*, **42**(1), 77–92.

Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hiker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML, *et al*. (2004) Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. *American Journal of Human Genetics*, **75**(6), 948–965.

Sieberts S, Thompson E and Wijsman E (2002) Relationship inference from trios of individuals in the presence of typing error. *American Journal of Human Genetics*, **70**, 170–180.

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al*. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**(5529), 489–493.

Sullivan P, Neale B, Neale M, van den Oord E and Kendler K (2003) Multipoint and single point non-parametric linkage analysis with perfect data. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics*, **121B**, 89–94.

Weber J and May P (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain-reaction. *American Journal of Human Genetics*, **44**, 388–396.

Wijsman E (2003) Summary of group 8: development and extension of linkage methods. *Genetic Epidemiology*, **25**(Suppl 1), S64–S71.

Wijsman EM and Amos C (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. *Genetic Epidemiology*, **14**, 719–735.

Specialist Review

Consequences of error

Derek Gordon

Rutgers University, Piscataway, NY, US

Stephen J. Finch

Stony Brook University, Stony Brook, NY, US

1. Introduction

In the field of statistical genetics, specifically in the areas of linkage and association methods (*see* Article 47, **Introduction to gene mapping: linkage at a crossroads**, Volume 1, Article 51, **Choices in gene mapping: populations and family structures**, Volume 1), "the skeleton in the closet" (Sobel *et al*., 2002) is the subject of misclassification errors. Misclassification occurs when an observed measurement is different than its true value. It may be one reason for the lack of replication in gene mapping studies (Page *et al*., 2003) (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2, Article 64, **Genetics of cognitive disorders**, Volume 2). Therefore, it is critically important for researchers to understand the consequences of error and whether their particular studies are subject to such error.

Misclassification errors may occur in either phenotype (diagnosis) or genotype. Phenotype (or diagnostic) misclassification has been consistently reported in diseases such as Alzheimer's (Lansbury, 2004), Multiple Sclerosis (Poser and Brinar, 2004), Parkinson's (Lansbury, 2004), and Inflammatory Bowel Disease (Silverberg *et al*., 2001) (*see* Article 62, **Inflammation and inflammatory bowel disease**, Volume 2). Genotype errors have been reported as a result of the technology used to determine genotypes, poor sample quality, or simply laboratory error (Bonin *et al*., 2004). Historically, microsatellite loci, which are based on nucleotide repeat sequences, were prone to error rates greater than 5% (Brzustowicz *et al*., 1993) (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1). New genotype technologies have focused on single nucleotide polymorphisms (SNPs) (single base-pair variation among individuals) because of their high abundance and ease in genotyping. Estimates of SNPs with minor allele frequency greater than 1% are that they occur about once every 290 base pairs, suggesting the existence of 11 million SNPs among the 3.2 billion base pairs in the human genome (Kruglyak and Nickerson, 2001) (*see* Article 53, **Information content in gene mapping**, Volume 1). SNPs are also reported to have lower misclassification rates, with estimates in the range of 0.1%–5% (Hao *et al*., 2004;

Tintle *et al.*, 2005). However, as Bonin *et al.* (2004) point out, genotype error rates will vary depending upon the quality of DNA, expertise of laboratory technicians, and other factors.

1.1. Nondifferential and differential misclassification

A critical distinction regarding the nature of misclassification is that of nondifferential versus differential misclassification error rates. Consider Tables 1a and 1b, which present phenotype and genotype misclassification probabilities, respectively. The conditional probabilities are the error model parameters. Nondifferential misclassification means that the error model parameters have the same value in the different cross-classified groups. For example, nondifferential genotype misclassification means that the error model parameters ϵ_{ij} (Table 1b) have the same values in the Affected and Unaffected groups. Similarly, nondifferential phenotype misclassification means that the error model parameters θ and ϕ (Table 1a) have the sample values in the AA, AB, and BB genotype groups. Mathematically, when:

$$\begin{aligned}&\Pr(\text{observed genotype} = \text{a}|\text{true genotype} = \text{b, true phenotype} = \text{Affected})\\&\quad = \Pr(\text{observed genotype}=\text{a}|\text{true genotype}=\text{b, true phenotype}=\text{Unaffected})\\&\quad = \Pr(\text{observed genotype} = \text{a}|\text{true genotype} = \text{b}),\end{aligned}$$

the genotype misclassification is non-differential.

Differential misclassification occurs when the assumption of nondifferential misclassification is false. The consequences of each type of error are documented below.

1.2. Terminology

We shall use the following terms throughout this work:

(GRR) = genotype relative risk. This term was first defined by Schaid and Sommer (1993). In general, the genotype relative risk is defined as the ratio: $R_i = f_i / f_0$, i = 1, 2. Here, f_i = Pr(*affected* | i copies of disease allele at the trait locus).

Table 1a Conditional probabilities for phenotype misclassification

	Observed Phenotype	
True Phenotype	Affected	Unaffected
Affected	$1 - \theta$	θ
Unaffected	ϕ	$1 - \phi$

In this table we present conditional probabilities Pr (observed phenotype = a | true phenotype = b) where a, b are either Affected or Unaffected. For example, θ = Pr (observed phenotype = Unaffected | true phenotype = Affected).

Table 1b Conditional probabilities for genotype misclassification assuming di-allelic

	Observed genotype		
True Genotype	AA	AB	BB
AA	$1 - \epsilon_{12} - \epsilon_{13}$	ϵ_{12}	ϵ_{13}
AB	ϵ_{21}	$1 - \epsilon_{21} - \epsilon_{23}$	ϵ_{23}
BB	ϵ_{31}	ϵ_{32}	$1 - \epsilon_{31} - \epsilon_{32}$

In this table we present conditional probabilities Pr (observed genotype = a | true genotype = b) where a, b are either AA, AB, or BB. For example, ϵ_{21} = Pr (observed genotype = AA | true genotype = AB).

(MSSN) = Minimum Sample Size Necessary. This term refers to the minimum sample size required to detect association when misclassification errors are present (Gordon *et al*., 2002; Kang *et al*., 2004b; Edwards *et al*., 2005). As is documented below, MSSN almost always increases in the presence of misclassification error.

1.3. Pedigree error

While we do not focus on this issue here, we do note that another source of error for genetic studies is pedigree error, a situation where unrelated individuals are labeled as being part of the same pedigree (*see* Article 78, **Genetic family history, pedigree analysis, and risk assessment**, Volume 2). Methods have been developed both to detect unlikely relationships and to determine the impact of pedigree error on statistical linkage and association statistics [e.g., (Goring and Ott, 1997; Sun *et al*., 2001; Sieberts *et al*., 2002; Sun *et al*., 2002)].

1.4. Missing informativeness

Another issue that we mention in passing regards the consequences of nonrandom missing data for linkage and association analysis. Specifically, we refer to the issue of "informative missingness". Allen *et al*. (2003) define informative missingness for genetic studies with pedigree data as the event that a parent's missing genotype is related to his or her genotype at a locus of interest. They further state, "Informative missingness can occur for several reasons. First, alleles at the locus may, in fact, cause or be proximal to a locus that causes the disease of interest, which may lead to differential missingness. For example, in a study of genetic factors in an aggressive form of cancer, parents carrying the disease-predisposing allele may be more likely to be missing. Second, alleles at the locus may cause or be proximal to a locus causing a different disease that results in parental missingness. In an era when the same candidate genes are tested for involvement in a variety of conditions, this coincidence cannot be ruled out. Alternatively, in genome scans, use of a large number of closely spaced markers increases the chance that a marker is linked to some locus that may cause parental missingness by association

with a disease other than the one under study. Finally, if there is population substructure and if the propensity to be missing is correlated with allele frequency in the subpopulations, then the genotype frequencies in the intact trios will not be representative of those among the missing parents." These authors present statistical methods to address informative missingness in case-parent designs (Allen *et al*., 2003). Other authors have also looked at this issue with regard to genetic association (Chen, 2004).

2. Error detection

Phenotype misclassification may be determined by the use of a "gold-standard" diagnosis (e.g., autopsy-proven diagnosis in diseases such as Alzheimer's, Parkinson's, and Multiple Sclerosis (Mayeux *et al*., 1998; Poser and Brinar, 2001; Lansbury, 2004)), by rigorous review of clinical records of all the patients involved in a study (Silverberg *et al*., 2001), or by diagnosis performed by two independent investigators. Autopsy-proven diagnosis will usually not be available on all the patients in a study unless that is part of the study design. Since rigorous review of clinical records is time-consuming and expensive, it may not be feasible to conduct it for all patients. If clinical measurement instruments are subject to higher error rates (say >10%), then even independent diagnoses still have a greater than 1% probability of both being incorrect.

Regarding genotype misclassification, methods of error detection for pedigree data have focused on detecting single-locus Mendelian inconsistencies (O'Connell and Weeks, 1998) or determination of unlikely genotypes through multipoint and other methods (Ehm *et al*., 1996; Douglas *et al*., 2000). Recent research suggests that error detection through Mendelian inconsistency has low and variable power (<70%), especially for SNP genotypes with larger minor allele frequencies (Gordon *et al*., 1999a; Gordon *et al*., 2000; Douglas *et al*., 2002; Geller and Ziegler, 2002). Furthermore, even if a pedigree demonstrates Mendelian inconsistency, unless one uses a gold-standard genotyping technology (which usually is expensive and is not performed on all the subjects) or makes restrictive assumptions about the nature of genotype errors (e.g., (Miller *et al*., 2002)), one cannot determine which genotype is true and which is incorrect. Simulation studies also indicate that a nonsignificant proportion of genotype errors are not detected by multipoint methods (Badzioch *et al*., 2003; Zou *et al*., 2003; Mukhopadhyay *et al*., 2004).

In population-based studies such as case/control genetic association analysis (*see* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1, Article 51, **Choices in gene mapping: populations and family structures**, Volume 1), methods for genotype error detection focus on either repeated sampling (Miller *et al*., 2002; Gordon *et al*., 2004b) or tests for deviation from the Hardy-Weinberg Equilibrium in control subjects (Hosking *et al*., 2004; Ott, 2004). Recent research suggests that tests for deviation from the Hardy-Weinberg Equilibrium in control subjects may require prohibitively large sample sizes (Leal, 2005).

3. Consequences of error

3.1. Association studies

3.1.1. Nondifferential misclassification

For nondifferential misclassification errors, the main consequence of errors is a loss in power to detect association or, equivalently, an increase in sample size necessary to maintain constant power for any given significance level. For phenotype misclassification using 2×2 contingency tables (e.g., case/control and exposure/nonexposure), Bross (1954) showed that there is no change in the level of significance for the chi-square test of independence, that parameter estimates are biased, and that there is a loss in power to detect association. In the field of epidemiology, as reviewed by Thomas *et al*. (1993), the issue of phenotype misclassification has been widely studied.

Two statistical applications for genetic association data are the linear trend statistic (Cochran, 1954; Armitage, 1955) and the chi-square test of independence on $2 \times c$ contingency tables, where c is the number of observed genotypes. Regarding the linear trend test, Zheng and Tian (2005) quantified the sample size requirements to detect association in the presence of phenotype (or diagnostic) misclassification as a function of the genetic model parameters including GRRs (Schaid and Sommer, 1993), disease allele frequency, and disease prevalence K. Edwards *et al*. (2005) considered this issue for the chi-square test of independence on contingency tables. Specifically, they computed sample size requirements in the presence of phenotype misclassification and used a linear Taylor series approximation to determine the relative costs of misclassifying a true affected (respectively, unaffected) as an observed unaffected (respectively, affected). A key finding was that the MSSN coefficient of misclassifying a true affected as an observed unaffected was proportional to:

$$K/(1 - K). \tag{1}$$

Similarly, the MSSN coefficient of misclassifying a true unaffected as an observed affected was proportional to:

$$(1 - K)/K. \tag{2}$$

From formula (2), as the prevalence K in the population being studied approaches 0 (*see* Article 59, **The common disease common variant concept**, Volume 2, Article 60, **Population selection in complex disease gene mapping**, Volume 2, Article 68, **Approach to common chronic disorders of adulthood**, Volume 2), the MSSN coefficient becomes indefinitely large. That is, the sample size requirements to maintain constant power become prohibitively large. Figure 1 illustrates the rapid increase in MSSN for nonzero error rate ϕ = Pr(observed affected | true unaffected) when the prevalence decreases. The sample size requirements go from 100 ($\log_{10} = 2.0$) for $K = 0.069$, $\phi = 0.001$ to 388

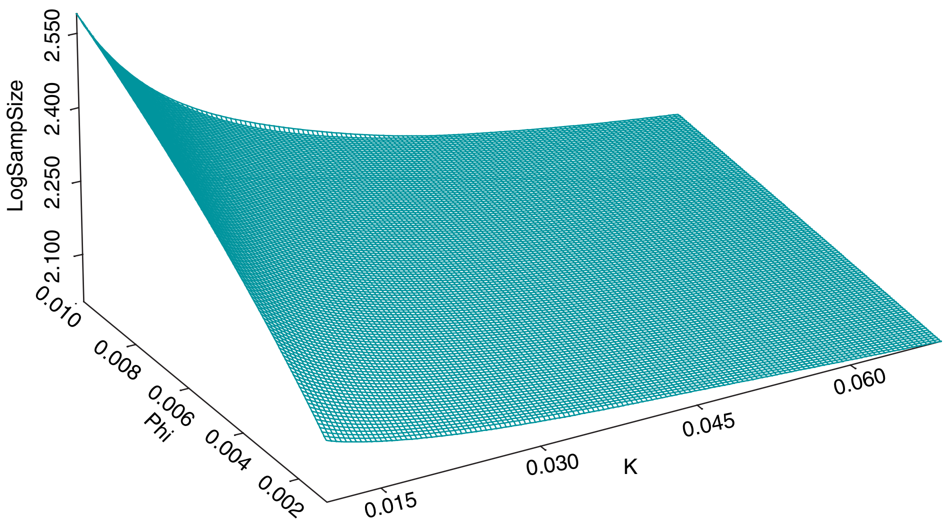


Figure 1 Sample size requirements (log transformed) as function of disease prevalence K and misclassification error rate θ. Total sample size (cases and controls; $\log_{10}$-transformed) is computed as implemented in the PAWE-3D webtool (**http://linkage.rockefeller.edu/pawe3 d/**) using the following genetic model parameter settings: genotype relative risks (GRR) $R_1 = 3$, $R_2 = 10$, Disease allele frequency = 0.25, SNP marker allele frequency = 0.25, r^2 = 1.0, power = 0.95, significance level = 0.05, chi-square test of independence on 2 degrees of freedom, Theta = Pr (observed unaffected | true affected) = 0.01, Phi = Pr (observed affected | true unaffected), and prevalence K . Sample sizes are plotted for Phi in the range [0.001, 0.01] and K in the range [0.01, 0.07]

($\log_{10} = 2.59$) for $K = 0.01$, ϕ=0.01. In general, power and sample size calculations for case/control studies of genetic association may be performed using the *P*ower for *A*ssociation *W*ith *E*rrors webtool (PAWE-3D) (Gordon *et al*., 2005).

Consequences of genotype error are that there is no change in the level of significance for the chi-square test of independence, that parameter estimates are biased, and that there is a loss in power to detect association. However, power loss does *not* depend upon prevalence, as with phenotype error. For the chi-square test of independence, power loss or MSSN for fixed power has been quantified (Mote and Anderson, 1965) and MSSN cost coefficients have been calculated for either a genetic model-based or model-free setting (Kang *et al*., 2004a; Kang *et al*., 2004b). This research indicates that the most costly error for a di-allelic locus such as SNP is misclassification of the more common homozygote as the less common homozygote, with unbounded MSSN coefficient as the minor allele frequency approaches 0. Another genotype error with unbounded MSSN coefficient is misclassification of the more common homozygote as the heterozygote. Kang *et al*. (2004a) demonstrated that for certain recessive genetic models, the MSSN coefficient for misclassification of the heterozygote as the less common homozygote was unbounded as the SNP minor allele frequency and the disease allele frequency approach 0. That is, as the minor allele frequency of SNP locus approaches 0, the sample size requirements to maintain constant power for even a 1% rate of these types of error approaches infinity. Therefore,

researchers should be particularly careful in calling genotypes when SNP minor allele frequencies are small (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1, Article 53, **Information content in gene mapping**, Volume 1).

Finally, genotype error can lead to erroneous inferences when measures of linkage disequilibrium between two samples are compared (Akey *et al*., 2001), to power loss for association when DNA pooling techniques are applied (Zou and Zhao, 2004), and to erroneous inference for population genetics studies based on individual identification (Bonin *et al*., 2004).

3.1.2. Differential misclassification

Differential misclassification is the most damaging type of misclassification from a statistical standpoint. If differential misclassification occurs, one does not know the null distribution of the test statistic unless one knows the error model parameters and can perform simulations under the null hypothesis of no association. So, for example, Bross's finding (1954) that the level of significance for the chi-square test of independence does not change under misclassification may no longer be true when differential misclassification occurs. In a recent publication, Clayton *et al*. (2005) documented a substantial inflation in type I error for a case-control genetic association study of type I diabetes due to differential genotyping misclassification. That is, individual SNP genotype error rates were different in the case and control populations.

How can one avoid such misclassification? One answer is randomization. For example, to avoid differential genotype misclassification, one should not perform genotyping of all cases at one-time and all controls at another time, but rather randomly assign genotyping platform runs independent of affection status.

3.2. Linkage studies

In this section we focus on nondifferential errors, as we are unaware of research that has considered differential misclassification in linkage methods. Phenotype misclassification has been consistently studied in linkage analysis over the past few decades (e.g., (Ott, 1977; Martinez *et al*., 1989; Burd *et al*., 2001; Silverberg *et al*., 2001)). Major findings are that there is loss in power to detect linkage and that estimates of the recombination fraction between disease locus and marker locus are too large, on average. Thus, the existence of phenotype misclassification can make gene mapping more difficult either because initial linkage analyses do not show significant evidence for linkage to disease loci and/or because the position of a disease locus may be several centiMorgans away from its estimated position (Martinez *et al*., 1989).

For genotype errors, a number of authors have looked at the effects in linkage studies. Results are summarized by Sobel *et al*. (2002). Major findings are that errors lead to inflation in genetic map distances (Buetow, 1991; Shields *et al*., 1991; Goldstein *et al*., 1997; Matise *et al*., 2003); an increase in type I error (false positive) rates (Gordon *et al*., 2001; Becker and

Knapp, 2003; Mitchell *et al*., 2003; Knapp and Becker, 2004; Seaman and Holmans, 2005); a decrease in power for statistical methods designed for gene localization (Gordon *et al*., 1999b; Abecasis *et al*., 2001; Gordon and Finch, 2005); biased estimates of parameters such as the recombination fraction among loci (Terwilliger *et al*., 1990); and biased estimates of population frequency parameters such as haplotype frequencies (Kirk and Cardon, 2002; Becker and Knapp, 2003).

3.2.1. Genotype errors in TDT analysis

Of particular importance is the consideration of genotype errors in transmission/disequilibrium test (TDT) linkage analysis (Spielman *et al*., 1993; Spielman and Ewens, 1996). A number of software programs that perform TDT analysis only do so for pedigrees showing Mendelian consistency. Thus, for researchers to run the software, they must "clean" the pedigrees. Cleaning is often done by removing (or "zeroing out") genotypes for individuals causing Mendelian inconsistency (O'Connell and Weeks, 1998; Mukhopadhyay *et al*., 1999). Still other software programs simply ignore those pedigrees that show Mendelian inconsistency. Unfortunately, especially for SNPs, many genotype errors will not result in Mendelian inconsistency (Gordon *et al*., 1999a; Gordon *et al*., 2000; Douglas *et al*., 2002; Geller and Ziegler, 2002). The result is that many pedigrees with genotype errors remain in the data set and are incorrectly considered in the calculation of the TDT statistic. An example is provided in Figures 2a and 2b. In figure 2a, we see that the true genotype status of all individuals is homozygous for the A allele. Since the TDT only considers those pedigrees in which at least one parent is heterozygous (Spielman *et al*., 1993), this pedigree should not be included in the calculation of the TDT statistic, because both parents are homozygous at the marker allele. Consider now the same pedigree after a genotype error is introduced, resulting in the father's marker genotype being recorded as the heterozygote AB. This error will not be detected by checking for Mendelian inconsistency. As a result, this pedigree will incorrectly be included in the calculation of the TDT statistic. The effect of such inclusions is a possibly substantial increase in the false-positive rate to detect linkage (Gordon *et al*., 2001; Mitchell *et al*., 2003). For example, Gordon *et al*. (2001) documented by simulation that, for 500 trio pedigrees, assuming a 5% misclassification of the B allele as the A allele, the false-positive rate at the 1% significance level was 11.8%, an eleven-fold increase. Furthermore, the inflation in false-positive rate increases as the sample size increases, because the absolute number of pedigrees with undetected genotype errors increases (even though the proportion of pedigrees with errors remains constant for a fixed genotype error rate). This issue of genotype error is likely to become more important rather than less for the foreseeable future, as many genetic analysis designs call for larger sample sizes (>500 trio pedigrees) with more than 100,000 genotyped SNPs across the genome (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1, Article 51, **Choices in gene mapping: populations and family structures**, Volume 1).



Figure 2 Example of pedigree that is incorrectly included in TDT analysis. These figures show genotypes for di-allelic locus with alleles A and B in a pedigree (father [open square], mother [open circle], affected daughter [dark circle]) before (a) and after (b) a genotype error is introduced into the pedigree. In this example, the father's genotype is changed from AA to AB

4. Statistical solutions to address challenges arising from phenotype and genotype errors

As noted above, the main challenges arising from nondifferential phenotype and genotype error rates are a loss in power for gene mapping studies, biased estimates of the position of a disease locus, biased estimates of population frequency parameters, and an increase in false-positive rates for TDT. A traditional approach (for nondifferential misclassification only) to address the issue of power loss is to increase sample size (Wong *et al*., 2003). However, methodological research has documented that the sample size requirements may be prohibitively high, especially if the prevalence of the disease is small, the effect size of the susceptibility gene is modest, and/or the misclassification rates are large (e.g., greater than 5%) (Martinez *et al*., 1989; Silverberg *et al*., 2001; Edwards *et al*., 2005; Zheng and Tian, 2005).

4.1. Incorporation of errors into design and analysis

We and others (Badzioch *et al*., 2003; Mukhopadhyay *et al*., 2004) recommend incorporation of misclassification error into the design and analysis of genetic linkage and association data. At the design stage, one can compute more realistic power and sample size values for genetic association studies by incorporation of nonzero phenotype and/or genotype misclassification error rates (Gordon *et al*., 2002, 2005). Given that low disease prevalence is a critical factor in power loss for phenotype misclassification (see equation (2)), a more robust study design is one in which the phenotype of interest is closer to 50% in the population being studied. For example, in pharmacogenetic association studies (*see* Article 73, **The clinical and economic implications of pharmacogenomics**, Volume 2), one can use as the phenotype responders or nonresponders to particular medications and/or treatments, which may have prevalences closer to 0.5. With genotype misclassification, one can perform sample size calculations assuming a larger rather than a smaller power (say, 99% vs. 80%) to protect against substantial power loss. A sample size chosen to have 99% power without errors for a trait locus that displays a dominant mode of inheritance will still have 95.5% power after a 5% genotyping error rate, whereas a sample size chosen to have 80% power without errors for the same mode of

Table 2 Software programs referenced in this work that consider phenotype and/or genotype misclassification error

Program	Description	Reference	URL
LRTae	Perform case/control genetic association analysis allowing for phenotype/genotype error by use of double-sampling	(Gordon *et al*., 2004b)	**ftp://linkage.rockefeller.edu/software/lrtae/**
Mendel/Simwalk	Check for unlikely genotypes in pedigree data using multipoint methods and assuming specific genotype error model	(Sobel and Lange, 1996; Sobel *et al*., 2002)	**http://www.genetics.ucla.edu/software/**
Merlin	Checks for genotype errors in pedigree data through Mendelian inconsistencies and multipoint methods	(Abecasis *et al*., 2002)	**http://www.sph.umich.edu/csg/abecasis/Merlin**
PAWE/PAWE-PH/PAWE-3D	Perform power/sample size calculations for case/control genetic association incorporating phenotype and genotype errors	(Gordon *et al*., 2002; Edwards *et al*., 2005; Gordon *et al*., 2005; Zheng and Tian, 2005)	**http://linkage.rockefeller.edu/pawe/**
PedCheck	Check for Mendelian inconsistencies (genotype errors) in pedigree data	(O'Connell and Weeks, 1998)	**http://watson.hgen.pitt.edu/register**
PseudoMarker	Perform multipoint linkage and association analysis on pedigree data allowing for genetic model misspecification and genotype errors	(Goring and Terwilliger, 2000a; Goring and Terwilliger, 2000b; Goring and Terwilliger 2000c; Goring and Terwilliger 2000 d)	**http://www.helsinki.fi/~tsjuntun/pseudomarker/**
TDTae	Perform TDT linkage analysis allowing for single-locus genotype errors	(Gordon *et al*., 2001; Gordon *et al*., 2004a)	**ftp://linkage.rockefeller.edu/software/tdtae2/**

In this table, we present a list of some currently available software methods to either detect phenotype and genotype errors or to integrate such errors into linkage and association analyses. Another program that is not listed in the table is one developed by Hao *et al*. (2004). Their method estimates genotype error rates from pedigree data. A software program is available by contacting the author (khao@hsph.harvard.edu).

inheritance will only have 65.4% power with the same 5% genotyping error rate (Gordon and Finch, 2005).

Statistical methods have also been developed that allow for genotyping error when estimating haplotype frequencies in either unrelated individuals or nuclear families (Zou and Zhao, 2003).

If a gold-standard phenotype or genotype measurement is available for a subset of individuals, then one can incorporate such information in a likelihood ratio test of association to increase power for gene mapping and also to determine unbiased estimates of genotype frequencies (Tenenbein, 1970; Tenenbein, 1972; Gordon *et al*., 2004b). Freely available software has been developed to compute the test statistic for multilocus genotype data on cases and controls (see Software). If one assumes a single-parameter error model (i.e., all $\epsilon_{ij} = \epsilon$ in Table 2), then one can genotype a subset of individuals for a given SNP and determine an estimate of the genotype error. This estimate can be used in a test of association to determine unbiased estimates of genetic model parameters such as GRRs (Rice and Holmans, 2003).

For linkage analysis with pedigree data, there are several methods to deal with phenotype misclassification that involve some form of averaging over genetic model parameters as a means of treating the bias in the parameters due to phenotype misclassification (Vieland, 1998; Goring and Terwilliger, 2000d). Software using a likelihood approach is available (Goring and Terwilliger, 2000d).

For TDT, statistical methods have been developed that incorporate pedigrees with genotype error into the analysis without inflation of false-positive rates (Bernardinelli *et al*., 2004; Gordon *et al*., 2004a; Morris and Kaplan, 2004). These methods have been successfully applied to gene mapping studies of diseases like psoriasis (Helms *et al*., 2003; Helms *et al*., 2005) and software is available.

5. Software

In table 2, we present a list of available software programs that consider the issue of phenotype and/or genotype misclassification error. These programs may be used at the design stage (power and sample size calculations), the processing stage (detecting genotype errors in pedigree genetic data), or the analysis stage (incorporating phenotype and genotype errors into the statistical analysis of genetic data).

Acknowledgments

The authors gratefully acknowledge grants K01-HG00055 and MH44292 from the National Institutes of Health.

References

Abecasis GR, Cherny SS and Cardon LR (2001) The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, **9**, 130–134.

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.
Akey JM, Zhang K, Xiong M, Doris P and Jin L (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *American Journal of Human Genetics*, **68**, 1447–1456.
Allen AS, Rathouz PJ and Satten GA (2003) Informative missingness in genetic association studies: case-parent designs. *American Journal of Human Genetics*, **72**, 671–680.
Armitage P (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.
Badzioch MD, DeFrance HB and Jarvik GP (2003) An examination of the genotyping error detection function of SIMWALK2. *BMC Genetics*, **4**(1), S40.
Becker T, Knapp M (2003) Comment on "The impact of genotyping error on haplotype reconstruction and frequency estimation". *European Journal of Human Genetics* **11**, 637; author reply 638.
Bernardinelli L, Berzuini C, Seaman S and Holmans P (2004) Bayesian trio models for association in the presence of genotyping errors. *Genetic Epidemiology*, **26**, 70–80.
Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C and Taberlet P (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.
Bross I (1954) Misclassification in 2 × 2 tables. *Biometrics*, **10**, 478–486.
Brzustowicz LM, Merette C, Xie X, Townsend L, Gilliam TC and Ott J (1993) Molecular and statistical approaches to the detection and correction of errors in genotype databases. *American Journal of Human Genetics*, **53**, 1137–1145.
Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. *American Journal of Human Genetics*, **49**, 985–994.
Burd L, Kerbeshian J and Klug MG (2001) Neuropsychiatric genetics: misclassification in linkage studies of phenotype-genotype research. *Journal of Child Neurology*, **16**, 499–504.
Chen YH (2004) New approach to association testing in case-parent designs under informative parental missingness. *Genetic Epidemiology*, **27**, 131–140.
Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE *et al*. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, **37**, 1243–1246.
Cochran WG (1954) Some methods for strengthening the common chi-squared tests. *Biometrics*, **10**, 417–451.
Douglas JA, Boehnke M and Lange K (2000) A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *American Journal of Human Genetics*, **66**, 1287–1297.
Douglas JA, Skol AD and Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics*, **70**, 487–495.
Edwards BJ, Haynes C, Levenstien MA, Finch SJ and Gordon D (2005) Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genetics*, **6**, 18.
Ehm MG, Kimmel M and Cottingham RW Jr (1996) Error detection for genetic data, using likelihood methods. *American Journal of Human Genetics*, **58**, 225–234.
Geller F and Ziegler A (2002) Detection rates for genotyping errors in SNPs using the trio design. *Human Heredity*, **54**, 111–117.
Goldstein DR, Zhao H and Speed TP (1997) The effects of genotyping errors and interference on estimation of genetic distance. *Human Heredity*, **47**, 86–100.
Gordon D and Finch SJ (2005) Factors affecting statistical power in the detection of genetic association. *The Journal of Clinical Investigation*, **115**, 1408–1418.
Gordon D, Finch SJ, Nothnagel M and Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human Heredity*, **54**, 22–33.
Gordon D, Haynes C, Blumenfeld J and Finch SJ (2005) PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits. *Bioinformatics*, **21**, 3935–3937.

Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM and Ott J (2004a) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *European Journal of Human Genetics*, **12**, 752–761.

Gordon D, Heath SC, Liu X and Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics*, **69**, 371–380.

Gordon D, Heath SC and Ott J (1999a) True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Human Heredity*, **49**, 65–70.

Gordon D, Matise TC, Heath SC and Ott J (1999b) Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genetic Epidemiology. Supplement*, **17**, S587–S592.

Gordon D, Leal SM, Heath SC and Ott J (2000) An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. *Pacific Symposium on Biocomputing*, **5**, 663–674.

Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM and Haroutunian V (2004b) Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. *Statistical Applications in Genetics and Molecular Biology*, **3**, 26.

Goring HH and Ott J (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. *European Journal of Human Genetics*, **5**, 69–77.

Goring HH and Terwilliger JD (2000a) Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *American Journal of Human Genetics*, **66**, 1095–1106.

Goring HH and Terwilliger JD (2000b) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *American Journal of Human Genetics*, **66**, 1107–1118.

Goring HH and Terwilliger JD (2000c) Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *American Journal of Human Genetics*, **66**, 1298–1309.

Goring HH and Terwilliger JD (2000 d) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *American Journal of Human Genetics*, **66**, 1310–1327.

Hao K, Li C, Rosenow C and Hung Wong W (2004) Estimation of genotype error rate using samples with pedigree information--an application on the GeneChip Mapping 10K array. *Genomics*, **84**, 623–630.

Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JA, Robarge J, Ott J *et al*. (2003) A putative RUNX1 binding site variant between SLC9A3 R1 and NAT9 is associated with susceptibility to psoriasis. *Nature Genetics*, **35**, 349–356.

Helms C, Saccone NL, Cao L, Daw JA, Cao K, Hsu TM, Taillon-Miller P, Duan S, Gordon D, Pierce B *et al*. (2005) Localization of PSORS1 to a haplotype block harboring HLA-C and distinct from corneodesmosin and HCR. *Human Genetics*, **118**, 446–476.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I and Xu CF (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, **12**, 395–399.

Kang SJ, Finch SJ, Haynes C and Gordon D (2004a) Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. *Human Heredity*, **58**, 139–144.

Kang SJ, Gordon D and Finch SJ (2004b) What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology*, **26**, 132–141.

Kirk KM and Cardon LR (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, **10**, 616–622.

Knapp M and Becker T (2004) Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *American Journal of Human Genetics,* **74**, 589–591; author reply 591–583.

Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**, 234–236.

Lansbury PT Jr. (2004) Back to the future: the 'old-fashioned' way to new medications for neurodegeneration. *Nature Reviews Neuroscience*, **5**, S51–S57.
Leal SM (2005) Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genetic Epidemiology*, **29**, 204–214.
Martinez M, Khlat M, Leboyer M and Clerget-Darpoux F (1989) Performance of linkage analysis under misclassification error when the genetic model is unknown. *Genetic Epidemiology*, **6**, 253–258.
Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S *et al*. (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *American Journal of Human Genetics*, **73**, 271–284.
Mayeux R, Saunders AM, Shea S, Mirra S, Evans D, Roses AD, Hyman BT, Crain B, Tang MX and Phelps CH (1998) Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. *The New England Journal of Medicine*, **338**, 506–511.
Miller CR, Joyce P and Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–366.
Mitchell AA, Cutler DJ and Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics*, **72**, 598–610.
Morris RW and Kaplan NL (2004) Testing for association with a case-parents design in the presence of genotyping errors. *Genetic Epidemiology*, **26**, 142–154.
Mote VL and Anderson RL (1965) An investigation of the effect of misclassification on the properties of chisquare-tests in the analysis of categorical data. *Biometrika*, **52**, 95–109.
Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP and Weeks DE (1999) Mega2, a data-handling program for facilitating genetic linkage and association analyses. *American Journal of Human Genetics*, **65**, A436.
Mukhopadhyay N, Buxbaum SG and Weeks DE (2004) Comparative study of multipoint methods for genotype error detection. *Human Heredity*, **58**, 175–189.
O'Connell JR and Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, **63**, 259–266.
Ott J (1977) Linkage analysis with misclassification at one locus. *Clinical Genetics*, **12**, 119–124.
Ott J (2004) Issues in association analysis: error control in case-control association studies for disease gene discovery. *Human Heredity*, **58**, 171–174.
Page GP, George V, Go RC, Page PZ and Allison DB (2003) "Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *American Journal of Human Genetics*, **73**, 711–719.
Poser CM and Brinar VV (2001) Diagnostic criteria for multiple sclerosis. *Clinical Neurology and Neurosurgery*, **103**, 1–11.
Poser CM and Brinar VV (2004) Diagnostic criteria for multiple sclerosis: an historical review. *Clinical Neurology and Neurosurgery*, **106**, 147–158.
Rice KM and Holmans P (2003) Allowing for genotyping error in analysis of unmatched cases and controls. *Annals of Human Genetics*, **67**, 165–174.
Schaid DJ and Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics*, **53**, 1114–1126.
Seaman SR and Holmans P (2005) Effect of genotyping error on type-I error rate of affected sib pair studies with genotyped parents. *Human Heredity*, **59**, 157–164.
Shields DC, Collins A, Buetow KH and Morton NE (1991) Error filtration, interference, and the human linkage map. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6501–6505.
Sieberts SK, Wijsman EM and Thompson EA (2002) Relationship inference from trios of individuals, in the presence of typing error. *American Journal of Human Genetics*, **70**, 170–180.
Silverberg MS, Daly MJ, Moskovitz DN, Rioux JD, McLeod RS, Cohen Z, Greenberg GR, Hudson TJ, Siminovitch KA and Steinhart AH (2001) Diagnostic misclassification reduces the ability to detect linkage in inflammatory bowel disease genetic studies. *Gut*, **49**, 773–776.

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Sobel E, Papp JC and Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496–508.

Spielman RS and Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics*, **59**, 983–989.

Spielman RS, McGinnis RE and Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, **52**, 506–516.

Sun L, Abney M and McPeek MS (2001) Detection of mis-specified relationships in inbred and outbred pedigrees. *Genetic Epidemiology*, **21**(Suppl 1), S36–S41.

Sun L, Wilder K and McPeek MS (2002) Enhanced pedigree error detection. *Human Heredity*, **54**, 99–110.

Tenenbein A (1970) A double sampling scheme for estimating from binomial data with misclassifications. *Journal of American Statistical Association*, **65**, 1350–1361.

Tenenbein A (1972) A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*, **14**, 187–202.

Terwilliger JD, Weeks DE and Ott J (1990) Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimates of the recombination fraction. *American Journal of Human Genetics*, **47**, A201.

Thomas D, Stram D and Dwyer J (1993) Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annual Review of Public Health*, **14**, 69–93.

Tintle NL, Ahn K, Mendell NR, Gordon D and Finch SJ (2005) Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and Center for Inherited Disease Research. *BMC Genetics*, **6**(Suppl 1), S154.

Vieland VJ (1998) Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage. *American Journal of Human Genetics*, **63**, 947–954.

Wong MY, Day NE, Luan JA, Chan KP and Wareham NJ (2003) The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology*, **32**, 51–57.

Zheng G and Tian X (2005) The impact of diagnostic error on testing genetic association in case-control studies. *Statistics in Medicine*, **24**, 869–882.

Zou G, Pan D and Zhao H (2003) Genotyping error detection through tightly linked markers. *Genetics*, **164**, 1161–1173.

Zou G and Zhao H (2003) Haplotype frequency estimation in the presence of genotyping errors. *Human Heredity*, **56**, 131–138.

Zou G and Zhao H (2004) The impacts of errors in individual genotyping and DNA pooling on association studies. *Genetic Epidemiology*, **26**, 1–10.

Short Specialist Review

Choices in gene mapping: populations and family structures

Toni I. Pollin and Jeffrey R. O'Connell
University of Maryland School of Medicine, Baltimore, MD, USA

1. Introduction

The goal of gene mapping is to find genes whose variation causes or contributes to a phenotype of interest. Thus, choice of population for gene mapping at its most fundamental level depends on the role of the genes in the underlying biochemical pathway(s) responsible for the phenotype. In choosing a population to study, we choose between specific ethnic groups, between populations with varying degrees of genetic isolation, and between sampling schemes (i.e., nuclear families, extended families, sibling-parent trios, sibling pairs, unrelated sets of affected cases and unaffected controls). The mathematical models used to describe genotype–phenotype relationships of susceptibility genes in complex diseases depend on factors such as the population history of the disease, disease prevalence, allele and genotype penetrance, phenocopy rate, gene-by-gene and gene-by-environment interaction, and allelic spectrum of the disease.

While these models may provide guidelines, practical issues such as the phenotyping and genotyping costs and recruitment feasibility often have a greater impact on the populations we choose to study. For example, regardless of the power of discordant sibling pairs to map quantitative traits, ascertaining sufficient numbers may just not be feasible if such pairs are very rare in the population.

Given that our models certainly do not capture the full complexity of the biology and that inferences made from the models depend on the values of the input parameters, it is not surprising that there is debate in the literature regarding how and in whom it is best to map genes for complex diseases. One central issue is whether the allelic architecture (number and diversity of disease-causing alleles) of susceptibility genes for complex diseases is the same as or similar to that for Mendelian diseases. The common disease/common variant (CDCV) hypothesis (Lander, 1996; Risch and Merikangas, 1996) states that many common diseases are caused by alleles that have been driven to high frequency under neutral selection pressure, as the diseases have late onset, or even positive selection pressure in past environments as considered by the thrifty gene hypothesis (Neel, 1961). Examples of genes with common variants associated with complex diseases include the *APOE* gene in Alzheimer's disease (Corder *et al*., 1993) and PPARγ in Type 2 diabetes (Yen *et al*., 1997; Altshuler *et al*., 2000). Under the CDCV hypothesis, linkage

analysis is expected to have limited power compared to testing for association between allelic variants and disease in population samples.

The prospect of genome-wide association mapping is fundamentally tied to the rapid advances in genotyping technology of single-nucleotide polymorphisms (SNPs). The number of SNPs available for high-throughput mapping has increased exponentially in the last several years. One goal of the HapMap Project is to provide a scaffold of SNPs based on the linkage disequilibrium (LD) patterns in the genome that adequately powers association mapping given that the risk allele may not be tested directly. Since required sample sizes depend on the effect size of the risk allele and amount of LD between the risk allele and the tested SNP, many companies, such as deCODE genetics, and governments, such as those of the United Kingdom and Japan, are turning to very large-scale population-based biobanks that contain DNA samples from hundreds of thousands of individuals (Wright *et al*., 2002; Triendl, 2003). The practical and ethical implications of these biobanks, such as the feasibility of obtaining informed consent from individuals for DNA banking for future, often unknown, research purposes, are subjects currently under discussion (MacWilliams, 2003).

2. Advantages of genetic isolates

Genetic isolates have been used quite successfully to map many Mendelian diseases, particularly, rare recessive disorders that become prevalent due to inbreeding, such as Ellis–van Creveld syndrome in the Amish (Ruiz-Perez *et al*., 2000) and Bardet–Biedl syndrome in the Bedouins (Nishimura *et al*., 2001). Although it is still unknown whether genetic isolates will prove to be equally valuable for mapping complex diseases, there have been a number of recent articles in the literature emphasizing the potential advantages of these populations (Ober and Cox, 1998; Wright *et al*., 1999; Peltonen *et al*., 2000; Shifman and Darvasi, 2001; Heutink and Oostra, 2002).

Some clear advantages are genetic homogeneity resulting from a small number of founder chromosomes and environmental homogeneity arising from common cultural or religious practices or geographical isolation. Genetic homogeneity reduces the number of susceptibility alleles segregating in the population, and environmental homogeneity reduces the proportion of the total variance in a trait attributable to nongenetic factors. Longitudinal studies, which can provide a clearer understanding of phenotypes that are influenced by developmental processes and gene-by-environment interactions, may be easier to perform in isolates that are also geographically localized. This geographical localization combined with other features such as cultural homogeneity has in some cases afforded researchers the opportunity to develop mutually beneficial, culturally sensitive, collaborative relationships with genetic isolate communities; perhaps the most notable example is the involvement in genetic research of the Old Order Amish over the past 40 years (McKusick *et al*., 1964; Francomano *et al*., 2003). Under the CDCV hypothesis, susceptibility alleles that are present in genetic isolates should be relevant to the outbred populations from which the founders came, as there has not been sufficient time for these alleles to be replaced by new variants. This hypothesis is supported

by recent empiric data showing that estimated frequencies of SNPs in candidate genes for cardiovascular risk were indeed similar between the Hutterites and CEPH families (Newman *et al*., 2004). Linkage disequilibrium generally extends over greater distances in younger genetic isolates, therefore requiring fewer SNPs for linkage disequilibrium (LD) mapping.

However, we are of the opinion that the true power of genetic isolates for studying complex disease has yet to be tapped, as their power lies not in the genotyping technologies of the HapMap Project to exploit LD mapping but rather in the sequencing technologies of the Personal Genome Project that promise to provide the $1000 sequence in 24 hours. Many of these revolutionary technologies are based on genotyping single molecules of DNA (Kwok and Xiao, 2004), and therefore will provide molecular haplotypes. Because haplotype phase can be at least partially determined within pedigrees when using standard diploid genotyping, interest in the molecular haplotype has been focused primarily on LD mapping in case-control studies. However, the molecular haplotype perhaps has its greatest power in eliminating the computational barriers faced in linkage and association analysis within genetic isolates.

The power of the molecular haplotype has actually already been used in genetic isolates such as the Bedouins, Finnish, and Old Order Amish for mapping rare recessive diseases that arise from inbreeding in the form of homozygosity mapping. Homozygosity mapping is based on the principle that an affected individual homozygous for a rare deleterious allele will be homozygous at nearby flanking polymorphic markers if the disease alleles were inherited identical by descent from a common ancestor (Lander and Botstein, 1987). Determining regions of homozygosity is equivalent to determining shared ancestral molecular haplotypes. The regions can be inferred with dense genotyping of polymorphic markers to detect sequences of homozygous genotypes whose length is greater than that expected by chance given the allele frequencies, with the heterozygous genotypes that interrupt the sequence defining the boundaries.

For complex diseases in genetic isolates, the principle is the same; that is, affected individuals carrying an ancestral risk allele will be enriched for a portion of the ancestral haplotype containing the risk allele, and the enrichment of the risk allele results from the multilineal paths of descent created by inbreeding. However, for complex diseases, the risk haplotype may be neither necessary nor sufficient to develop disease. Given that two haploid chromosomes differ by, on average, one nucleotide per kilobase, single-molecule genotyping will allow us to directly infer regions of identity by descent (IBD) between any two chromosomes on the basis of the length of sequence similarity.

This technology will eliminate the major computational obstacle of computing mulitipoint IBD sharing probabilities in pedigrees from young genetic isolates, such as the Old Order Amish, with 10–14 generations, thousands of individuals, and often hundreds of inbreeding loops (O'Connell, 2003). Direct IBD between two haploids is inferred from matching sequence rather than from a likelihood calculation on a complex ancestral pedigree structure, often comprising 5–7 generations with no genotype information. The genealogy is only required for the unconditional IBD sharing probabilities, which are straightforward to compute. Haploid genotyping will potentially allow such pedigrees to be analyzed in their

entirety, thereby incorporating a much broader and deeper recombination history to reduce the length of shared ancestral haplotypes that contain susceptibility genes to complex disease.

Certainly, no single population or gene mapping approach will provide all the answers to our questions about the genetics of complex diseases, but the prospect of single-molecule technologies providing large-scale affordable sequencing on the haploid background will unlock the true richness of genetic isolates in the study of complex diseases. Thus, we would advocate establishing biobanks specific to genetic isolates, a resource likely enhanced by the close-knit nature of these populations and our consequent ability to work closely with them as collaborators, to complement the growing number of outbred population-based biobanks.

Related articles

Article 58, **Concept of complex trait genetics**, Volume 2; Article 59, **The common disease common variant concept**, Volume 2; Article 60, **Population selection in complex disease gene mapping**, Volume 2

References

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, *et al*. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, **26**, 76–80.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL and Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923.

Francomano CA, McKusick VA and Biesecker LG (2003) Medical genetic studies in the Amish: historical perspective. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, **121**, 1–4.

Heutink P and Oostra BA (2002) Gene finding in genetically isolated populations. *Human Molecular Genetics*, **11**, 2507–2515.

Kwok PY and Xiao M (2004) Single-molecule analysis for molecular haplotyping. *Human Mutation*, **23**, 442–446.

Lander ES (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.

Lander ES and Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

MacWilliams B (2003) Banking on DNA: Estonia's genetic database promises medical advances–maybe. *The Chronicle of Higher Education*, **49**, A16, A18.

McKusick VA, Hostetler JA and Egeland JA (1964) Genetic studies of the Amish, background and potentialities. *Bulletin of the Johns Hopkins Hospital*, **115**, 203–222.

Neel JV (1961) The hemoglobin genes: a remarkable example of the clustering of related genetic functions on a single mammalian chromosome. *Blood*, **18**, 769–777.

Newman DL, Hoffjan S, Bourgain C, Abney M, Nicolae RI, Profits ET, Grow MA, Walker K, Steiner L, Parry R, *et al*. (2004) Are common disease susceptibility alleles the same in outbred and founder populations? *European Journal of Human Genetics*, **12**, 584–590.

Nishimura DY, Searby CC, Carmi R, Elbedour K, Van Maldergem L, Fulton AB, Lam BL, Powell BR, Swiderski RE, Bugge KE, *et al*. (2001) Positional cloning of a novel gene on chromosome 16q causing Bardet-Biedl syndrome (BBS2). *Human Molecular Genetics*, **10**, 865–874.

Ober C and Cox NJ (1998) The genetics of asthma. Mapping genes for complex traits in founder populations. *Clinical and Experimental Allergy*, **28**(Suppl 1), 101–105; discussion 108–110.

O'Connell JR (2003) Solving the multipoint likelihood problem using haploid genotyping and likelihood factorization. *American Journal of Human Genetics*, **73**, A193.

Peltonen L, Palotie A and Lange K (2000) Use of population isolates for mapping complex traits. *Nature Reviews. Genetics*, **1**, 182–190.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Ruiz-Perez VL, Ide SE, Strom TM, Lorenz B, Wilson D, Woods K, King L, Francomano C, Freisinger P, Spranger S, *et al*. (2000) Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodental dysostosis. *Nature Genetics*, **24**, 283–286.

Shifman S and Darvasi A (2001) The value of isolated populations. *Nature Genetics*, **28**, 309–310.

Triendl R (2003) Japan launches controversial biobank project. *Nature Medicine*, **9**, 982.

Wright AF, Carothers AD and Campbell H (2002) Gene-environment interactions–the biobank UK study. *The Pharmacogenomics Journal*, **2**, 75–82.

Wright AF, Carothers AD and Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23**, 397–404.

Yen CJ, Beamer BA, Negri C, Silver K, Brown KA, Yarnall DP, Burns DK, Roth J and Shuldiner AR (1997) Molecular scanning of the human peroxisome proliferator activated receptor gamma (hPPAR gamma) gene in diabetic Caucasians: identification of a Pro12Ala PPAR gamma 2 missense mutation. *Biochemical and Biophysical Research Communications*, **241**, 270–274.

Short Specialist Review

Algorithmic improvements in gene mapping

Gonçalo R. Abecasis and Yu Zhao
University of Michigan, Ann Arbor, MI, USA

1. Introduction

The analysis of human gene mapping data has generated many challenging computational problems. The challenges arise because in most gene-mapping studies the DNA sequence of each individual is only measured imperfectly. For some individuals, these measurements are the result of genotyping assays at specific loci or chromosomal regions. For other individuals, there might be even greater uncertainty: their DNA sequence might only be measured indirectly through information obtained on the genotypes of their relatives. In either situation, there can be a very large number of DNA sequences compatible with observed data, and identifying the most likely DNA sequence configuration(s) might require many individuals to be considered jointly.

2. Analysis of human pedigrees

To survey algorithmic challenges in gene mapping, we will focus on the analysis of pedigree data. These data are often used in linkage studies of discrete or quantitative traits, in the construction of genetic linkage maps (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1), in quality assessments for genotyping data, or to identify individual haplotypes. Maximum likelihood can solve these and other problems related to the analysis of pedigree data, and many algorithms have been developed to calculate likelihoods for human pedigrees. Briefly, the likelihood of interest can be written as:

$$L(\text{data}) = \sum_{G_l} \ldots \sum_{G_n} \prod_i P(X_i|G_i) \prod_{\text{founder}} P(G_{\text{founder}}) \prod_{\{o,f,m\}} P(G_o|G_f, G_m) \quad (1)$$

This likelihood involves a nested summation over the set of possible haplo-genotypes, G_i, for each individual. The likelihood of each possible configuration is a product with factors denoting (1) the probability of observed phenotypes conditional on each individual haplo-genotype, $P(X_i|G_i)$; (2) the probability of

founder haplo-genotypes, $P(G_{\text{founder}})$; and (3) the probability of offspring haplo-genotypes conditional on parental haplo-genotypes, $P(G_o|G_f, G_m)$, calculated for all parent-offspring triples.

Direct evaluation of this nested sum is only possible in the simplest of cases, involving a very small number of loci and individuals. The number of summation terms to be evaluated increases exponentially with both the number of individuals in the pedigree (which add extra levels to the nested sum) and the number of markers being considered (which increase the number of possible haplo-genotypes for each individual).

In early gene-mapping studies, investigators painstakingly evaluated likelihoods for each pedigree examined, using careful algebra to factor the calculation and identify repeated terms. Gene mapping was a new field, laboratory methods were rudimentary allowing the use of only small amounts of data, and this laborious approach was adequate.

3. The Elston–Stewart and related algorithms

Elston and Stewart (1971) developed the first general algorithm for rapid pedigree likelihood calculation. They showed that the likelihood could be updated gradually, one nuclear family at a time. Each update required iterating over possible genotypes for individuals in the nuclear family, resulting in a relatively small nested sum. Their strategy proved highly effective, and their algorithm is still the method of choice for the analysis of large, noninbred pedigrees. Their method was later implemented in LIPED, the first widely available automated software for pedigree likelihood calculation (Ott, 1976), which played a crucial role in enabling the gene-mapping revolution.

Many improvements to the basic algorithm have been proposed. For example, Cannings *et al*. (1978) showed how the method could – in theory – be applied to complex pedigrees, even with inbreeding. Lange and Boehnke (1983) showed that the likelihood could be updated one individual, rather than one nuclear family at a time, and that different sequences of updates could produce dramatically different computing time and memory requirements. With these improved formulations, the complexity of calculating likelihoods for most noninbred pedigrees increases linearly with the number of individuals in a family, and likelihoods can be calculated for very large pedigrees, including hundreds of individuals. Another important enhancement was the development of algorithms for identifying sets of haplo-genotypes for each individual compatible with the observed genotypes for each family (Lange and Goradia, 1987).

In parallel to these algorithmic improvements, more sophisticated computer implementations of the Elston–Stewart algorithm were developed. The LINKAGE computer package (Lathrop *et al*., 1984; Lathrop *et al*., 1985) enabled geneticists to analyze multiple marker loci jointly. Together with the discovery of highly polymorphic VNTR and microsatellite markers, LINKAGE enabled the localization of genes for many Mendelian disorders through multilocus linkage analysis in relatively large pedigrees.

In the 1990s, further enhancements to the Elston–Stewart algorithm were discovered. Cottingham *et al*. (1993) used improved software engineering techniques, such as caching and replacement of floating point with integer operations, to speed up LINKAGE by about one order of magnitude. O'Connell and Weeks (1995) showed that combining alleles that do not appear in an individual's descendants in a single meta-allele could dramatically reduce the number of distinct haplo-genotypes and further speed up calculation.

Despite these enhancements, the complexity of likelihood calculations using the Elston–Stewart algorithm grows exponentially with the number of marker loci considered. State-of-the-art implementations of the Elston–Stewart algorithm in the VITESSE (O'Connell and Weeks, 1995) and FASTLINK (Cottingham *et al*., 1993) computer packages cannot handle more than 5–10 marker loci at a time. The ability of geneticists to rapidly collect data for hundreds of microsatellite markers and the interest in complex disease gene mapping using large collections of small pedigrees shifted the focus to a different collection of algorithms.

4. The Lander–Green and related algorithms

Lander and Green (1987) proposed a very different strategy for pedigree likelihood calculations. Their approach is based on the use of inheritance vectors, which summarize inheritance at specific genomic location. They showed that the probability of observed genotypic or phenotypic data can be calculated for any particular inheritance vector and that, in the absence of genetic interference, inheritance vectors form a Markov Chain along the chromosome. Using a Hidden Markov Model, they proposed an algorithm for the calculation of pedigree likelihoods whose complexity increased only linearly with the number of markers. The algorithm is suitable for very large numbers of markers, but limited to relatively small pedigrees because the number of possible inheritance vectors increases exponentially with pedigree size.

As with the Elston–Stewart algorithm, many enhancements were later discovered, and progressively more powerful computer implementations contributed to the success of countless gene-mapping studies. One important enhancement resulted from the observation that there are many redundancies within inheritance vector space so that inheritance vectors can be grouped to speed up calculation. Over the years, progressively more sophisticated strategies were developed for identifying these redundancies, first focusing on symmetries resulting from the transmission of alleles from single founders (Kruglyak *et al*., 1996), then founder couples (Gudbjartsson *et al*., 2000), and later from other individuals in the pedigree (Markianos *et al*., 2001; Abecasis *et al*., 2002). Another important series of improvements focused on the manipulation of transition matrices, used for the calculation of conditional inheritance vector distributions at neighboring locations, a key step in the Markov Chain. Two distinct approaches have proved very successful at speeding up this step of the calculation: either a divide-and-conquer algorithm (Idury and Elston, 1997) or Fast Fourier Transform (Kruglyak and Lander, 1998) can reduce the computational cost of generating these conditional distributions by several orders of magnitude.

Popular implementations of the Lander–Green algorithm include the computer packages GENEHUNTER (Kruglyak *et al*., 1996; Markianos *et al*., 2001), ALLEGRO (Gudbjartsson *et al*., 2000), and MERLIN (Abecasis *et al*., 2002). All these packages can handle very large numbers of markers and allow the estimation of individual haplotypes or the analysis of quantitative and discrete traits, providing parametric and nonparametric linkage tests. They also provide more specialized algorithms including, for example, algorithms that estimate information content along the genome. Relative information content can highlight areas where genotyping additional markers would provide the greatest information gain (*see* Article 53, **Information content in gene mapping**, Volume 1). In addition to these standard features, the newer packages can generate simulated datasets (commonly used for calculating empirical significance levels), carry out more accurate linkage tests (Kong and Cox, 1997; Sham *et al*., 2002), and even identify likely genotyping errors (Abecasis *et al*., 2004).

Although current implementations of the Lander–Green algorithm can comfortably handle hundreds and even thousands of genetic markers, advances in laboratory technology are already highlighting a need for even more powerful methods. The shift to SNP (single-nucleotide polymorphism) markers and very large scale genotyping has generated datasets with hundreds of thousands of markers measured per individual, with substantial amounts of linkage disequilibrium between neighboring markers (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1). It is likely that further enhancements to gene-mapping algorithms will be forthcoming to allow the analysis of these new datasets.

5. Markov-chain Monte-Carlo algorithms

While the Elston–Stewart and related algorithms can handle a small number of markers in very large noninbred pedigrees and the Lander–Green and related algorithms can handle very large numbers of markers in small pedigrees, neither approach can handle a large number of markers in a large pedigree. Very large pedigrees arise in many interesting settings, most often in the study of isolated populations (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1). It is often desirable to analyze multiple genetic markers in these pedigrees to clarify inheritance patterns when genotype data are not available for individuals in the early generations. The analysis of these most challenging datasets has motivated the development of Monte-Carlo-based methods, which try to identify the most important terms in the pedigree likelihood but avoid summing over all possible terms.

A variety of Monte-Carlo approaches have been employed successfully in linkage analysis, including Simulated Annealing (Sobel and Lange, 1996, implemented in the SIMWALK2 computer program), the Gibbs sampler (Heath, 1997, implemented in the LOKI computer program), and Sequential Imputation (Irwin *et al*., 1994). In addition to the ability to handle very large datasets, these software packages often provide capabilities not currently available in packages based on the Elston–Stewart or Lander–Green algorithms. For example, LOKI (Heath, 1997) can model the contributions of multiple susceptibility loci simultaneously and

SIMWALK2 (Sobel and Lange, 1996; Sobel *et al*., 2002) can model genotyping error explicitly.

6. Outlook for the future

While this is an incomplete account of all the algorithms developed for the linkage analysis of human pedigrees, we have attempted to emphasize those algorithms and developments that entered widespread use through the availability of easy-to-use computer programs. We have certainly missed some packages and ideas that deserve credit, as well as some research paths and strategies that were tried along the way but never become popular in practice. Currently, one promising avenue appears to be the use of Bayesian Networks (Jensen, 1996). These allow complex likelihoods to be evaluated gradually, and provide for a more flexible updating scheme than the Lander–Green or Elston–Stewart algorithms, which conduct updates considering either all individuals (for one locus) or all loci (for one or more individuals) at a time.

The past 20–30 years have produced many algorithmic advances in the analysis of human pedigrees, and these have enabled geneticists to extract the full benefits of new laboratory methods that allow the collection of increasing amounts of genetic information on increasing samples of individuals. Whereas initial methods focused on the analysis of single genetic markers and simple Mendelian traits, more modern methods can analyze very large numbers of genetic markers and individuals and have led to some promising results in the analysis of even complex traits such as diabetes, asthma, and psychiatric disorders. It is tempting to speculate that, with the increasing emphasis on genetic association studies and fine-mapping data (Cardon and Abecasis, 2003), the next decade will produce similar advances in algorithms for the estimation and analysis of haplotypes . . . , but we will leave that story for the 2nd edition!

Further reading

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, **7**, 111–122.

References

Abecasis GR, Burt RA, Hall D, Bochum S, Doheny KF, Lundy SL, Torrington M, Roos JL, Gogos JA and Karayiorgou M (2004) Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *American Journal of Human Genetics*, **74**, 403–417.

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Cannings C, Thompson EA and Skolnick MH (1978) Probability functions on complex pedigrees. *Advances in Applied Probability*, **1**, 26–61.

Cardon LR and Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, **19**, 135–140.

Cottingham RW Jr, Idury RM and Schaffer AA (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics*, **53**, 252–263.

Elston RC and Stewart J (1971) A general model for the genetic analysis of pedigree data. *Human Heredity*, **21**, 523–542.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

Idury RM and Elston RC (1997) A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. *Human Heredity*, **47**, 197–202.

Irwin M, Cox NJ and Kong A (1994) Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 11684–11688.

Jensen FV (1996) *An Introduction to Bayesian Networks*, University College Press: London.

Kong A and Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, **61**, 1179–1188.

Kruglyak L, Daly MJ, ReeveDaly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Kruglyak L and Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, **5**, 1–7.

Lander ES and Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363–2367.

Lange K and Boehnke M (1983) Extensions to pedigree analysis. V. optimal calculation of Mendelian likelihoods. *Human Heredity*, **33**, 291–301.

Lange K and Goradia TM (1987) An algorithm for automatic genotype elimination. *American Journal of Human Genetics*, **40**, 250–256.

Lathrop GM, Lalouel J, Julier C and Ott J (1984) Strategies for multilocus linkage in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3443–3446.

Lathrop GM, Lalouel JM, Julier C and Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics*, **37**, 482–498.

Markianos K, Daly MJ and Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, **68**, 963–977.

O'Connell JR and Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

Ott J (1976) A computer program for general linkage analysis of human pedigrees. *American Journal of Human Genetics*, **26**, 588–597.

Sham PC, Purcell S, Cherny SS and Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics*, **71**, 238–253.

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Sobel E, Papp JC and Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496–508.

Short Specialist Review

Information content in gene mapping

Dan L. Nicolae
The University of Chicago, Chicago, IL, USA

1. Introduction

All genome-wide linkage screens carried out on qualitative and quantitative traits as well as most of the association studies extract only part of the underlying information. Missing information can be the result of different sources, such as absence of DNA samples, missing genotypes, spacing between markers, noninformativeness of the markers, or unknown haplotype phase. The information is never complete in linkage studies but, at least in theory, the amount of missing information can be made arbitrarily low by increasing the number of markers in a region. Knowing how much information is missing from the dataset is important for selecting follow-up strategies and assessing the chances of increasing the significance by reducing the amount of missing information. A related question in linkage studies is the relative efficiency between two ways of increasing information–collecting more families or adding more markers.

There are several measures of information that are used in practice. They can be divided into two main categories. Averaged measures are a function of the data structure and allele frequencies (Botstein *et al*., 1980; Teng and Siegmund 1998; Guo and Elston 1999), and are used to evaluate how informative the data would be. The measures that are conditional on the observed data are a function of the data structure, allele frequencies, and observed genotypes (Kruglyak *et al*., 1996; Nicolae and Kong 2004), and are used to determine the amount of information extracted by the available data. The former measures can be used in designing linkage studies. The latter measures can be used in both the design of the study and of the follow-ups, and in the interpretation of the results.

2. Measures of information

The Polymorphism Information Content (Botstein *et al*., 1980) is defined as the probability to deduce which allele was inherited from a heterozygous parent. They proposed to use this in localizing genes involved in a rare dominant disease, and

they showed that, under Hardy–Weinberg equilibrium,

$$\mathrm{PIC} = \sum_{i=1}^{k} p_i^2 - 2\sum_{i=1}^{k-1}\sum_{j=1}^{k} p_i^2 p_j^2 \tag{1}$$

where p_i is the frequency of the ith allele.

The Linkage Information Content (LIC), introduced by Guo and Elston 1999, is defined as the probability of knowing the number of alleles shared identical by descent (IBD) by a particular pair at a given marker. LIC values are derived for full sib, half sib, grandparent–grandchild, first cousin, and avuncular pairs as a function of the allelic frequencies (Guo and Elston 1999; Guo *et al*., 2002).

In many applications, it is important to measure the information contained in the data, which is relevant for deciding whether certain regions have to be enriched by additional genotyping. Also, the majority of the current genome-wide linkage scans use multipoint calculations (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1) to extract the IBD information from multiple markers. Because the information measures described above are not adequate for these cases, different approaches have been introduced (Kruglyak *et al*., 1996; Nicolae and Kong 2004). In general, the marker information of all the loci on the chromosome is used to calculate a probability distribution on the space of inheritance vectors. For each locus, an inheritance vector is a binary vector that specifies, for all the nonfounding members of the pedigree, which grandparental alleles are inherited. Under the assumption of no linkage of phenotype to the chromosome, all inheritance vectors are equally likely, which leads to a uniform prior distribution on their space. Assuming no interference, a Hidden Markov Model (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1) can be used to calculate the inheritance distribution conditional on the genotypes at all marker loci. The distribution of the inheritance vectors conditional on the observed data is the basis of the statistical inference. In particular, this distribution is used to determine the conditional distribution of the number of alleles IBD at a given location.

In general, the data consist of n pedigrees with genotype data on some markers for some pedigree members. For a position t and for pedigree i, let $\omega_i = \omega_i(t)$ denote the inheritance vector. The information content of inheritance vectors was introduced by Kruglyak *et al*. (1996) based on the entropy (Shannon 1948) as

$$I_{\mathrm{E}}^{i} = 1 - \frac{E^i(t)}{E_0^i} = 1 - \frac{-\sum_{\omega_i} P_0(\omega_i|\mathrm{data}) \log_2 P_0(\omega_i|\mathrm{data})}{-\sum_{\omega_i} P_0(\omega_i) \log_2 P_0(\omega_i)} \tag{2}$$

where probabilities are calculated under Mendelian inheritance. This measure of information always lies between 0 and 1, where 1 indicates perfect information and 0 indicates total uncertainty. The definition of the overall information content of a dataset is based on the global entropy, which, summed over all n pedigrees is

defined as

$$I_{\mathrm{E}} = 1 - \frac{E(t)}{E_0} = 1 - \frac{\sum_{i=1}^{n} E^i(t)}{\sum_{i=1}^{n} E_0^i} = \frac{\sum_{i=1}^{n} E_0^i I_{\mathrm{E}}^i}{\sum_{i=1}^{n} E_0^i} \quad (3)$$

Thus, the global information is a weighted average of the family information with weights proportional to E_0^i. These weights depend on the family size; they are equal to the dimension of the inheritance vector for the corresponding family. Note that this measure does not depend on the linkage test used. This might be a disadvantage when the linkage test statistic is a function of the IBD process as I_{E} can be smaller than 1 even if the IBD pattern is known with certainty.

The "model-free" linkage analysis (*see* Article 48, **Parametric versus non-parametric and two-point versus multipoint: controversies in gene mapping**, Volume 1) usually involves a scoring function that is defined on the basis of the IBD sharing among the affected individuals at locus t, $S_i(t)$. In general, the scoring function $S_i = S_i(t)$ can be any function that has a higher expected value under linkage than no linkage. The standardized form of S_i is defined as

$$Z_i = \frac{S_i - \mu_i}{\sigma_i} \quad (4)$$

where $\mu_i = E(S_i|H_0)$ and $\sigma_i^2 = \mathrm{Var}(S_i|H_0)$ are calculated under the null hypothesis of no linkage. The test for linkage can be done using a linear combination

$$Z = \frac{\sum_{i=1}^{n} \gamma_i Z_i}{\sqrt{\sum_{i=1}^{n} \gamma_i^2}} \quad (5)$$

where $\gamma_i \geq 0$ are the weighting factors assigned to the individual families.

In general, information on descent is incomplete and the Z_i's, and hence Z, are not fully determined by the data. The test statistics are calculated on the basis of the probability distribution over the space of inheritance vectors conditional on the genotypes. The missing information can be handled using an artificial likelihood construct, namely, exponential tilting (Kong and Cox 1997)

$$P_\delta(Z_i = z) = P_0(Z_i = z)c_i(\delta)\exp(\delta\gamma_i z) \quad (6)$$

where $c_i(\delta)$ is the renormalization constant. The evidence for linkage is summarized by a lod score that is defined as

$$\mathrm{lod} = \frac{l(\hat{\delta}) - l(0)}{\log(10)} \quad (7)$$

where $\hat{\delta}$ denotes the maximum likelihood estimator for δ.

Let Z(δ) denote the value of the complete information standardized score 5 that leads to a maximum likelihood estimate equal to δ. Note that, in the case of complete IBD information, there is a monotonic one-to-one correspondence between the sufficient statistic Z, the maximum likelihood estimate of δ, and the lod score (*see* Article 56, **Computation of LOD scores**, Volume 1). Let $\text{lod}^*(Z = Z(\hat{\delta}))$ be the exponential lod score calculated assuming complete information exists and assuming that the maximum likelihood estimate of δ is the same as the one obtained with partial information. This lod score can be interpreted as the predicted lod score in the case of complete information. The difference between the observed lod score and the predicted lod score can be interpreted as the amount of missing information in the data

$$\text{RI}_\text{P} = \frac{\text{lod}}{\text{lod}^*(Z = Z(\hat{\delta}))} \tag{8}$$

and can be used as a measure of relative information (Nicolae and Kong 2004). It is bounded between 0 and 1, equaling 1 in the case of complete information, and it is natural to define it as 0 in the case of total uncertainty (note that both lods are equal to zero when the data are totally uninformative and the value of the measure in this case is obtained by taking a limit). This measure is conditional on the observed data and is dependent on the test statistic used in the linkage test. Follow-up strategies can be designed using this measure. For example, suppose we are interested in a given location with $\text{RI}_\text{P} = 0.5$. The lod score will double if we add markers that make the IBD process known, and the value of the maximum likelihood estimate of δ does not change. Also, the lod score will double if we double the number of families such that the sharing signal and the information is the same in the additional families as it was in the original families.

Likelihood-based measures of relative information can be applied to other areas of gene mapping. For example, consider haplotype-based case-control association studies (*see* Article 12, **Haplotype mapping**, Volume 3). The test for identifying haplotype groups that confer different risks can be done using likelihood ratio statistics on well-chosen models (Gretarsdottir *et al*., 2003). Let Λ denote the likelihood ratio statistic calculated from the genotype data and Λ^* denote the likelihood ratio statistic that would have been obtained assuming no uncertainty in the haplotype counts and that the maximum likelihood estimates stay the same. To measure the relative information, one can use (Nicolae and Kong 2004)

$$\text{RI}_\text{H} = \frac{\Lambda}{\Lambda^*} \tag{9}$$

The value of $1 - \text{RI}_\text{H}$ is a measure of the proportion of information missing due to uncertainty with phase and/or missing genotypes.

The LIC values are implemented in a software called POLYMORPHISM (Niu *et al*., 2001). The entropy-based measure I_E can be calculated using GENEHUNTER (Kruglyak *et al*., 1996). The RI_P measure is implemented in the software ALLEGRO (Gudbjartsson *et al*., 2000). The information content in haplotype association studies, RI_H, can be calculated with NEMO (Gretarsdottir *et al*., 2003).

References

Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **3**, 314–331.

Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, Jonsdottir T, Gudmundsdottir T, Bjarnadottir SM, Einarsson OB, Gudjonsdottir HM, *et al*. (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genetics*, **35**, 131–138.

Gudbjartsson, DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Guo X and Elston RC (1999) Linkage information content of polymorphic markers. *Human Heredity*, **49**, 2: 112–118.

Guo X, Olson JM, Elston RC, Niu T (2002) The linkage information content of polymorphic genetic markers in model-free linkage analysis. *Human Heredity*, **53**, 1: 45–48.

Kong A and Cox NJ (1997) Allele-Sharing Models: LOD Scores and Accurate Linkage Tests. *American Journal of Human Genetics*, **61**, 1179–1188.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Nicolae Dl and Kong A (2004) Measuring the relative information in allele-sharing linkage studies. *Biometrics*, **60**, 368–375.

Niu T, Struk B and Lindpaintner K (2001) Statistical considerations for genome-wide scans: design and application of a novel software package POLYMORPHISM. *Human Heredity*, **52**, 102–109.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.

Teng J and Siegmund DO (1998) Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, **54**, 1247–1265.

Short Specialist Review

Sex-specific maps and consequences for linkage mapping

Solveig K. Sieberts
Rosetta Inpharmatics LLC, Seattle, WA, USA

Daniel F. Gudbjartsson
deCODE Genetics, Reykjavík, Iceland

1. Introduction

From the beginning of linkage mapping, differential recombination rates have been observed in the sexes (Haldane, 1922; Huxley, 1928). In the extreme case, Drosophila males exhibit no recombination during meiosis (Morgan, 1914), while females do. In most other species for which linkage maps exist, this sex difference is more moderate. For mammals, excluding humans, published genome-wide female-to-male ratios of genetic map distances vary between 1.55 : 1.0 and 1.0 : 1.2, with a typical lengthening of female map with respect to the male map (Barendse *et al*., 1994; Dietrich *et al*., 1996; Mikawa *et al*., 1999; Neff *et al*., 1999; Maddox *et al*., 2001). The exceptions include bovine, for which there appears to be no sex difference (Barendse *et al*., 1994), and sheep, for which there is a higher recombination rate in males than in females (Maddox *et al*., 2001).

2. Sex-specific maps in humans

In humans, females also tend to show an increase in recombination relative to males. This was first demonstrated by Renwick and Schulze (1965) and has been verified through the construction of genome-wide sex-specific genetic maps (Donis-Keller *et al*., 1987; Broman *et al*., 1998; Kong *et al*., 2002). The deCODE map, which is the most recent human genetic map, to date, estimates the overall female-to-male map distance ratio to be 1.65 : 1.0. The distance ratio varies by chromosome from 1.34 : 1.0 to 1.80 : 1.0 for chromosomes 19 and 11, respectively (Kong *et al*., 2002), with typically higher values on longer chromosomes.

Relative sex-specific distances are not constant throughout the genome. Figure 1 shows the relative map difference as a function of location for chromosomes 4, 10, 13, and 16. The relative map difference (Matise *et al*., 1993) is a transformation

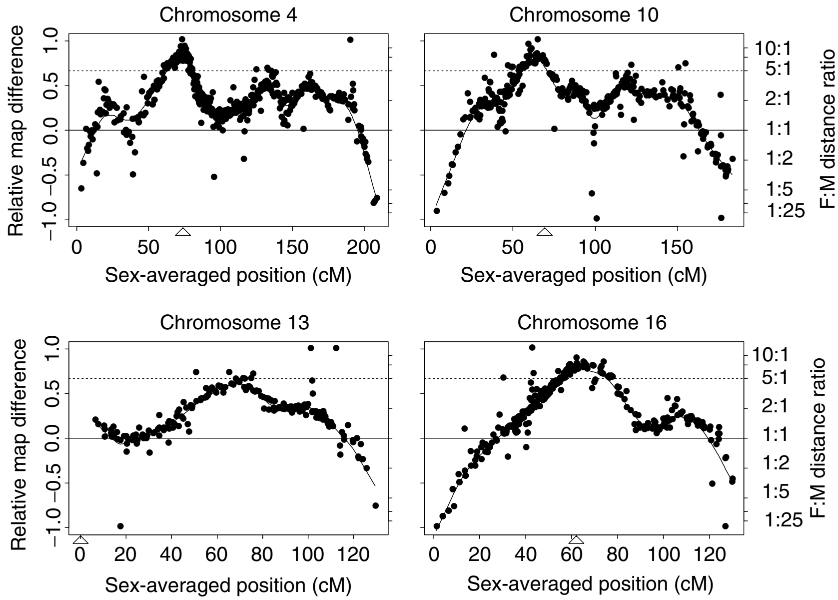


Figure 1 For chromosomes 4, 10, 13, and 16, the relative difference (q) between female and male map distances for all possibly overlapping marker intervals with lengths between 0 and 25 cM. The relative difference is plotted versus the sex-averaged midpoint of the marker interval. The deCODE framework markers are used for interval length calculation. These consist of 66, 55, 34, and 41 markers on chromosomes 4, 10, 13, and 16 respectively. The solid line is a smoothing of the points and the triangle indicates the approximate location of the centromere

of the map distance ratio and is defined by

$$q = \frac{d_f - d_m}{d_f + d_m} = \frac{d_f/d_m - 1}{d_f/d_m + 1}$$

where d_f and d_m are the female and male distances respectively. This measure is used because it always takes values in the range $[-1, 1]$, whereas the ratio, d_f/d_m can take values in the range $[0, \infty)$. Positive values of q indicate female-to-male ratios greater than 1 and negative values indicate female-to-male ratios less than 1. In Figure 1, the relative differences are plotted versus the sex-averaged position for the center of the interval, for all possible intervals with length between 0 and 25 cM, in order to reveal the pattern of distance ratio along the chromosome. A general trend toward greater distances in females than in males is evident. The relative map difference varies greatly along the chromosome and patterns of sex differences vary between chromosomes. Typically, the telomeric regions tend to exhibit more recombination in males than in females. The regions surrounding the centromeres often exhibit the highest female recombination rates, relative to the male rates. Broman *et al*. (1998) speculate that peaks in relative female recombination rate, which do not coincide with the presence of a centromere, may be due to the presence of suppressed latent centromeres. The patterns of

sex-specific recombination may explain the trend toward higher chromosome-wide female-to-male distance ratios on longer chromosomes, because the telomeric regions of increased male recombination comprise relatively smaller portions of the chromosome. These patterns also suggest that true genome-wide female-to-male ratios may be overestimated because of the lack of markers, and thus map distance estimates, near the telomeres where male recombination is relatively more frequent.

Map distance ratios in smaller regions can be quite extreme. For example, the deCODE genetic map shows marker intervals with sex-averaged map distances up to 3.98 cM (female distance 7.95 cM) for which the male distance is estimated to be zero. Similarly, there are regions with sex-averaged distances up to 3.55 cM (male distance 7.09 cM) for which the female distance is estimated to be zero. If we restrict ourselves to regions greater than 5 cM, there are still several instances of female-to-male ratios greater than 25 : 1 and less than 1 : 25, and most chromosomes exhibit at least one region with ratios more extreme than 10 : 1 or 1 : 10. On chromosome 20, this region is almost 25-cM long. Many chromosomes exhibit large segments with ratio greater than 5 : 1, typically surrounding the chromosome's centromere. These segments, which are identified by q values greater than 2/3, can be observed in Figure 1 and can be over 40-cM long.

3. Consequences for linkage analysis

Despite the differences in sex-specific maps, most linkage analyses assume sex-averaged maps. Daw *et al*. (2000) showed that for backcross data (i.e., fully informative meioses at all loci, including the trait), falsely assuming a sex-averaged map caused a negative bias in LOD score when linkage truly exists. This bias is typically quite small. For example, for a trait located midway between two markers, spaced 10 cM apart, falsely assuming a sex-averaged map causes an LOD score bias of −0.001 when the true female-to-male distance ratio is 5 : 1. This is quite small relative to the expected LOD (ELOD) under the true map, which is 0.265. The amount of bias increases as the true sex-ratios become more extreme and as the intermarker distances increase.

For an unlinked trait, the bias is positive. In this case, the bias is much larger both in absolute value and relative to the ELOD. For example, for a 10-cM interval with sex ratio 5 : 1, the ELOD under the correct map at a locus midway between the two markers is −1.028 and the bias that results from assuming a sex-averaged map is 0.118. In the unlinked case, the bias increases as the true sex-ratios become more extreme but decreases with increasing marker interval. The authors concluded that assuming a sex-averaged map causes only a slight reduction in power. However, it causes a modest increase in false-positive rate. These authors assumed a special case (i.e., fully informative meioses at all loci, including the trait), which is typically not the type of data available for humans.

In practice, meioses at marker loci are often not fully informative and meioses at the trait locus are rarely fully informative. In this case, the genetic map is especially important because it indicates the dependence between meioses at different loci. When meioses at a particular locus are not completely informative, multipoint linkage analyses incorporate information from other linked loci to increase the

Table 1 Distribution of parametric LODs[a], assuming either the true map or the sex-averaged map when the true map is a sex-specific map with F : M ratio 5 : 1

		Linked			Unlinked		
Pedigree	Map[b]	ELOD (SD)	$P(>3)$	$P(<1)$	ELOD (SD)	$P(<-2)$	$P(>1)$
Sibs	SS	2.79 (1.49)	0.44	0.12	−3.01 (1.68)	0.73	8.4×10^{-3}
	SA	2.81 (1.50)	0.45	0.11	−3.03 (1.69)	0.73	8.5×10^{-3}
Hsib (f)[c]	SS	1.23 (0.98)	0.04	0.41	−1.35 (1.31)	0.28	0.02
	SA	1.09 (1.16)	0.05	0.47	−1.95 (1.33)	0.48	0.01
Hsib (m)[d]	SS	2.19 (1.28)	0.26	0.18	−2.51 (1.56)	0.63	0.01
	SA	2.25 (1.10)	0.25	0.13	−1.77 (1.34)	0.43	0.01

[a]LODs are computed for a fully dominant trait, with disease allele frequency 0.1. Reported values are for 100 affected relative pairs in the absence of parental data. Probabilities are computed using a normal approximation.
[b]Assumed map is either sex-specific (SS) or sex-averaged (SA).
[c]Half sibs whose common parent is the female parent.
[d]Half sibs whose common parent is the male parent.

Table 2 Allele sharing Z_{LR} and NPL assuming either a sex-specific or sex-averaged map for data simulated under a sex-specific map with F : M ratio 5 : 1[a]

		Linked		Unlinked	
Pedigree	Map[b]	Z_{LR}[c] (SD)	NPL (SD)	Z_{LR}[c]	NPL (SD)
Sibs	SS	3.51 (0.97)	2.56 (0.70)	0.00 (1.00)	0.00 (0.73)
	SA	3.56 (0.97)	2.57 (0.71)	0.01 (1.00)	0.01 (0.72)
Hsib (f)[d]	SS	2.40 (0.96)	1.05 (0.41)	0.00 (1.00)	0.00 (0.46)
	SA	2.31 (0.96)	1.16 (0.47)	−0.08 (1.00)	−0.05 (0.53)
Hsib (m)[e]	SS	3.20 (0.93)	1.81 (0.50)	0.01 (1.00)	0.00 (0.60)
	SA	3.28 (0.92)	1.65 (0.44)	0.10 (1.00)	0.05 (0.53)

[a]Estimated from 10^5 sets of 100 pedigrees in the absence of parental data.
[b]Assumed map is either sex-specific (SS) or sex-averaged (SA).
[c]Z_{LR} calculated using the exponential model (Kong and Cox, 1997).
[d]Half sibs whose common parent is the female parent.
[e]Half sibs whose common parent is the male parent.

information. Misspecifying the genetic map causes the dependence between the markers to be misspecified and, thus, affects the imputations at each locus. As a result, assuming a sex-averaged map can have different consequences than in the backcross case, where the map is used only to specify the expected number of recombinations against which the observed number is compared.

Tables 1 and 2 demonstrate the effects of assuming a sex-averaged map in cases in which meioses are not fully informative. Three different affected relative pairs are used: full sibs, half sibs where the common parent is a female [hsib (f)], and half sibs where the common parent is a male [hsib (m)]. In each case, the parental data was assumed to be missing. Marker data was assumed at two loci, each with four equifrequent alleles. The sex-averaged intermarker spacing was 10 cM,

and the female-to-male map distance ratio was 5 : 1. A fully dominant trait locus with allele frequency 0.1 was assumed to be positioned either midway between the markers or unlinked. Table 1 shows exact ELODs and standard deviations of LODs for parametric computations under the true map and sex-averaged map. These computations show that assuming sex-averaged maps can cause positive or negative bias in the LOD score. When the pedigree contains equal numbers of male and female meioses, as is the case for full sibs, falsely assuming a sex-averaged map can cause a small inflation in the ELOD, when the trait is linked, and a small decrease in the ELOD when the trait is unlinked. In both the linked and unlinked cases, for a single sibpair, the median LOD score is slightly larger under the sex-specific map than under the sex-averaged map. The probability of a significant or suggestive LOD is slightly increased under the sex-averaged map, both when the trait is linked or unlinked. The probability of excluding linkage,with LOD scores less than −2, is also slightly increased under the sex-averaged map in both the linked and unlinked case. In the case of half sibs, the number of male and female meioses are different, so the effects can be more extreme. In fact, the relevant meioses in these pedigrees are either solely the male meioses or solely the female meioses, depending on which parent is shared by the affected individuals. As a result, assuming a sex-averaged map has the same effect as assuming a map that is strictly too long or too short. When the affected relative pairs are half sibs with a common mother, all meioses, which are informative for linkage, are female meioses and the sex-averaged map is too short. This causes a negative bias in the LOD score in both the linked and unlinked cases, which results in reduced power to detect suggestions of linkage when the trait is linked. When the affected pair share a common father, the sex-averaged map is too long. The resulting sex-averaged LOD score shows positive bias, but because the variance is reduced, the probability of a significant LOD score is actually slightly reduced.

Table 2 shows the effects of assuming a sex-averaged map on NPL scores (Kruglyak *et al*., 1996) and allele-sharing Z_{LR} scores (Kong and Cox, 1997). The Z_{LR} is the signed square root of the allele-sharing likelihood ratio, and thus is equal to approximately 2.15 times the signed square root of the allele-sharing LOD score. This statistic, like the NPL score, is asymptotically normally distributed. In the linked case, the expectation of the Z_{LR} behaves similar to the LOD in the parametric case, but the standard deviation is barely affected. The NPL score shows different patterns. When the affected individuals are half sibs with a common mother, the sex-averaged map causes an increase in NPL over the sex-specific map NPL, and when the half sibs share a common father, the NPL under the sex-averaged map causes a decreased NPL score. Note that difference in the sign of the bias between the NPL and the other test statistics is driven by the bias in the standard deviation. If the NPL statistic were normalized, so that standard deviation were truly 1 under the null hypothesis, as assumed, the bias sign would match the bias sign of the other two statistics. When the trait is unlinked, both the Z_{LR} and NPL exhibit small amounts of bias under sex-averaged maps. However, the standard deviation of the NPL is affected by the map misspecification, while the standard deviation of the Z_{LR} is not. In the case of half sibs sharing a mother, the standard deviation is higher under the sex-averaged map. Because NPL tests falsely assume a null distribution with variance 1, in this case, the use of the

sex-averaged map results in a test that is less conservative than under the true map. When the shared parent is a father, the standard deviation is smaller under the sex-averaged map, which causes an already conservative test to be more conservative.

The amount of bias caused by map misspecification is affected by several factors. The bias increases with marker spacing. For example, using the hsib (m) pedigrees, the bias in Z_{LR} for a pair of markers spaced 2 cM apart is 0.03 for an unlinked trait. For a marker at a 10-cM spacing, the bias is 0.09. The bias also increases with marker number and marker density. Six markers with 2-cM intermarker spacing, thus spanning 10 cM, yield a bias of 0.27 when the trait is unlinked. Six markers at a 10-cM intermarker spacing yield a bias of 0.41 when the trait is unlinked. The number of alleles, and thus the marker information, also affects the amount of bias. The more informative the markers, the greater the bias. For two markers with eight equifrequent alleles, the bias is 0.23, when the trait is unlinked and the markers are spaced 10 cM apart. Finally, the sample size affects the amount of bias. Parametric LOD scores are additive across families, so the amount of bias increases linearly with the number of families. Doubling the number of families doubles the bias. Both the NPL and the Z_{LR} are constructed to have fixed variance under the null hypothesis. The NPL is scaled by the complete-data standard deviation and the Z_{LR} is constructed to have an asymptotic variance of 1. As a result, the bias of both of these statistics scales as the square root of the number of families. Thus, quadrupling the sample size doubles the bias.

Typically, the difference in number of male and female meioses will be more moderate than in the half sib case. In most practical cases, the use of sex-averaged maps will not significantly affect the results of linkage analyses. Still, the danger of bias in multipoint linkage analysis exists in certain situations. This bias becomes more extreme as the number of linked markers and marker informativeness increase. Analyses with sex-specific maps should be done to verify signals that occur in areas of extreme sex-distance ratios or when the signal is driven by pedigrees with unequal numbers of male and female meioses. Loss of power can also be an issue, but strategies to prevent this situation are more difficult. Another instance, when the use of sex-specific maps is particularly important, is when modeling parent-of-origin effects, such as in imprinting models (Hanson *et al*., 2001; Shete and Amos, 2002; Karason *et al*., 2003). Mukhopadhyay and Weeks (2003) found that using sex-specific maps could cause significant change in variance component LOD scores for an imprinting model. In most instances, the use of a sex-specific map increased the LOD score, although they reported one instance of a decrease in LOD score. At several loci, the LOD score assuming the sex-averaged map was insignificant, but was significant under the sex-specific map. Linkage analyses assuming sex-specific maps can be performed in many linkage packages including ALLEGRO (Gudbjartsson *et al*., 2000), FASTLINK (Cottingham *et al*., 1993), LINKAGE (O'Connell and Weeks, 1995), Loki (Heath, 1997), S.A.G.E. (S.A.G.E., 2002), and VITESSE (O'Connell and Weeks, 1995). These packages span a variety of different types of linkage analyses. As of the current date, sex-specific maps are not implemented in GENEHUNTER (Kruglyak *et al*., 1996).

Further reading

Lathrop GM, Lalouel JM, Julier C and Ott J (1984) Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3443–3446.

References

Barendse W, Armitage SM, Kossarek LM, Shalom A and Kurkpatrick BW (1994) A genetic linkage map of the bovine genome. *Nature Genetics*, **6**, 227–235.

Broman KW, Murray JC, Sheffield VC, White RL and Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics*, **63**, 861–869.

Cottingham RW, Idury RM and Scháffer AA (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics*, **53**, 252–263.

Daw EW, Thompson EA and Wijsman EM (2000) Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology*, **19**, 366–380.

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ, *et al*. (1996) A comprehensive genetic map of the mouse genome. *Nature*, **380**, 149–152.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, *et al*. (1987) A genetic linkage map of the human genome. *Cell*, **51**, 319–337.

Gudbjartsson D, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Haldane JBS (1922) Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, **12**, 101–109.

Hanson RL, Kobes S, Lindsay RS and Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *American Journal of Human Genetics*, **68**, 951–962.

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

Huxley JS (1928) Sexual difference of linkage Grammarus chereuxi. *Journal of Genetics*, **20**, 145–156.

Karason A, Gudjonsson JE, Upmanyu R, Antonsdottir AA, Hauksson VB, Runasdottir EH, Jonsson HH, Gudbjartsson DF, Frigge ML, Kong A, *et al*. (2003) A susceptibility gene for psoriatic arthritis maps to chromosome 16q. *American Journal of Human Genetics*, **72**(1), 125–131. Epub 2002 Dec 9.

Kong A and Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, **61**, 1179–1188.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, **31**, 241–247.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**(6), 1347–1363.

Maddox JF, Davies KP, Crawford AM, Hulme DJ, Vaiman D, Cribiu EP, Freking BA, Beh KJ, Cockett NE, Kang N, *et al*. (2001) An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Research*, **11**(7), 1275–1289.

Matise TC, Blaschak JE, Kompanek AJ, Weeks DE and Chakravarti A (1993) Patterns of sex-difference and interference in the human genome. *American Journal of Human Genetics* **53**(Suppl 262).

Mikawa S, Akita T, Hisamatsu N, Inage Y, Ito Y, Kobayashi E, Kusumoto H, Matsumoto T, Mikami H, Minezawa M, *et al*. (1999) A linkage map of 243 DNA markers in an intercross of Góttingen miniature and Meishan pigs. *Animal Genetics*, **30**(6), 407–417.

Morgan TH (1914) No crossing over the male drosophila of genes in the second and third pair of chromosomes. *Biological Bulletin*, **26**, 195–204.

Mukhopadhyay N and Weeks DE (2003) Linkage analysis of adult height with parent-of-origin effects in the framingham heart study. *BMC Genetics*, **4**(Suppl. 1), S76.

Neff MW, Broman KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander EA and Rine J (1999) A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics*, **151**(2), 803–820.

O'Connell JR and Weeks DE (1995) The vitesse algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

Renwick J and Schulze J (1965) Male and female recombination fraction for the nail-patella: ABO linkage in man. *Annals of Human Genetics*, **28**, 379–392.

S.A.G.E. (2002) *Statistical analysis for genetic epidemiology*, Computer program package available from Statistical Solutions Ltd: Cork.

Shete S and Amos CI (2002) Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *American Journal of Human Genetics*, **70**, 751–757.

Polymorphic inversions, deletions, and duplications in gene mapping

Susan L. Christian
University of Chicago, Chicago, IL, USA

One of the under recognized complexities of gene mapping is the presence of chromosomal anomalies (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1) within the human genome. Inversions, deletions, and duplications have the effect of disrupting linkage or association analyses by complicating genetic mapping of a region (inversions) or overlooking a disease gene caused by either a deletion or duplication. An understanding of the underlying architecture of the human genome that causes these anomalies is needed to accurately assess the data acquired through linkage or association analyses.

One unexpected structural feature identified following sequencing of the human genome was the presence of segmental duplications (SDs) (*see* Article 26, **Segmental duplications and the human genome**, Volume 3) that mediate chromosomal rearrangements. Segmental duplications are defined as stretches of genomic DNA that are >1 kb in length with >90% sequence identity that map to multiple regions (Eichler *et al.*, 2004). These regions are normal stretches of genomic DNA containing transcriptionally active genes that have undergone duplication as part of evolution of the human genome (Bailey and Eichler, 2003). SDs comprise ~5% of the genome with a frequency of 1–14% among the 24 chromosomes (Zhang *et al.*, 2005). The distribution of SDs varies between chromosomes with preferential location in pericentromeric and subtelomeric locations (Eichler *et al.*, 2004; Zhang *et al.*, 2005).

The presence of these SDs has posed a major problem in mapping and sequencing of the human genome. The latest release of the human sequence in October 2004 stills shows 341 gaps, many of which either contain or are flanked by SDs (IHGSC, 2004). Additionally, the larger SDs (>100 kb) and those with >99% sequence identity complicated the sequencing of the human genome such that regions containing SDs were merged together or otherwise misassembled (Cheung *et al.*, 2003; Eichler *et al.*, 2004). This problem was exacerbated by the whole-genome shotgun sequencing approach used by Celera that relied on sequencing of small stretches of DNA followed by assembly (She *et al.*, 2004).

The presence of SDs within the human genome has also complicated the development of single nucleotide polymorphisms (SNPs) (*see* Article 50, **Gene**

mapping and the transition from STRPs to SNPs, Volume 1). The goal in the development of SNPs was to find sequence variation at a single nucleotide position within the genome. With the presence of SDs, a putative SNP may actually represent sequence variation between two different copies of an SD present at separate loci within the genome. An examination of SNPs validated for a single locus showed an underrepresentation in SDs where 3.75% of validated SNPs versus 13.1% of nonvalidated SNPs were identified in the 4.5% of the genome composed of SDs (Fredman *et al*., 2004). Therefore, validation of SNPs is essential to prevent usage of variants located in SDs that represent paralogous copies from other chromosomes rather than a true SNP.

An additional problem with the presence of SDs is the propensity of these regions to misalign during meiosis, producing chromosomal inversions, deletions, and duplications (*see* Article 13, **Meiosis and meiotic errors**, Volume 1 and Article 17, **Microdeletions**, Volume 1). When SDs arranged in an inverted orientation align with each other, the single copy region between them is inverted. Inversions do not change the copy number of genomic DNA but may disrupt genes at the breakpoints. Several disorders have been identified on Xq28 that are thought to arise during male meiosis where the single X folds back on itself and rearranges at the sites of SDs causing inversions. Additionally, carriers of inversions may produce offspring with unbalanced karyotypes associated with severe abnormalities. One example is a submicroscopic 8p inversion that is present in 26% of a population of European descent (Giglio *et al*., 2001). Females with this inversion had children with three different types of abnormalities of 8p including an interstitial deletion associated with a heart defect, a derivative marker 8p associated with a distinct phenotype plus an inverted duplication 8p chromosome with minor anomalies (Giglio *et al*., 2001).

The alignment of a chromosomal region containing an inversion during meiosis forces one homolog to form a loop to pair with the other homolog. Any rearrangement within this loop will produce unbalanced offspring. Only the unchanged normal and inverted homologs will produce a normal chromosomal complement. Thus, individuals heterozygous for an inversion would appear to show reduced recombination because of the dosage imbalance produced in the gametes.

As marker maps used in linkage mapping become denser, polymorphic inversions such as the 8p inversion described above may add to the challenges of linkage mapping. In the case of polymorphic inversions, different individuals actually have different genetic maps at the site of the inversion. Conducting linkage analysis assuming the marker order of the most common of the inversion orientations can lead to a falsely inflated estimate of local recombination frequencies and/or increases in the genotyping error rates. This is because a single recombination that occurs between informative markers within the inverted segment in an individual homozygous for the rarer orientation will be interpreted as a set of three recombination events in genetic intervals that are very close together. Of course, the apparent increase in recombination frequency and genotyping error would be even worse if the analysis is conducted assuming a marker order consistent with the less common orientation of the inversion. The basic problem is that either orientation will create apparent recombination events (often interpreted as genotyping error) because there are actually multiple maps within the population.

Reciprocal deletions and duplications occur when SDs arranged in a direct orientation undergo homologous recombination. In this case, one copy of the region between the SDs is translocated from one homolog to the other, producing a loss on one homolog and a gain on the other. Many human diseases are now recognized that are caused by these rearrangements between SDs including Charcot-Marie-Tooth disease type 1A and hereditary neuropathy with liability to pressure palsies due to a reciprocal duplication/deletion rearrangement on 17p11.2 (Shaw and Lupski, 2004).

Genotyping data of a region containing a deletion will appear as a contiguous stretch of homozygous markers including some with non-Mendelian inheritance from a single parent. Also, a deletion cannot be distinguished from a case of uniparental isodisomy (*see* Article 19, **Uniparental disomy**, Volume 1) where there are two identical copies of a chromosomal region present from a single parent. Duplications are sometimes more difficult to identify. In linkage analysis involving microsatellite markers, a duplication may appear as the presence of three alleles with equal intensity or two alleles with double dosage of one allele. Most software programs designed to assign microsatellite genotypes do not anticipate this situation and read only the first two alleles. In cases with three alleles, the data will be read as the presence of two maternal alleles in 2/3 of the instances rather than reading all three alleles. Therefore, microsatellite data across a duplication would show the presence of a mixture of uninformative single alleles plus some with biparental inheritance and others with apparent uniparental heterodisomy showing two different alleles from a single parent. With the dense SNP maps being used in linkage analysis, such regions may be characterized by clusters of markers with higher than average numbers of Mendelian incompatibilities.

Whole-genome association studies using SNP arrays also have the potential to identify genomic imbalances. In one study, cell lines with 1–5 X chromosomes were analyzed using a 10K SNP array to detect DNA copy number (Zhao *et al*., 2004). Comparisons were made with a pool of normal DNA and the log 2 ratios computed. In all cases, the inferred copy number matched the known X chromosome copy number. Additional studies involving cancer cell lines also detected both deletions and duplications that were confirmed by quantitative PCR.

With SDs comprising ~5% of the genome, the potential for chromosomal rearrangements between them is vast. Many benign polymorphisms (*see* Article 10, **Measuring variation in natural populations: a primer**, Volume 1) are likely to be present in addition to anomalies that will provide susceptibility to human disease. The identification of chromosomal imbalances that are part of the normal variation within humans is currently under study. Sebat *et al*. (2004) surveyed 50 Mb in 20 normal individuals using representational oligonucleotide microarray analysis (ROMA). In this method, a reduced complexity library of Bgl II DNA was developed to identify 85 000 probes in a size range of 200–1200 bases that were spaced 1/35 kb across the genome. Seventy-six unique copy number polymorphisms were identified with an average of 11 per individual.

In another study using array comparative genomic hybridization (*see* Article 23, **Comparative genomic hybridization**, Volume 1), Iafrate *et al*. (2004) surveyed 300 Mb in 55 unrelated individuals. Thirty-nine were unrelated apparently healthy controls and 16 had known abnormalities. Two hundred and twenty-six copy number polymorphisms were identified. One hundred and two were found in two or

more individuals and 24 in >10%. On average, 12.4 copy number polymorphisms were observed per individual.

The question then becomes whether the genomic imbalances detected represent benign polymorphisms or susceptibility to human disease. The answer will most likely be a mixture of both. For inheritance of common diseases such as diabetes or psychiatric disorders, the presence of multiple susceptibility loci that provide small effects (*see* Article 58, **Concept of complex trait genetics**, Volume 2) has long been a hypothesis for the genetic mechanism of disease. The primary effort has focused on the identification of variation at the nucleotide level in specific genes as the bases for these susceptibilities. However, the presence of these larger deletion/duplication variants mediated by SDs may also represent another class of small effect loci that affect susceptibility to common disease.

References

Bailey JA and Eichler EE (2003) Genome-wide detection and analysis of recent segmental duplications within mammalian organisms. *Cold Spring Harbor Symposia on Quantitative Biology*, **68**, 115–124.

Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui L-C and Scherer SW (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology*, **4**, R25.

Eichler EE, Clark RA and She X (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Reviews. Genetics*, **5**, 345–354.

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT and Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, **36**, 861–866.

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, *et al*. (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *American Journal of Human Genetics*, **68**, 874–883.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **36**, 949–951.

International Human Genome Sequencing Consortium (IHGSC) (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shaw CJ and Lupski JR (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human Molecular Genetics*, **13**(1), R57–R64.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL and Eichler EE (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.

Zhang L, Lu HH, Chung WY, Yang J and Li WH (2005) Patterns of segmental duplication in the human genome. *Molecular Biology and Evolution*, **22**, 135–141.

Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen T-H, Girard L, Minna J, Christiani D, Leo C, *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*, **64**, 3060–3071.

Short Specialist Review

Impact of linkage disequilibrium on multipoint linkage analysis

Christopher I. Amos

U.T.M.D. Anderson Cancer Center, Houston, TX, US

Qiqing Huang

Johnson & Johson Pharmaceutical Research and Development, Raritan, NJ, US

In multipoint linkage analysis, when there is unresolved phase information for multiple heterozygous individuals, equal probabilities are usually assigned to all possible phases that are compatible with the data (O'Connell and Weeks, 1995; Kruglyak *et al*., 1996). Currently, all implementations of the Lander-Green algorithm, which are commonly used for most multipoint linkage applications, such as Allegro (Gudbjartsson *et al*., 2000), and Genehunter (Kruglyak *et al*., 1996), assume linkage equilibrium among the markers. This is a reasonable assumption for sparsely spaced markers (e.g. ∼10 cM microsatellite markers) since the markers would not be expected to show LD at this distance. However, this assumption can be problematic when there is strong LD among tightly linked markers, because LD causes the observed haplotype frequencies to deviate from the expected haplotype frequencies, and results in inaccurate phase probabilities.

When parental genotype data are available, linkage calculations use the observed genotypes rather than specified genotype frequencies. In this situation, linkage analysis is robust to misspecification of phase probabilities (Ott, 1999). However, when some of the parental data are missing, for example, for late-onset diseases (LOD), the linkage equilibrium assumption can lead to a severe bias for tightly linked markers (Huang *et al*., 2004; Boyles *et al*., 2005).

Schaid *et al*. (2002) showed with an example that assuming linkage equilibrium in pedigree-based haplotype inference produced highly inaccurate haplotype frequencies from population based EM estimation. They found that haplotype frequencies inferred by pedigree analysis programs are close to the expected frequencies under an assumption of equal phase probabilities. Huang *et al*. (2004) studied the impact of linkage equilibrium assumption on multipoint linkage statistics when there is strong LD between markers. They showed both analytically and by simulation that assuming linkage equilibrium between tightly linked markers could lead to false- positive linkage signals when some of the parental data are

missing. Boyles *et al*. (2005) further extended the results of Huang *et al*. (2004) by simulation and found that the squared correlation between markers is a better indication of potential for spurious inflation of LOD scores than the standardized disequilibrium coefficient. In addition, they found that markers that show negative LD do not lead to inflation of the LOD score (in part because the squared correlation is low between such markers).

LD between tightly linked markers causes certain haplotypes to be more frequent than expected underlinkage equilibrium. The accrual of those haplotypes in families selected through affected relatives may be misinterpreted as oversharing of multipoint IBD, if the phases have been incorrectly specified (for example, because the linkage program assumes equal phase probabilities). The biased interpretation of IBD sharing among relatives will generate false- positive evidence of linkage for affected sib-pair analysis using model-free linkage methods. Similarly, for parametric linkage analysis, if there are only affected sib-pairs and parental data are missing, the accrual of certain haplotypes within a subset of families may also appear to indicate evidence of linkage to the disease. Linkage analysis under a heterogeneity model will lead to an excess of false-positive results (Figure 1 of Huang *et al*., 2004).

Single point linkage analysis (parametric and model-free) does not suffer from this bias, although it does require accurate marker allele frequency estimates at each locus if the families have been selected through affected sibling pairs (Williamson and Amos, 1995). Bias can also be eliminated with parental data and reduced with additional unaffected sibs.

Recently, there is increasing interest in using high-density single-nucleotide-polymorphism (SNP) markers for linkage analysis instead of traditional panels of microsatellite markers (Middleton *et al*., 2004; John *et al*., 2004; Schaid *et al*., 2004). Evans and Cardon (2004) showed that a SNP map of 1 SNP/cM is sufficient to extract most of the inheritance information (>95% in most cases) when parental genotypes are available and much higher density SNP markers are needed when some of the parental data are missing.

Although high-density SNP markers can provide more information than sparse microsatellites, caution should be exercised when evaluating evidence of linkage with dense markers since strong LD may exist between some of the markers. The apparent evidence of linkage may reflect an excess of false-positive linkage results due to LD between the tightly linked markers. To reduce false-positive evidence for linkage, Webb *et al*., (2005) have introduced software that evaluates the LD among sets of markers and then drops any markers that show LD among a user-defined threshold. But this may decrease the information content and reduce the power. In addition, with very dense panels of markers, we have found that just applying a pairwise-test for LD among markers may not completely protect against false-positive results. Bacanu (2005) suggested dividing dense mapping panels into sets of disjoint but overlapping markers. Results of linkage on these disjoint panels can then be averaged and the estimate of variance among panels provides a statistic that can be used to construct hypothesis tests.

Alternatively, a version of Merlin has been developed (Abecasis *et al*., 2002; Abecasis and Wigginton, 2005) that uses haplotypes rather than alleles at clusters of tightly linked markers. The definitions of the clusters can be either user-defined

or automatically generated, by using an R-squared criterion for LD among the marker loci. Within a cluster, haplotype frequencies are estimated by an application of the E-M algorithm and then the linkage analysis uses haplotypes, weighted by their probabilities, rather than genotypes. Application of this approach to simulated data showed a mild loss of information by clustering markers into haplotypes but protection from false-positive linkage findings. Many existing parametric linkage procedures that use the Elston-Stewart algorithm such as Linkage (Lathrop and Lalouel, 1988) can incorporate haplotype probabilities but implementation is currently awkward when considering more than a few loci jointly.

References

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Abecasis GR and Wigginton JS (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *American Journal of Human Genetics*, **77**, 754–767.

Bacanu SA (2005) Multipoint linkage analysis for a very dense set of markers. *Genetic Epidemiology*, **29**, 195–203.

Boyles AL, Scott WK, Martin ER, Schmidt S, Li Y-J, Ashley-Koch A, Bass MP, Schmidt M, Pericak-Vance MA, Speer MC, *et al*. (2005) Linkage disequilibrium inflates Type I error rates in multipoint linkage analysis when parental genotypes are missing. *Human Heredity*, **59**, 220–227.

Evans DM and Cardon LR (2004) Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *American Journal of Human Genetics*, **75**, 687–692.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Huang Q, Shete S and Amos CI (2004) Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *American Journal of Human Genetics*, **75**, 1106–1112.

John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, *et al*. (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *American Journal of Human Genetics*, **75**, 54–64.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Lathrop GM and Lalouel JM (1988) Efficient computations in multilocus linkage analysis. *American Journal of Human Genetics*, **42**, 498–505.

Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H, *et al*. (2004) Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *American Journal of Human Genetics*, **74**, 886–897.

O'Connell JR and Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

Ott J (1999) *Analysis of Human Genetic Linkage*, Third Edition, Johns Hopkins University Press: Baltimore, MD, p. 251.

Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hilker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML, *et al*. (2004) Comparison of microsatellites versus

single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility Loci. *American Journal of Human Genetics*, **75**, 948–965.

Schaid DJ, McDonnell SK, Wang L, Cunningham JM and Thibodeau SN (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *American Journal of Human Genetics*, **71**, 992–995.

Webb EL, Sellick GS and Houlston RS (2005) SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. *Bioinformatics*, **21**, 3060–3061.

Williamson JA and Amos CI (1995) Guess LOD approach: sufficient conditions for robustness. *Genetic Epidemiology*, **12**, 163–176.

Basic Techniques and Approaches

Computation of LOD scores

Anthony L. Hinrichs and Brian K. Suarez
Washington University in St. Louis, St. Louis, MO, USA

1. Introduction

In this section, we will demonstrate detailed examples of linkage calculations for various types of data. Our initial examples will be of parametric analysis for Mendelian traits. This will be followed by an example using Affected Sibling Pair (ASP) methods. Finally, we will compute a nonparametric linkage (NPL) score using Whittemore–Halpern statistics for the same dataset. We will assume that the reader is already familiar with recombination, genetic map distances, identity by descent measurement, and likelihood (*see* Article 49, **Gene mapping, imprinting, and epigenetics**, Volume 1, Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1, and Article 51, **Choices in gene mapping: populations and family structures**, Volume 1).

2. Parametric linkage – dominant disorder

We will begin with an example adapted from Suarez and Cox (1985) (Figure 1). In this case, the sample phenotype is fully penetrant Mendelian dominant. Since individual I-1 is affected, we know that individual I-1 carries at least one copy of the disease allele. Since there is at least one unaffected offspring, we know that individual I-1 does not carry two copies of the disease allele. In the case of no unaffected offspring, the following calculations would be weighted by the probability that the affected founder carries two copies of the disease allele.

As presented in Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1, a parametric LOD score (Barnard, 1949) for a recombination rate of θ is computed as

$$\log_{10}\left(\frac{L(\theta)}{L(0.5)}\right)$$

Because phase between the marker alleles and the disease allele is unknown in individual I-1, we must condition on the phase using a uniform prior. There are 19 descendants of individual I-1 who had a chance of receiving the disease allele (since individual II-5 is unaffected, her offspring carry no information). Since all pairs of

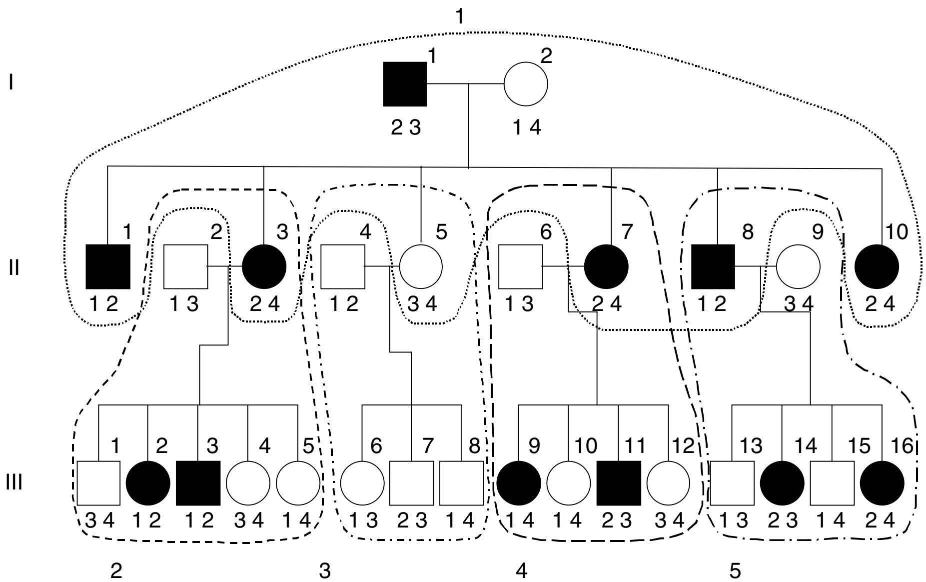


Figure 1 Fully penetrant Mendelian dominant phenotype. Pedigree members are numbered by generation and sequentially within a generation. The numbers below the pedigree symbols are the individuals' genotype. The dashed and dotted lines separate the five nuclear families that comprise the larger pedigree

parents have 4 unique markers (so all meioses can be directly observed), we directly observe 19 recombination/nonrecombination events for each of the two possible phases. If the disease allele is linked to marker 2 in individual I-1, then we observe 1 recombination (in individual III-9) and 18 nonrecombinations. If the disease allele is linked to marker 1 in individual I-1, then there are 18 recombinations and 1 nonrecombination. Then for any recombination fraction θ, we have

$$L(\theta) = \frac{1}{2}\left(\theta(1-\theta)^{18} + \theta^{18}(1-\theta)\right) \tag{1}$$

The LOD score for disease locus at the marker locus is undefined; we observe a recombination so the likelihood is 0. However, we can compute the maximum likelihood estimator (MLE) for θ by solving

$$\frac{\mathrm{d}L}{\mathrm{d}\theta} = 0 \tag{2}$$

Solving numerically, we find the maximum at $\theta = 0.0526$. For this recombination fraction, the LOD score is $\log_{10}(L(.0526)/L(0.5)) = 3.717$.

We will next demonstrate the advantage of extended pedigrees. If we divide the pedigree in Figure 1 into nuclear pedigrees, there are four pedigrees with affected individuals with recombination/nonrecombination counts of 6/0, 5/0, 3/1, and 4/0 or 0/6, 0/5, 1/3, and 0/4. For a given recombination fraction, the overall likelihood

of all of the pedigrees is

$$L^*(\theta) = \left(\tfrac{1}{2}\right)^4 \left((1-\theta)^6 + \theta^6\right)\left((1-\theta)^5 + \theta^5\right) \times \left(\theta(1-\theta)^3 + \theta^3(1-\theta)\right)\left((1-\theta)^4 + \theta^4\right) \quad (3)$$

This is simply the product of the likelihoods for each pedigree considered separately. The leading coefficient is $(1/2)^4$ since for each of the four pedigrees we must assume a uniform prior on phase. If we take the derivative as before, we find that the maximum occurs at $\theta = 0.0530$. For this recombination fraction, the LOD score is $\log_{10}(L^*(.0530)/L^*(0.5)) = 2.815$. We note that although $L(0.5) = L^*(0.5)$, the numerator of L^* has decreased owing to the extra ambiguity of phase. For small θ, the difference is the equivalent of three phase-known meioses:

$$L(\theta) - L^*(\theta) \approx 3\log_{10} 2 \approx 0.9$$

This cost of $\log_{10}2$ is paid for each ambiguous phase. By using the entire pedigree at once, we only have one conditional phase. In this case, the whole is greater than the sum of its parts.

3. Parametric linkage – recessive disorder

Our next example is a fully penetrant Mendelian recessive trait (Figure 2). Since the parents are unaffected but have affected children, we know that each carries one copy of the disease allele. As before, we look at recombinations but in this case we are concerned with the phase separately for two disease allele–bearing haplotypes (one in each parent). Furthermore, we cannot be sure whether the unaffected

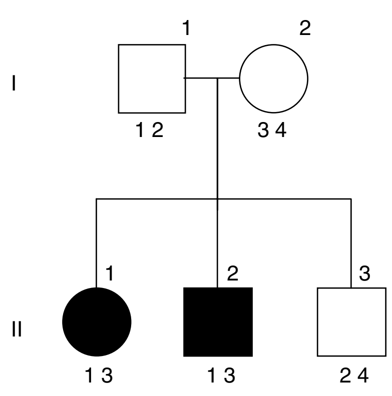


Figure 2 Fully penetrant Mendelian recessive phenotype. Pedigree members are numbered by generation and sequentially within a generation. The numbers below the pedigree symbols are the individuals' genotype

offspring (II-3) carries 0 or 1 copy of the disease allele. We, therefore, must put a uniform prior on the possibilities for recombinations for the unaffected offspring. Out of the four possibilities, one would give the unaffected offspring two disease alleles. Thus, we have 12 possibilities for the six recombination/nonrecombination events:

$$L(\theta) = \frac{1}{12}\left[\begin{array}{c}(1-\theta)^6 + 2\theta(1-\theta)^5 + 2\theta^2(1-\theta)^4 + 2\theta^3(1-\theta)^3 \\ +2\theta^4(1-\theta)^2 + 2\theta^5(1-\theta) + \theta^6\end{array}\right] \tag{4}$$

This function achieves a maximum at $\theta = 0$, with an LOD score of $\log_{10}(L(0)/L(0.5)) = 0.727$. Since parametric LOD scores are additive, we would require five or more pedigrees of this configuration to meet or exceed the standard for demonstrating linkage (a LOD of 3) (Morton, 1955).

4. Affected sibling pair linkage

We will next consider non-Mendelian disorders. Since there is substantially less power in a single pedigree, we will consider a set of nuclear pedigrees (two genotyped parents and two genotyped affected siblings) detailed in Table 1. We will use the Affected Sibling Pair (ASP) method of Risch (1990), as implemented by Kruglyak and Lander (1995) in MAPMAKER/SIBS. Let $\pi_j(i)$ denote the probability that sibling pair i shares j alleles identical-by-descent (IBD). Let (z_0, z_1, z_2) represent a specific hypothesis for average sibling pair sharing of IBD 0, IBD 1, and IBD 2; let (α_0, α_1, α_2) represent the random expectation under the null hypothesis of no linkage. Then, we compute the LOD score by summing over all sibling pairs:

$$\text{LOD} = \sum_i \log_{10}\left(\frac{z_0\pi_0(i) + z_1\pi_1(i) + z_2\pi_2(i)}{\alpha_0\pi_0(i) + \alpha_1\pi_1(i) + \alpha_2\pi_2(i)}\right) \tag{5}$$

For sibling pairs, the null hypothesis will be $(\alpha_0, \alpha_1, \alpha_2) = (1/4, 1/2, 1/4)$. In this example, IBD status is known with certainty at the locus in question for each affected sibling pair, so we can compute the MLEs for (z_0, z_1, z_2) through a simple proportion of pedigrees:

$$(z_0, z_1, z_2) = \left(\frac{18}{126}, \frac{60}{126}, \frac{48}{126}\right) = \left(\frac{1}{7}, \frac{10}{21}, \frac{8}{21}\right) \tag{6}$$

Table 1 IBD status at candidate locus for 126 quartets (2 parents, 2 affected offspring)

IBD status	0	1	2
Number of pedigrees	18	60	48

We do not require $z_1 = 0.5$; that is, we allow dominance deviation. These MLEs also lie within the "possible triangle" (Holmans, 1993): $z_1 \leq 0.5$ and $z_1 \geq 2z_0$. If the proportions resulted in MLEs outside of the possible triangle or dominance deviation was disallowed, then an EM algorithm could be used to find the parameters (Kruglyak and Lander, 1995). We then compute the LOD score:

$$\begin{aligned}
\text{LOD} &= \sum_i \log_{10}\left(\frac{z_0\pi_0(i) + z_1\pi_1(i) + z_2\pi_2(i)}{\alpha_0\pi_0(i) + \alpha_1\pi_1(i) + \alpha_2\pi_2(i)}\right) \\
&= \sum_i \log_{10}\left(\frac{1/7\pi_0(i) + 10/21\pi_1(i) + 8/21\pi_2(i)}{1/4\pi_0(i) + 1/2\pi_1(i) + 1/4\pi_2(i)}\right) \\
&= 18\log_{10}(4/7) + 60\log_{10}(20/21) + 48\log_{10}(32/21) \\
&= 3.135 \qquad (7)
\end{aligned}$$

5. Nonparametric linkage

Finally, we will consider Nonparametric Linkage (NPL) scores using Whittemore–Halpern statistics (Whittemore and Halpern, 1994; Kruglyak *et al*., 1996), implemented in GENEHUNTER. These statistics are based on a scoring function $S(v,\Phi)$, where Φ represents the phenotypic configuration of a particular pedigree and v is an inheritance vector – the inheritance pattern of segregation for a pedigree at a particular locus. Typically, the true inheritance pattern is unknown, so one computes the average over all possible inheritance vectors weighted by the probability of the vector:

$$\overline{S}(\Phi) = \sum_{w \in V} S(w, \Phi) P(v = w) \qquad (8)$$

There are several possible scoring functions. We will consider S_{pairs}, defined to be the number of pairs of alleles from distinct affected pedigree members that are IBD. We will consider only the case where the parents are unaffected so each pedigree will have one pair of affected individuals (namely, the sibling pair used in the previous example). Since segregation is unambiguous for the 126 nuclear families listed in Table 1, each affected sibling pair has a score of 0, 1, or 2. For each pedigree, mean and standard deviation of the scoring function (based on a uniform distribution of all possible inheritance vectors) are computed. For each pedigree in this example, the mean is 1 and the standard deviation is $1/\sqrt{2}$. The observed scores are then normalized on the basis of these values:

$$Z(i) = \frac{S(i) - \mu}{\sigma} = \begin{cases} -\sqrt{2}, & \text{if IBD} = 0 \\ 0, & \text{if IBD} = 1 \\ \sqrt{2}, & \text{if IBD} = 2 \end{cases} \qquad (9)$$

Finally, we sum the normalized score over all pedigrees, with a weight of $1/\sqrt{N}$, so that the final sum has mean 0 and variance 1 under the null:

$$Z = \sum_i \frac{1}{\sqrt{N}} Z(i) = 18\left(\frac{-\sqrt{2}}{\sqrt{126}}\right) + 60\left(\frac{0}{\sqrt{126}}\right) + 48\left(\frac{\sqrt{2}}{\sqrt{126}}\right) = 3.78 \quad (10)$$

Note that this is an observation from a normal distribution and not an LOD score as in the previous examples. For many likelihood ratio tests, twice the natural log of the likelihood ratio is asymptotically distributed as a chi-squared statistic, so we can compute an approximate equivalence since a normal squared is a chi-square with one degree of freedom:

$$\frac{Z^2}{2 \ln 10} = \frac{3.78^2}{4.6} \approx \text{LOD} = 3.10 \quad (11)$$

The ASP test for this example is slightly more significant, but this is only an approximation since the NPL is not based on maximization of a likelihood estimator.

References

Barnard GA (1949) Statistical inference. *Journal of the Royal Statistical Society*, **B11**, 115–135.

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, **52**, 362–374.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Kruglyak L and Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, **57**, 439–454.

Morton NE (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics*, **7**, 277–318.

Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics*, **46**, 229–241.

Suarez BK and Cox NJ (1985) Linkage analysis for psychiatric disorders. I. Basic concepts. *Psychiatric Developments*, **3**, 219–243.

Whittemore AS and Halpern J (1994) A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

Introductory Review

Genetics of complex diseases: lessons from type 2 diabetes

Leif Groop and Peter Almgren
Lund University, Malmö, Sweden

1. Introduction

A disease can be inherited or acquired or both. While cystic fibrosis is an example of an inherited disease, most infectious diseases are acquired. But susceptibility to an infectious disease can also be influenced by genetic factors. Heterozygous carriers of the mutation causing sickle-cell anemia are resistant against malaria (Miller *et al*., 1975). A 32-bp deletion in the gene encoding for the lymphoblastoid chemokine receptor (CCR5) was introduced in Europe during the plague by *Yersina pestis* in the fourteenth century (Stephens *et al*., 1998). Carriers of this deletion are today less susceptible to HIV infections.

Cystic fibrosis is caused by mutations in one gene, *CTFR*, and represents a monogenic disorder with early onset, usually from birth. The segregation of the disease follows a clear Mendelian recessive inheritance; in Europe is one out 2000 children affected. In contrast, a polygenic disease such as type 2 diabetes (T2D) is caused by "mild" variations in several genes and has a late onset. A polygenic disease is also referred to as complex because of its complex inheritance pattern. A complex disease often appears to be acquired; the development of obesity and T2D is triggered by environmental factors such as intake of dense caloric food and lack of exercise in genetically susceptible individuals.

2. Genetic risk

The *relative genetic risk (λs)* of an inherited disease is defined as the recurrence risk for a sibling of an affected person divided by the risk for the general population. The higher the λs, the easier is it to map the genetic cause of the disease. The λs value for cystic fibrosis is approximately 500, for type 1 diabetes (T1D) 15 and for T2D 3. It is, therefore, not surprising that cystic fibrosis was the first inherited disease to be mapped by linkage analysis (Kerem *et al*., 1989), whereas success for T2D has been limited (Florez *et al*., 2003).

The relative genetic risk should not be mixed by the *population attributable risk (PAR)*. This is important from a public health perspective, but does not tell anything

about the individual risk. It describes the fraction of a disease, which would be eliminated if the genetic risk factor were removed from the population. PAR is high in monogenic rare disorders such as cystic fibrosis (around 50) but low for rare alleles in complex diseases. If the disease-associated allele is common, PAR increases. This is illustrated by the role of the Apo $\varepsilon 4$ allele in the Alzheimer's disease. The PAR for Apo $\varepsilon 4$ in Alzheimer's disease is 0.2 because of the high frequency of the Apo $\varepsilon 4$ allele in the population (16%).

3. Genetic variability

Mapping of an inherited disease requires the identification of the genetic variability contributing to the disease. Such variability is often a change in a single nucleotide in the genome, a single-nucleotide polymorphism (SNP). There are about 10 million SNPs in the human 3 billion bp genome, which means one SNP at about 300 bp intervals. SNPs in coding sequences (exons) are seen at 1250 bp intervals. Microsatellites are short tandem repeats of nucleotide sequences (e.g., CA) found at about 5000 bp intervals. Whereas SNPs are frequently biallelic, microsatellites have multiple alleles and are thus much more polymorphic than SNPs. Several public databases provide information on SNPs (e.g., **www.ncbi.nlm.nig.gov/**SNP.

A SNP can either be the cause of the disease (causative SNP) or it can be a marker of the disease. This occurs when the disease susceptibility allele and the marker allele are so close to each other that they are inherited together, a situation called *linkage disequilibrium (LD or allelic association).* Such a combination of tightly linked alleles on a discrete chromosome is called a haplotype. While this region is characterized by little or no recombination (haplotype block), regions with high recombination rate usually separate haplotype blocks. An international effort to create a genome-wide map of LD and haplotype blocks is called the HapMap project (**http://www.hapmap.org/groups.html**). The hope is that by knowing the haplotype block structure of the genome, one would only need to genotype a few representative SNPs for the haplotype block (tag SNPs) rather than all SNPs.

4. Mapping genetic variability

The ultimate goal of mapping genetic variability is to identify the mutation causing a monogenic disease or the SNPs increasing susceptibility to a polygenic disease.

5. Linkage

The traditional way of mapping a disease gene has been to search for linkage between a chromosomal region and a disease by genotyping a large number (about 400–500) of polymorphic markers (microsatellites) in affected family members. If the affected family members would share an allele more often than expected by nonrandom Mendelian inheritance, there is evidence of excess allele sharing.

Ideally, such a genome-wide scan would be carried out in large pedigrees where mode of inheritance and penetrance is known. Since these parameters are not known and parents are rarely available in a complex disease with late onset, most genome-wide scans are performed in affected siblings with no assumptions of mode of inheritance and penetrance (nonparametric linkage).

The probability test of linkage is called the *LOD score* (logarithm of odds). Two loci are considered linked when the probability of linkage as opposed to the probability against linkage is equal to or greater than the ratio of 1000/1. An LOD score of 3 corresponds to an odds ratio of 1000/1 ($p < 10^{-4}$). In a study of affected sib pairs, a nonparametric linkage (nonparametric linkage score (NPL)) score is presented. Although this threshold was developed for monogenic disorders with complete information of genotype and phenotype, the situation in mapping complex disorders is much more complex. Lander and Kruglyak (1995) have proposed that the LOD threshold for *significant genome-wide linkage* should be raised to 3.6 ($p < 2 \times 10^{-5}$), while that for *suggestive linkage* (would occur one time at random in a genome-wide scan) can be set at 2.2 ($p < 7 \times 2^{-4}$). In addition, they suggest to report all nominal p-values < 0.05 without any claim for linkage. In reality, these thresholds differ from scan to scan depending upon information content. Therefore, these thresholds should be simulated using the existing data set before any claims of linkage can be made.

Accurate definition of the phenotype is a prerequisite for success but this may not always be easy and dichotomizing variables may result in loss of power. One alternative is, therefore, to search linkage to a quantitative trait, for example, blood glucose, blood pressure body mass index instead of diabetes, hypertension, and obesity.

Linkage in complex diseases will only identify relatively large chromosomal regions (often >20 cM) with often more than 100 genes. Fine mapping with additional markers can narrow the region further but at the end the causative SNP or a SNP in LD with the causative SNP has to be identified by an association study. Several approaches have been described to estimate whether an observed association can account for linkage (Li *et al*., 2004). Without the functional support, it is not always possible to know whether linkage and association represent the genetic cause of the disease. This can for many complex disorders require a series of in vitro and in vivo studies.

6. Calpain 10 and type 2 diabetes

In the first successful genome-wide scan of a complex disease such as T2D, Graeme Bell and coworkers reported in 1996 that significant linkage (LOD 4.1; $p < 10^{-4}$) of T2D diabetes to a locus on chromosome 2q37 (Hanis *et al*., 1996). Still, this region was quite large (12 cM) encompassing a large number of putative genes. A reexamination of the data suggested an interaction (*epistasis*) with another locus on chromosome 15 (LOD 1.5). This enabled the researchers to narrow the region down to 7 cM. Luckily, the 7 cM genetic map only represented 1.7 Megabases of physical DNA. To clone the underlying gene, they genotyped a number of SNPs in this interval and identified a three-marker haplotype, which was associated with

T2D. It turned out that three intronic SNPs in the gene encoding for calpain 10 (*CAPN10*) could explain most of the linkage (Horikawa *et al*., 2000). Calpain 10, a cystein protease with largely unknown functions in glucose metabolism, was no obvious candidate gene for T2D. Despite a number of subsequent negative studies, several meta-analyses have shown consistent association of CAPN10 with T2D (Parikh and Groop, 2004). Carriers of the risk alleles are associated with decreased expression of the gene in skeletal muscle and insulin resistance. How this translates into increased risk of T2D is not known and will require functional studies.

7. Gene expression

Since genes are transcribed to RNA and RNA is translated into proteins and defects in proteins cause disease, the ultimate goal would be to carry out a random search of expressed proteins in target tissues. This may not yet be feasible but the study of transcript profiles is. This approach has been successful in defining prognosis of cancers but for complex diseases affecting many target tissues it may not be that simple. Also, defining what is differentially expressed among >20 000 gene transcripts on a chip is a statistical challenge. Despite these problems, analysis of gene expression in skeletal muscle of patients with T2D and prediabetic individuals has provided new insights into the pathogenesis of the disease. It, however, required analysis of coordinated gene expression in metabolic pathways rather than of individual genes. Genes regulating oxidative phosphorylation in mitochondria showed a 20% coordinated downregulation in muscle from prediabetic and diabetic individuals (Mootha *et al*., 2003; Patti *et al*., 2003). Furthermore, a similar down regulation of the gene encoding for a master regulator of oxidative phosphorylation, the PPARγ coactivator, PGC-1α was observed. These findings suggest that impaired mitochondrial function and impaired oxidation of fat may predispose to T2D through a "thrifty gene" mechanism (see below).

8. Association studies

If there is a prior strong candidate gene for the disease, the best approach is to search for association between an allelic variant of the gene and the disease. This can either be a case control or nested cohort study. In a case–control study, the inclusion criteria for the cases are predefined, and thereafter, matched individual controls are searched representing the same ethnic group as the cases. In a cohort study, affected and unaffected groups are matched. Ideally, cohorts should be population based and older than cases to exclude the possibility that they still will develop the disease. The question of matching is crucial for the results, matching for a parameter influenced by the genetic variant (e.g., BMI) might influence its effect on a disease such as T2D. If cases and controls are not drawn from the same ethnic group, a spurious association can be detected due to ethnic stratification.

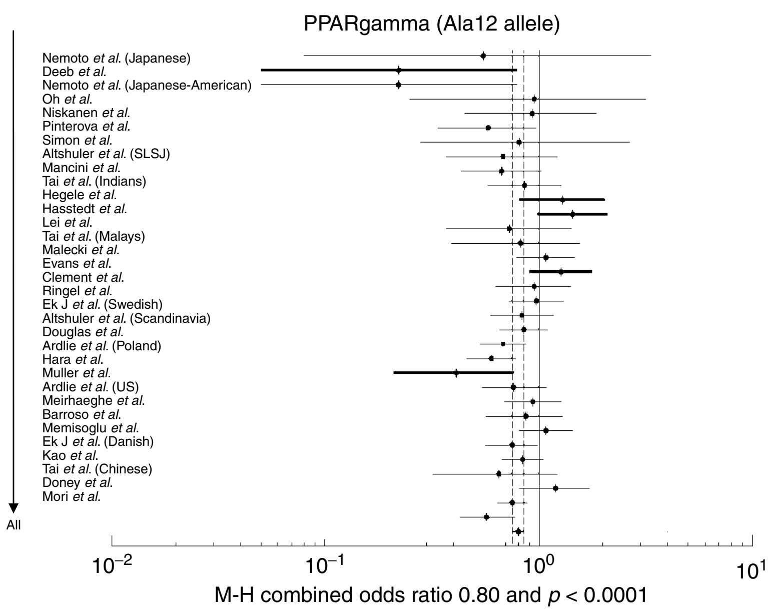
One way to circumvent this problem is to perform family-based association studies. Excess transmission of the disease-associated allele from heterozygous

parents to the affected offspring would indicate association in the face of linkage. This transmission disequilibrium test (TDT) represents the most unbiased association study approach but suffers from low power as only transmissions from heterozygous parents are informative. The prerequisite of DNA from parents usually enrich for individuals with an earlier onset of the disease.

Even screening only one gene for SNPs can represent a huge and expensive undertaking. The PPARγ gene on the short arm of chromosome 3 spans 83 000 nucleotides with 231 SNPs in public databases.

The gene encodes for a nuclear receptor, which is predominantly expressed in adipose tissue where it regulates transcription of genes involved in adipogenesis.

In the 5′ untranslated end of the gene is an extra exon B that contains a SNP changing a proline in position 12 of the protein to alanine. The rare Ala allele is seen in about 15% of Europeans and was in an initial study associated with increased transcriptional activity, increased insulin sensitivity, and protection against T2D (Deeb *et al*., 1998). A number of subsequent studies could not replicate the initial finding. However, using the TDT approach we could show excess transmission of the Pro allele to the affected offspring (Altshuler *et al*., 2000). A meta-analysis combining the results from all published studies showed a highly significant association with T2D ($p < 2 \times 10^{-10}$) (Figure 1). The individual risk reduction conferred by the Ala allele is only 15% but since the risk allele Pro is so common, it translates into a population attributable risk of 25%. There is also a strong interaction



Meta-analysis of the association between the Ala12 allele in the PPARγ gene and type 2 diabetes without Deeb *et al*., Hegele *et al*., Hasstedt *et al*., Evans *et al*. and Hara *et al*. studies (thick lines) to obtain homogeneity by Mantel–Haenszel Test of homogeneity.

Figure 1 A meta-analysis demonstrating the strong risk of the Pro12Ala polymorphism in the PPARγ2 gene and type 2 diabetes

with nutritional factors and the protective effect of the Ala allele is enhanced with a high intake of unsaturated free fatty acids (Luan *et al*., 2001). In fact, free fatty acids have been proposed as natural ligands for PPARγ.

It is still debated whether common or rare variants are the cause of common complex diseases. The common variant-common disease hypothesis assumes that relatively ancient common variants increase susceptibility to common diseases such as obesity, hypertension, T2D, and so on. These variants would be enriched in the population as they have been associated with survival advantage during the evolution, so-called thrifty genes (Neel, 1962). Storage of surplus energy during periods of famine may have been beneficial for survival, while in the Westernized society we rather need genetic variants which would waste energy.

9. Why is it difficult to replicate a finding of an association with a complex disease?

The literature on the genetics of complex diseases has been enriched with papers not being able to replicate the initial findings. There are several reasons for this. There is a clear tendency that the first study reports the strongest association, as researchers and editors prefer strong positive findings ("winners curse"). False-positive findings are unfortunately common. In an analysis of 301 published studies covering 25 different reported associations, only half showed significant replication in a meta-analysis (Lohmueller *et al*., 2003). The most important reason is lack of power. The odds ratio for a complex disease is often below 1.5. The sample size is dependent not only upon the odds ratio but also on the frequency of the at-risk genetic variant. For an odds ratio of 1.8 and a frequency of the at-risk allele of 25%, at least 1000 cases and controls are required (Figure 2).

10. Why do not linkage studies detect all associations?

Despite initial linkage, it has often been difficult to identify the underlying genetic variation. This is particularly difficult if the disease-causing allele has a high frequency in the population. If too many individuals will be homozygous for the disease allele, in which case one will not observe linkage between the disease allele and an allele at a nearby locus, because either of the homologous chromosomes can be observed as transmitted to an affected offspring. This was the case for the Pro12Ala polymorphism in the PPARγ gene. No linkage has been observed between T2D and the region for PPARγ on chromosome 3p, since the Pro allele will typically be transmitted from both parents. A simulation indicated that 3 million sib pairs would be required to detect such a linkage (Altshuler *et al*., 2000).

11. Future directions

The rapid improvement in high-throughput technology for SNP genotyping and decreasing costs per genotype (in 10 years the cost has decreased by a factor

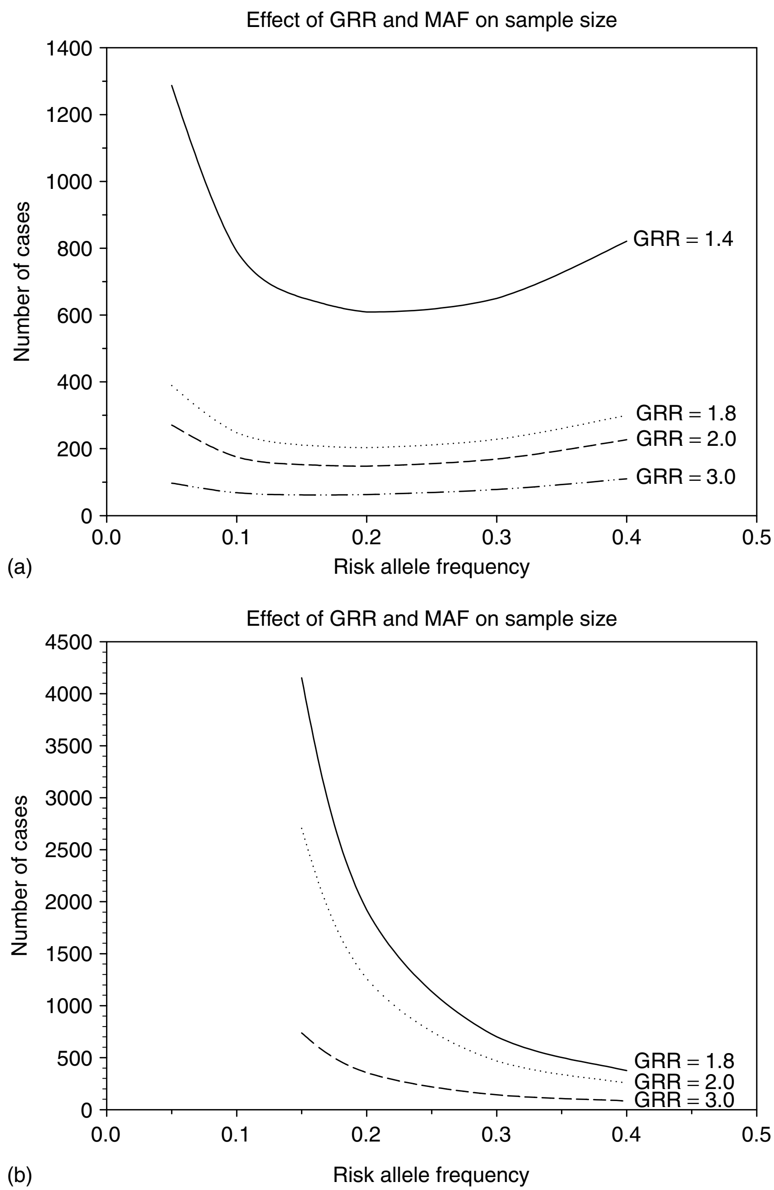


Figure 2 Number of cases and controls required for association studies are dependent upon the estimated odds ratio and the frequency of the at-risk allele. (a) Dominant model; (b) recessive model

of 10) open new possibilities for both linkage and association studies. The use of DNA chips containing >11 000 SNPs for genome-wide scans is estimated to be much more powerful than previously used 450 microsatellites (Middleton *et al*., 2004).

Such high-density DNA chips may be useful for detecting rare genes with a strong effect in large pedigrees, but for the detection of the genetic variation of complex diseases, association studies are needed. In the near future, it will be possible to perform genome-wide association studies using SNPs. If the common variant–common disease hypothesis holds, it may be possible to obtain an atlas of disease-associated genetic variants using approximately 500 000–1 000 000 SNPs. This will not be a cheap undertaking and it is obvious that while the tools are there, money is still the limiting factor for dissecting the genetics of complex diseases.

References

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, *et al*. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, **26**, 76–80.

Deeb SS, Fajas L, Nemoto M, Pihlajamaki J, Mykkanen L, Kuusisto J, Laakso M, Fujimoto W and Auwerx J (1998) A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genetics*, **20**, 284–287.

Florez JC, Hirschhorn J and Altshuler D (2003) The inherited basis of diabetes mellitus. Implications for the genetic analysis of complex traits. *Annual Review of Genomics and Human Genetics*, **4**, 257–291.

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, Wapelhorst B, Spielman RS, Gogolin-Ewens KJ, *et al*. (1996) A genome–wide search for human non–insulin–dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genetics*, **13**, 161–166.

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, *et al*. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics*, **26**, 163–175.

Kerem B, Rommens JM, Buchanan JA, Markiewics D, Cox TK, Chakravati A, Buchwald M and Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, 1073–1080.

Lander E and Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

Li C, Scott LJ and Boehnke M (2004) Assessing whether an allele can account in part for a linkage signal: the Genotype-IBD Sharing Test (GIST). *American Journal of Human Genetics*, **74**, 418–431.

Lohmueller KE, Pearce CL, Pike M, Lander ES and Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**, 177–182.

Luan J, Browne PO, Harding AH, Halsall DJ, O'Rahilly S, Chatterjee VK and Wareham NJ (2001) Evidence for gene-nutrient interaction at the *PPAR*gamma locus. *Diabetes*, **50**, 686–689.

Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H, *et al*. (2004) Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *American Journal of Human Genetics*, **74**, 886–897.

Miller LH, Mason SJ, Dvorak JA, McGinniss MH and Rothman IK (1975) Erythrocyte receptors for (*Plasmodium knowlesi*) malaria. *Duffy Blood Group Determinants Science*, **189**, 561–563.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267–273.

Neel V (1962) Diabetes mellitus: a ”thrifty” genotype rendered detrimental by progress. *American Journal of Human Genetics*, **14**, 352–362.

Parikh H and Groop L (2004) Candidate genes for type 2 diabetes. *Reviews in Endocrine and Metabolic Disorders*, **5**, 151–176.

Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, Miyazaki Y, Kohane I, Costello M, Saccone R, *et al*. (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8466–8471.

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, *et al*. (1998) Dating the origin of the CCR-5-Delta 32 AIDS resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics*, **62**, 1507–1515.

Introductory Review

Importance of complex traits

Joseph H. Lee
Columbia University, New York, NY, US

1. What is a complex trait?

Complex traits refer to traits that do not follow simple Mendelian inheritance. These include most common chronic diseases (e.g., Alzheimer's dementia, coronary heart diseases, obesity), infectious diseases (e.g., hepatitis, malaria), and common physiological traits (e.g., blood pressure, obesity). These common diseases or traits contribute much to the public health burden in both the developed and the developing countries. For example, the three most common types of diseases – cancer, cardiovascular diseases, and neuropsychiatric disorders – account for approximately 56 and 25% of the years of life lost to premature death and years lived with a disability in the developed countries and the developing countries, respectively (WHO, 1999). To develop effective therapeutic and preventive measures, it is desirable to identify the underlying genetic and environmental factors to understand the biology of these diseases. However, identification of the underlying causes is difficult for these traits because there are multiple genetic or environmental factors, and each individual factor is likely to contribute only little toward the disease or the trait. Further, these factors can interact to produce different effects. A number of different approaches have been applied to reduce the complexity and to understand the underlying components, but challenges still remain.

2. Genetic architecture of complex traits

Most traits are caused by a network of genetic and environmental risk factors (Figure 1). Obviously it is difficult to characterize any causal relations in a one-to-one manner, since a single factor among a collection of interrelated factors is likely to explain only a small proportion of variation in phenotype. Moreover, when the distribution of these cofactors differs in different samples, the observed association between any one potential putative factor and the trait is likely to vary, leading to inconsistent observations across studies. For example, in one population, Gene 1 may be the predominant form leading to the disease; while in another population, Gene 3 may be most prevalent, leading to a conclusion of "nonreplication" of the

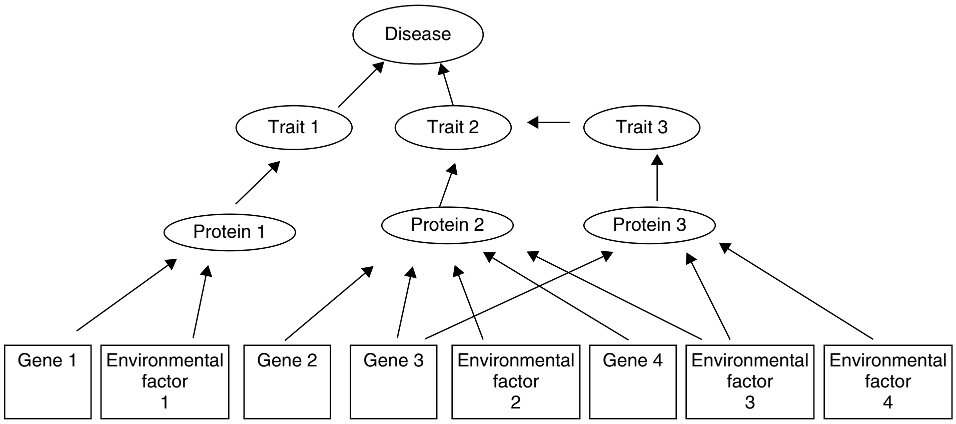


Figure 1 This simplified figure shows how genetic and environmental factors contribute to causal pathways toward a common disease

first association. Thus, depending on the population sampling, the observed support for a causal relation with a genetic factor can vary.

In complex disorders, genes not only confer small to modest effects on the disease, but are likely to contribute to variable expression via many different means, as in earlier age at onset, varying levels of severity, or some other means. Moreover, mechanisms by which these genes influence the trait may deviate from our traditional understanding of genetic inheritance. Imprinting affects the phenotypic expression of a genetic variant by its parent of origin, and this phenomenon has been reported in some cancers and neuropsychiatric diseases (Bjornsson *et al.*, 2004). Others have reported that gene duplications and deletions may alter normal physiological variations and eventually the risk of cancers, depending on the copy number polymorphism (Buckland, 2003). Some reports have shown that variants in mitochondrial DNA contribute a wide range of phenotypes, including aging, neurological disorders, and diabetes (Wallace, 2005). These illustrate that determination of genetic mechanism, even after the gene has been identified, can be difficult. Moreover, some environmental factors can have major contributions, while those of others may be minor. Because some behaviors (e.g., smoking, diet, academic achievement) run in families and can mimic genetic transmission, disentangling the multitude of etiologic factors requires a careful study design and multiple approaches.

3. Different study designs for different purposes

To understand the nature of the complex trait, researchers have used a number of different study designs to delve into the network of causation. They have attempted to determine evidence of genetic effect, identify the gene, measure the effect size of the gene, and then assess its functions. To assess the role of genetic factors directly or indirectly, researchers have used twin studies, adoption studies, genetic isolate studies, and family studies. Detailed discussion of the strengths and weaknesses

of these study designs are discussed elsewhere (Rao and Province, 2001; Strachan and Read, 1999; Terwilliger and Goring, 2000). It is hotly debated as to what an optimal study for complex traits would be (Botstein and Risch, 2003; Hirschhorn and Daly, 2005; Weiss and Terwilliger, 2000; Wright *et al*., 1999). In the end, each disease or trait will require a somewhat different study design, depending on the hypothesis to be tested and the available study samples. Fundamentally, to identify genes, it is generally desirable to simplify the complex relations among causal factors, and this is best achieved in a family study design. This approach can be further simplified by examining genetically isolated populations. On the other hand, to characterize population parameters (e.g., a gene's effect size and its impact in the population), it is desirable to include all the factors that are believed to be significant contributors. For this purpose, a population-based epidemiologic study design is more suitable.

Traditionally, large extended families with multiple affected individuals were used to identify genes. It is reasoned that if a gene causes a disease, it is likely that the disease will "run" in that family, and will lead to many affected family members, who share the risk alleles from some common ancestor. Thus, the gene can be localized to some chromosomal region by linkage analysis. Linkage analysis is based on the observation that chromosomal loci that are physically close will tend to be coinherited more often than will two random loci in the genome. A disease-causing or a disease-influencing variant will cosegregate with nearby genetic markers. Thus, by identifying genetic markers that are shared by affected individuals in a given family, it will be possible to identify chromosomal regions that may harbor the disease gene. Using this principle, researchers have searched the human genome to identify many causative genes, mainly for Mendelian disorders (McKusick, 1998). In general, the ascertainment using large families increases the predictive power of phenotype for the underlying genotypes. Because a limited number of large families are examined and the environment within the family is more similar than in random samples, this approach reduces the number of genetic variants and environmental risk factors, thereby reducing the complexity. Examples of genes for complex traits identified to date using large families include Breast Cancer Gene 1 and 2 (*BRCA1*, *2*) (Hall *et al*., 1990; Miki *et al*., 1994; Tavtigian *et al*., 1996; Wooster *et al*., 1995) for breast cancer, and presenilin 1 and 2 (*PS1*, *2*) (Levy-Lahad and Bird, 1996; Rogaev *et al*., 1995; Sherrington *et al*., 1995) and apolipoprotein E (*APOE*) (Strittmatter *et al*., 1993) for Alzheimer's disease (AD).

Other researchers have argued against using large families because such large families with multiple affected individuals are difficult to find and the identified genes tend to be rare so that they explain only a small and unique proportion of the disease (i.e., the "familial" variant of a given common disorder). As a result, many opted to recruit a greater number of smaller families, such as affected sibling pairs, trios (parents and affected child), or even cases and controls. However, these approaches have had limited success. As a large number of sibling pairs, trios, or cases and controls are used, the number of genetic variants in affected individuals as well as the number of environmental risk factors will increase, so each risk factor will explain proportionately less of the phenotypic variance in the samples studied.

Recently, many investigators have combined existing large population-based studies with high-throughput and high-resolution genotyping technologies to search

their way through the genome in genome-wide association studies. This approach uses the principle that the allelic association can be made directly with single-nucleotide polymorphisms (SNPs) that cause an amino acid change (Botstein and Risch, 2003) or with SNPs that are close to the true disease-causing variant (Hinds *et al*., 2005; Maraganore *et al*., 2005). Several issues need to be considered when evaluating allelic associations. First, population stratification can lead to false-positive findings. This happens when cases are recruited from an ancestry that is different from that of the controls, thereby leading to different allele frequencies between the two groups, independent of the disease status. Consequently, the allele frequency difference between ethnic groups will be falsely attributed to the allele frequency difference between cases and controls. This problem will need to be addressed using genomic control methods (Devlin *et al*., 2004). Second, because of the nature of linkage disequilibrium (see Article 000, **Impact of Linkage Disequilibrium on Multipoint Linkage Analysis**, Volume 0), allelic association requires a great many more markers compared to linkage analysis. For example, because of multiple markers tested, a SNP needs to achieve a nominal p value of 5×10^{-7} to be comparable to the Bonferroni corrected p value of 0.05, when a "100K SNP chip" (i.e., 100 000 markers) is used for the genotyping. This is only for one phenotype. To circumvent this problem, some have advocated a two-stage design in which one dataset is used to screen for candidate SNPs associated with the disease, and then the finding confirmed or refuted in another dataset (van den Oord and Sullivan, 2003). Moreover, even when 100 000 SNPs are tested, they cover $<50\%$ of the common variants (Hirschhorn and Daly, 2005); therefore, there is a risk of overlooking a significant number of common variants that may contribute to the disease. Third, perhaps the most significant difficulty lies in allelic heterogeneity, which cannot be overcome by statistical means. To identify susceptibility genes using the allelic association approach, it is necessary to identify not only the correct genetic marker but also the correct allele. For example, using large extended pedigrees, it was found that PS1 mutations cause familial early onset Alzheimer's disease (EOAD). However, even for this 'less complex' subset of Alzheimer's disease, researchers have identified 155 allelic variants from 315 families (**http://www.molgen.ua.ac.be/ADMutations/**) thus far, and the number will only increase. Thus, if a SNP chip is used for diagnostic purpose, the results from such tests may still be less than conclusive. Lastly, these susceptibility alleles confer only small effects; thus, the effect size will vary considerably depending on the set of genetic and environmental risk factors present.

Despite these difficulties, some have reported successes. Klein and colleagues examined 116 204 SNPs using 96 cases and 50 controls (Klein *et al*., 2005). The authors identified that a common variant in an intron of complement factor *H* gene was strongly associated with age-related macular degeneration. However, it is of interest to note that this locus had been reported by linkage analysis. Herbert and colleagues (2006) reported a common allele near the *INSIG2* gene that is associated with adult and childhood obesity. For this purpose, the authors examined multiple datasets, including families, trios, and cases–controls, and found the association to be significant in four of five datasets. However, to minimize the number of statistical tests, the authors used a two-stage approach in which they screened 86 604 SNPs in parents, and then tested the top 10 SNPs with the greatest power

to detect allelic association in their offspring. They identified one SNP from this analysis and confirmed it in four different datasets. Although these two studies report positive findings based on a genome-wide association approach, it remains to be seen whether this approach can be fruitful even after "low hanging fruits" have been picked.

Estimation of genetic parameters, once a susceptibility gene has been identified, requires large randomly selected samples. This avoids the ascertainment bias that lead to overestimation of gene parameters (e.g., allele frequencies, attributable risk) arising from the extreme samples that identified the gene. For example, *BRCA1* identified from the large pedigrees with multiple breast cancer cases was thought to be highly prevalence and penetrant. In one study, 84% of the families with four or more cases that had onset prior to 60 years of age were associated with either *BRCA1* or *BRCA2* mutations (Thompson and Easton, 2004). In the population-based studies of randomly selected population samples, however, the prevalence of mutations in this gene was $<5\%$, highlighting the strong influence of ascertainment on genetic parameters.

4. Endophenotype

As a means to simplify the complexity, biologically relevant endophenotypes can help us to understand the biology better. A genetic variant can have a cascading effect on protein structure, protein synthesis, and gene expression levels, while environmental factors can modulate the effect of the genetic variant each step along the way. Common diseases are far removed from direct actions of the underlying genetic and environmental factors (Figure 1). To identify the genetic or environmental risk factors, it makes sense to study intermediate phenotypes that are closely associated with metabolism, physiology, structural integrity, and related traits, which are themselves risk factors for the disease. These traits tend to be continuous (rather than affected vs. unaffected), so that they explain normal variations in the population. For example, endophenotypes can include cholesterol levels for coronary heart disease and amyloid β level and brain structure for AD. Essentially, the use of endophenotypes strengthens genotype–phenotype relations, and increases the predictive power of phenotypes for genotype. Consequently, it reduces the likelihood of falsely concluding an effect of phenocopy.

Currently many studies use endophenotypes instead of clinical endpoints to better understand complex traits. For example, using amyloid β protein level as the phenotype, Ertekin-Taner *et al*. (2000) localized a susceptibility gene for AD to 10q, which was confirmed in two independent studies (Bertram *et al*., 2000; Myers *et al*., 2000) that used AD as the phenotype. Others have used endophenotypes – such as P50 and P300 event-related potential wave measures for schizophrenia (Blackwood *et al*., 2001; Freedman *et al*., 1997) and blood coagulation for the risk of thromboembolic disease (Souto *et al*., 2003)–to localize the susceptibility locus for the clinical endpoints. In addition, these endophenotypes can be used to disentangle the complexity of the clinical endophenotype. Lee *et al*. (2006) showed, using memory scores as the phenotypes, that one candidate gene was more likely to contribute to AD via memory impairment, while another was more

likely to contribute to variation in memory in healthy elderly. From the practical standpoint, the several advantages of studying endophenotypes are as follows: (1) ascertainment is simpler; (2) greater information can be obtained as everyone has measurable values as part of normal human variation; and (3) there is efficiency as multiple quasi-independent quantitative traits (e.g., body mass index, lipid profile, intraocular pressure, bone density, vision, etc.) can be obtained from a single study.

In sum, complex traits present a major challenge for biomedical scientists, as current approaches have yielded limited success. No single approach can provide the best answer to our understanding of complex traits. It is necessary to choose an optimal study design for the hypothesis of interest and apply multiple integrated approaches.

References

Bertram L, Blacker D, Mullin K, Keeney D, Jones J, Basu S, Yhu S, McInnis MG, Go RC, Vekrellis K, *et al* (2000) Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. *Science*, **290**(5500), 2302–2303.

Bjornsson HT, Fallin MD and Feinberg AP (2004) An integrated epigenetic and genetic approach to common human disease. *Trends in Genetics*, **20**(8), 350–358.

Blackwood DH, Fordyce A, Walker MT, St Clair DM, Porteous DJ and Muir WJ (2001) Schizophrenia and affective disorders–cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *American Journal of Human Genetics*, **69**(2), 428–433.

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl), 228–237.

Buckland PR (2003) Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Annals of Medicine*, **35**(5), 308–315.

Devlin B, Bacanu SA and Roeder K (2004) Genomic Control to the extreme. *Nature Genetics*, **36**(11), 1129–1130; author reply 1131.

Ertekin-Taner N, Graff-Radford N, Younkin LH, Eckman C, Baker M, Adamson J, Ronald J, Blangero J, Hutton M and Younkin SG (2000) Linkage of plasma Abeta42 to a quantitative locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Science*, **290**(5500), 2303–2304.

Freedman R, Coon H, Myles-Worsley M, Orr-Urtreger A, Olincy A, Davis A, Polymeropoulos M, Holik J, Hopkins J, Hoff M, *et al* (1997) Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(2), 587–592.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B and King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, **250**(4988), 1684–1689.

Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, *et al* (2006) A common genetic variant is associated with adult and childhood obesity. *Science*, **312**(5771), 279–283.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA and Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**(5712), 1072–1079.

Hirschhorn JN and Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, **6**(2), 95–108.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, *et al* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.

Lee JH, Lee H-S, Santana V, Williamson J, Lantigua R, Medrano M, *et al* (2006) Genetic dissection of the Alzheimer disease phenotype via memory traits. *Neurology*, **66**(5), A380, Suppl. 2.

Levy-Lahad E and Bird TD (1996) Genetic factors in Alzheimer's disease: a review of recent advances. *Annals of Neurology*, **40**(6), 829–840.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR and Bellinger DF (2005) High-resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics*, **77**(5), 685–693.

McKusick VA (1998) *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*, Twelfth Edition, Johns Hopkins University Press: Baltimore.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, *et al* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**(5182), 66–71.

Myers A, Holmans P, Marshall H, Kwon J, Meyer D, Ramic D, Shears S, Booth J, DeVrieze FW, Crook R, *et al* (2000) Susceptibility locus for Alzheimer's disease on chromosome 10. *Science*, **290**(5500), 2304–2305.

van den Oord EJ and Sullivan PF (2003) False discoveries and models for gene discovery. *Trends in Genetics*, **19**(10), 537–542.

Rao DC and Province MA (2001) *Genetic Dissection of Complex Traits*. Academic Press: San Diego, CA.

Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, Chi N, Lin C, Holman K, Tsuda T, *et al* (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature*, **376**(6543), 775–778.

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi N, Lin C, Li G, Holman K, *et al* (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, **375**(6534), 754–760.

Souto JC, Almasy L, Soria JM, Buil A, Stone W, Lathrop M, Blangero J and Fontcuberta J (2003) Genome-wide linkage analysis of von Willebrand factor plasma levels: results from the GAIT project. *Thrombosis and Haemostasis*, **89**(3), 468–474.

Strachan T and Read AP (1999) *Human Molecular Genetics 2*, Second Edition, Wiley-Liss: New York.

Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS and Roses AD (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, **90**(5), 1977–1981.

Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, *et al* (1996) The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genetics*, **12**(3), 333–337.

Terwilliger JD and Goring HH (2000) Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Human Biology*, **72**(1), 63–132.

Thompson D and Easton D. (2004) The genetic epidemiology of breast cancer genes. *Journal of Mammary Gland Biology and Neoplasia*, **9**(3), 221–236.

Wallace DC (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annual Review of Genetics*, **39**, 359–407.

Weiss KM and Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nature Genetics*, **26**(2), 151–157.

WHO (1999) *The World Health Report 1999: Making a Difference*. WHO: Geneva, Switzerland.

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C and Micklem G (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.

Wright AF, Carothers AD and Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23**(4), 397–404.

Specialist Review

Concept of complex trait genetics

Kenneth M. Weiss and Anne V. Buchanan
Pennsylvania State University, University Park, PA, USA

1. Complex traits and common disease

In genetics, the term "complex" is not standardized but usually refers to traits for which there is no single cause with uniformly high predictive power. Complex traits are thought to involve both multiple genes and environmental risk factors, as well as interaction between a gene and other genes and/or the environment. However, it has generally not been possible to enumerate an exhaustive set of associated genes or environmental risk factors for such traits. This is in part because the relationship between genotype and phenotype is in most cases not one to one, though there may be some instances, usually rare, that follow Mendelian patterns of inheritance or that seem to have a single major cause. In fact, an enumerable, exhaustive set of causes may not even exist because the factors are dynamic, changing, or variable over time or space within and among populations.

Complex diseases are currently the focus of much attention in human genetics and epidemiology because they are the most common causes of morbidity and mortality in industrialized nations. In addition to efforts to understand causation, they are considered to be potentially good targets for pharmacogenomic intervention. Complex diseases are a mixed set of traits, however; some, including cardiovascular diseases, hypertension, diabetes, cancers, asthma and other pulmonary diseases, allergies, psychiatric disorders, and Alzheimer's disease, are common chronic complex diseases in the fullest sense defined above. Others, such as Hirschsprung disease (Gabriel *et al*., 2002) or Bardet–Biedl syndrome (Eichers *et al*., 2004), are rare and seem to be due to several genes, while still others, such as autoimmune disorders, are relatively rare but apparently involve major contributions from both genes and environment. Even some traits with the simplest genetic etiology involve many of the same elements of complexity (Scriver and Waters, 1999).

Complex diseases also differ in age at onset; some, such as Alzheimer's or coronary heart disease, develop only after decades of healthy life, while others are diseases of childhood, for example, allergic asthma and many psychiatric disorders. And the prevalence of many of them has risen sharply in recent decades (NIDDM, autism, and asthma), or even fallen dramatically (cardiovascular disease in North Karelia, Finland, for example, where heart disease mortality rates were once the highest in the world but which have fallen by 75% in the last several decades (Neroth, 2004)). These changes clearly show the involvement of environmental

factors. However, the prevalence of other complex traits has remained fairly stable over time (e.g., some of the psychiatric disorders), and may be similar in very different populations.

Many complex traits, particularly those that arise at older ages, are best viewed not as biological entities per se, but as the tail ends (extremes) of the distribution of normal variation. In such cases, risk can be described statistically in terms of the deviation of an individual's underlying risk from the population mean. Cholesterol, weight, blood pressure, and glucose tolerance are such factors, for which everyone in the population has a value, but for which the probability of associated pathology increases when these measures fall in the upper tail of the distribution. Many of these traits are intermediate variables in a pathway from normal to disease, as illustrated in Figure 1(a). In fact, conditions such as "high" blood pressure or glucose levels are often defined not by their own pathological symptoms, but by an arbitrarily decided level of statistical risk for some future outcome (stroke, diabetes), based on population-level correlations. This is important to keep in mind as we try to understand the biological nature of complex traits.

There clearly is not a one-size-fits-all explanation for this diversity of traits. An example that illustrates some of these points, and shows the difficulty of attributing causation, is allergic asthma (for a more in depth discussion, *see* Article 61, **Allergy and asthma**, Volume 2). Asthma has become a very common disease in industrialized countries, most often arises early in childhood, and may involve a critical period during development, perhaps of the immune system, during which exposure to an environmental insult is the trigger. The condition aggregates in families (that is, risk is higher for children with affected sibs than for random children in the population), but does not segregate in a Mendelian way as if caused by a single gene. Recent reports of genes "for" asthma (*ADAM33*, or different interleukin genes, see, for example Wills-Karp and Ewart, 2004) have been equivocal, replicated by some studies but not by others.

Noninsulin dependent diabetes mellitus (NIDDM), or type II diabetes, is another complex disease whose epidemiology suggests a genetic component, although prevalence rises significantly in all populations that "westernize", that is, are exposed to an overabundance of calories, with reduced levels of physical activity (Stern *et al*., 1991; Trowell and Burkitt, 1981). Unlike asthma, NIDDM usually develops gradually at later ages, although age at onset has been falling in high-risk populations and now even children are developing what used to be an adult onset disease. The history of the disease in the last century suggests the involvement of exposure to some environmental trigger, perhaps in childhood, or even in utero. The factor(s) seem likely to involve diet and exercise in some way. Comparative epidemiology of this disease in different populations suggests the possibility of a relatively simple genetic risk, at least in some populations, but even so, the rise in prevalence with lifestyle changes strongly implicates environmental risk factors; NIDDM may be one of the most significant examples of a common complex disease due to gene-environment interaction. Numerous searches for diabetes genes in different populations, most notably Native American groups in which NIDDM prevalence is very high, have mainly yielded suggestive or even contradictory results, or genes of only minor contributory effect. Few genes are known in which

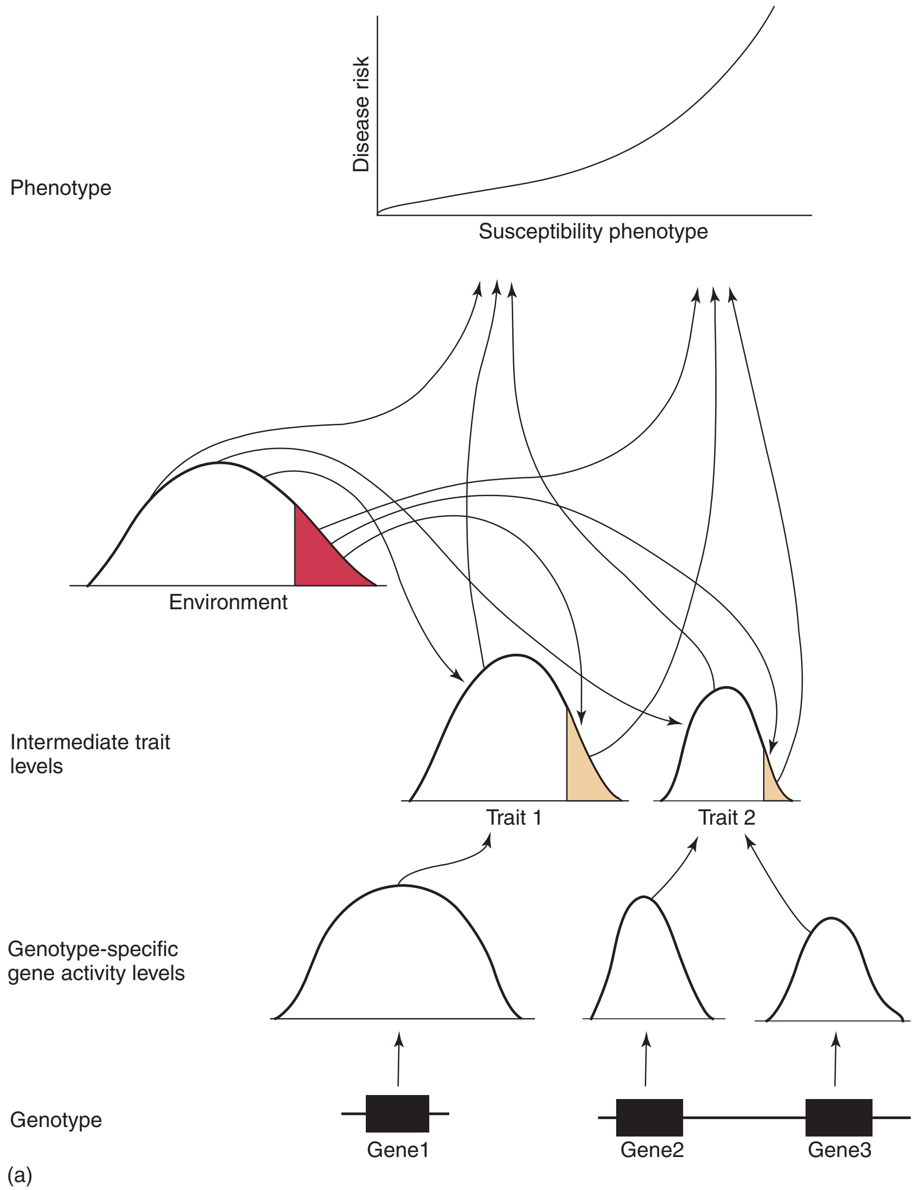


Figure 1 (a) Pathway from genotype to disease phenotype, showing genes 1, 2, and 3 yielding distributions of allele frequencies, which, singly or with other genes and environmental factors, lead to the distribution of traits (cholesterol or HDL levels are possible examples) that, at their extremes and in interaction with environmental risk factors, can increase risk of disease; (b) Hourglass figure showing that many genotypes may produce intermediate traits that can be risk factors for disease, alone, with other genes, and/or environmental risk factors

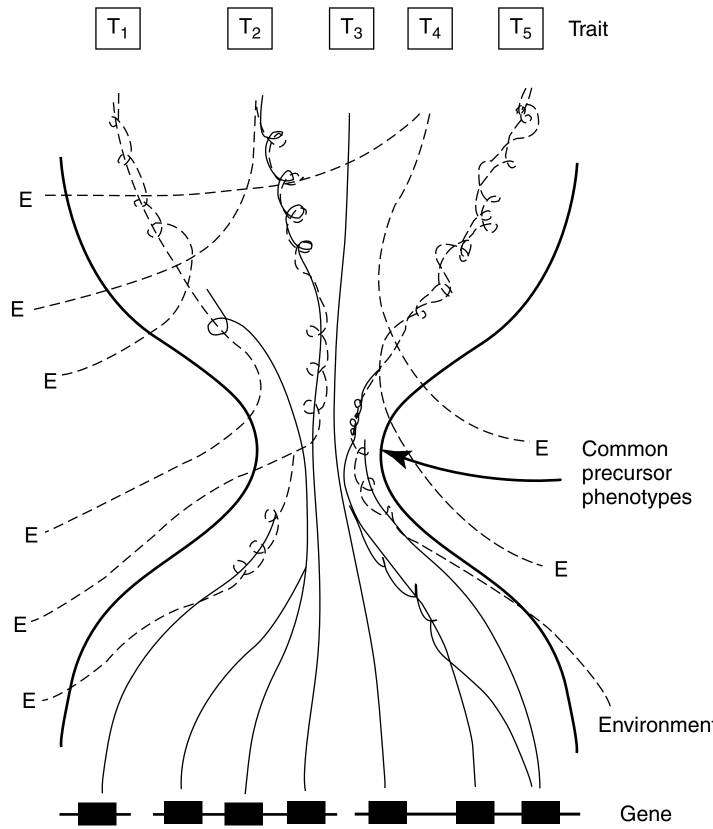


Figure 1 (*continued*)

mutations confer very elevated risk, and these only account for a small fraction of all cases.

Psychiatric disorders are usually considered genetic diseases when there is evidence that cases aggregate in families, at least to some extent (*see* Article 64, **Genetics of cognitive disorders**, Volume 2), and when the biochemistry seems to be suggestive of genetic involvement, but as of yet, none has been substantially explained in genetic terms. Autism has an early age at onset, and siblings of affected children have a higher risk than the general population, suggesting a genetic component, but the phenotype varies considerably, even within affected families, and this has made it particularly difficult to study. Again, incidence has been rising rapidly in the industrialized world, and changing diagnostic criteria, infectious or environmental agents such as preservatives used in childhood vaccines, as well as genes, have all been implicated, but the etiology is scarcely better understood now than when incidence began to rise more than a decade ago, and the search for genes involved has yielded some suggestive loci but little in the way of definitive results (Veenstra-VanderWeele and Cook, 2004).

2. Analytic issues

The above examples are biologically very different, but share the same general epidemiological features – clustering of risk in families that implicates genetic risk factors, and rapid rise in prevalence that suggests a role for environmental factors. The basic theoretical model for the *genetics* of such traits has been known since early in the twentieth century. Complex traits, even including quantitative traits such as blood pressure or stature, reflect the combined effects of alleles at many independently segregating "Mendelian" loci. Individuals inheriting different combinations of alleles at the loci can have a similar phenotype; that is, even under the simplest of circumstances, and only considering genes, there is a many-to-one relationship between genotype and phenotype. If only a modest number of loci contribute substantial effects, the trait is sometimes called *oligogenic*, whereas if a large number of loci contribute, the trait is called *polygenic*. Even most oligogenic traits have a polygenic "background", that is, many alleles of individually very small effect modifying the phenotype. Polygenic effects are so numerous and small that they are not individually identifiable in practice.

In any given population, most loci contain at most a small number of common alleles, and many rare ones (frequency $<1\%$ or even much less). In some instances or in some populations, rare alleles at some loci may have strong effect on the trait (high "penetrance"). These effects will manifest Mendelian segregation in families and the genes can be mapped by existing methods. Such genes are sometimes referred to as *Quantitative Trait Loci* (QTL). QTL alleles are usually individually rare so that any given family will be affected by only one of them, but the specific allele or locus will differ among families; this is known as *multiple unilocus* etiology.

Common alleles with strong effect can be mapped, just as rare ones can, although an allele that is very common essentially defines "normal" rather than disease. For this reason, there is typically a confounding of size or strength of effect of an allele and its frequency. Strong effects are usually rare, because strong effects usually mean far from normal. This leads to serious analytic challenges, especially if weaker effects are more responsive to environmental conditions, as seems likely.

Overall, the genetic contribution to a complex phenotype (P) can be viewed very schematically as the result of some mean or scaling value (μ) plus contributions from various risk factors: $P = \mu + Q + PG + SN$, where Q represents genetic variants with mappably strong effect (QTL), PG from polygenes, and statistical error or "noise" (SN) due to uncertainties of various sorts including measurement error (some models do not include a mean term). For more details on mapping strategies and methods (*see* Article 9, **Genome mapping overview**, Volume 3).

Another reason harmful alleles are generally rare is that they usually will be eliminated from the population by natural selection, so that harmful alleles in the population at any given time are typically on their way out. Alleles with major effect that are at more than very low frequency may have (or have had in the past) some balancing benefit that keeps them in the population. It is a matter of interest to understand what the selective factors may be, although selection cannot be relied upon to have simplified a complex trait. As examples routinely show, a variety of loci and alleles can respond to any given selective factor. The classic example is

the many alleles, differing in different parts of the world at the loci coding for the proteins in hemoglobin, that are associated with resistance to malaria. Malaria is a serious disease that will favor any alleles present in a given local area that provide resistance. Malaria is a strong selective agent, if an allele's effect is also strong. But alleles with small effect, or with effect only in association with others or with environmental factors – likely attributes of complex disease alleles – are less susceptible to selection. In particular, if a disease arises only at late ages, causal genotypes are likely beyond the reach of natural selection, which works only through differential reproduction.

For this reason, genes that contribute to complex diseases of late (postreproductive) onset probably have not been targets of natural selection (at least not in relation to the modern chronic disease in question), so their frequency distribution will generally be as described earlier, that is, the loci will have a small number of common, and many rare, alleles. At present, there are only a few reliable examples of common variants with substantial effect on common disease. This issue of how many we should expect to find is contentious, however, for two reasons. First, common alleles with small effect on risk are very difficult to find by the usual genetic mapping study designs. For example, a common low-penetrance allele will be present in many unaffected people, requiring large case-control studies to detect its excess presence in affecteds. Some case-only study designs have been developed that can be used under some circumstances to circumvent this problem, which is exacerbated even further when there is causal heterogeneity, as typifies complex traits. Secondly, small effects can be important to discover because, while perhaps only of minor effect on the risk to any individual carrier, if they are common enough they can have noticeable effect on the rates of disease in the population (*see* Article 59, **The common disease common variant concept**, Volume 2).

Genetic complexity can arise in another way that is invisible to natural selection – and to linkage or association mapping studies. During an organism's life, every mitosis produces somatic daughter cells that can carry *somatically* arising mutations that are then transmitted to those cells' somatic descendants. Individuals are somatic mosaics, and if enough descendant cells are affected, somatic mutation can lead to disease, but because the mutations are not in the germ line, such causation will be invisible to mapping studies that are based on parent–offspring transmission. Cancer is one major example of disease affected or caused by somatic mutation, but there may be many others.

Overall, the nature of genetic effects on complex traits can be summarized by an hourglass metaphor of causation (Figure 1b). Many genotypes in a population constitute a wide base of ultimate risk factors that, often in interaction with environmental factors, converge to the same proximate risk factor, which in turn may have diverging phenotype effects. For example, many diverse genotypes contribute to high circulating cholesterol or triglyceride levels, but high cholesterol levels can have divergent disease effects, such as hypertension, obesity, heart attack, stroke, and others. One can focus therapeutic measures on the narrow waist of the hourglass – elevated cholesterol – without concern for the *specific* genotype involved in each case. Thus, statin drugs inhibit cholesterol synthesis, helping control circulating cholesterol levels and reducing their pathologic consequences, but with no effect on the physiologic pathway that caused the elevated cholesterol.

3. Complex traits are due to interacting factors

The model as discussed so far only considers the genetic aspects of complex traits, but few traits, including few human diseases, are due to genetic factors alone – cholesterol levels, of course, are also affected by diet, and can be influenced by dietary changes – and the genetic factors themselves may not contribute additively, but interactively. An additive contribution by an allele is the same regardless of other alleles in the person's genotype. When there is *epistasis* (gene–gene interaction), the effect of an allele at one locus depends on the genotype at one or more other variants in the person's genome (at the same or some other locus). Epistatic interactions are represented symbolically by a product term, conventionally denoted *GxG*. The impact of epistasis on disease risk is debatable; most studies indicate that the majority of effects are additive, but for technical reasons, there has been much less effort made to date to search for epistatic effects, and other statistical approaches may be used to find them when they exist. The rules of genetic inheritance are basically clear, but there are no a priori rules on genetic *effects*.

Perhaps far more importantly, it seems clear that the prevalence of most complex traits is heavily affected by nongenetic factors usually referred to collectively as *environmental*. These can include the various effects of lifestyle, such as diet and physical activity levels, or exposure to toxins or radiation, or many other factors. The problematic success record of environmental epidemiology suggests that environmental risk factors with small or interactive effect are unknown or at least the details of their effects are not understood. Thus, "diet" contributes to many diseases, but it is rarely clear how the specific components, subcomponents, combinations, or exposure histories actually affect risk. This uncertainty is reflected in the oscillations in public health recommendations in regard to many nutrients (e.g., unsaturated vs. saturated fats), therapeutic exposures (diagnostic mammograms), drugs (e.g., hormone replacement), and many others.

Like gene–gene relationships, environmental exposures can interact or contribute additively and there is no single rule. More importantly, environments and genotypes clearly interact in major ways. This phenomenon is symbolized in genetic models as *GxE.* The clearest evidence is the rapid recent rise in prevalence of many complex diseases, such as the examples described earlier, that seem to have an elusive genetic component. With many environmental agents, many possible genotypes, and many exposure regimes, it is very difficult to evaluate most *GxE* interactions or to replicate findings among studies, even if environmental factors could be measured with sufficient accuracy, which is often not the case.

In the end, we can summarize the logic of complex causation as

$$P = \mu + G + PG + E + GxG + GxE + ExE + SN \tag{1}$$

Each single term symbolically represents what may be an aggregate of many individual terms (for example, *G* may represent additive effects of genotypes at several loci, and *E* may represent several measured environmental factors) that, when known, appear as individual terms in a real application. Risk of each variable is quantified by some form of *regression*, or dose-response parameter, for example,

$\beta x Smoking$, where *Smoking* is measured in cigarettes consumed per day and β is the effect on the phenotype per cigarette per day.

An important thing about such models is that they are *statistical* and based on the existing variation in the population. An epidemiological study is necessarily a *sample* of that variation, rather than a perfect representation – and the population may never exactly replicate its own genetic or environmental exposure history. The models are *parametric* in that, except for the modification of each person's risk by the other measured factors in the risk equation, the same dose effect, β, is assumed to apply no matter who is doing the smoking or how many cigarettes are smoked. This is an assumption about the nature of the causal world that is analytically tractable, but except for strong effects, not easy to confirm as an accurate representation. For example, an average dose effect may be the combination of a few people at very high risk from that dose and other people who are immune to it.

4. Why so many inconclusive findings?

The identification of genes associated with disease has largely been based on the point-causation worldview made possible by the discovery of Mendelian segregation of particulate risk factors: a person inherits a genotype as a point-source exposure, that is, at conception. Over the past half-century this model worked with increasing ease to find genes responsible for hundreds of diseases, best-known examples being cystic fibrosis, sickle-cell hemoglobin, phenylketonuria, and Huntington's disease. Although most cases of these diseases are due to effects of variation at a single locus, the more intensely studied even these diseases were, the more complex we learn that they are (see Table 1), involving many different mutations and other factors (Scriver and Waters, 1999; Weiss and Buchanan, 2003).

Epidemiology has a similar history of success with infectious disease and disease caused by environmental exposures, such as smoking or asbestos. The effect of the risk factor is strong. But, as with Mendelian reasoning in genetics, epidemiological reasoning that works so well for diseases with strong main-effect causation does not work as well when such causation does not pertain. Although there have been major advances in the detection of important genetic mutations for complex diseases, it generally has not yet been possible to break these traits down into enough reliable genetic causal elements to generate genetically-based interventions on a public health scale. Most cases remain unexplained. Similarly, it has generally not been possible to identify major environmental risk factors for chronic disease that would make behavioral or policy interventions an effective way to prevent these diseases. It is known that diet and exercise can prevent some heart disease or diabetes, but not always, and this is primarily because obesity and inactivity are intermediate factors in the pathway to disease. And although there is plenty of suggestive evidence, much of it contradictory, we do not yet know which environmental factors trigger most of these diseases, so that prevention is not yet an option.

The usual finding for complex human diseases is that most cases arise sporadically (without specific known cause), but that a few families segregate the disease in a monogenic Mendelian way. Genes associated with asthma, diabetes,

Table 1 Numbers of alleles at selected disease-related genes

Gene (trait)	Mutations
CFTR (cystic fibrosis)	1011
LDL receptor (heart disease)	793
BRCA1 (breast, ovarian cancer)	603
DMD (muscular dystrophy)	597
PAH (phenylketonuria)	447
BRCA2 (breast, ovarian cancer)	388
TP53 (colon, other cancers)	264[a]
Pax6 (eye problems)	189
GBA (Gaucher's disease)	187
TSD (Tay-Sachs disease)	105
Pax3 (hearing, pigmentation)	47

Source: TP53 data from IARC database, **http://www.iarc.fr/p53/**.
[a]264 germ-line variants; nearly 20 000 somatic mutations in TP53 have been found in tumors. Human Genome Mutation Database 2004 (Stenson *et al*., 2003); see **http://www.hgmd.org/**.

Alzheimer's disease, some cancers, autism, epilepsy and bipolar disorder, among others, have been identified but typically the findings are restricted to a small number of families or isolated populations; there is often evidence that risk can be attenuated or exacerbated by environmental exposures, and identified genes generally explain only a small fraction of cases.

Two of the primary criteria used to test hypotheses in science are *induction* and *falsification*. Epidemiology and human genetics rely on the inductive approach of *replication* to confirm hypotheses, viewing failure to replicate as a form of falsification. However, it is not easy to apply these criteria in practice, because contradictory results are so common in the study of complex diseases, and difficult to interpret. When two studies disagree, for example as to whether a given nutrient is a risk factor for a type of cancer, it may indicate that one study is wrong, with the possible reasons ranging from chance sampling effects to poor study design, poorly chosen controls, confounding, bias, artifact or inadequate sample size or technology, or measurement errors. On the other hand, the discrepant results might accurately reflect true differences between the samples. There is no clear, single explanation.

One tactical response has been to study a complex trait in a small population isolate, where causation may be simpler because bottlenecks in ancestry may reduce genetic heterogeneity. One risk is that a causative allele found in such an isolate may be rare or absent in other populations. Another response is to greatly augment sample size by combining results of many studies in a meta-analysis (Ioannidis *et al*., 2001; Lohmueller *et al*., 2003; Munafo and Flint, 2004). The typical result is an increase in heterogeneity and the diminution or disappearance of evidence of the effect of a given genetic factor. One reason for this is that genome-wide mapping studies that initially report an effect have inherent bias in terms of sampling high-risk families or individuals, the simultaneous testing of hundreds of independent markers, and other artifacts (Goring *et al*., 2001; Terwilliger and Weiss, 2003).

The risk for some complex diseases may eventually be largely accountable in terms of specific genetic or environmental factors, while others will remain poorly understood and unpredictable, particularly if the ultimate goal is individual risk prediction. Complex diseases are caused by many risk factors, and the nongenetic ones may have strong secular trends (as most risk factors for common diseases have had over the past half-century). If individual factors have individually small effect, and causation is heterogeneous, and if risk is a probability assessed on population data yet every individual is unique in terms of his or her genome and lifetime environmental exposures, the task of predicting risk on a case by case basis becomes manifestly more complex, if not close to impossible.

5. Current issues and prospects

Causal complexity seems to be well understood in principle, but each case presents an epistemological quandary when it does not easily parse into replicable causal components. If the reality of complexity is acknowledged, and, for example, each family or population isolate has a unique pathway to disease, it becomes difficult to know what biomedical approaches would be effective for preventing disease, and the goal of identifying individual risk with rigor or accuracy may be unachievable, or too costly for the gain in knowledge.

It is very expensive in terms of the number of genetic markers, sample sizes, meta-analyses and the like even to *detect* genetic factors and probably even more difficult and costly to identify environmental factors or interactions with small effect on risk.

An inherited genotype is present from birth, but exposure to environmental effects that may interact with that genotype is often gradual. Therapeutic or public health decisions are based on prospective risk, to the *next* generation, but based on studies of the effect of *past* exposures. Yet it is very clear from the past century of epidemiological research that major, relevant environmental changes are always occurring and the *future* regimes of exposure are largely unpredictable. Not only does this cause problems in practice but it also clearly shows the elusive nature of the basic concepts of causality in this field.

The evidence of global prevalence distributions and secular trends shows that for many complex diseases, such as diabetes, cancer, cardiovascular disease, obesity, and others, trends to “westernization”' of lifestyle have similar general effects worldwide. This does not mean that genetic variation is uninvolved, but from a public health point of view, focus on the physiological funnel points is likely to have greater effect than focus on individual genetic causes. Genotypes that lead to early onset or that are less sensitive to the environment will be suitable targets for intervention, but will generally be much rarer.

The problems associated with decomposing complex traits into their genetic and other causes are challenging, but this does *not* always reflect a lack of scientific understanding of such traits. Indeed, the general role of genes in complex traits has been known for almost a century. Advances in molecular genetic technology over the past 20 years have made it possible to document the general validity of that theory, and the results of thousands of studies of the genetics of complex traits

are also consistent with what we know about the way evolution works (Weiss and Buchanan, 2003). This is a triumph of modern science, even if it does not provide easy answers. Complex traits are what they seem: complex.

Further reading

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl 3), 228–237.

Lander E and Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

Lander ES and Schork NJ (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Tabor HK, Risch NJ and Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews. Genetics*, **3**, 391–397.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, **18**, 19–24.

Zondervan KT and Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nature Reviews. Genetics*, **5**, 89–100.

References

Eichers ER, Lewis RA, Katsanis N and Lupski JR (2004) Triallelic inheritance: a bridge between Mendelian and multifactorial traits. *Annals of Medicine*, **36**, 262–272.

Gabriel SB, Salomon R, Pelet A, Angrist M, Amiel J, Fornage M, Attie-Bitach T, Olson JM, Hofstra R, Buys C, *et al*. (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nature Genetics*, **31**, 89–93.

Goring HH, Terwilliger JD and Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, **69**, 1357–1369.

Ioannidis JP, Ntzani EE, Trikalinos TA and Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nature Genetics*, **29**, 306–309.

Lohmueller KE, Pearce CL, Pike M, Lander ES and Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**, 177–182.

Munafo MR and Flint J (2004) Meta-analysis of genetic association studies. *Trends in Genetics*, **20**, 439–444.

Neroth P (2004) Fat of the land. *Lancet*, **364**, 651–653.

Scriver CR and Waters PJ (1999) Monogenic traits are not simple: lessons from phenylketonuria. *Trends in Genetics*, **15**, 267–272.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M and Cooper DN (2003) Human gene mutation database (HGMD): 2003 update. *Human Mutation*, **21**, 577–581.

Stern MP, Knapp JA, Hazuda HP, Haffner SM, Patterson JK and Mitchell BD (1991) Genetic and environmental determinants of type II diabetes in Mexican Americans. Is there a "descending limb" to the modernization/diabetes relationship? *Diabetes Care*, **14**, 649–654.

Terwilliger JD and Weiss KM (2003) Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Annals of Medicine*, **35**, 532–544.

Trowell HC and Burkitt DP (1981) *Western Diseases, their Emergence and Prevention*, Harvard University Press: Cambridge.

Veenstra-VanderWeele J and Cook EH Jr (2004) Molecular genetics of autism spectrum disorder. *Molecular Psychiatry*, **9**, 819–832.

Weiss KM and Buchanan AV (2003) Evolution by phenotype: a biomedical perspective. *Perspectives in Biology and Medicine*, **46**, 159–182.

Wills-Karp M and Ewart SL (2004) Time to draw breath: asthma-susceptibility genes are identified. *Nature Reviews. Genetics*, **5**, 376–387.

Specialist Review

The common disease common variant concept

Deborah S. Cunninghame Graham and Timothy J. Vyse
Imperial College, London, UK

1. Introduction

Unraveling the genetic basis of human diseases represents a major challenge in human genetics. This is especially the case for complex diseases, which comprise the bulk of the disease burden in industrialized societies. Complex genetic disease traits arise as a consequence of genetic and environmental contributions to disease susceptibility (*see* Article 58, **Concept of complex trait genetics**, Volume 2). The genetic component is split across many loci, each contributing a small effect to the overall susceptibility. Thus, the identification of the causal genetic variants in these diseases presents a major challenge to human geneticists. The problems of achieving this goal are further magnified by the complexities that result from the gene–gene and gene–environment interactions.

Linkage mapping has been used to great effect with simple Mendelian disease traits, such as cystic fibrosis (Pritchard and Cox, 2002) where genetic variants in a single gene increase the risk of disease dramatically. The nature of complex traits is such that the identification of disease susceptibility genes has relied heavily on association methods. Whether of case–control design, or family-based design, the successful outcome from association studies is dependent on the presence of linkage disequilibrium (LD) in the genome. When a novel variant first appears, it will create a new haplotype, the constituent markers of which will be coinherited for a number of generations. During this stage, it is possible to predict the identity of the other alleles by identifying the allele at one marker on the haplotype. It is this genetic predictability that is exploited in LD mapping, before recombination events occur, which will, over time, cause the overall pattern of LD in this new haplotype to start to break down, thereby reducing the predictable nature of the pattern of inheritance. The pattern of LD across the genome, therefore, has a major impact on the interpretation and design of association studies. The efficacy of an association study to identify a given disease-susceptibility locus is also influenced by the number of alleles at that locus that contribute to disease risk. How many susceptibility alleles there are at any given disease locus depends on their allele frequencies and the allelic heterogeneity. The number of different variants present

within a locus that predispose to a disease phenotype may be referred to as its allelic diversity or the allelic spectrum. The term genetic architecture has been used to refer to the range of allelic diversity in the genome that will contribute to a phenotype. Obviously, the genetic architecture will have an impact on the statistical power of mapping studies to detect positive associations. In single-gene disorders, the allelic spectrum can be wide, but the penetrance of the alleles is maintained. In polygenic disease traits, the complexity of the allelic spectra and the genetic architecture will influence the feasibility of identifying disease susceptibility alleles. A large number of loci, each with complex allelic spectra would severely compromise the probability of finding susceptibility alleles in polygenic disorders. Success would require both larger study cohorts and very large numbers of polymorphisms to be typed (Wang and Pike, 2004). However, it can be argued from a theoretical standpoint that more common allelic variants may predispose to complex polygenic diseases: the common disease–common variant hypothesis. Preliminary data from association studies in complex traits suggest that commoner variants do make some contribution to disease. The arguments behind the common disease–common hypothesis will be described and some examples provided. We will also discuss alternative models and suggest that the true nature of the architecture of genetic disease will reflect more than one single concept.

2. The concept of genetic architecture

When considering the penetrance and number of alleles that predispose to disease, it is useful to consider the influence that quantitative variation in these parameters will exert on the ability to find and identify causal disease polymorphisms (*see* Article 10, **Measuring variation in natural populations: a primer**, Volume 1). Genetic disease can be divided into monogenic disorders and polygenic or complex disorders. Monogenic diseases are usually rarer and result from the inheritance of highly penetrant genetic variants at one or a very limited number of loci. More than 1500 monogenic disease genes have been identified, and it is well established that the majority of such diseases exhibit marked allelic diversity (On-line Mendelian Inheritance in Man **http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM**). The degree of allelic heterogeneity in any population will reflect a balance between the processes of spontaneous mutation and selection against the inheritance and propagation of any allele. The rate of spontaneous mutation in the genome is highly variable: estimates vary from 10^{-4} to 10^{-7} per locus per generation (Bellus *et al*., 1995; Crow, 2000; Eyre-Walker and Keightley, 1999). There is no compelling evidence, however, that simple Mendelian disorders arise in genes that exhibit particularly high-mutation rates. However, in comparison with complex genetic conditions, monogenic diseases are more likely to exert some influence on survival, and hence, selection is more likely to operate in a population in relation to monogenic disease when compared with complex genetic disease. How might the allelic diversity in monogenic disease be explained? It is imperative in the consideration of such issues to incorporate the fact of the rapid expansion of the human population over the preceding millennia. The exact number of millennia is open to debate; best estimates suggest that this population explosion has occurred over the preceding

700–6000 generations; that is, in the range of 18 to 150 millennia. This impressive demonstration of fecundity has had a major impact on the genetic architecture of the modern human population. The dramatic increase in population size would be expected to increase the numerical representation of disease alleles, even rarer ones, in modern humans. However, two factors will serve to reduce this effect and lead to marked allelic diversity in monogenic disease. First, as new mutations are generated they are less likely to arise on a preexisting disease-associated haplotype, if that haplotype itself is uncommon. Thus, for rarer disease alleles, the accrual of mutations over generations is likely to simply expand the number of disease susceptibility alleles. Second, the presence of selection will serve to reduce the frequency of causal alleles over time.

Not all monogenic diseases demonstrate marked allelic diversity. Several factors may explain this situation. Most simply is the case of heterozygous advantage. In this scenario, alleles whose pathologic effect is biased toward recessivity can confer a selective advantage when present as a single copy. There are several well-characterized examples of polymorphisms that exhibit some degree of protection from malaria, for example, G6PD deficiency at the *G6PD* locus, and HbC and HbS at the *HBB* locus in West African populations; the estimated allele frequencies being 0.20, 0.09, and 0.10, respectively. Selection may operate during the phase of population expansion, and it may increase the frequency of certain alleles in the population in the preexpansion phase. For example, the mutations that give rise to cystic fibrosis (CF) do not immediately appear to comply with the arguments set out above. More than 900 alleles of the CF transmembrane conductance (*CFTR*) gene have been associated with CF, however, approximately 70% of cases are due to one single deletion, ΔF508. It has been postulated that heterozygotes carrying *CFTR* mutations may have some selective resistance to *Salmonella typhi* (Pier *et al*., 1998). Irrespective of the mechanism, it can be shown that for an allele as frequent as ΔF508 in the preexpanded population this simple ancestral spectrum will persist following expansion with a half-life of 39 000 to 390 000 years, depending on the assumed mutation rate (Reich and Lander, 2001).

Do the effects of recent population expansion differ in complex traits when compared with monogenic disorders? If one assumes that the frequency of the disease susceptibility alleles in more common complex genetic traits is higher than in rare monogenic disease, then essentially the answer is yes. And not only is the answer in the affirmative, but it was shown by Reich and Lander that this situation will remain so for tens if not hundreds of thousands of years, depending on the frequency of the disease allele. Unsurprisingly, more common alleles persist at a similar frequency in an expanded population for longer. The implications of this model have given rise to the common disease–common variant hypothesis.

3. Current hypotheses to explain genetic architecture

3.1. Common Disease Common Variant (CDCV) hypothesis

The CDCV, or interactive, model of genetic architecture was first proposed in 2001 (Reich and Lander, 2001). It states that complex diseases are caused by the

interaction of common alleles at a small group of susceptibility loci (Smith and Lusis, 2002; *see also* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2). These common alleles are not population specific, but are present at >1% minor allele frequency in multiple populations.

By taking into account the rapid, recent expansion of the human population, as described above, it is apparent that the allelic spectrum at all loci will gradually increase with time. However, the frequency of the allele in the preexpansion phase is critical to the rate at which allelic diversity is generated. Mathematically, the main influences on the allelic spectrum, in the absence of selection, can be summarized as a balance between the forward mutation rate, $4N_e u_s$, and the reverse mutation rate, $4N_e u_n$, where N_e is the effective population size (usually estimated to be 10 000) and u_s is the mutation rate per generation per locus, the n subscript denoting the reverse mutation rate. That is, in any stable population, where the genetic variants are in equilibrium, there is a balance between the rate at which new mutations arise, which for most genes is at a rate of approximately $1 \times 10^{-5} - 1 \times 10^{-6}$ per gene per generation (Bellus *et al*., 1995; Crow, 2000; Eyre-Walker and Keightley, 1999; Peltomaki, 2001; Sankaranarayanan, 1998) and their loss by negative selection. However, in a rapidly expanded population, rare alleles in the original founder pool are reduced in frequency as the result of strong selective pressure, and will become swamped by the generation of new alleles (Reich and Lander, 2001). By contrast, common alleles, which have a large population reservoir, and are under little selective pressure, have a lower turnover, so will take longer to be diluted out (diversified) by the presence of these new mutations (Smith and Lusis, 2002). New disease-susceptibility mutations might arise but they would still be carried on the ancestral haplotype. The higher the initial allele frequency, the longer this will take (Wright and Hastie, 2001). Allele diversification will continue in this way, until the alleles reach equilibrium in the expanded population. Reich and Lander state that the diversification of the alleles in the expanded human population is in an intermediate stage, so that these selectively neutral common alleles that were important markers of disease in the preexpanded population are predicted to have retained a low level of allelic diversity in the postexpansion human population. Therefore, they can still be useful markers of common disease in the modern human population (Smith and Lusis, 2002). These concepts are illustrated in Figure 1.

Another important factor in formulating the CDCV hypothesis is the strength of the selection pressure on allele diversification. For alleles to attain a high-equilibrium frequency in a population, that is, retain low diversity and be useful markers in LD mapping of complex diseases, they must be selectively neutral or be under low-selective pressure. This leads to a lower turnover in the population and decreases the effect of new mutations. However, selectively neutral alleles are likely to make a small contribution to the overall disease risk, giving weak associations to disease, which will be another factor in the requirement for large samples size. This situation might be the case for alleles implicated in disorders such as for late onset diseases such as type 2 diabetes (T2D) or hypertension (Wright and Hastie, 2001). Furthermore, it is necessary to take into account stratification issues when interpreting association study results. An allele may be at a high frequency in a subpopulation, due to urbanization or geographical isolation, so a positive or negative association may not be applicable to the population as a whole. The

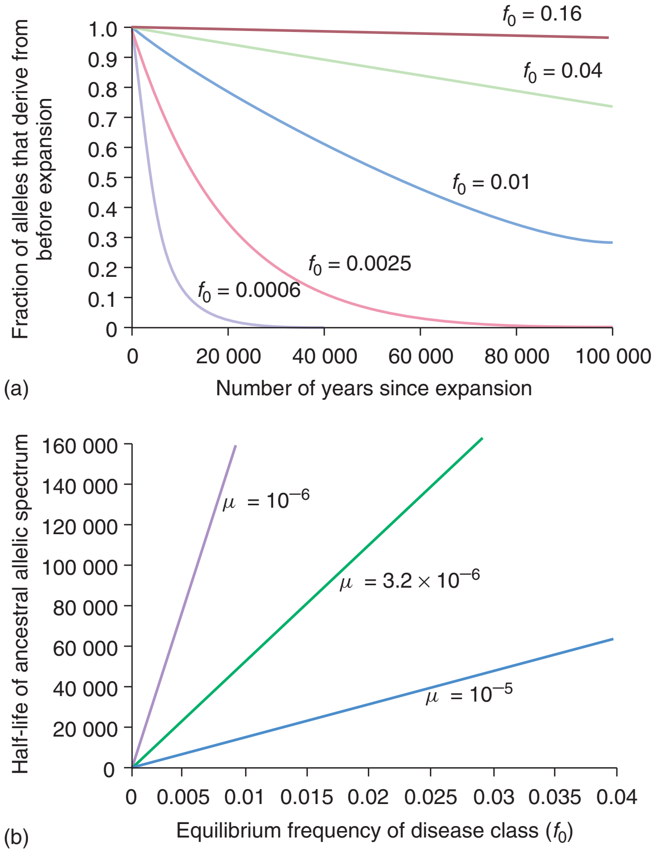


Figure 1 This Figure is taken from the paper by Riech and Lander. It illustrates the change over time accompanying the human population expansion of the frequency of the disease-susceptibility alleles, f_0. How the fraction of the total proportion of disease-susceptibility alleles changes with time is shown in (a). The curves are calculated assuming a population increase from 10^5 to 6×10^9 and a fixed mutation rate (μ) of 3.2×10^{-6} per generation. As is shown, more frequent classes of disease alleles will have persisted over the preceding 100 000 years and shown only a slight reduction in frequency. The same point is illustrated in (b). In this graph, the time taken for the original disease allele class frequency to fall by half is shown. Thus, for a disease allele frequency of 0.005, the frequency will have fallen to 0.0025 after approximately 20 000 years. It can also be seen that at higher-mutation rates (μ), the half-life is substantially shorter. The probability that the two alleles randomly picked both belong to the disease-susceptibility class is denoted by ϕ_{disease}. Hence, the reciprocal of this term is an index of the allelic spectrum of any disease. In (c), an arbitrary threshold for simpler disease spectra is shown by a cutoff of $1/\phi_{\text{disease}} = 10$. The simpler disease spectra are maintained for at least 100 000 years with disease allele frequencies between 0.01 and 0.05 (Reprinted from *Trends in Genetics*, vol. 17, Reich DE, *et al*, On the allelic spectrum of human disease, pp. 502–510, Copyright 2001, with permission from Elsevier)

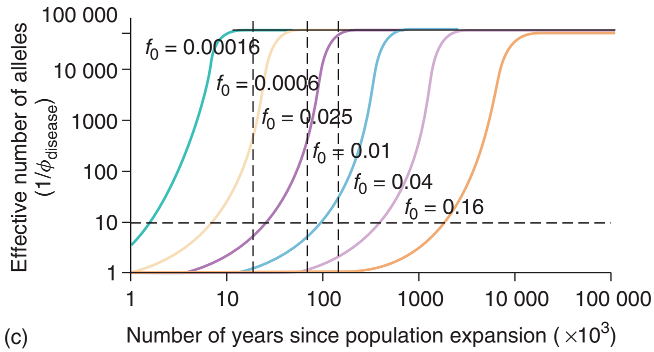


Figure 1 (*continued*)

influence of the rapid expansion of human populations may also be seen in relatively isolated populations. In such circumstances, there may be a marked influence of the chromosomal composition of the founders of that population. Indeed, as can be seen in Figure 1, if this "founding" effect occurred relatively recently, then even low-frequency disease-susceptibility alleles may be disproportionately represented in the contemporary population.

The CDCV model predicts that there is an increased risk for common diseases from common variants of low individual risk, across multiple populations. Three examples of common alleles supporting this hypothesis are: a 32-bp deletion in the coding region of the *CCR5* gene, which decreases the transmission of HIV-1, has an allele frequency of at least 9% in Caucasians (Wright and Hastie, 2001), and is common in a number of other populations (Lu *et al*., 1999); the E4 allele of Apolipoprotein E (*APOE*), which has a prevalence of 15% in Caucasians, over 20% in Africa and Scandinavia, and 6–12% in Japan and China (Ritchie and Dupuy, 1999) has been associated with Alzheimer's disease (Saunders *et al*., 1993) and coronary artery disease. A common variant, P12A at the *PPAR*γ locus is associated with type 2 diabetes (Altshuler *et al*., 2000). The causal allele is at high frequency in the population (0.85) and as expected the risk associated with it is small. The genetic basis of complex disease is insufficiently established, however, to provide full empirical support to the CDCV hypothesis. It should also be appreciated that association studies have a bias toward identifying more frequent variants.

3.2. *Multilocus-multiallele hypothesis*

However, the CDCV hypothesis does not provide the complete answer to the nature of the genetic architecture of human populations. The CDCV hypothesis states that the common diseases are due to common alleles, but the multilocus-multiallele (MLMA) or genetic heterogeneity hypothesis describes complex diseases as having a low probability of carrying a given susceptibility allele, given the fact that these diseases have a contribution from multiple susceptibility alleles and environmental influences. This gives a low chance, or low detectance, of detecting a single deleterious allele during an association study (Weiss and Terwilliger, 2000).

Furthermore, the high frequency of some complex diseases could be due to the prevalence of environmental triggers that form part of the natural aging process. For some complex diseases, for example hypertension, there is a high-level contribution of environmental/natural aging factors implicated in the disease process, since the heritability decreases with increasing age of onset, due to an increasing contribution from the aging process and other environmental factors.

One more factor supporting the MLMA hypothesis is a predictor that low-frequency variants have a role to play in the genetics of a disease, rather than the high-frequency variants proposed by the CDCV hypothesis. This predictor arises from the inverse relationship between allele frequency and genetic effect (Morton, 1996; Wright, 1968) and proposes that the lower the allele frequency, the higher the genetic effect. Evidence from breast cancer and familial hypercholesterolaemia (FH) are two examples that support the idea of rare alleles being important risk factors for disease. In the *BRCA2* gene, there are greater than 404 rare alleles with increased risk of disease, but only 1 out of 6 common alleles have any effect. Likewise, in FH, there is an increased risk of premature coronary artery disease attributed to more than 435 rare alleles in LDL receptor, but no known common alleles (Wright and Hastie, 2001).

These examples also serve to raise the question over what proportion of weakly associated, selectively neutral common alleles actually have a role to play in the genetic aetiology of complex diseases. Since it has been proposed that neutral alleles either tend to be lost to a population or be fixed in that population (Pritchard and Przeworski, 2001) and common alleles are thought to attain their high-allele frequency by virtue of being selectively neutral, they are unlikely to have a significant effect on risk of disease (Pritchard and Przeworski, 2001).

3.3. Selection in complex traits

As may be predicted by the CDCV hypothesis, the E4 allele of *APOE* appears to be the ancestral allele (Fullerton *et al*., 2000) and is not uncommon the population. The situations at *PPAR*γ locus and in the VNTR alleles at the *INS* (Insulin) locus are analogous. Why should disease-susceptibility alleles be so prevalent in the population? The model of recent population expansion predicts their high frequency in the ancestral population. It could be argued that their individual disease risk is small, often complex disease traits pertain to diseases of adult onset that may impact less on selection. It is also possible that there are some selective advantages to alleles that predispose to complex traits (Fullerton *et al*., 2000). It was suggested more than 40 years ago that diabetes mellitus might be so common in the modern world as a consequence of the selection of alleles conferring metabolic efficiency on the bearer (Neel, 1962). Similarly, it can be argued that autoimmunity may arise as a consequence of the selective pressures imposed by infection.

3.4. Common variant/multiple disease hypothesis

The CDCV hypothesis also lends itself to extension, raising the possibility that common disease variants may predispose to multiple diseases. That is, if a restricted

number of common genetic variants predispose to a disease state, would this mechanism be disease specific? The overlap between many complex disease traits is well established at an epidemiological level, although often less well understood pathologically. It would not be surprising to learn that coronary artery disease (*see* Article 63, **Hypertension genetics: under pressure**, Volume 2), obesity, and type 2 diabetes shared some common aetiology. It is apparent that multiple autoimmune diseases appear to be more prevalent in some families (Broadley *et al*., 2000). Indeed, there is genetic evidence to support the concept of shared genetic risk factors in different, but related disease states. For example, it has been shown in autoimmune diseases that the loci mapped in genome-wide linkage studies tend to overlap more than would be anticipated by chance. This observation has also been made in animal models of autoimmunity (Becker, 2004). Similar observations have been made in schizophrenia and in bipolar disorder, and in type 2 diabetes and obesity. Linkage analysis defines large genomic regions that do not necessarily harbor identical disease alleles. The lack of well-substantiated disease associations in complex traits does not allow clear support of commonality of genetic factors. However, there are some tantalizing clues. The E4 allele of *APOE* has been cited in support of the CDCV hypothesis, above, the association with Alzheimer's disease and coronary artery disease suggest the possibility of shared genetic factors, a conclusion strengthened by the association of E4 with other related disease states (Smith, 2000). Similarly, *CTLA4* polymorphisms have been reported to be associated with multiple autoimmune diseases including type 1 diabetes, multiple sclerosis, and autoimmune thyroid disease. The heterogeneity of the published data and the complexity of *CTLA4* haplotypes are such that it is unclear at present whether these *CTLA4* associations represent the influence of a common or restricted number of causal variants.

4. Conclusions

Much of the progress in LD mapping of complex diseases has been made using the major assumptions of the CDCV hypothesis, that is, that common alleles cause common diseases. However, the results are often inconclusive, with reports of negative or borderline associations, and positive associations, which are frequently population dependent. There are many reasons for these findings. There is probably an overall lack of power in the individual studies, so that association is biased toward the identification of more common, but weaker haplotypes. However, the failure to find widespread, reproducible association using LD mapping is not a reason to discard the CVDV model as a working hypothesis. The CDCV model does not have to be universally applicable to justify its use in LD mapping studies (Goldstein and Chikhi, 2002). However, the results of such studies need to be interpreted with caution. If negative results in an LD-based mapping study, that is, no common alleles are associated with disease, then it is worth going back to the data and looking for rare variants. For positive associations with common alleles, it will be necessary to first gain replication of the results in an independent sample set and then look for rarer SNPs, with potentially greater penetrance, which may have been overlooked in the initial study. It is likely that the true genetic architecture includes

a combination of both rare variants and common alleles, potentially in the same locus, making a genetic contribution to disease, since that is the pattern of allele frequencies across the rest of the genome (Wang and Pike, 2004). Furthermore, it seems unlikely that there are multiple common alleles interacting with one or more common environmental triggers to cause common diseases and that these disorders are not even more prevalent. One solution to this problem is that common diseases require an additional contribution from rarer, more penetrant alleles, with increased disease risk to "tip the balance". For the future, the full potential of the CDCV hypothesis and LD mapping is unlikely to be realized unless the contribution of rare alleles is assessed through fine mapping in large collaborative projects.

References

Altshuler D, Daly M and Kruglyak L (2000) Guilt by association. *Nature Genetics*, **26**(2), 135–137.

Becker KG (2004) The common variants/multiple disease hypothesis of common complex genetic disorders. *Medical Hypotheses*, **62**(2), 309–317.

Bellus GA, Hefferon TW, Ortiz de Luna RI, Hecht JT, Horton WA, Machado M, Kaitila I, McIntosh I and Francomano CA (1995) Achondroplasia is defined by recurrent G380R mutations of FGFR3. *American Journal of Human Genetics*, **56**(2), 368–373.

Broadley SA, Deans J, Sawcer SJ, Clayton D and Compston DAS (2000) Autoimmune disease in first-degree relatives of patients with multiple sclerosis: a UK survey. *Brain*, **123**(6), 1102–1111.

Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nature Reviews. Genetics*, **1**(1), 40–47.

Eyre-Walker A and Keightley PD (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**(6717), 344–347.

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, *et al*. (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *American Journal of Human Genetics*, **67**(4), 881–900.

Goldstein DB and Chikhi L (2002) Human migrations and population structure: what we know and why it matters. *Annual Review of Genomics and Human Genetics*, **3**, 129–152.

Lu Y, Nerurkar VR, Dashwood WM, Woodward CL, Ablan S, Shikuma CM, Grandinetti A, Chang H, Nguyen HT, Wu Z, *et al*. (1999) Genotype and allele frequency of a 32-base pair deletion mutation in the CCR5 gene in various ethnic groups: absence of mutation among Asians and Pacific Islanders. *International Journal of Infectious Diseases*, **3**(4), 186–191.

Morton NE (1996) Logarithm of odds (lods) for linkage in complex inheritance. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(8), 3471–3476.

Neel JV (1962) Diabetes Mellitus: a "Thrifty" genotype rendered detrimental by "Progress". *American Journal of Human Genetics*, **14**, 353–362.

Peltomaki P (2001) Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics*, **10**(7), 735–740.

Pier GB, Grout M, Zaidi T, Meluleni G, Mueschenborn SS, Banting G, Ratcliff R, Evans MJ and Colledge WH (1998) Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature*, **393**(6680), 79–82.

Pritchard JK and Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant ... or not? *Human Molecular Genetics*, **11**(20), 2417–2423.

Pritchard JK and Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**(1), 1–14.

Reich DE and Lander ES (2001) On the allelic spectrum of human disease. *Trends in Genetics*, **17**(9), 502–510.

Ritchie K and Dupuy AM (1999) The current status of apo E4 as a risk factor for Alzheimer's disease: an epidemiological perspective. *International Journal of Geriatric Psychiatry*, **14**(9), 695–700.

Sankaranarayanan K (1998) Ionizing radiation and genetic risks IX. Estimates of the frequencies of Mendelian diseases and spontaneous mutation rates in human populations: a 1998 perspective. *Mutation Research*, **411**(2), 129–178.

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, *et al*. (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, **43**(8), 1467–1472.

Smith JD (2000) Apolipoprotein E4: an allele associated with many diseases. *Annals of Medicine*, **32**(2), 118–127.

Smith DJ and Lusis AJ (2002) The allelic structure of common disease. *Human Molecular Genetics*, **11**(20), 2455–2461.

Wang WY and Pike N (2004) The allelic spectra of common diseases may resemble the allelic spectrum of the full genome. *Medical Hypotheses*, **63**(4), 748–751.

Weiss KM and Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nature Genetics*, **26**(2), 151–157.

Wright S (1968) *Evolution and the Genetics of Populations*. *Vol 1: Genetic and Biometric Foundations*, University of Chicago Press.

Wright AF and Hastie ND (2001) Complex genetic diseases: controversy over the Croesus code. *Genome Biology*, **2**(8), COMMENT2007.1–COMMENT2007.9.

Specialist Review

Population selection in complex disease gene mapping

Teppo Varilo and Leena Peltonen

National Public Health Institute and University of Helsinki, Biomedicum Helsinki, Helsinki, Finland

The Broad Institute at MIT and Harvard, Boston, MA, USA

1. Introduction

Perhaps one single most critical choice in any complex disease gene mapping efforts is the selection of the study population. Nonetheless, geneticists often tend to neglect the optimal study design, and too often turn to population samples offered to them more or less by chance. Following the path of the feasibility and least resistance has often translated to the use of case-control samples from heterogeneous, less-ideally characterized study populations. Thanks to the human genome project, the scientific community has now excellent tools for the identification of specific alleles associated with any disease process. High throughput genotyping of multiallelic Short Tandem Repeat (STR) or biallelic (SNP) markers can be used to define the genetic profiles of DNA samples with a high accuracy. However, the exponentially advancing technical capacity is wasted if blindly utilized for inadequately characterized study samples. The detailed information of the character and history of the target population has a major impact on the study design and on the interpretation of the obtained results. Differences in demographic history of the isolated populations demand divergent research strategies and also dictate the selection of markers and statistical methods to be used in initial genome scans and in the subsequent positional cloning aimed at the identification of allelic variants. A failure to take into account in association studies the extended-familial relationships, typical for isolated populations, can produce false–positive results, missing true associations while overestimating the level of significance (Newman *et al*., 2001). Most initial findings of predisposing genes of complex traits have been substantiated only in extremely well-designed study samples, often representing families or study samples of special populations with exceptionally high loading of a disease, exposing distinct genetic character of a trait (Hugot *et al*., 2001; Mira *et al*., 2004).

Population isolates offer an avenue for a cost-effective initial identification of complex disease loci, and if wisely utilized, have tremendous potential also for the final identification of specific disease alleles. The isolates simplify the genetic

background of polygenic diseases since they show a restricted number of involved genes and their alleles. In typical isolates, a founder effect, genetic drift, and isolation of varying period of time have molded the gene pool, resulting in the enrichment of some disease alleles. Some population isolates provide the research conditions comparable to those accomplished with inbred animal strains. In an ideal case, families gathered from population isolates show, in addition to a higher genetic homogeneity, also social cohesion and restricted amount of environmental and cultural diversity, not achieved in outbred populations. Further benefits of ideal population isolates include accurate genealogical records, standardized health care, and reliable medical information. All these features greatly facilitate the collection and assembly of powerful study samples, identification of cases with shared ancestors (shared genetic background), and systematic compilation of clinical data, including longitudinal quantitative information of various phenotypic features.

2. The concept of an isolate

Multiple bottlenecks and repeated periods of genetic drift characterize the history of human populations, and the diversity of the human genome is significantly less than intuitively anticipated. In early times, the human population expanded in size as it inhabited the wilderness. The waves of migration involved typically small groups, first out of Africa about 100 000 years ago (Cavalli-Sforza, 1998) and some 50 000 to 30 000 years ago, into new regions, such as the Americas and Australia. A third wave of migration and growth occurred about 10 000 years ago after the last glacial period with the spread of agriculture from the Middle East. New subpopulations isolated by distance were eventually established even in the most peripheral regions of the world during the written historical times. In addition, the extreme environmental changes such as famines, wars, and epidemics of infectious diseases, and even social factors such as religion or sources of livelihood have created recent, efficient bottlenecks, restricting the regional allelic diversity. The subsequent population expansion of some global isolates is often accompanied by relative inbreeding lasting for tens of generations, causing a reduction in the effective population size and further reducing allelic diversity.

Unfortunately, for most global isolates, we lack reliable information on their initial genetic makeup, their total number of founders, and the extent and duration of their isolation. Only a handful of isolates such as Quebec, Iceland, Northern Sweden, or Finland, and some special religious populations have easily accessible and reliable genealogical records, which date back to the establishment of the population, offering a helpful tool for complex disease gene hunt based on the assumed sharing of disease alleles among affecteds. We can roughly infer the extent of a population's isolation by its trademark monogenic diseases. Younger isolates tend to have distinct spectrum of such diseases like exemplified by the Finnish disease heritage (**www.findis.org**). As expected, the genealogy of older isolates is more complicated and the enrichment of specific diseases less remarkable.

A very small number of founders for a population provide a significant advantage for disease gene mapping efforts (Kruglyak, 1999; Wright *et al*., 1999; Johnson and Todd, 2000). Linkage disequilibrium (LD), a strong association between

markers on the alleles originating from the same historical ancestral chromosome, is detectable over wide genetic intervals in young isolates. Using multiallelic markers, the disease alleles of such isolates can expose LD intervals reaching 10 Mb, whereas in older isolates, the extent of LD intervals is not different from more heterogeneous populations, reaching just a few kilobases. For rare monogenic diseases, the standard genome-wide scans with some 400 multiallelic markers have typically exposed a significant linkage disequilibrium in disease alleles of isolated populations. This has facilitated an efficient identification of disease loci in very small study samples (Houwen *et al*., 1994; Nikali *et al*., 1995). The subsequent haplotype sharing strategy with a denser set of markers has provided powerful restriction of the critical chromosomal region to manageable intervals for sequencing. The same strategy has also worked for some common traits in genetic isolates. In lactose intolerance, a single nucleotide polymorphism (SNP), a C to T change residing some 14 kilobases upstream of the lactase (LCT) gene on 2q21-22, was found to exhibit a complete association with lactase persistence (Enattah *et al*., 2002). This variant, interfering a cis-acting regulatory element, was uncovered in an exemplary set of Finnish families descending from an isolate where LD and haplotype analysis could be relied upon. The result has been subsequently confirmed in outbred populations, enabling a global diagnostic test. We, among others, have shown that in general (nondisease) alleles of young (10–20 generations old) subisolates, the size of LD intervals reaches up to 1 Mb, and this feature most probably offers an avenue for the initial locus positioning for complex traits in these populations.

So far, most findings of complex disease genes have been based on the linkage studies in rare families with multiple disease cases. These studies have yielded initial evidence for disease loci, but for most cases, verification of the locus finding in different populations and the final identification of the gene has remained a problem. As expected from the anticipated allelic and locus heterogeneity of these diseases, differentiating the genuine loci from the statistical ghosts has been very complicated even within the best-characterized study samples. However, some population isolates have disclosed encouraging allelic associations for complex diseases, often since they have facilitated efficient sample collection and careful phenotyping (see Table 1).

3. Considerable dissimilarity of isolates

Not all isolates are alike – even Finland has subpopulations emerging from various founder populations. Globally, the isolates vary in crucial aspects. The number of founders, the age of the population, the expansion rate, the historical bottlenecks (famines, wars), and amount of immigration, all affect the extent of the diversity in disease (or any) alleles of the given population. Furthermore, successful genetic studies of complex traits, with a multigenetic background, require a sufficient population size to offer adequate study samples and to expose the complete genetic background of any trait. However, even very small isolates originating from a handful of ancestral chromosomes might turn out to be extremely valuable for the initial identification of individual, important, but most probably family-specific,

Table 1 Some promising gene associations in complex diseases

Population	Age of population	Reported genome scans	Loci showing linkage	Gene identified	References
Finland, late settlement	330 years	Asthma and atopy	Chr 7p 14-15	*GPRA* (G protein-coupled receptor for asthma susceptibility)	Laitinen *et al*., 2004
Australia, UK	Mixed	Asthma and atopy	Chr 2q14	*DPP10* (dipeptidyl peptidase 10)	Allen *et al*., 2003
(1) Whole Finland	2000 years	Diabetes (type 2)	Chr 20q13	*HNF4A* (hepatocyte nuclear factor-4 alpha)	Silander *et al*., 2004
(2) Ashkenazi Jewish					Love-Gregory *et al*., 2004
Finland (Botnia) and Mexican Americans	2000 years	Diabetes (type 2)	Chr 2q27	*CAPN10* (Calpain-10)	Horikawa *et al*., 2000
Whole Finland?	2000 years	Diabetes (type 2)	Chr 3p25	*PPARG* (Peroxisome proliferator-activated receptor gamma)	Deeb *et al*., 1998
		Body mass index			
Oji-Cree Indians		Diabetes (type 2)	Chr 12q22-qter	*HNF-1A* (Hepatic nuclear factor-1alpha)	Triggs-Raine *et al*., 2002
Finland, early settlement	2000 years	Dyslexia	Chr 15q21	*DYX1C1*, gene of unknown function	Taipale *et al*., 2003
Whole Finland	2000 years	Familial combined hyperlipidemia	Chr 1q21	*USF1* (upstream transcription factor 1)	Pajukanta *et al*., 2004
Iceland	About 1000 years	Ischemic stroke	Chr 5q12	*PDE4D* (Phospho-diesterase 4D)	Gretarsdottir *et al*., 2003
Finland, late settlement	330 years	Schizophrenia	Chr 1q	*DISC1* (Disrupted in Schizophrenia 1)	Hennah *et al*., 2003
Iceland	About 1000 years	Schizophrenia	Chr 8p12-21	*NRG1* (Neuregulin-1)	Stefansson *et al*., 2002
US (of Northern European origin)	Mixed	Schizophrenia	Chr 14q32	*AKT1*	Emamian *et al*., 2004
Ireland	Mixed	Schizophrenia	Chr 6p22.3	*DTNBP1* (Dysbindin)	Straub *et al*., 2002
Japan		Rheumatoid arthritis	Chr 1p36	*PADI4* (peptidylarginine deiminase 4)	Suzuki, 2003
Mennonite	About 200 years	Hirschsprung disease	Chr 13q22	Endothelin B receptor	Puffenberger *et al*., 1994
Caucasians	Mixed	Crohn's disease	Chr 16q12	*NOD2*	Hugot *et al*., 2001 Ogura *et al*., 2001
Vietnam	Mixed	Leprosy	Chr 6p25	*PARK2*, *PACRG*	Mira *et al*., 2004

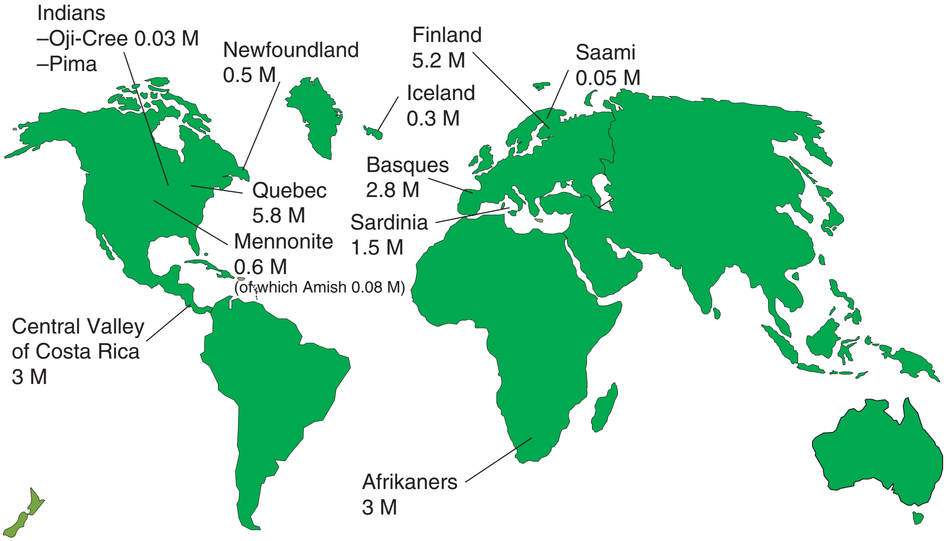


Figure 1 A global view to some interesting, well-characterized population isolates. The biggest populations here are in Quebec, Finland, Costa Rica, and South Africa

high-impact disease alleles. They will expose pathways of general significance for the clinical phenotype and thus disclose new targets for research. The better the characteristics of the populations and their history can be established, the better are the opportunities to design the optimal strategy for the disease-gene identification. An extra bonus is obtained from an access to the detailed population records, which facilitate construction of mega pedigrees and identification of families with shared ancestral alleles. Self-evidently, homogeneity in life style and environment facilitate the control of these intervening factors. Often, geneticists tend to focus on populations with an extraordinary high prevalence of the disease of interest, but for complex diseases, this might reflect not only genetic but also environmental effect. On the basis of the good experiences of gene identification for rare diseases, one might actually argue that a population with an exceptionally low prevalence of a complex disease might present a special opportunity for gene identification. In Figure 1, we provide a global view to some population isolates with a reasonable size, most of them already utilized in complex disease gene mapping efforts with promising results, some still representing a highly potential option for future studies.

4. Statistical methods

For the initial mapping of Mendelian traits, linkage analysis has proved remarkably successful (Wright *et al*., 1999). Linkage analysis evaluates the recombination events occurring within a sample of pedigrees under a precise model of the inheritance for the trait. The evidence in favor of linkage is summarized by the lod (logarithm of the odds) score plotted against the recombination fraction separating the disease locus and a nearby marker locus. The position of the linkage peak is

usually taken as an estimate of the genetic distance between the marker and the actual disease gene. In isolated populations, this initial positioning of the disease locus can be followed by fine mapping positioning accurately the disease gene based on the shared haplotype among affected individuals, originating from one founder.

With common diseases, the situation is less favorable both for initial positioning using linkage analysis and for final positioning of the disease locus linkage analysis. The genetic model for the trait cannot be well defined in linkage analysis, and genetic heterogeneity quickly erodes the statistical evidence for linkage for any single locus (Greenberg *et al*., 1998). Also, the initial positioning of the disease locus is not accurate. The position of the linkage peak is highly dependent on the parameters applied, and the incomplete penetrance of the disease genes results in broad linkage peaks, providing very little guidance of the actual position of the gene (Hovatta *et al*., 1998; Terwilliger, 2001). However, most studies have used linkage analysis in exceptional muliplex families, assuming a monogenic disease model with reduced penetrance (Clerget-Darpoux *et al*., 1986). The extremes of dominant and recessive models for inheritance are thought to cover most cases of family-ascertained study samples. However, nonparametric allele-sharing methods, avoiding explicit assumptions about disease transmission, are favored over parametric linkage analyses (Kruglyak, 1996; Sobel and Lange, 1996). If the affected individuals within a collection of selected pedigrees or sibships show an excessive sharing of marker alleles in a particular chromosomal region, that region is likely to harbor a disease gene. The power of parametric versus nonparametric methods depends strongly on how strong is the genetic element of the disease. For both parametric and nonparametric methods, isolated populations obviously offer an advantage because isolation limits genetic heterogeneity.

Association studies can be divided into case-control studies and familial tests of transmission distortion. In genome-wide scans even with an atypically high density of the markers, the association is the by-product of linkage disequilibrium. By contrast, the alleles at a candidate locus may directly influence the risk of disease. In a genetically homogeneous population, case-control studies involve straightforward comparison of allele or haplotype frequencies in random samples of cases and controls, both ascertained from the same (sub)population (Risch and Merikangas, 1996). In heterogeneous populations, with inbuilt problems of the population stratification, association studies rely heavily on the transmission-disequilibrium test (TDT) (Terwilliger and Ott, 1992; Spielman *et al*., 1993). TDT compares how parental alleles are transmitted or not transmitted to affected children and evaluates whether one or more alleles are preferentially transmitted to the affected children. Although the TDT method was originally recommended for independent trios of both parents and an affected child, versions of the TDT exist for sibships and even pedigrees (Allison *et al*., 1999; Boehnke and Langefeld, 1998; Spielman and Ewens, 1998; Schaid and Rowland, 1998; Teng and Risch, 1999). With parametric versions of the TDT, it is possible to combine case-control data and familial transmission data into a single analysis that checks both differences in allele frequencies and distorted transmission to children (Sinsheimer *et al*., 2000).

Analogous to linkage mapping, population isolates have certain advantages over outbred populations in association studies. The uniform genetic and environmental background of a population isolate makes case-control studies less suspect.

Population isolates and outbred populations depart from genetic equilibrium in different ways. Hardy–Weinberg equilibrium, which is more likely in a population isolate than in a population with considerable racial admixture, helps in testing differences in allele frequencies between case and controls.

Haplotype mapping has the advantage of revealing ancient recombination events in addition to the contemporary recombination events detectable by linkage mapping. Linkage mapping operates best on a coarse scale where haplotype signatures are lost. Thus, it is sensible to view linkage and haplotype mapping as complementary techniques. The usual strategy is to use linkage mapping to identify a candidate region and then to follow up with haplotype mapping to pinpoint the disease locus within that region (Jorde *et al*., 2000; Peltonen, 2000). If a good candidate gene, based on the known biology, is suspected in advance, it is possible to reverse the standard procedure and test first for association between the disease and alleles at the candidate gene or nearby markers. In practice, most geneticists hedge their candidate gene bets by including these genes as markers in a genome scan. New methods providing efficient SNP genotyping and rapidly increasing information of the LD intervals and SNP haploblocks in different populations are paving the way toward genome-wide haplotype-association studies (see later).

Efforts have been made to quantify the extent of LD around a disease gene and using this as a guide to positional cloning. Inference from any single marker tends to be unreliable. Because of differences in the information content of the markers and the high probability of a by-chance association of a disease-causing variant with the marker allele, single-marker measures of disequilibrium do not greatly contribute to the final positioning of a disease gene. This fact has prompted the development of multimarker measures of disequilibrium and model-based methods of analysis that exploit these measures (Collins and Morton, 1998; Devlin *et al*., 1996; Kaplan *et al*., 1995; Lazzeroni and Lange, 1998; McPeek and Strahs, 1999; Rannala and Slatkin, 1998; Terwilliger, 1995; Xiong and Guo, 1997). Furthermore, informativeness of the marker has a distinct impact on the detection of intermarker LD. In the Finnish subisolate, the marker pairs of most informative microsatellite markers exposed intermarker LD over wide intervals reaching 7.5 cM, whereas less-informative SNPs revealed significant LD over very limited intervals, typically less than 0.1 cM (Varilo *et al*., 2003).

Much enthusiasm and debate surrounds the potential of the international high-profile Haplotype Map (**www.hapmap.org**) project and its ability to guarantee a high resolution of allelic variants for complex disease gene mapping even in association-based studies (Daly *et al*., 2001; Weiss and Clark, 2002; Wall and Pritchard, 2003; Terwilliger and Weiss, 2003). So far, the studies addressing the diversity of haploblocks have been carried out in a highly limited number of populations. It is not a surprising finding that the extent of intermarker LD with SNPs is much less and the haploblocks more diverse in African populations than in the younger global populations. The average interval showing strong evidence of recombination was found to be only 8 kb in Yoruban (from Nigeria) and African–American samples compared to 22 kb in the European and Asian samples (Gabriel *et al*., 2002). Similarly, LD was observed over 5 kb in Yorubans versus 60 kb in US population of North-European ancestry (Reich *et al*., 2001). The size of

haploblocks constructed using common SNPs varies from a few to several hundreds of kilobases and the sizes are greatly variable in different chromosomal regions. Although the statistical methods addressing the true significance of the observed haploblocks are still under construction, the number of haploblocks is typically higher. Gabriel *et al*. (2002) reported 5 common haploblocks ($>5\%$) in Africans versus 3.5 in other populations. It is important to realize that the outcome of the Hapmap based on the information collected from common SNPs (exposing common haploblocks) will not reveal the full variety of the allelic spectrum in different global populations. More detailed analysis of the distribution of the allelic diversity will be needed to identify deviations in this distribution in disease-associated alleles, and this will require a significant amount of extensive sequencing of different alleles.

5. Diversity of complex disease alleles

Evaluating the amount and nature of the genetic component in complex diseases is problematical. Heritability estimates of any given trait only express the proportion of the phenotypic variance in a study population that can be explained by familial factors. It is often forgotten that the obtained values are influenced by all genetic and shared environmental factors, and their interactions. Thus, the estimates for the heritability do not necessarily provide meaningful information of the feasibility of the genetic mapping studies of a given trait.

Let us consider two extreme models of the genetic background of complex traits: Common alleles of a handful of loci could interact to predispose to a disease. Such alleles would show low penetrance but could impose a significant risk at the population level. Alternatively, rare alleles at several loci can each significantly increase the predisposition to a disease. Such alleles could be family specific, show high penetrance, and would be relatively nonsignificant at the population level. In both models, the locus and allelic heterogeneity would be more restricted in an isolate. However, it is obvious that population isolates are especially valuable for the isolation of rare, high-impact genes since the founder effect and/or bottlenecks have dramatically restricted the number of alleles, making the genetic background closely resemble that of any monogenic disease.

It is plausible that population isolates originating from a small number of founders will expose "population specific" alleles or even disease genes. However, little experimental data exists to support this. Most linkage results disclosed in families from population isolates have been replicated in other populations and when this is not the case, differences in the ascertainment strategy or applied clinical parameters can equally contribute to the obtained results. Gene mapping efforts using multiple populations not only hold the promise of confirmation but also provide more information (more genotyped individuals, more meiotic information), allowing a better definition of the mapped interval containing the gene. It is worth stressing here the advantages of older outbred populations since they expose a shorter range of linkage disequilibrium intervals.

6. Genetic makeup of population isolates

With regard to multifactorial diseases presumably caused by old mutations, some population isolates are suggested to be more valuable than others for genetic studies (Peltonen *et al*., 1999; Wright *et al*., 1999). Isolates such as the Saami (Lapps of Scandinavia) or the Basques of southern Europe are well established, 200–400 generations old, and demographically stable (Terwilliger *et al*., 1998). The Saami forager population of 50 000 is estimated to have been of constant size throughout history, regulated and spread to large land areas by their source of livelihood, whereas agriculture led to a population expansion among the major European populations. However, because of their small size, complicated accessibility, and the paucity of recessive diseases, both Basques and Saamis have so far been largely ignored by geneticists.

For studies of complex traits, conscious geneticists have tended to target recent founder populations (100–200 generations old) such as Finland, Sardinia or Iceland, or even younger population isolates (10–20 generations old) that originate from a small number of founders and have undergone rapid population expansion. Examples are populations of late-settled parts of Finland (reaching northern Sweden), Central Valley of Costa Rica, Mennonite, French-speaking Quebec, Newfoundland, Afrikaners, and some regions of Holland. A modification of this strategy has been to send a field group to a remote, usually very small, rural subisolate to measure all quantitative information of various interested phenotypic features (Williams-Blangero *et al*., 2002; Bulayeva *et al*., 2000). Because of the accelerative internationalization, such populations are gradually vanishing, and it should be our minimum responsibility to gather adequate samples for the complex disease projects in future.

We have scrutinized the characteristics of the alleles of Finnish subpopulations, representing internal isolates of this relatively homogeneous population and found them striking. Although the roots of the population of Finland reach over 2000 years, the internal migration within Finland in the 1500s created regional subisolates. Villages were often established by only a few founder couples, which was a potentially ideal setting for complex gene mapping efforts. The effects of multiple bottlenecks can be demonstrated even today by marked differences in the prevalence of rare disease mutations among regional subisolates (Pastinen *et al*., 2001). Historical routes of many internal migrations can actually be reconstructed using the distribution of these disease alleles and the historical time when the mutation was introduced to the population determined based on the interval over which the markers show LD in the disease alleles (Varilo *et al*., 1996b; Varilo *et al*., 1996a).

In fine mapping of complex disease genes, LD and haplotype sharing should be of special value in population isolates. The LD intervals reflect the history of (sub)populations and emphasize the importance of the detailed population history for the study design in the disease gene hunt. Small, constant-size populations like the Saami can be expected to exhibit LD over large genomic intervals and greatly reduce allelic and haplotype diversity, both owing to genetic drift (Laan

and Paabo, 1997; Laan and Paabo, 1998; Terwilliger *et al*., 1998); such populations may be especially powerful for the initial phase of mapping common trait loci if only adequate study samples are available. The advantages of rapidly expanded population isolates have been based on the research on the 100–200 generations old founder populations such as Sardinia or Western early settlement regions of Finland (Zavattari *et al*., 2000; Mohlke *et al*., 2001). The results might be somewhat misleading since population admixture on these coastal areas could easily have dampened the amount of LD observed (Heutink and Oostra, 2002; Angius *et al*., 2001). There are practical implications of this type of information of specific regional subisolates for the prospect and process of LD-based disease gene mapping in study samples from one population. The long-range LD required for the initial genome-wide mapping exists in young subpopulations, with a well-established population history. Conversely, the LD decay, associated with the older main population, represents an efficient tool for finer-scale gene localization.

7. Isolates provide an avenue for efficient collection of pedigrees

Most successful gene identifications in complex traits (Table 1) have so far been initiated in large pedigrees with carefully defined and often extreme clinical phenotypes and at least to some extent shared environment. This has probably resulted in the identification of somewhat exceptional, "high impact" genes, most of them having a low attributable fraction at the population level. The most extensive example of the systematic use of population history, genealogical records, and nation-wide health care registries in genetic research is the Decode project in Iceland. Although not representing an ideal population isolate (Arnason *et al*., 2000; Helgason *et al*., 2001; Heutink and Oostra, 2002), the excellent genealogical records uncovering extensive pedigrees with shared ancestors for several complex diseases simultaneously have gained momentum for the disease locus mapping in Iceland. This is illustrated in the identification of multiple complex disease loci and potentially involved alleles (**www.decode.com**). Also in the Decode project, most of the identified alleles represent rare, most probably high-impact alleles, and their significance at the population level remains unknown. Nevertheless, exploring the biochemical pathways behind rare, high-impact genes identified in special populations will without doubt contribute to our understanding of the molecular pathogenesis of these traits. The initial findings concerning specific alleles and their significance for the disease phenotype need to be followed by genetic analyses in other populations and epidemiological study samples for elucidating their general significance and their population attributable fraction.

References

Allen M, Heinzmann A, Noguchi E, Abecasis G, Broxholme J, Ponting CP, Bhattacharyya S, Tinsley J, Zhang Y, Holt R, *et al*. (2003) Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nature Genetics*, **35**, 258–263.

Allison DB, Heo M, Kaplan N and Martin ER (1999) Sibling-based tests of linkage and association for quantitative traits. *American Journal of Human Genetics*, **64**, 1754–1763.

Angius A, Melis PM, Morelli L, Petretto E, Casu G, Maestrale GB, Fraumene C, Bebbere D, Forabosco P and Pirastu M (2001) Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Human Genetics*, **109**, 198–209.

Arnason E, Sigurgislason H and Benedikz E (2000) Genetic homogeneity of Icelanders: fact or fiction? *Nature Genetics*, **25**, 373–374.

Boehnke M and Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *American Journal of Human Genetics*, **62**, 950–961.

Bulayeva KB, Leal SM, Pavlova TA, Kurbanov R, Coover S, Bulayev O and Byerley W (2000) The ascertainment of multiplex schizophrenia pedigrees from Daghestan genetic isolates (Northern Caucasus, Russia). *Psychiatric Genetics*, **10**, 67–72.

Cavalli-Sforza LL (1998) The DNA revolution in population genetics. *Trends in Genetics*, **14**, 60–65.

Clerget-Darpoux F, Bonaiti-Pellie C and Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, **42**, 393–399.

Collins A and Morton NE (1998) Mapping a disease locus by allelic association. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1741–1745.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Deeb SS, Fajas L, Nemoto M, Pihlajamaki J, Mykkanen L, Kuusisto J, Laakso M, Fujimoto W and Auwerx J (1998) A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genetics*, **20**, 284–287.

Devlin B, Risch N and Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, **36**, 1–16.

Emamian ES, Hall D, Birnbaum MJ, Karayiorgou M and Gogos JA (2004) Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia. *Nature Genetics*, **36**, 131–137.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L and Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, **30**, 233–237.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Greenberg DA, Abreu P and Hodge SE (1998) The power to detect linkage in complex disease by means of simple LOD-score analyses. *American Journal of Human Genetics*, **63**, 870–879.

Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, Jonsdottir T, Gudmundsdottir T, Bjarnadottir SM, Einarsson OB, Gudjonsdottir HM, *et al*. (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genetics*, **35**, 131–138.

Helgason A, Hickey E, Goodacre S, Bosnes V, Stefansson K, Ward R and Sykes B (2001) mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *American Journal of Human Genetics*, **68**, 723–737.

Hennah W, Varilo T, Kestila M, Paunio T, Arajarvi R, Haukka J, Parker A, Martin R, Levitzky S, Partonen T, *et al*. (2003) Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. *Human Molecular Genetics*, **12**, 3151–3159.

Heutink P and Oostra BA (2002) Gene finding in genetically isolated populations. *Human Molecular Genetics*, **11**, 2507–2515.

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, *et al*. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics*, **26**, 163–175.

Houwen RH, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA and Freimer NB (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics*, **8**, 380–386.

Hovatta I, Lichtermann D, Juvonen H, Suvisaari J, Terwilliger JD, Arajärvi R, Kokko-Sahin M-L, Ekelund J, Lönnqvist J and Peltonen L (1998) Linkage analysis of putative schizophrenia gene

candidate regions on chromosomes 3p, 5q, 6p, 8p, 20p and 22q in a population-based sampled Finnish family set. *Molecular Psychiatry*, **3**, 452–457.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, *et al*. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.

Johnson GC and Todd JA (2000) Strategies in complex disease mapping. *Current Opinion in Genetics and Development*, **10**, 330–334.

Jorde LB, Watkins WS, Kere J, Nyman D and Eriksson AW (2000) Gene mapping in isolated populations: new roles for old friends? *Human Heredity*, **50**, 57–65.

Kaplan NL, Hill WG and Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. *American Journal of Human Genetics*, **56**, 18–32.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139–144.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Laan M and Paabo S (1997) Demographic history and linkage disequilibrium in human populations. *Nature Genetics*, **17**, 435–438.

Laan M and Paabo S (1998) Mapping genes by drift-generated linkage disequilibrium. *American Journal of Human Genetics*, **63**, 654–656.

Laitinen T, Polvi A, Rydman P, Vendelin J, Pulkkinen V, Salmikangas P, Makela S, Rehn M, Pirskanen A, Rautanen A, *et al*. (2004) Characterization of a common susceptibility locus for asthma-related traits. *Science*, **304**, 300–304.

Lazzeroni LC and Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity*, **48**, 67–81.

Love-Gregory LD, Wasson J, Ma J, Jin CH, Glaser B, Suarez BK and Permutt MA (2004) A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4 alpha gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an ashkenazi jewish population. *Diabetes*, **53**, 1134–1140.

McPeek MS and Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics*, **65**, 858–875.

Mira MT, Alcais A, Nguyen VT, Moraes MO, Di Flumeri C, Vu HT, Mai CP, Nguyen TH, Nguyen NB, Pham XK, *et al*. (2004) Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature*, **427**, 636–640.

Mohlke KL, Lange EM, Valle TT, Ghosh S, Magnuson VL, Silander K, Watanabe RM, Chines PS, Bergman RN, Tuomilehto J, *et al*. (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Research*, **11**, 1221–1226.

Newman DL, Abney M, McPeek MS, Ober C and Cox NJ (2001) The importance of genealogy in determining genetic associations with complex traits. *American Journal of Human Genetics*, **69**, 1146–1148.

Nikali K, Suomalainen A, Terwilliger J, Koskinen T, Weissenbach J and Peltonen L (1995) Random search for shared chromosomal regions in four affected individuals: the assignment of a new hereditary ataxia locus. *American Journal of Human Genetics*, **56**, 1088–1095.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, *et al*. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, Duan XJ, Soro-Paavonen A, Naukkarinen J, Saarela J, *et al*. (2004) Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nature Genetics*, **36**, 371–376.

Pastinen T, Perola M, Ignatius J, Sabatti C, Tainola P, Levander M, Syvanen AC and Peltonen L (2001) Dissecting a population genome for targeted screening of disease mutations. *Human Molecular Genetics*, **10**, 2961–2972.

Peltonen L (2000) Positional cloning of disease genes: advantages of genetic isolates. *Human Heredity*, **50**, 66–75.

Peltonen L, Jalanko A and Varilo T (1999) Molecular genetics of the Finnish disease heritage. *Human Molecular Genetics*, **8**, 1913–1923.

Puffenberger EG, Hosoda K, Washington SS, Nakao K, deWit D, Yanagisawa M and Chakravart A (1994) A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease. *Cell*, **79**, 1257–1266.

Rannala B and Slatkin M (1998) Likelihood analysis of disequilibrium mapping, and related problems. *American Journal of Human Genetics*, **62**, 459–473.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Schaid DJ and Rowland C (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *American Journal of Human Genetics*, **63**, 1492–1506.

Silander K, Mohlke KL, Scott LJ, Peck EC, Hollstein P, Skol AD, Jackson AU, Deloukas P, Hunt S, Stavrides G, *et al*. (2004) Genetic variation near the hepatocyte nuclear factor-4 alpha gene predicts susceptibility to type 2 diabetes. *Diabetes*, **53**, 1141–1149.

Sinsheimer JS, Blangero J and Lange K (2000) Gamete-competition models. *American Journal of Human Genetics*, **66**, 1168–1172.

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Spielman RS and Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics*, **62**, 450–458.

Spielman RS, McGinnis RE and Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, **43**, 166–172.

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, *et al*. (2002) Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics*, **71**, 877–892.

Straub RE, Jiang Y, MacLean CJ, Ma Y, Webb BT, Myakishev MV, Harris-Kerr C, Wormley B, Sadek H, Kadambi B, *et al*. (2002) Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *American Journal of Human Genetics*, **71**, 337–348.

Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, Nagasaki M, Nakayama-Hamada M, Kawaida R, Ono M, *et al*. (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nature Genetics* **34**, 395–402.

Taipale M, Kaminen N, Nopola-Hemmi J, Haltia T, Myllyluoma B, Lyytinen H, Muller K, Kaaranen M, Lindsberg PJ, Hannula-Jouppi K, *et al*. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11553–11558.

Teng J and Risch N (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Research*, **9**, 234–241.

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, **56**, 777–787.

Terwilliger JD (2001) On the resolution and feasibility of genome scanning approaches. *Advances in Genetics*, **42**, 351–391.

Terwilliger JD and Ott J (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity*, **42**, 337–346.

Terwilliger JD and Weiss KM (2003) Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Annals of Medicine*, **35**, 532–544.

Terwilliger JD, Zöllner S, Laan M and Pääbo S (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: "Drift Mapping" in small populations with no demographic expansion. *Human Heredity*, **48**, 138–154.

Triggs-Raine BL, Kirkpatrick RD, Kelly SL, Norquay LD, Cattini PA, Yamagata K, Hanley AJ, Zinman B, Harris SB, Barrett PH, *et al*. (2002) HNF-1alpha G319S, a transactivation-deficient mutant, is associated with altered dynamics of diabetes onset in an Oji-Cree community. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4614–4619.

Varilo T, Nikali K, Suomalainen A, Lonnqvist T and Peltonen L (1996a) Tracing an ancestral mutation: genealogical and haplotype analysis of the infantile onset spinocerebellar ataxia locus. *Genome Research*, **6**, 870–875.

Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD and Peltonen L (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Human Molecular Genetics*, **12**, 51–59.

Varilo T, Savukoski M, Norio R, Santavuori P, Peltonen L and Järvelä I (1996b) The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. *American Journal of Human Genetics*, **58**, 506–512.

Wall JD and Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews. Genetics*, **4**, 587–597.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, **18**, 19–24.

Williams-Blangero S, VandeBerg JL, Subedi J, Aivaliotis MJ, Rai DR, Upadhayay RP, Jha B and Blangero J (2002) Genes on chromosomes 1 and 13 have significant effects on Ascaris infection. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5533–5538.

Wright AF, Carothers AD and Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23**, 397–404.

Xiong M and Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, **60**, 1513–1531.

Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, Loddo M, Eaves I, Mastio G, Todd JA and Cucca F (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Human Molecular Genetics*, **9**, 2947–2957.

Short Specialist Review

Allergy and asthma

Miriam F. Moffatt and William O.C. Cookson
University of Oxford, Oxford, UK

1. Introduction

Asthma is an inflammatory disease of the small airways of the lung. There is a modern epidemic of asthma, which now affects more than 10% of children in many westernized societies. The skin rash of infantile eczema (atopic dermatitis, AD) is also increasingly common in the developed world, affecting up to 15% of children in some countries. Asthma is present in 60% of children with severe AD, a significant proportion of which continue with problems into adult life. Both diseases are familial, and arise from the interaction between strong genetic and environmental factors (*see* Article 58, **Concept of complex trait genetics**, Volume 2).

AD, asthma, and hay fever are often considered to be part of a common syndrome of atopic diseases. The term "atopy" is variously used, but it most consistently refers to the presence of IgE-mediated skin test responses to common allergens. Atopic individuals are also typified by the presence of elevated levels of total and allergen-specific IgE in serum.

Childhood-onset asthma runs strongly in families, and studies of twins with asthma or AD show a heritability of approximately 60%. Consequently, many genome screens and candidate gene studies have been carried out to search for genetic effects on these illnesses.

2. The genetic basis of asthma

2.1. Genome screens

At least 11 full genome screens have been reported for asthma and its associated phenotypes (reviewed in (Cookson, 2003; Wills-Karp and Ewart, 2004)). These have identified 10 regions of linkage that were reproducible between screens and 4 regions that were statistically significant but not replicated by other groups (Cookson, 2003).

Asthma consistently shows linkage to the Major Histocompatibility Complex (MHC) (Cookson, 2002), in common with many other diseases, and linkage loci for asthma also overlap with loci for ankylosing spondylitis on chromosomes 1p31–36,

7p13, and 16q23; with type 1 diabetes on 1p32–34, 11q13, and 16q22–24; and with multiple sclerosis and rheumatoid arthritis on 17q22–24 (Cookson, 2002) (Figure 1).

These findings may mean that the susceptibility to different diseases arising from these loci is influenced by individual genes in various forms (alleles). Alternatively, as in the case of the MHC, disease susceptibility may be modified by physical clusters of genes that have a variety of effects on immune responses.

Three genome screens for AD have been carried out and have identified significant linkage to AD in four regions (reviewed in (Bowcock and Cookson, 2004)). These regions do not overlap with known regions of linkage to asthma but are closely coincident with psoriasis susceptibility loci (Bowcock and Cookson, 2004) (Figure 1). This suggests that particular genes or families of genes have general effects on immune reactions in the skin.

Wiskott–Aldrich syndrome, Hyper IgE syndrome, and Netherton's disease are Mendelian (single gene) disorders that show strong features of atopy.

Children with Netherton's disease consistently develop symptoms of atopic disease (hay fever, food allergy, urticaria, and asthma) and high levels of serum IgE. Mutations in the gene encoding a serine protease inhibitor known as SPINK5 or LEKTI was shown to be causing disease in these patients. Subsequent work has shown that common polymorphism in SPINK5 (particularly Glu420→Lys) modifies the risk of developing AD, asthma, and elevated serum IgE levels (Walley *et al*., 2001).

The SPINK5 protein contains 13 active protease inhibitor domains that provide a polyvalent action against multiple substrates. It is expressed in the outer epidermis, suggesting that AD and asthma may result from the failure to inhibit environmental proteases such as those that arise from bacteria or allergens.

Genetic linkage studies of AD and psoriasis have both highlighted the importance of chromosome 1q21, which contains the epidermal differentiation complex (EDC). Many of the EDC genes have shown increased expression in AD (Nomura *et al*., 2003) and in psoriasis (Nomura *et al*., 2003; Zhou *et al*., 2003).

Several gene families are present within the complex: these encode small proline-rich proteins (SPRRs), S100A calcium-binding proteins, and late envelope proteins (LEPs). The *SPRR* and *LEP* genes encode precursor proteins of the cornified cell envelope, whereas *S100* genes have antimicrobial and immune signaling functions. Expression of the EDC proteins occurs late during the maturation of epidermal cells, and, like SPINK5, EDC proteins are mainly localized just beneath the cornified envelope.

2.2. *Postionally cloned genes for asthma*

Four completely novel genes underlying asthma have been identified by positional cloning, from chromosomes 2q14 (*DPP10*) (Allen *et al*., 2003), 7p14 (*GPRA*) (Laitinen *et al*., 2004), 13q14 (*PHF11*) (Zhang *et al*., 2003), and 20p (*ADAM33*) (Van Eerdewegh *et al*., 2002). The functions and activities of these genes are as yet poorly understood, but they certainly do not fit into classical pathways of asthma pathogenesis.

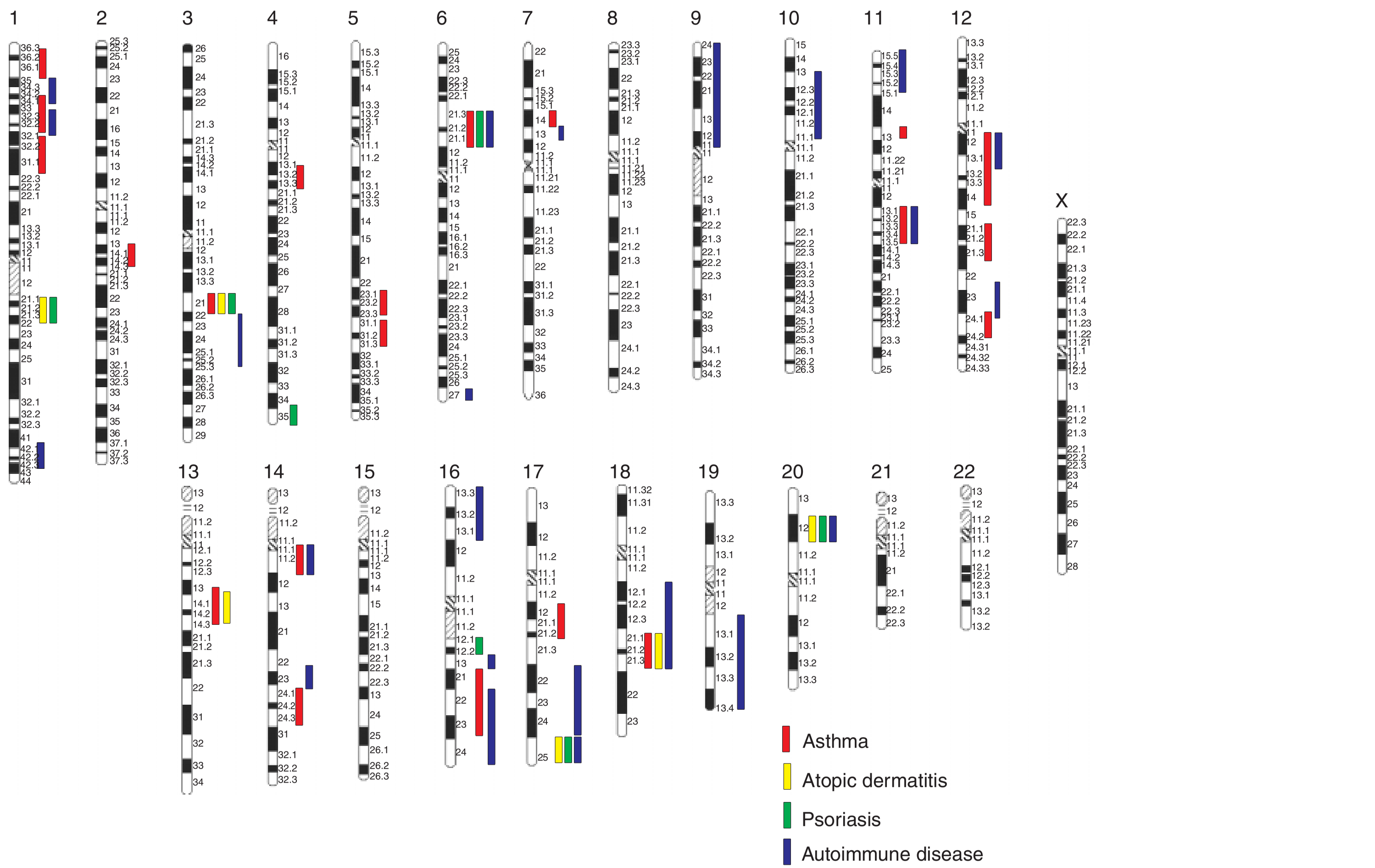


Figure 1 Genome screens for asthma, AD, and other immune disorders. The results are shown schematically, and only include the most significant linkages. The length of the bars indicates imprecision of localization. Clustering of disease susceptibility genes is found for the MHC on chromosome 6p21, and in several other genomic regions

ADAM33 is expressed in bronchial smooth muscle, and is thought to alter the hypertrophic response of bronchial smooth muscle to inflammation (a component of a process known as airway remodeling) (Van Eerdewegh *et al*., 2002). *PHF11* encodes a nuclear receptor that is part of a complex containing a histone methyl transferase (*SETDB2*), a regulator of HDAC (*RCBTB1*) and a nuclear transport molecule (*karyopherin α3*). *DPP10* encodes a prolyl dipeptidase, which may remove the terminal two peptides from certain inflammatory chemokines (Allen *et al*., 2003), and *GPRA* encodes an orphan G protein-coupled receptor named GPRA (G protein-coupled receptor for asthma susceptibility).

It is of interest that DPP10 and GPRA are both concentrated in the terminally differentiating bronchial epithelium. The homologous layer in the epidermis is the site of maximal expression of *SPINK5* and the genes of the EDC.

2.3. Candidate genes

2.3.1. The MHC

The MHC is the longest studied locus influencing atopy. It is well known that *HLA-DR* alleles restrict the IgE response to particular allergens, usually with a relative risk less than 2 (Young *et al*., 1994). In addition, the MHC Class II associated *TNF -308* promoter single nucleotide polymorphism (SNP) shows robust associations with asthma, independently of association to particular allergens (Moffatt and Cookson, 1997).

The MHC has, however, been little studied in individuals with AD. Genome screens do not show linkage of AD to the region, and no convincing associations with Human Leucocyte Antigen (HLA) Class I or Class II alleles have been established. Definitive studies are however lacking.

The ability to react to particular allergens has also been linked to the *TCR-α/δ* locus (but not *TCR-β*), and *HLA-DR* and *TCR-α/δ* alleles interact in the susceptibility to house dust mite allergens (Moffatt *et al*., 1997). The importance of γ/δ T-cells in dermal immunity suggests that the *TCR-α/δ* locus should also be explored in individuals with AD.

2.3.2. *FcϵRI-β* (chromosome 11q13)

Chromosome 11q13 was originally linked to atopy and was subsequently shown to contain the β-chain of the high-affinity receptor for IgE. Polymorphisms in *FcϵRI-β* are associated with asthma, allergy, bronchial hyper-responsiveness, and AD. These variations seem to be associated with severe atopic disease.

FcϵRI-β acts as an approximately sevenfold-amplifying element of the high-affinity IgE receptor response to activation and stabilizes the expression of the receptor on the mast cell surface. *FcϵRI-β* may therefore modify nonspecifically the strength of the response to allergens. Coding polymorphisms have been identified within the gene, but do not appear to alter its function. The actions of other polymorphisms within regulatory elements of the gene are currently under investigation.

2.3.3. The IL-4 cytokine cluster (chromosome 5q34)

The cytokine cluster on chromosome 5q34 contains many candidates that might influence atopic processes, including *IL-4*, *IL-13*, *GM-CSFH*, and *IL-9*. Polymorphisms in *IL-4* may be weakly associated with asthma, but a far stronger association has been established between *IL-13* polymorphisms and increased serum IgE levels, atopy, and asthma. The coding polymorphism Arg130->Gln seems to show the strongest effect (Graves *et al*., 2000).

These polymorphisms have not yet been explored for a role in AD. An association between *GM-CSF* and the severity of AD has been suggested but not yet confirmed.

IL4-Ra on chromosome 16 is a shared component of the receptor for both IL-4 and IL-13, and polymorphisms in this gene are also associated with asthma and atopy. It is of interest that different asthma-associated traits are associated with individual polymorphisms that affect splicing of IL4-Ra, illustrating the complexity of mechanisms that may vary the actions of a single gene.

2.3.4. Genes that interact with the microbial environment

CD14 is a receptor for bacterial lipopolysacchride (LPS, also known as endotoxin). This molecule is part of the innate immune response against bacterial infection. Polymorphism in the *CD14* gene is also associated with asthma, perhaps providing some of the structural explanation for the hygiene hypothesis. The intracellular pattern-recognition receptors *NOD1* and *NOD2* are also associated with asthma, through unknown mechanisms.

2.3.5. Pharmacogenetics and asthma

β-adrenergic drugs are the first line of treatment for asthma, and act through the β-adrenergic receptor (ADRB). Functional variants in *ADRB* modify the response of individual asthmatics to β-adrenergic therapy. A variant within the promoter of the 5-lipoxygenase gene predicts the response of asthmatics to antileukotriene therapy.

Acknowledgments

WOCC and MFM are funded by the Wellcome Trust.

References

Allen M, Heinzmann A, Noguchi E, Abecasis G, Broxholme J, Ponting CP, Bhattacharyya S, Tinsley J, Zhang Y, Holt R, *et al*. (2003) Positional cloning of a novel gene influencing asthma from Chromosome 2q14. *Nature Genetics*, **35**(3), 258–263.

Bowcock AM and Cookson WO (2004) The genetics of psoriasis, psoriatic arthritis and atopic dermatitis. *Human Molecular Genetics*, **13**(Spec No 1), R43–R55.

Cookson W (2002) Genetics and genomics of asthma and allergic diseases. *Immunological Reviews*, **190**, 195–206.

Cookson W (2003) A new gene for asthma: would you ADAM and Eve it? *Trends in Genetics*, **19**(4), 169–172.

Graves PE, Kabesch M, Halonen M, Holberg CJ, Baldini M, Fritzsch C, Weiland SK, Erickson RP, von Mutius E and Martinez FD (2000) A cluster of seven tightly linked polymorphisms in the IL-13 gene is associated with total serum IgE levels in three populations of white children. *Journal of Allergy and Clinical Immunology*, **105**(3), 506–513.

Laitinen T, Polvi A, Rydman P, Vendelin J, Pulkkinen V, Salmikangas P, Makela S, Rehn M, Pirskanen A, Rautanen A, *et al*. (2004) Characterization of a common susceptibility locus for asthma-related traits. *Science*, **304**(5668), 300–304.

Moffatt MF and Cookson WO (1997) Tumour necrosis factor haplotypes and asthma. *Human Molecular Genetics*, **6**(4), 551–554.

Moffatt MF, Schou C, Faux JA and Cookson WO (1997) Germline TCR-A restriction of immunoglobulin E responses to allergen. *Immunogenetics*, **46**(3), 226–230.

Nomura I, Gao B, Boguniewicz M, Darst MA, Travers JB and Leung DY (2003) Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. *Journal of Allergy and Clinical Immunology*, **112**(6), 1195–1202.

Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, *et al*. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature*, **418**(6896), 426–430.

Walley AJ, Chavanas S, Moffatt MF, Esnouf RM, Ubhi B, Lawrence R, Wong K, Abecasis GR, Jones EY Harper JI, *et al*. (2001) Gene polymorphism in Netherton and common atopic disease. *Nature Genetics*, **29**(2), 175–178.

Wills-Karp M and Ewart SL (2004) Time to draw breath: asthma-susceptibility genes are identified. *Nature Reviews Genetics*, **5**(5), 376–387.

Young RP, Dekker JW, Wordsworth BP and Cookson WOCM (1994) HLA-DR and HLA-DP genotypes and Immunoglobulin E responses to common major allergens. *Clinical Experimental Allergy*, **24**, 431–439.

Zhang Y, Leaves NI, Anderson GG, Ponting CP, Broxholme J, Holt R, Edser P, Bhattacharyya S, Dunham A, Adcock IM, *et al*. (2003) Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nature Genetics*, **34**(2), 181–186.

Zhou X, Krueger JG, Kao M-CJ, Lee E, Du F, Menter A, Wong WH and Bowcock AM (2003) Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiological Genomics*, 69–78.

Short Specialist Review

Inflammation and inflammatory bowel disease

Christopher G. Mathew
King's College London, London, UK

1. Introduction

Inflammatory bowel disease (IBD) is associated with chronic inflammation of the gastrointestinal tract, and has two main forms, Crohn's disease (CD) and ulcerative colitis (UC). CD can affect any part of the intestine, with discontinuous, transmural lesions of the gut wall, whereas in UC, inflammation is confined to the colon and rectum, and lesions are continuous and superficial. The molecular basis of pathogenesis in IBD is not yet clear, but may involve persistent bacterial infection, a defective mucosal barrier, and an imbalance in the regulation of the intestinal immune response. There is strong evidence from both epidemiological and genetic studies for the existence of genetic determinants of susceptibility to IBD (Bouma and Strober, 2003; Mathew and Lewis, 2004).

2. The discovery of *NOD2* (*CARD15*)

Genetic linkage studies in IBD families identified a region of linkage on chromosome 16 (the *IBD1* locus), which was later shown by positional cloning to contain a gene called *NOD2* (now called *CARD15*) (Hugot *et al*., 2001). Mutations in *CARD15* were shown to be associated with susceptibility to CD (but not to UC), and the association was reported independently by two other groups (Hampe *et al*., 2001; Ogura *et al*., 2001), and has since been widely replicated. The predicted structure of the CARD15 protein (Figure 1) contains two caspase recruitment domains (CARDs), a central nucleotide-binding oligomerization domain, and 10 carboxy-terminal leucine-rich repeats (LRRs), which offers some clues as to its function (see below). Three common mutations have been described (Figure 1); a frameshift (L1007fs) that results in loss of the C-terminal LRR, a missense mutation in the LRR (G908 R), and a missense mutation between the nucleotide-binding domain and the LRR (R702 W). Numerous other rare variants have been detected, some of which may also affect function (Lesage *et al*., 2002). Heterozygosity for the mutations confers a two- to threefold increase in risk of CD, but in mutation homozygotes, the risk is increased by about 30-fold. The fact that only about 35–40% of CD patients have mutations in *CARD15* and that the mutations

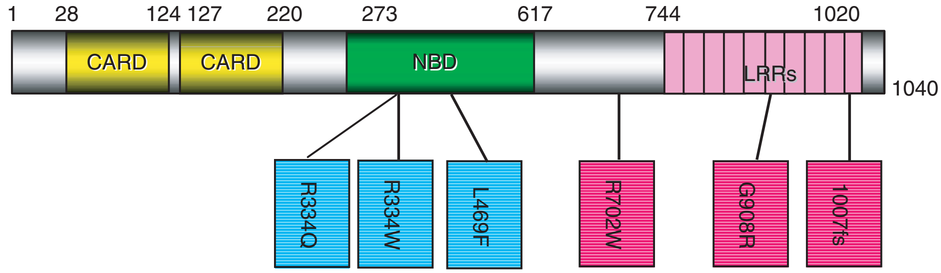


Figure 1 Predicted structure of the CARD15 protein and associated mutations. (CARD = caspase recruitment domain, NBD = nucleotide-binding domain, LRR = leucine-rich repeats). Crohn's disease mutations are in pink and Blau syndrome mutations in blue

are relatively common in the general population, shows that, as might be expected for a complex trait, they are neither necessary nor sufficient for the development of the disease. *CARD15* mutations have also been found in Blau syndrome, a rare autosomal dominant condition associated with early-onset granulomatous arthritis (Miceli-Richard *et al*., 2001). The Blau syndrome mutations are all located in the nucleotide-binding domain of CARD15, which may explain the differences in phenotype and penetrance.

3. *NOD2/CARD15* function

The identification of *NOD2/CARD15* was a very important discovery since it provided proof of the principle that positional cloning of susceptibility genes for complex diseases was possible. Knowledge of the identity of the encoded protein has also offered insight into a possible pathway to pathogenesis. The presence of LRRs suggested that, like the toll-like receptors (TLRs), it might be involved in the recognition of microbial components, and subsequent work showed that transfection of *CARD15* into HEK293 cells led to activation of NF-κB after stimulation with the peptidoglycan muramyl dipeptide (MDP) (Girardin *et al*., 2003; Inohara, Ogura *et al*., 2003). Patient-derived mutations of *CARD15* lead to a loss or reduction of MDP-mediated NF-κB activation in transfected and in primary cells (Ogura *et al*., 2001; Van Heel, *et al*., 2005; Li, Moran *et al*., 2004). This is puzzling, since CD is characterized by increased inflammation via an NF-κB-dependent T helper type (T_H1) response. Part of the resolution to this paradox may lie in the recent finding that *Card15*$^{-/-}$ mice have an enhanced TLR2-dependent response to peptidoglycan (Watanabe *et al*., 2004), which suggests that NOD2 downregulates the TLR2 response and that this control is deficient in Crohn's patients with *CARD15* mutations, leading to increased inflammation.

4. Other IBD genes

Numerous genetic linkage studies (reviewed in Mathew and Lewis, 2004) and a recent meta-analysis of genome scans (Van Heel, *et al*., 2004) have pointed to the existence of other susceptibility genes for IBD, including the MHC region

on chromosome 6p21 (see Table 1). Linkage around the cytokine gene cluster on chromosome 5q31-33 (*IBD5*) to CD led to the identification of a common risk haplotype over a region of 250 kb, which contained multiple SNPs that were in almost complete linkage disequilibrium (LD) with each other, and were strongly associated with CD (Rioux *et al*., 2001). The extensive LD did not permit the identification of a causative gene or mutation, but a recent report suggests that two variants in the *SLC22A4* and *SLC22A5* genes that encode organic cation transporters are the causative mutations (Peltekova *et al*., 2004). The missense substitution L503F in *SLC22A4* was associated with reduced carnitine uptake by the encoded OCTN1 transporter, and the −207G>C variant in *SLC22A5* (encoding OCTN2) occurs within a heat-shock transcription factor binding element that reduced transcriptional activation of a reporter gene by the mutated version of the promoter. It is argued that these variants may cause intestinal inflammation by, for example, defective metabolism of bacterial toxins (Peltekova *et al*., 2004), but the functional link between the OCTN transporters and CD is not immediately obvious. Confirmation of the role of these variants in the pathogenesis of Crohn's disease will require replication of their association with disease in the absence of the background 250 kb risk haplotype, and evidence of altered expression or activity of these genes in cells or tissues from CD patients.

Evidence for a third IBD susceptibility gene comes from a systematic search for association in a 40-cM region of linkage on chromosome 10. Transmission disequilibrium tests revealed strong association of a SNP with both CD and the broader IBD phenotype, with the signal being confined to a haplotype block of 85 kb across the *DLG5* gene at 10q23 (Stoll *et al*., 2004). Haplotype tagging SNPs were undertransmitted to IBD cases in a primary and a replication set of families, and a missense mutation (R30Q) was associated with disease in an independent sample of CD cases and controls, with an odds ratio of 1.6. *DLG5* is a member of the membrane-associated guanylate kinase gene family. It is expressed in the colon and intestinal epithelium (Stoll *et al*., 2004), and has been implicated in the regulation of cell growth and shape, and in the maintenance of epithelial cell integrity. Thus, it is proposed that genetic variants in *DLG5* may affect epithelial barrier function in the colon (Stoll *et al*., 2004). As with the *SLC22A4* and *SLC22A5* genes at the *IBD5* locus, independent replication of the genetic findings are needed, as well as functional studies to demonstrate the altered expression or activity of the DLG5 protein in cells or tissue from IBD patients.

5. Challenges for the future

What lessons can be drawn from the genetic findings at this point? First, the *CARD15* discovery emerged from a well-replicated linkage signal on chromosome 16, and, in the context of complex disorders, has a large effect size in mutation homozygotes. Also, since all three of the main susceptibility alleles arose on the same more ancient haplotype, common SNPs that tagged the haplotype all showed strong association with disease (Hugot *et al*., 2001). This discovery, impressive and important though it is, may therefore represent low-hanging fruit in the search for other genetic determinants of IBD. The challenge of identifying the causal

Table 1 Chromosomal regions showing significant/suggestive genome-wide evidence for linkage, from genome-wide linkage studies, and from two large consortium analyses (follow-up of chromosomes 12 and 16; IBD International Genetics Consortium, 2001), and a meta-analysis of 10 genome-wide linkage studies (Van Heel *et al*., 2004))

Region	Marker at max LOD score	Study	Phenotype	p-value (*significant)[b]	No. of families[a]	Association/genes identified
1p	D1 S236	Hugot *et al*. (1996)	CD	0.0006	42 ASPs	
1p	D1 S552	Cho *et al*. (1998)	IBD	2.4×10^{-4}	297 ARPs	
2q	MA[c]	Van Heel *et al*. (2004)	IBD	0.0043	1952 ARPs	
3p	D3 S1573	Satsangi *et al*. (1996)	IBD	0.00021	186 ASPs	
	D3 S1766	Rioux *et al*. (2000)	IBD	0.0004	170 ASPs	
3q	D3 S3053	Cho *et al*. (1998)	IBD	5.7×10^{-4}	297 ARPs	
5q	D5 S2497	Rioux *et al*. (2000)	CD	1.1×10^{-5}*	50 early-onset CD families	250-kb region of association, including *OCTN1* and *OCTN2*
5q	D5 S673	Ma *et al*. (1999)	CD	0.0003	26 ASPs (Jewish)	
6p	D6 S461	Hampe *et al*. (1999a)	IBD	5.5×10^{-6}	428 ASPs	Multiple MHC associations
6p	D6 S1281	Rioux *et al*. (2000)	IBD	6×10^{-4}	170 ASPs	
6p	MA[c]	Van Heel *et al*. (2004)	IBD	0.00012*	1952 ARPs	
7	D7 S669	Satsangi *et al*. (1996)	IBD	8.2×10^{-5}	186 ASPs	
10	D10 S458	Hampe *et al*. (1999b)	CD	5.7×10^{-4}	162 ASPs	Association at *DLG5* locus
12	D12 S83	Satsangi *et al*. (1996)	IBD	2.7×10^{-7}*	186 ASPs	
12	D12 S345	Ma *et al*. (1999)	CD	0.0004	65 ASPs	
12	D12 S364	van Heel *et al*. (2003)	CD	0.0005	25 ARPs (5q31−ve)	
14	D14 S261	Duerr *et al*. (2000)	CD	2.3×10^{-5}	127 ARPs	
	D14 S261	Ma *et al*. (1999)	CD	0.0002	65 ASPs	
15	D15 S128	Satsangi *et al*. (1996)	IBD	0.0004	89 ASPs	
16	D16 S409	Hugot *et al*., 1996	CD	1.5×10^{-5}*	78 families	*NOD2/CARD15*
16	D16 S411	IBDIGC *et al*. (2001)	CD	1.2×10^{-7}	386 ASPs	
16	MA[c]	Van Heel *et al*. (2004)	CD	0.0032	1068 ARPs	
19	D19 S591	Rioux *et al*. (2000)	IBD	2.1×10^{-6}*	170 ASPs	

Table 1 (*continued*)

Region	Marker at max LOD score	Study	Phenotype	p-value (*significant)[b]	No. of families[a]	Association/genes identified
19	D19 S217	van Heel *et al*. (2003)	CD	0.0001	64 ASPs (CARD15−ve)	
19	MA[3]	Van Heel *et al*. (2004)	CD	0.0066	1068 ARPs	

[a]ASPs: affected sib pairs; ARPs: affected relative pairs.
[b]Genome-wide studies: significant, $p < 2.2 \times 10^{-5}$; suggestive, $p < 7.4 \times 10^{-4}$. Meta-analysis: significant, $p < 0.0004$; suggestive, $p < 0.01$.
[c]MA: chromosomal region identified from meta-analysis (Van Heel *et al*., 2004).

gene and variants in large regions of extensive LD is exemplified by the *IBD5* locus on chromosome 5q31, and is likely to complicate efforts to identify the IBD gene or genes in the region of strong LD around the MHC locus on chromosome 6p21. Also, if the effect sizes for other loci are relatively small, such as appears to be the case for *DLG5*, very large sample sizes may be required to confirm their existence. The availability of the immense database of SNPs has led to a large expansion in the number of reported associations in IBD, many of which have not been replicated. This is due in part to methodological problems in study design (Colhoun *et al*., 2003), and also due to the large numbers of tests being conducted. Theoretical and practical issues relating to multiple testing are about to escalate dramatically with the introduction of genome-wide screens for association with panels of 100 000–300 000 SNPs. Finally, in those instances in which strong and well-replicated evidence of association has been obtained, there will be the challenge of demonstrating credible functional links between putative susceptibility genes and pathogenesis.

If the prospect of these challenges seems daunting, many of them are common to other complex disorders, and the attempt to resolve them will be extremely interesting. A few more successes may have a major impact on our understanding of pathogenesis in IBD, and may provide novel targets for improvements in therapy.

References

Bouma G and Strober W (2003) The immunological and genetic basis of inflammatory bowel disease. *Nature Reviews Immunology*, **3**, 521–533.

Cho JH, Nicolae DL, Gold L, Fields C, LaBuda MC, Rohal PM, Pickles M, Qin L, Fu Y, Mann JS, *et al*. (1998) Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 7502–7507.

Colhoun HM, McKeigue PM and Smith GD (2003) Problem of reporting genetic associations with complex outcomes. *Lancet*, **361**, 865–872.

Duerr RH, Barmada MM, Zhang L, Pfutzer R and Weeks DE (2000) High-density genome scan in Crohn disease shows confirmed linkage to chromosome 14q11-12. *American Journal of Human Genetics*, **66**, 1857–1862.

Girardin SE, Boneca IG, Viala J, Chamaillard M, Labigne A, Thomas G, Philpott DJ and Sansonetti PJ (2003) Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *Journal of Biological Chemistry*, **278**, 8869–8872.

Hampe J, Cuthbert A, Croucher PJ, Mirza MM, Mascheretti S, Fisher S, Frenzel H, King K, Hasselmeyer A, MacPherson AJ, *et al*. (2001) Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet*, **357**, 1925–1928.

Hampe J, Shaw SH, Saiz R, Leysens N, Lantermann A, Mascheretti S, Lynch NJ, MacPherson AJ, Bridger S, van Deventer S, *et al*. (1999a) Linkage of inflammatory bowel disease to human chromosome 6p. *American Journal of Human Genetics*, **65**, 1647–1655.

Hampe J, Schreiber S, Shaw SH, Lau KF, Bridger S, Macpherson AJ, Cardon LR, Sakul H, Harris TJ, Buckler A, *et al*. (1999b) A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. *American Journal of Human Genetics*, **64**, 808–816.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, *et al*. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.

Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, *et al*. (1996) Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature*, **379**, 821–823.

IBD International Genetics Consortium (2001) International collaboration provides convincing linkage replication in complex diseases through analysis of a large pooled data set: Crohn disease and chromosome 16. *American Journal of Human Genetics*, **68**, 1165–1171.

Inohara N, Ogura Y, Fontalba A, Gutierrez O, Pons F, Crespo J, Fukase K, Inamura S, Kusumoto S, Hashimoto M, *et al*. (2003) Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. *Journal of Biological Chemistry*, **278**, 5509–5512.

Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V, *et al*. (2002) CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *American Journal of Human Genetics*, **70**, 845–857.

Li J, Moran T, Swanson E, Julian C, Harris J, Bonen DK, Hedl M, Nicolae DL, Abraham C and Cho JH (2004) Regulation of IL-8 and IL-1β expression in Crohn's disease associated *NOD2/CARD15* mutations. *Human Molecular Genetics*, **13**, 1715–1725.

Ma Y, Ohmen JD, Li Z, Bentley LG, McElree C, Pressman S, Targan SR, Fischel-Ghodsian N, Rotter JI and Yang H (1999) A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflammatory Bowel Disease*, **5**, 271–278.

Mathew CG and Lewis CM (2004) Genetics of inflammatory bowel disease: progress and prospects. *Human Molecular Genetics*, **13**, R161–R168.

Miceli-Richard C, Lesage S, Rybojad M, Prieur AM, Manouvrier-Hanu S, Hafner R, Chamaillard M, Zouali H, Thomas G and Hugot JP (2001) *CARD15* mutations in Blau syndrome. *Nature Genetics*, **29**, 19–20.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karalluskas R, Duerr RH, *et al*. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.

Peltekova VD, Wintle RF, Rubin LA, Amos CI, Huang Q, Gu X, Newman B, Van Oene M, Cescon D, Greenberg G, *et al*. (2004) Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nature Genetics*, **36**, 471–475.

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, *et al*. (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics*, **29**, 223–228.

Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, *et al*. (2000) Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *American Journal of Human Genetics*, **66**, 1863–1870.

Satsangi J, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K, Terwilliger JD, Lathrop GM, Bell JI and Jewell DP (1996) Two stage genome-wide search in inflammatory bowel disease

provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nature Genetics*, **14**, 199–202.

Stoll M, Corneliussen B, Costello CM, Waetzig GH, Mellgard B, Koch WA, Rosenstiel P, Albrecht M, Croucher PJ, Seegert D, *et al*. (2004) Genetic variation in *DLG5* is associated with inflammatory bowel disease. *Nature Genetics*, **36**, 476–480.

van Heel DA, Dechairo BM, McGovern DP, Negoro K, Carey AH, Cardon LR, Mackay I, Jewell DP and Lench NJ (2003) The IBD6 Crohn's disease locus demonstrates complex interactions with CARD15 and IBD5 disease-associated variants. *Human Molecular Genetics*, **12**, 2569–2575.

Van Heel DA, Fisher SA, Kirby A, Daly MJ, Riouk JD and Lewis CM (2004) Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. *Human Molecular Genetics*, **13**, 763–770.

Van Heel DA, Ghosh S, Butler M, Hunt KA, Lundberg A, Ahmad T, McGovern DPB, Onnie C, Negoro K, Goldthape S, *et al*. (2005) Muramyl dipeptide and Toll-like receptor sensitivity in NOD2 associated Crohn's disease. *Lancet*, in press.

Watanabe T, Kitani A, Murray PJ and Strober W (2004) NOD2 is a negative regulator of Toll-like receptor 2-mediated T helper type 1 responses. *Nature Immunology*, **5**, 800–808.

Short Specialist Review

Hypertension genetics: under pressure

Fadi J. Charchar, Maciej Tomaszewski and Anna F. Dominiczak

University of Glasgow, Glasgow, UK

1. Introduction

Human essential hypertension (EH) is a typical example of a complex, multifactorial, and polygenic trait with a significant effect on public health. Although effective treatment is available, hypertension remains as a major risk factor for cardiovascular disease. Some mutations in genes responsible for the severe Mendelian forms of hypertension have been successfully identified (Lifton *et al*., 2001); in comparison, the search for the genes involved in EH (hypertension with unknown cause) has been less productive. The question that remains unanswered, as put forward by Luft (2004), is whether we can find such genes? The answer may lie in the methods and species that we use to discover these genes.

2. Genome-wide scans

The search for hypertension genes has mainly utilized genome screens or the candidate gene approach. The genome-wide scan is best defined as a search for quantitative trait loci across the entire genome. The quantitative trait locus (QTL) is a chromosomal region containing a gene or genes that may influence a trait of interest such as hypertension. A genome scan is designed primarily to detect causative regions without making presumptions to the physiological or pathological relevance of the genes in the region. Finding chromosomal regions containing one or more QTLs is usually the first step followed by saturation of the region with further markers or SNPs (single nucleotide polymorphisms) to grid-tighten the region (Mein *et al*., 2004) and to determine the causative gene(s). False-positive and false-negative results are, however, inevitable because the genotype–phenotype associations with QTLs are weaker than those observed with the more direct and severe Mendelian traits, and because nongenomic environmental factors can easily obscure any linkage. Additionally, even when the linkage is strong, the possibility remains that the cause of an increase or decrease in BP (blood pressure) is a linked but unrecognized genetic difference rather than the recognized polymorphic factor. Examples of this include publication of the two largest genome scans: the

Family Blood Pressure Programme (FBPP) and British Genetics of Hypertension (BRIGHT) study (Caulfield *et al*., 2003; Province *et al*., 2003). The FBPP showed no significant genome-wide linkage, except for a 2.96 LOD score on chromosome 1q for diastolic BP. In contrast, the BRIGHT study showed a locus for hypertension on chromosome 6q (LOD score of 3.21) that attained genome-wide significance and three further loci with suggestive evidence of linkage on chromosomes 2q, 5q, and 9q. The use of different markers, varying study designs, and characteristically broad linkage peaks resulting from by definition incorrect parameters in linkage analyses (the inheritance pattern of hypertension is not known) complicate the interpretation of the results of genome-wide scans.

3. Candidate gene studies

The candidate approach that associates EH with polymorphic variations in pathophysiologically relevant genes has not been successful in identification of consistent candidate(s) for further verification in functional studies. Consequently, despite high expectations, the contribution of most association studies to the understanding of hypertension and its genetic determinants has been modest. Lack of robust phenotyping, hidden population stratification or lack of relative power, has undoubtedly contributed to this. Recently, criteria have been suggested for the high-quality association studies (Ioannidis *et al*., 2001). These criteria would include really significant small *P*-values (e.g., surviving corrections for multiple testing), biologic plausibility, large sample size, functional significance, independent replication in several populations, and confirmation in family-based studies.

4. Rodent models

Analytical approaches using animal models have the potential to overcome some of the genetic and environmental complexities of human studies (Jacob and Kwitek, 2002). Inbred rat strains that display hypertension as an inherited trait have long been used as a means for identifying genes that can give rise to EH. These strains include spontaneously hypertensive rats (SHRs), stroke-prone spontaneously hypertensive rats (SHRSP), Dahl salt-sensitive rats, Sabra hypertensive rats, Milan, Lyon, fawn-hooded, and Prague hypertensive rats. To eliminate the variability arising from the heterogeneous genetic background of these animals, congenic and consomic rat strains have been developed and used to identify QTLs for hypertension. Congenic strains have a chromosomal region transferred to an inbred strain by backcrossing. Consomic strains have a single chromosome transferred to an inbred strain by backcrossing (McBride *et al*., 2004). Indeed, at least one QTL for BP has been identified on almost all chromosomes in the rat genome, confirming the complex and polygenic nature of the disease (Dominiczak *et al*., 2000). However, hundreds of genes can be present in the region encompassing the QTL(s) adding to the difficulty in identifying the causative gene(s).

While congenic mapping is an important approach, the narrowing of QTL regions to smaller than 200 kb and cloning of the gene responsible has only been achieved once (Cicila *et al*., 2001). Knockout, knockin, and transgenic models have yet to be utilized to help determine the contribution of individual genes within a congenic region to hypertension. Likewise, RNA interference is another method that can alter gene expression from a congenic region, and is yet to be used. Development of these approaches and others in a targeted systematic way will be required to understand the role of individual genes within congenic intervals. We also predict that more than one model or species will be needed to verify genes. Indeed, evolutionary studies may come to the rescue here, since real polymorphisms in genes that are highly conserved between more than one species may be the real culprits in EH. On the controversial side, it may be that the rodent model will only be truly valuable for further physiological interrogation of human candidate genes for EH (as in pharmacological studies) rather than the discovery of genes. However, findings from other diseases and lessons learned from other model organisms show that this is not the case.

5. Transcript profiles

Recent advances in molecular biology and technology have made it possible to examine the expression levels of virtually all genes (mRNA or proteins). As the tools for gene expression profiling using microarray have become more widely available, the number of investigators applying this technology in hypertension research, as in other fields of biomedical research, has grown rapidly, in particular using animal models. A combination of congenic mapping and transcription profiling was successfully used to implicate glutathione S-transferase μ-type 1 in hypertension (McBride *et al*., 2003). In humans, microarrays obviously require cells or tissues for analysis. A direct application to the genetics of human hypertension requires biopsies of samples from relevant tissues. Such approaches are technically feasible and will undoubtedly soon be applied to EH.

6. How diverse is the genetic background of hypertension?

It is clear from data published so far in both human and animal models of hypertension that hypertension is complex polygenic disorder. There is no real consensus on the number or the actual genes involved. The lower statistical power of linkage screens means that the number of human hypertension loci is likely to be much more than what was thought previously by researchers in the field (even in the order of tens of genes suggest Mein *et al*., 2004). Recently reported genome screens suggest that there are no genes with a major effect. Mein *et al*. (2004) remark that “we are looking for many genes with a genotype relative risk of 1.2–1.5”. In addition, there is still no evidence that BP genes identified in rodents will be the same as the causative genes in humans. The present challenge is to confirm linkage peaks (in both human and the rat) and identify disease predisposing variants using new resources that are becoming available, for example, the SNP haplotyping,

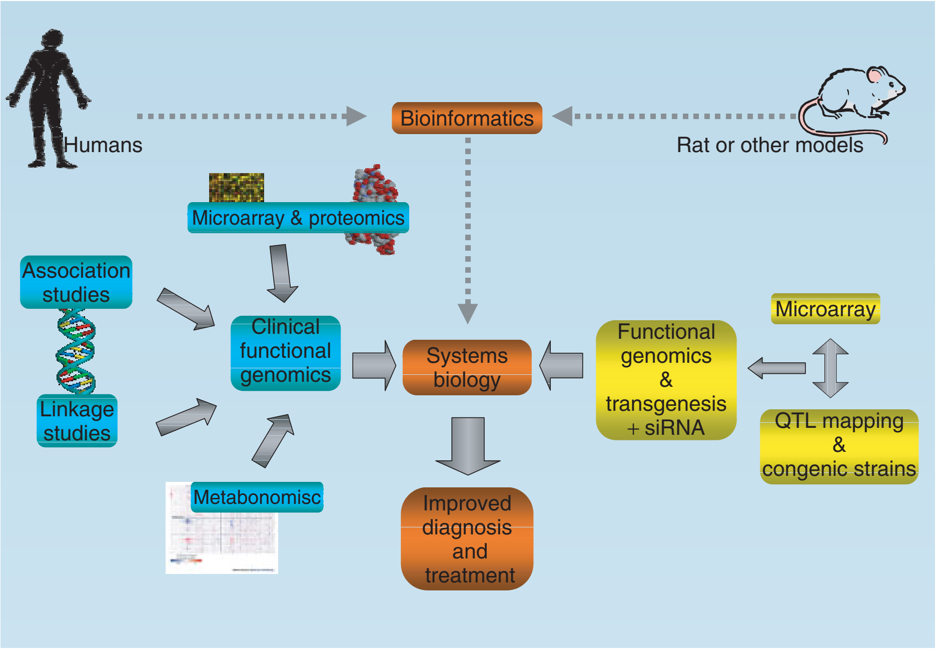


Figure 1 There is a need to improve treatment and diagnosis of EH in humans. Here we present a feasible schema on the future direction of hypertension research to meet this challenge. In humans, association and linkage studies, microarrays, proteomics, and metabonomics can be used to identify molecular targets that must be tested in both functional genomic and clinical experiments. In the animal models, the use of QTL and congenic mapping can be used to identify genes that can be tested by transgenesis and functional experiments. To identify disease predisposing variants, we will rely on comparative biology with the aid of bioinformatics between humans and animal models. Robust candidates can then be studied using systems biology in animal models. Ultimately, the development of better diagnosis and treatment from pharmacogenetics will lead to the health benefits

microarrays, bioinformatics, proteomics, and metabonomics (Figure 1). Future efforts to find associations in humans may rely upon dense collections of SNP screens to identify hypertension susceptibility loci. But these need to have enough power to detect genes with a lower relative risk. Factors that are yet to be determined fully in such studies are SNP selection, analysis methods, study size, heterogeneity in sample collections, linkage disequilibrium, and haplotype tagging. It may be too early or pessimistic to say that “genetics might never contribute to the diagnosis of EH” (Harrap, 2003) simply because its cause is too complex. It is easy to forget that our current explosion of genetic and genomic tools is only recent, and whether EH is caused by common or rare variants, the plummeting cost for these emerging technologies will have an impact in the future.

7. Future challenges

Although major advances in our understanding of the mechanism of disease development and in the treatment of EH have been achieved over the past few decades,

substantial gaps in our genetic knowledge remain. Bringing together scientists representing multiple disciplines and expertise will give us the opportunity to narrow these gaps of knowledge in the future. We predict that the postgenome era, with its ability to study functions and interactions of all the genes in the genome, including their interactions with environmental factors, will bring improved understanding of EH as a complex trait. Furthermore, we predict that the traditional genetic paradigm with its central dogma of gene to function will increasingly dominate mechanistic studies in hypertension in both animal models and humans (Figure 1). The increased prevalence of hypertensive complications, high costs of antihypertensive therapy along with its social implications in both the developed and developing world (Whitworth, 2003), puts researchers "under pressure" to dissect the genetic pathophysiology of EH. Ultimately, the success of the genetics of EH will be best measured by the progress achieved in the better understanding and treatment of EH in all patients over the next few years.

Acknowledgments

This study was funded jointly by a British Heart Foundation Programme Grant (98001) and the Wellcome Trust Cardiovascular Functional Genomics Initiative (066780).

References

Caulfield M, Munroe P, Pembroke J, Samani N, Dominiczak A, Brown M, Benjamin N, Webster J, Ratcliffe P, O'Shea S, *et al*. (2003) Genome-wide mapping of human loci for essential hypertension. *Lancet*, **361**(9375), 2118–2123.

Cicila GT, Garrett MR, Lee SJ, Liu J, Dene H and Rapp JP (2001) High-resolution mapping of the blood pressure QTL on chromosome 7 using Dahl rat congenic strains. *Genomics*, **72**(1), 51–60.

Dominiczak AF, Negrin DC, Clark JS, Brosnan MJ, McBride MW and Alexander MY (2000) Genes and hypertension: from gene mapping in experimental models to vascular gene transfer strategies. *Hypertension*, **35**(1 Pt 2), 164–172.

Harrap SB (2003) Where are all the blood-pressure genes? *Lancet*, **361**(9375), 2149–2151.

Ioannidis JP, Ntzani EE, Trikalinos TA and Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nature Genetics*, **29**(3), 306–309.

Jacob HJ and Kwitek AE (2002) Rat genetics: attaching physiology and pharmacology to the genome. *Nature Reviews. Genetics*, **3**(1), 33–42.

Lifton RP, Gharavi AG and Geller DS (2001) Molecular mechanisms of human hypertension. *Cell*, **104**(4), 545–556.

Luft FC (2004) Geneticism of essential hypertension. *Hypertension*, **43**(6), 1155–1159.

McBride MW, Carr FJ, Graham D, Anderson NH, Clark JS, Lee WK, Charchar FJ, Brosnan MJ and Dominiczak AF (2003) Microarray analysis of rat chromosome 2 congenic strains. *Hypertension*, **41**(3 Pt 2), 847–853.

McBride MW, Charchar FJ, Graham D, Miller WH, Strahorn P, Carr FJ and Dominiczak AF (2004) Functional genomics in rodent models of hypertension. *The Journal of Physiology*, **554**(Pt 1), 56–63.

Mein CA, Caulfield MJ, Dobson RJ and Munroe PB (2004) Genetics of essential hypertension. *Human Molecular Genetics*, **13**, Spec No 1, R169–R175.

Province MA, Kardia SL, Ranade K, Rao DC, Thiel BA, Cooper RS, Risch N, Turner ST, Cox DR, Hunt SC, *et al*. (2003) A meta-analysis of genome-wide linkage scans for hypertension: the National Heart, Lung and Blood Institute Family Blood Pressure Program. *American Journal of Hypertension*, **16**(2), 144–147.

Whitworth JA (2003) 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. *Journal of Hypertension*, **21**(11), 1983–1992.

Short Specialist Review

Genetics of cognitive disorders

Brett S. Abrahams and Daniel H. Geschwind
David Geffen School of Medicine, UCLA, Los Angeles, CA, USA

1. Cognition and genetic modulation

Disorders of cognition impair mental processes mediating awareness, perception, reasoning, or judgment, and consequently interfere with thinking. Although interindividual differences in human cognition are easily observed, little is known about the genetic factors underlying such phenotypic variation. Phenotypes associated with cognition range from what are currently considered the more measurable, such as language and memory, to spatial and social abilities, which have been less aptly defined. Progress in understanding the genetics of cognition and cognitive disorders necessarily relies on progress in understanding brain-behavior relationships, a discipline that is still in its infancy. We are likely to gain significant insight into the molecular basis of normal variation in human performance from the study of Mendelian and complex genetic disorders of cognition.

2. Brain structure differences are heritable

One important factor to consider is that genetic factors influence cognitive phenotypes only indirectly, through modulation of brain structure and function. At the same time, heritability for differences in brain structure is high (Thompson *et al*., 2001; Pfefferbaum *et al*., 2004), with estimates ranging from 0.65 to 0.95 (Thompson *et al*., 2001; Geschwind *et al*., 2002). Heritability for differences in brain structure is similar to that for common neuropsychiatric disorders involving the development of cognition and behavior, including dyslexia (0.4–0.6; DeFries and Alarcon, 1996; Wadsworth *et al*., 2000), attention-deficit hyperactivity disorder (0.6–0.9; Gjone *et al*., 1996; Levy *et al*., 1997), and autism (0.6–0.9; Folstein and Rutter, 1988; Bailey *et al*., 1995). A more complete understanding of the alterations in brain structure that underlie these disorders will provide important phenotypes that are directly amenable to genetic analyses. One example of recent progress in our understanding of genetic modulation of brain structure comes from the work linking a common variant within brain-derived neurotrophic factor (BDNF) to both differences in hippocampal and prefrontal gray matter volumes, as well as performance on an episodic memory task in normal subjects (Egan *et al*., 2003; Hariri *et al*., 2003; Pezawas *et al*., 2004). It will likely be worthwhile to explore the extent to which variation at *BDNF* influences diseases of memory.

3. Spearman's "g": a unitary intelligence?

Spearman's "g" parallels the concept of IQ and attempts to capture the finding that individuals who do well on some types of cognitive tests generally do well on others. Twin data suggest that heritability of g varies as a function of age, with estimates of 40% in childhood and upward of 60% later on in life (McClearn *et al*., 1997; Plomin and Craig, 1997; Alarcon *et al*., 1998; Alarcon *et al*., 2005). Although initially a purely theoretical construct, recent work has begun to define the structures and circuitry that may underlie interindividual differences in general cognitive performance as measured by standardized tests. Such work suggests that structural variation within the frontal cortex, specifically the dorsolateral prefrontal cortex and medial frontal gyrus, are important modulators of g-related tasks (Duncan *et al*., 2000; Thompson *et al*., 2001). The involvement of these well-studied frontal regions suggests that g may be measuring some form of executive function and working memory, rather than general intelligence. That g is heritable is likely to provide one window, albeit imprecise from the standpoint of cognitive neuroscience, to link brain, genes, and behavior.

4. Disease genes and normal variation in performance

A large body of work now demonstrates that loss-of-function mutations in a variety of brain-expressed genes can give rise to profound and relatively generalized cognitive dysfunction (Ross and Walsh, 2001). Less clear, however, is whether differences in cognition between normal individuals may be due to common variants within these same genes (Plomin and Rutter, 1998), one paradigm driving research in this area. Similarly, it remains to be seen whether genetic variation underlying disease-related endophenotypes may help explain corresponding phenotypic differences within the normal population. To some extent, both will depend on whether particular clinical entities are qualitatively different from the corresponding "normal state" or, alternatively, an extreme within the normal distribution. At present, neuropsychiatric disease status is based on practical definitions of function within society, representing an arbitrary cutoff, rather than a qualitatively distinct state. Finding ways to identify and measure heritable endophenotypes rather than using clinically defined disease status, which may have little direct relationship to genetic factors, represents one of the most promising avenues for defining the genetic basis of cognitive disorders.

5. Autism and broader phenotype definitions

Along these lines, the study of individuals with autism may increase our understanding of several different aspects of cognition and behavior, as it is defined by deficits in three major domains: language, social interaction, and repetitive restrictive behavior (Folstein and Rosen-Sheidley, 2001). Although some monogenic syndromes (Tuberous Sclerosis, Fragile X, Joubert, and Rett Syndrome) show phenotypic overlap or increased co-occurrence with autism (Holroyd *et al*., 1991; Feinstein and Reiss, 1998; Ozonoff *et al*., 1999; Mount *et al*., 2003; Wiznitzer, 2004),

Table 1 Gene/locus findings for language-related disorders including autism, dyslexia, and specific language impairment

Disorder	(Endo)phenotypes considered[a]	Locus[b] (References)
Autism	Delayed onset of phrase speech	2q (Buxbaum *et al*., 2001; Shao *et al*., 2002)
	Autism-spectrum disorders	3q25-27 (Auranen *et al*., 2002; Auranen *et al*., 2003)
	Autism diagnosis only	7q (IMGSA-Consortium, 1998; IMGSA-Consortium, 2001)
	Language, repetitive behaviors	7q (Alarcon *et al*., 2002; Alarcon *et al*., 2005)
	Autism diagnosis only	7q36 – *Engrailed 2* gene (Petit *et al*., 1995; Gharani *et al*., 2004; also see Zhong *et al*., 2003)
	Autism diagnosis only	15q11-q13 (Baker *et al*., 1994; Cook *et al*., 1997; Cook *et al*., 1998; Schroer *et al*., 1998; Nurmi *et al*., 2003; also see Salmon *et al*., 1999)
	Insistence on sameness	15q11-q13 (Shao *et al*., 2002)
	Autism diagnosis only	15q11-q13 – *GABA* receptor genes (Martin *et al*., 2000; Menold *et al*., 2001; Buxbaum *et al*., 2002; also see Maestrini *et al*., 1999)
	Autism diagnosis only	Xp22 – *Neuroligin4* gene – (Thomas *et al*., 1999; Jamain *et al*., 2003; also see Gauthier *et al*., 2004; Laumonnier *et al*., 2004; Vincent *et al*., 2004)
	Autism diagnosis only	Xq13 – *Neuroligin3* gene – (Jamain *et al*., 2003; also see Gauthier *et al*., 2004; Vincent *et al*., 2004)
Dyslexia	Dyslexia diagnosis only	1p34-36 (Rabin *et al*., 1993; Grigorenko *et al*., 2001)
	Spelling performance	1p34-36 (Tzenova *et al*., 2004)
	Dyslexia diagnosis only	2p11 (Kaminen *et al*., 2003; Peyrard-Janvid *et al*., 2004)
	Dyslexia diagnosis only	2p15-16 (Fagerheim *et al*., 1999; Petryshen *et al*., 2002)
	Spelling performance	2p15-16 (Petryshen *et al*., 2002)
	Reading disability measures	6p21 (Cardon *et al*., 1994; Grigorenko *et al*., 1997; Fisher *et al*., 1999; Gayan *et al*., 1999; Grigorenko *et al*., 2000; Kaplan *et al*., 2002; Turic *et al*., 2003)
	Dyslexia diagnosis only	15q21 – *DYX1C1* gene – (Taipale *et al*., 2003; also see Scerri *et al*., 2004)
	Reading disability measures	18p11.2 (Fisher *et al*., 2002)
Other language impairment	Specific language impairment (language and reading scores)	16q (SLI-Consortium, 2002; SLI-Consortium, 2004)
	Specific language impairment (low standard language scores)	19q (SLI-Consortium, 2002; SLI-Consortium, 2004)
	Speech and language disorder with orofacial dyspraxia	7q31 – *Forkhead 2* gene (Fisher *et al*., 1998; Lai *et al*., 2001; Liegeois *et al*., 2003; see O'Brien *et al*., 2003 for evidence suggesting region may also be involved in specific language impairment)

[a]Endophenotypes considered but without significant linkage/association are not included.
[b]Obtained by linkage or association.

the genetics of idiopathic, nonsyndromic autism are complex and heterogeneous (see Table 1). Quantitative analysis of language performance amongst patients points to the presence of modulation of language performance by specific genetic factors (Alarcon *et al*., 2002; Alarcon *et al*., 2005). Further support for the use of quantitative methods for analysis of abnormal social interaction in autistic subjects

comes from data that suggests that such behavior may simply represent an extreme within the normal range of phenotypic variation in social cognition and behavior (Constantino and Todd, 2003). Despite some work suggesting reduced heterogeneity after clustering by repetitive behavior presentation (Buxbaum *et al*., 2004), it is not yet clear whether such behaviors in patients similarly represent part of the normal continuum, and if so, whether when observed in isolation such abnormalities might present as obsessive compulsive disorder. The extent to which domain-specific measures, as compared to a categorical diagnosis of autism, may increase power for gene identification remains to be determined, but appears promising. Following the observation that males are at increased risk for autism, stratification of families into those containing males only has also proven to be a powerful way to reduce heterogeneity (Stone *et al*., 2004; Cantor *et al*., 2005). Such an approach may be broadly applicable to other neurodevelopmental disorders with sex-dependent differences in incidence. Despite the clear utility of quantitative approaches and population stratification, important advances are also being made from the joint consideration of broadly related, yet clinically distinct, autism-spectrum disorders (Auranen *et al*., 2002; Auranen *et al*., 2003). Linkage to 3q25-27, after scoring family members with any of infantile autism, Asperger syndrome, or developmental dysphasia as affected, suggests that at least a subset of phenotypes shared amongst autism-spectrum disorders may possess a common genetic etiology. This again highlights the limited utility of categorical clinical definitions of complex neuropsychiatric disease in genetic studies. All of these methodologies will likely profit from an increased understanding of the genetic basis of developmental factors fundamental to human cognition and language, such as brain asymmetry (Geschwind, 1979; Geschwind and Miller, 2001; Geschwind *et al*., 2002; Herbert *et al*., 2004; Herbert *et al*., 2005). While some of these genetic influences may lie on the sex chromosomes (Skuse *et al*., 1997), most are likely to be autosomal (Stone *et al*., 2004).

6. Dissociating cognitive and behavioral abnormalities

Cognition and behavior can be strongly linked even within single-gene disorders. A single gene may effect the development of multiple brain structures, a single brain structure may be involved in several distinct cognitive and behavioral functions, or a single gene may affect a process, such as neurotransmission, that is widely distributed in different brain circuits that serve distinct functions. A good example of this linkage is observed in a rare Dutch family in which males hemizygous for a null allele of the X-linked Monoamine Oxidase A (*MAOA*) gene show cognitive impairment and violent behavior (Brunner *et al*., 1993). Although converging lines of evidence (Cases *et al*., 1995; Caspi *et al*., 2002) strongly support the idea that allelic variation at *MAOA* does in fact modulate behavior, it is not yet clear whether the extent to which the behavior abnormalities are secondary to cognitive dysfunction, or simply reflect dysfunction in multiple distributed brain circuits. At the same time, the existence of disorders in which cognitive processes are impaired but behavior remains intact (Fisher and DeFries, 2002; Good *et al*., 2003) demonstrates that certain elements of behavior and cognition can be clearly dissociated.

In a parallel vein, it is important to consider that although "mental retardation" or "generalized cognitive dysfunction" may be useful clinical conceptualizations, each is insufficiently precise for a detailed understanding of how genetic variation influences brain structure, and secondarily, cognitive function. Salient examples of this fact are found in the myriad of mental retardation syndromes, in which subjects have the same low IQ, but show dramatic differences in specific cognitive or behavioral domains (Bellugi *et al*., 1999; Folstein and Rosen-Sheidley, 2001; Whittington *et al*., 2004). Similarly, cognitive impairment in some domains may accompany real strengths in others, as is the case in certain neurodevelopmental syndromes (Plaisted *et al*., 1998; Happe, 1999; Heaton, 2003), dementias (Miller *et al*., 2000; Miller and Hou, 2004), or more circumscribed impairments such as developmental dyslexia (Nation, 1999; Richman and Wood, 2002). These cases emphasize the need for broad and detailed neurocognitive and behavioral assessment to carefully delineate specific phenotypes that comprise different cognitive disorders.

7. Successes from the study of language disorders

Disorders of language represent one area where cognitive deficits are typically distinct from disturbances in behavior. Additionally, language is a uniquely human cognitive ability that while genetically complex, can be broken down into underlying cognitive and linguistic components. Also, because a good deal is known about the structural basis of language in the brain (Galaburda *et al*., 1978; Geschwind, 1979), the interpretation of how allelic variation modulates specific structure-function relationships is facilitated. Independent genome-wide scans for developmental dyslexia, or specific reading impairment, has highlighted multiple genomic loci (see Table 1). Similarly, the study of specific language impairment, defined as difficulty in the acquisition of language skills amongst individuals with normal intelligence and access to education, supports the involvement of at least two loci (SLI-Consortium, 2002; SLI-Consortium, 2004). While no gene for either dyslexia or specific language impairment has been conclusively identified, a translocation disrupting *DYX1C1* segregated with dyslexia in one family, and two common variants were overrepresented amongst other affected individuals (Taipale *et al*., 2003). However, the importance of this gene in other dyslexic populations remains unclear (Scerri *et al*., 2004), highlighting the difficulties of replication in complex and heterogeneous disorders of cognition (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2 and Article 60, **Population selection in complex disease gene mapping**, Volume 2).

The only gene clearly linked to human speech and language is *FOXP2,* identified by the study of a family segregating a rare Mendelian disorder involving speech dyspraxia, language and grammatical impairment, and other aspects of cognitive dysfunction (Lai *et al*., 2001). Interestingly, subsequent evolutionary analysis of *FOXP2* shows positive selection in humans, supporting the notion that it has been adapted for language-related cognitive specialization in humans (Enard *et al*., 2002a). At the same time, conserved *FOXP2* expression between birds and humans may suggest a more general role in vocal-motor learning (Teramitsu *et al*., 2004),

emphasizing the substantial complexity of even single-gene disorders of cognition and the utility of animal models and evolutionary comparisons. Other monogenic syndromes that are associated with language impairment, for example, Klinefelter Syndrome, Floating-Harbor Syndrome, and Neurofibromatosis type 1 (Robinson *et al*., 1988; Hofman *et al*., 1994; Geschwind *et al*., 2000), may also prove useful in our understanding of language development and dysfunction.

8. Promise for the future

As intriguing as all these data are, few would argue that they represent only a glimpse of what is likely to be uncovered in the next 20 years. Widespread application of emerging genomic approaches (Karsten and Geschwind, 2002; Geschwind, 2003), and the integration of phenotype data with copy number polymorphisms (Sebat *et al*., 2004) and single-nucleotide variation (Sachidanandam *et al*., 2001; Gabriel *et al*., 2002) will provide a more comprehensive understanding of the molecular biology underlying the regulation of human cognition. Efforts to limit false-positives (Ioannidis *et al*., 2001), and at the same time allow for multilocus, gene–gene and gene–environment interactions will be valuable. Because neither the human genome nor human environment is amenable to experimental manipulation, promising associations must be followed up in experimental systems. Although some aspects of cognition may be difficult to translate to the laboratory, many others will benefit from cellular or in vivo assays. Interspecies comparisons, particularly those calling on multiple primate species (Enard *et al*., 2002b; Caceres *et al*., 2003; Khaitovich *et al*., 2004; Preuss *et al*., 2004), or in the case of speech, vocal learners such as the songbird (Teramitsu *et al*., 2004) will add value to these endeavors. Such work will not only provide important insights into human disease but may also provide empirical answers to fundamental and long-standing philosophical questions that motivate much of the research in this area.

Acknowledgments

Work on autism and the genetics of human cognitive specializations in the Geschwind laboratory is supported by grants from the NIH (NIMH and NINDS) and the James S. McDonnell Foundation. We gratefully acknowledge numerous insights we have gained from a number of ongoing collaborations and our collaborators.

References

Alarcon M, Plomin R, Fulker DW, Corley R and DeFries JC (1998) Multivariate path analysis of specific cognitive abilities data at 12 years of age in the Colorado adoption project. *Behavior Genetics*, **28**, 255–264.

Alarcon M, Cantor RM, Liu J, Gilliam TC and Geschwind DH (2002) Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *American Journal of Human Genetics*, **70**, 60–71.

Alarcon M, Yonan AL, Gilliam TC, Cantor RM and Geschwind DH (2005) Quantitative genome scan and Ordered-Subsets Analysis of autism endophenotypes support language QTLs. *Molecular Psychiatry*, Advance Online Publication: April 12, 2005, doi:10.1038/sj.mp.4001666, PMID: 15824743.

Auranen M, Vanhala R, Varilo T, Ayers K, Kempas E, Ylisaukko-Oja T, Sinsheimer JS, Peltonen L and Jarvela I (2002) A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. *American Journal of Human Genetics*, **71**, 777–790.

Auranen M, Varilo T, Alen R, Vanhala R, Ayers K, Kempas E, Ylisaukko-Oja T, Peltonen L and Jarvela I (2003) Evidence for allelic association on chromosome 3q25-27 in families with autism spectrum disorders originating from a subisolate of Finland. *Molecular Psychiatry*, **8**, 879–884.

Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E and Rutter M (1995) Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, **25**, 63–77.

Baker P, Piven J, Schwartz S and Patil S (1994) Brief report: duplication of chromosome 15q11-13 in two individuals with autistic disorder. *Journal of Autism and Developmental Disorders*, **24**, 529–535.

Bellugi U, Lichtenberger L, Mills D, Galaburda A and Korenberg JR (1999) Bridging cognition, the brain and molecular genetics: evidence from Williams syndrome. *Trends Neurosci*, **22**, 197–207.

Brunner HG, Nelen M, Breakefield XO, Ropers HH and van Oost BA (1993) Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science*, **262**, 578–580.

Buxbaum JD, Silverman J, Keddache M, Smith CJ, Hollander E, Ramoz N and Reichert JG (2004) Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. *Molecular Psychiatry*, **9**, 144–150.

Buxbaum JD, Silverman JM, Smith CJ, Kilifarski M, Reichert J, Hollander E, Lawlor BA, Fitzgerald M, Greenberg DA and Davis KL (2001) Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *American Journal of Human Genetics*, **68**, 1514–1520.

Buxbaum JD, Silverman JM, Smith CJ, Greenberg DA, Kilifarski M, Reichert J, Cook EH Jr, Fang Y, Song CY and Vitale R (2002) Association between a GABRB3 polymorphism and autism. *Molecular Psychiatry*, **7**, 311–316.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM and Barlow C (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13030–13035.

Cantor RM, Kono N, Duvall JA, Alvarez-Retuerto A, Stone JL, Alarcon M, Nelson SF and Geschwind DH (2005) Replication of Autism Linkage: Fine-Mapping Peak at 17q21. *American Journal of Human Genetics*, Advance Online Publication: April 1, 2005, **76**(6), PMID: 15806440.

Cardon LR, Smith SD, Fulker DW, Kimberling WJ, Pennington BF and DeFries JC (1994) Quantitative trait locus for reading disability on chromosome 6. *Science*, **266**, 276–279.

Cases O, Seif I, Grimsby J, Gaspar P, Chen K, Pournin S, Müller U, Aguet M, Babinet C, Shih JC, *et al*. (1995) Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Science*, **268**, 1763–1766.

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, Taylor A and Poulton R (2002) Role of genotype in the cycle of violence in maltreated children. *Science*, **297**, 851–854.

Constantino JN and Todd RD (2003) Autistic traits in the general population: a twin study. *Archives of General Psychiatry*, **60**, 524–530.

Cook EH Jr, Lindgren V, Leventhal BL, Courchesne R, Lincoln A, Shulman C, Lord C and Courchesne E (1997) Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *American Journal of Human Genetics*, **60**, 928–934.

Cook EH Jr, Courchesne RY, Cox NJ, Lord C, Gonen D, Guter SJ, Lincoln A, Nix K, Haas R, Leventhal BL, *et al*. (1998) Linkage-disequilibrium mapping of autistic disorder, with 15q11-13 markers. *American Journal of Human Genetics*, **62**, 1077–1083.

DeFries JC and Alarcon M (1996) Genetics of specific reading disability. *Mental Retard Dev Disabil Res Rev*, **2**, 39–47.

Duncan J, Seitz RJ, Kolodny J, Bor D, Herzog H, Ahmed A, Newell FN and Emslie H (2000) A neural basis for general intelligence. *Science*, **289**, 457–460.

Egan MF, Kojima M, Callicott JH, Goldberg TE, Kolachana BS, Bertolino A, Zaitsev E, Gold B, Goldman D, Dean M, *et al*. (2003) The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell*, **112**, 257–269.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP and Paabo S (2002a) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, *et al*. (2002b) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.

Fagerheim T, Raeymaekers P, Tonnessen FE, Pedersen M, Tranebjaerg L and Lubs HA (1999) A new gene (DYX3) for dyslexia is located on chromosome 2. *Journal of Medical Genetics*, **36**, 664–669.

Feinstein C and Reiss AL (1998) Autism: the point of view from fragile X studies. *Journal of Autism and Developmental Disorders*, **28**, 393–405.

Fisher SE and DeFries JC (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nature Reviews Neuroscience*, **3**, 767–780.

Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP and Pembrey ME (1998) Localisation of a gene implicated in a severe speech and language disorder. *Nature Genetics*, **18**, 168–170.

Fisher SE, Marlow AJ, Lamb J, Maestrini E, Williams DF, Richardson AJ, Weeks DE, Stein JF and Monaco AP (1999) A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics*, **64**, 146–156.

Fisher SE, Francks C, Marlow AJ, MacPhie IL, Newbury DF, Cardon LR, Ishikawa-Brush Y, Richardson AJ, Talcott JB, Gayan J, *et al*. (2002) Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia. *Nature Genetics*, **30**, 86–91.

Folstein SE and Rutter ML (1988) Autism: familial aggregation and genetic implications. *Journal of Autism and Developmental Disorders*, **18**, 3–30.

Folstein SE and Rosen-Sheidley B (2001) Genetics of autism: complex aetiology for a heterogeneous disorder. *Nature Reviews Genetics*, **2**, 943–955.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Galaburda AM, LeMay M, Kemper TL and Geschwind N (1978) Right-left asymmetrics in the brain. *Science*, **199**, 852–856.

Gauthier J, Bonnel A, St-Onge J, Karemera L, Laurent S, Mottron L, Fombonne E, Joober R and Rouleau GA (2004) NLGN3/NLGN4 gene mutations are not responsible for autism in the Quebec population. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, **132B**, 74–75.

Gayan J, Smith SD, Cherny SS, Cardon LR, Fulker DW, Brower AM, Olson RK, Pennington BF and DeFries JC (1999) Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics*, **64**, 157–164.

Geschwind DH (2003) DNA microarrays: translation of the genome from laboratory to clinic. *Lancet Neurology*, **2**, 275–282.

Geschwind DH and Miller BL (2001) Molecular approaches to cerebral laterality: development and neurodegeneration. *American Journal of Medical Genetics*, **101**, 370–381.

Geschwind DH, Boone KB, Miller BL and Swerdloff RS (2000) Neurobehavioral phenotype of klinefelter syndrome. *Mental Retardation and Developmental Disabilities Research Reviews*, **6**, 107–116.

Geschwind DH, Miller BL, DeCarli C and Carmelli D (2002) Heritability of lobar brain volumes in twins supports genetic models of cerebral laterality and handedness. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3176–3181.

Geschwind N (1979) Specializations of the human brain. *Scientific American*, **241**, 180–199.

Gharani N, Benayed R, Mancuso V, Brzustowicz LM and Millonig JH (2004) Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. *Molecular Psychiatry*, **9**, 474–484.

Gjone H, Stevenson J and Sundet JM (1996) Genetic influence on parent-reported attention-related problems in a Norwegian general population twin sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, **35**, 588–596.

Good CD, Lawrence K, Thomas NS, Price CJ, Ashburner J, Friston KJ, Frackowiak RS, Oreland L and Skuse DH (2003) Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. *Brain*, **126**, 2431–2446.

Grigorenko EL, Wood FB, Meyer MS and Pauls DL (2000) Chromosome 6p influences on different dyslexia-related cognitive processes: further confirmation. *American Journal of Human Genetics*, **66**, 715–723.

Grigorenko EL, Wood FB, Meyer MS, Pauls JE, Hart LA and Pauls DL (2001) Linkage studies suggest a possible locus for developmental dyslexia on chromosome 1p. *American Journal of Human Genetics*, **105**, 120–129.

Grigorenko EL, Wood FB, Meyer MS, Hart LA, Speed WC, Shuster A and Pauls DL (1997) Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Human Genetics*, **60**, 27–39.

Happe F (1999) Autism: cognitive deficit or cognitive style? *Trends in Cognitive Sciences*, **3**, 216–222.

Hariri AR, Goldberg TE, Mattay VS, Kolachana BS, Callicott JH, Egan MF and Weinberger DR (2003) Brain-derived neurotrophic factor val66met polymorphism affects human memory-related hippocampal activity and predicts memory performance. *The Journal of Neuroscience*, **23**, 6690–6694.

Heaton P (2003) Pitch memory, labelling and disembedding in autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, **44**, 543–551.

Herbert MR, Ziegler DA, Makris N, Filipek PA, Kemper TL, Normandin JJ, Sanders HA, Kennedy DN and Caviness VS Jr (2004) Localization of white matter volume increase in autism and developmental language disorder. *Annals of Neurology*, **55**, 530–540.

Herbert MR, Ziegler DA, Deutsch CK, O'Brien LM, Kennedy DN, Filipek PA, Bakardjiev AI, Hodgson J, Takeoka M, Makris N, *et al*. (2005) Brain asymmetries in autism and developmental language disorder: a nested whole-brain analysis. *Brain*, **128**, 213–226.

Hofman KJ, Harris EL, Bryan RN and Denckla MB (1994) Neurofibromatosis type 1: the cognitive phenotype. *The Journal of Pediatrics*, **124**, S1–S8.

Holroyd S, Reiss AL and Bryan RN (1991) Autistic features in Joubert syndrome: a genetic disorder with agenesis of the cerebellar vermis. *Biological Psychiatry*, **29**, 287–294.

IMGSA-Consortium (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Human Molecular Genetics*, **7**, 571–578.

IMGSA-Consortium (2001) Further characterization of the autism susceptibility locus AUTS1 on chromosome 7q. *Human Molecular Genetics*, **10**, 973–982.

Ioannidis JP, Ntzani EE, Trikalinos TA and Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nature Genetics*, **29**, 306–309.

Jamain S, Quach H, Betancur C, Rastam M, Colineaux C, Gillberg IC, Soderstrom H, Giros B, Leboyer M, Gillberg C, *et al*. (2003) Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature Genetics*, **34**, 27–29.

Kaminen N, Hannula-Jouppi K, Kestila M, Lahermo P, Muller K, Kaaranen M, Myllyluoma B, Voutilainen A, Lyytinen H, Nopola-Hemmi J, *et al*. (2003) A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. *Journal of Medical Genetics*, **40**, 340–345.

Kaplan DE, Gayan J, Ahn J, Won TW, Pauls D, Olson RK, DeFries JC, Wood F, Pennington BF, Page GP, *et al*. (2002) Evidence for linkage and association with reading disability on 6p21.3-22. *American Journal of Human Genetics*, **70**, 1287–1298.

Karsten S Geschwind D (2002) Gene expression analysis using cDNA microarrays. In *Current Protocols in Neuroscience*, Supplement 20, Section (4), Unit 4.28, Crawley J, Gerfen C, Rogawski M, Sibley D, Skolnick P, Wray S, (eds.); John Wiley & Sons: New York.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, *et al*. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Research*, **14**, 1462–1473.

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F and Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, **413**, 519–523.

Laumonnier F, Bonnet-Brilhault F, Gomot M, Blanc R, David A, Moizard MP, Raynaud M, Ronce N, Lemonnier E, Calvas P, *et al*. (2004) X-linked mental retardation and autism are associated with a mutation in the NLGN4 gene, a member of the neuroligin family. *American Journal of Human Genetics*, **74**, 552–557.

Levy F, Hay DA, McStephen M, Wood C and Waldman I (1997) Attention-deficit hyperactivity disorder: a category or a continuum? Genetic analysis of a large-scale twin study. *Journal of the American Academy of Child and Adolescent Psychiatry*, **36**, 737–744.

Liegeois F, Baldeweg T, Connelly A, Gadian DG, Mishkin M and Vargha-Khadem F (2003) Language fMRI abnormalities associated with FOXP2 gene mutation. *Nature Neuroscience*, **6**, 1230–1237.

Maestrini E, Lai C, Marlow A, Matthews N, Wallace S, Bailey A, Cook EH, Weeks DE and Monaco AP (1999) Serotonin transporter (5-HTT) and gamma-aminobutyric acid receptor subunit beta3 (GABRB3) gene polymorphisms are not associated with autism in the IMGSA families. The International Molecular Genetic Study of Autism Consortium. *American Journal of Medical Genetics*, **88**, 492–496.

Martin ER, Menold MM, Wolpert CM, Bass MP, Donnelly SL, Ravan SA, Zimmerman A, Gilbert JR, Vance JM, Maddox LO, *et al*. (2000) Analysis of linkage disequilibrium in gamma-aminobutyric acid receptor subunit genes in autistic disorder. *American Journal of Medical Genetics*, **96**, 43–48.

McClearn GE, Johansson B, Berg S, Pedersen NL, Ahern F, Petrill SA and Plomin R (1997) Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science*, **276**, 1560–1563.

Menold MM, Shao Y, Wolpert CM, Donnelly SL, Raiford KL, Martin ER, Ravan SA, Abramson RK, Wright HH, Delong GR, *et al*. (2001) Association analysis of chromosome 15 gabaa receptor subunit genes in autistic disorder. *Journal of Neurogenetics*, **15**, 245–259.

Miller BL and Hou CE (2004) Portraits of artists: emergence of visual creativity in dementia. *Archives of Neurology*, **61**, 842–844.

Miller BL, Boone K, Cummings JL, Read SL and Mishkin F (2000) Functional correlates of musical and visual ability in frontotemporal dementia. *The British Journal of Psychiatry*, **176**, 458–463.

Mount RH, Charman T, Hastings RP, Reilly S and Cass H (2003) Features of autism in Rett syndrome and severe mental retardation. *Journal of Autism and Developmental Disorders*, **33**, 435–442.

Nation K (1999) Reading skills in hyperlexia: a developmental perspective. *Psychological Bulletin*, **125**, 338–355.

Nurmi EL, Amin T, Olson LM, Jacobs MM, McCauley JL, Lam AY, Organ EL, Folstein SE, Haines JL and Sutcliffe JS (2003) Dense linkage disequilibrium mapping in the 15q11-q13 maternal expression domain yields evidence for association in autism. *Molecular Psychiatry*, **8**, 570, 624–634,

O'Brien EK, Zhang X, Nishimura C, Tomblin JB and Murray JC (2003) Association of specific language impairment (SLI) to the region of 7q31. *American Journal of Human Genetics*, **72**, 1536–1543.

Ozonoff S, Williams BJ, Gale S and Miller JN (1999) Autism and autistic behavior in Joubert syndrome. *Journal of Child Neurology*, **14**, 636–641.

Petit E, Herault J, Martineau J, Perrot A, Barthelemy C, Hameury L, Sauvage D, Lelord G and Muh JP (1995) Association study with two markers of a human homeogene in infantile autism. *Journal of Medical Genetics*, **32**, 269–274.

Petryshen TL, Kaplan BJ, Hughes ML, Tzenova J and Field LL (2002) Supportive evidence for the DYX3 dyslexia susceptibility gene in Canadian families. *Journal of Medical Genetics*, **39**, 125–126.

Peyrard-Janvid M, Anthoni H, Onkamo P, Lahermo P, Zucchelli M, Kaminen N, Hannula-Jouppi K, Nopola-Hemmi J, Voutilainen A, Lyytinen H, *et al*. (2004) Fine mapping of the 2p11 dyslexia locus and exclusion of TACR1 as a candidate gene. *Human Genetics*, **114**, 510–516.

Pezawas L, Verchinski BA, Mattay VS, Callicott JH, Kolachana BS, Straub RE, Egan MF, Meyer-Lindenberg A and Weinberger DR (2004) The brain-derived neurotrophic factor val66met polymorphism and variation in human cortical morphology. *The Journal of Neuroscience*, **24**, 10099–10102.

Pfefferbaum A, Sullivan EV and Carmelli D (2004) Morphological changes in aging brain structures are differentially affected by time-linked environmental influences despite strong genetic stability. *Neurobiology of Aging*, **25**, 175–183.

Plaisted K, O'Riordan M and Baron-Cohen S (1998) Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual learning task. *Journal of Child Psychology and Psychiatry*, **39**, 765–775.

Plomin R and Craig I (1997) Human behavioural genetics of cognitive abilities and disabilities. *Bioessays*, **19**, 1117–1124.

Plomin R and Rutter M (1998) Child development, molecular genetics, and what to do with genes once they are found. *Child Development*, **69**, 1223–1242.

Preuss TM, Caceres M, Oldham MC and Geschwind DH (2004) Human brain evolution: insights from microarrays. *Nature Reviews Genetics*, **5**, 850–860.

Rabin M, Wen XL, Hepburn M, Lubs HA, Feldman E and Duara R (1993) Suggestive linkage of developmental dyslexia to chromosome 1p34-p36. *Lancet*, **342**, 178.

Richman LC and Wood KM (2002) Learning disability subtypes: classification of high functioning hyperlexia. *Brain and Language*, **82**, 10–21.

Robinson PL, Shohat M, Winter RM, Conte WJ, Gordon-Nesbitt D, Feingold M, Laron Z and Rimoin DL (1988) A unique association of short stature, dysmorphic features, and speech impairment (Floating-Harbor syndrome). *The Journal of Pediatrics*, **113**, 703–706.

Ross ME and Walsh CA (2001) Human brain malformations and their lessons for neuronal migration. *Annual Review of Neuroscience*, **24**, 1041–1070.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Salmon B, Hallmayer J, Rogers T, Kalaydjieva L, Petersen PB, Nicholas P, Pingree C, McMahon W, Spiker D, Lotspeich L, *et al*. (1999) Absence of linkage and linkage disequilibrium to chromosome 15q11-q13 markers in 139 multiplex families with autism. *American Journal of Medical Genetics*, **88**, 551–556.

Scerri TS, Fisher SE, Francks C, MacPhie IL, Paracchini S, Richardson AJ, Stein JF and Monaco AP (2004) Putative functional alleles of DYX1C1 are not associated with dyslexia susceptibility in a large sample of sibling pairs from the UK. *Journal of Medical Genetics*, **41**, 853–857.

Schroer RJ, Phelan MC, Michaelis RC, Crawford EC, Skinner SA, Cuccaro M, Simensen RJ, Bishop J, Skinner C, Fender D, *et al*. (1998) Autism and maternally derived aberrations of chromosome 15q. *American Journal of Medical Genetics*, **76**, 327–336.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shao Y, Raiford KL, Wolpert CM, Cope HA, Ravan SA, Ashley-Koch AA, Abramson RK, Wright HH, DeLong RG, Gilbert JR, *et al*. (2002) Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *American Journal of Human Genetics*, **70**, 1058–1061.

Skuse DH, James RS, Bishop DV, Coppin B, Dalton P, Aamodt-Leeper G, Bacarese-Hamilton M, Creswell C, McGurk R and Jacobs PA (1997) Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. *Nature*, **387**, 705–708.

SLI-Consortium (2002) A genomewide scan identifies two novel loci involved in specific language impairment. *American Journal of Human Genetics*, **70**, 384–398.

SLI-Consortium (2004) Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *American Journal of Human Genetics*, **74**, 1225–1238.

Stone JL, Merriman B, Cantor RM, Yonan AL, Gilliam TC, Geschwind DH and Nelson SF (2004) Evidence for sex-specific risk alleles in autism spectrum disorder. *American Journal of Human Genetics*, **75**, 1117–1123.

Taipale M, Kaminen N, Nopola-Hemmi J, Haltia T, Myllyluoma B, Lyytinen H, Muller K, Kaaranen M, Lindsberg PJ, Hannula-Jouppi K, *et al*. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11553–11558.

Teramitsu I, Kudo LC, London SE, Geschwind DH and White SA (2004) Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. *The Journal of Neuroscience*, **24**, 3152–3163.

Thomas NS, Sharp AJ, Browne CE, Skuse D, Hardie C and Dennis NR (1999) Xp deletions associated with autism in three females. *Human Genetics*, **104**, 43–48.

Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M, Lonnqvist J, Standertskjold-Nordenstam CG, Kaprio J, Khaledy M, *et al*. (2001) Genetic influences on brain structure. *Nature Neuroscience*, **4**, 1253–1258.

Turic D, Robinson L, Duke M, Morris DW, Webb V, Hamshere M, Milham C, Hopkin E, Pound K, Fernando S, *et al*. (2003) Linkage disequilibrium mapping provides further evidence of a gene for reading disability on chromosome 6p21.3-22. *Molecular Psychiatry*, **8**, 176–185.

Tzenova J, Kaplan BJ, Petryshen TL and Field LL (2004) Confirmation of a dyslexia susceptibility locus on chromosome 1p34-p36 in a set of 100 Canadian families. *American Journal of Medical Genetics B Neuropsychiatr Genet*, **127**, 117–124.

Vincent JB, Kolozsvari D, Roberts WS, Bolton PF, Gurling HM and Scherer SW (2004) Mutation screening of X-chromosomal neuroligin genes: no mutations in 196 autism probands. *American Journal of Medical Genetics*, **129B**, 82–84.

Wadsworth SJ, Olson RK, Pennington BF and DeFries JC (2000) Differential genetic etiology of reading disability as a function of IQ. *Journal of Learning Disabilities*, **33**, 192–199.

Whittington J, Holland A, Webb T, Butler J, Clarke D and Boer H (2004) Cognitive abilities and genotype in a population-based sample of people with Prader-Willi syndrome. *Journal of Intellectual Disability Research: JIDR*, **48**, 172–187.

Wiznitzer M (2004) Autism and tuberous sclerosis. *Journal of Child Neurology*, **19**, 675–679.

Zhong H, Serajee FJ, Nabi R and Huq AH (2003) No association between the EN2 gene and autistic disorder. *Journal of Medical Genetics*, **40**, e4.

Short Specialist Review

Complexity of cancer as a genetic disease

Tea Vallenius and Tomi P. Mäkelä
University of Helsinki, Helsinki, Finland

1. Cancer is a complex genetic disease

Although cancer is clearly a genetic disease, only a small fraction of cancer is hereditary, and instead the majority of genetic alterations accumulate during disease progression. These somatic mutations are caused by DNA-damaging agents either arising from endogenous metabolic sources or external harmful agents. Occasionally, and fortunately rarely, mutations occur in an oncogene or a tumor-suppressor gene facilitating continued growth and viability of cells, leading to clonal expansion. Accumulation of subsequent mutations can provide the cells with further growth-promoting characteristics, and thus progression of tumorigenesis (Hanahan and Weinberg, 2000), which is generally believed to require several independent genetic insults (Hahn and Weinberg, 2002). Another classic hallmark of cancer cells are the frequent abnormalities in chromosome numbers (aneuploidy). Both the mechanisms leading to aneuploidy, and the role of aneuploidy in tumor progression remain debated issues. One of the models has presumed that aneuploidy simply reflects a general increased propensity of tumor cells to acquire mutations (mutator phenotype). Interestingly, it was recently noted that mutations in the *CDC4* gene are sufficient to lead to aneuploidy of colon cancer cells, suggesting that aneuploidy could be the result of mutations in specific genes (Rajagopalan and Lengauer, 2004). On the basis of the current knowledge of signaling pathways targeted in cancer, it is clear that a number of significant pathways and their interplay are poorly understood. The increasing number of tools enabling us to analyze cancer cells and tissues from a systems level viewpoint has set the stage for important new discoveries in this field. From the viewpoint of treating cancer, the most important question is how these results translate to therapy.

2. Targeted cancer therapy

Identification of specific oncogenic lesions common in some forms of cancer has spurred development of drugs targeted against these to support or even replace traditional chemo- and radiotherapies. The first success was the discovery of a

selective tyrosine kinase inhibitor imatinib mesylate (STI-571, Gleevec) in the treatment of chronic myeloid leukemia (CML) (Druker *et al*., 2001). Although imatinib was initially used in CML as an inhibitor of the BCR-ABL fusion protein, subsequently, it has been shown to inhibit additionally c-kit and PDGFR in a variety of tumor types (Druker, 2004). These success stories have motivated attempts to identify compounds that selectively block the activity of certain critical molecule in tumor cells, and hundreds of potential drugs, both specific small molecule inhibitors as well as monoclonal antibodies, are under different stages of development (**www.nci.nih.gov/clinicaltrials**). One very interesting class is epidermal growth factor receptor inhibitors gefitinib (Iressa) and erlotinib (Tarceva), which block the signal transduction pathways implicated in cancer cell growth and survival (Ross *et al*., 2004). Gefitinib trials in non-small-cell lung cancer (NSCLC) were initially disappointing due to a large variation seen in its ability to reduce lung carcinomas. However, more recent studies indicate that patients who respond to gefitinib treatment carry a mutation in a kinase domain of EGFR, pinpointing the need for diagnostic tools to select patients who will benefit from this treatment (Lynch *et al*., 2004; Paez *et al*., 2004).

These encouraging results reflect a changing trend in cancer therapy. At the same time, the trials with these drugs have been educating in the way of understanding the possibilities of killing cancer. Although there is a wealth body of evidence indicating that cancer is caused by the accumulation of mutations, the clinical responses attained from first targeted kinase inhibitors suggest that at least certain tumors are highly dependent on activation of a single kinase. This in turn increases the hope that other tumor types might be dependent on other kinases, which could be used in developing cancer-specific drugs. Obviously, this means that it will be fundamental to identify kinases or other molecules and their signaling mechanisms that maintain the progression of different tumors. To this end, integrated research data is required from different large-scale analysis including transcriptomics, and proteomics (Figure 1). In addition to the rapidly increasing expression and proteomics profiling data, it is important to identify specific gene mutations as demonstrated by, for example, the NSCLC and GIST experiences described above; mutations in specific genes are more indicative than changes in expression levels (Medeiros *et al*., 2004; Paez *et al*., 2004). Furthermore, the rapidly emerging field of biomarker identification through mass spectrometry and the ability to analyze posttranslational modifications in large scale provide information that should be integrated in the drug target discovery process.

Although clinical experience on targeted therapies is still in its early days, it is already known that relapses occur at some frequency, and, for example, in CML more frequently with patients having more advanced stages of disease (Gorre *et al*., 2001). In some cases of CML, the relapse is due to clonal expansion of a cell that acquires additional mutations in the targeted BCR-ABL kinase leading to reactivation of the damaged signaling pathway. It is also important to note that in the clinical setting, new promising candidate drugs are generally tested initially in patients with end-stage disease and usually as an added component to a treatment regimen of one or more traditional anticancer drugs. Results might be both better and easier to interpret if trials could be performed at earlier disease stages. In this

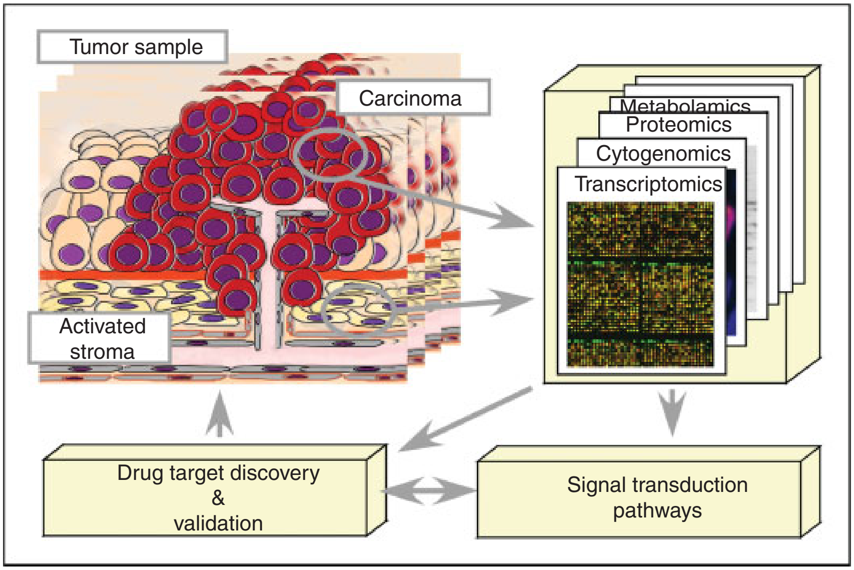


Figure 1 Integrated research data from different large-scale analyses is likely to speed up the understanding of signal transduction pathways leading to cancer and the development of cancer-specific drugs

respect, one current limitation to study drug effects is appropriate in vivo tumor models, particularly mouse models mimicking human cancer.

3. Stromal cells in tumors as therapy targets

Along with studies trying to identify the deregulated molecular circuits leading to the growth of neoplastic cells, a lot of attention has been recently focused on the interplay between neoplastic epithelial cells and the surrounding normal cells (Bhowmick *et al*., 2004b; Mueller and Fusenig, 2004). The neoplastic epithelial cells evolve ways in which to utilize several types of normal cells, for example, through production of growth factors, proteases, and so on. The ensuing stromal reaction is characterized by the presence of activated fibroblasts, proliferating blood vessel cells, altered extracellular matrix, and inflammatory response, all of which help the tumor to grow further and finally metastasize (Mueller and Fusenig, 2004). A dramatic example of this was provided by the observation that lack of TGF-signaling in fibroblasts resulted in epithelial neoplasia (Bhowmick *et al*., 2004a). This and other studies have led to the proposal that modified tumor contributes to tumor formation and progression (Bhowmick *et al*., 2004b). These types of observations are increasing interest in attempting to target the supporting cells of the carcinoma with therapy. On this line, the established strategy is antiangiogenic therapy, where endothelial cells growing into the tumors and providing tumor vasculature are targeted. Several antiangiogenic compounds are under development of which bevacizumab (Avastin) was the first to be approved for patients having

metastatic colorectal cancer (Ross *et al*., 2004). Bevacizumab is a monoclonal antibody neutralizing VEGF, which in turn loses it ability to activate its receptor on endothelial cells, and thus prevents the formation of new blood vessels. Another potential therapeutic strategy is to inhibit inflammatory cells and cytokines by nonsteroidal anti-inflammatory drugs, such as COX-2 inhibitors (Mueller and Fusenig, 2004). The future will show how these compounds survive in anticancer therapy, but it has been speculated that one benefit to target stroma is that these cells are genetically more stable than epithelial neoplastic cells and thus less subjected to developing drug resistance (Mueller and Fusenig, 2004). However, it is clear that further investigation into the tissue-specific interactions between neoplastic and stromal cells will be an essential component of rational drug development.

4. Conclusions

The recent significant achievements in cancer research accompanied by the available postgenomic era tools offer a lot of optimism to understand the enormous heterogeneity in genetic alterations and their implications in cancer. This does not mean that the primary therapy for a solid tumor is surgery when possible or that chemo- and radiotherapy remain as major forms of cancer therapy. However, several recent breakthroughs in the targeted cancer therapy indicate that they have arrived to stay as important tools in treating cancer patients. During the next several years, numerous new, targeted drug products will be introduced accompanied by diagnostic assays designed to identify right patients for each compound.

References

Bhowmick NA, Chytil A, Plieth D, Gorska AE, Dumont N, Shappell S, Washington MK, Neilson EG and Moses HL (2004a) TGF-beta signaling in fibroblasts modulates the oncogenic potential of adjacent epithelia. *Science*, **303**, 848–851.

Bhowmick NA, Neilson EG and Moses HL (2004b) Stromal fibroblasts in cancer initiation and progression. *Nature*, **432**, 332–337.

Druker BJ (2004) Imatinib as a paradigm of targeted therapies. *Advances in Cancer Research*, **91**, 1–30.

Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, *et al*. (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, **344**, 1031–1037.

Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, Rao PN and Sawyers CL (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, **293**, 876–880.

Hahn WC and Weinberg RA (2002) Modelling the molecular circuitry of cancer. *Nature Reviews Cancer*, **2**, 331–341.

Hanahan D and Weinberg RA (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, *et al*. (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, **350**, 2129–2139.

Medeiros F, Corless CL, Duensing A, Hornick JL, Oliveira AM, Heinrich MC, Fletcher JA and Fletcher CD (2004) KIT-negative gastrointestinal stromal tumors: proof of concept and therapeutic implications. *The American Journal of Surgical Pathology*, **28**, 889–894.

Mueller MM and Fusenig NE (2004) Friends or foes – bipolar effects of the tumour stroma in cancer. *Nature Reviews Cancer*, **4**, 839–849.

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, *et al*. (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, **304**, 1497–1500.

Rajagopalan H and Lengauer C (2004) Aneuploidy and cancer. *Nature*, **432**, 338–341.

Ross JS, Schenkein DP, Pietrusko R, Rolfe M, Linette GP, Stec J, Stagliano NE, Ginsburg GS, Symmans WF, Pusztai L, *et al*. (2004) Targeted therapies for cancer 2004. *American Journal of Clinical Pathology*, **122**, 598–609.

Short Specialist Review

The mitochondrial genome

Douglas C. Wallace
University of California, Irvine, CA, USA

1. The nature of the mitochondrial genome

The mitochondria are semiautonomous bacteria that reside within the cytoplasm of virtually all eukaryotic cells. These bacteria formed a symbiotic relationship with a nucleated host cell over 2 billion years ago. At the time of the symbiosis, the bacterial (mitochondrial) DNA contained all of the genes necessary for a free-living organism. However, over time, most of the genes in the mtDNA were transferred to the host cell's nuclear DNA (nDNA). Consequently, today the mitochondrial genome is distributed between the mtDNA and the nDNA.

The mtDNAs of humans and mammals retain the same genes. These include the small (12 S) and large (16 S) ribosomal RNAs (rRNA), the 22 transfer RNAs (tRNAs) necessary to translate the modified mitochondrial genetic code, and 13 proteins. The 13 proteins are all key components of the mitochondrial energy generating pathway oxidative phosphorylation (OXPHOS), and include seven (ND1, ND2, ND3, ND4L, ND4, ND5, ND6) of the approximately 46 polypeptides of OXPHOS complex I, one (cytochrome b, cytb) of the 11 proteins of complex III, three (COI, COII, COIII) of the 13 proteins of complex IV, and two (ATP6, ATP8) of the approximately 16 proteins of complex V. All of the remaining proteins of the mitochondrial genome, currently estimated to be about 1500 and including those for the polymerases, ribosomal proteins, structural proteins, and enzymes, are now located in the nucleus (Wallace and Lott, 2002).

While the nDNA-encoded mitochondrial genes are inherited, replicated, transcribed, and translated like other nDNA genes, the biology and genetics of the mtDNA genes is totally different. Consequently, the genetics of the mitochondrion is unique and complex.

The human mtDNA is a circular molecule of approximately 16 569 nucleotide pairs (nps) (Figure 1). In addition to the 37 structural genes, the mtDNA encompasses an approximately 1000-np control region (CR).

2. Mitochondrial function

Mitochondria generate energy by burning hydrogen derived from carbohydrates and fats with oxygen to generate water (H_2O) (Wallace and Lott, 2002) (Figure 2).

As a toxic by-product of OXPHOS, the mitochondria generate most of the reactive oxygen species (ROS or oxygen radicals) of eukaryotic cells. This results

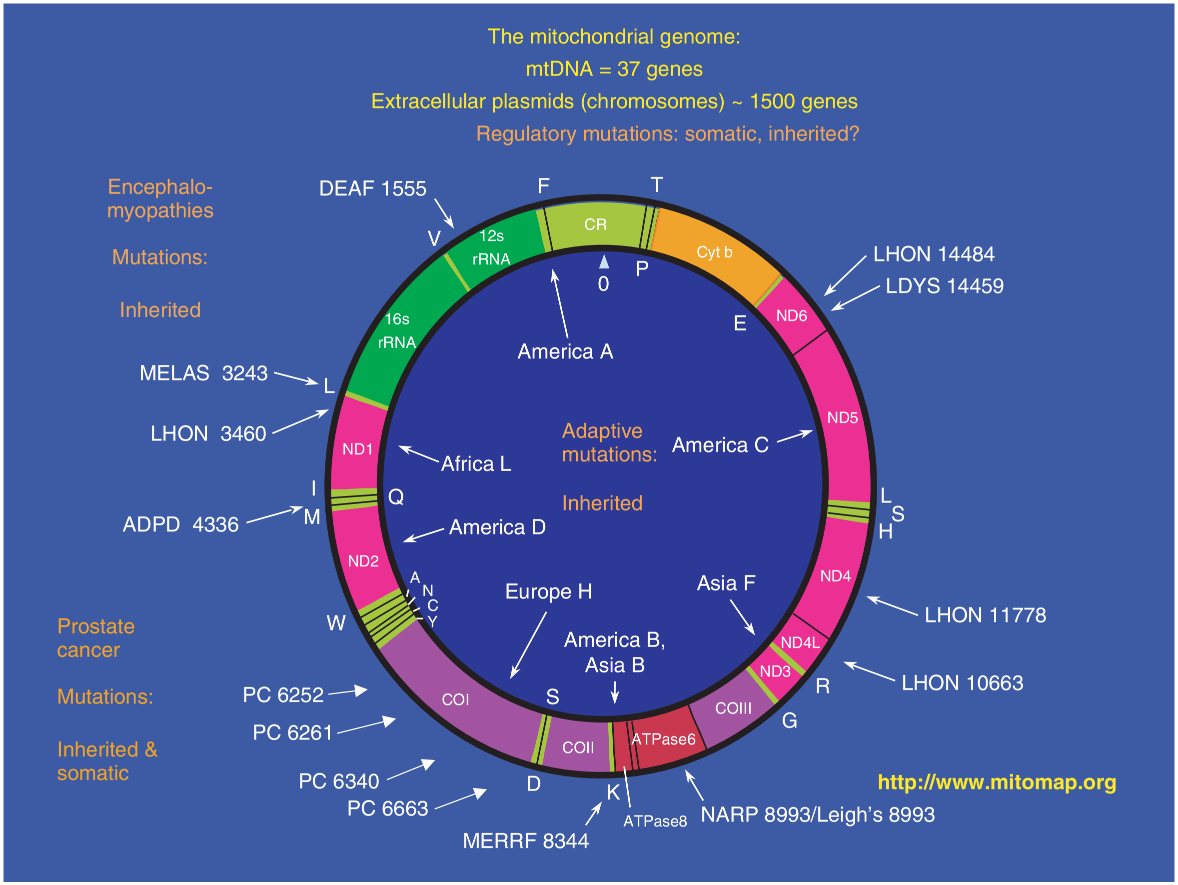


Figure 1 Human mitochondrial DNA Map. D-loop = control region. Letters around the outside perimeter indicate cognate amino acids of the tRNA genes. Other gene symbols are defined in the text. Arrows followed by continental names and associated letters on the inside of the circle indicate the position of defining polymorphisms of selected region-specific mtDNA lineages. Arrows associated with abbreviations followed by numbers around the outside of the circle indicate representative pathogenic mutations, the number being the nucleotide position of the mutation. The abbreviations are DEAF = deafness; MELAS = mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes; LHON = Leber's hereditary optic neuropathy; ADPD = Alzheimer's and Parkinson's disease; MERRF = myoclonic epilepsy and ragged red fiber disease; NARP = neurogenic muscle weakness, ataxia, retinitis pigmentosum; LDYS = LHON + dystonia

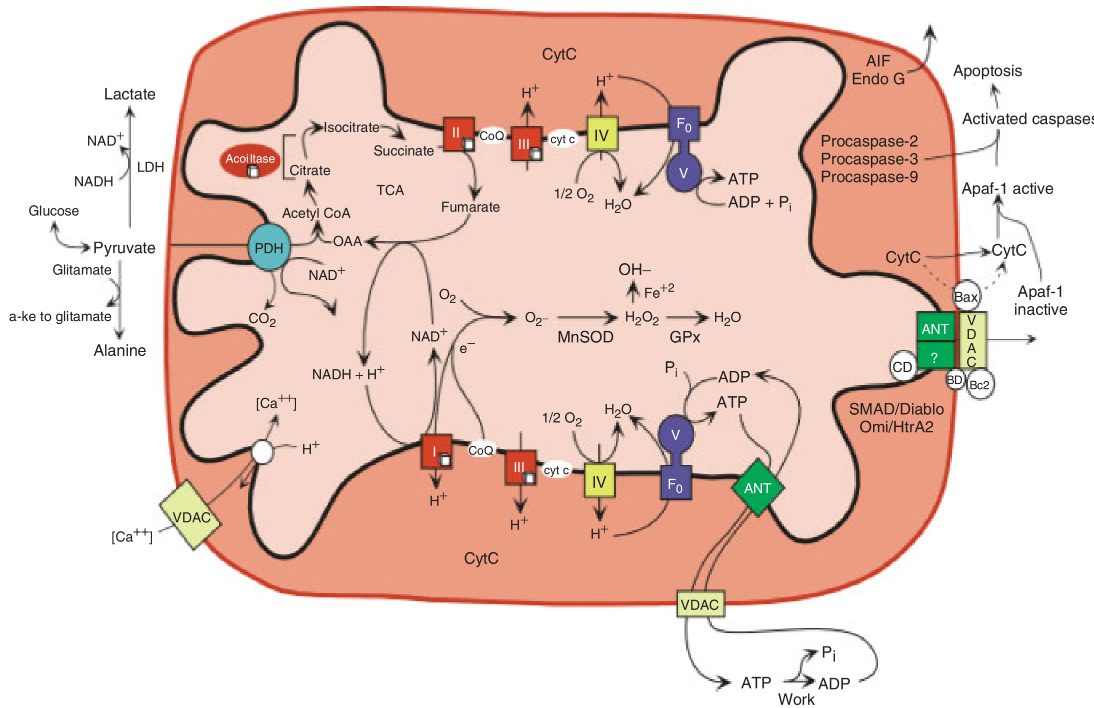


Figure 2 Three features of mitochondrial metabolism relevant to the pathophysiology of disease: (1) energy production by oxidative phosphorylation (OXPHOS), (2) reactive oxygen species (ROS) generation as a toxic by-product of OXPHOS, and (3) regulation of apoptosis through activation of the mitochondrial permeability transition pore (mtPTP). The energy that is released is used to maintain body temperature and to generate ATP. The reducing equivalents (electrons) from dietary calories are collected by the tricarboxylic acid (TCA) cycle and β-oxidation and transferred to either NAD^+ to generate NADH or FAD to give $FADH_2$. The electrons are then oxidized by the electron transport chain (ETC). NADH transfers electrons to complex I (NADH dehydrogenase) and succinate from the TCA cycle transfers electrons to complex II (succinate dehygrogenase) and both complexes transfer their electrons to ubiquinone (coenzyme Q_{10} or CoQ) to generate ubisemiquinone ($CoQH^{\cdot}$), and then ubiquinol ($CoQH_2$). From $CoQH_2$, the electrons are transferred to complex III, then cytochrome c, then complex IV, and finally to $\frac{1}{2}$ O_2 to give water. The energy that is released during electron transport is used to pump protons from the matrix out across the mitochondrial inner membrane to create an electrochemical gradient ($\Delta P = \Delta\Psi + \Delta\mu^{H+}$). This biological capacitor serves as a source of potential energy to drive complex V to condense ADP + Pi to give ATP. The mitochondrial ATP is then exchanged for cytosolic ADP by the ANT. The efficiency by which ΔP is converted to ATP is known as the coupling efficiency. Tightly coupled OXPHOS generates the maximum ATP and minimum heat, while loosely coupled OXPHOS generates less ATP but more heat. I, II, III, IV, and V = OXPHOS complexes I–V; ADP or ATP = adenosine di- or triphosphate; ANT = adenine nucleotide translocator, cytc = cytochrome c; GPx = glutathione peroxidase-1; LDH = lactate dehydrogenase; MnSOD = manganese superoxide dismutase or SOD2; NADH = reduce nicotinamide adenine dinucleotide; TCA = tricarboxylic acid cycle; VDAC = voltage-dependent anion channel

from of the misdirection of electrons from the early stages of the electron transport chain (ETC) surrounding $CoQH^{\cdot}$ directly to O_2 to generate superoxide anion ($O_2^{\cdot -}$). Superoxide anion is highly reactive, but can be detoxified by mitochondrial MnSOD to generate hydrogen peroxide (H_2O_2), which is relatively stable. However, in the presence of transition metals, H_2O_2 can be reduced to hydroxyl radical ($^{\cdot}OH$), which is the most toxic ROS. H_2O_2 can be detoxified by reduction to water by glutathione peroxidase (GPx1) or conversion to O_2 and H_2O by catalase (Figure 2). ROS damage mitochondrial proteins, lipids, and nucleic acids, which inhibit OXPHOS and further exacerbate ROS production. Ultimately, the mitochondria become sufficiently energetically impaired that they malfunction and must be removed from the tissue.

Cells with faulty mitochondria are removed by apoptosis (programmed cell death) through the activation of the mitochondrial permeability transition pore (mtPTP). The mtPTP is thought to be composed of the outer membrane voltage-dependent anion channel (VDAC) proteins, the inner membrane ANT, the pro- and antiapoptotic Bcl2-Bax gene family proteins, and cyclophillin D. This complex senses changes in the mitochondrial electrochemical gradient (ΔP), adenine nucleotides, ROS, and Ca^{++}, and when these factors get sufficiently out of balance, the mtPTP becomes activated and opens a channel between the cytosol and the mitochondrial matrix depolarizing ΔP and causing the mitochondrion to swell. Proteins from the mitochondrial intermembrane space are then released into the cytosol. Among these are cytochrome c that activates cytosolic Apaf-1 to convert procaspases 2, 3, and 9 to active caspases that degrade cellular and mitochondrial proteins. Apoptosis initiating factor (AIF) and Endonuclease G are transported to the nucleus where they degrade the nDNA (Figure 2).

3. Genetics of the nDNA mitochondrial genes and disease mutations

A variety of mitochondrial diseases result from mutations in the nDNA-encoded genes of the mitochondrial genome, and thus exhibit Mendelian inheritance. These can affect structural and assembly genes of OXPHOS, genes involved in mtDNA maintenance, and genes affecting mitochondrial fusion and mobility (Wallace and Lott, 2002; DiMauro, 2004) (Table 1).

4. Genetics of the mtDNA and degenerative diseases, cancer, and aging

In sharp contrast to the Mendelian inheritance of diseases resulting from nDNA mitochondrial mutations, diseases resulting from mutations in the mtDNA have a totally unique inheritance. In vertebrates, the mtDNA is inherited exclusively through the mother. However, each cell can harbor thousands of copies of the mtDNA, which can either be pure normal (homoplasmic wild type), a mixture of mutant and normal (heteroplasmic), or homoplasmic mutant. The percentage

Table 1 Diseases due to Mutations in nDNA Mitochondrial genes

Clinical disease	Mitochondrial defect	Genetic defect	References
Leigh syndrome	Complex I	Various structural genes	Procaccio and Wallace (2004), Smeitink and van den Heuvel (1999)
Leigh syndrome	Complex IV	SURF1 assembly	Tiranti *et al*. (1998), Zhu *et al*. (1998)
AD-CPEO	Multiple mtDNA deletions	ANT1 (14q34)	Kaukonen *et al*. (2000)
		Twinkle helicase (10q21)	Spelbrink *et al*. (2001)
		Polymerase γ (15q25)	Van Goethem *et al*. (2003), Van Goethem *et al*. (2001)
Variable organ failure	mtDNA depletion	Deoxyguanosine kinase (2q13)	Mandel *et al* (2001)
		Thymidine kinase 2 (16q21)	Saada *et al*. (2001)
MNGIE	MtDNA deletions and depletion	Thymidine phosphorylase (22q13)	Nishino *et al*. (1999)
AD-OPA	Mitochondrial fusion and fission	Dynamin-related GTPase (OPA1)	Delettre *et al*. (2000)
Peripheral neuropathy	Mitochondrial fusion and fission	Mitofusin 2 (MFN2)	Zuchner *et al*. (2004)

AD-CPEO: autosomal dominant – chronic progressive external ophthalmoplegia; MNGIE: mitochondrial neurogastrointestinal encephalomyopathy; AD-OPA: autosomal dominant optic atrophy.

of mutant mtDNAs determines the relative energy deficiency of the cell and thus the probability of apoptosis. The tissues most sensitive to the adverse effects of mitochondrial decline are the central nervous system, heart, muscle, renal, and endocrine systems, the tissues most affected in aging (Wallace and Lott, 2002).

Because of its chronic exposure to ROS, mammalian mtDNAs have a very high mutation rate. Hence, even though the human mtDNA is small, mtDNA diseases are common. Clinically relevant mtDNA mutations come in three classes: (1) recent maternal germline mutations frequently resulting in disease; (2) ancient polymorphisms, a subset of which permitted our ancestors to adapt to new environmental conditions; and (3) somatic mtDNAs that accumulate in postmitotic tissue mtDNAs with age (Wallace and Lott, 2002).

4.1. Recent pathogenic mutations

Recent pathogenic mutations include both base substitution and rearrangement mutations. Base substitution mutations can alter proteins or protein synthesis (rRNA and tRNA) genes. An example of a mtDNA missense mutation is the np 11778

mutation in ND4 (R340H), which causes Leber's Hereditary Optic Neuropathy (LHON), a form of sudden-onset, midlife blindness. An example of a protein synthesis mutation is the np 8344 in the $tRNA^{Lys}$ gene that causes Myoclonic Epilepsy and Ragged Red Fiber disease (MERRF). An extensive list of the known pathogenic mtDNA mutations is provided at our website "**www.mitomap.org**", and the characteristics of the diseases are given in two large reviews (Wallace and Lott, 2002; Wallace *et al*., 2001).

Rearrangement mutations can occur throughout the mtDNA, and can include both maternally inherited and spontaneous cases. Maternally inherited rearrangement mutations are most commonly associated with Type II diabetes and deafness, and are thought to be inherited as insertion mutations (Ballinger *et al*., 1992; Ballinger *et al*., 1994). Deletions in the mtDNA, with or without attendant insertions, are primarily spontaneous and result in multisystem disorders encompassing a continuum of phenotypes ranging from Chronic Progressive External Ophthalmoplegia (CPEO) to the more severe Kearns–Sayre syndrome. The most severe mtDNA rearrangement disorder is the Pearson's marrow pancreas syndrome, which frequently leads to death in childhood from pancytopenia. The range of mtDNA rearrangements is cataloged at **www.mitomap.org**.

mtDNA diseases can affect the central nervous system, diminishing vision, hearing, movement, balance, and memory; muscle causing progressive weakness; heart leading to hypertrophic and dilated cardiomyopathy; the endocrine system resulting in diabetes mellitus, and the renal system. Hence, it is now clear that mitochondrial defects represent one of the most common forms of human inborn errors of metabolism (Wallace and Lott, 2002).

The pathogenicity of mildly deleterious mtDNA mutations can also be augmented when the mutation occurs on regional mtDNAs that harbor ancient adaptive mtDNA polymorphisms. For instance, LHON can be caused by a number of mtDNA point mutations in the *ND* genes. The more biochemically severe mutations cause LHON independent of the background mtDNA, but the milder ND mutations such as np 11778, 14484, and 10663 cause LHON when they co-occur with the European mtDNA lineage J.

4.2. *Ancient adaptive mtDNA mutations*

The mtDNAs of indigenous human populations around the world harbor region-specific mtDNAs that encompass groups of related haplotypes, called *haplogroups*. These haplogroups arose a people migrated out of African into Eurasia and ultimately into the Americas.

The mtDNA history is remarkable for the striking discontinuities that exist between the extensive mtDNA variation of Africa from which only two mtDNA lineages (M and N) occupied all of Eurasia and the plethora of Asian mtDNA types of which only three haplogroups (A, C, and D), and much later G came to occupy the extreme northeastern Chukotka Peninsula of Siberia. These discontinuities strikingly correlate with latitude and thus the climate. This led to the hypothesis that certain mtDNA haplogroups acquired functional mutations that decreased the coupling efficiency, increasing mitochondrial heat production that permitted people to adapt to the increasing cold of the more northern latitudes.

This hypothesis is supported by the fact that missense mutations in mtDNA protein genes show regional specificity. Missense mutations are prevalent in the *ATP6* gene in the arctic, in the *cytb* gene in Europe, and in the *COI* gene in Africa. Mutations in different *ND* genes also show regional correlation (Mishmar *et al*., 2003). Moreover, many of the ancient missense mutations change amino acids that are as highly evolutionarily conserved as are most known pathogenic mutations, yet have been retained in the human population for tens of thousands of years. Hence, they could not be pathogenic in the environment in which they reside, but rather must be adaptive and thus beneficial. Furthermore, an analysis of the cytb missense mutations for which there is a complex III crystal structure revealed that many of these missense mutations affected CoQ binding sites, which would alter the Q-cycle and thus proton pumping (Ruiz-Pesini *et al*., 2004).

Finally, when the European mtDNA haplogroups were correlated with longevity and predisposition to Alzheimer disease (AD) and Parkinson disease (PD), it was found that mtDNAs harboring uncoupling variants were enriched in the elderly and deficient in AD and PD patients. This led to the conclusion that the uncoupling mutations must enhance the flow of electrons through the ETC, keeping the ETC carriers oxidized. This, in turn, reduces the spurious transfer of electrons to O_2 to generate ROS and associated mitochondrial and cellular damage.

These same uncoupling mutations would also reduce the efficiency of ATP generation that would then exacerbate the ATP deficiencies resulting from the milder mtDNA mutations. This could account for the predilection of LHON patients harboring the milder mtDNA mutations to also have haplogroup J mtDNAs, which harbor either the np 14798 or np 15257 cytb missense mutations. Thus, ancient adaptive mtDNA variants are affecting individual predisposition to degenerative diseases and aging today.

4.3. Somatic mtDNA mutations

Postmitotic tissues also accumulate somatic mtDNA mutations with age. These can be rearrangement, coding region base substitution, and CR mutations. Since mtDNA diseases generally have a delayed-onset and progressive course and affect the same tissues and generate the same symptoms as aging, it has been hypothesized that the accumulation of somatic mtDNA mutations results in the age-related decline in mitochondrial function that ultimately results in degenerative diseases and the senescence of aging (Wallace and Lott, 2002).

Consistent with these concepts, mtDNA deletions have been found to accumulate with age in those tissues most prone to age-related decline. Also, mtDNA CR mutations have been found to accumulate in various tissues including fibroblasts, muscle, and brain (Michikawa *et al*., 1999; Coskun *et al*., 2003)

Somatic mtDNA mutations have also been found in a variety of solid tumors including prostate, breast, colon, bladder, head, and neck tumors. In prostate cancer, a study of the *COI* gene revealed that both germline and somatic mtDNA mutations were associated with prostate cancer. Moreover, the substitution of a mtDNA harboring a known pathogenic mutation for the resident mtDNAs of a prostate cancer cell line increased tumor growth, while substitution of the same mtDNA but with the

normal base inhibited tumor growth. This increased tumorigenicity was associated with increased ROS production. Therefore, mtDNA mutations that increase ROS production may also be important factors in tumorigenicity (Petros *et al*., 2005).

Somatic mtDNA mutations may also be a major contributor to age-related neurodegenerative diseases including AD. A survey of the mtDNA CRs of AD brains revealed that 65% harbored the T414G mutations in the mtTFA binding site for P_L, while none of the controls harbored this mutation. Moreover, the AD brains had an overall 73% increase in CR mutations. Finally, these mutations were preferentially located in CR regulatory elements and were associated with a 50% reduction in mtDNA copy number and a 50% reduction in the L-strand ND6 transcript (Coskun *et al*., 2004). Hence, mtDNA CR mutations may be an important factor in the development of neurodegenerative disease.

5. Conclusion

The distribution of the genes of the mitochondrial genome between the high copy-number, maternally inherited mtDNA and the low copy-number, Mendelian nDNA results in a complex genetics with both quantitative and quantized genetic elements. As a result, the mitochondrial genome provides the necessary features to explain the complex inheritance, delayed onset, and progressive course of the age-related degenerative diseases, aging, and cancer. Only the mtDNA is present in sufficient copies within each cell to permit a progressive accumulation of somatic mtDNA mutations, thus providing an aging clock. Moreover, the central role of the mitochondria in energy production, ROS generation, and apoptosis regulation links dietary intake with genetic predisoposition, a common hallmark of degenerative diseases and cancer. Hence, if we are to understand the common degenerative diseases, aging, and cancer, we must understand the genetics and pathophysiology of the mitochondrion.

Acknowledgments

This work was supported by NIH grants NS21328, HL64017, AG24373, AG13154, and NS41850.

References

Ballinger SW, Shoffner JM, Hedaya EV, Trounce I, Polak MA, Koontz DA and Wallace DC (1992) Maternally transmitted diabetes and deafness associated with a 10.4 kb mitochondrial DNA deletion. *Nature Genetics*, **1**, 11–15.

Ballinger SW, Shoffner JM, Gebhart S, Koontz DA and Wallace DC (1994) Mitochondrial diabetes revisited. *Nature Genetics*, **7**, 458–459.

Coskun PE, Beal MF and Wallace DC (2004) Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10726–10731.

Coskun PE, Ruiz-Pesini EE and Wallace DC (2003) Control region mtDNA variants: longevity, climatic adaptation and a forensic conundrum. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 2174–2176.

Delettre C, Lenaers G, Griffoin JM, Gigarel N, Lorenzo C, Belenguer P, Pelloquin L, Grosgeorge J, Turc-Carel C, Perret E, *et al*. (2000) Nuclear gene OPA1, encoding a mitochondrial dynamin-related protein, is mutated in dominant optic atrophy. *Nature Genetics*, **26**, 207–210.

DiMauro S (2004) Mitochondrial medicine. *Biochimica et Biophysica Acta*, **1659**, 107–114.

Kaukonen J, Juselius JK, Tiranti V, Kyttala A, Zeviani M, Comi GP, Keranen S, Peltonen L and Suomalainen A (2000) Role of adenine nucleotide translocator 1 in mtDNA maintenance. *Science*, **289**, 782–785.

Mandel H, Szargel R, Labay V, Elpeleg O, Saada A, Shalata A, Anbinder Y, Berkowitz D, Hartman C, Barak M, *et al*. (2001) The deoxyguanosine kinase gene is mutated in individuals with depleted hepatocerebral mitochondrial DNA. *Nature Genetics*, **29**, 337–341.

Michikawa Y, Mazzucchelli F, Bresolin N, Scarlato G and Attardi G (1999) Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science*, **286**, 774–779.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, *et al*. (2003) Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 171–176.

Nishino I, Spinazzola A and Hirano M (1999) Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science*, **283**, 689–692.

Petros JA, Baumann AK, Ruiz-Pesini E, Amin MB, Sun CQ, Hall J, Lim S, Issa MM, Flanders WD, Hosseini SH, *et al*. (2005) mtDNA mutations increase tumorigenicity in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 719–724.

Procaccio V and Wallace DC (2004) Late-onset Leigh syndrome in a patient with mitochondrial complex I NDUFS8 mutations. *Neurology*, **62**, 1899–1901.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V and Wallace DC (2004) Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, **303**, 223–226.

Saada A, Shaag A, Mandel H, Nevo Y, Eriksson S and Elpeleg O (2001) Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. *Nature Genetics*, **29**, 342–344.

Smeitink J and van den Heuvel L (1999) Protein biosynthesis ′99. Human mitochondrial complex I in health and disease. *American Journal of Human Genetics*, **64**, 1505–1510.

Spelbrink JN, Li FY, Tiranti V, Nikali K, Yuan QP, Tariq M, Wanrooij S, Garrido N, Comi G, Morandi L, *et al*. (2001) Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nature Genetics*, **28**, 223–231.

Tiranti V, Hoertnagel K, Carrozzo R, Galimberti C, Munaro M, Granatiero M, Zelante L, Gasparini P, Marzella R, Rocchi M, *et al*. (1998) Mutations of SURF-1 in Leigh disease associated with cytochrome c oxidase deficiency. *American Journal of Human Genetics*, **63**, 1609–1621.

Van Goethem G, Dermaut B, Lofgren A, Martin JJ and Van Broeckhoven C (2001) Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nature Genetics*, **28**, 211–212.

Van Goethem G, Martin JJ and Van Broeckhoven C (2003) Progressive external ophthalmoplegia characterized by multiple deletions of mitochondrial DNA: unraveling the pathogenesis of human mitochondrial DNA instability and the initiation of a genetic classification. *Neuromolecular Medicine*, **3**, 129–146.

Wallace DC and Lott MT (2002) Mitochondrial genes in degenerative diseases, cancer and aging. In *Emery and Rimoin's Principles and Practice of Medical Genetics*, Rimoin DL, Connor JM, Pyeritz RE and Korf BR (Eds.), Churchill Livingstone: London, pp. 299–409.

Wallace DC, Lott MT, Brown MD and Kerstann K (2001) Mitochondria and neuro-ophthalmological diseases. In *The Metabolic and Molecular Basis of Inherited Disease*, Vol.

II, Scriver CR, Beaudet AL, Sly WS and Valle D (Eds.), McGraw-Hill: New York, pp. 2425–2512.

Zhu Z, Yao J, Johns T, Fu K, De Bie I, Macmillan C, Cuthbert AP, Newbold RF, Wang J, Chevrette M, *et al*. (1998) SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. *Nature Genetics*, **20**, 337–343.

Zuchner S, Mersiyanova IV, Muglia M, Bissar-Tadmouri N, Rochelle J, Dadali EL, Zappia M, Nelis E, Patitucci A, Senderek J, *et al*. (2004) Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A. *Nature Genetics*, **36**, 449–451.

Introductory Review

Approach to rare monogenic and chromosomal disorders

Marc S. Williams
Gundersen Lutheran Medial Center, La Crosse, WI, USA

1. Introduction

Medical geneticists are asked to evaluate a wide variety of conditions. These include congenital malformations (single and multiple), mental retardation, neurologic disorders, inborn errors of metabolism, chromosomal abnormalities, hereditary cancer syndromes, disorders of hearing and vision, teratogenic exposures, and families with multiple affected members. It is the medical geneticist's job to ascertain and analyze information from a variety of sources in order to establish a diagnosis coordinate testing initiate treatment explain inheritance, recurrence risk, and prognosis; and ultimately assist in the coordination of care for the individual and family.

The purpose of this chapter is to introduce the reader to the genetic evaluation of rare monogenic and chromosomal disorders. The author invites the reader to "observe" a hypothetical clinic visit, with explication, to illustrate this concept.

2. Patient evaluation

2.1. Chief complaint

The patient is a 45-year-old female referred by her internist to a multidisciplinary clinic for evaluation of mental retardation and unusual behaviors.

The central element of any medical encounter is the history and physical. Information gathered from this process allows generation of a differential diagnosis that becomes the basis for subsequent investigation to establish a definite diagnosis. Experienced practitioners recognize that this is a dynamic process that begins with the chief complaint and is constantly modified throughout the encounter. As information is acquired, subsequent questions are added, subtracted, or modified to refine the differential. While this process is hardly unique to the genetics clinic, the range of information required is generally much broader. Typically, this will involve review of primary medical records as well as obtaining history from the patient, or, in the case of this patient, from a parent or reliable caretaker. While

the differential is of necessity broad at this point, we know that the diagnosis must be restricted to congenital disorders and involve mental retardation. This could include chromosome disorders (although Trisomies 13 and 18 can likely be ruled out due to the patient's age, and so is Down Syndrome, given that this is a highly recognizable disorder), inborn errors of metabolism (including Phenylketonuria (PKU) and congenital hypothyroidism, as the patient was born before routine newborn screening), multiple congenital anomaly syndromes, congenital infection, perinatal accident, and teratogenic exposures.

2.2. Patient history

Preclinic records review. Pregnancy was complicated by possible exposure to "Asian flu". Delivery was described as "delayed", without additional explanation. She was described as a poor feeder with a poor suck, but no other neonatal information was available. She was institutionalized in a center for handicapped children at age 3, where she remained until discharge to a group home at age 33. She was noted to fall easily and had several fractures relating to falls. Of note was a neck injury secondary to a fall that resulted in a central cord syndrome, which was thought to explain her predisposition to falls. Diagnoses included profound mental retardation, impulse control disorder with self-injurious behavior, episodic depression, and possible seizure disorder. Previous evaluation included a head CT scan, which was normal; MRI of the cervical spine showing degenerative changes of C3-C6 vertebral bodies, but no changes in the cord; normal Thyroid stimulating hormone (TSH) on several occasions; an EEG (just prior to this evaluation) that showed multifocal epileptiform discharges without a clear clinical correlate; normal hearing; myopia; and exotropia. IQ testing done 10 years previously gave a mental age estimate between 9 and 14 months. Review of the previous intellectual assessments ruled out deterioration of function over time. She never developed speech. She was on multiple medications, including several psychotropic medications to attempt to control behavior and two anticonvulsants. The anticonvulsants were being tapered, given the lack of clinically evident seizures.

On the basis of record review, congenital hypothyroidism can be ruled out. There was history of an illness during the pregnancy. If the "Asian flu" was, in fact, one of the TORCH infections (Toxoplasmosis, Other, Rubella, Cytomegalovirus, Herpes) this could explain the patient's problems. Other acquired causes of retardation such as CNS infection can be ruled out. Disorders with short stature (height reported was 165 cm) or structural CNS anomalies can also be eliminated. The degree of mental retardation is profound, but the effects of institutionalization on ultimate intellectual function, while subject to debate, could have had an effect (Russell et al., 1986; Silverstein, 1962, 1969; Sternlicht and Siegel, 1968). Progressive neurologic disorders can be eliminated from the differential, as the patient has a chronic static encephalopathy. The question of seizures is important given the history of frequent falls and the abnormal EEG findings, even in the absence of clinically apparent seizures. Additional information will be needed to characterize the behavioral issues. Attempts to obtain family history will be critical to knowing whether this

is an isolated case in the family. A specific diagnosis, Angelman syndrome, had moved to the top of the author's differential diagnosis at this point.

2.3. Office visit

The patient was accompanied to the office visit by her group home worker. Her parents were her legal guardians, but lived in another state. The worker was unable to provide any additional information about past medical or family history. She did indicate that the patient was generally happy and that there were no specific behavioral issues currently. She confirmed that the patient had an unsteady gait, but did not walk with arms extended (something that can be seen in Angelman syndrome). No obvious seizure manifestations, including staring spells, unresponsiveness, or incontinence, were noted. The patient is generally healthy with the exception of injuries from falls. A complete review of systems did not reveal any other concerns. (NB: The genetic review of systems includes the usual medical information, but adds information specific to genetic disorders such as unusual odors, intolerance to certain foods or fasting, response to illness and cyclical patterns of illness. Presence of some or all of these findings strongly suggests an inborn error of metabolism, with buildup of toxic compounds leading to symptoms.)

Given the lack of early information, the parents were contacted by phone. The mother indicated that her pregnancy was normal, and she did not have any complications, exposures, or illnesses. She indicated that her physician explained that the retardation was due to a problem with the pregnancy or delivery, despite the lack of any pregnancy complications, a normal vaginal delivery at term, and no neonatal difficulties other than the feeding difficulties. (It should be noted that in virtually all patients born before 1970, and in many born after, the cause is given as problems with pregnancy or delivery, or high fever. Rarely is there evidence to implicate either. This reflects a lack of information about the etiology of birth defects and mental retardation at that time that persists to some degree even to the present day). The parents were encouraged by their physician to institutionalize the patient.

2.4. Family history

No information was available in the medical record. The parents indicated that the family history was unremarkable for any other individual affected with mental retardation, seizures, congenital malformations, or multiple miscarriages. The patient's siblings were all healthy, but several expressed concerns about having a child similar to the patient.

In this case, the family history is not helpful in the differential diagnosis. In many cases, however, the family history provides the information crucial to diagnosis. The genetic family history (or pedigree) is a structured ascertainment of disease in at least three generations of both the maternal and paternal families, which, in addition, may require confirmation of diagnosis, age at diagnosis, pathologic

diagnosis, and ethnicity, as well as other information not collected in a traditional family history (Aase, 1990; Scheuner et al., 1997; see also Article 78, ***Genetic family history, pedigree analysis, and risk assessment****, Volume 2). This information is frequently subjected to formal probability analysis in order to appropriately quantify risk for the consultand (Gelehrter and Collins, 1990; see also Article 78,* ***Genetic family history, pedigree analysis, and risk assessment****, Volume 2). In many monogenic disorders, the family history is the best "genetic test" that can be done.*

2.4.1. Caveat #1: beware the mythology of the medical record

The reader will note several inconsistencies in the history of the case under discussion. Most notable is the history from the mother that she did not experience illness during the pregnancy. While this does not completely eliminate intrauterine infection from the differential diagnosis, it makes it less likely. Having reviewed thousands of medical records, it is eye-opening to see how frequently information is inaccurate. In many cases, an error can be identified that is propagated by the practice of copying what is in the last note. Because of this practice, many records resemble the garbled message that results from the old party game, "Telephone". Confirmation of information from primary sources, when available, is absolutely critical to arriving at an accurate diagnosis, not to mention preventing inappropriate treatment decisions based on the inaccurate information.

2.5. Physical examination

The patient appeared well. Current weight was 55 kg (30th percentile) and height was 165 cm (75th percentile). Head circumference was 53.3 cm (3rd percentile). She had upslanting palpebral fissures but no epicanthic folds. The nose was long with a pointed tip. The mandible was prominent. Cardiothoracic and abdominal exams were unremarkable. She had a moderate scoliosis of the spine. Extremities were unremarkable, except for scarring from biting. Neurologic examination was remarkable for increased muscle tone, and hyperreflexia. Strength was normal. Her gait was wide-based and unsteady, although she did not use her arms for balance. Measurements of facial features (eye spacing, ear length, philtrum length), hands, digits, and feet were within normal limits.

The genetic examination includes not only a full physical and neurologic examination but also examination for the so-called minor anomalies (see Article 79, ***The physical examination in clinical genetics****, Volume 2). Minor anomalies are defined as unusual morphologic features (usually defined as present in less than 3% of the referent population) of no serious medical or cosmetic significance (Jones, 1997; Merks et al., 2003). Patients must be compared to the relevant ethnic group (e.g., upslanting palpebral fissures and epicanthic folds are minor anomalies in Caucasians but not in Asians). Also, some minor anomalies (e.g., preauricular pits or polydactyly) can be inherited in families in an autosomal dominant fashion. Other minor anomalies, such as eye spacing, short fingers, and so on, require measurement for confirmation. Standards for comparison are available (Saul et al., 1998; Hall et al., 1989). For further information on the dysmorphologic history and physical*

examination, the reader is referred to Diagnostic Dysmorphology (Aase, 1990; see also Article 79, ***The physical examination in clinical genetics****, Volume 2). This patient has the following minor anomalies: microcephaly, upslanting palpebral fissures, unusual nose, and prominent mandible.*

Another important aspect of the genetic examination is the concept of the gestalt. The best example of this is Down syndrome, a condition familiar to physicians and lay people alike. While one can make a list of the multiple minor anomalies present in a given individual with Down syndrome, one look at the face allows the diagnosis to be made in virtually all patients, regardless of age or ethnicity. This recognition of the pattern represents the gestalt, or whole is greater than the sum of the parts. Many other genetic conditions are equally distinctive and recognizable (in fact, the most used text of dysmorphologists is titled Recognizable Patterns of Human Malformation (Jones, 1997)). However, because they are much less frequent, most nongeneticists do not recognize the specific disorders. This patient's gestalt is consistent with Angelman syndrome, the top differential diagnosis at this point.

2.6. Synthesis

Table 1 shows the relevant points of this patient's history and physical examination. Genetic diagnosis lends itself to hierarchical search strategies using a variety of public and proprietary tools (Bankier and Danks, 2000; NCBI PubMed, 2004; OMIM, 2000; Winter and Baraitser, 2002; *see also* Article 86, **Uses of databases**, Volume 2). Before demonstrating this approach, it is important to discuss the use of "handles". A handle is a slang term for a relevant item of history or physical examination. In this patient, any of the items listed in Table 1 could be used as a handle. Handles can be used singly, but are more frequently used in combination to develop and prioritize the differential diagnosis.

2.6.1. Caveat #2: choose your handles wisely

Handles vary in their utility for this purpose. Using OMIM (Online Mendelian Inheritance in Man) as an example, consider the following searches. As might

Table 1 Relevant clinical information

History	Physical
Normal pregnancy	Microcephaly
Congenital	Upslant palpebral fissures
Mental retardation–profound	Long, pointed nose
Abnormal EEG	Prominent mandible
Possible clinical seizures	Hypertonicity
No speech	Ataxia
Absence of major birth defects	
Normal neuroimaging	
Susceptible to falls	
Episodic depression	
Negative family history	

be expected, a search on "mental retardation" would yield thousands of possible diagnoses (1157 entries). Using modifiers such as profound can limit the diagnoses some (97 entries) but not to the point of generating a useable list. Searching on the term ataxia would yield a slightly smaller list of diagnoses (552 entries). Ataxia is a bit more specific than mental retardation. Similarly, through experience, geneticists recognize that upslanting palpebral fissure is not a good handle, while prominent mandible is less common, therefore more specific. The true power lies in combining more specific search terms. By searching on mental retardation and ataxia, 165 entries are identified. If one adds seizures to these two, the list is further pared to 72 entries, with Angelman syndrome appearing second on the list. Adding the physical feature "prominent mandible" to the search produces one entry in OMIM – the gene for Angelman syndrome. The problem with being too specific is that other potential diagnoses can be missed. For example, if PubMed (NCBI PubMed, 2004) is searched on that specific combination, there are no "hits". It is always appropriate to try different combinations of search terms and identify common diagnoses that appear. Table 2 shows the results of the search "mental retardation + seizures + ataxia" in several different tools. Individual diagnoses are grouped

Table 2 Top search results for mental retardation + seizures + ataxia

PubMed-limit Human (93 citations)[a]	OMIM (72 hits first 10 listed)[b]	POSSUM (84 hits)[c]
Rett syndrome (14)	Ataxia-telangiectasia	Chromosomal (11)
Angelman syndrome (12)	Partington X-linked MR syndrome	Metabolic (18)
Mitochondrial (6)	Rett syndrome	Teratogen (2)
Dentatorubral-pallidoluysian atrophy (4)	Disorders of pyruvate metabolism	Angelman syndrome
Other known single gene disorders (16)	Leukoencephalopathy vanishing white matter	Rett syndrome
Metabolic disorders (15)	Angelman syndrome	Single gene disorders/unique cases (51)
Unique case reports/new syndromes (7)	Polymicrogyria frontoparietal	
Chromosome disorders excluding Angelman (1)	ATP synthase 6	
Acquired disorders (7)	Neimann–Pick Type C	
Not relevant (9)	Dentatorubral-pallidoluysian atrophy	

[a]PubMed search produces a citation list. Individual disorders with more than four citations are listed separately, while the rest are grouped into larger categories.
[b]Online Mendelian Inheritance in Man (OMIM) searches are ranked according to "relevance". The database decisions that determine relevance are not explicated. This leads to differences in hierarchy when compared to the other two tools. Chromosome abnormalities are not listed in OMIM for the most part.
[c]POSSUM Mental retardation was specified as moderate-severe. Dentatorubral-pallidoluysian atrophy did not appear. This syndrome is not included in this database. If microcephaly and prognathism are added, only five disorders have all five traits, Angelman syndrome, Carbohydrate deficient glycoprotein syndrome, deletion 18q, microcephaly with chorioretinopathy, and X-linked mental retardation-Christianson type.

Table 3 Differential diagnosis

Rett syndrome
Angelman syndrome
Dentatorubral-pallidoluysian atrophy
Other chromosomal anomalies
Metabolic disorder
Monogenic disorder

into major diagnostic categories. These are combined into the differential diagnosis in Table 3.

2.7. Confirmation

The purpose of the preceding exercise is to generate and rank order a differential diagnosis. This allows for a rational approach to diagnostic testing. There are so many available tests, with more coming on line each day that a nondirected approach will result in great expenditure of resources with minimal likelihood of success. Put another way, there is no genetic "chem-20". Let us examine the differential diagnosis from Table 3.

Chromosome deletion or duplication: This is a very common consideration in patients such as this one. High-resolution chromosome analysis is a general screen for these disorders. If the deletion or duplication is large enough, a cytogeneticist will be able to identify the affected chromosomal segment, which will allow comparison to other cases reported in the literature. It is now known that many of these abnormalities are smaller than the resolution of standard microscopy (*see* Article 17, **Microdeletions**, Volume 1). Molecular genetics is providing tools such as fluorescence-in-situ-hybridization (FISH) (*see* Article 22, **FISH**, Volume 1) and comparative genomic hybridization (CGH) (*see* Article 23, **Comparative genomic hybridization**, Volume 1) to identify smaller abnormalities. Some of these techniques require a suspected diagnosis (e.g., Velo-Cardio-Facial syndrome can be diagnosed by using a specific FISH probe for chromosome 22q11 that will detect the relevant deletion (*see* Article 87, **The microdeletion syndromes**, Volume 2)), while others may be used for whole-genome screening (although the resolution of these techniques currently is not sufficient to supplant standard chromosome analysis).

2.8. Monogenic disorder

The tests to confirm these disorders, for the most part, depend on making a tentative diagnosis. Exceptions would include biochemical analysis of amino acids (*see* Article 70, **Advances in newborn screening for biochemical genetic disorders**, Volume 2) and organic acids (*see* Article 70, **Advances in newborn screening for biochemical genetic disorders**, Volume 2) that can screen for a variety of disorders. The tests can include simple biochemical assays (serum lactate and pyruvate for disorders of pyruvate metabolism (*see* Article 70, **Advances in newborn screening for biochemical genetic disorders**, Volume 2)), protein electrophoresis (transferrin

electrophoresis for Carbohydrate Deficient Glycoprotein syndrome (*see* Article 70, **Advances in newborn screening for biochemical genetic disorders**, Volume 2)), cell-based functional assays (mitochondrial and respiratory chain disorders (*see* Article 70, **Advances in newborn screening for biochemical genetic disorders**, Volume 2)), and direct mutation analysis (MECP2 sequencing for Rett syndrome (*see* Article 89, **Familial adenomatous polyposis**, Volume 2) or measurement of the size of the trinucleotide repeat for dentatorubro-pallidoluysian atrophy (*see* Article 89, **Familial adenomatous polyposis**, Volume 2)).

2.9. Angelman syndrome

The pathogenesis of this disorder demonstrates causal heterogeneity. It can be caused by deletions, both macroscopic and microscopic, in the maternal chromosome 15q11-13, paternal uniparental disomy 15 (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1), abnormal imprinting of chromosome 15 (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1), and mutations in the *UBE3A* gene (*see* Article 89, **Familial adenomatous polyposis**, Volume 2). Thus, Angelman syndrome can be a chromosomal disorder (most common) or a monogenic disorder. To further complicate matters, similar clinical features have been seen in females with mutations in the *MECP2* gene on the X chromosome that is usually associated with Rett syndrome (*see* Article 89, **Familial adenomatous polyposis**, Volume 2). This phenocopy has features of both Angelman and Rett syndrome. On the basis of the differential diagnosis, the best first test would be high-resolution chromosome analysis, followed by specific testing for Angelman syndrome if the chromosomes were negative. In this patient, chromosome analysis revealed a visible deletion of 15q11-13, confirming the diagnosis of Angelman syndrome. (For the reader interested in more information on testing for Angelman syndrome, please see Diagnostic Testing for Prader-Willi and Angelman Syndromes: Report of the ASHG/ACMG Test and Technology Transfer Committee **http://www.acmg.net/resources/policies/pol-024.asp**).

2.10. Rationale

The core tenet of clinical genetics is that establishing a specific and accurate diagnosis confers significant benefit to the patient and family. These may include elimination of unnecessary tests, information about recurrence risks for reproductive decision making, initiation of therapies to improve health, condition-specific anticipatory guidance (Cassidy and Allanson, 2001), and insight into prognosis for the disorder. In this case, the diagnosis led to several tangible benefits. The anticonvulsant taper was stopped, as seizures are a universal feature of Angelman syndrome. Additionally, a medication being used for management of behavior that lowers the seizure threshold was changed to one that does not have that effect. The patient's siblings were informed that they were not at increased risk for having an affected child with Angelman syndrome. Finally, and perhaps most importantly, at the informing session with the parents, both parents wept with joy that a diagnosis

had been made. The mother expressed how the guilt she experienced because of the diagnosis of birth injury had never left her. A simple explanation that this occurred by accident and was outside of anyone's control provided release of this guilt that 45 years had not diminished.

2.10.1. Caveat #3 (aka Bryan Hall's rule #1): no diagnosis is better than the wrong diagnosis

In this case, the "diagnosis" of birth injury stifled clinical investigation of the etiology of this patient's problems for over 40 years. She and her family were unable to benefit from the information a specific diagnosis would have provided for this entire time. It should be noted, however, that Angelman syndrome was not described until this patient was 10 years old (Angelman, 1965).

2.10.2. Caveat #4: there is no statute of limitations on diagnostic investigations

It is important to recognize that in many cases serial evaluation of patients is necessary to establish a diagnosis. Frequently, the changing phenotype with age provides the necessary clues to a diagnosis. As important is the rapid increase of knowledge about the genetic cause of disorders and description of new disorders that will, as in this case, allow for a diagnosis to be made in time.

In conclusion, the evaluation of chromosomal and monogenic disorders requires a systematic approach to gather information, identify handles, develop a differential diagnosis, prioritize diagnostic investigations, and, hopefully, arrive at a specific diagnosis.

Further reading

Garrod A (1909) *Inborn Errors of Metabolism*, First Edition, Frowde: London.

Gorlin RJ, Cohen MM and Hennekam RCM (2001) *Syndromes of the Head and Neck*, Fourth Edition, Oxford University Press: New York.

Gorlin RJ and Pindborg JJ (1964) *Syndromes of the Head and Neck*, First Edition, McGraw-Hill: New York.

LeJeune J, Gauier M and Turpin R (1959) Etude des chromosomes somatiques de enfants mongoliens. *Comptes rendus de l'Academie des sciences*, **248**, 1721–1722.

McKusick VA (1966) *Mendelian Inheritance in Man*, First Edition, Johns Hopkins University Press: Baltimore.

Saul RA, Seaver LH, Sweet KM, Geer JS, Phelan MC and Mills CM (1998) *Growth References Third Trimester to Adulthood*, Second Edition, Greenwood Genetic Center: Greenwood.

Scriver CR, Beaudet AL, Sly WL, Valle D (Eds.) Childs B, Kinzler KW, Vogelstein B (Assoc. Eds.) (2001) *The Metabolic & Molecular Basis of Inherited Disease*, McGraw-Hill: New York.

Smith DW (1970) *Recognizable Patterns of Human Malformation*, First Edition, W. B. Saunders Co: Philadelphia.

Stanbury JB, Wyngaarden JB, Fredrickson DS (Eds.) (1960) *The Metabolic Basis of Inherited Disease*, First Edition, McGraw-Hill Book Co: New York.

References

Aase JM (1990) *Diagnostic Dysmorphology*, First Edition, Plenum Medical Book Co: New York.

Angelman H (1965) "Puppet" children: A report on three cases. *Developmental Medicine and Child Neurology*, **7**, 681–688.

Bankier A and Danks D (2000) *POSSUM* ™ *Version 5.5.* Telemedia Software Laboratories and The Murdoch Institute. **http://www.possum.net.au**.

Cassidy SB and Allanson JE (2001) *Management of Genetic Syndromes*, First Edition, Wiley-Liss: New York.

Gelehrter TD and Collins FS (1990) *Principles of Medical Genetics*, First Edition, Williams and Wilkins: Baltimore.

Hall JG, Froster-Iskenius UG and Allanson JE (1989) *Handbook of Normal Physical Measurements*, First Edition, Oxford University Press: Oxford.

Jones KL (1997) *Smith's Recognizable Patterns of Human Malformation*, Fifth Edition, W. B. Saunders Co: Philadelphia.

Merks JHM, Van Karnebeek CDM, Caron HN and Hennekam RCM (2003) Phenotypic abnormalities: terminology and classification. *American Journal of Medical Genetics*, **123A**, 211–230.

NCBI PubMed National Library of Medicine (2004) **http://www.ncbi.nlm.nih.gov/PubMed/medline.html**.

Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), (2000). World Wide Web URL: **http://www.ncbi.nlm.nih.gov/omim/**.

Russell T, Dillon A and Bryant CA (1986) The relationship between chronological age, length of institutionalization and measured intellectual functioning among moderately mentally retarded adults. *Journal of Mental Deficiency Research*, **30**, 331–339.

Saul RA, Seaver LH, Sweet KM, Geer JS, Phelan MC and Mills CM (1998) *Growth References Third Trimester to Adulthood*, Second Edition, Greenwood Genetic Center: Greenwood.

Scheuner MT, Wang SJ, Raffel LJ, Larabell SK and Rotter JI (1997) Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. *American Journal of Medical Genetics*, **71**, 315–324.

Silverstein AB (1962) Length of institutionalization and intelligence test performance in mentally retarded adults. *American Journal of Mental Deficiency*, **67**, 618–620.

Silverstein AB (1969) Changes in the measured intelligence of institutionalized retardates as a function of hospital age. *Developmental Psychology*, **1**, 125–127.

Sternlicht M and Siegel L (1968) Institutional residence and intellectual functioning. *Journal of Mental Deficiency Research*, **12**, 119–127.

Winter RM and Baraitser M (2002) *The London Dysmorphology Database v.3.0*, Oxford University Press: Oxford.

Introductory Review

Approach to common chronic disorders of adulthood

Maren T. Scheuner
UCLA School of Public Health, Los Angeles, CA, USA

1. Introduction

Common disease genetics is the study of genetic aspects of diseases that are major public health concerns, such as coronary heart disease, stroke, cancer, and diabetes (American Heart Association, 2002; American Cancer Society, 2003). These are diseases of affluence that have increased in prevalence since industrialization of our society because of inactivity, excess calories, processed foods, tobacco use, radiation, and pollution. These diseases are chronic conditions that develop over decades, usually occurring in adulthood. Early disease detection and prevention are possible because of the chronic nature of these disorders.

Common chronic diseases are considered complex genetic conditions because of genetic and nongenetic/environmental risk factors. For any given disorder, in some individuals, the interaction of multiple genetic factors may explain most of the disease susceptibility (i.e., polygenic inheritance), whereas in others, environmental factors predominate. However, in most cases, disease is due to the interaction of genetic and environmental risk factors (i.e., multifactorial inheritance). This is in contrast to the classical paradigm of genetic disease, where a single gene mutation is usually sufficient for disease expression. Rarely, common chronic diseases of adulthood can occur primarily as the manifestation of a single gene mutation. However, even in these cases, there is substantial clinical heterogeneity (Scheuner *et al*., 2004).

The ability to identify genetic susceptibilities to common chronic diseases has the potential to improve our efforts in diagnosing, treating, and preventing these conditions. Because the environmental risk factors for common chronic diseases have become so pervasive in our society, our genetic susceptibilities are readily revealed. This is especially true for most individuals diagnosed with a common chronic disease at an early age (about 10 to 20 years earlier than the typical age of diagnosis). In addition, because a variety of chronic diseases can be linked to each environmental risk factor (e.g., coronary artery disease, lung cancer, and emphysema are all associated with smoking), genetic susceptibilities become paramount in determining which individuals have the greatest risk for developing a specific disease, given a particular exposure or lifestyle.

2. Genetic risk assessment strategies

Genetic susceptibility to common disease can be assessed by several approaches, including DNA-based testing, phenotypic assessment of biochemical and physiological traits, as well as personal and family history collection and interpretation. Each approach may be used in the clinical setting when performing individualized risk assessment. However, in population-based settings, the approach used to identify individuals with a genetic risk is influenced by the prevalence of genetically determined disease and the accuracy, reliability, and acceptability of the screening method.

Currently, family history collection and interpretation is the most practical population-based strategy for identifying individuals with a genetic susceptibility to many common chronic diseases (Scheuner *et al*., 1997a). It represents complex interactions of genetic, environmental, cultural, and behavioral factors shared by family members. For many common diseases, a positive family history is quantitatively significant with relative risks ranging from 2 to 5 times those of the general population, and this risk generally increases with an increasing number of affected relatives and earlier ages of disease onset (Table 1). Family history characteristics that increase disease risk include early age at diagnosis, two or more close relatives affected with a disease or a related condition, a single family member with two or more related diagnoses, multifocal or bilateral disease, and occurrence of disease in the less-often affected sex. By recognizing the magnitude of risk associated with these familial characteristics, stratification into different familial risk groups (e.g., low, intermediate, and high) is possible (Table 2), which can guide risk-specific recommendations for disease management and prevention (Figure 1). Referral for genetic evaluation by a geneticist or other specialist should be considered for individuals with high familial risk.

Family history is a prevalent and relatively accurate predictor of risk for several chronic conditions, including many forms of cancer, coronary heart disease, stroke,

Table 1 Risk estimates due to family history for selected common chronic diseases of adulthood

Disease	Risk due to family history
Coronary heart disease	OR = 2.0 (one first-degree relative) (Ciruzzi *et al*., 1997)
	OR = 5.4 (two or more first-degree relatives) (Silberberg *et al*., 1998)
Type 2 diabetes	RR = 2.4 (mother) (Klein *et al*., 1996)
	RR = 4.0 (maternal and paternal relatives) (Bjornholt *et al*., 2000)
Osteoporosis	OR = 2.0 (female first-degree relative) (Keen *et al*., 1999)
	RR = 2.4 (father) (Fox *et al*., 1998)
Asthma	OR = 3.0 (mother) (Tariq *et al*., 1998)
	RR = 7.0 (father and mother) (*American Journal of Respiratory and Critical Care Medicine*, 1997)
Breast cancer	RR = 2.1 (one first-degree relative) (Pharoah *et al*., 1997)
	RR = 3.9 (three or more first-degree relatives) (Collaborative Group, 2001)
Colorectal cancer	OR = 1.7 (one first-degree relative) (Fuchs *et al*., 1994)
	OR = 4.9 (two first-degree relatives) (Sandhu *et al*., 2001)
Prostate cancer	RR = 3.2 (one first-degree relative) (Cerhan *et al*., 1999)
	RR = 11.0 (three first-degree relatives) (Steinberg *et al*., 1990)

Table 2 Suggested guidelines for risk stratification based on family medical history of common chronic diseases of adulthood

Low familial risk

1. No affected relatives.
2. Family medical history unknown.
3. Adopted person with unknown family medical history.
4. Only one affected second-degree relative with any age of disease onset from one or both sides of the pedigree if the disease is not sex-limited.
5. If the disease is sex-limited, only one affected second-degree relative with any age of disease onset from the parental lineage of the more-often affected sex.
6. If the disease is sex-limited, only one affected second-degree relative with a late or unknown age of disease onset from the parental lineage of the less-often affected sex.

Intermediate familial risk

1. Only personal history with a later age of disease onset.
2. Only one affected first-degree relative with late or unknown age of disease onset.
3. Only two affected second-degree relatives from one lineage with a late or unknown age of disease onset.
4. If the disease is sex-limited, only one affected second-degree relative with an early age of disease onset from the parental lineage of the less-often affected sex.

High familial risk

1. Personal history with an early age of disease onset.
2. Personal history with a late or unknown age of onset and personal history of a related condition at any age of onset.
3. Personal history with a late or unknown age of disease onset and one affected first- or second-degree relative with an early age of disease onset, or one affected first- or second-degree relative with a late or unknown age of disease onset and a related condition in the same relative.
4. One affected first-degree relative with an early age of disease onset.
5. If the disease is sex-limited, one affected second-degree relative with an early age of disease onset in the less-often affected sex.
6. Two affected first-degree relatives with a late or unknown age of disease onset.
7. One affected first-degree relative with a late or unknown age of disease onset and one affected second-degree relative with an early age of disease onset, or one affected second-degree relative with a late or unknown age of disease onset and a related condition in the same relative.
8. Two affected second-degree relatives from the same lineage with at least one having an early age of disease onset, or both having a late or unknown age of disease onset and at least one having a related condition.
9. Three or more affected first- and/or second-degree relatives from one lineage with any age of disease onset.
10. One first- or second-degree relative with an early age of disease onset and two first- and/or second-degree relatives from the same lineage with related conditions.
11. "Intermediate familial risk" on both sides of the pedigree.

Adapted from Scheuner *et al*. (1997a). Early age of disease onset refers to disease that occurs about 10 to 20 years earlier than typical. Examples of sex-limited diseases include coronary heart disease, osteoporosis, breast cancer, ovarian cancer, and prostate cancer. Examples of related conditions include coronary heart disease, diabetes and stroke, or colorectal cancer, endometrial cancer, and ovarian cancer.

and diabetes. About 43% of healthy, young adults will have a family history for one of these disorders. Depending upon the specific disorder, approximately 5–15% will have a moderate familial risk (about 2 times the population risk) and 1–10% a high familial risk (about 3–5 times the population risk, approaching risks associated with Mendelian disorders) (Scheuner *et al*., 1997b). Most sensitivity values for a positive

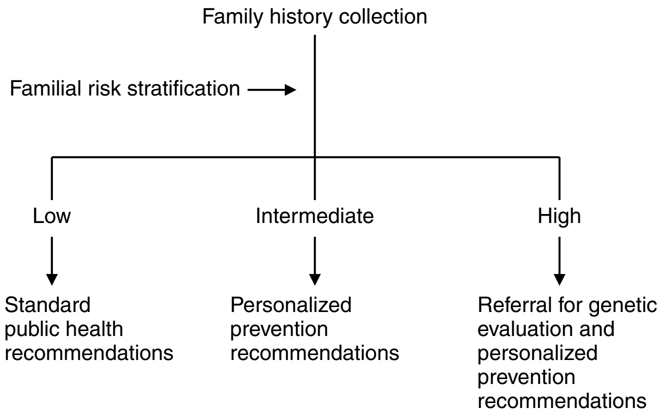


Figure 1 Family history collection followed by risk stratification that recognizes family history characteristics that increase disease risk (e.g., early age at diagnosis, two or more close relatives affected with a disease or a related condition, a single family member with two or more related diagnoses, multifocal or bilateral disease, and occurrence of disease in the less often affected sex) can guide risk-specific recommendations for disease management and prevention. Standard public health messages would be appropriate for individuals with a low familial risk. Personalized prevention recommendations should be provided to individuals with intermediate and high familial risk such as lifestyle changes if indicated, earlier and more frequent screening if appropriate, and use of chemoprevention when available. Referral for genetic evaluation by a geneticist or other specialist should be considered for individuals with high familial risk

family history of these diseases in a first-degree relative range from 70 to 85% and specificity is usually 90% or greater (Love *et al*., 1985; Hunt *et al*., 1986; Acton *et al*., 1989; Kee *et al*., 1993; Bensen *et al*., 1999). Overall, the available studies suggest that a positive family history can generally be used with a high degree of confidence for the identification of individuals at increased risk for developing many common chronic diseases.

Although numerous susceptibility alleles for common chronic diseases have been discovered, there has been limited progress in the discovery of genes for non-Mendelian forms of these diseases that have meaningful clinical relevance (Glazier *et al*., 2002). Thus, DNA-based testing for common chronic diseases is not a practical population-based approach for genetic risk assessment. Before testing for low-risk susceptibility genes has widespread clinical application, additional studies are needed to assess the prevalence and penetrance of these genotypes, as well as the effect of other genes and environmental factors on their expression. Furthermore, the clinical utility of DNA-based testing for disease susceptibility compared to other risk assessment strategies, including familial risk assessment and assessment of biochemical risk factors, must be proven.

Currently, genetic testing for chronic disease susceptibility is generally only available for rare Mendelian disorders. Personal and family history characteristics are crucial for identifying Mendelian disorders (Scheuner *et al*., 2004); therefore, genetic testing will likely remain a clinical intervention based on familial characteristics for many years to come. Thus, use of family history is central to providing access to genetic testing services that are currently available, and it is likely the

paradigm of familial risk assessment will inform future genetic testing of less penetrant susceptibility alleles.

3. Genetic evaluation for common chronic diseases

Clinical genetic evaluation for common disease should be performed for individuals with a high familial risk or when a Mendelian disorder is suspected. Genetic evaluation is composed of several components. The process includes (1) genetic counseling and education, (2) risk assessment and diagnosis using personal and family medical history, physical examination, and genetic testing, and (3) recommendations for management and prevention options appropriate for a genetic risk.

3.1. Genetic counseling and education

An important goal of genetic evaluation for common chronic diseases is the development of individualized preventive strategies based on the genetic risk assessment, the patient's personal medical history, lifestyle, and preferences. Genetic counseling is critical for delineating a patient's motivation and likely responses to learning of a genetic risk. Through genetic counseling, patients will be educated about the role of behavioral and genetic risk factors for disease, the mode of inheritance of genetic risk factors, and the options for prevention and risk factor modification.

This is a necessary process for individuals at risk for common chronic diseases since there is evidence that awareness of increased risk because of genetic or familial factors does not automatically translate to spontaneous improvement in lifestyle choices (Kip *et al*., 2002; West *et al*., 2003). For example, the occurrence of a heart attack or stroke in an immediate family member did not lead to self-initiated, sustained change in modifiable risk factors in young adults (Kip *et al*., 2002). Among low-income, rural African–American women who had not had mammogram recently, knowledge of family history of breast cancer was not associated with perceived risk or screening (West *et al*., 2003). These results argue that counseling and education are needed to actively intervene in people with a family history of common chronic disease, where the opportunities for prevention are substantial (Tavani *et al*., 2004a,b; Slattery *et al*., 2003).

Genetic counseling ensures the opportunity to provide informed consent, including discussion of the potential benefits, risks, and limitations of genetic risk assessment. Knowledge of genetic susceptibility to a common chronic disease has the potential to improve diagnosis, management, and prevention efforts. Psychological benefits can also result from knowledge of a genetic risk (Lerman *et al*., 1996; Croyle *et al*., 1997). Confirming a suspected genetic risk can be empowering and may relieve anxiety related to not knowing. In some cases, genetic risk assessment can also reassure individuals for whom a familial susceptibility can be excluded. On the other hand, potentially harmful psychological effects, such as increased anxiety, can result from knowing of a genetic risk, particularly if there are no proven interventions available for management or prevention, or if such interventions are inaccessible or unacceptable (e.g., prophylactic oophorectomy for a woman who has

not completed childbearing). Family dynamics may change from knowledge of a genetic risk for disease. For example, a parent may feel guilt about passing on a disease predisposition, or a sibling for whom genetic susceptibility has been excluded may experience survivor guilt if the susceptibility is identified in another sibling. Family members may experience loss of privacy when asked to share their medical history and medical records, and if labeled as having a genetic risk for disease, family, friends, or society may stigmatize them. There is also the possibility of misuse of genetic risk information by third parties such as employers, educators, and insurers that could exclude individuals from employment or educational opportunities, or from obtaining health, life or disability insurance, although the evidence of genetic discrimination against otherwise healthy individuals is minimal (Billings *et al*., 1992; Geller *et al*., 1996; Epps, 2003). These potential harms should be considered and weighed against the potential benefits when providing genetic risk assessment.

3.2. Pedigree analysis

Pedigree analysis is typically the first step in genetic risk assessment and diagnosis. The pedigree structure is created and usually includes all first- and second-degree relatives, spanning 3–4 generations. Demographic information for each family member is documented that typically includes each relative's current age or age at death. Medical history is documented for each family member, including age at diagnosis, cause of death if deceased, and known interventions or procedures, which can help clarify a diagnosis. For example, questioning regarding coronary artery bypass surgery, angioplasty, heart transplant, or pacemaker placement may help clarify a relative's diagnosis of heart disease. Information is also collected regarding important risk factors for a disease, such as use of hormone replacement therapy or chest irradiation for breast cancer, and smoking, asbestos exposure, and coal mining for lung cancer. Medical records are reviewed when possible to verify the medical history of each family member or at least those who are critical to the genetic risk assessment and diagnosis. The family history should include ethnicity and country of origin of grandparents since certain conditions might be more prevalent in certain ethnic groups. For example, the prevalence of insulin resistance is high among individuals of Native American admixture (Arnoff *et al*., 1977) and Asian Indian origin (Sharp *et al*., 1987), and there are common *BRCA* gene founder mutations in Ashkenazi Jewish families with breast and ovarian cancer (Struewing *et al*., 1997).

Once this information is collected, pedigree analysis is performed to determine the most likely mode of inheritance (i.e., Mendelian versus multifactorial) and the risk of disease to the patient and to unaffected relatives based on their position in the pedigree. This analysis also helps to elucidate the differential diagnosis through pattern recognition (Scheuner *et al*., 2004). For example, when considering an inherited form of breast cancer, there are at least seven different Mendelian disorders to consider, including hereditary site-specific breast cancer, hereditary breast–ovarian cancer syndrome, Li–Fraumeni syndrome, Cowden syndrome, Peutz–Jeghers syndrome, hereditary nonpolyposis colon cancer, and ataxia telangiectasia (Hoskins *et al*., 1995; Scheuner *et al*., 2004). The types of cancers

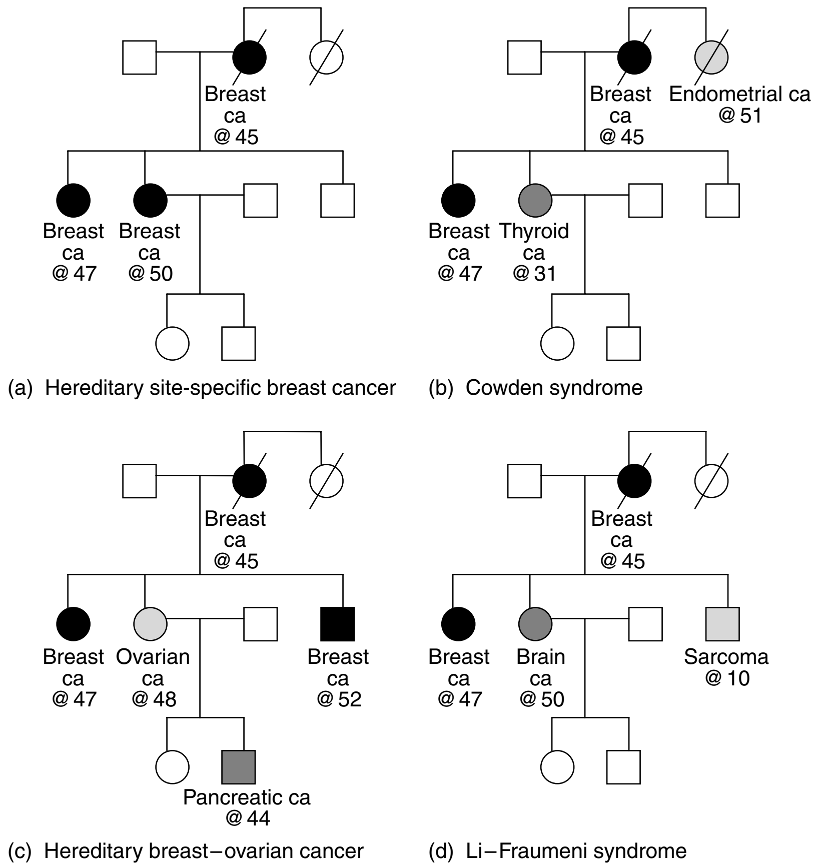


Figure 2 Each pedigree shown has a high familial risk for breast cancer. However, by recognition of the patterns of cancer in the family, a more accurate diagnosis can be made. Pedigree (a), which features early onset breast cancer and no other cancers is most consistent with hereditary site-specific breast cancer, which is often due to *BRCA1* or *BRCA2* gene mutations. Pedigree (b) features early onset breast, thyroid and endometrial cancer and is most consistent with Cowden syndrome due to *PTEN* gene mutations. Pedigree (c) features early onset breast and ovarian cancer and is most consistent with hereditary breast–ovarian cancer syndrome, which is almost always due to *BRCA1* or *BRCA2* gene mutations. In this case, a *BRCA2* gene mutation is likely given the family history of male breast cancer and pancreatic cancer. The history of early onset breast cancer, brain tumor, and childhood sarcoma in pedigree (d) is most consistent with Li–Fraumeni syndrome due to *TP53* gene mutations. Multiple primary cancers are common among individuals with Li–Fraumeni syndrome

and other conditions reported in the family help distinguish each of these syndromes (Figure 2). Mutations in different genes underlie the genetic susceptibility in these syndromes and genetic testing can help to confirm a suspected diagnosis.

For pedigrees consistent with multifactorial inheritance or that lack convincing evidence of Mendelian inheritance, quantitative risk assessment can be performed for specific conditions using mathematical models or published estimates (Amos *et al*., 1992; Claus *et al*., 1993, 1994; St John *et al*., 1993). For example, a

woman's absolute risk for breast cancer by age 80 or in the next 10 years can be provided on the basis of the family history of breast or ovarian cancer in first- and second-degree relatives and their age at diagnosis (Claus *et al*., 1993, 1994), and she can contrast this to the population risk of breast cancer by age 80 or in the next 10 years. Because family history is often the most significant predictor of risk, these models provide good estimates of risk. However, most estimates using these models have limitations, and they should not be the only means for risk assessment. In particular, most do not account for significant exposures or behaviors that might influence disease risk.

3.3. *Personal history*

In addition to review of past medical history and medical records for confirmation, assessment of signs and symptoms of the disease of concern should be performed to more accurately assess risk for the patient. For example, when evaluating a genetic risk for heart disease, review of systems should include questions regarding angina, shortness of breath, dyspnea on exertion, paroxysmal nocturnal dyspnea, pedal edema, claudication, and exercise tolerance. In the case of risk assessment for colorectal cancer, questions should be asked regarding frequency of bowel movements, caliber of the stool, color of the stool, and presence of blood in the bowel movement. If symptoms are present, follow-up confirmatory testing should be recommended. For example, exercise treadmill testing or echocardiogram to evaluate cardiovascular symptoms, or colonoscopy to assess change in bowel habits, or blood in the stool.

3.4. *Physical examination*

The physical examination should be performed to identify signs of the disease of concern as well as characteristic manifestations of Mendelian forms of a disease. For example, an evaluation for cardiovascular risk should include auscultation of the heart, lungs, and major vessels in the neck, abdomen, and groin, and palpation of the aorta and distal pulses. Any abnormalities can be followed up with additional studies, such as ultrasound. Blood pressure in the upper and lower extremities can identify hypertension, and these measurements can be used to calculate the ankle/brachial blood pressure index (ABI). Values <0.9 are correlated with atherosclerosis. Weight and height should be obtained and body mass index should be calculated to identify overweight and obese patients, and follow-up measurements can help monitor diet and exercise interventions. Waist circumference should be obtained, as increased values are associated with the metabolic syndrome (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001), a common cause of cardiovascular disease (Park *et al*., 2003). Evaluation of possible lipid disorders should include examination of the eyes, assessing corneal arcus and lipemia retinalis. Examination of the skin should include assessment for xanthelasma and tendonous xanthomas. Physical signs of Mendelian disorders that feature cardiovascular disease should

be assessed, such as dolichostenomelia and arachnodactyly associated with Marfan syndrome, abnormal scarring, and translucent skin associated with Ehlers–Danlos syndrome type IV, papular skin lesions and plaques in flexural creases and angioid streaks on the retina associated with pseudoxanthoma elasticum, and angiokeratomas (vascular cutaneous lesions), and corneal and lenticular opacities associated with Fabry disease.

3.5. Genetic testing

Identification of appropriate tests that may be helpful in confirming a genetic diagnosis and refining genetic risk is the initial step in the genetic testing process. Genetic tests for assessing common diseases usually include DNA-based tests and biochemical tests. DNA-based testing is typically performed when assessing a Mendelian form of cancer susceptibility, whereas biochemical testing is more often available to assess a susceptibility to cardiovascular disease or diabetes (Scheuner *et al*., 2004).

Most commercial laboratories performing DNA-based testing utilize polymerase chain reaction (PCR)-based methods to amplify the patient's DNA to be followed by sequencing or other methodologies. In most cases, sequencing is considered the gold standard for identification of a mutation. However, even sequencing may miss rearrangements or large deletions of DNA or mutations in regulatory regions (Yan *et al*., 2000; Gad *et al*., 2001). Rarely, errors may occur with sample handling, contamination by airborne particles in the laboratory, or failure of the PCR technique.

When considering genetic testing, clinicians should choose the testing strategy that would be most informative for a patient and their family members. Ideally, genetic testing should begin in an affected family member to identify the specific genetic determinant(s) of disease susceptibility and to recommend appropriate management. If an abnormality or abnormalities are identified, then at-risk relatives can be tested for those familial factors. Normal results can provide reassurance regarding disease risk due to those factors; however, there will always be a background or population risk for disease development. At-risk relatives with abnormal results can begin surveillance and prevention activities appropriate for their risk.

It is not always possible to test an affected family member, as they are often deceased and therefore unable to provide a sample, or not interested in testing, or unable to participate because of financial reasons. In this case, unaffected family members may participate in testing. Identification of a genetic risk factor or factors may help with the risk assessment and in planning appropriate screening and preventive strategies. Though unaffected individuals can initiate the testing process, normal test results cannot be entirely reassuring since the available clinical testing is not comprehensive. A normal test result can exclude the genetic risk factors that have been tested but not the possibility of an inherited susceptibility. In such a case, medical management would typically depend upon the empirical risks associated with the family history. To further clarify the situation, testing additional family members, both affected and unaffected, may be helpful.

Once a genetic testing strategy has been formulated, laboratories that can provide the testing are identified. Obtaining, processing, and shipping the sample may also be necessary, as well as facilitation of the billing process (providing the most appropriate diagnostic codes and documenting medical necessity for testing when requested). Issues to consider when choosing a laboratory include methodology, analytic sensitivity and specificity, technical support, cost, and turn-around time. Understanding which methodology is used is important since standards for many genetic tests currently do not exist. There is great variability in testing procedures for many conditions. As a result, clinicians need to be personally familiar with a given laboratory's protocol and the test limitations.

An individual's genetic test results should be interpreted within the context of their family history. The interpretation should incorporate a discussion of the impact of the results on their risk assessment, diagnosis, and plans for management and prevention. Family members who might benefit from risk assessment and genetic testing should also be identified, and the clinician should facilitate communication between the index case and at-risk relatives as well as provide referrals to genetics professionals or other specialists for relatives.

Interpretation of genetic test results for disease susceptibility alleles should consider the variable effects of an allele. For example, polymorphisms in the methylenetetrahydrofolate reductase (MTHFR) gene are associated with cardiovascular disease, neural tube defects, and unexplained recurrent embryo losses in early pregnancy (Isotalo *et al*., 2000; Eikelboom *et al*., 1999). The apolipoprotein E genotypes contribute to serum cholesterol differences in the population and have been associated with cardiovascular disease risk as well as risk for Alzheimer disease (Wilson *et al*., 1996; Farrer *et al*., 1997). Recognizing these associations has important implications regarding informed consent when offering specific tests for the purpose of refining a diagnosis or for presymptomatic detection of disease.

3.6. Management and prevention

Identifying individuals with genetic susceptibilities to common chronic diseases may have profound implications not only for their health but the health of their family. Knowledge of a genetic risk can influence the clinical management and prevention of a disease. Prevention strategies include targeted lifestyle changes, screening at earlier ages, more frequently and with more intensive methods than used for average risk individuals, use of chemoprevention, and for those at highest risk, prophylactic procedures and surgeries (see Table 3 for examples). Screening and prevention guidelines are available for many chronic disorders (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001; Diabetes Prevention Program Research Group, 2002; Straus *et al*., 2002; Walsh and Terdiman, 2003; Smith *et al*., 2002; Scheuner *et al*., 2004), and data are accumulating regarding the effectiveness of these strategies in high-risk individuals (Järvinen *et al*., 2000; Rebbeck *et al*., 1999, 2002; Narod *et al*., 1998, 2000). Because many behavioral risk factors for disease aggregate in families, a family-based approach to risk factor modification ought to be an effective strategy, and this has been demonstrated in a few studies (Knutsen and Knutsen, 1991; *British*

Table 3 Examples of early detection and prevention strategies according to familial risk level

Colorectal cancer

Low familial risk: Beginning at age 50, annual FOBT and/or flexible sigmoidoscopy every 5 years, or double-contrast barium enema every 5 to 10 years, or colonoscopy every 10 years (Winawer *et al*., 1997; Rex *et al*., 2000; Smith *et al*., 2001; US Preventive Services Task Force, 2002b).

Intermediate familial risk: Beginning at age 40, annual FOBT and/or flexible sigmoidoscopy every 5 years, or double-contrast barium enema every 5 to 10 years, or colonoscopy every 10 years (Winawer *et al*., 1997; Rex *et al*., 2000).

High familial risk: Beginning at age 40 or 10 years before the earliest age of diagnosis in the family, colonoscopy every 3 to 5 years (Rex *et al*., 2000). Consider daily use of aspirin (Baron *et al*., 2003; Sandler *et al*., 2003), and daily intake of supplemental folic acid 400 mcg (Giovannucci *et al*., 1998) and calcium 1200 mg (Holt *et al*., 1998; Baron *et al*., 1999).

Coronary heart disease

Low familial risk: Beginning at age 45 for women and at age 35 for men, screen for lipid disorders every 5 years or shorter intervals for people with lipid levels close to those warranting therapy (US Preventive Services Task Force, 2001).

Intermediate familial risk: Beginning at age 20 for women and men, screen for lipid disorders every 5 years or shorter intervals for people with lipid levels close to those warranting therapy (US Preventive Services Task Force[a], 2003). Consider use of aspirin daily or every other day (US Preventive Services Task Force[a], 2003).

High familial risk: Beginning at age 20 for women and men, screen for lipid disorders every 5 years or shorter intervals for people with lipid levels close to those warranting therapy (US Preventive Services Task Force[a], 2001). Assess emerging risk factors (e.g., homocysteine, lipoprotein(a), C-reactive protein, LDL cholesterol density) (Scheuner, 2003). Consider early detection strategies (e.g., carotid artery duplex scanning to measure intima-media thickness, ankle-brachial index, electron beam CT to detect coronary artery calcification) (Smith *et al*., 2000; Scheuner, 2003). Consider use of aspirin daily or every other day (US Preventive Services Task Force[a], 2002a).

[a]Based on United States Preventive Services Task Force recommendations for individuals with increased risk.
For all individuals, regardless of familial risk level, assess relevant exposures, lifestyle choices and habits, and recommend changes if indicated. Consider referral for a genetic evaluation for individuals with high familial risk.

Medical Journal, 1994; Pyke *et al*., 1997). Lifestyle changes, such as dietary modification, weight control, and smoking cessation are likely to be more effective when delivered to the family than to an individual because family members can influence each other and provide ongoing support to one another.

The plan for management and prevention should be tailored to an individual's genetic risk while taking into consideration their personal medical history, family history, habits, and preferences. Other disease risks that may influence decision making regarding options for risk reduction should be identified as part of the management and prevention plan. For example, use of tamoxifen for breast cancer prevention might be contraindicated, given a personal or family history of thrombophilia. In addition, as with any treatment plan, potential risks or complications that may result from the recommended management and prevention strategies must be recognized.

Finally, the recommendation for management and prevention options specific to an individual's genetic risk should be communicated in writing, so the patient and referring physician can incorporate these recommendations into a plan for future health management. Plans for continued contact or follow-up with a geneticist may

be appropriate to review the individualized plan for management of genetic risk, and updated personal and family history information can be obtained with revision of the management and prevention plan as indicated. New information and technology available for genetic risk assessment or management and prevention of disease can also be discussed.

Acknowledgments

This work was supported in part by a Career Development Award sponsored by the Centers for Disease Control and Prevention and the Association of Teachers of Preventive Medicine Cooperative Agreement U50/CCU300860.

References

Acton RT, Go R and Roseman J (1989) Strategies for the Utilization of Genetic Information to Target Preventive Interventions. *Proceedings of the 25th Annual Meeting of the Society of Prospective Medicine, Indianapolis*, 88–100.

American Cancer Society (2003) *Cancer Facts & Figures 2003*, American Cancer Society: Atlanta.

American Heart Association (2002) *Heart and Stroke Statistics-2003 Update*, American Heart Association: Dallas.

Amos CI, Shaw GL, Tucker MA and Hartge P (1992) Age at onset for familial epithelial ovarian cancer. *Journal of the American Medical Association*, **268**, 1896–1899.

Arnoff SL, Bennett PH, Gorden P, Rushforth N and Miller M (1977) Unexplained hyperinsulinemia in normal and 'prediabetic' Pima Indians compared with normal Caucasians. *Diabetes*, **26**, 827–840.

Baron JA, Beach M, Mandel JS, van Stolk RU, Haile RW, Sandler RS, Rothstein R, Summers RW, Snover DC, Beck GJ, *et al*. for the Calcium Polyp Prevention Study Group (1999) Calcium supplements for the prevention of colorectal adenomas. *New England Journal of Medicine*, **340**, 101–107.

Baron JA, Cole BF, Sandler RS, Haile RW, Ahnen D, Bresalier R, McKeown-Eyssen G, Summers RW, Rothstein R, Burke CA, *et al*. (2003) A randomized trial of aspirin to prevent colorectal adenomas. *New England Journal of Medicine*, **348**, 891–899.

Bensen JT, Liese AD, Rushing JT, Province M, Folsom AR, Rich SS and Higgins M (1999) Accuracy of proband reported family history: The NHLBI Family Heart Study (FHS). *Genetic Epidemiology*, **17**, 141–150.

Billings PR, Kohn MA, de Cuevas M, Beckwith J, Alper JS and Natowicz MR (1992) Discrimination as a consequence of genetic testing. *American Journal of Human Genetics*, **50**, 476–482.

Bjornholt JV, Erikssen G, Liestol K, Jervell J, Thaulow E and Erikssen J (2000) Type 2 diabetes and maternal family history: an impact beyond slow glucose removal rate and fasting hyperglycemia in low-risk individuals? Results from 22.5 years of follow-up of healthy nondiabetic men. *Diabetes Care*, **23**, 1255–1259.

Cerhan JR, Parker AS, Putnam SD, Chiu BC, Lynch CF, Cohen MB, Torner JC and Cantor KP (1999) Family history and prostate cancer risk in a population-based cohort of Iowa men. *Cancer Epidemiology, Biomarkers & Prevention*, **8**, 53–60.

Ciruzzi M, Pramparo P, Rozlosnik J, Zylberstjn H, Delmonte H, Haquim M, Abecasis B, de la Cruz Ojeda J, Mele E, La Vecchia C, *et al*. (1997) Frequency of family history of acute myocardial infarction in patients with acute myocardial infarction. Argentine FRICAS (Factores de Riesgo Coronario en America del Sur) Investigators. *American Journal of Cardiology*, **80**, 122–127.

Claus EB, Risch N and Thompson WD (1993) The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Research and Treatment*, **28**, 115–120.

Claus EB, Risch N and Thompson WD (1994) Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer*, **73**, 643–651.

Collaborative Group on Hormonal Factors in Breast Cancer (2001) Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*, **358**, 1389–1399.

Croyle RT, Smith K, Botkin JR, Baty B and Nash J (1997) Psychological responses to *BRCA1* mutation testing. *Health Psychology*, **16**, 63–72.

Diabetes Prevention Program Research Group (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine*, **346**, 393–403.

Eikelboom JW, Lonn E, Genest J, Hankey G and Yusuf S (1999) Homocyste(e)ine and cardiovascular disease: A critical review of the epidemiologic evidence. *Annals of Internal Medicine*, **131**, 363–375.

Epps PG (2003) Policy before practice: genetic discrimination reviewed. *American Journal of Pharmacogenomics*, **3**, 405–418.

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001) Executuve summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *Journal of the American Medical Association*, **285**, 2486–2497.

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N and van Duijn CM (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *Journal of the American Medical Association*, **278**, 1349–1356.

Fox KM, Cummings SR, Powell-Threets K and Stone K (1998) Family history and risk of osteoporotic fracture: Study of Osteoporotic Fractures Research Group. *Osteoporosis International*, **8**, 557–562.

Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE and Willett WC (1994) A prospective study of family history and the risk of colorectal cancer. *New England Journal of Medicine*, **331**, 1669–1674.

Gad S, Scheuner M, Pages-Berhouet S, Caux-Moncoutier V, Bensimon A, Aurias A, Pinto M and Stoppa-Lyonnet D (2001) Identification of a large rearrangement of the BRCA1 gene using colour bar code on combed DNA in an American breast-ovarian cancer family previously studied by direct sequencing. *Journal of Medical Genetics*, **38**, 388–391.

Geller LN, Alper JS, Billings PR, Barash CI, Beckwith J and Natowicz MR (1996) Individual, family, and societal dimensions of genetic discrimination: a case study analysis. *Science and Engineering Ethics*, **2**, 71–88.

Giovannucci E, Stampfer MJ, Colditz GA, Hunter DJ, Fuchs C, Rosner BA, Speizer FE and Willett WC (1998) Multivitamin use, folate, and colon cancer in women in the Nurses' Health Study. *Annals of Internal Medicine*, **129**, 517–524.

Glazier AM, Nadeau JH and Aitman TJ (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.

Holt PR, Atillasoy EO, Gilman J, Guss J, Moss SF, Newmark H, Fan K, Yang K and Lipkin M (1998) Modulation of abnormal colonic epithelial cell proliferation and differentiation by low-fat dairy foods. *Journal of the American Medical Association*, **280**, 1074–1079.

Hoskins KF, Stopfer JE, Calzone KA, Merajver SD, Rebbeck TR, Garber JE and Weber BL (1995) Assessment and counseling for women with a family history of breast cancer. *Journal of the American Medical Association*, **273**, 577–585.

Hunt SC, Williams RR and Barlow GK (1986) A comparison of positive family history definitions for defining risk of future disease. *Journal of Chronic Diseases*, **39**, 809–821.

Isotalo PA, Wells GA and Donnelly JG (2000) Neonatal and fetal methylenetetrahydrofolate reductatse genetic polymorphisms: An examination of C677T and A1298C mutations. *American Journal of Human Genetics*, **67**, 986–990.

Järvinen HJ, Aarnio M, Mustonen H, Aktan-Collan K, Aaltonen LA, Peltomaki P, De La Chapelle A and Mecklin JP (2000) Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology*, **118**, 829–834.

Kee F, Tiret L, Robo JY, Nicaud V, McCrum E, Evans A and Cambien F (1993) Reliability of reported family history of myocardial infarction. *British Medical Journal*, **307**, 1528–1530.

Keen RW, Hart DJ, Arden NK, Doyle DV and Spector TD (1999) Family history of appendicular fracture and risk of osteoporosis: a population-based study. *Osteoporosis International*, **10**, 161–166.

Kip KE, McCreath HE, Roseman JM, Hulley SB and Schreiner PJ (2002) Absence of risk factor change in young adults after family heart attack or stroke. The CARDIA Study. *American Journal of Preventive Medicine*, **22**, 258–266.

Klein BE, Klein R, Moss SE and Cruickshanks KJ (1996) Parental history of diabetes in a population-based study. *Diabetes Care*, **19**, 827–830.

Knutsen SF and Knutsen R (1991) The Tromso Survey: the family intervention study – the effect of intervention on some coronary risk factors and dietary habits, a 6-year follow-up. *Preventive Medicine*, **20**, 197–212.

Lerman C, Narod S, Schulman K, Hughes C, Gomez-Caminero A, Bonney G, Gold K, Trock B, Main D, Lynch J, *et al*. (1996) BRCA1 testing in families with hereditary breast-ovarian cancer. A prospective study of patient decision making and outcomes. *Journal of the American Medical Association*, **275**, 1885–1892.

Love RR, Evans AM and Josten DM (1985) The accuracy of patient reports of a family history of cancer. *Journal of Chronic Diseases*, **38**, 289–293.

Narod SA, Risch H, Moslehi R, Dorum A, Neuhausen S, Olsson H, Provencher D, Radice P, Evans G, Bishop S, *et al*., for the Hereditary Ovarian Cancer Clinical Study Group (1998) Oral Contraceptives and the risk of hereditary ovarian cancer. *New England Journal of Medicine*, **339**, 424–428.

Narod SA, Brunet J, Ghadirian P, Robson M, Heimdal K, Neuhausen S, Stoppa-Lyonnet D, Lerman C, Passini B, De Los Rios P, *et al*., for the Hereditary Breast Cancer Study Group (2000) Tamoxifen and the risk of contralateral breast cancer in BRCA1 and BRCA2 mutation carriers: A case-control study. *Lancet*, **356**, 1876–1881.

British Medical Journal, Randomised controlled trial evaluating cardiovascular screening and intervention in general practice: principal results of British family heart study. 1994, **308**, 313–320.

American Journal of Respiratory and Critical Care Medicine, Genes for asthma? An analysis of the European Community Respiratory Health Survey. 1997, **156**, 1773–1780.

Park YW, Zhu S, Palaniappan L, Heshka S, Carnethon MR and Heymsfield SB (2003) The metabolic syndrome: prevalence and associated risk factor findings in the US population from the Third National Health and Nutrition Examination Survey, 1988-1994. *Archives of Internal Medicine*, **163**, 427–436.

Pharoah PD, Day NE, Duffy S, Easton DF and Ponder BA (1997) Family history and the risk of breast cancer: a systematic review and meta-analysis. *International Journal of Cancer*, **71**, 800–809.

Pyke SDM, Wood DA, Kinmonth AL and Thompson SG (1997) Change in coronary risk and coronary risk factor levels in couples following lifestyle intervention. *Archives of Family Medicine*, **6**, 354–360.

Rebbeck TR, Levin AM, Eisen A, Snyder C, Watson P, Cannon-Albright L, Isaacs C, Olopade O, Garber JE, Godwin AK, *et al*. (1999) Breast cancer risk after bilateral prophylactic oophorectomy in BRCA1 carriers. *Journal of the National Cancer Institute*, **91**, 1475–1479.

Rebbeck TR, Lynch HT, Neuhausen SL, Narod SA, van't Veer L, Garber JE, Evans G, Isaacs C, Daly MB, Matloff E, *et al*. (2002) Prophylactic oophorectomy in carriers of BRCA1 or BRCA2 mutations. *New England Journal of Medicine*, **346**, 1616–1622.

Rex DK, Johnson DA, Lieberman DA, Burt DW and Sonnenberg A (2000) colorectal cancer prevention 2000: screening recommendations of the American College of Gastrenterology. *American Journal of Gastroenterology*, **95**, 868–877.

Sandhu MS, Luben R and Khaw KT (2001) Prevalence and family history of colorectal cancer: implications for screening. *Journal of Medical Screening*, **8**, 69–72.

Sandler RS, Halabi S, Baron JA, Budinger S, Paskett E, Keresztes R, Petrelli N, Pipas JM, Karp DD, Loprinzi CL, *et al*. (2003) A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer. *New England Journal of Medicine*, **348**(10), 883–890.

Scheuner MT, Raffel LJ, Wang S-J and Rotter JI (1997a) The family history: A comprehensive genetic risk assessment method for the chronic conditions of adulthood. *American Journal of Medical Genetics*, **71**, 315–324.

Scheuner MT, Tran LT, Henderson LB and Goldstein DR (1997b) Utilization of preventive strategies by genetically susceptible individuals. *American Journal of Human Genetics*, **61**, A57.

Scheuner MT (2003) Genetic evaluation for coronary artery disease. *Genetics in Medicine*, **5**, 269–285.

Scheuner MT, Yoon P and Khoury MJ (2004) Contribution of Mendelian disorders to common chronic disease: Opportunities for recognition, intervention and prevention. *American Journal of Medical Genetics*, **125 C**, 50–65.

Sharp PS, Mohan V, Levy JC, Mather HM and Kohner EM (1987) Insulin resistance in patients of Asian Indian and European origin with non-insulin dependent diabetes. *Hormone and Metabolic Research*, **19**, 84–85.

Silberberg JS, Wlodarczyk J, Fryer J, Robertson R and Hensley MJ (1998) Risk associated with various definitions of family history of coronary heart disease: the Newcastle Family History Study II. *American Journal of Epidemiology*, **147**, 1133–1139.

Slattery ML, Levin TR, Ma K, Goldgar D, Holubkov R and Edwards S (2003) Family history and colorectal cancer: predictors of risk. *Cancer Causes and Control*, **14**, 879–887.

Smith RA, Cokkinides V, von Eschenbach AC, Levin B, Cohen C, Runowicz CD, Sener S, Saslow D and Eyre HJ (2002) American Cancer Society guidelines for the early detection of cancer. *CA: A Cancer Journal for Clinicians*, **52**, 8–22.

Smith RA, von Eschenbach AC, Wender R, Levin B, Byers T, Rothenberger D, Brooks D, Creasman W, Cohen C, Runowicz C, *et al*. (2001) American Cancer Society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers; also update 2001-testing for early lung cancer detection. *CA: A Cancer Journal for Clinicians*, **51**, 38–75.

Smith SC, Greenland P and Grundy SM (2000) AHA Conference Proceedings. Prevention Conference V. Beyond secondary prevention: Identifying the high-risk patient for primary prevention. Executive Summary. *Circulation*, **101**, 111–116.

St John DJ, McDermott FT, Hopper JL, Debney EA, Johnson WR and Hughes ES (1993) Cancer risk in relatives of patients with common colorectal cancer. *Annals of Internal Medicine*, **118**, 785–790.

Steinberg GD, Carter BS, Beaty TH, Childs B and Walsh PC (1990) Family history and the risk of prostate cancer. *Prostate*, **17**, 337–347.

Straus SE, Majumdar SR and McAlister FA (2002) New evidence for stroke prevention. *Journal of the American Medical Association*, **288**, 1388–1395.

Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Brody LC and Tucker MA (1997) The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine*, **336**, 1401–1408.

Tariq SM, Matthews SM, Hakim EA, Stevens M, Arshad SH and Hide DW (1998) The prevalence of and the risk factors for atopy in early childhood: a whole population birth cohort study. *Journal of Allergy and Clinical Immunology*, **101**, 587–593.

Tavani A, Augustin L, Bosetti C, Giordano L, Gallus S, Jenkins DJA and La Vecchia C (2004a) Influence of selected lifestyle factors on risk of acute myocardial infarction in subjects with familial predisposition for the disease. *Preventive Medicine*, **38**, 468–472.

Tavani A, Bosetti C, Dal Maso L, Giordano L, Franceschi S and La Vecchia C (2004b) Influence of selected hormonal and lifestyle factors on familial propensity to ovarian cancer. *Gynecologic Oncology*, **92**, 922–926.

United States Preventive Services Task Force (2001) *Lipid Disorders: Screening*. **http://www.ahcpr.gov/clinic/uspstf/uspschol.htm**, Accessed March 1, 2004.

United States Preventive Services Task Force (2002a) *Aspirin for Primary Prevention for Cardiovascular Events: Preventive Medication*. **http://www.ahrq.gov/clinic/uspstf/uspsasmi.htm**, Accessed March 1, 2004.

United States Preventive Services Task Force (2002b) Clinical guidelines: screening for colorectal cancer: recommendation and rationale. *Annals of Internal Medicine*, **137**, 129–131.

Walsh JME and Terdiman JP (2003) Colorectal cancer screening. *Journal of the American Medical Association*, **289**, 1288–1296.

West DS, Greene PG, Kratt PP, Pulley L, Weiss HL, Siegfried N and Gore SA (2003) The impact of a family history of breast cancer on screening practices and attitudes in low-income, rural, African American women. *Journal of Women's Health (Larchmt)*, **12**, 779–787.

Wilson PWF, Schaefer EJ, Larson MG and Ordovas JM (1996) Apolipoprotein E alleles and risk of coronary disease: a meta-analysis. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **16**, 1250–1255.

Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, Woolf SH, Glick SN, Ganiats TG, Bond JH, *et al*. (1997) Colorectal cancer screening; Clinical guidelines, evidence and rationale. *Gastroenterology*, **112**, 594–642.

Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, Lynch HT, Chadwick RB, de la Chapelle A, Berg K, *et al*. (2000) Conversion of diploidy to haploidy. *Nature*, **403**, 723–724.

Specialist Review

Current approaches to prenatal screening and diagnosis

Paula C. Cosper

University of Alabama at Birmingham, Birmingham, AL, USA

1. Introduction

Prenatal diagnosis of genetic disorders was first accomplished in 1966 when Steele and Breg performed chromosome studies on cells from an unborn fetus obtained from the amniotic fluid. Since that time, the number of disorders diagnosable prenatally has grown exponentially. Today, several methods are available for screening and detection of genetic disorders in the unborn fetus.

The most common reasons for prenatal diagnosis are:

1. maternal age 35 or over
2. previous child with a chromosome abnormality
3. parent, a known carrier of a rearrangement of chromosomes
4. family history of a detectable biochemical abnormality
5. mother, a known carrier of an X-linked disorder
6. family history of an inherited disorders detectable by ultrasound
7. abnormal first- or second-trimester screen.

Physicians who care for pregnant women should obtain information regarding any family history of genetic disorders or learning problems. Patients with a positive family history should be referred to a medical geneticist for evaluation, counseling, and possible testing. There are also screening tests that should be offered to patients of particular ethnic backgrounds. Jewish individuals, especially of Ashkenazi background, should be offered a panel of testing including Tay-Sachs disease, cystic fibrosis, Canavan's disease, and familial dysautonomia; some centers include additional tests in this panel. Caucasians should be offered screening for cystic fibrosis. Individuals of African-American descent should have sickle cell screening. Asian, Mediterranean, and Middle Eastern individuals are at increased risk for thalassemia and should be screened using a complete blood count (MCV, MCH, and hemoglobin) to determine if further testing is indicated.

2. Noninvasive screening

2.1. Second trimester

Noninvasive screening such as ultrasound and a maternal serum screen for Down syndrome, neural tube defects (NTDs), and trisomy 18 should be offered to all pregnant women regardless of family history or age. Maternal serum screening is routinely performed between 14 and 20 weeks gestation. Either three or four different analytes associated with pregnancy are measured in the maternal serum. The triple screen examines the level of alpha fetoprotein (AFP), unconjugated estriol (uE3), and human chorionic gonadotrophin (HCG) in the maternal serum. The quad screen adds Inhibin A. To take into account population differences in the screen components and laboratory differences in assay methods, the results are expressed as multiples of the median (MOM). Each laboratory that performs the testing has determined levels of each of these components for each week of pregnancy in hundreds of women who had a normal pregnancy and calculated median values for each week of pregnancy. The patient's value is compared to the median for the appropriate week of gestation and expressed as multiples of that median. Since the median values may vary significantly from week to week, it is essential that the correct gestational age be given for the patient being screened. Also taken into account in the calculation of risks from the screen are the patient's age, race, and weight. Multiple fetuses can make the screen much less accurate in predicting Down syndrome and trisomy 18. Women who receive a positive screen should be referred for further genetic testing. It should be emphasized to the patient that the maternal marker screen is not a diagnostic test, but only an estimate of risk.

AFP is a protein produced by the fetal liver and is excreted normally into the amniotic fluid through the fetal urine. If there is an opening that allows fetal fluids to escape into the amniotic fluid, the concentration of AFP will be much higher than normal in the fluid and thus in the maternal serum. Any maternal serum AFP value of 2.5 MOM or above is considered positive for open NTDs. Since insulin-dependent diabetic patients are at increased risk to have babies with NTDs, the cut-off for a positive screen in diabetic mothers is 2 MOM. The maternal serum AFP will detect at least 80% of the cases of open neural tube defects. If combined with a detailed ultrasound scan, the detection rate increases to at least 95% for open spina bifida and 100% for anencephaly. For individuals who have a family history of a first- or second-degree relative with a neural tube defect, maternal serum AFP at 16 weeks and a detailed ultrasound scan at 18 weeks are recommended. Folic acid taken before and during pregnancy has been shown to reduce the risk of having a child with a neural tube defect by 75% (Nussbaum *et al*., 2001) Since 95% of babies with neural tube defects are born to individuals with no family history of NTDs, it is recommended that all women of childbearing age take folic acid supplementation of at least 400 μg daily and higher for women who have a positive family history of NTDs. Maternal serum AFP is elevated by abnormalities other than NTDs. Abdominal wall defects and kidney problems such as congenital nephrosis can cause elevated levels of maternal serum AFP. The most commonly seen abdominal wall defects are omphalocele and gastroschisis. An ultrasound

scan can usually distinguish between these two defects. Placental problems are a common cause of elevated maternal serum AFP. When placental problems allow fetal blood to get into the maternal circulation, the levels of AFP can be greatly elevated. Fetal death and twins may also elevate serum AFP. Elevated maternal serum AFP in the absence of fetal abnormalities is associated with an increased risk for preeclampsia and premature delivery.

Maternal serum HCG and uE3 are more predictive of an increased risk for Down syndrome and trisomy 18. Both HCG and uE3 are hormones produced by the placenta. HCG is the strongest predictor of Down syndrome. An elevation in the maternal serum level of HCG will almost always increase the risk of Down syndrome in the fetus over the normal maternal age risk. Fetuses with Down syndrome are also typically associated with a low AFP and low uE3. When the values of all three components of the screen are low, there is an increased risk for trisomy 18. The maternal serum triple screen will detect at least 60% of the fetuses with Down syndrome and 50% of the fetuses with trisomy 18 with a 5% false-positive rate (Wapner *et al*., 2003). Very low levels of uE3 may be associated with steroid sulfatase deficiency (X-linked ichthyosis) or the Smith–Lemli–Opitz syndrome (Nussbaum *et al*., 2001).

Inhibin A is now a part of the maternal serum screen in many laboratories. Inhibin A is a hormone produced by the granulosa cells of the female follicles and also by the placenta. It inhibits the synthesis and secretion of FSH and LH by the pituitary. Inhibin A may be elevated in the maternal serum if the fetus has Down syndrome. The addition of Inhibin A to the triple screen, thus making it the quad screen, increases the detection rate of Down syndrome to 80% (Malone *et al*., 2003).

2.2. *First trimester*

Screening for chromosome abnormalities and other birth defects may be done at 11–13 weeks of pregnancy, with a combination of ultrasound and biochemical markers in the maternal serum. An ultrasound scan is performed to measure the nuchal translucency (NT), which is the area of translucency between the skin and underlying tissue at the base of the neck of the fetus. An increase in the NT is caused by an accumulation of fluid under the skin. Fetuses with an increased NT are more likely to have genetic problems or other birth defects such as cardiac abnormalities. Measuring the NT requires special training and experience to obtain consistent results, and quality control is a major issue with this technology. The first-trimester screen also includes the determination of the level of free beta HCG and pregnancy-associated placental protein A (PAPP-A) in the maternal serum. High amounts of free beta HCG and low levels of PAPP-A are associated with an increased risk for the fetus to have Down syndrome. Low levels of both free beta HCG and PAPP-A are associated with an increased risk of trisomy 18 in the fetus. Using a mathematical formula taking into consideration the NT measurement, maternal serum levels of free beta HCG and PAPP-A, gestational age, and maternal age, the risk for Down syndrome, and trisomy 18 can be calculated. Multicenter studies in the United States have indicated that at least 80% of the fetuses with

Down syndrome can be identified using this technology. The false-positive rate is about 5%. The screening identified about 90% of the cases of trisomy 18 with a false-positive rate of 2% (Malone *et al*., 2003; Malone and D'Alton, 2003). First-trimester evaluation of nasal bones does not add significantly to the detection of Down syndrome (Malone and D'Alton, 2003). One limitation of the first-trimester screen is the lack of testing for neural tube defects. It is recommended that patients who have first-trimester screening have a maternal serum alpha fetoprotein and targeted ultrasound in the second trimester for the detection of neural tube defects.

2.3. *Integrated screen*

The integrated screening that combines both first-trimester nuchal translucency plus serum screening and second-trimester quad screening appears to be the most efficient testing method with almost a 90% detection rate for Down syndrome with a false-positive rate of 5% (Malone *et al*., 2003). A major drawback to this approach is that the results would not be available to the patient until about the 16th week of pregnancy.

2.4. *Ultrasound*

Ultrasound has become a very important noninvasive tool for prenatal diagnosis of certain fetal abnormalities. The high resolution of today's ultrasound machines allows detection of many types of birth defects. We have just discussed its use in screening for fetal aneuploidy in the first trimester of pregnancy and the detection of fetal abnormalities such as neural tube defects and abdominal wall defects. There are a number of single-gene disorders that are associated with abnormalities that can be detected by ultrasound including Holt–Oram syndrome, Meckel–Gruber syndrome, Beckwith–Weideman syndrome, osteogenesis imperfecta, and many types of dwarfism. Multifactorial disorders such as cleft lip and heart defects may also be detected. Certain findings on ultrasound such as hydrops, cystic hygroma, polydactyly, diaphragmatic hernia, ventriculomegaly, duodenal atresia, and renal abnormalities may warrant further testing such as amniocentesis since those abnormalities may be associated with a chromosome abnormality.

3. Invasive testing

3.1. *Chorionic villus sampling*

For diagnostic testing, we must obtain cells representative of the fetus using the technique of chorionic villus sampling (CVS) or amniocentesis. The most common reasons for diagnostic testing are maternal age of 35 or above and abnormal first- or second-trimester screening tests. The procedure used most commonly in the first trimester is chorionic villus sampling. This procedure may be done either trans-abdominally or transcervically. The position of the placenta can determine which

approach is used. The transabdominal approach involves the insertion of a 22-gauge spinal needle through the abdomen into the placenta and aspiration of a sample of tertiary villi. For the transcervical approach, a flexible cannula is introduced through the cervix to the placenta where the villi are aspirated. The tissue sampled in CVS is tertiary villi. The villi are derived from the extraembryonic cells of the blastocyst called the trophoblast. Since these cells are from extraembryonic tissue, there may be instances where abnormalities present in these cultured cells may not represent the fetus. About 1–2% of the patients who have CVS are recommended to have a follow-up amniocentesis because of some ambiguous result (usually mosaicism). There is also a greater risk for maternal cell contamination in CVS since the maternal decidua is so closely associated with the chorionic villi. Laboratory personnel must be very diligent in cleaning the sample of any contaminating decidua. The major advantage of CVS is that it can be performed in the 10th to 12th week of pregnancy, thus giving a result earlier than the second-trimester testing. Neural tube defects cannot be detected with CVS, so second-trimester screening with maternal serum AFP and ultrasound are recommended. Fetal loss rate after CVS is about 1%. Amniocentesis may be done during the first trimester, but usually not until the 12th week. Early amniocentesis has a miscarriage risk of about 2% and has been associated with an increased risk for club foot in the fetus (Canadian Early and Mid-Trimester Amniocentesis Trial (CEMAT) Group, 1998).

3.2. Amniocentesis

The preferred method of obtaining fetal cells for prenatal diagnosis in the second trimester is the amniocentesis. It most commonly is done between 16 and 18 weeks. An ultrasound scan is performed before the amniocentesis to verify the number of fetuses and viability, look for structural abnormalities in the uterus, determine the position of the placenta and the best position to introduce the needle for removal of fluid, and examine the fetus for any abnormalities. The procedure involves the removal of amniotic fluid from the amniotic sac with a 22-gauge spinal needle. It is done transabdominally under ultrasound guidance. In experienced hands, the fetal loss rate after amniocentesis is 1in 300–400. Fortunately, most women do not have any problems following the procedure. The most common complaint is mild cramping, which usually subsides within a few hours after the tap. Bleeding or leakage of amniotic fluid vaginally may occur in about 0.2% of the cases (Brumfield *et al*., 1996).

The amniotic fluid contains cells from the fetus that can be cultured and used for cytogenetic, biochemical, or molecular testing. The amniotic fluid itself may be used to test for certain abnormalities. At least 99% of the neural tube defects can be detected by assaying the amniotic fluid for AFP (Milunsky, 1986). As discussed previously, the amount of AFP is elevated in the amniotic fluid when there is an opening such as a neural tube defect or abdominal wall defect. Other causes of an elevated amniotic fluid AFP include fetal death, congenital nephrosis, and the presence of fetal blood.

Cytogenetic, biochemical, and molecular testing usually take from 1 to 3 weeks to complete. Using the fluorescence in situ hybridization (FISH) technique, there

is now a quick screen for trisomy 13, 18, 21 and detection of the number and type of sex chromosomes. By using DNA probes for these chromosomes, abnormalities in number for these chromosomes can be detected within 24 h after obtaining the amniotic fluid sample and is 98–99% accurate in detecting those abnormalities. This is particularly useful if the pregnancy is advanced and a quick result is needed. This study does not detect any structural abnormalities or changes in chromosome number involving any of the other chromosomes. A full cytogenetic study is recommended, and no decision about pregnancy termination should be based on the FISH result alone.

3.3. Periumbilical blood sampling (PUBS)

Occasionally, it is necessary to obtain a fetal blood sample for diagnosis. To do this, a procedure similar to an amniocentesis is performed and a small amount of blood is taken from the umbilical cord. Since the fetal loss rate after PUBS is 2–3%, it is used only when CVS or amniocentesis cannot provide the information needed (Nussbaum *et al*., 2001). Some situations that might warrant PUBS include decreased amount of amniotic fluid, ambiguous results from an amniocentesis or CVS procedure that may be clarified by studying fetal blood, or diagnosis of an abnormality that cannot be done from any other tissue.

3.4. Preimplantation genetics

Recent advances in genetics and reproductive technologies have made it possible to determine for some abnormalities the genetic makeup of an embryo before introduction into the uterus. This allows the implantation of only the unaffected embryos. Preimplantation genetic diagnosis (PGD) starts with routine in vitro fertilization. At the 6–8-cell blastocyst stage of the embryo, one or two cells are removed for genetic testing. These cells are then analyzed using polymerase chain reaction (PCR) or FISH to detect the abnormality of interest. Using FISH, PGD can detect chromosome abnormalities such as aneuploidies, deletions, translocations, and other rearrangements. Candidates for PGD using FISH include:

1. Women who are over 35 and at increased risk to have babies with aneuploidy.
2. Couples where one carries a rearrangement of chromosomes or deletion such as the DiGeorge syndrome deletion on chromosome 22.
3. Women who carry an abnormal X-linked gene where there is no molecular testing available (for sex determination).

Couples who are carriers of a detectable recessive, dominant, or X-linked disorder may elect to have PGD using the PCR technology. This, of course, requires a definitive diagnosis and known mutations in the parents. Obtaining PCR results from one cell is a daunting task and requires a great deal of experience and care in the laboratory.

There are many advantages of PGD. Having the diagnosis made before implantation takes away the necessity of making a decision about termination of a pregnancy

when an abnormality is detected during the first or second trimester of pregnancy. Individuals who are known carriers of a genetic disease will sometimes decide not to have children to avoid passing the abnormal gene to them. PGD affords these people the opportunity to have children free of the disease. As PGD becomes a more widely used technology, the incidence of these serious detectable diseases potentially will decline.

The disadvantages in PGD lie mainly in the technology of the procedures required to accomplish it. Normally fertile women must undergo IVF procedures in order to get pregnant. The technical difficulty of embryo biopsy, PCR, and FISH on a single cell may cause problems or failure of the process. Not all chromosome or single gene disorders can be detected by PGD because only a certain number of chromosomes or genes may be examined during one process. Many ethical concerns have been raised about the possible use of the technology and the fate of the unused embryos.

In the future, it is possible that comparative genomic hybridization (CGH) may be used to analyze all the chromosomes in the complement rather than a selected few. Also, the technique of interphase conversion where an interphase nucleus is converted to metaphase by fusing it with an enucleated oocyte may allow the analysis of all chromosomes.

3.5. Fetal cells in maternal serum

It was demonstrated over 35 years ago that fetal cells exist in maternal serum (Walknowska *et al*., 1969). Since that time, investigators have been trying to develop technologies to concentrate enough fetal cells from the maternal serum for prenatal diagnostic testing. Limited success has been reported. At this point, no consistent, effective protocols are available for prenatal diagnosis using these cells. Research in this area is continuing.

References

Brumfield CG, Lin S, Conner W, Cosper P, Davis RO and Owen J (1996) Pregnancy outcome following genetic amniocentesis at 11-14 versus 16-19 weeks' gestation. *Obstetrics and Gynecology*, **88**(1), 114–118.

Canadian Early and Mid-Trimester Amniocentesis Trial (CEMAT) Group (1998) Randomized trial to assess safety and fetal outcome of early and midtrimester amniocentesis. *Lancet*, **351**, 242–247.

Malone FD and D'Alton ME (2003) First-trimester sonographic screening for Down syndrome. *Obstetrics and Gynecology*, **102**(5), 1066–1079.

Malone FD, Wald NJ, Canick JA, Ball RH, Nyberg DA, Comstock CH, Bukowski R, Berkowitz RL, Gross SJ, Dugoff L, *et al*. (2003) First and second trimester evaluation of risk (FASTER) trial: principal results of the NICHD multicenter Down syndrome screening study. *American Journal of Obstetrics and Gynecology*, **189**, S56, [SMFM abstracts].

Milunsky A (1986) Neural tube and other congenital defects. *Genetic Disorders and the Fetus: Diagnosis, Prevention, and Treatment*, *Second Edition*, Plenum Publishing Company: New York, p. 465.

Nussbaum RL, McInnes RR and Willard HE (2001) *Thompson and Thompson Genetics in Medicine*, *Sixth Edition*, WB Saunders Company: Philadelphia, pp. 364–365.

Walknowska J, Conte FA and Grumbach MM (1969) Practical and theoretical implications of fetal/maternal lymphocyte transfer. *Lancet*, **I**, 1119–1122.

Wapner R, Thom E, Simpson JL, Pergament E, Silver R, Filkins K, Plat L, Mahoney M, Hohnson A, Hogge WA, *et al*. (2003) First trimester maternal serum biochemistry and fetal nuchal translucency screening (BUN) study group. *The New England Journal of Medicine*, **349**(15), 1405–1413.

Specialist Review

Advances in newborn screening for biochemical genetic disorders

Inderneel Sahai and Harvey L. Levy
Harvard Medical School, Boston, MA, USA

1. Introduction

Newborn screening (NBS) has evolved significantly since its introduction in 1962 for one biochemical genetic disorder, phenylketonuria (PKU). It now includes endocrinopathies, hemoglobinopathies, congenital infections, cystic fibrosis, and congenital deafness as well as many biochemical genetic disorders. The wide coverage of biochemical genetic disorders has been possible because of the recent addition of tandem mass spectrometry (MS/MS) to the technology of newborn screening. This article briefly reviews newborn screening and then focuses on the advances in neonatal identification of the biochemical genetic disorders.

2. Background

Soon after Robert Guthrie developed newborn screening as a means of preventing mental retardation in PKU, he realized that several other biochemical genetic disorders fit the same mold, that is, the disorders are capable of producing mental retardation but are potentially preventable by newborn identification and dietary treatment. Among these disorders were homocystinuria, maple syrup urine disease (MSUD), and galactosemia - disorders that can also result in somatic complications. Consequently, he modified his bacterial assay for phenylalanine, the analytical marker for PKU, to render it responsive to increases in other metabolites and to allow for the detection of these disorders (Guthrie, 1968).

The keys to these additions and further developments in newborn screening are the dried blood filter paper specimen, now known as the Guthrie specimen, and the small amount of blood needed for the assays. Consequently, tests can be added to newborn screening without the requirement for another specimen, provided the test can be performed on whole blood eluted from a small disk punched from the specimen. Over the years, many investigators have become aware of the possibility that this system could be used to screen newborns for disorders in their areas of interest, and several have succeeded in modifying existing assays to meet the specifications of newborn screening. The results have been newborn screening

for congenital hypothyroidism (Dussault *et al*., 1975), sickle cell disease (Garrick *et al*., 1973), and congenital adrenal hyperplasia (Pang *et al*., 1977). However, until very recently, the only biochemical genetic disorder among these additions has been biotinidase deficiency (Heard *et al*., 1984).

This omission was not by choice. Guthrie and his group tried in vain to expand biochemical genetic coverage. Several programs adopted neonatal urine screening using paper or thin-layer chromatography to expand coverage (Bamforth *et al*., 1999). This resulted in additional coverage but required collection of the specimen at several weeks of age rather than within the first day of life and examined only amino acids and a very limited number of organic acids. Consequently, many serious disorders were not covered, or not covered in time for effective therapy, and the disorders that were most frequently identified were benign (Wilcken *et al*., 1980). Thus, coverage of many important biochemical genetic disorders by screening the newborn blood specimen had to await MS/MS technology.

3. Overview of tandem mass spectrometry (MS/MS)

The mass spectrometer is an analytical instrument that detects ions by virtue of their mass and charge. It has been used for many years to identify metabolites that are separated on columns in techniques such as capillary electrophoresis, gas chromatography, and HPLC. This requirement for prior separation of metabolites, however, meant that mass spectrometry could not be used in screening since separation would be much too time consuming for the rapid throughput that routine screening requires. However, MS/MS is applicable to newborn screening because it directly analyzes the individual components of complex mixtures, eliminating the need for prior chemical separation. Developed in the late 1970s (McLafferty, 1981), MS/MS employs two mass spectrometers in tandem separated by a collision cell. It takes advantage of the principle that molecules fragment into predictable species depending on their structural properties. Millington *et al*. (1990) adapted this technology to identify metabolites in the Guthrie dried blood spots (DBS) of newborn infants. Naylor subsequently used this system to greatly expand the number of biochemical genetic disorders identifiable in routine newborn screening (Naylor and Chace, 1999).

In its current application to newborn screening, MS/MS measures amino acids and acylcarnitine conjugates of organic and fatty acids in the Guthrie blood specimen. Figure 1 depicts the elements of MS/MS and the chemical characteristics of derivatization and fragmentation of the amino acids and the acylcarnitines. For a detailed description of the MS/MS methods and analysis, the reader is referred to articles by Chace *et al*. (1999, 2002, 2003). The process of newborn screening begins with methanol extraction of blood from the specimen, followed by butanol derivatization of amino acids and acylcarnitines. The butyl ester derivatives are ionized by electrospray. The ions enter the first MS, where the parent molecules are sorted by mass and charge and are then directed into the collision chamber. In this chamber, collision-induced disassociation (CID) produces characteristic fragments because of collision between an inert gas and the ionized particles. These fragments are then guided into the second MS where they are sorted by mass and charge (m/z)

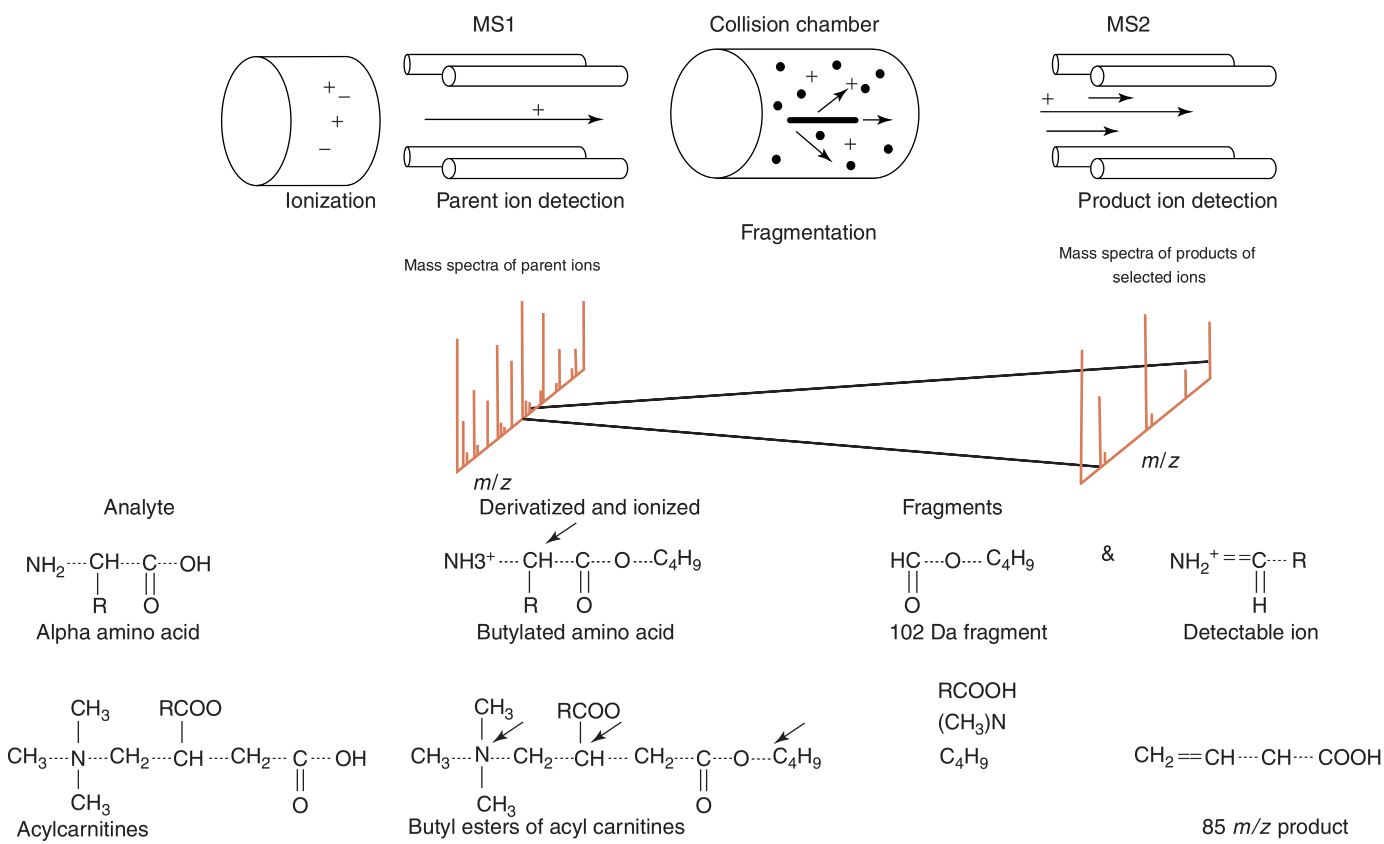


Figure 1 Schematic diagram of a tandem mass spectrometer and the path of the molecules and ions. Depicted below the diagram is the spectrum obtained from MS1 and MS2, which is later correlated by the analyzer. Also shown are amino acid and acylcarnitine molecules and their fragments

and detected as signals. Computer algorithms enable correlation between the mass spectra of the fragments in the second MS and the spectra of the intact parent molecules in the first MS. The data thus acquired can be scanned for characteristic fragments so that parent molecules corresponding with these fragments can be selectively analyzed. In general, amino acids are identified by their loss of a 102-Da fragment, while acylcarnitines are always identified by the presence of their common m/z 85 fragment ion. Different scanning modes can analyze the same data for other specific fragments, thereby resulting in identification of multiple classes of compounds in the same specimen. Quantitative analysis is performed using the ratio of the metabolite to its isotopic analog (internal standard is added at the beginning of the process). Software programs are formatted to analyze the spectra, perform the calculations to determine the concentrations, and flag the abnormal values. Though the automated system detects metabolite abnormalities, personnel familiar with metabolic disorders are essential for recognizing the metabolite profiles that indicate a disorder.

MS/MS analysis requires only about 2 minutes per specimen, making it ideal for high-volume programs such as newborn screening. Nevertheless, the system has certain technical limitations: (1) The equipment needs proper maintenance and tuning, making trained technicians vital for its operation. (2) The volume of blood used in calculating the metabolite values is the volume assumed to have been extracted from the standard-sized disk of the Guthrie specimen used in the analysis, but the actual volume can vary in different specimens, depending on the extraction efficiencies. This probably accounts for most of the variation seen with MS/MS analysis. (3) The derivatization can produce hydrolysis of acylcarnitines, falsely reducing the estimate of its endogenous quantity (and falsely increasing the estimated level of endogenous carnitine).

4. Expanded newborn screening

This is the term generally applied to the increased number of biochemical genetic disorders identifiable in newborn screening that is made possible by MS/MS technology. They are broadly classified into amino acid, organic acid, and fatty acid oxidation disorders.

4.1. Prevalence

The reported frequency of identified cases in expanded newborn screening has varied from 1:2400 among 250,000 screened infants in Baden-Württemberg, Germany (Schulze *et al*., 2003) to 1:4000 among 1.1 million infants who were screened primarily in Pennsylvania (Chace *et al*., 2002) and 1:5000 among 164,000 of those who were screened in Massachusetts (Zytkovicz *et al*., 2001). In New South Wales, Australia, a frequency of 1:6400 cases among 362 000 screened infants has been reported, but it excludes PKU (Wilcken *et al*., 2003), one of the most frequent of the detected disorders.

4.2. Amino acid disorders

This group of disorders is marked by the accumulation of characteristic amino acids; each disorder is the result of an inherited defect in an enzyme that is responsible for conversion of an amino acid into its product. Disorders traditionally screened, such as, PKU, MSUD, and homocystinuria, plus disorders that were not previously screened, such as, tyrosinemia, hyperglycinemia, hydroxyprolinemia, and several urea cycle defects, are identifiable by MS/MS. The incidence of the amino acid disorders detected by MS/MS screening has been reported as 1:3800 in Germany (Schulze *et al*., 2003), 1:7400 in Pennsylvania (Chace *et al*., 2003), and approximately 1:8000 in Massachusetts (Zytkovicz *et al*., 2001).

The clinical phenotype of the amino acid disorders is usually insidious in development. In PKU, for instance, the neonate is clinically normal and remains so until developmental delay, irritability, and autistic features appear during the latter half of the first year. Homocystinuria may present with a similar picture, subsequently producing somatic features that include ectopia lentis, skeletal changes, and thromboembolism. Tyrosinemia type I can produce liver disease in early infancy, whereas tyrosinemia type II can produce developmental delay without liver involvement. MSUD, however, can cause profound neonatal disease characterized by metabolic acidosis. This phenotype is characteristic of the organic acid disorders (in which category MSUD is often included). Other exceptions are the urea cycle disorders, in which the affected neonate may have severe hyperammonemia, producing lethargy that can rapidly progress to coma. Nevertheless, in all of the amino acid disorders, clinical signs of disease may not have appeared during the first few days of life. Expanded newborn screening, revealing a specific amino acid elevation, can identify these infants and preventive therapy can be effective.

The major confirmatory test following any newborn screening abnormality is a plasma amino acid analysis. Table 1 lists the amino acid disorders that can be identified in expanded screening according to the screening abnormalities that bring them to attention and highlights their biochemical and clinical phenotypes.

4.3. Fatty acid oxidation disorders (FAOD)

Fatty acid oxidation primarily occurs in the mitochondria and is critical for energy generation in the fasting state or during exercise. The sequence of fatty acid oxidation from transport into mitochondria through β-oxidation is depicted in Figure 2. It begins with the mobilization of very long chain and long chain fatty acids from adipose tissue. In the cytosol of cells, they link to coenzyme A (CoA) under the influence of fatty acyl-CoA synthetase to yield fatty acyl-CoA. The fatty acyl-CoA moves to the mitochondrion, as does carnitine, which has entered the cell via a carnitine transporter. On the outer mitochondrial membrane, carnitine palmitoyltransferase I (CPT I) catalyzes an exchange of carnitine with the CoA to yield fatty acylcarnitine, releasing CoA. The fatty acylcarnitine is transported to the inner mitochondrial membrane by a translocase where CPT II decouples and

Table 1 Amino acid disorders

Major screening analyte(s)	Disorder	Enzyme defect	Clinical phenotype	General treatment
Phenylalanine	Phenylketonuria (PKU)	Phenylalanine hydroxylase	Mental retardation	Dietary restriction of phenylalanine
			Autism Hyperactivity Seizures	
Leucine/isoleucine/ hydroxyproline	Maple syrup urine disease (MSUD)	Branched-chain a-ketoacid dehydrogenase complex	Lethargy	Dietary restriction of branched-chain amino acids
			Failure to thrive Coma Seizures Maple syrup odor in urine and cerumen	
Leucine/isoleucine/ hydroxyproline	Hydroxyprolinemia	Hydroxyproline oxidase	Asymptomatic	None
Methionine	Homocystinuria	Cystathionine β-synthase	Mental retardation	Dietary restriction of methionine
			Arachnodactyly	Supplemental B6 (for B6 responsive)
			Osteoporosis Ectopia lentis Thromboembolism	
Methionine	Hypermethioninemia (MAT I/III deficiency)	Methionine adenosyl transferase (MAT I/III)	Asymptomatic	None known
			(Possibly rare cognitive reduction)	

Tyrosine	Tyrosinemia type I	Fumarylacetoacetate hydrolyase	Hepatic disease	2-(2-nitro-4-trifluoromethylbenzoyl)-1,3-cyclohexanedione (NTBC); dietary restriction of phenylalanine and tyrosine
			Hypoglycemia Hypophosphatemic rickets	Liver transplant
Tyrosine	Tyrosinemia type II	Tyrosine aminotransferase	Keratoconjunctivitis	Dietary restriction of phenylalanine and tyrosine
			Palmer/plantar keratosis Cognitive reduction	
Citrulline	Citrullinemia	Argininosuccinic acid synthetase	Hyperammonemia	Dietary restriction of protein
	(Urea cycle defect)		Mental retardation Failure to thrive Lethargy, coma	Supplements: L-arginine Sodium benzoate Sodium phenylacetate
Citrulline	Argininosuccinic acidemia (Urea cycle defect)	Argininosuccinic acid lyase	Hyperammonemia Failure to thrive Lethargy, coma	Dietary restriction of protein Supplements: L-arginine Sodium benzoate Sodium phenylacetate
Arginine	Arginase deficiency	Arginase	Mild-moderate hyperammonemia	Dietary restriction of arginine and protein
	(Urea cycle defect)		Mental retardation Spastic diplegia	
Glycine	Nonketotic hyperglycinemia (NKH)	Glycine cleavage enzyme	Seizures	Supportive care
			Hypotonia Marked lethargy	Sodium benzoate

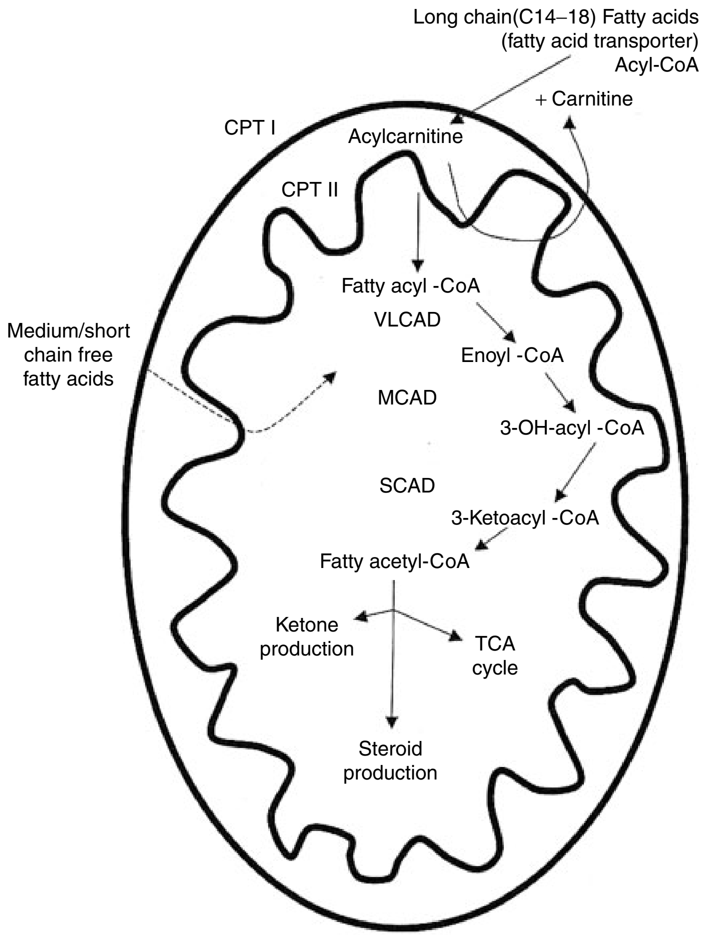


Figure 2 Metabolism of fatty acids illustrating entrance into mitochondrion and intramitochondrial β-oxidation. Abbreviations: CPT, Carnitine palmitoyltransferase; VLCAD, very long chain acyl-CoA dehydrogenase; MCAD, medium chain acyl-CoA dehydrogenase; SCAD, short chain acyl-CoA dehydrogenase; TCA, tricarboxylic acid

releases the carnitine and links the fatty acid to CoA again. The resulting fatty acyl-CoA moves into the mitochondrial matrix where it is β-oxidized sequentially by six enzymes, including very long chain acyl-CoA dehydrogenase (VLCAD), long chain hydroxyacyl-CoA dehydrogenase (LCHAD), medium chain hydroxyacyl-CoA dehydrogenase (MCHAD), medium chain acyl-CoA dehydrogenase (MCAD), short chain hydroxyacyl-CoA dehydrogenase (SCHAD), and short chain acyl-CoA dehydrogenase (SCAD). The successive removal of two carbons via the β-oxidation cycle results in acetyl-CoA that is used to generate energy through the TCA cycle and the synthesis of ketones. Acetyl-CoA is also required for the synthesis of steroids.

Disorders have been described for each of these transporters and enzymes except fatty acyl-CoA synthetase. The reported frequency of the FAODs in expanded newborn screening is 1:10 000 in Germany (Schulze *et al*., 2003) and Massachusetts (Zytkovicz *et al*., 2001), and 1:13 000 in Australia (Wilcken *et al*., 2003). The most frequently detected in expanded newborn screening has been medium chain acyl-CoA dehydrogenase deficiency (MCADD), with an incidence of 1:20 000, which is similar to that for PKU.

The most striking clinical feature in the FAOD phenotype is intolerance to fasting or to vigorous exercise. Normally, reduction of glucose in the fasting state or the requirement for additional glucose during exercise triggers the compensatory production of ketones to supply energy. In the FAOD, the inability to meet this requirement for energy through the β-oxidation of fatty acids results in muscle weakness; lethargy; fatty liver; inhibition of gluconeogenesis, exacerbating the hypoglycemia; and a high risk of sudden death. Cardiomyopathy is a particular feature of carnitine transporter defect and of defects in the β-oxidation of very long chain and long chain fatty acids. A key laboratory feature is hypoketotic hypoglycemia. Other laboratory findings during acute episodes include metabolic acidosis, hyperammonemia, and elevations of the liver enzyme levels. Early detection by expanded newborn screening combined with treatment that includes avoidance of fasting, adherence to a low-fat diet, and carnitine supplementation has been effective in preventing sudden death and other complications in these disorders (Waisbren *et al*., 2003).

In expanded newborn screening by MS/MS, the carnitine and acylcarnitine patterns reveal the presence of an FAOD and, usually, the specific defect. Since carnitine transport defect limits both the transport of free carnitine from the intestinal lumen and its reabsorption in the kidneys, the blood carnitine concentration is very low. In CPT-I deficiency, acylcarnitines are not formed, so the concentration of free carnitine is high but that of the acylcarnitines are low. In CPT-II deficiency, the acylcarnitines are unchanged, so the long chain acylcarnitine levels are increased. For the β-oxidation defects, the levels of long chain acylcarnitines are increased in very long chain acyl-CoA dehydrogenase deficiency (VLCADD), long chain hydroxyacylcarnitines levels are increased in long chain hydroxyacyl-CoA dehydrogenase deficiency (LCHADD), medium chain acylcarnitines, particularly octanoylcarnitine (C8), are increased in MCADD, and the short chain acylcarnitines, primarily butyrylcarnitine (C4), are increased in short chain acyl-CoA dehydrogenase deficiency (SCADD). After an acylcarnitine abnormality has been ascertained, usually by repeat screening, confirmatory testing is performed by analyses of plasma acylcarnitines and urine acylglycines. The individual fatty acid oxidation disorders are summarized in Table 2.

In contrast to amino acid concentrations, which increase in a time-dependent manner, the acylcarnitines decrease significantly with passage of time and with initiation of feeding. This occurs within the first week of life. The rate of decrease of the long chain acylcarnitines is more rapid than that of the short chain acylcarnitines (Chace *et al*., 2003). Effective screening, therefore, requires collection of the Guthrie specimen within 24–48 h of life and lowering of the cutoff value for repeat specimens. The concentrations of multiple acylcarnitine markers and relative molar ratios are also important in the interpretation of a positive result.

Table 2 Fatty acid oxidation disorders

Major screening analyte(s)	Disorder	Enzyme defect	Clinical phenotype	General treatment
C4 (butyrylcarnitine)	Short chain acyl-CoA dehydrogenase deficiency (SCADD)	Short chain acyl-CoA dehydrogenase (SCAD)	Variable presentation: may be asymptomatic	Avoid fasting
				Diet: low fat, high carbohydrate (CHO)
C8 (octanoylcarnitine)	Medium chain acyl-CoA dehydrogenase deficiency (MCADD)	Medium chain acyl-CoA dehydrogenase (MCAD)	Asymptomatic when well	Avoid fasting
			With fasting and/or illness: vomiting, lethargy, seizures, coma, hypoketotic hypoglycemia, metabolic acidosis, hepatomegaly	Frequent CHO feedings (q4 hours until 4 months; q6 hours until 8 months; q8 hours thereafter)
			Can cause sudden death	Possible supplementation with carnitine
C8 (octanoylcarnitine)	Multiple acyl-CoA dehydrogenase deficiency (MADD)	Multiple acyl-CoA dehydrogenation defects due to decreased electron transfer flavoprotein (ETF), or decreased electron transfer flavoprotein ubiquinone oxidoreductase (ETF-Q0)	Hypoketotic hypoglycemia	Avoid fasting
	[Also known as glutaric acidemia II (GA II)]		Metabolic acidosis	Dietary restriction of fat
C8:1 (octenoylcarnitine)			"Sweaty feet" odor	Frequent CHO feedings
			Muscle weakness	Supplement with riboflavin and carnitine
			Neonatal form has occasional dysmorphic features and hypotonia	
C14 (tetradecanoylcarnitine)	Very long chain acyl-CoA dehydrogenase deficiency (VLCADD)	Very long chain acyl-CoA dehydrogenase (VLCAD)	Variable presentation:	Avoid fasting
			Cardiomyopathy (hypertrophic)	Low fat, high CHO diet

C14:1 (tetradecenoylcarnitine)			Sudden death	Frequent CHO feedings, medium chain triglycerides (MCT)
			With illness: vomiting, lethargy, seizures, coma	Carnitine
			Hypoketotic hypoglycemia	
C16OH (hydroxydecanoylcarnitine)	Long chain hydroxyacyl-CoA dehydrogenase deficiency (LCHADD)	Long chain hydroxyacyl-CoA dehydrogenase (LCHAD)	Hypoketotic hypoglycemia	Avoid fasting
			Mild hyperammonemia	Low fat, high CHO diet
C18:1OH (hydroxyoctadecenoylcarnitine)			Cardiomyopathy	Supplement with carnitine
			Sudden death	
C18OH (hydroxyloctadecanoylcarnitine)			Associated maternal syndrome	Early severe hypotonia
			Renal tubular acidosis	Hemolysis, Elevated liver enzymes and Low Platlets (HELLP)
CN (carnitine)	Carnitine palmitoyltransferase I (CPT I) deficiency	Carnitine palmitoyltransferase I (CPT I)	Infantile form:	Dietary restriction of high fat foods
			Fasting hypoketotic hypoglycemia	Supplement: Frequent CHO feedings, MCT supplements, carnitine
			Lethargy	
			Hepatomegaly	
			Seizures	
C16 (palmitoylcarnitine)	Carnitine palmitoyltransferase II (CPT II) deficiency	Carnitine palmitoyl-transferase II (CPT II)	Infantile presentation:	Dietary restriction of high fat foods
	or	or	Hypoketotic hypoglycemia	Supplement: Frequent CHO feedings, carnitine
C18:1 (octadecenoylcarnitine)	Carnitine/acylcarnitine translocase deficiency	Carnitine/acylcarnitine translocase	Lethargy	
			Seizures	
			Cardiomyopathy	

Table 3 Organic acid disorders

Major screening analyte	Disorder	Enzyme defect	Clinical phenotype	General treatment
C3 (propionylcarnitine)	Propionic acidemia	Propionyl-CoA carboxylase	Metabolic acidosis	Dietary restriction of threonine, isoleucine, valine, methionine
			Hyperammonemia Vomiting Failure to thrive	
C3 (propionylcarnitine)	Methylmalonic acidemia	Methylmalonyl-CoA mutase	Metabolic acidosis	Dietary restriction of isoleucine, valine, methionine (supplemental B_{12} for B_{12}-responsive form)
			Hyperammonemia Vomiting Failure to thrive	
C5 (isovalerylcarnitine)	Isovaleric acidemia	Isovaleryl-CoA dehydrogenase	Metabolic acidosis "Sweaty feet" odor	Dietary restriction of leucine Supplemental glycine and L-carnitine
			Vomiting Lethargy Failure to thrive	
C5 (2-methylbutyrylcarnitine)	2-methylbutyryl-CoA dehydrogenase deficiency	2-methylbutyryl-CoA dehydrogenase	Possibly Asymptomatic	Dietary restriction of protein
		(short/branched-chain acyl-CoA dehydrogenase)	Possibly Failure to thrive, hypoglycemia, metabolic acidosis	Supplementation with L-carnitine

C5DC (glutarylcarnitine)	Glutaric Acidemia I	Glutaryl-CoA dehydrogenase	Dystonia	Dietary restriction of lysine and tryptophan
			Acidosis and vomiting	Supplemental L-carnitine, riboflavin
			Macrocephaly	
C5OH (3-hydroxyisovalerylcarnitine)	3-Hydroxy-3-methylglutaryl (HMG)-CoA lyase deficiency	3-Hydroxy-3-methylglutaryl-CoA lyase	Vomiting	Intravenous (IV) glucose
			Lethargy Hypotonia Seizures Hypoglycemia Metabolic acidosis	Restrict protein, fat
C5OH (3-hydroxyisovalerylcarnitine)	3-Methylcrotonyl-CoA carboxylase (3-MCC) deficiency	3-Methylcrotonyl-CoA carboxylase	Asymptomatic, or:	IV glucose
			Lethargy Hypotonia Hypoglycemia Metabolic acidosis	Restrict protein

4.4. Organic acid disorders (OAD)

The defect in an OAD is usually in an enzyme within an amino acid degradative pathway that is responsible for conversion of an organic acid derivative of the amino acid to another organic acid. The result of the defect is an increased concentration of the unconverted organic acid, some of which can conjugate with free carnitine to produce an increased acylcarnitine level detectable by MS/MS. The reported frequency of OADs is 1:15 000 in Germany (Schulze *et al*., 2003), 1:30 000 in Australia (Wilcken *et al*., 2003), and 1:55 000 in Massachusetts (Zytkovicz *et al*., 2001). The unconjugated organic acid (and secondary organic acid metabolites) accumulate to very high levels and produce metabolic acidosis. This can present in the neonate as poor feeding, lethargy, vomiting, tachypnea, and seizures. During such acute episodes, substantial hyperammonemia is also often present. Later complications include developmental delay, mental retardation, and other neurologic abnormalities. A notable exception to this dramatic presentation is glutaric acidemia type I (GA I), in which the neonate appears to be normal (except for macrocephaly) but suddenly becomes dystonic later in the first or second year of life, often during an acute febrile illness or after an immunization. Expanded newborn screening by MS/MS has identified these neonates presymptomatically by an acylcarnitine abnormality and has improved their outcome through appropriate dietary therapy (Waisbren *et al*., 2003). Confirmation of an acylcarnitine screening abnormality that suggests an OAD includes urine organic acid analysis. The individual organic acid disorders are summarized in Table 3.

5. Considerations in MS/MS screening

5.1. Amino acid disorders

MSUD: An elevation of levels of leucine and variable increases in levels of valine, isoleucine, and alloisoleucine are the characteristics of MSUD. Leucine, isoleucine, and alloisoleucine have the same molecular weight and, therefore, give the same signal in the MS/MS analysis. Nevertheless, since all these metabolites are elevated in MSUD, an increase in the signal intensity of "leucine", especially when accompanied by increased valine levels, is suggestive of this disorder. The false-positive rate for MSUD can be reduced by considering the ratios of "leucine" to phenylalanine and "leucine" to alanine. Increased levels of hydroxyproline also increases the signal intensity of "leucine" since its molecular weight is the same as leucine and the isoleucines. Consequently, hydroxyprolinemia, which is a benign disorder (Kim *et al*., 1997), can suggest MSUD in MS/MS screening. An important difference, however, is the likely presence of increased valine levels in MSUD and normal valine levels in hydroxyprolinemia.

Homocystinuria: Homocystinuria results in a primary elevation of homocysteine levels, with a secondary increase in methionine. Homocysteine cannot be detected by MS/MS, so methionine is the marker metabolite for this disorder.

As a secondary metabolite, however, the accumulation of methionine lags behind that of homocysteine. Consequently, during the first days of life, the methionine level can remain normal or below the cutoff level for identifying it as being abnormal even if the infant has homocystinuria (Peterschmitt *et al*., 1999). This is especially likely when the Guthrie screening sample is collected at less than 24 hours after birth.

Tyrosinemias: Tyrosinemia I results in only a moderate elevation of tyrosine. Thus, in order to increase the sensitivity of the screening, a low threshold for identifying increased tyrosine must be set. Nevertheless, since many or most affected infants have normal blood tyrosine levels during the first days of life, they are still missed. This was recognized several years ago in Quebec and the newborn screening program was stimulated to apply an enzyme assay screening method for succinylacetone, the specific marker metabolite for tyrosinemia I, to the Guthrie specimen (Grenier *et al*., 1982), resulting in almost complete ascertainment. Unfortunately, succinylacetone is currently not identifiable by MS/MS. Tyrosinemia II may more likely be identified on the basis of increased tyrosine. When increased tyrosine is flagged in newborn screening, the vast majority of infants have transient neonatal tyrosinemia, a benign finding (Wilcken *et al*., 2003).

Urea cycle defects: Three of the five urea cycle disorders (i.e., citrullinemia, argininosuccinic acidemia, and arginase deficiency) can be identified by expanded newborn screening. Citrullinemia and argininosuccinic acidemia are identified on the basis of increased citrulline levels, more pronounced in citrullinemia than in argininosuccinic acidemia. Arginase deficiency is identified by increased arginine levels. The other two urea cycle disorders, carbamyl phosphate synthetase deficiency (CPSD) and ornithine transcarbamylase deficiency (OTCD), could be determined by decreased citrulline levels, but, at present, this identification is limited by an excessively high false-positive rate.

5.2. *Fatty acid oxidation disorders*

MCADD: The levels of the primary analyte of MCADD, octanoylcarnitine (C8), can also be elevated because of valproate therapy or medium chain triglycerides (MCT) oil supplementation and is also an indicator of the fatty acid disorder multiple acyl-CoA dehydrogenase deficiency (MADD). The C8/C10 ratio, which is usually elevated in MCADD, is useful in distinguishing MCADD from these other causes of elevated C8. MCADD is also usually accompanied by elevations in C6, C10, and C10:1.

SCADD: Elevations in butyrylcarnitine (C4) levels, characteristic of this disorder, are also seen in the fatty acid disorder, MADD.

VLCADD: Elevations of C14 (tetradecanoylcarnitine) and C14:1 (tetradecenoylcarnitine) are the primary abnormalities of this disorder, though they can also be present in MADD, CPT-II, and carnitine-acylcarnitine translocase deficiency. VLCADD may also be accompanied by elevations of C16, C18, and C18:1.

LCHADD: Elevations of the primary analytes of LCHADD, C16OH (hydoxy-hexadecanoylcarnitine), C18OH (hydroxyoctanoylcarnitine), and C18:1OH (hydroxyoctadecenoylcarnitine), can be minimal, making LCHADD difficult to diagnose. Elevations of C16, C14, C14:1, and C14OH may also be present.

CPT-I deficiency: This is the only FAOD with elevated levels of free carnitine. Also present in CPT-I deficiency are decreased levels of C16, C18, and C18:1.

CPT-II deficiency and carnitine-acylcarnitine translocase deficiency: The levels of the primary analytes for these disorders, C16 (hexadecanoylcarnitine), C18 (octadecanoylcarnitine), and C18:1(octadecenoylcarnitine), can also be elevated in VLCADD and LCHADD but are more prominent here. Additionally, free carnitine is decreased, and the ratio of C16 to carnitine is a useful diagnostic marker.

Carnitine transporter defect: In this disorder, free carnitine level is decreased to less than 5% of normal levels.

MADD (Glutaric Acidemia II): The acylcarnitines levels are only mildly increased in MADD, making the detection of this disorder difficult. MADD is characterized by increased C8 and C8:1, but variable elevations of C6, C5, C5DC, C4, C14, and C14:1 are also seen.

5.3. *Organic acid disorders*

Propionic acidemia (PA) and methylmalonic acidemia (MMA): C3 (propionyl-carnitine) level is elevated in both these conditions, but the level is typically higher in PA. In fact, in MMA the C3 elevation may be just at or even below the screening cutoff, rendering MMA very difficult to detect. Since cobalamin (B12) is a cofactor for methylmalonyl-CoA mutase, vitamin B12 deficiency can also result in increased methylmalonic acid and C3. The ratio of C3 to C2 increases the sensitivity and specificity for detection.

Isovaleric acidemia (IVA): The indicating metabolite of isovaleric acidemia in newborn screening is C5 (isovalerylcarnitine). 2-methylbutyryl-CoA dehydrogenase deficiency (2-MBDD) also produces an elevation of C5 but is not as striking an elevation as does IVA. However, pivalic acid, a constituent of certain antibiotics, produces a metabolite that is also detected at the m/z of C5. Neonates of mothers who ingested pivalic acid can thus mistakenly be thought to have IVA or 2-MBDD on newborn screening (Abdenur *et al*., 1998).

Glutaric acidemia I (GA I): The primary analyte in this disorder, C5DC (glutarylcarnitine), a dicarboxylic acid, has two negative charges that need to be neutralized by derivatization before analysis in MS/MS. Elimination of the derivatization step can thus decrease the sensitivity for C5DC.

Hydroxymethylglutaryl (HMG) CoA lyase and 3-methylcrotonyl carboxylase (3-MCC) deficiencies:The level of the marker metabolite for these two disorders, C5OH (3-hydroxyisovalerylcarnitine), can also be elevated in 3-methylglutaconic aciduria I. Furthermore, a peak at m/z 318 indicating C5OH can also be produced by methylbutyrylhydroxy acylcarnitine, which is the marker metabolite for methyl acetoacetyl-CoA thiolase (β-ketothiolase) deficiency.

6. Missed cases

Most infants with biochemical genetic disorders are not identified by newborn screening. These infants usually have disorders that are currently not detectable by screening, including diseases in categories such as the lysosomal disorders (e.g., Gaucher disease and Pompe disease), the glycogen storage diseases (e.g., von Gierke disease), the mitochondrial disorders (e.g., Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke like episodes (MELAS), Myoclonic Epilepsy with Ragged Red Fibers (MERRF)), and others. Attempts are being made to include these categories in newborn screening, the most notable of which is the development of feasible screening methods to detect lysosomal disorders in the newborn (Chamoles *et al*., 2001), but, so far, methodologies suitable for routine screening have not emerged.

Much less frequent are infants with metabolic disorders that are identifiable by newborn screening but are undetected. These "missed" infants can be placed into two categories, true negatives and false negatives. In the former classification are infants such as those with tyrosinemia I who did not have increased tyrosine when the Guthrie specimen was collected, those with homocystinuria and a normal methionine level during screening (Wagstaff *et al*., 1991), and those with carnitine transporter defect but a carnitine level in the Guthrie specimen that was above the cutoff level for identification (Wilcken *et al*., 2003). In the false-negative category is the rare infant who had an abnormal level of the metabolite marker for a disorder but was missed because of program or laboratory error (Smith *et al*., 2001).

The important message is that any patient in whom there is clinical suspicion of a biochemical genetic disorder could have a disorder, either because it was not included in the newborn screening schedule or because it was missed in the screening process. Thus, all such patients should have laboratory testing for the disorders. These tests may include plasma amino acid analysis, urine organic acid testing, plasma acylcarnitine profiling, urine acylglycine analysis, tests for lysosomal disorders, and/or additional tests, depending on the clinical and general laboratory findings.

7. Molecular genetic techniques

The application of molecular genetic techniques to screening for a wide spectrum of disorders has been made possible by the stability of DNA in the Guthrie card (McCabe *et al*., 1987). Segments of genes can be amplified directly from the filter paper matrix using the polymerase chain reaction (PCR) without prior extraction. These amplified segments can be examined for mutations relevant to the disorder(s) screened using identification techniques such as restriction analysis or single-nucleotide polymorphism (SNP) methodology. Molecular testing is not currently used in primary newborn screening for multiple reasons, including the many mutations responsible for each of the biochemical genetic disorders, the greatly increased "noise" that would be generated by heterozygote detection, and the prohibitive expense. The plethora of mutations could be met by multiplex PCR coupled with microarrays using chip technology (Caggana *et al*., 1998),

but heterozygote detection and perhaps the expense would continue to be major issues. Nevertheless, candidates for primary molecular screening include those with treatable disorders with few molecular etiologies such as factor V Leiden, sickle cell disease, and the mitochondrial MERRF and MELAS syndromes. Primary screening for severe combined immunodeficiency (SCID), fragile X syndrome, type I diabetes, acute lymphoblastic anemia, and hereditary hemochromatosis might also be considered.

Currently, molecular testing in newborn screening is used as a second tier, through examination of Guthrie cards from infants in whom the primary screening result was abnormal. The purpose of this is to sort the affected infants from those whose primary screening abnormality was due to some reason other than the targeted disorder. The most notable use of molecular testing is in screening for cystic fibrosis, wherein the primary screen, immunoreactive trypsinogen (IRT), can be increased for many reasons other than cystic fibrosis. Second tier testing of the abnormal specimens sorts out those infants who harbor one or more CFTR mutations known to be associated with cystic fibrosis. In expanded newborn screening, second tier molecular testing is also used to test the Guthrie specimen for the A985G mutation in MCAD, the major mutation associated with MCADD, when the primary screen reveals an increased level of C8. In both these examples, second tier molecular testing substantially reduces the number of false-positive results.

8. Postmortem specimens

Inherited disorders of fatty acid oxidation account for an estimated 3–6 % of sudden unexpected deaths in infancy (Boles *et al*., 1998). Among this group, MCADD is the most frequent, and death during the initial presentation has been noted in 25–33% of the cases (Iafolla *et al*., 1994). Sudden death as a result of an inherited metabolic disorder may remain unexplained even after an autopsy and, hence, may get classified as sudden infant death syndrome (SIDS). To examine the frequency of biochemical genetic disorders in such infants and children, MS/MS analysis was performed on postmortem Guthrie specimens obtained from over 7000 cases of sudden unexplained death. In 66 specimens, diagnosis of a biochemical genetic disorder was considered likely (Chace *et al*., 2001). The disorders included MCADD, VLCADD, glutaric acidemia I and II, CPT-II/translocase deficiency, carnitine deficiency, isovaleric acidemia/2-methylbutyryl-CoA dehydrogenase deficiency, and LCHADD. This suggests that protocols should include metabolic screening by MS/MS as part of the postmortem investigation of all cases of sudden death.

Postmortem blood specimens from all infants, even those without a metabolic disorder, are characterized by increased carnitine, acylcarnitines, and amino acids. Consequently, identification of biochemical genetic disorders from postmortem specimens must utilize criteria that are different from those used for the interpretation of metabolic profiles in newborn screening.

References

Abdenur JE, Chamoles NA, Guinle AE, Schenone AB and Fuertes ANJ (1998) Diagnosis of isovaleric acidaemia by tandem mass spectrometry: False-positive result due to pivaloylcarnitine in a newborn screening programme. *Journal of Inherited Metabolic Disease*, **21**, 624–630.

Bamforth FJ, Dorian V, Vallance H and Wishart DS (1999) Diagnosis of inborn errors of metabolism using 1H NMR spectroscopic analysis of urine. *Journal of Inherited Metabolic Disease*, **22**, 297–301.

Boles RG, Buck EA, Blitzer MG, Platt MS, Cowan TM, Martin SK, Yoon H, Madsen JA, Reyes-Mugica M and Rinaldo P (1998) Retrospective biochemical screening of fatty acid oxidation disorders in postmortem livers of 418 cases of sudden death in the first year of life. *Journal of Pediatrics*, **132**, 924–933.

Caggana M, Conroy JM and Pass KA (1998) Rapid, efficient method for multiplex amplification from filter paper. *Human Mutation*, **11**, 404–409.

Chace DH, DiPerna JC and Naylor EW (1999) Laboratory integration and utilization of tandem mass spectrometry in neonatal screening: A model for clinical mass spectrometry in the next millennium. *Acta Paediatrica Supplementum*, **88**, 45–47.

Chace DH, DiPerna JC, Mitchell BL, Sgroi B, Hofman LF and Naylor EW (2001) Electrospray tandem mass spectrometry for analysis of acylcarnitines in dried postmortem blood specimens collected at autopsy from infants with unexplained cause of death. *Clinical Chemistry*, **47**, 1166–1182.

Chace DH, Kalas TA and Naylor EW (2002) The application of tandem mass spectrometry to neonatal screening for inherited disorders of intermediary metabolism. *Annual Review of Genomics and Human Genetics*, **3**, 17–45.

Chace DH, Kalas TA and Naylor EW (2003) Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clinical Chemistry*, **49**, 1797–1817.

Chamoles NA, Blanco MB, Gaggioli D and Casentini C (2001) Hurler-like phenotype: Enzymatic diagnosis in dried blood spots on filter paper. *Clinical Chemistry*, **47**, 2098–2102.

Dussault JH, Coulombe P, Laberge C, Letarte J, Guyda H and Khoury K (1975) Preliminary report on a mass screening program for neonatal hypothyroidism. *Journal of Pediatrics*, **86**, 670–674.

Garrick MD, Dembure P and Guthrie R (1973) Sickle-cell anemia and other hemoglobinopathies. Procedures and strategy for screening employing spots of blood on filter paper as specimens. *New England Journal of Medicine*, **288**, 1265–1268.

Grenier A, Lescault A, Laberge C, Gagne R and Mamer O (1982) Detection of succinylacetone and the use of its measurement in mass screening for hereditary tyrosinemia. *Clinica Chimica Acta*, **123**, 93–99.

Guthrie R (1968) Screening for “inborn errors of metabolism” in the newborn infant – a multiple test program. *Birth Defects Original Article Series*, **IV**, 92–98.

Heard GS, Secor McVoy JR and Wolf B (1984) A screening method for biotinidase deficiency in newborns. *Clinical Chemistry*, **30**, 125–127.

Iafolla AK, Thompson RJ Jr and Roe CR (1994) Medium-chain acyl-coenzyme A dehydrogenase deficiency: Clinical course in 120 affected children. *Journal of Pediatrics*, **124**, 409–415.

Kim SZ, Varvogli L, Waisbren SE and Levy HL (1997) Hydroxyprolinemia: Comparison of a patient and her unaffected twin sister. *Journal of Pediatrics*, **130**, 437–441.

McCabe ER, Huang SZ, Seltzer WK and Law ML (1987) DNA microextraction from dried blood spots on filter paper blotters: Potential applications to newborn screening. *Human Genetics*, **75**, 213–216.

McLafferty FW (1981) Tandem mass spectrometry. *Science*, **214**, 280–287.

Millington DS, Kodo N, Norwood DL and Roe CR (1990) Tandem mass spectrometry: A new method for acylcarnitine profiling with potential for neonatal screening for inborn errors of metabolism. *Journal of Inherited Metabolic Disease*, **13**, 321–324.

Naylor EW and Chace DH (1999) Automated tandem mass spectrometry for mass newborn screening for disorders in fatty acid, organic acid, and amino acid metabolism. *Journal of Child Neurology*, **14**, S4–S8.

Pang S, Hotchkiss J, Drash AL, Levine LS and New MI (1977) Microfilter paper method for 17 alpha-hydroxyprogesterone radioimmunoassay: Its application for rapid screening for congenital adrenal hyperplasia. *Journal of Clinical Endocrinology and Metabolism*, **45**, 1003–1008.

Peterschmitt MJ, Simmons JR and Levy HL (1999) Reduction of false negative results in screening of newborns for homocystinuria. *New England Journal of Medicine*, **341**, 1572–1576.

Schulze A, Lindner M, Kohlmüller D, Olgemöller K, Mayatepek E and Hoffmann GF (2003) Expanded newborn screening for inborn errors of metabolism by electrospray ionization-tandem mass spectrometry: Results, outcome, and implications. *Pediatrics*, **111**, 1399–1406.

Smith WE, Millington DS, Koeberl DD and Lesser PS (2001) Glutaric acidemia, type I, missed by newborn screening in an infant with dystonia following promethazine administration. *Pediatrics*, **107**, 1184–1187.

Wagstaff J, Korson M, Kraus JP and Levy HL (1991) Severe folate deficiency and pancytopenia in a nutritionally deprived infant and homocystinuria caused by cystathionine beta-synthase deficiency. *Journal of Pediatrics*, **118**, 569–572.

Waisbren SE, Albers S, Amato S, Ampola M, Brewster TG, Demmer L, Eaton RB, Greenstein R, Korson M, Larson C, *et al*. (2003) Effect of expanded newborn screening for biochemical genetic disorders on child outcomes and parental stress. *Journal of the American Medical Association*, **290**, 2564–2572.

Wilcken B, Smith A and Brown DA (1980) Urine screening for aminoacidopathies: Is it beneficial? Results of a long-term follow-up of cases detected by screening one million babies. *Journal of Pediatrics*, **97**, 492–497.

Wilcken B, Wiley V, Hammond J and Carpenter K (2003) Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *New England Journal of Medicine*, **348**, 2304–2312.

Zytkovicz TH, Fitzgerald EF, Marsden D, Larson CA, Shih VE, Johnson DM, Strauss AW, Comeau AM, Eaton RB and Grady GF (2001) Tandem mass spectrometric analysis for amino, organic, and fatty acid disorders in newborn dried blood spots: A two-year summary from the New England Newborn Screening Program. *Clinical Chemistry*, **47**, 1945–1955.

Specialist Review

Advances in cytogenetic diagnosis

Daynna J. Wolff
Medical University of South Carolina, Charleston, SC, USA

1. Introduction

Long before the human genome project made the idea of studying genomes alluring, the original "genome screen", a karyotype, was used to detect genetic aberrations. Just three years after Tjio and Levan (1956) reported that human cells contained 46 chromosomes, Lejeune *et al*. (1959), Ford *et al*. (1959), and Jacobs and Strong (1959) described the numerical chromosome abnormalities associated with Trisomy 21, Turner syndrome, and Klinefelter syndrome, respectively. Thus, the importance of clinical cytogenetics in medical practice was established. Since the early days of assessing the number of solid-stained chromosomes, clinical cytogenetics has been, and continues to be transformed by technological advances such as high-resolution banding, fluorescence in situ hybridization (FISH), and comparative genomic hybridization (CGH). This chapter presents a short overview of clinical cytogenetics and highlights recent advances in the field.

2. Overview of clinical cytogenetics

The study of chromosomes has become standard of care for many constitutional, as well as acquired, disorders. Pregnant women over the age of 35 and those with aberrant serum screening studies or abnormal ultrasound findings are routinely offered prenatal diagnosis for chromosome disorders. Newborns with multiple congenital abnormalities, children with developmental delays or mental retardation, and couples with reproductive difficulties have chromosomal analyses as part of the routine laboratory evaluation (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1 and Article 20, **Cytogenetics of infertility**, Volume 1). In addition, cytogenetic analysis of hematologic malignancies and solid tumors provides important information for diagnosis, prognosis, and treatment decisions (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1).

Generation of a banded karyotype using traditional cytogenetic methods requires that cells be in the metaphase component of the cell cycle. Spontaneously dividing cells may be obtained directly from bone marrow, lymph nodes, and chorionic villi trophoblasts. However, cell culture, with or without mitogens, is necessary to assess

cells from amniotic fluid, solid tissues, and peripheral blood. The basic steps for obtaining cells in metaphase and preparing for cytogenetic analysis are the same for all cell types: arresting cells in metaphase by disrupting spindle fiber attachments using colcemid, swelling cells by placing in a hypotonic solution, fixing the cells using a methanol:acetic acid treatment, releasing the chromosomes from cell membrane onto slides or coverslips, and staining the chromosomes to achieve the light and dark banding patterns. Once these procedures have been performed, skilled microscope technologists and the laboratory director evaluate metaphase cells and the arranged karyotypes to assess for chromosomal aberrations. Standard karyotypic studies typically detect chromosomal abnormalities that involve approximately 5–10 Mb of DNA. High-resolution studies of cells captured in the prometaphase stage of the cell cycle can distinguish greater than 1000 separate bands and allow for the identification of alterations involving 3–5 Mb (Yunis, 1981).

Three additional methods have been particularly useful in achieving an even higher resolution, such that small alterations (<3 Mb) in DNA copy number or location can be detected. These methods are fluorescence in situ hybridization (FISH), multicolor painting/banding (SKY, M-FISH, and M-banding), and comparative genomic hybridization (CGH). These techniques enable cytogeneticists to go from the chromosomal level to the molecular level and to resolve the most complex karyotypes.

3. FISHing for abnormalities

Incorporation of FISH (also known as molecular cytogenetics) has greatly increased the sensitivity of chromosome studies (*see* Article 22, **FISH**, Volume 1). The process of FISH involves hybridization of cloned fluorochrome-labeled DNA fragments (probes) to target chromosomal sequences. Probe DNA is denatured in the presence of Cot-1 DNA that binds highly and moderately repetitive genomic sequences to allow for select hybridization of the unique sequence DNA to the target. Probe detection is accomplished by UV-light excitement of a fluorochrome, such as fluorescein-5-thiocyanate (FITC) or rhodamine that is either directly attached to the probe DNA or attached to a hapten (biotin or digoxygenin)-labeled probe.

Using FISH technology, subtle aberrations in chromosome content can be readily assessed and linked with clinical disorders. FISH has not only been used to detect submicroscopic alterations in metaphase cells, it may also be performed on interphase nuclei. Importantly, FISH probes have been mapped to chromosomal locations and molecular sequence maps, allowing for the direct connection of data being generated from the human genome project to clinical phenotypes caused by cytogenetic abnormalities. These links provide clues to the molecular basis of disease and the genes involved.

One of the first clinical applications of FISH was the development of definitive diagnostic tests for microdeletion and microduplication syndromes that result from alterations of submicroscopic amounts of genetic material, typically <2 Mb (*see* Article 17, **Microdeletions**, Volume 1). For example, FISH is routinely used to identify microdeletions of 22q11.2 in patients with DiGeorge and Velocardiofacial (VCF) syndromes (Figure 1a). Because these relatively common syndromes have

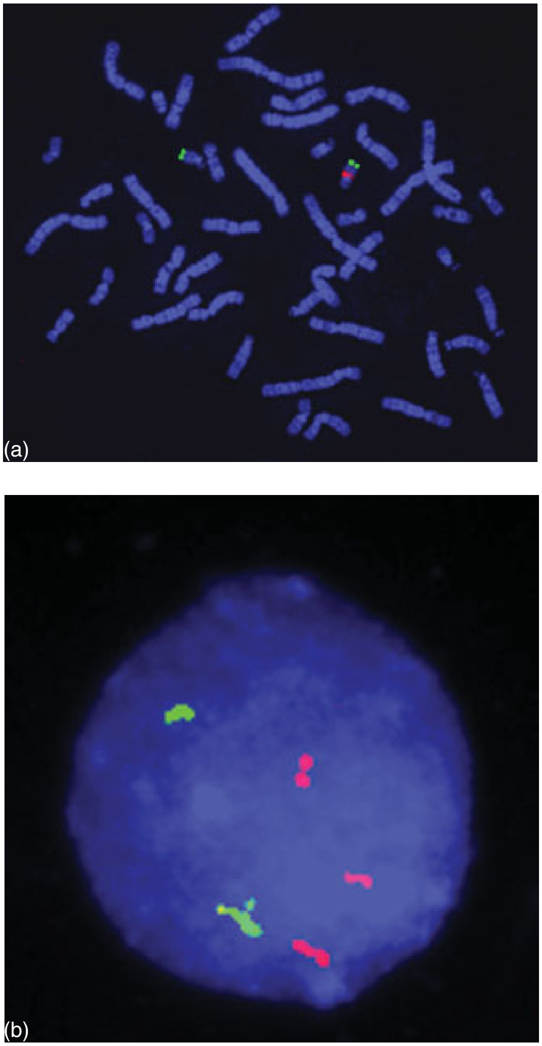


Figure 1 FISH with a probe for TUPLE1 on a metaphase cell from a patient with DiGeorge syndrome revealing a deletion of 22q11.2 (a), and on an interphase cell from a patient with Dup(22)(q11.2q11.2) syndrome (9) (b) (Figure supplied by S. Jalal, Mayo Clinic.)

relatively high frequencies, they are associated with congenital heart disease and may be inherited in a significant number of cases (Swillen *et al*., 1998), FISH studies are also performed for fetuses and newborns with a heart defect, and parents of children diagnosed with DiGeorge or VCF. The majority of the deletions, which are mediated by repetitive duplicated regions, have the same proximal and distal breakpoints (Lindsey, 2001; Lupski, 1998). The nonallelic homologous recombination events that generate deletions may also result in duplications of 22q11.2 (Figure 1b), which has been associated with dysmorphic features, growth failure, cognitive deficits, hearing loss, and velopharyngeal insufficiency

Table 1 Common genetic abnormalities in malignancies that are routinely studied using FISH

Chromosomal aberration[a]	Chromosome – gene(s) involved	Disease association
t(9;22)(q34;q11.2)	9 – *ABL*1 22 – *BCR*	Acute lymphocytic leukemia Acute myeloblastic leukemia Chronic myelogenous leukemia
t(15;17)(q22;q21.1)	15 – *PML* 17 – *RARA*	Acute myeloblastic leukemia
t(*;11)(*.*; q23)	11 – *MLL*	Acute lymphocytic leukemia Acute myeloblastic leukemia
t(8;21)(q22;q22)	8 – *CBFA2T*1 (*ETO*) 21 – *RUNX*1 (*AML*1)	Acute myeloblastic leukemia
inv(16)(p13q22) or t(16;16)(p13;q22)	16q22 – *CBFB*	Acute myeloblastic leukemia
t(12;21)(p13;q22)	12 – *ETV*6 (*TEL*) 21 – *RUNX*1 (*AML*1)	Acute lymphocytic leukemia
Trisomy 8	8 – 8*cen*	Acute lymphocytic leukemia Chronic myelogenous leukemia
t(8;14)(q24;q32)	8 – *MYC* 14 – *IGH*	Acute lymphocytic leukemia Non-Hodgkin's lymphoma
t(11;14)(q13;q34)	11 – *CCND*1 14 – *IGH*	Non-Hodgkin's lymphoma Multiple myeloma
t(14;18)(q32;q21)	14 – *IGH* 18 – *BCL*2	Non-Hodgkin's lymphoma
t(*;14)(*.*;q32)	14 – *IGH*	Non-Hodgkin's lymphoma Multiple myeloma
Del(13)(q14) or −13	Unknown tumor suppressor	Chronic lymphocytic leukemia Multiple myeloma
Trisomy 12	12 – 12cen unknown gene(s)	Chronic lymphocytic leukemia
del(11)(q23)	*ATM*	Chronic lymphocytic leukemia
HER-2 amplifaction	*ERBB*2 (*HER*2)	Breast cancer

[a] An asterisk (*) is used to delineate multiple loci or breakpoints.

(Ensenauer *et al*., 2003). FISH is also useful for studying recombination at low copy repeats resulting in deletion/duplication of 15q11.2 (Prader–Willi/Angelman/Dup(15)(q11.2q11.2) syndromes), 17p11.2 (Smith–Magenis/dup(17)(p11.2p11.2) syndromes), and 17p12 (hereditary neuropathy with liability to pressure palsies (HNPP)/Charcot–Marie–Tooth disease type 1A (CMT1A) syndromes) (Lupski, 1998).

An area that has been advanced greatly by FISH is the study of chromosomal abnormalities associated with cancer. Probes have been developed for the majority of recurrent translocations, as well as for some deletions, duplications, and amplifications found in hematologic malignancies and solid tumors (Table 1) (Wolff and Schwartz, 2004). Commercially available *BCR/ABL*1 probe sets are used to identify *BCR/ABL*1 fusion events in diagnostic samples of patients with chronic

myelogenous leukemia (CML) and are also used for posttreatment residual disease assessment. In addition, FISH is valuable for detecting the nearly 20% of CML patients with atypical FISH patterns resulting from microdeletions at the ABL1 breakpoint (Sinclair *et al*., 2000). Patients with these microdeletions have a relatively poor prognosis and reduced response to treatment (Huntly *et al*., 2001; Cohen *et al*., 2001; Smoley *et al*., 2004). As the deletions are not visible with routine cytogenetics and cannot be readily detected by polymerase chain reaction (PCR), FISH is the optimal methodology for accessing this prognostic marker.

One of the most important advantages of FISH over routine cytogenetic studies is the ability to enumerate chromosomes and to detect structural aberrations in interphase cells. For acute lymphocytic leukemia (ALL), routine cytogenetic studies often produce suboptimal results; therefore, FISH is a valuable adjunct to routine testing. Many clinical laboratories offer an ALL FISH screening panel assessment that may include probes for t(9;22) (*BCR/ABL*1), 11q23 (*MLL)*, t(12;21) (*ETV* 6/*RUNX* 1), *MYC*, and a common deletion in 9p21-22 (Andreasson *et al*., 2000; Nordgren *et al*., 2002). The panel of probes will detect the majority of genetic abnormalities associated with ALL, particularly since the screen is also useful for unmasking hidden hyperdiploidy when multiple signals are seen for various probes. In chronic lymphocytic leukemia (CLL), the disease cells have a low mitotic index; therefore, the more sensitive interphase FISH assay with a battery of probes (Figure 2) has largely replaced conventional cytogenetics (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1). FISH studies reveal that the most common abnormalities include deletions of 13q14, trisomy 12, deletions of 11q22.3-q23.1, deletions of 17p13, and deletions of 6q21-q23 (Dohner *et al*., 2000). These genetic abnormalities are important independent predictors of CLL progression and survival.

FISH provides the framework for monitoring numerous loci in a single test format. One of the best examples is a probe panel used to detect subtle aberrations near the ends of chromosomes (Knight *et al*., 1997, Figure 3). Homologous recombination within repetitive telomeric regions can result in submicroscopic deletions or duplications of the adjacent unique DNA sequence (subtelomeric sequences). Because the subtelomeres are extremely gene rich (Saccone *et al*., 1992), abnormalities in these regions are likely to have clinical significance. FISH studies with probes for the subtelomeric DNA sequences that are found approximately 100–300 kb from the physical ends of the chromosomes has increased the aberration detection sensitivity for the subtelomeric sequences by at least 10-fold over conventional chromosome studies (National Institutes of Child Health and Institute of Molecular Medicine Collaboration, 1996). Abnormalities have been documented by FISH in approximately 5% of patients with idiopathic mental retardation and previous normal karyotypic analysis (Knight *et al*., 1999; Joyce *et al*., 2001; Popp *et al*., 2002). Detection rates are highest in cases with moderate to severe mental retardation (Knight *et al*., 1999) and in cases with prenatal onset of growth retardation, dysmorphic features, and/or a positive history of mental retardation (deVries *et al*., 2001).

Another widely used FISH panel, consisting of probes for chromosomes 13, 18, 21, and the sex chromosomes, identifies numerical abnormalities, as well as triploidy, in uncultured amniotic fluid cells for prenatal diagnostic studies (Klinger

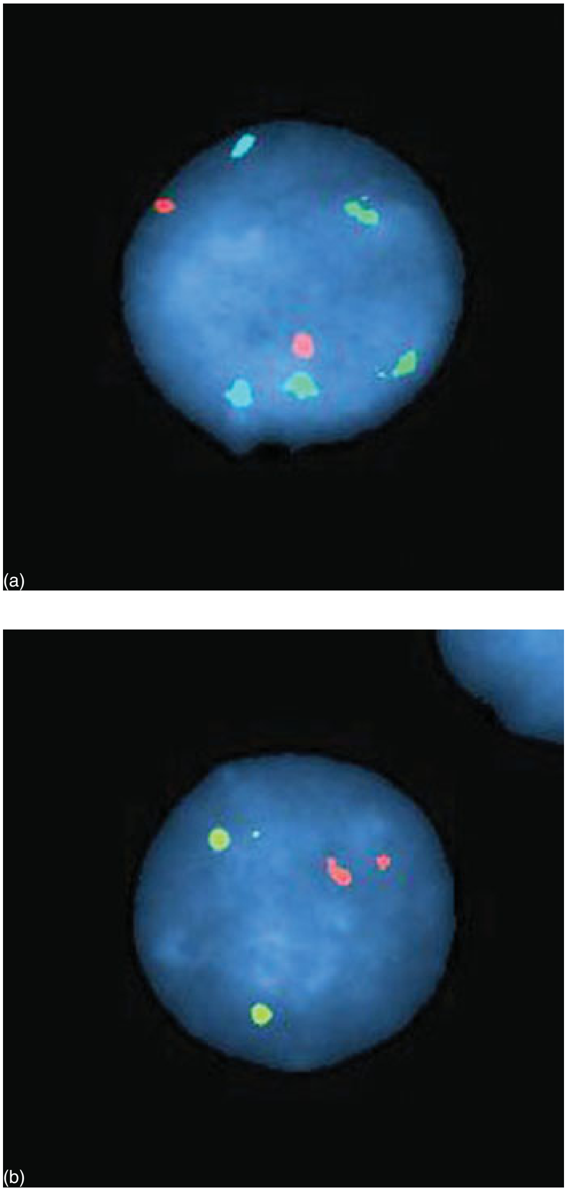


Figure 2 Results of a CLL probe panel (Vysis, Inc, Downers Grove, IL) analysis on interphase cells from a patient diagnosed with chronic lymphocytic leukemia. The analysis revealed three green signals for the probe for the centromere of chromosome 12, consistent with trisomy 12, two red and two aqua signals for the probes for 13q14 and the 13 centromere, respectively (a), two red signals for the p53 gene, and two green signals for the ATM gene (b)

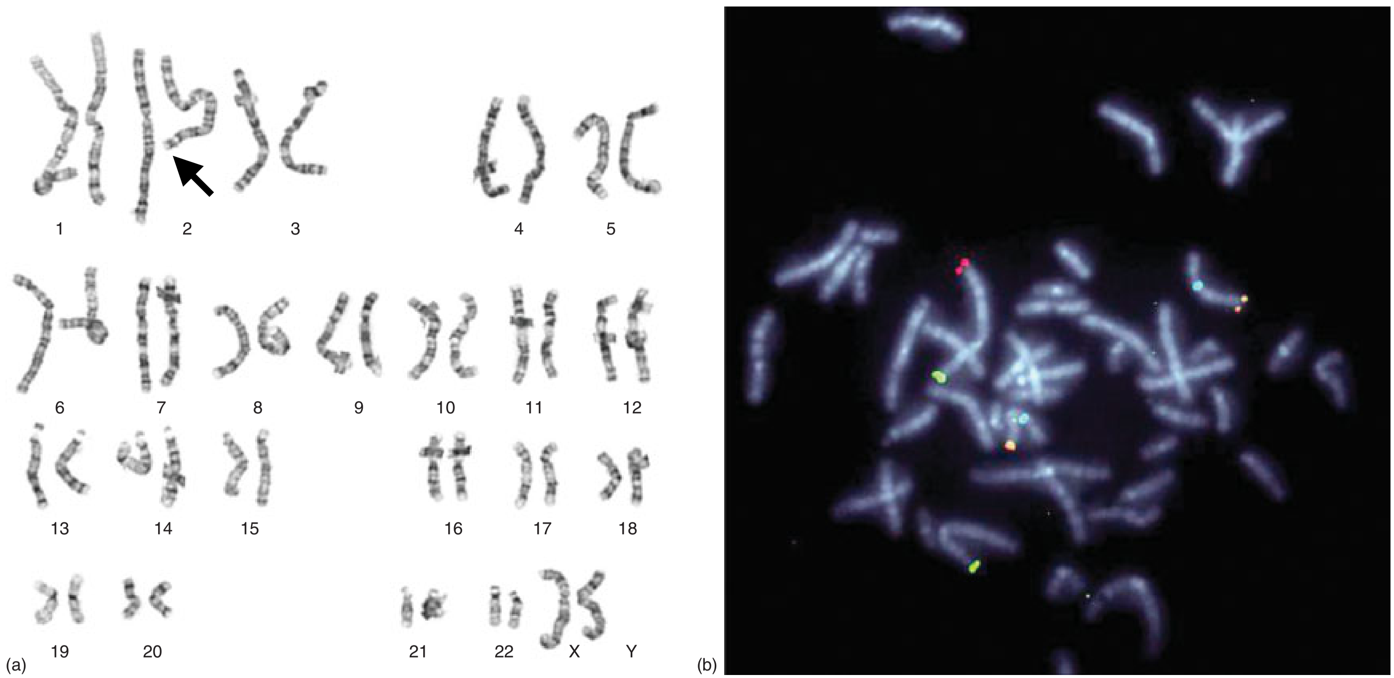


Figure 3 Karyotype (a) and partial results of a subtelomeric probe panel (ToTelVysion, Vysis, Inc., Downers Grove, IL) analysis (b) for a female patient with developmental delay and autism. The probe cocktail shown contains probes for chromosomal regions 2p (green), 2q (red), X centromere (blue) and Xq/Yq (yellow). This patient had a subtle deletion of the subtelomeric region on chromosome 2q (arrows)

et al., 1992). This technology allows results to be generated within 24 h for cases with advanced maternal age or pregnancies indicated to be at increased risk due to maternal screening results or abnormal ultrasound findings. Enumeration of signals for these five probes in uncultured amniotic fluid detects the majority (65–80%) of chromosomal abnormalities (Test and technology transfer committee of the American College of Medical Genetics, 2000). For couples at high risk for genetic disease, FISH is also valuable for assessing embryo sex and chromosome number prior to implantation [preimplantation genetic diagnosis (PGD)] so that the birth of an affected child can be avoided (Findley, 2000). As with prenatal diagnosis, the common probes used for ploidy evaluation are for chromosomes 13, 18, 21, X, and Y. For PGD of translocations when one of the parents is a known carrier, FISH with subtelomeric probes provides diagnostic information.

4. The SKY is the limit

FISHing got more interesting with the ability to simultaneously visualize all 24 human chromosome pairs with different colors (Schrock *et al*., 1996; Speicher *et al*., 1996). Spectral karyotyping (SKY, Figure 4) or Multiplex-FISH (M-FISH) techniques incorporate combinatorial or ratio-based fluorochrome labeling schemes to produce 24 colors, one for each chromosome. To accomplish this, whole chromosome probes, composed of unique and moderately repetitive sequences from an entire chromosome, are labeled with specific combinations of fluorescent dyes. The probes are generated from DNA from a particular chromosome that is isolated from the rest of the genome by flow sorting, creation of somatic cell hybrids containing a single human chromosome or area of a chromosome, or microdissection of chromosomes and subsequent amplification of the dissected DNA sequences via the PCR (Gray *et al*., 1992). Hybridization of the collection of whole chromosome probes to metaphase cells produces continuous fluorescent signals on each of the chromosomes. For SKY, the spectral characteristics of each pixel are scored by an inferometer (Schrock *et al*., 1996); whereas for M-FISH, images are collected through a series of excitation and emission filters and computer software is used to classify each chromosomal segment (Speicher *et al*., 1996). The computer then applies a distinct pseudocolor for each chromosome allowing the cytogeneticist to view a karyotype with each chromosome "painted" a different color.

This type of analysis has been especially useful for defining complex rearrangements, such as those seen in neoplastic disorders and solid tumors. For example, in a study of 30 children with acute lymphoblastic leukemia whose leukemic blast cells lacked chromosomal abnormalities detectable by conventional cytogenetics or whose blast cells had multiple chromosomal abnormalities that could not be completely characterized by G-banding analysis, SKY identified three cryptic translocations in the previously normal category and was also successful in defining the nature of the chromosomal abnormalities in 4 of the 10 patients with marker and derivative chromosomes (Matthew *et al*., 2001). SKY and M-FISH have also been used to determine the origin of de novo duplications and marker chromosomes in pre- and postnatal cases (Haddad *et al*., 1998; Uhrig *et al*., 1999).

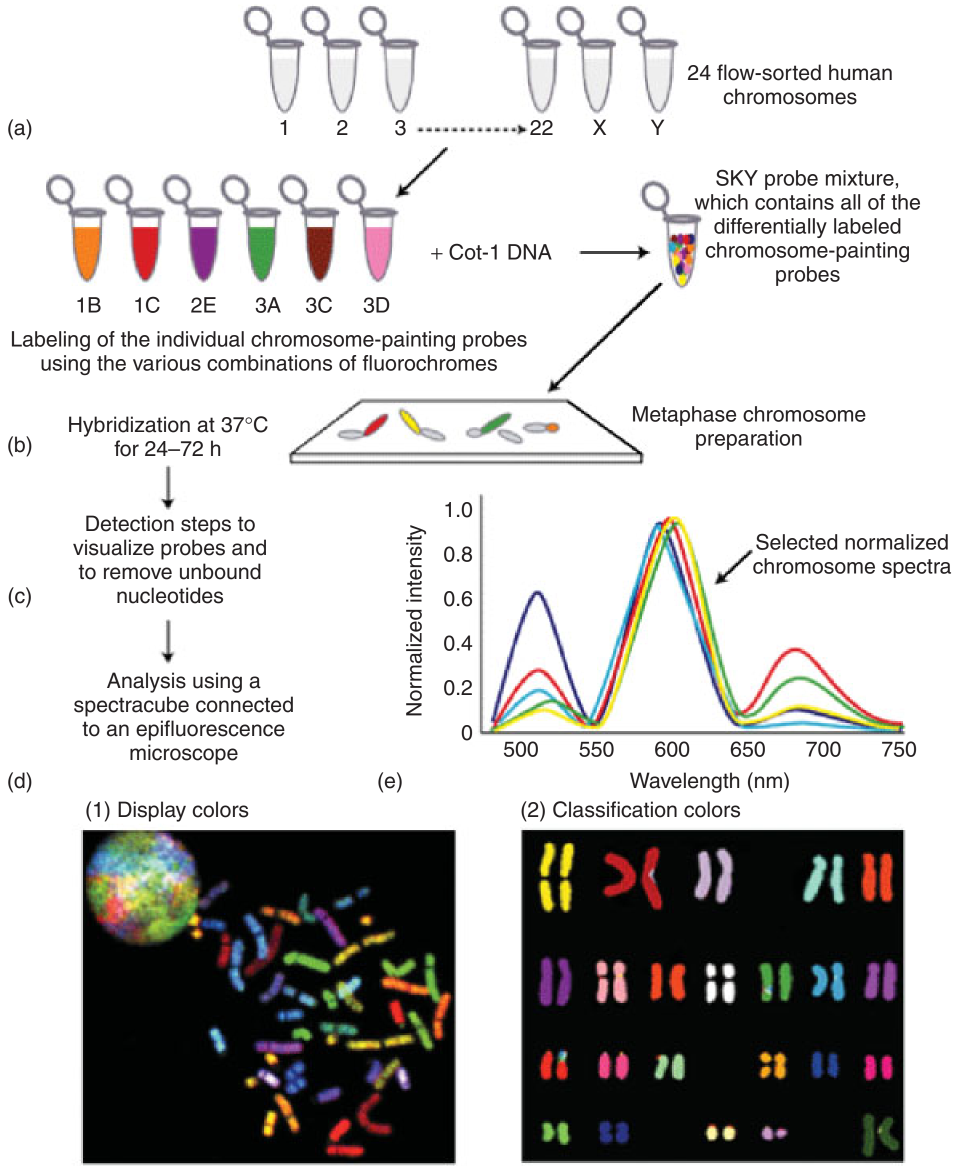


Figure 4 Schematic of cytogenetic analysis using spectral karyotyping (SKY). (a) Flow-sorted human chromosomes are combinatorially labeled with at least one, and as many as five, fluorochrome combinations to create a unique spectral color for each chromosome pair. (b) The SKY probe mixture is hybridized to a metaphase chromosome preparation. (c) Detection of the hybridized sample. (d) Visualization of the SKY hybridization using a Spectracube connected to an epifluorescence microscope. (e) Spectral classification assigns a discrete color to all pixels, which is converted to a display image that is based on the fluorescence intensities of all of the chromosome-painting probes: (1) classification colors are assigned to the chromosomes; (2) this measurement forms the basis for automated chromosome identification. (Reproduced by permission from Thomas Reed and Cambridge University Press)

Multicolor banding (Chudoba *et al*., 1999) uses similar technology to produce differential painting of specific sections of the chromosomes. Mixtures of partial-chromosome paints are labeled with various fluorochromes and a computer program analyzes metaphase chromosome data producing a pseudocolored, banded karyotype with an estimated resolution of 550-bands, regardless of chromosome length. This methodology is advantageous for the determination of breakpoints and the analysis of intrachromosomal rearrangements, and has been used to identify and classify markers and derivative chromosomes (Muller, 1997).

5. FISH and CGH chips

The introduction of comparative genomic hybridization (CGH), in which two genomes are directly compared for DNA content differences (Kallioniemi, 1992), allowed cytogenetics to identify chromosomal imbalances without directly studying the patient's chromosomes (*see* Article 23, **Comparative genomic hybridization**, Volume 1). For CGH, DNA from a test genome and DNA from a reference genome are differentially labeled in two distinct fluorochromes (red and green, respectively, for example) and are allowed to compete for hybridization sites on normal metaphase chromosomes (Figure 5). The ratio of the red to green is measured and sequences that are overrepresented or amplified in the test DNA appear more red, deleted regions appear greener, and an equal representation of test and reference appears yellow. CGH has been particularly useful for studying solid tumors and has led to the identification of several oncogenes important for ovarian cancer (Shayesteh *et al*., 1999) and pancreatic cancer (Wallrapp *et al*., 1999). CGH is limited however by its reliance on metaphase chromosomes for the detection of subtle rearrangements with an ability to distinguish chromosomal losses on the order of 5–10 Mb and amplifications of 2 Mb.

The highest resolution analysis of the genome has been promised by the recent development of CGH applied to genomic microarrays. For array CGH, vectors such as bacterial artificial chromosomes (BACs), containing cloned DNA fragments, are substituted for metaphase chromosomes as targets for the test DNA hybridization (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998; Geschwind *et al*., 1998) (Figure 5). The genomic resolution achievable with CGH array analysis depends solely upon the map distances between markers and the length of the clones used (Solinas-Toldo *et al*., 1997). Thus, this methodology has the potential to identify chromosomal alterations on the gene level.

Most CGH array studies have been applied to malignancies, such as breast cancer (Albertson *et al*., 2000; Daigo *et al*., 2001), ovarian cancer (Schraml *et al*., 2003) and gastric cancer (Weiss *et al*., 2003). These analyses provide a high-resolution screen for detecting copy number changes of putative oncogenes and tumor suppressor genes. For constitutional studies, arrays designed to cover the entire chromosome 22 have been used to detect abnormalities that result from the gain or loss of a chromosomal region (Buckley *et al*., 2002). Focused arrays have also been developed for the detailed analysis of the subtelomeric regions (Veltman *et al*., 2002) and 1p36 (Yu *et al*., 2003). Recently, a clone set composed of >32,000 BACs with an average resolution of 75 kb has been developed

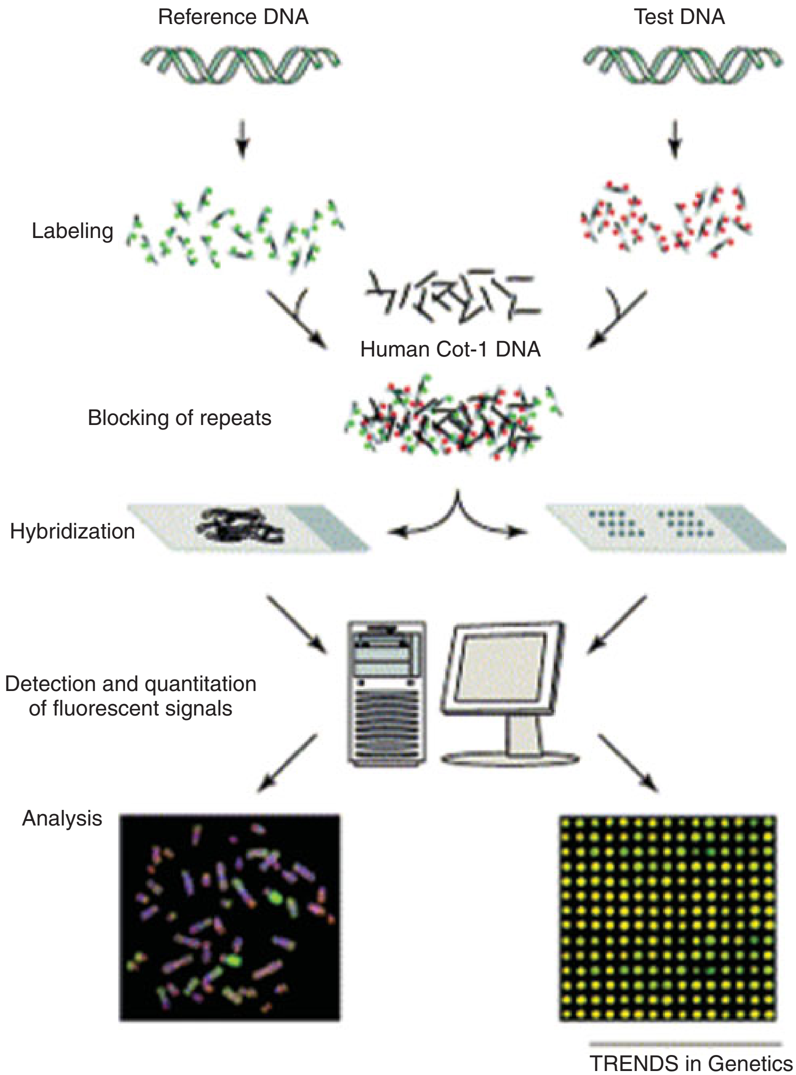


Figure 5 Comparative genomic hybridization (CGH) on metaphase spreads and arrays. Test and genomic DNAs are differentially labeled and mixed with Cot-1 DNA to block repetitive sequences. The probe hybridization mixture is placed on a slide containing metaphase chromosomes or array targets and hybridized. Fluorescent signals are detected and image-processing software identifies regions with an abnormal test: reference ratio, indicating regions with a loss or gain of DNA sequence. (Reproduced by permission from Elsevier)

to cover approximately 99% of the current human genome sequence assembly (**www.bcgsc.ca/lab/mapping/bacrarray/human/**). While a wonderful research tool, the whole-genome approach of aberration detection may not be the best clinical assessment tool as it is likely to uncover many polymorphic regions with no clinical significance.

Targeted array CGH has the potential to be the major technique used for diagnosis of chromosome imbalances, due to the reduced labor compared with conventional cytogenetic analysis, its high-resolution assessment, and the promise of automation. However, array CGH will not detect balanced rearrangements such as inversions, insertions, and translocations. Thus, the original genome scanning tool, karyotypic analysis, will still be necessary for parental and family studies when an abnormality is detected with array CGH and for assessment of all cases with normal microarray results.

6. New cytogenetic horizons

While advances in cytogenetic technologies have greatly increased the ability to diagnose genetic imbalances, limitations exist. Assay sensitivities are determined by the content encompassed by the FISH probe or BAC used on an array. Thus, owing to the large size of the target sequences, small deletions and duplications within those sequences will still be overlooked. One approach to overcome this limitation is to use single copy DNA probes as short as most PCR products. ScFISH, based upon the design and synthesis of single copy DNA probes for FISH (Rogan *et al*., 2001) has detected tiny deletions in the Prader–Willi/Angelman imprinting control region and has differentiated chromosomal breakpoints for a variety of genomic rearrangements (e.g., t(9;22) in CML and the atypical deletion in Smith–Magenis syndrome) at a resolution similar to that of genomic Southern blot analysis (Knoll and Rogan, 2003). The properties of scFISH probes not only allow for detection of aberrations less than 10 kb but also reveal clinically significant heterogeneity of chromosomal breakpoints among affected individuals that have apparently identical molecular cytogenetic findings using conventional genomic FISH probes (Knoll and Rogan, 2003).

Another limitation of cytogenetic and most molecular techniques is the lack of ability to assess clinically important characteristics of the target cells. For example, routine FISH assessment of a bone marrow sample from a patient with multiple myeloma may yield normal results owing to the low percentage of diseased plasma cells. Thus, a method to enrich for or to label the cells of interest is needed to offer a sensitive and specific clinical test (Ahmann *et al*., 1998). An easy way to identify the cells is to use a specific fluorescence antibody to label the cells concurrently with the FISH probe assessment (Weber-Matthiesen *et al*., 1995).

An exciting technique with potential clinical cytogenetic application involves a liquid-based hybridization method and analysis using the Luminex flow cytometric platform. For this methodology, fluorescent tagged microspheres are coupled with single copy DNA sequences complementary to clinically relevant chromosomal regions (capture probes). To quantitate the genetic content of the specific genomic regions of interest, the oligonucleotide capture probes immobilized on the fluorescent microspheres are mixed with patient DNA samples. Hybridization of probe and target occurs in the liquid suspension and, following a wash step, the hybridized microspheres are passed through a Luminex counting device that records the fluorescence intensity of each microsphere. The fluorescence intensity

is related to copy number of the bound target DNA and deviations from a normalized fluorescence corresponding to two copies of sequence is recorded as gains or losses. Up to 100 distinct microbeads with different fluorescent color codes are available from Luminex, allowing for the detection and quantitation of 100 single copy DNA regions of interest in a single assay. The beauty of this technology is the ability to inexpensively manufacture appropriate microbeads and to analyze hundreds of targets within several hours. Similar analyses have been applied to gene expression studies (Yang *et al*., 2001) and single nucleotide polymorphism identification (Chen *et al*., 2000; Wedemeyer *et al*., 2001), revealing the utility of coupling bead array DNA hybridization with flow cytometry.

7. The end of the beginning

The medical importance of scanning the human genome for abnormalities associated with clinical phenotypes was recognized when the first numerical chromosome aberrations were described (Lejeune *et al*., 1959; Ford *et al*., 1959; Jacobs and Strong, 1959). From their unique study of the genome through the microscope, clinical cytogeneticists have been uncovering the links between patients' abnormal phenotypes and changes in chromosome content or arrangement. Recently, this original form of genome assessment has greatly benefited from technological advances and information being generated from the Human Genome project. Routine chromosome assessment is now coupled with FISH, CGH, and BAC array data to provide high-resolution, sensitive evaluation of human genomes. These techniques have linked traditional cytogenetic banding techniques, capable of identifying aberrations on the megabase scale, with the basic sequence data of molecular biology. Recent developments and future applications will offer unprecedented opportunities to study the relationship between chromosomes and phenotype that will undoubtedly have a significant impact on the practice of medicine.

References

Ahmann GJ, Jalal SM, Juneau AL, Christensen ER, Hanson CA, Dewald GW and Greipp PR (1998) A novel three-color, clone-specific fluorescence in situ hybridization procedure for monoclonal gammopathies. *Cancer Genetics and Cytogenetics*, **1010**, 7–11.

Albertson DG, Ylstra B, Segraves R, Collins C, Dairkee SH, Kowbel D, Kuo WL, Gray JW and Pinkel D (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genetics*, **25**, 144–146.

Andreasson P, Hoglund M, Bekassy AN, Garwicz S, Heldrup J, Mitelman F and Johansson B (2000) Cytogenetic and FISH studies of a single center consecutive series of 152 childhood acute lymphoblastic leukemias. *European Journal of Haematology*, **65**, 40–51.

Buckley PG, Mantripragada KK, Benetkiewicz M, Tapia-Paez I, Diaz De Stahl T, Rosenquist M, Ali H, Jarbo C, De Bustos C, Hirvela C, *et al*. (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Human Molecular Genetics*, **1**, 3221–3229.

Chen J, Iannone MA, Li MS, Taylor JD, Rivers P, Nelsen AJ, Slentz-Kesler KA, Roses A and Weiner MP (2000) A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Research*, **10**, 549–557.

Chudoba I, Plesch A, Lorch T, Lemke J, Claussen U and Senger G (1999) High resolution multicolor-banding: a new technique for refined FISH analysis of human chromosomes. Cytogenet. *Cell Genetics*, **84**, 156–160.

Cohen N, Rozenfeld-Granot G, Hardan I, Brok-Simoni F, Amariglio N, Rechavi G and Trakhtenbrot L (2001) Subgroup of patients with Philadelphia-positive chronic myelogenous leukemia characterized by a deletion of 9q proximal to ABL gene: expression profiling, resistance to interferon therapy, and poor prognosis. *Cancer Genetics and Cytogenetics*, **128**, 114–119.

Daigo Y, Chin SF, Gorringe KL, Bobrow LG, Ponder BA, Pharoah PD and Caldas C (2001) Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers: a new approach for the molecular analysis of paraffin-embedded cancer tissue. *American Journal of Pathology*, **158**, 1623–1631.

de Vries BB, White SM, Knight SJ, Regan R, Homfray T, Young ID, Super M, McKeown C, Splitt M, Quarrell OW, *et al*. (2001) Clinical studies on submicroscopic subtelomeric rearrangements: a checklist. *Journal of Medical Genetics*, **38**, 145–150.

Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, Bullinger L, Dohner K, Bentz M and Lichter P (2000) Genomic aberrations and survival in chronic lymphocytic leukemia. *The New England Journal of Medicine*, **343**, 1910–1916.

Ensenauer RE, Adeyinka A, Flynn HC, Michels VV, Lindor NM, Dawson DB, Thorland EC, Pham-Lorentz C, Goldstein JL, McDonald MT, *et al*. (2003) Microduplication 22q11.2 an emeging syndrome: clinical, cytogenetic and molecular analysis of 12 patients. *American Journal of Human Genetics*, **73**, 1027–1040.

Findley A (2000) Pre-implantation genetic diagnosis. *British Medical Bulletin*, **56**, 672–690.

Ford CE, Miller OJ, Polani PE, de Almedda JC and Briggs JH (1959) A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet*, **1**, 711–713.

Geschwind DH, Gregg J, Boone K, Karrim J, Pawlikowska-Haddal A, Rao E, Ellison J, Ciccodicola A, D'Urso M, Woods R, *et al*. (1998) Klinefelter's syndrome as a model of anomalous cerebral laterality: testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray. *Development of Genetics*, **23**, 215–229.

Gray JW, Kallioniemi A, Kallioniemi O, Pallavicini M, Waldman F and Pinkel D (1992) Molecular cytogenetics: Diagnosis and prognostic assessment. *Current Opinion in Biotechnology*, **3**, 623–631.

Haddad BR, Schrock E, Meck J, Cowan J, Young H, Ferguson-Smith MA, du Manoir S and Ried T (1998) Identification of de novo chromosomal markers and derivatives by spectral karyotyping. *Human Genetics*, **103**, 619–625.

Huntly BJ, Reid AG, Bench AJ, Campbell LJ, Telford N, Shepherd P, Szer J, Prince HM, Turner P, Grace C, *et al*. (2001) Deletions of the derivative chromosome 9 occur at the time of the Philadelphia translocation and provide a powerful and independent prognostic indicator in chronic myeloid leukemia. *Blood*, **98**, 1732–1738.

Jacobs PA and Strong JA (1959) A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature*, **183**, 302–303.

Joyce CA, Dennis NR, Cooper S and Browne CE (2001) Subtelomeric rearrangements: results from a study of selected and unselected probands with idiopathic mental retardation and control individuals by using high-resolution G-banding and FISH. *Human Genetics*, **109**, 440–451.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Klinger K, Landes G, Shook D, Harvey R, Lopez L, Locke P, Lerner T, Osathanondh R, Leverone B, Houseal T, *et al*. (1992) Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence in situ hybridization (FISH). *American Journal of Human Genetics*, **51**, 55–65.

Knight SJ, Horsley SW, Regan R, Martin Lawrie N, Maher EJ, Cardy DLN, Flint J and Kearney L (1997) Development and clinical application of an innovative fluorescence in situ hybridization technique which detects submicroscopic rearrangements involving telomeres. *European Journal of Human Genetics*, **5**, 1–8.

Knight SJ, Regan R, Nicod A, Horsley SW, Kearney L, Homfray T, Winter RM, Bolton P and Fint J (1999) Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet*, **354**, 1676–1681.

Knoll JH and Rogan PK (2003) Sequence-based, in situ detection of chromosomal abnormalities at high resolution. *American Journal of Medical Genetics*, **121**, 245–257.

Lejeune J, Gautler M and Turpin MR (1959) Etude des chromosomes somatiques de neuf enfants mongoliens. *Comptes rendus de l'Academie des sciences (Paris)*, **248**, 1721–1722.

Lindsey EA (2001) Chromosomal microdeletions: dissecting del 22q11 syndrome. *Nature Reveiws. Genetics.*, **2**, 858–868.

Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, **14**, 417–422.

Matthew S, Rao PH, Dalton J, Downing JR and Raimondi SC (2001) Multicolor spectral karyotyping identifies novel translocations in childhood acute lymphoblastic leukemia. *Leukemia*, **15**, 468–872.

Muller S, Rocchi M, Ferguson-Smith MA and Wienberg J (1997) Toward a multicolor chromosome bar code for the entire human karyotype by fluorescence in situ hybridization. *Human Genetics*, **100**, 271–278.

National Institutes of Child Health and Institute of Molecular Medicine Collaboration (1996) A complete set of human telomeric probes and their clinical applications. *Nature Genetics*, **14**, 86–89.

Nordgren A, Heyman M, Sahlen S, Schoumans J, Soderhall S, Nordenskjold M and Blennow E (2002) Spectral karyotyping and interphase FISH reveal abnormalities not detected by conventional G-banding. Implications for treatment stratification of childhood acute lymphoblastic leukaemia: detailed analysis of 70 cases. *European Journal of Haematology*, **68**, 31–41.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Popp S, Schulze B, Granzow M, Keller M, Holtgreve-Grez H, Schoell B, Brough M, Hager HD, Tariverdian G, Brown J, *et al*. (2002) Study of 30 patients with unexplained developmental delay and dysmorphic features or congenital abnormalities using conventional cytogenetics and multiplex FISH telomere (M-TEL) integrity assay. *Human Genetics*, **111**, 31–39.

Rogan PK, Cazcarro PM and Knoll JH (2001) Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization. *Genome Research*, **11**, 1086–1094.

Saccone S, De Sario A, Della Valle G and Bernardi G (1992) The highest gene concentrations in the human genomce are in telomeric bands of metaphase chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 4913–4917.

Schraml P, Schwerdtfeger G, Burkhalter F, Raggi A, Schmidt D, Ruffalo T, King W, Wilber K, Mihatsch MJ and Moch H (2003) Combined array comparative genomic hybridization and tissue microarray analysis suggest PAK1 at 11q13.5-q14 as a critical oncogene target in ovarian carcinoma. *American Journal of Pathology*, **163**, 985–992.

Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, *et al*. (1996) Multicolor spectral karyotyping of human chromosomes. *Science*, **26**, 494–497.

Shayesteh L, Lu Y, Kuo WL, Baldocchi R, Godfrey T, Collins C, Pinkel D, Powell B, Mills GB and Gray JW (1999) PIK3CA is implicated as an oncogene in ovarian cancer. *Nature Genetics*, **21**, 99–102.

Sinclair PB, Nacheva EP, Leversha M, Telford N, Chang J, Reid A, Bench A, Champion K, Huntly B and Green AR (2000) Large deletions at the t(9;22) breakpoint are common and may identify a poor-prognosis subgroup of patients with chronic myeloid leukemia. *Blood*, **95**, 738–743.

Smoley SA, Brockman SR, Paternoster SF, Meyer RG and Dewald GW (2004) A novel tricolor, dual fusion fluorescence in situ hybridization method to detect BCR/ABL fusion in cells with t(9;22)(q34;q11.2) asssociated with deletion of DNA on the derivative 9 chromosome in chronic myelogenous leukemia. *Cancer Genetics and Cell Genetics*, **148**, 1–6.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, Chromosomes and Cancer*, **20**, 399–407.

Speicher MR, Gwyn Ballard S and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, **12**, 368–375.

Swillen A, Devriendt K, Vantrappen G, Vogels A, Rommel N, Fryns JP, Eyskens B, Gewillig M and Dumoulin M (1998) Familial deletions of chromosome 22q11: the Leuven experience. *American Journal of Medical Genetics*, **80**, 531–532.

Test and technology transfer committee of the American College of Medical Genetics (2000) Technical and clinical assessment of fluorescence in situ hybridization. An ACMG/ASHG position statement. I. Technical considerations. *Genetics in Medicine*, **2**, 356–361.

Tjio HJ and Levan A (1956) The chromosome numbers of man. *Heriditas*, **42**, 1–6.

Uhrig S, Schuffenhauer S, Fauth C, Wirtz A, Daumer-Haas C, Apacik C, Cohen M, Muller-Navia J, Cremer T, Murken J, *et al*. (1999) Multiplex-FISH for pre- and postnatal diagnostic applications. *American Journal of Human Genetics*, **65**, 448–462.

Veltman JA, Schoenmakers EF, Eussen BH, Janssen I, Merkx G, van Cleef B, van Ravenswaaij CM, Brunner HG, Smeets D and van Kessel AG (2002) High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *American Journal of Human Genetics*, **70**, 1269–1276.

Wallrapp C, Muller-Pillasch F, Micha A, Wenger C, Geng M, Solinas-Toldo S, Lichter P, Frohme M, Hoheisel JD, Adler G, *et al*. (1999) Novel technology for detection of genomic and transcriptional alterations in pancreatic cancer. *Annals of Oncology*, **10** Suppl 4, 64–68.

Weber-Matthiesen K, Pressi S, Schlegelburger B and Grote W (1995) Combined immunophenotyping and in situ hybridization (FICTION) – a rapid method to study cell lineae involvement in myelodysplastic disorder. *British Journal of Hematology*, **90**, 701–706.

Wedemeyer N and Potter T (2001) Flow cytometry: an 'old' tool for novel applications in medical genetics. *Clinical Genetics*, **60**, 1–8.

Weiss MM, Snijders AM, Kuipers EJ, Ylstra B, Pinkel D, Meuwissen SG, van Diest PJ, Albertson DG and Meijer GA (2003) Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of Pathology*, **200**, 320–326.

Wolff DJ and Schwartz S (2004) Fluorescence *in situ* Hybridization (FISH). In *The Principles of Clinical Cytogenetics*. (Gerson S and Keagle M, eds), Humana Press, Inc, Totowa, NJ, pp 455–490.

Yang L, Tran DK and Wang X (2001) BADGE, BeadsArray for the Detection of Gene Expression, a high-throughput diagnostic bioassay. *Genome Research*, **11**, 1888–1898.

Yu W, Ballif BC, Kashork CD, Heilstedt HA, Howard LA, Cai WW, White LD, Liu W, Beaudet AL, Bejjani BA, *et al*. (2003) Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. *Human Molecular Genetics*, **12**, 2145–2152.

Yunis JJ (1981) Mid-prophase human chromosomes. The attainment of 2000 bands. *Human Genetics*, **56**, 293–298.

Specialist Review

Current approaches to molecular diagnosis

O. Thomas Mueller
University of South Florida, St. Petersburg, FL, USA

1. Introduction

Of the approximately 30 000 different genes that have been identified by numerous researchers in the Human Genome Project, relatively few have been associated with specific pathogenic conditions (*see* Article 24, **The Human Genome Project**, Volume 3 and Article 31, **Overlapping genes in the human genome**, Volume 3). It will be a continuing challenge to correlate gene with disease in ways that are clinically useful and meaningful, especially in disorders of complex etiology. Current approaches to the identification of mutations in inherited human disorders encompass a variety of methods, including traditional procedures such as Southern blotting and detection by DNA hybridization as well as novel highly efficient techniques that use real-time polymerase chain reaction (PCR) or gene microarrays. Recognition of the most effective and efficient diagnostic approach for a patient with a suspected diagnosis requires an understanding of the molecular pathology of the suspected syndrome(s), as well as a recognition of the limitation of the diagnostic testing available. In some disorders, the molecular pathology is well defined and an efficient diagnostic protocol is in place; however, for most conditions, the most effective protocols will emerge as research and experience dictate and will require a close interaction of clinical and laboratory professionals. This review will summarize the current approaches to diagnostic techniques in a molecular genetic laboratory, emphasizing the complexities in disorders with trinucleotide repeat expansions and those with genetic imprinting mechanisms.

2. Single base-pair detection methods

2.1. *Allele-specific oligonucleotides (ASO)*

This term describes a variety of techniques that have been used for single nucleotide polymorphism (SNP) detection and are based on the hybridization specificity of short oligonucleotides (*see* Article 71, **SNPs and human history**, Volume 4).

These synthesized DNA sequences of approximately 20 bp are constructed to match specific mutant and wild-type base-pair changes and are used to detect specific known mutations. The detection method of hybridized mutant or wild-type oligonucleotide varies with the different applications (Beckmann, 1988).

2.2. *Allele-specific amplification procedure (ASAP) or allele refractory mutation system (ARMS)*

These techniques are variations of SNP detection techniques that are based on the hybridization specificity of oligonucleotides used as PCR primers. Primers are designed to precisely match a known sequence variation, and a second primer pair that matches the wild-type sequence at the same locus is used in a parallel reaction. The matching/mismatching base is positioned at the 3′ or next to the 3′ end of one of the PCR primers. Successful discrimination of the normal and mutant sequences can usually be achieved by optimizing the PCR conditions. Detection can be as simple as reading the fluorescence of ethidium bromide–stained amplification fragments without electrophoresis or as sophisticated as automated detection of simultaneous fluorescent dye–labeled amplification products.

2.3. *CR detection using fluorescence resonance energy transfer (FRET)*

In the Amplifluor modification of the ARMS (allele refractory mutation system) technique, marketed by Intergen (ITK Diagnostics BV), the mutant and wild-type primers are labeled with different fluorescent dyes, and are used in combination to allow competitive detection. Fluorescence is derived only from amplification products, since the fluorescence of the unincorporated primers is inhibited by a quenching molecule held in proximity by a hairpin structure (Newton, 1989; Wu, 1989; Myakishev, 2001). Amplification correlates directly with fluorescence generated and can be detected either as real time or as an endpoint determination.

Another application of SNP detection is utilized by the Invader reaction, marketed by Third Wave Technologies, Inc. This technology uses two short DNA probes that hybridize with the target at the site of the variation or mutation. The structure that forms from a mismatched hybridization is recognized and released into solution by a Cleavase® enzyme, which cuts DNA at single/double strand junctions. The released DNA fragment is detected by real-time fluorescent resonance energy transfer procedures. In the absence of any sequence variation, the enzyme fails to cleave and no fluorescence is generated.

2.4. *Ligase chain reaction*

This technique detects single-base substitutions at specified base sequences in a targeted DNA sequence. A ligase enzyme links two adjacent oligonucleotides that

hybridize with the target sequence. If a mismatch exists at the site where the two oligonucleotides meet, the ligation of the adjacent oligonucleotides fails, and this sequence will not amplify in subsequent cycles. Oligonucleotides hybridizing with the wild-type sequence will be joined and continue to participate in subsequent cycles. Detection of the SNP mutation, therefore, depends on designing oligos specific for the expected base changes (Barany, 1991).

2.5. Real-time PCR

Real-time PCR is a relatively new procedure, which adapts the sensitivity of fluorescent detection methods to single nucleotide polymorphism (SNP) detection. This method utilizes a system of two different labeled probes that are designed to bind and release a fluorescent emission on the basis of single-base differences. The two probes, designated donor and acceptors, bind adjacent to the site of the mutation or SNP and release fluorescent signals only when there is a perfect hybridization match. Real-time PCR offers several technical advantages over other SNP detection procedures. Since it is based on hybridization of two different probes, there is increased detection sensitivity. In addition, the probe melting temperatures can be determined, allowing mutations to be characterized on the basis of their sequence differences. Finally, the sensitivity of real-time PCR allows more accurate quantitative determination than traditional PCR, allowing applications involving gene expression and viral load measurement.

2.6. Dye-labeled oligonucleotide ligation or template-directed dye-terminator incorporation (TDI) assays

This method is based on SNP detection using the real-time generation of fluorescent signals but incorporates a ligation reaction. Two alternate upstream primers and a universal downstream primer are annealed with the template. The SNP allows the hybridization of only the matching upstream primer, and the ligase connects the upstream and downstream primers to form a new molecule having both fluorescent labels. Specific real-time signal detection is based on fluorescence resonance energy transfer (FRET) (Chen and Kwok, 1997).

The TaqMan assay was developed by (Holland, 1991) and is marketed by Applied Biosystems. This procedure is a Real-time PCR adaptation that uses the 5′-exonuclease activity of the Taq polymerase to detect single nucleotide sequence variations. A DNA probe labeled with donor and fluorescent dyes anneals specifically to the accumulating amplification product, but only when sequence-specific amplification occurs. Exonuclease cleavage of the probe dissociates the donor and quenching acceptor dyes and causes an exponential increase of the fluorescence during amplification, which can be measured spectrophotometrically. Because fluorescence only increases when particular sequences are generated by PCR, this

technique is especially useful in diagnostic applications, such as the detection of pathogens and disease-associated genes (Livak, 1995).

2.7. DNA chips, genome chips, and microarrays

Several novel applications variously termed DNA chips, gene or genome chips, and microarrays are expected to have a major impact on genome research (*see* Article 90, **Microarrays: an overview**, Volume 4 and Article 91, **Creating and hybridizing spotted DNA arrays**, Volume 4). These are different variations of orderly arranged samples of DNA or oligonucleotides that allow automation of gene expression and gene mutation–screening applications. These arrays are assembled using robotics and have DNA or oligonucleotide samples arranged on a nylon or glass backing. Each individual sample is extremely small (less than 200 µm for microarrays and greater than 300 µm for macroarrays), which allows simultaneous screening of a large number of genome loci based on the DNA hybridization principle using imaging technology. Microarrays generally refer to longer (500–5000 bp) DNA sequences that were developed by Stanford University (Ekins and Chu, 1999). Arrays with shorter oligonucleotide probes (20–80 bp) are generally termed DNA-, gene-, or genome chips, with the GeneChip designation a trademark registered by Affymetrix, Inc. Microarrays and gene chips have numerous applications including gene discovery, disease diagnosis, pharmacogenetics, and toxicogenomics (Shi, 1998).

3. Mutation screening methods

Several methods have been described that allow rapid screening of multiple exons or gene regions for mutations. Many of these utilize the characteristics of DNA heteroduplexes, which are formed by the annealing of a wild-type sequence with one that has one or more base-pair changes (presumably pathogenic). In the case of a dominant mutation, where both normal and a mutated sequence are expected, the samples need to be denatured and allowed to reanneal. Both of the native homoduplexes would be expected to reform, but a significant amount of heteroduplex formation containing one or more base mismatches are also formed and are the basis of these methods. In recessive disorders, where both copies of a gene or segment may be identical, these detection methods require that specimens to be analyzed are mixed with control DNA samples.

3.1. Denaturing gradient gel electrophoresis (DGGE)

This method can be used to rapidly screen for mutations in PCR-amplified exons suspected to have a mutation. This method uses the different denaturation properties of mismatched heteroduplex and homoduplex DNA fragments in carefully defined temperature and denaturant conditions. These melting differences cause them to migrate differently in polyacrylamide gels. Although this method is useful to screen

a large number of exons for suspected mutations, the detection rate is estimated to be approximately 70–95%, since not all base-pair changes are identified. The specific base-pair change must be identified by DNA sequencing and discriminated from nonpathogenic DNA polymorphisms (Myers *et al*., 1985).

3.2. Denaturing detergent gradient gel electrophoresis (DDGE)

DDGE is based on the different melting properties of heteroduplex and homoduplex DNA sequences, but uses detergent gradients to denature them. Single- versus double-stranded segments have different electrophoretic migration easily detected on acrylamide gels (Sheffield, 1989).

3.3. Single-stranded conformational polymorphism (SSCP)

SSCP is a mutation screening procedure that is based on secondary and tertiary structure differences that occur between single-strand DNA segments that have one or more base-pair changes. DNA segments are denatured, rapidly cooled, and resolved on nondenaturing polyacrylamide gels. Differently migrating SSCP alleles suggest a base-pair change that must be identified by DNA sequence analysis (Orita, 1989; Antolin, 1996).

3.4. Denaturing high-pressure liquid chromatography (DHPLC)

This method is proving to be a valuable tool used for identifying DNA sequence variations based on slight differences in their duplex melting temperatures. High-pressure liquid chromatography has had numerous applications in medicine, but when ion-paired reversed-phase columns were developed by Peter Oefner and Christian Huber, their application expanded to the detection of DNA variations. Mutations in over 350 different disease-associated genes have been analyzed using this method. At present, the DHPLC technology is available in two different formats. In the WAVE system marketed by Transgenomic, Inc., the columns are packed with alkylkated poly(styrene-divinylbenzene) particles (Bonn, 1996). Varian offers the Helix DHPLC system that uses columns containing 1000-Å alkylated silica. Methods can be developed for screening unknown mutations or detecting known SNP sites. The sensitivity of mutation detection protocols has been reported over 90%, and can depend on the estimated melting temperature parameters (Ravnik-Glavac, 2002).

3.5. Cleavase fragment length polymorphism (CFLP)

This method is a mutation screening method based on sequence-dependant secondary structure variations, as in SSCP. DNA fragments are amplified, denatured by heating, and cooled to a predetermined temperature. However, unlike SSCP, these sequences are resolved by using an endonuclease, Cleavase I, to make specific cuts between single- and double-stranded regions. Following electrophoresis, any

changes in the specific bonding pattern suggests a base change that must be identified by DNA sequence analysis (Brow, 1996).

4. DNA sequencing methods

In a diagnostic laboratory, the mutation screening methods described above are useful tools to help identify small DNA alterations; however, the gold standard remains in identifying specific alterations by DNA sequence analysis (*see* Article 4, **Sequencing templates – shotgun clone isolation versus amplification approaches**, Volume 3). A number of different DNA sequencing methods based on the di-deoxy termination chemistry of Sanger (1977) are used, as well as several that use alternate methods.

4.1. Polyacrylamide-based sequencing systems

Several DNA sequencing systems utilize polyacrylamide gel electrophoresis to separate DNA fragments. LI-COR, Inc, offers a DNA sequencing system that uses an infrared detection system to detect two different dye-labeled dideoxy nucleotides, which allows relatively long reads and simultaneous bidirectional sequencing. Visible Genetics, Inc. has developed multiple systems, including MicroGene Blaster™, the MicroGene Clipper™, the Long-Read Tower™, and the OpenGene DNA Sequencing System, which uses disposable MicroCel cassettes with different read lengths. Their TRUGENE HIV-1 genotyping kit including the OpenGene DNA Sequencing System was the first DNA sequencing–based test to receive FDA approval.

4.2. Sequencing sytems based on capillary electrophoresis

Several other manufacturers offer systems that incorporate capillaries containing renewable polymers to separate the DNA sequencing fragments. Amersham Biosciences markets several DNA sequence analysis platforms that are intended for either medium-throughput usage (ALFexpress™, a single dye system that uses polyacrylamide gels) or high-throughput usage (MegaBACE™, which provides for simultaneous use of up to 348 capillaries). Beckman Coulter offers the CEQ 8800 Genetic Analysis System, which is a capillary-based system that features linear polyacrylamide gels, coated capillaries, and cyanine-based infrared dyes that also allows fragment analysis, SNP detection, and amplified fragment length polymorphism (AFLP) fingerprinting. Applied Biosystems has several DNA sequencers that utilize 4 or 5 color dye–based sequencing with capillary electrophoresis. They vary in capability from the single capillary ABI Prism 310 Genetic Analyzer, the 4- or 16-capillary ABI Prism 3100-Avant and 3100

Genetic Analyzers, to the Applied Biosystems 3730 DNA Analyzer, with up to 96 capillaries.

4.3. *Novel methodologies*

Other corporations have developed novel alternatives to the Sanger di-deoxy terminator–based chemistry. The Pyrosequencing™ technology, marked by Biotage AB, is based on a series of enzymatic reactions, whereby the nucleotides are detected as they are sequentially added to a DNA template. A sulfurylase reaction detects the release of pyrophosphate, converting luciferin to oxyluciferin, and a luciferase catalyzes the release of light signals with each addition of nucleotides. The DNA sequence is determined by the pattern and intensity of light signals generated as nucleotides are added to the template. One report claims 96 sequence analyses within 20–30 min (Alderborn, 2000). Several other alternate DNA sequencing methods have been in development; however, they have not come into common usage in a DNA diagnositic laboratory. These include matrix-assisted laser desorption/ionization DNA sequencing (Fitzgerald, 1993) and DNA sequencing by oligonucleotide hybridization (Dramanac, 1993).

5. Current approaches for specific disorders

5.1. *Disorders caused by trinucleotide repeat expansion*

Over a dozen disorders have been identified that are caused by expansions of trinucleotide repeat tracts. The trinucleotide repeat varies with the specific disorder, although $(CAG)_n$ is most common, and occurs most often in the 5′-UTR or the first exon of the expanded gene. Interestingly, many of these disorders are late-onset neuropathies that are progressive and are inherited as autosomal dominant conditions. Since the mutation in these disorders is strictly a size abnormality, laboratory detection of these disorders is accomplished by PCR amplification of the specific gene region containing the repeat followed by accurate measurement of the fragment size generated. Polyacrylamide gel or capillary electrophoresis is commonly used in order to maximize sizing accuracy, with Southern blotting used only to confirm the presence of exceptionally large gene expansions, and DNA sequencing is used to identify mutations in nonexpansion cases. In several disorders, fragile X syndrome, myotonic dystrophy, and spinocerebellar ataxia type X, the extent of the gene expansion is substantially more than the others (hundreds of trinucleotide repeats in affected individuals), and a Southern blot strategy is usually employed in order to detect and size the expanded alleles. The former disorders are associated with a total loss of gene function associated with failure of gene transcription, whereas those disorders with more subtle gene expansions are associated with a gain of abnormal deleterious function. For instance, those disorders with CAG expansions are transcribed into polyglutamine runs, which migrate to the nucleus of affected cells and form insoluable protein deposits that cause damage to the peripheral neurons.

5.1.1. Intermediate size alleles

The length of the trinucleotide repeat is extremely polymorphic in normal individuals as well as in affected persons. This size variability is most likely associated with replication errors of these repetitive tracts during evolution. For many of these conditions, there is a significant gap between the normal size alleles range and those that are pathogenic, which is termed the intermediate size range or premutation range. The accurate detection of the exact repeat length is often of importance since there are different risks associated with different sizes of intermediate alleles in most of these conditions. Intermediate size gene expansions are known to be unstable, and the risk of expansion into pathogenic alleles varies with the size of the premutation. In addition, there may be mild symptoms associated with intermediate size genes, as in Huntington disease. The American College of Medical Genetics has issued a policy statement with specific recommendations for Fragile X Syndrome: Fragile X Syndrome: Diagnostic and Carrier Testing (July 22, 1994) and Technical Standards and Guidelines for Fragile X (Maddalena, 2001). They recommend diagnostic testing to include accurate determination of the size of the smaller normal and premutation alleles (e.g., polyacrylamide electrophoresis), as well as a second method (Southern blotting) targeted for detection of the large expansions.

5.1.2. Genetic anticipation

Many of the trinucleotide expansion disorders demonstrate extremely variable clinical expressivity. This is termed genetic anticipation when symptoms appear to become dramatically more severe within a single generation of a family. Genetic anticipation is more often associated with maternal transmission of the affected gene in some disorders (Fragile X syndrome, myotonic dystrophy) and with paternal transmission of the affected gene expansion in others (Huntington disease, most forms of spinocerebellar ataxia). These changes are often associated with dramatic changes in the length of the trinucleotide repeat, and laboratory testing of extended family members are often useful for accurate risk prediction and genetic counseling. An example is myotonic dystrophy, whose severity varies dramatically with the extent of expansion. Adults can have mild symptoms and are undiagnosed until a dramatic gene expansion occurs in their congenitally affected child. DNA testing measures the length of the repeated CTG pattern in the 3′-untranslated region of the *DMPK* gene. Normal individuals have from 5 to 30 repeat copies; mildly affected persons have from 50 to 80; and congenital onset forms have been described with 2000 or more copies. A combination of methods that allow both accurate sizing of the normal and smaller affected alleles (PCR and polyacrylamide gels) as well as sensitive detection of large expansions (Southern blotting with probe hybridization detection) is the recommended protocol.

5.2. Prader–Willi and Angelman syndromes: cytogenetics versus molecular analysis

Prader–Willi syndrome (PWS) is characterized by severe hypotonia and feeding difficulties in early infancy, followed by uncontrolled appetite (hyperphagia) and

severe obesity in childhood. All patients have mild to moderate mental retardation and, often, behavioral problems. Patients have hypogonadism and infertility. Angelman syndrome (AS) presents in early childhood with hypotonia followed by motor and intellectual retardation. Affected children are ataxic, epileptic, have absence of speech, and an unusual facies characterized by a large mandible and an open-mouthed expression. Patients demonstrate excessive laughter, an occipital groove, abnormal choroidal pigmentation, and characteristic electroencephalogram (EEG) discharges. PWS and AS are caused by a variety of etiologies, including specific 4-megabase deletions of chromosome 15q11-q12, uniparental disomy of chromosome 15, chromosomal translocations, imprinting center mutations, or specific gene mutations (Table 1). Their different clinical presentations are explained by the absence of specific imprinted genes in this chromosomal region that are either paternal in origin (PWS) or maternal (AS).

The presence of imprinted genes, which are either active or inactive depending on their parent of origin, is unusual on an autosome, although examples of other genes on chromosomes 7, 11, 14, and others exist (*see* Article 49, **Gene mapping, imprinting, and epigenetics**, Volume 1). However, when the normal biallelic inheritance of these genes is disrupted by chromosomal deletions or the inheritance of both chromosomal copies from only one parent (uniparental disomy), then the lack of one set of imprinted genes causes a clinical phenotype: PWS, when the absent genes are paternal in origin, or AS when they originated from the mother. DNA methylation analysis of the *SNRPN* gene in the PWS/AS critical region is performed as the preliminary screen in addition to a routine karyotype. If the methylation test is positive, the diagnosis is confirmed by testing for a deletion by fluorescent in situ hybridization (FISH) and for uniparental disomy (UPD) by microsatellite analysis. Specimens from the parents of the proband are required to establish either isodisomy (same chromosome from one parent that is present in duplicate) or heterodisomy (different chromosomes inherited from the same parent).

The American Society of Human Genetics and the American College of Medical Genetics has issued a position statement recommending the testing protocol for suspected PWS or AS patients (American Society of Human Genetics/American College of Medical Genetics Test and Transfer Committee, 1996). There are two strategies that differ somewhat depending on the specific circumstances of the patient. If the symptoms are not clearly suggestive of PWS or AS, then they recommend cytogenetic testing to identify the chromosome 15 deletion or another cytogenetic abnormality. If these results are normal, then molecular genetic testing to detect methylation or uniparental disomy is indicated. However, if the patient fits the diagnosis of PWS or AS more closely, then molecular methods could be used as the initial screen. If the methylation results are positive, then FISH analysis (for a deletion), microsatellite analysis (for unipatental disomy), or imprinting mutations are done subsequently. The type of abnormality is crucial for accurate risk assessment and genetic counseling. Although both deletion-type PWS and AS and those caused by uniparental disomy are de novo and there is a low recurrence risk, imprinting center mutations have a high recurrence risk.

Table 1 Current approaches to molecular genetic diagnosis of selected syndromes

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Achondrogenesis type 1B	SLC26A2 5q32-q33.1	Sulfate transporter	Common mutations detect 60%: R279W, IVS1+2T>C, delV340, R178X; Other base changes: 30%	Sequence analysis	Autosomal recessive inheritance
Albright hereditary osteodystrophy, pseudohypopara-thyroidism, IA	GNAS 20q13.2	Guanine nucleotide binding protein G(S), alpha subunit	Point mutations, deletions, and insertions	Sequence analysis	Inheritance is autosomal dominant but sex-influenced
Alport syndrome	COL4A5 Xq22.3	Collagen alpha 5(IV) chain	Base-pair substitutions in glycine codons: > 60%	Sequence analysis	80% of Alport sydrome are X-linked
	COL4A3 2q36-q37	Collagen alpha 3(IV) chain	Assorted base-pair mutations	Sequence analysis	15% of Alport sydrome are autosomal recessive
	COL4A4 2q36-q37	Collagen alpha 4(IV) chain	Unknown	Sequence analysis	5% of Alport sydrome are autosomal dominant
Alzheimer disease, early onset familial	PSEN1, 14q24.3	Presenilin 1	PSEN1 mutations account for 30–70%	Sequence analysis	Assorted base-pair changes: 6% with no family history
	APP, 21q21	Amyloid beta A4 protein	APP mutations account for 10–15%	Sequence analysis	Assorted base-pair changes
	PSEN2, 1q31-q42	Presenilin 2	PSEN2 mutations account for <5%	Sequence analysis	Assorted base-pair changes
Androgen insensitivity syndrome	AR xq11-q12	Androgen receptor	Point mutations, deletions, and insertions	Sequence analysis	Complete AIS: 95% Partial AIS: ~50% Mild AIS: unknown
Angelman syndrome	UBE3A 15q11.2-q13	Ubiquitin-protein ligase E3A	Methylation abnormality: 78%	Southern blot methylation-sensitive PCR	Deletions and UPD are usually de novo mutations, with a low recurrence risk. UBE3A and imprinting center defects are inherited as sex-limited autosomal dominant
			Deletion of 3–5 Mb: 68%	FISH	

			Uniparental disomy of chromosome 15: ~7%	Microsatellite analysis	
			UBE3A point mutations: ~11%	Sequence analysis	
			Imprinting defects: 3%	Sequence analysis	
Aniridia, isolated	PAX6, 11p13	Paired boxprotein Pax-6	Mutations found in 90% of familial aniridia cases	Sequence analysis	Autosomal dominant inheritance
Atelosteogenesis type 2	SLC26A2 5q32-q33.1	Sulfate transporter	Common mutations detect 60%: R279W, IVS1+2T>C, delV340, R178X, Other base changes: 30%	Sequence analysis	Autosomal recessive inheritance
Bardet–Biedl syndrome (BBS)	BBS1, 11q13 BBS2, 16q21 BBS4, 15q22 MKKS, 20p12 BBS7, 4q27	Bardet–Biedl syndrome proteins 1, 2, 4, 7, McKusick–Kaufman Sydrome	Assorted base-pair changes: common mutation in BBS1M390R is present in 18–46%	Mutation scanning, sequencing	Autosomal recessive inheritance
Becker muscular dystrophy	DMD Xp21.1	Dystrophin	Intergenic deletions (rare) intragenic deletions: 85%	Southern blot, Multiplex PCR, FISH	Most deletions preserve the reading frame; mutations are missense substitutions
			Intragenic duplications: 6%	Southern blot, Multiplex PCR, FISH	
			Point mutations: 10%	Sequence analysis	
Beckwith–Wiedemann syndrome	KCNQ1OT1, 11p15	Potassium channel antisense transcript	Methylation abnormalities in KCNQ1OT1: 50%; H19: 2–7%	Methylation analysis	Uniparental disomy of chromosome 11 in 10–20% of cases
	CDKN1C 11p15	Cyclin-dependent kinase inhibitor	Mutations in CDKN1C in 5–10% of simplex and 40% of familial cases	Sequence analysis	
Berardinelli–Seip congenital lipodystrophy	AGPAT2 9q34.3	1-acyl-sn-glycerol-3-phosphate acyltransferase beta	Assorted base-pair changes: 39% in AGPAT2	Mutation scanning, sequence analysis	Autosomal recessive inheritance
	BSCL2 11q13	Seipin	Assorted base-pair changes: 55% in BSCL2		

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Biotinidase deficiency	BTD 3p25	Biotinidase	Assorted base-pair changes: G98:d7i3, Q456H, R538C, D444H, and D444H/A171T	Sequence analysis	Partial biotinidase deficiencies are compound heterozygous with D444H
Branchiootorenal syndrome	EYA1 8q13.3	Eyes absent homolog 1	Assorted base-pair changes detected in 40% of cases	Sequence analysis	Autosomal dominant inheritance
Breast cancer	BRCA1 17q21	Breast cancer type 1 susceptibility protein	Assorted base-pair changes detected in *BRCA1* or *BRCA2* genes in 63% of cases	Mutation scanning, sequence analysis	Autosomal dominant inheritance
	BRCA2 13q12.3	Breast cancer type 2 susceptibility protein			
Childhood ataxia with central nervous system hypomyelination	EIF2B1 12	Translation initiation factor eIF-2B alpha subunit	Assorted base-pair changes: detected in 1% of cases	Sequence analysis	
	EIF2B2 14q24	Beta subunit	Assorted base-pair changes: detected in 14% of cases		
	EIF2B3	Gamma subunit	Assorted base-pair changes: detected in 5% of cases		
	EIF2B4, 2p23.3	Delta subunit	Assorted base-pair changes: detected in 10% of cases		
	EIF2B5 3q27	Epsilon subunit	Assorted base-pair changes: detected in 70% of cases		
Charcot-Marie-Tooth disease, CMT1A	PMP22 17p11.2	Peripheral myelin protein 22	PMP22 gene duplication: 76%	FISH	Autosomal dominant inheritance
			PMP22 gene point mutations: unknown	Sequence Analysis	
Charcot-Marie-Tooth disease, CMT1B	MPZ 1q22	Myelin P0 protein	*MPZ* gene point mutations: 5–10%	Sequence Analysis	Autosomal dominant inheritance
Charcot-Marie-Tooth disease, CMT1D	EGR2 10q21.1-q22.1	Early growth response protein 2	*EGR2* gene point mutations: unknown	Sequence analysis	Autosomal dominant inheritance
Charcot-Marie-Tooth disease, CMT2E	NEFL 8p21	Neurofilament triplet l protein	*NEFL* gene point mutations: unknown	Sequence analysis	Autosomal dominant inheritance

Charcot-Marie-Tooth disease, CMT4A	GDAP1 8q13-q21.1	Ganglioside-induced differentiation protein 1	*GDAP1* gene point mutations: unknown	Sequence analysis	Autosomal recessive inheritance
Charcot-Marie-Tooth disease, CMT4E	EGR2 10q21.1-q22.1	Early growth response protein 2	*EGR2* gene point mutations: unknown	Sequence analysis	Autosomal recessive inheritance
Charcot-Marie-Tooth disease, CMT4F	PRX 19q13.1-q13.2	Periaxin	*PRX* gene point mutations: unknown	Sequence analysis	Autosomal recessive inheritance
Charcot-Marie-Tooth disease, CMTX	GJB1 Xq13.1	Gap junction beta-1 protein (connexin 32)	*GJB1* gene point mutations: 90% of CMTX	Sequence analysis	X-linked dominant inheritance
Choroideremia	CHM Xq21.2	Rab proteins geranyl-geranyltransferase component A 1	Assorted base-pair changes: 60–95%; Common mutation: exon C, insT in the Finnish population	Sequence analysis	X-linked recessive inheritance
Cockayne syndrome, Xeroderma Pigmentosum	CKN1 Chr 5	Cockayne syndrome WD-repeat protein CSA	25% of patients have CKN1 mutations or deletions	Sequence analysis, deletion detection	Autosomal recessive Inheritance
	ERCC6 10q11	Excision repair protein ERCC-6	75% of patients have ERCC6 mutations or deletions		
Coffin–Lowry syndrome	RPS6KA3 Xp22.2-p22.1	Ribosomal protein S6 kinase alpha 3	Base-pair changes: 60–70% by protein truncation; 35–40% by sequence analysis	Mutation scanning; protein truncation	X-linked dominant Inheritance
Congenital adrenal hyperplasia	CYP21A2 6p21	Cytochrome P450 XXIB	Base-pair changes: Nine common substitutions and deletions: 90–95%	Sequence analysis	Autosomal recessive Inheritance
Congenital muscular dystrophy	LAMA2 6q22-q23	Laminin alpha 2	Point mutations: 50%	Sequence analysis	Autosomal recessive Inheritance
Congenital muscular dystrophy	ITGA7 12q13	Integrin alpha-7	Point mutations: ∼3%	Sequence analysis	Autosomal recessive Inheritance
Congenital muscular dystrophy	FKRP 19q13.3	Fukutin-related protein	A retrotransposal insertion in the 3′UTR: 87%	Southern blot	Autosomal recessive Inheritance

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Congenital myasthenic syndrome	CHRNA1 2q24-q32	Acetylcholine receptor protein, alpha chain	Assorted mutations found in 50–70% of patients. One common mutation found in 50% of European patients: 1267delG	Sequence analysis	Autosomal recessive inheritance; occasionally autosomal dominant
	CHRNB1 17p12-p11	Acetylcholine receptor protein, beta chain			
	CHRND 2q33-q34	Acetylcholine receptor protein, delta chain			
	CHRNE 17p13-p12	Acetylcholine receptor protein, epsilon chain			
	RAPSN 11p11.2-p11.1	43 kDa receptor-associated protein of the synapse	Assorted mutations found in 20–25% of patients		
	COLQ 3p24.2	Acetylcholinesterase collagenic tail peptide	Assorted mutations found in 15% of patients		
	CHAT 10q11.2	Choline *O*-acetyltransferase	Assorted mutations found in 5% of patients		
Cystic fibrosis	CFTR 7q31.2	Cystic fibrosis transmembrane conductance regulator	900 assorted base-pair changes are known. Common mutation: Delta F508: 66% (in Caucasians)	Assorted multiplex PCR techniques	Most common autosomal recessive disorder in Caucasians
Cystinosis	CTNS 17p13	Cystinosin	Assorted base-pair changes; common mutations: 57 kb deletion exon 1-10; W138X	Mutation scanning, sequencing	Autosomal recessive
Deafness-dystonia-optic neuronopathy syndrome	TIMM8A Xq22	Mitochondrial import inner membrane translocase subunit TIM8 A	TIMM8A deletions, missense, and nonsense mutations at unknown rates.	Sequence and deletion analysis	X-linked recessive inheritance

Dentatorubral – pallidoluysian atrophy	DRPLA 12p13.31	Atrophin-1 related protein	Trinucleotide (CAG) repeat expansion	PCR, PAGE	CAG expansion ranges: Normal: < 35 Affected: 48–93
DiGeorge syndrome	DiGeorge critical region, [UFD1L] 22q11.3	Contiguous genes, [Ubiquitin fusion degradation 1–like]	Deletions found in >95%; Mutations in UFD1L found in <5%	FISH, Sequence analysis	93% of deletions are denovo; 7% are inherited as autosomal dominant
Duchenne muscular dystrophy	DMD Xp21.1	Dystrophin	Intergenic deletions (rare); intragenic deletions: 65%	Southern blot Multiplex PCR, FISH	Intergenic deletions may have retinitis pigmentosa, chronic granulomatous disease, and McLeod red cell phenotype or glycerol kinase deficiency and adrenal hypoplasia
			Intragenic duplications: 6%	Southern blot Multiplex PCR, FISH	
			Point mutations: 30%	Sequence analysis	
Dystonia	DYT1 9q34	Torsin A	Codon 310 GAG deletion: 100%	Sequence analysis	Autosomal dominant inheritance
Episodic ataxia, (EA1)	KCNA1,12p13	Voltage-gated potassium channel protein Kv1.1			Autosomal dominant inheritance
Episodic ataxia, (EA2)	CACNA1A 19p13	Voltage-dependent P/Q-type calcium channel alpha-1A subunit			
Episodic ataxia, (EA2)	CACNB4 2q22-q23	Dihydropyridine sensitive L-type, calcium channel beta-4 subunit			Autosomal dominant inheritance
Facioscapulohumeral muscular dystrophy (FSHD)	Unidentified 4q35	Unknown	D4Z4 repeat region deletion: 95–100%	PFGE with restriction digestion	Repeat size ranges: Normal: 42 kb Grey zone: 35–41 kb Affected: <35 kb
Factor V Leiden thrombophilia	F5 1q23	Coagulation factor V	G1691A (R806G): 100%	PCR, restriction digestion	Population freq: 3–8%

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
					Patients with deep vein thrombosis: 15–20%
Familial adenomatous polyposis coli, Gardner syndrome	APC 5q21-q22	Adenomatous polyposis coli protein	Over 800 base-pair changes: Most cause premature protein termination	Protein truncation, sequence analysis	Autosomal dominant inheritance
Fragile X syndrome	FMR1 Xq27.3	Fragile X mental retardation 1 protein (FMRP)	Trinucleotide (CGG) repeat expansion	PCR, PAGE, Southern blot for large expansions	Expansion ranges: Normal: 4–44 Intermediate: 45–58 Premutation: 59–330 Affected: over 200 to several thousand
Friedreichs ataxia	FRDA 9q13	Frataxin	Trinucleotide (GAA) repeat expansion	Long-range PCR, Southern blot	Expansion ranges: Normal: 5–33 Premutation: 34–60 Affected: 66–1700
Fukuyama congenital muscular dystrophy	FCMG 9q31	Fukutin	A 3-kb insertion in the 3′ noncoding region: 87%	Sequence analysis	
Hemochromatosis	HFE 6p21.3	Hereditary hemochromatosis protein	Specific base-pair substitutions: C282Y and H63D	PCR and restriction digestion	C282Y homozygotes: 85%; C282Y/H63D compound heterozygotes: <10% Others: <12%
Hereditary hearing loss, nonsyndromic, autosomal dominant (DNFA)	GJB2 13q11-q12	Gap junction beta-2 protein (connexin 26)	Assorted base-pair changes: 100%; common mutations are W44C and C202F		DNFA mutation are heterozygous
	GJB6 13q12	Gap junction beta-6 protein (connexin 30)	Assorted base-pair changes: unknown		
	COCH 14q12-q13	Cochlin			
Hereditary hearing loss, nonsyndromic, autosomal recessive (DNFB)	GJB2 13q11-q12	Gap junction beta-2 protein (connexin 26)	Common mutations: 35delG; 167delT; 235delC Large 342-kb deletions including part of GJB6: 2%	Sequence analysis; deletion analysis	DFNB mutations are homozygous or compound heterozygous

	GJB6 13q12	Gap junction beta-6 protein (connexin 30)	Large 342-kb deletions including part of GJB2: 2%		
	SLC26A4 7q31	Pendrin			
Hereditary hearing loss, nonsyndromic, mitochondrial	MTRNR1	Mitochondrial 12S rRNA			Mitochondrial (maternal) inheritance
	MTTS1	Mitochondrial tRNA serine 1			
Hereditary hearing loss, nonsyndromic, X-linked	POU3F4 [BRN4] Xq21.1	Pou domain, class 3, transcription factor 4			X-linked inheritance
Hereditary multiple exostoses	EXT1, 8q24	Exostosin-1	Assorted base changes detected in >70%	Sequence analysis	Autosomal dominant inheritance
	EXT2 11p12-p11	Exostosis-2			
Hereditary neuropathy with liability to pressure palsies (HNPP)	PMP22 17p11.2	Peripheral myelin protein 22	*PMP22* gene deletions	FISH	Autosomal dominant inheritance
Holoprosencephaly, nonsyndromic	SHH 7q36	Sonic hedgehog protein	Mutations detected: 30–40%	Sequence analysis	Autosomal dominant inheritance
	ZIC2 13q32	Zinc-finger protein ZIC2	Mutations detected: 5%		
	SIX 2p21	Homeobox protein SIX3	Mutations detected: 1.3%		
	TGIF 18p11.3	5′-TG-3′ interacting factor	Mutations detected: 1.3%		
Huntington disease	IT-15 4p13	Huntington	Trinucleotide (CAG) repeat expansion in >99%	PCR, PAGE	Expansion ranges: Normal: 9–25 Mutable: 26–34 Intermediate: 35–38 Affected: 39–120
Jervell and Lange–Nielsen syndrome	KCNQ1 11p15.5	I_{Ks} K+ channel alpha subunit	Assorted base-pair substitutions: >90%	Mutation scanning, sequence analysis	Autosomal recessive inheritance

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
	KCNE1 21q22.1-q22.2	I_{Ks} K+ channel beta subunit	Assorted base-pair substitutions: <10%		
Juvenile polyposis syndrome	MADH4 (SMAD4) 18q21.1	Mothers against decapentaplegic homolog 4	Assorted MADH4 mutations: 20%	Sequence analysis	Autosomal dominant inheritance
	BMPR1A 10q22.3	Bone morphogenetic protein receptor type IA	Assorted BMPR1A mutations: 23%		
Leber hereditary optic neuropathy (LHON)	mtDNA genes: MTND1, MTND4, MTND5, MTND6	mtDNA NADH-ubiquinone oxidoreductase chain 1, 4, 5, 6	Specific mtDNA base-pair substitutions: G11778A, T14484C, G3460A: 95% of cases	Sequence analysis	Mitochondrial (maternal) inheritance; genotype/phenotype correlations: G11778A is most severe, G3460A is intermediate, and T14484C is mildest.
Leigh syndrome, mtDNA-associated	MTATP6 gene of mtDNA	ATP synthase A6	Base-pair substitutions: MTATP6: T8993G or T8993C: 10–20%	Sequence analysis	Mitochondrial (maternal) inheritance; other genes involved: MTTL1, MTTK, MTND1, MTND3, MTND4, MTND5, MTND6, MTCO3, MTTW, MTTV
Li–Fraumeni syndrome	TP53 17p13.1	Cellular tumor antigen P53	Assorted alterations in P53 gene detected in ~95%; most are missense mutations in exons 5–8.	Sequence analysis, DNA chip detection	Autosomal dominant inheritance
	CHEK2 22q12.1	Serine/threonine-protein kinase Chk2			
Limb–Girdle muscular dystrophy, autosomal dominant	TTID 5q31	Myotilin	Assorted base-pair changes	Sequence analysis	Autosomal dominant inheritance
	LMNA 1q21.2	Lamin A/C			
	CAV3 3p25	Caveolin-3			

	COL6A1, 21q22.3	Collagen alpha 1(VI) chain			
	COL6A2 21q22.3	Collagen alpha 2(VI) chain			
	COL6A3 2q37	Collagen alpha 3(VI) chain			
Limb–Girdle muscular dystrophy, autosomal recessive	SGCA 17q12-q21.3	Alpha-sarcoglycan	Assorted base-pair changes	Sequence analysis	Autosomal recessive inheritance
	SGCB 4q12	Beta-sarcoglycan			
	SGCG 13q12	Gamma-sarcoglycan			
	SGCD 5q33	Delta-sarcoglycan			
	CAPN3 15q15	Calpain 3			
	DYSF 2p13	Dysterlin			
	TCAP 17q12	Telethonin			
	TRIM32 9q31-q34.1	Zinc-finger protein HT2A	D487N homozygous mutation: 100%		
	FKRP 19q13	Fukutin-related protein			
Long QT syndrome, Romano–Ward syndrome	KCNQ1 11p15.5	I_{Ks} K+ channel alpha subunit	LQT1, Assorted base-pair substitutions: 55–60%	Sequence analysis	Autosomal dominant: other genes involved: LQT4 (unknown), LQT5 (KCNE1), LQT6 (MiRP1/KCNE2), LQT7 (unknown)
	HERG 7q35-q36	I_{Ks} K+ channel alpha subunit	LQT2, Assorted base-pair substitutions: 35–40%		
	SCN5A 3p21-p24	I_{Na} Na+ channel alpha subunit	LQT3, Assorted base-pair substitutions: 3–5%		
Machado–Joseph disease	MJD 14q24.3-q31	Machado–Joseph disease protein	Trinucleotide (CAG) repeat expansion	PCR, PAGE	Autosomal dominant; expansion ranges: Normal: up to 47 Intermediate: 48–51 Affected: 53–86
Malignant hyperthermia susceptibility	RYR1 19q13.1	Ryanodine receptor type 1	Base-pair changes: 70% Mutation panel detects 25–30%	Sequence analysis	

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Marfan syndrome	FBN1 15q21.1	Fibrillin 1	Point mutations: 30–90% Detection rate depends on strength of diagnostic criteria met	Mutation screening, sequence analysis	Autosomal dominant inheritance
Medium-chain acyl-coenzyme A dehydrogenase deficiency	ACADM 1p31	Medium-chain acyl-CoA dehydrogenase (MCAD)	One common mutation, K304E present in 76%; Assorted base-pair changes: 95%	Sequence analysis	Autosomal recessive inheritance
Megalencephalic leukoencephalopathy with subcortical cysts	MLC1 22qter	Membrane protein MLC1	Assorted base-pair changes detected in 60–70%	Sequence analysis	Autosomal recessive inheritance
MELAS (Mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes)	MTTL1 gene of mtDNA	tRNA $^{Leu(UUR)}$	Base-pair substitutions: MTTL1 A3243G: 80% MTTL1T3271C: ~7.5% MTTL1 A3252G: ~7.5 – 10%	Sequence analysis	Mitochondrial (maternal) inheritance
McCune–Albright syndrome	GNAS 20q13.2	Guanine nucleotide-binding protein G(S), alpha subunit	Point mutations, deletions, and insertions	Sequence analysis	Few familial cases; GNAS mutations are mosaic
Familial medullary carcinoma of the thyroid (FMTC)	RET 10q11.2	RET oncogene	Point mutations in 88% of FMTC; especially, cysteine codons in exons 10, 11, 13–15:	Sequence analysis	Autosomal dominant inheritance
Menkes disease	ATP7A Xq13-q13	Copper-transporting ATPase 1	Point mutations and deletions: 95%	Mutation scanning: Sequence analysis	X-linked recessive inheritance
Methylenetetrahydro-folate reductase thrombophilia	MTHFR 1p36.3	Methylenetetrahydro-folate Reductase	Base-pair substitution: C677T base change causes a heat-labile MTHFR	PCR, restriction digestion	Homozygous C677T is risk factor for thrombophilia
Muenke nonsyndromic coronal craniosynostosis	FGFR3 4p16.3	Fibroblast growth factor receptor 3	Base-pair changes: unknown frequency	Sequence analysis	Inheritance pattern unclear

Mucopolysaccharidosis type 1, Hurler and Scheie syndromes	IDUA 4p16	Alpha-L-iduronidase	Assorted base-pair changes	Sequence analysis	Autosomal recessive inheritance
Multiple endocrine neoplasia, 2A	RET 10q11.2	RET oncogene	Mutations in exon 10 and 11 cysteine codons found in 95%	Sequence analysis	Autosomal dominant inheritance
Multiple endocrine neoplasia, 2B	RET 10q11.2	RET oncogene	Exon 16 M918T mutations found in 95%	Sequence analysis	Autosomal dominant inheritance
Multiple epephyseal dysplasia, dominant	COMP 19p13.1	Cartilage oligomeric matrix protein	COMP mutations account for 25% of MED	Sequence analysis	Autosomal dominant inheritance
	COL9A1 6q13	Collagen alpha 1(IX) chain	COL9A mutations account for <5% of MED		
	COL9A2 1p33-p32.2	Collagen alpha 2(IX) chain			
	COL9A3 20q13.3	Collagen alpha 3(IX) chain			
	MATN3 2p24-p23	Matrilin-3	MATN3 mutations account for 10% of MED		
Multiple epephyseal dysplasia, recessive	SLC26A2 5q32-q33.1	Sulfate transporter	Assorted base-pair changes	Sequence analysis	Autosomal recessive inheritance
Myoclonic epilepsy with ragged red fibers (MERRF)	MTTK gene of mtDNA	$tRNA^{Lys}$	Base-pair substitutions: A8344G: 80% T8356C and G8363A: 10% Others: 10%	Sequence analysis	Mitochondrial (maternal) inheritance
Myotonic dystrophy	DMPK 19q13.2-q13.3	Myotonic dystrophy protein kinase	Trinucleotide (CTG) repeat expansion	PCR, PAGE, Southern blot for large expansions	Allele expansion ranges: Normal: 5–37 Intermediate: 38–49 Mild: 50–150 Classical: 100–1500 Congenital: over 1000
NARP (Neurogenic muscle weakness, ataxia, retinitis pigmentosa)	MTATP6 gene of mtDNA	ATP synthase A6	Base-pair substitutions: MTATP6: T8993G or T8993C: > 50%	Sequence analysis	Other genes involved: MTTL1, MTTK, MTND1, MTND3, MTND4, MTND5, MTND6, MTCO3, MTTW, MTTV

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Neurofibromatosis 1	NF1 17q11	Neurofibromin	Point mutations: ~ 80%	Protein truncation, Sequence analysis	Autosomal dominant inheritance; half of all cases are de novo mutations
			Gene deletions: 5–10%	FISH	
Neurofibromatosis 2	NF2 22q12.1	Merlin (Schwannomin)	Assorted base-pair changes detected in ~65%	Mutation scanning; sequence analysis	Autosomal dominant inheritance; half of all cases are de novo mutations
Noonan syndrome	PTPN11 12q24.1	Protein-tyrosine phosphatase, nonreceptor type 11	Assorted base-pair changes found in 50%	Sequence analysis	Autosomal dominant inheritance
Norrie disease	NDP Xp11.4	Norrin	Base-pair changes: 85% Submicroscopic deletions: 15%	Sequence analysis; Southern blot	X-linked recessive inheritance
Oculocutaneous albinism, type 1 (OCA1A and OCA1B)	TYR 11q14-q21	Tyrosinase	Assorted base-pair changes: OCA1 mutations completely inactivate TYR; OCA1B mutations partially inactivate TYR	Mutation scanning, sequencing	Autosomal recessive inheritance
Oculocutaneous albinism, type 2 (OCA2)	P Gene 15q11.2-q12	P Protein	Assorted base-pair changes: Common mutation: 2.7-kb deletion in Africans; A481T in Japanese; V443I in Europeans	Mutation scanning, sequencing	Autosomal recessive inheritance
Oculopharyngeal muscular dystrophy (OPMD)	PABPN1 14q11.2-q13	Polyadenylate binding protein nuclear 1	Trinucleotide (GCG) repeat expansion	PCR, PAGE	Expansion ranges: Normal: $(GCG)_6$ Dominant: $(GCG)_{8-13}$ Recessive: $(GCG)_7$
Pallister–Hall syndrome	GLI3 7p13	Zinc-finger protein GLI3	Assorted base-pair changes in 50%; common mutations 2023delG and 2012 delG	Sequence analysis	Autosomal dominant inheritance

Peroxisome biogenesis disorders, Zellweger syndrome	PEX1 7q21-q22	Peroxisome biogenesis factor 1	Common PEX1 mutations detect 80%: I700fs and G843D	Sequence analysis	Recessive inheritance; Research testing for PEX6, PEX26, PEX10, PEX12, PXMP3 (PEX2), PEX3, PEX13, PEX16, PXF (PEX19)
Peutz–Jeghers syndrome	STK11 19p13.3	Serine/threonine-protein kinase 11	Assorted base-pair changes: 73% with positive history 65% with negative history	Sequence analysis	Autosomal dominant inheritance
Pendred syndrome	SLC26A4 7q22-q31	Pendrin	Assorted base-pair changes: 75% detected common mutations: L236P, T416P, and 1001 + 1 G-to-A	Sequence analysis	Autosomal recessive inheritance
Phenylalanine hydroxylase deficiency, phenylketonuria	PAH 12q23.2	Phenylalanine hydroxylase	Panels of 4–15 common mutations detect 30–50%; Sequencing entire gene detects 99%	Mutation scanning; sequence analysis	Autosomal recessive inheritance
Polycystic kidney disease, autosomal dominant	PKD1 16p13.3-p13.12	Polycystin	Assorted base-pair changes in PKD1: 85%	Sequence analysis	Autosomal dominant inheritance
	PKD2 4q21-q23	Polycystin 2	Assorted base-pair changes in PKD2: 15%	Sequence analysis	Autosomal dominant inheritance
Polycystic kidney disease, autosomal recessive	PKHD1 6p21		Mutation detection rate: 40–80%	Mutation scanning; sequencing	Autosomal recessive inheritance
Prader–Willi syndrome	SNRPN, others 15q11.2-q13		Methylation abnormality: 99%	Southern blot Methylation-sensitive PCR	Deletions and UPD are usually de novo mutations; imprinting center defects are inherited as sex-limited autosomal dominant.
			Chromosome deletion: 70%	FISH	
			Uniparental disomy of chromosome 15: ~25%	Microsatellite analysis	

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
			Imprinting center defects: less than 5%	Southern blot Methylation-sensitive PCR	
			Chromosomal rearrangements: 1%	Karyotype FISH	
Prothrombin gene thrombophilia	Factor 2	Prothrombin	G20210A in 100%	PCR, restriction digestion	Population incidence: 2%
					In patients with deep vein thrombosis: 6%
PTEN Hamartoma tumor syndrome (PHTS)	PTEN 10q23.31	Dual-specificity phosphatase PTEN	Assorted PTEN mutations: Cowden syndrome: 80% BRRS: 60% Proteus-like syndrome: 50% BRRS: 11% PTEN deletion	Sequence analysis, quantitative PCR	Autosomal dominant inheritance
Retinitis pigmentosa, autosomal dominant (ADRP)	RHO 3q22.1	Rhodopsin	Accounts for 25–30% of ADRP	Sequence analysis	Autosomal dominant inheritance
	RDS 6p21.2	Peripherin 2	Accounts for 5–10% of ADRP		
	RP1 8q12.1	Oxygen-regulated protein 1	Accounts for 5–10% of ADRP		
	IMPDH1 7q32.1	Inosine monophosphate dehydrogenase 1	Accounts for 3–5% of ADRP		
	PIMIK 7p14.3	Pim-1 kinase	Unknown %		
Retinitis pigmentosa, autosomal recessive (ARRP)	RPE65 1p31.2	Retinal pigment epithelium-specific 61 kDA protein	Accounts for 2% of ARRP	Sequence analysis	Autosomal recessive inheritance
	ABCA4 1p22.1	Retinal-specific ATP-binding cassette transporter, retinal recessive	Rare		
	CRB1 1q31.3	Crumbs protein homolog 1	Rare		
	USH2A 1q41	Usherin	Accounts for 4–5% of ARRP		

	MERTK 2q13	C-mer protooncogene receptor tyrosine kinase	Unknown		
	SAG 2q37.1	S-arrestin	Rare		
	RHO 3q22.1	Rhodopsin	Rare		
	PDE6B, 4p16.3	Rod cGMP-specific 3′, 5′-cyclic phosphodiesterase alpha subunit	Accounts for 3–4% of ARRP		
	CNGA1 4p12	cGMP-gated cation channel alpha 1 subunit	Rare		
	LRAT 4q32.1	Lecithin reitnol acyltransferase	Unknown %		
	PDE6A 5q33.1	Rod cGMP-specific 3′, 5′-cyclic phosphodiesterase alpha subunit	Accounts for 3–4% of ARRP		
	TULP1 6p21.3	Tubby-related protein 1	Rare		
	RGR 10q23.1	RPE-retinal G protein-coupled receptor	Accounts for 10–20% ARRP		
	NR2E3 15q23	Nuclear receptor subfamily 2 group E3	Rare		
	RLBP1 15q26.1	Cellular retinaldehyde-binding protein	Unknown %		
	CNGB1 16q13	Rod cGMP-gated channel beta subunit	Unknown %		
Retinitis pigmentosa X-linked (XLRP)	RPGR Xq21.1	Retinitis pigmentosa GTPase regulator	Accounts for 70% or XLRP	Sequence analysis	X-linked dominant or recessive inheritance
	RP2 Xp11.2	XRP2 protein	Accounts for 8% of XLRP		
Retinitis pigmentosa mitochondrial	MTTS2 MtDNA	Mitochondrial sevine tRNA2	Rare	Sequence analysis	Mitochondrial (maternal) inheritance
Retinoblastoma	RB1 13q14.1-q14.2	Retinoblastoma-associated protein	Assorted base-pair changes detected in 70–75%	Sequence analysis	Autosomal dominant inheritance

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
			Submicroscopic deletions detected in 8%	FISH, Southern blot	
Rett syndrome	MECP2 Xq28	Methyl-CpG-binding protein 2	Assorted base-pair changes detected in 80%	Mutation screening, sequence analysis	Affects females only; 99% cases are de novo
Rothmund–Thomson syndrome	RECGL4 8q24.3	ATP-dependent DNA helicase Q4	Assorted base-pair changes detected in ~70%	Sequence analysis	RECGL4 mutations cause absent or truncated proteins
Rubinstein–Taybi syndrome	CREBBP 16p13.3	CREB-binding protein	35% with assorted base-pair changes; Microdeletions detected in 10%	Sequence analysis; FISH	Autosomal dominant inheritance
Smith–Lemli–Opitz syndrome	DHCR7 11q12-q13	7-dehydrocholesterol reductase	Assorted base-pair changes: Common mutations are intron 8 splice acceptor, W151X, and R404C	Sequence analysis	Autosomal recessive inheritance
Spinal and bulbar muscular atrophy (SBMA, Kennedy disease)	AR Xq11-q12	Androgen receptor	$(CAG)_n$ repeat expansions: Normal: up to 34 Intermediate: 35–37 Affected: 38 or more	PCR, PAGE	Autosomal dominant inheritance
Spinal muscular atrophy	SMN1 5q12.2-q13.3	Spinal motor neuron	Homozygous deletions of exons 7 and 8: 95% compound heterozygotes for deletions and point mutations: 5%	PCR and restriction digestion, Sequence analysis	Autosomal recessive inheritance
Spinocerebellar ataxia, type I	SCA1 6p23	Ataxin-1	$(CAG)_n$ repeat expansions: Normal: 6–44 Intermediate: 36–38 Affected: 39–91	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type II	SCA2 12q24	Ataxin-2	$(CAG)_n$ repeat expansions: Normal: up to 31 Affected: 32–200	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type IV	SCA4 16q22.1	Unknown	Unknown		Autosomal dominant inheritance

Spinocerebellar ataxia, type V	SCA5 11p11-q11	Unknown	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type VI	CACNA1A 19p13	Voltage-dependent P/Q-type calcium channel alpha-1A subunit	$(CAG)_n$ repeat expansions: Normal: up to 18 Intermediate: 19 Affected: 20–33	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type VII	SCA7 3p21.1-p12	Ataxin-7	$(CAG)_n$ repeat expansions: Normal: 7–35 Intermediate: 28–35 Affected: 36–300	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type VIII	SCA8 13q21	Unknown	$(CTG)_n$ repeat expansions: Normal: 15–50 Intermediate: 50–70 Affected: 71–800	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type X	SCA10 22q13	Ataxin-10	$(ATTCT)_n$ repeat expansions: Normal: 10–22 Affected: 280–4500	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type XI	SCA11 15q14-q21.3	Unknown	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type XII	SCA12 5q31-33	Phosphatase P2a regulatory subunit	$(CAG)_n$ repeat expansions: Normal: 7–31 Affected: 55–78	PCR, PAGE	Autosomal dominant inheritance
Spinocerebellar ataxia, type XIII	SCA13 19q13.3-q13.4	Unknown	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type XIV	PKC8 19q13.4-qter	Protein kinase C gamma	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type XV	SCA15	Unknown	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type XVI	SCA16 8q22.1-q24.1	Unknown	Unknown		Autosomal dominant inheritance
Spinocerebellar ataxia, type XVII	TBP, SCA17	TATA binding protein	$(CAG)_n$ repeat expansions: Normal: 25–42 Intermediate: 42–44 Affected: 46–63	PCR, PAGE	Autosomal dominant inheritance
Spondyloepiphyseal dysplasia tarda, X-linked	SEDL Xp22.2-p22.1	Sedlin	Assorted base-pair changes detected in 80%	Sequence analysis	X-linked recessive inheritance

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
Stickler syndrome	COL2A1 12q13.11-q13.2	Collagen alpha 1(II) chain	Unknown	Mutation scanning, sequence analysis	Autosomal dominant inheritance
	COL11A1 1p21	Collagen alpha 1(XI) chain	Assorted base-pair changes detected in ~70–80%		
	COL11A2 6p21.3	Collagen alpha 2(XI) chain	Unknown		
Transthyretin amyloidosis	TTR 18q11.2-q12.1	Transthyretin	Assorted base-pair changes detect 99%; common mutation is V30M	Sequence analysis	Autosomal dominant inheritance
Tuberous sclerosis complex	TSC1 9q34	Harartin	Mutations detected in 30% of familial cases	Sequence analysis	Autosomal dominant inheritance
	TSC2 16p13.3	Tubrnin	Mutations detected in 50% of familial cases		
Usher syndrome, type 1	USH1A 14q32	Unknown	2% Detected	Mutation scanning, sequence analysis	Autosomal recessive inheritance
	MYO7A 11q13	Myosin VIIa	60% Detected		
	USH1C 11p15	Harmonin	5% Detected		
	CDH23 10q21	Cadherin-23	10% Detected		
	USH1E 21q21	Unknown	Unknown		
	PCDH15 10q21	Protocadherin	Unknown		
	SANS 17q24	Sans	Unknown		
Usher syndrome, type 2	USH2A 1q41	Usherin	USH2A base-pair changes detected in 80%; Most common: 2299delG	Sequence analysis	Autosomal recessive inheritance
	USH2B 3p24 USH2C 5q USH2D	Unknown	USH2B type is rare; USH2C mutations account for 15%	Sequence analysis	
Usher syndrome, type 3	USH3 3q24	Usher syndrome type 3 protein	Unknown	Mutation analysis	Autosomal recessive inheritance

Von Hippel–Lindau syndrome	VHL 3p26-p25	Von Hippel–Lindau disease tumor suppressor	Assorted base-pair changes detected in >90%; some mutations confer genotype/phenotype correlations	Sequence analysis	Autosomal dominant inheritance; 80% inherited and 20% de novo mutations
Waardenburg syndrome, type 1	PAX3 2q35	Paired box protein Pax-3	Assorted base-pair changes: detected in >90%	Sequence analysis	Autosomal dominant inheritance
Waardenburg syndrome, type 2	MITF 3p14-p12.3	Microphthalmia-associated transcription factor	Assorted base-pair changes: detected in 10–20%	Sequence analysis	Autosomal dominant inheritance
Werner syndrome	WRN 8p12-p11.2	Werner syndrome helicase	Assorted base-pair changes detected in ~90%	Sequence analysis	Autosomal recessive inheritance
Wilms tumor	WT1 11p13	Wilms tumor protein	10–15% of Wilms tumor patients are inherited	Deletions, sequence analysis	A second Wilms tumor locus at 11p15 is suspected
Wilson disease	ATP7B 13q14.3-q21.1	Copper-transporting ATPase 2	Assorted base-pair changes; common mutations: H1069Q, R778L, H714Q, delC2337	Mutation scanning; sequence analysis	Autosomal recessive inheritance
Xeroderma pigmentosum	XPA 9q22.3	DNA-repair protein complementing XP-A cells	25% of patients with XP	Sequence analysis	Autosomal recessive inheritance: Complementation groups not listed (mutation percentage): XPE (rare), XPF (6%), and XPG (6%)
	ERCC3 2q21	TFIIH basal transcription factor complex helicase XPB subunit	Rare		

(*continued overleaf*)

Table 1 (*continued*)

Disorder or syndrome	Gene, map location	Protein product	Mutation type(s): % detected	Diagnostic methods	Inheritance or mutation comments
	XPC 3p25	DNA-repair protein complementing XP-C cells	25% of patients with XP		
	ERCC2 19q13.2-q13.3	FIIH basal transcription factor complex helicase subunit	15% of patients with XP		
	DDB2 11p12-p11	DNA damage binding protein 2	Rare		
XX male syndrome; XY gonadal dysgenesis	SRY Yp11.3	Sex-determining region Y protein	Assorted base-pair changes: detected in 80% of XXMS	Sequence analysis, FISH	XX male syndrome is de novo; XY; if SRY translocated onto an autosome, inheritance is sex-limited dominant

For each syndrome, the associated genes and their chromosomal map location, mutation detection rates and most methods, and inheritance patterns are listed, where known. This listing is not comprehensive due to space limitations, but focuses on the more common syndromes where molecular genetic testing is diagnostically useful (GeneTests, 1993-2004). Mutation-scanning techniques include heteroduplex analysis, SSCP, DGGE, DDGE, DHPLC, or CFLP. Deletions are detected by FISH (fluorescent in situ hybridization) or Southern blotting.

Further reading

Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research*, **17**, 2503–2516.

Myers RM, Maniatis T and Lerman LS (1987) Detection and localization of single base changes by DGGE. *Methods in Enzymology*, **155**, 501–527.

References

Alderborn A, Kristofferson A and Hammerling U (2000) Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Research*, **10**, 1249–1258.

American Society of Human Genetics/American College of Medical Genetics Test and Transfer Committee (1996) Diagnostic testing for Prader-Willi and Angleman syndromes: Report of the ASHG/ACMG Test and Technology Transfer Committee. *American Journal of Human Genetics*, **58**, 1085–1088.

Antolin MF, Bosio CF, Cotton J, Sweeney W, Strand MR and Black WC 4th (1996) Intensive linkage mapping in a wasp (*Bracon hebetor*) and a mosquito (*Aedes aegypti*) with single-strand conformation polymorphism analysis of random amplified polymorphic DNA markers. *Genetics*, **143**, 1727–1738.

Barany F (1991) Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 189–193.

Beckmann JS (1988) Oligonucleotide polymorphisms: A new tool for genomic genetics. *Biotechnology*, **6**, 161–164.

Bonn G, Huber C and Oefner P (1996) *Nucleic Acid Separation on Alkylated Nonporous Polymer Beads*. U.S. Patent No. 5, 585, 236.

Brow MA, Oldenburg MC, Lyamichev V, Heisler LM, Lyamicheva N, Hall JG, Eagan NJ, Olive DM, Smith LM, Fors L, *et al*. (1996) Differentiation of bacterial 16S rRNA genes and intergenic regions and Mycobacterium tuberculosis katG genes by structure-specific endonuclease cleavage. *Journal of Clinical Microbiology*, **34**, 3129–3137.

Chen X and Kwok P-Y (1997) Template-directed dye-terminator incorporation (TDI) assay: A homogeneous DNA diagnostic method based on fluorescence resonance energy transfer. *Nucleic Acids Research*, **25**, 2347–2353.

Dramanac R, Dramanac S, Strzoska Z, Paunesku T, Labat I, Zeremski M, Snoddy J, Funkhouser WK, Koop B, Hood L, *et al*. (1993) DNA sequencing determination by hybridization: a strategy for efficient large-scale sequencing. *Science*, **260**, 1649–1652.

Ekins R and Chu FW (1999) Microarrays: their origins and applications. *Trends in Biotechnology*, **17**, 217–218.

Fitzgerald MC, Zhu L and Smith LM (1993) The analysis of mock DNA sequencing reactions using matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **7**, 895–897.

GeneTests: Medical Genetics Information Resource (database online). (1993-2004) Copyright, University of Washington and Children's Health System, Seattle. Updated weekly. Available at **http://www.genetests.org**. Accessed January, 2004.

Holland PM, Abramson RD, Watson R and Gelfland DH (1991) Detection of specific polymerase chain reaction product by utilizing the $5' \rightarrow 3'$ exonuclease activity of Thermus aquaticus DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 7276–7280.

Livak K, Marmaro J and Todd JA (1995) Towards fully automated genome-wide polymorphism screening. *Nature Genetics*, **9**, 341–342.

Maddalena A, Richards CS, McGinniss MJ, Brothman A, Desnick RJ, Grier RE, Hirsch B, Jacky P, McDowell GA, Popovich B, *et al*. (2001) Technical standards and guidelines for fragile X: the first of a series of disease-specific supplements to the Standards and Guidelines

for Clinical Genetics Laboratories of the American College of Medical Genetics. Quality Assurance Subcommittee of the Laboratory Practice Committee. *Genetics in Medicine*, **3**, 200–205.

Myakishev MV, Khripin Y, Hu S and Hamer DH (2001) High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. *Genome Research*, **11**, 163–169.

Myers RM, Fischer SG, Maniatis T and Lerman LS (1985) Modification of the melting properties of duplex DNA by attachment of a GC-rich DNA sequence as determined by denaturing gradient gel electrophoresis. *Nucleic Acids Research*, **13**, 3111–3129.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC and Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research*, **17**, 2503–2516.

Orita M, Suzuki Y, Sekiya T and Hayashi K (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using polymerase chain reaction. *Genomics*, **5**, 874–879.

Ravnik-Glavac M, Atkinson A, Glavac D and Dean M (2002) DHPLC screening of cystic fibrosis gene mutations. *Human Mutation*, **19**, 374–383.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Sheffield VC, Cox DR, Lerman LS and Myers RM (1989) Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 232–236.

Shi L (1998) *DNA Microarray (Genome Chip) - Monitoring the Genome on a Chip*. On the World Wide Web: **www.Gene-Chips.com**.

Wu DY, Ugozzoli L, Pal BK and Wallace RB (1989) Allele-specific enzymatic amplification of beta-globin genomic DNA for diagnosis of sickle cell anemia. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 2757–2760.

Specialist Review

The clinical and economic implications of pharmacogenomics

David L. Veenstra
University of Washington, Seattle, WA, USA

1. Introduction

The rapid advance of the Human Genome Project and the development of new genetic analysis technologies promise to bring a new era of genomics to medicine (Collins, 1999; Friend, 1999). Among the applications of genetics to medicine, pharmacogenomics will likely be one of the first tangible benefits from the Human Genome Project (Evans and Relling, 1999).

It has been suggested that the use of pharmacogenomics will be widespread and lead to overall decreased health care costs (Hodgson and Marshall, 1998; Kleyn and Vesell, 1998; Marshall, 1997; Persidis, 1998; Regalado, 1999; Sadee, 1999). However, this issue remains unclear. The objective of this article is to present a cost-effectiveness framework for evaluating the incremental clinical and economic benefits of pharmacogenomic-based therapies, apply this framework to noted pharmacogenomic examples, and evaluate the potential impact of pharmacogenomics on research and development in the pharmaceutical industry and the delivery of health care.

2. Economic evaluation of health care technologies

The integration of genetic technologies into medical practice will be a complex process (Holtzman and Marteau, 2000). Such an undertaking encompasses multiple technologies (molecular biology, informatics) and systems (pharmaceutical and health care industries), and society as a whole. Because of the different nature of using genetic information to improve patients' health, there has been increased interest in the broad-ranging impact of genomics – commonly referred to the ethical, legal, and social implications (ELSI). Although the implementation of pharmacogenomics will be guided by ethical and legal issues, it will be driven largely by economic factors – which are determined by a complex interaction of social utilities.

The impact of pharmacogenomics will be determined to a significant extent by incentives for researchers to develop targeted therapies, health care systems, and clinicians to provide them, and patients to accept them. Over the past several decades, economic evaluation in the health care field has evolved to study these questions. Drawing on methods and concepts from economics, clinical epidemiology, psychology, and the decision sciences, the field of "cost-effectiveness" research has synthesized a set of tools and a theoretical framework for evaluating the complex issues in health care (Garber and Phelps, 1997; Weinstein *et al*., 1996).

2.1. Types of economic evaluation

There are a variety of methods used in the economic evaluation of health care technologies: (1) cost-minimization, (2) cost-consequences analysis, cost-benefit analysis, (3) cost-effectiveness analysis, and (4) cost-utility analysis (Table 1). These methods vary primarily in the way drug effectiveness is valued. For example, in cost-minimization, it is assumed there is no difference in drug effectiveness. In both cost-effectiveness and cost-consequences analysis, effectiveness is measured in natural, clinical units such as heart attacks or infections avoided. In cost-benefit analysis, a monetary value is assigned to effectiveness (e.g., a heart attack might be "valued" at $100 000). And in cost-utility analysis, effectiveness is measured

Table 1 Types of economic evaluations in health care

Study design	Costs measured?	Effects measured?	Strengths	Weaknesses
Cost-minimization	Yes	No	• Easy to perform	• Only useful if effectiveness is assumed to be the same
Cost-consequences	Yes	Yes, typically in clinical terms	• Data presented in a straightforward fashion	• A ratio is not calculated, thus making comparisons of health interventions difficult
Cost-benefit	Yes	Yes, in economic terms	• Good theoretical foundation • Can be used within health care and across sectors of the economy	• Less commonly accepted by health care decision makers • Evaluation of benefits methodologically challenging
Cost-effectiveness	Yes	Yes, in clinical terms	• Relevant for clinicians • Easily understandable	• Cannot compare interventions across disease areas
Cost-utility	Yes	Yes, in quality-adjusted life years (QALYs)	• Incorporates quality of life • Comparable across disease areas and interventions	• Requires evaluation of patient preferences • Can be difficult to interpret

in quality-adjusted life years (QALYs), which account for improvements in both life expectancy and quality of life. Another characteristic that differentiates these methods is that cost-minimization and cost-consequences analysis do not involve the calculation of a ratio. Of note, although "cost-effectiveness analysis" is a specific type of economic evaluation (Table 1), the term also refers generally to all types of economic evaluation in health care.

2.2. *Cost-effectiveness calculation*

In any economic evaluation, it is important that the technology being evaluated is compared to current medical practice. Weinstein and Staason (1977) defined the incremental cost-effectiveness ratio (ICER) as

$$ICER = \frac{C2 - C1}{E2 - E1} \qquad (1)$$

where $C2$ and $E2$ are the cost and effectiveness of the new intervention being evaluated and $C1$ and $E1$ are the cost and effectiveness of the standard therapy. The costs and effects that are included in equation (1) depend on the perspective of the analysis. From a societal perspective, indirect costs and effects such as patient time away from work, downstream medical care costs years or decades after the intervention, and the quality of life of the patient and even their family need to be considered. Because of the all-encompassing nature of the societal perspective, it is generally best suited for national health care plans. More relevant to health care plans or providers in the United States is the payer perspective, which addresses primarily the direct medical care costs incurred by the payer (e.g., drug cost, professional fees, hospital stay).

Where are the data for cost-effectiveness calculations obtained? Because there is such diversity in the data requirements, it is unusual to obtain all estimates from one study. In some cases, cost-effectiveness studies can be based on a single randomized clinical trial. However, because clinical trials are usually in a controlled setting, the costs incurred are not representative of utilization in a real-world setting. The "efficacy" observed in a controlled setting can also be distinguished from the "effectiveness" that could be expected in practice. Finally, the time frame of clinical trials for chronic conditions are generally not sufficient to evaluate long-term outcomes. Because of these reasons, modeling techniques such as decision analysis are often used to extrapolate the results from clinical studies using primarily epidemiologic and economic data from other sources (Detsky *et al*., 1997).

2.3. *Cost-effectiveness criteria*

How can cost-effectiveness information guide health policy decisions? The favored approach for formal decision-making rules is to utilize cost-utility analysis because it allows for comparisons across interventions and diseases, accounts for impact on life expectancy and quality of life, and has theoretical foundations in welfare

economics. Medical interventions are considered to be cost-effective when they produce health benefits at a cost comparable to other commonly accepted treatments. A general guide is that interventions that produce one quality-adjusted life year (QALY, equivalent to 1 year of perfect health) for under $50 000 are considered cost-effective, those between $50 000 and $100 000 per QALY are of questionable cost-effectiveness, and above $100 000 per QALY is not considered cost-effective. The cutoff of $50 000 per QALY was derived loosely from the cost of providing dialysis for a patient for one year – a service paid for by Medicare for any US citizen.

2.4. *Strengths and weaknesses*

Cost-minimization has probably played the biggest role in drug-selection decisions in the past because it is the most straightforward approach, and decisions are often driven by drug budget impact. For the most part, this is unfortunate, as the assumption that different drugs have the same effect rarely holds true. Cost-benefit calculations are useful because the results are directly applicable to budget impact assessments, but the process of assigning monetary value to health is both technically and politically challenging. Clinicians are often more accepting of cost-effectiveness analysis because results such as "cost per heart attack avoided" seem intuitive; however, such calculations make evaluation across interventions and disease areas difficult (e.g., is it better to prevent a heart attack or a stroke?). Cost-utility analysis has been advocated because it measures benefit in patient-oriented terms (quality of life) and permits comparison between different interventions by standardizing the denominator. Cost-utility analysis works particularly well in settings with a single health care payer, as the overall objective of such a health care system is to improve society's health in a cost-effective manner. In the United States, however, health care plans are not structured to provide all-encompassing, lifetime health care, and cost-utility analysis does not fit easily within the decision-making framework. For this reason, many cost-effectiveness practitioners and health plans are relying more often on cost-consequences analysis, which presents economic and clinical information in an evidence-based format.

2.5. *Demand for economic information*

The application of cost-effectiveness analysis has increased dramatically in the past decade as a result of increasing health care costs and the desire to deliver the greatest health value for the money. The formal application of cost-effectiveness analysis to drug coverage decisions has its origins in countries with single-payer health care systems (e.g., government sponsored). Recently, multiple countries and health care systems have begun to adopt requirements for pharmacoeconomic information. These requirements formalize an otherwise implicit demand for health care technologies that are cost-effective, and will influence, to a certain extent, "go, no-go" decisions in drug research and development.

The United Kingdom, Canada, and Australia all have formal requirements in place for cost-effectiveness information and programs in place for evaluating cost-effectiveness data – the National Institute for Clinical Excellence (NICE, United Kingdom), the Canadian Coordinating Office for Health Technology Assessment (CCOHTA), and the Pharmaceutical Benefits Advisory Committee (PBAC, Australia) (CCOHTA, 1997; PBAC, 1999; NICE, 1999). To varying degrees, European countries other than the United Kingdom utilize economic analyses in decision making, and several countries such as The Netherlands have indicated that formal requirements will be introduced in the near future (Drummond *et al*., 1999). In Japan, the Ministry of Health and Welfare modified the guidelines for submission of pharmacoeconomic data were revised in September 1994 to include specific data requirements and submission format, although there has not been much response to this directive on the part of the pharmaceutical industry (Hisashige, 1997).

In the United States, cost-effectiveness information is most often used in support of drug formulary listing in managed care settings. Several managed care organizations and pharmacy benefits managers currently have or are considering the inclusion of guidelines that require outcomes and economic information for formulary evaluation (Langley, 1999; Mather *et al*., 1999). In addition, the Academy of Managed Care Pharmacy (AMCP) has recently adopted guidelines for the submission of information, including outcomes and cost-effectiveness data, to support formulary consideration (Fry *et al*., 2003).

2.6. Cost-effectiveness drivers

Although the determination of the incremental costs and effects of a new health care intervention can be complicated, the incremental cost-effectiveness of almost all interventions is usually driven by a few important factors: the cost and efficacy of the intervention, the morbidity, mortality, and prevalence of the disease that is being prevented, and the cost of treating the disease. On the basis of these factors, Veenstra and colleagues developed a cost-effectiveness framework for evaluating pharmacogenomic interventions (Veenstra *et al*., 2000a; Phillips *et al*., 2001a).

3. Cost-effectiveness framework

3.1. Economic costs

The cost of a genetic test will go beyond the procurement cost of the test itself, as with any diagnostic procedure. There will be induced costs, including direct costs such as additional medical care follow-up, and indirect costs such as patient time away from work. These costs are potentially of greater magnitude than the direct cost of purchasing the test. If test results are not available at the point of care, the additional clinical, administrative, and patient time required to respond to the test results may negate any efficiency gained by providing the test. For conditions such as acute infectious processes, a delay in obtaining test results may

have serious clinical consequences. In contrast, for chronic diseases, the availability of test results within a week's time frame may have only a minimal impact on overall treatment costs if additional office visits, and so on, are not required.

However, these induced costs will be offset to a certain extent. One of the benefits of pharmacogenomic testing is that the information can be used throughout the lifetime of the patient. For example, rather than measuring serum drug levels to infer the metabolic capability of a patient every time a novel drug is introduced to a patient, an assay to identify variations in the genes that encode drug metabolizing enzymes could be used throughout the lifetime of the patient, and for a variety of medications. The induced costs with pharmacogenomic testing also will likely be less than those associated with genetic testing for disease risk.

Finally, a cost that must be considered is the impact on patients of knowing information about their genetic makeup. Knowledge about genetic variations may lead to anxiety and a decreased quality of life in some patients, in addition to behavior such as avoiding all drug therapies. In contrast, patients without any major genetic variations may adopt a careless attitude with regard to drug compliance and consumption. Although initial work in the area of breast cancer (Lerman *et al*., 1996) suggests that patients benefit from knowledge of their genetic status, there is a critical need for additional studies in this area.

3.2. Effectiveness of genetic tests

As with data from controlled clinical trials, it is important to distinguish the "efficacy" of a genetic test from its "effectiveness". The efficacy of a test can be viewed as the diagnostic ability of the assay – that is, the ability of the test to accurately detect the genetic variation it was designed to identify. Diagnostic test performance is typically evaluated on the basis of sensitivity and specificity, or receiver-operator curves (ROCs). Because genetic tests have high sensitivity and specificity when direct sequencing or restriction site assays (>90%), they are often viewed as being highly accurate. But in a cost-effectiveness or clinical perspective, it is the prognostic significance of the test result that is important (its effectiveness).

The prognostic significance of a test is determined by the degree of association between the identified genetic variation and its physical manifestation(s). The association between genotype and phenotype, known as *gene penetrance*, will drive both clinical and economic outcomes. For example, if half of all patients with a gene variant experience a severe side effect from a drug (gene penetrance of 50%), avoiding the use of that drug in all patients with the variant would unnecessarily deprive the other half of the patients (the "false positives") of medication. The issue of "false-positives" will be important for almost all applications of pharmacogenomics, and the consequence of labeling patients as having a genetic variation despite the fact that not all of them will have clinically relevant effects must be considered. Genes with high penetrance will be better candidates for cost-effective pharmacogenomic strategies. Note that the term "false positives" does not refer to patients that were falsely identified as having a variant gene, but patients *with* a variant gene that do not express the clinical phenotype.

3.3. Outcomes: clinical and economic

Providing individualized drug therapy is an inherently appealing concept, and has driven much of the excitement about pharmacogenomics. The clinical and economic benefits of doing so, however, must be considered. In the case of pharmacokinetic strategies, avoiding adverse drug reactions (ADRs) will be the most likely benefit, as serious ADRs have been associated with significant morbidity, mortality, and economic costs. A recent analysis by Phillips *et al*. (2001b) suggests that genetic variations in drug metabolizing enzymes are associated with serious ADRs.

Testing costs for pharmacodynamic strategies, on the other hand, will be offset by avoiding unnecessary drug expenditures for patients who are unlikely to respond, or by providing beneficial treatment to patients that would otherwise not have been treated. Thus, pharmacodynamic-based strategies will likely be more cost-effective for expensive or chronic medications.

3.4. Drug monitoring and individualization

The *incremental* cost-effectiveness of pharmacogenomics will depend on the current ability to accurately monitor patients for toxic effects or drug response and individualize their therapy accordingly. Plasma drug levels are often used to monitor toxic drugs, while surrogate markers such as blood pressure for hypertension, lipid levels for hypercholesteremia, and HbA1c for diabetes are used to measure drug response for chronic diseases. When there are readily available, inexpensive, and validated means of monitoring drug response, pharmacogenomics may offer little incremental benefit.

3.5. Gene prevalence

Finally, the cost-effectiveness of any preventative screening strategies, such as pharmacogenomics, is highly dependent on the underlying prevalence of disease. In the case of pharmacogenomics, it is the frequency of the variant allele in the population being tested that will be a critical factor. For example, if the frequency of a variant allele is 0.5%, on average only one patient with a variant allele would be detected for every 200 patients tested. The importance of gene variant prevalence is highlighted in several of the examples below.

3.6. Summary of criteria

The considerations above provide a set of "cost-effectiveness criteria" for evaluating the potential cost-effectiveness of pharmacogenomics: (1) genetic factors – gene prevalence and gene penetrance; (2) test factors – sensitivity, specificity, and cost; (3) disease factors – severity of disease outcomes (clinical and economic); and (4) treatment factors – current ability to monitor drug response, cost of current therapy (Table 2). Prior to conducting a formal cost-effectiveness analysis,

Table 2 Factors that influence the cost-effectiveness of pharmacogenomic strategies

	Factors to assess	Features that favor cost-effectiveness
Gene	Prevalence	• Variant allele is relatively common.
	Penetrance	• Gene penetrance is high.
Test	Sensitivity; specificity; cost	• High specificity and sensitivity. • A rapid and relatively inexpensive assay is available.
Disease	Prevalence	• High disease prevalence in the population.
	Outcomes and economic impacts	• High untreated mortality. • Significant impact on quality of life. • High costs of disease management using conventional methods.
Treatment	Outcomes and economic impacts	• Reduction in adverse effects that significantly impact quality of life or survival. • Significant improvement in quality of life or survival due to differential treatment effects. • Monitoring of drug response is currently not practiced or difficult. • No or limited incremental cost of treatment with pharmacogenomic strategy.

Adapted from Flowers C and Veenstra DL (2000) Will pharmacogenomics in oncology be cost-effective? *Oncology Economics*, **1**, 26–33

these criteria can be useful indicators as to which interventions warrant a full cost-effectiveness analysis. These criteria also can assist scientists in designing basic research strategies that will be more likely to result in clinically useful and economically viable improvements in patient care. Below we review several examples of pharmacogenomics and evaluate their potential cost-effectiveness.

4. Examples

4.1. Pharmacodynamic-based strategies

4.1.1. Hyperlipidemia

A recent study reported the association of a variant allele of cholesteryl ester transferase protein (CETP), which is involved in cholesterol metabolism, with clinical response to pravastatin (Kuivenhoven *et al*., 1998). Drug response as measured by coronary vessel intraluminal diameter was correlated with CETP genotype, but not with lipid levels, suggesting that drug response may be predictable based on CETP genotype but not the typically used surrogate marker, lipid levels. This finding, if verified in subsequent long-term studies, could have a significant impact on the management of hyperlipidemia given that the prevalence of the non-responder genotype is 16%. Although the outcome of administering pravastatin to a nonresponder in the short term may simply be hyperlipidemia for a month or two,

traditional monitoring methods would not detect the lack of response, and in the long term the patient would be a increased risk of outcomes that are clinically severe and expensive (e.g., myocardial infarction), and expenditures on antihyperlipidemia agents would be significant as patients are often on therapy for decades.

4.1.2. HIV

Highly Active Anti-Retroviral Therapy (HAART) has resulted in dramatic reduction in HIV-related morbidity and mortality, but only 46% of patients are able to reach <50 copies/mL with 48 weeks of HAART (Hirsch *et al*., 2000). Most patients eventually "fail" initial therapy because of problems with drug adherence, drug toxicities, or virological resistance to HIV protease and reverse transcriptase inhibitors.

The evaluation of viral genotype appears to be clinically useful for individualizing HIV treatment cocktails in patients who have failed an initial regimen of HAART (secondary resistance). The Community Programs for Clinical Research on AIDS (CPCRA) trial 046 showed that among patients who failed HAART, viral suppression <500 copies/mL) with a subsequent regimen was significantly more common for patients randomized to receive genetic testing (34%) than for patients treated on the basis of clinical judgment alone (22%) (Baxter *et al*., 2000). The AntiretroVIRal ADAPTtation (VIRADAPT) trial reported similar results (Durant *et al*., 1999). These studies can be viewed as a "gold-standard" for the clinical evaluation of the effectiveness of pharmacogenomic testing strategies because patients were randomized to testing versus no testing, although long-term outcomes were not evaluated.

Several of the factors associated with HIV secondary resistance testing would suggest that it may be cost-effective: clinically severe and expensive outcomes for patients who fail HAART, identifying the drug(s) responsible for poor response is difficult, a genetic test is commercially available (approximately US$ 355), and the effectiveness of genotyping has been established in a pragmatic randomized trial. However, with regard to primary testing, the prevalence of resistance in treatment naive patients is fairly low, 1–10%, suggesting that primary resistance testing may be cost-effective only in higher-risk populations.

Weinstein *et al*. (2001) recently evaluated the cost-effectiveness of both primary and secondary HIV resistance testing using cost-effectiveness and decision-analysis techniques to model life expectancy, quality of life, and lifetime costs. The authors estimated that secondary resistance testing was cost-effective compared to no testing ($17 900 per QALY gained). The cost-effectiveness of primary testing was highly dependent on the prevalence of resistance in a treatment naive population: $22 300 per QALY at 20% prevalence and $69 000 per QALY at 4% prevalence. The results support current guidelines for the use of secondary testing, but additional studies are needed to evaluate the effectiveness of primary testing and elucidate the prevalence of HIV resistance in different settings.

4.2. Pharmacokinetic-based strategies

Many of the examples of pharmacodynamic-based strategies are preliminary and implementation in clinical practice may be years to decades in the future. In

contrast, there has been extensive research on the genetic variation of enzymes involved in drug metabolism. Many of the first applications of pharmacogenomics will likely be in this area because of the extensive basic research conducted over the past several decades.

Adverse drug reactions are a significant cause of morbidity and mortality. Although many adverse drug reactions are considered "nonpreventable", pharmacogenomics may help to prevent these reactions. In a recent study, Phillips *et al*. (2003) evaluated the potential role of pharmacogenomics in reducing the incidence of adverse drug reactions. Using a systematic review of ADR studies and reviews on polymorphisms of drug metabolizing enzymes, the authors identified 27 drugs frequently cited in adverse drug reaction studies. Among these drugs, 59% were metabolized by at least one enzyme with a variant allele known to cause poor metabolism. Conversely, only 7–22% of randomly selected drugs is known to be metabolized by enzymes with this genetic variability ($p < .01$). These results suggest that genetic variability in drug metabolizing enzymes plays an important role in adverse drug reactions. Below, we review the potential cost-effectiveness of preventing ADRs from one of the drugs included in their study, the anticoagulant warfarin.

4.2.1. Warfarin

The anticoagulant warfarin exhibits large variability in drug response, primarily because of disease, diet, and drug interactions. However, part of the variability has been attributed to polymorphisms of the enzyme that metabolizes warfarin, the cytochrome P450 enzyme CYP2C9 (Rettie *et al*., 1994). Individuals who are deficient in CYP2C9 activity may be at a higher risk for severe bleeding episodes and require lower starting doses or more frequent monitoring. The use of genetic information may thus assist clinicians in initiating and monitoring warfarin dosing.

The prevalence of heterozygotes is relatively high, approximately 30%, but patients with a null genotype are rare (<1%). In addition, serious bleeding events are rare in patients followed in anticoagulation clinics because warfarin therapy is closely monitored and individualized. Genetic testing will have to facilitate this process in a cost-effective manner. Initial studies suggested an association between CYP2C9 genotype and bleeding risk (Aithal *et al*., 1999). However, a subsequent study failed to detect an association between CYP2C9 genotype and anticoagulation variability (Taube *et al*., 2000). Additional epidemiologic studies are needed before assessing the appropriateness of CYP2C9 genetic analysis for patients on chronic warfarin therapy.

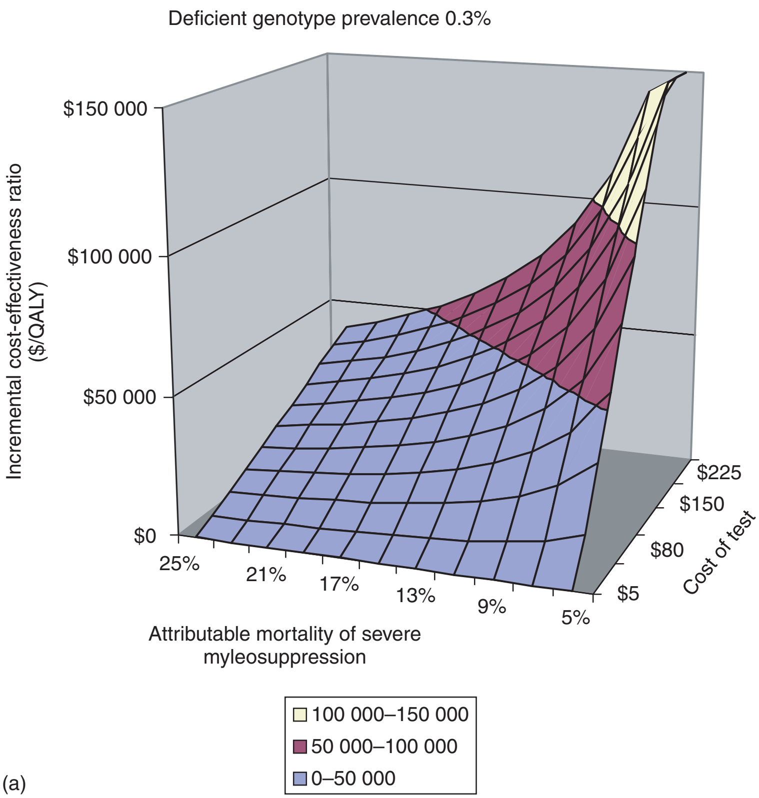
4.2.2. Childhood leukemia

Because of their narrow therapeutic index, oncolytics will also be prime candidates for pharmacogenomics (Iyer and Ratain, 1998). An often cited example is dosage individualization for 6-mercaptopurine (6-MP), which is used for treatment of acute lymphoblastic leukemia (ALL) in children (Krynetski and Evans, 1999). Thiopurine *S*-methyltransferase (TPMT) is responsible for the inactivation of 6-MP, and TPMT

deficiency is associated with severe hematopoietic toxicity when deficient patients are treated with standard doses of 6-MP.

Because the implications of overdosing 6-MP are serious, and because of the significant costs involved in treatment of ALL, testing of children to establish their TPMT genotype may be cost-effective. Veenstra *et al*. (2000b) developed a decision analytic model to evaluate the potential cost-effectiveness of genotyping children before administering 6-MP. The authors assumed that patients not dying from myleosuppression had a quality-adjusted life expectancy of 10 years (10 QALYs), the cost of treating myleosuppression was \$5000, and the probability of severe myleosuppression for a patient deficient in TPMT was 90% without testing and 10% with testing. Three parameters representative of the dimensions that affect the cost-effectiveness of pharmacogenomics were varied: (1) *economic* (cost of test, \$5 to \$250), (2) *genetic* (allele frequency, 0.3, 0.5, and 1.0%), and (3) *clinical* (mortality of myleosuppression, 5–25%).

The importance of gene prevalence is immediately apparent (Figure 1). At a null allele genotype frequency of 1.0%, the incremental cost-effectiveness of genetic



(a)

Figure 1 (a) Influence of cost of genetic test and severity of clinical outcome on the hypothetical cost-effectiveness of TPMT genotyping with a deficient genotype prevalence of 0.3%. (b) Influence of cost of genetic test and severity of clinical outcome on the hypothetical cost-effectiveness of TPMT genotyping with a deficient genotype prevalence of 1.0%

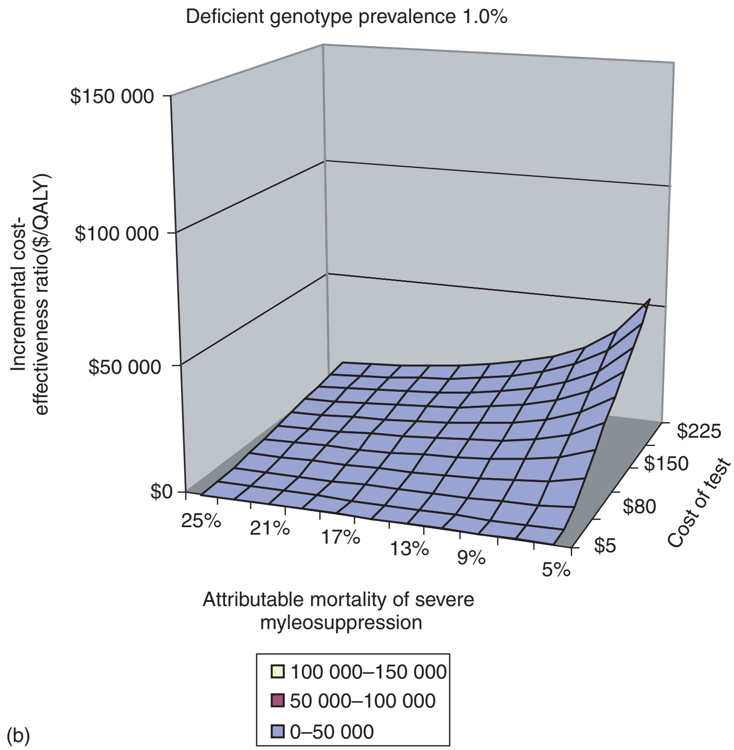


Figure 1 (*continued*)

testing falls below the commonly cited $50 000/QALY cutoff for essentially all of the parameter combinations tested. But for a frequency of 0.3%, the actual frequency of the null allele genotype for TPMT, the cost-effectiveness of genetic testing is not clear. In this scenario, the cost of the test has a significant impact on the cost-effectiveness ratio because approximately 300 children must be tested, on average, in order to identify the one that is TPMT deficient. However, with a high attributable mortality of severe myleosuppression (e.g., >20%), genetic testing is cost-effective for all scenarios.

5. Implications for the health care industry

5.1. Research areas

5.1.1. Adverse drug reactions

Drugs with a narrow therapeutic index will be some of the top candidates for individualization using pharmacogenomics. Drugs such as warfarin immediately come to mind, as well as oncolytics such as 6-MP. However, the widespread application of pharmacogenomics to prevent ADRs in the majority of currently marketed

drugs is less clear. Almost all drugs on the market have been developed for use in the general population. Drugs that have caused serious ADRs (e.g., seldane) – potentially as a result of genetic variation in drub metabolizing enzymes – have typically been removed from market. Thus, the incidence of serious ADRs related to genetic variation in drug metabolizing enzymes, although not known, may be low. Pharmacogenomics will thus be more cost-effective for secondary testing in patients that have had idiosynchratic ADRs as opposed to primary testing in the general population.

Where pharmacogenomics will play a more important role is in the development and approval of new drugs. The economic and clinical loss to society when a drug fails in the drug development stage because of rare, serious ADRs is significant. Through genetic profiling of the DMEs of patients enrolled in clinical trials, it will be possible to identify, to a certain extent, patients at higher risk for ADRs, and to achieve regulatory approval for the majority of patients who are able to metabolize the drug efficiently. Use of the drug in patients with poor metabolism would not necessarily be contraindicated, as appropriate dosage adjustments could be made.

5.1.2. Chronic diseases

The cost-effective application of pharmacodynamic-based strategies to chronic diseases such as hypertension, diabetes, and hypercholesteremia will depend on one critical aspect – the validity of the surrogate markers that are commonly used to measure disease progression and drug response (Temple, 1999). For example, pharmacogenomics may be useful for the treatment of hypertension. Clinicians are able to adjust dosing and modify drug selection on the basis of blood pressure response and side effects over the span of several office visits. However, recent studies suggest that blood pressure control may not be a good surrogate marker for clinical outcomes (Pahor *et al*., 2000; ALLHAT, 2000). In the ALLHAT study, treatment with doxazosin was terminated early because of a significantly higher incidence of combined cardiovascular disease events in patients on doxazosin compared to the diuretic chlorthalidone, even though higher systolic blood pressure was only 3 mmHg higher in the doxazosin group.

The outlook for diabetes, in contrast, is not as clear. Several studies have established the correlation between HbA1c levels and the complications of diabetes such as neuropathy, nephropathy, and retinopathy. However, with the classes of agents being introduced that work through a variety of mechanisms, it may be more difficult to rely on HbA1c as a surrogate marker in the future. Pharmacogenomics also may be cost-effective for disease states such as asthma, where outcomes are acute and expensive (e.g., emergency room visits), and attaining control of symptoms can be a trial and error process, and clinicians now have a multitude of drug therapies to choose from (Liggett, 1998; Israel *et al*., 2000). As more therapeutic alternatives become available for many disease states, pharmacogenomics will become increasingly important to assist in drug selection.

Drugs for which it is difficult to measure response in a relatively short time frame (e.g., 1 month) are also good candidates for individualization (Rioux, 2000). For example, patients can require up to 6 weeks of treatment before responding to antidepressant therapy (Jonsson *et al*., 1998). Another example is Alzheimer's

Disease – it is difficult to measure clinical response, and if no improvement or slowing of progression is seen over 6 months to 1 year, valuable time may be lost if other medications are available (Farlow *et al*., 1996; Poirier, 1999).

5.1.3. Nonhost genomics

The application of genomic analysis to infectious agents and tumor cell lines offers a useful "bridge" from traditional means of individualizing drug therapy to pharmacogenomics. Although the viral genome is analyzed rather than the patient's, many of the issues will be similar. Antibacterial sensitivity analysis is routinely used throughout the world to guide antibiotic drug selection. As genetic testing for HIV resistance exemplifies, analysis of viral genomes will likely become more commonplace and analogous to antibiotic sensitivity testing.

The use of genetic testing to identify viral genotype in the treatment of hepatitis C with interferon and ribivirin (combination therapy) provides an excellent case study in pharmacodynamic-based testing strategies. Patients with viral genotype 1 respond significantly better to 48 weeks of treatment versus 24 weeks of treatment, while patients with nongenotype 1 respond similarly to 24 or 48 weeks of therapy (McHutchinson *et al*., 1998). Several studies suggest that it is cost-effective to perform such testing. Younossi *et al*. (1999) reported that genotyping was the most cost-effective strategy among a variety of treatment options. Veenstra *et al*. (2000b) estimated that genotyping added 0.33 QALYs per patient compared to empiric 6-month therapy, at an incremental cost-effectiveness ratio of $3500/QALY. Thus, pharmacogenomics results in treating some patients for an extended duration, but in a cost-effective manner.

Pharmacogenomics may not play as significant a role in the analysis of bacterial genomes, as current phenotypic tests are readily available. However, a genomic analysis that would allow clinicians to prescribe appropriate therapy at the point of care would be an immense improvement over current practices of empirical antibiotic drug therapy.

Flowers and Veenstra (2000) reviewed the potential cost-effectiveness of pharmacogenomics in oncology. As noted above, there is a real and current need for individualized drug dosing in oncology. In addition, evaluation of tumor cell line genetics will become increasingly important as predictors of drug response are identified.

5.1.4. Minority populations

Another consideration is the desire to provide appropriate drug therapy to minority populations. To understand the impact of pharmacogenomics on minority populations, it will be necessary to understand two issues. First, will minority populations benefit clinically from pharmacogenomics? A recent study reported that enalapril is associated with a significant reduction in the risk of hospitalization for heart failure among white patients with left ventricular dysfunction, but not among similar black patients (Exner *et al*., 2001). Another consideration is the safety of pharmaceuticals. Minorities have not always been well represented in clinical trials, and thus may be at higher risk for ADRs because of

differences in genetic makeup, particularly with regard to the drug-metabolizing enzymes. If future studies indicate that pharmacogenomics can benefit minority populations, cost-effectiveness considerations may guide whether such strategies are pursued.

5.2. Incentives for research investment

What are the implications of pharmacogenomics for the pharmaceutical and biotech industries? The promise of new drug targets, decreased development time and cost, and quicker FDA approval all suggest that the integration of genomics with drug therapy will offer multiple opportunities for investment in research and development. But a key issue must be faced: will decreased market sizes with individualized drugs be offset by better market penetration and added value?

If we examine the interaction between CETP genotype and pravastatin, we can see how economic incentives might drive the development of targeted drug therapies. Suppose a company wished to develop and market a drug specifically for patients with the $\beta2\beta2$ genotype. Assume the captured market share is all patients with the $\beta2\beta2$ genotype, the effectiveness of the drug compared to nontargeted therapy is the inverse of the prevalence of the $\beta2\beta2$ genotype (i.e., if 50% of the population are non-responders and are eliminated, the drug's effectiveness will essentially be doubled), and the annual cost of standard therapy is $1000.

We examine three scenarios: (1) drug price remains unchanged regardless of effectiveness; (2) drug price is a function of effectiveness, so that the incremental cost-effectiveness ratio is constant; and (3) drug price is some unknown function of effectiveness (and gene prevalence). In the first scenario, total revenue is a linear function of gene prevalence (Figure 2a). In this instance, the only incentive for the development of a pharmacogenomic test is if the gene prevalence is higher than the company's current market share. In the second scenario, there is neither incentive nor disincentive for the development of individualized therapy – because companies can demand higher prices for better drugs, the decrease in market share is offset (Figure 2b) However, neither fixed drug prices nor drug prices that are a solely a function of cost-effectiveness are realistic. More likely is a scenario in which high prices can be charged for very effective medications that target a select few patients, but not sufficiently to generate the equivalent revenue from a larger market share (Figure 2c). In addition, as genotype prevalence approaches 100%, any gain in revenue from drug price would likely be lost in the extra development cost. Thus, we are left with a situation in which the most likely targets for pharmacogenomics are genetic variants that are neither too rare (e.g., $<10\%$) nor too common (e.g., $>90\%$) (Lindpainter, 1999).

Society's willingness to pay for individualized drug therapy potentially will be one of the most important factors driving or inhibiting the transition of pharmacogenomics from the laboratory to clinical practice (Lichter and Kurth, 1997; Danzon and Towse, 2002). In one of the first articles to quantitatively examine the potential impact of pharmacogenomics on health care payers and the pharmaceutical industry, Danzon and Towse concluded that the increased social gain from targeted therapy will need to be reflected in the value of pharmacogenomic-based drugs, and that

for drugs targeted to small populations, regulatory incentives such as the Orphan Drug Act may be needed to encourage investment in research and development.

Another potential benefit of pharmacogenomics to the industry is the opportunity for drug salvage – that is, obtaining approval for drugs that have been pulled from the market because of adverse events. Danzon and Towse examined the potential salvage of the drug Centoxin. Centoxin was developed to treat gram-negative bacteremia sepsis, which comprise one-third of all sepsis cases. However, Centoxin can cause harm to patients with non-gram-negative bacteremia and was withdrawn from the market. The authors calculated that if a (hypothetical) test had been developed to identify patients who would benefit from Centoxin, the test, surprisingly, could have cost $11 000 and still been cost-effective.



Figure 2 Effect of genotype prevalence on revenue generated for three pricing assumptions: fixed pricing, cost-effective pricing, and "real-world" pricing

5.3. Health care systems

Health care systems will also face significant challenges incorporating pharmacogenomics into health care delivery. In the short term, pharmacogenomics likely will lead to increased drug budgets because of higher drug prices and increased drug utilization (driven in part by patient demand), as well as increased diagnostic test costs. Because of these challenges, health care plans and payers should be prepared to evaluate the clinical and economic value of pharmacogenomic testing in a systematic and evidence-based fashion. Pharmacogenomics will necessitate improved clinician and patient education, more complex drug formulary structures, and advanced information systems at the point of care and drug dispensing. Despite these challenges, pharmacogenomics will offer health plans a mechanism for improving appropriate drug utilization, long-term improvement in their patients' health, and economic cost offsets in the future.

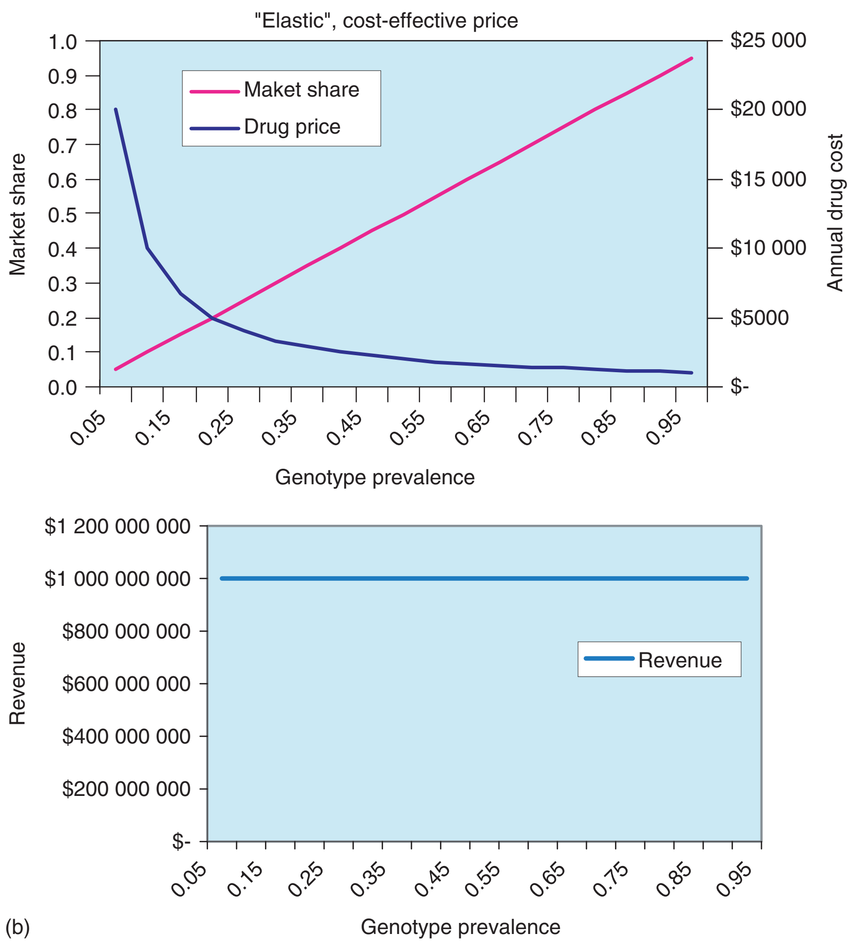


Figure 2 (*continued*)

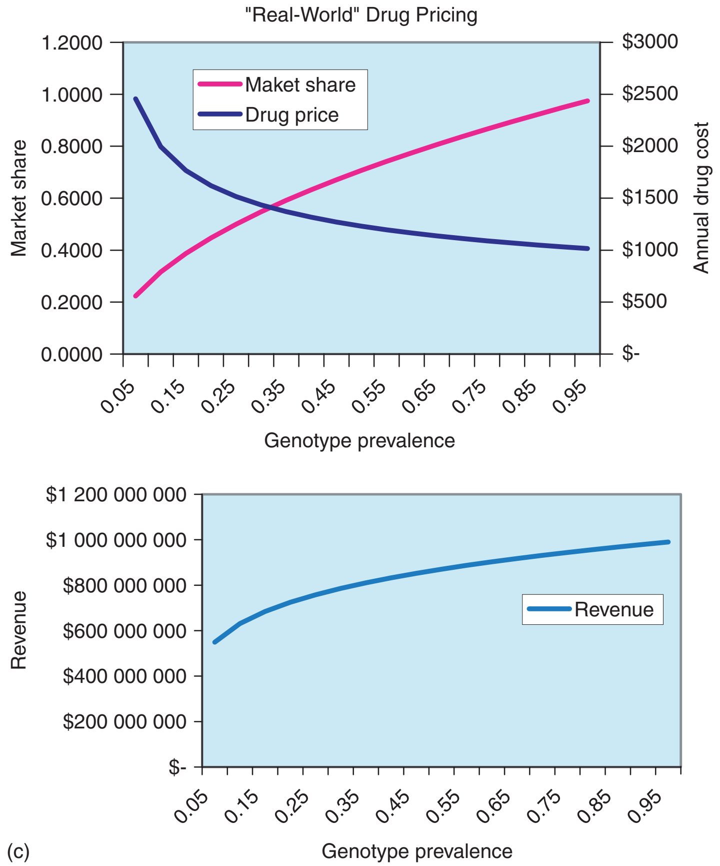


Figure 2 (*continued*)

6. Conclusion

The use of pharmacogenomics to individualize drug therapy will offer clinical and economic benefits. However, these benefits must be weighed against the additional cost of genotyping all patients in order to adjust therapy in a few. Our analysis suggests that pharmacogenomics will likely be cost-effective only for certain combinations of disease, drug, gene, and test characteristics, and that the cost-effectiveness of pharmacogenomic-based therapies needs to be evaluated on a case-by-case basis. Cost-effectiveness analysis can help scientists, clinicians, and decision makers invest in technologies and develop drug utilization guidelines that will benefit patients in an economically viable fashion.

Further reading

Richard F, Helbecque N, Neuman E, Guez D, Levy R and Amouyel P (1997) APOE genotyping and response to drug treatment in Alzheimer's disease. *Lancet*, **349**, 539.

Shak S and the Herceptin Multinational Investigator Study Group (1999) Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. *Seminars in Oncology*, **26**(4 Suppl 12), 71–77.

References

Aithal GP, Day CP, Kesteven PJ and Daly AK (1999) Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. *Lancet*, **353**, 717–719.

ALLHAT Collaborative Research Group (2000) Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: The antihypertensive and lipid–lowering treatment to prevent heart attack trial (ALLHAT). *The Journal of the American Medical Association*, **283**, 1967–1975.

Baxter JD, Mayers DL, Wentworth DN, Neaton JD, Hoover ML, Winters MA, Mannheimer SB, Thompson MA, Abrams DI, Brizz BJ, *et al*. (2000) A randomized study of antiretroviral management based on plasma genotypic antiretroviral resistance testing in patients failing therapy. CPCRA 046 Study Team for the Terry Beirn Community Programs for Clinical Research on AIDS. *AIDS*, **14**(9), F83–F93.

CCOHTA. Canadian Coordinating Office for Health Technology Assessment (1997) *Guidelines for Economic Evaluations of Pharmaceuticals*, *Second Edition*, Canadian Co-ordinating Office for Health Technology Assessment: Canada.

Collins FS (1999) Genetics: An explosion of knowledge is transforming clinical practice. *Geriatrics*, **54**, 41–47.

Danzon P and Towse A (2002) The economics of gene therapy and of pharmacogenetics. *Value Health*, **5**(1) 5–13.

Detsky AS, Naglie G, Krahn MD, Naimark D and Redelmeier DA (1997) Primer on medical decision analysis: Part 1–Getting started. *Medical Decision Making*, **17**, 123–125.

Drummond M, Dubois D, Garattini L, Horisberger B, Jönsson B, Kristiansen IS, Le Pen C, Pinto CG, Bo Poulsen P, Rovira J, *et al*. (1999) Current trends in the use of pharmacoeconomics and outcomes research in Europe. *Value in Health*, **2**, 323–332.

Durant J, Clevenbergh P, Halfon P, Delgiudice P, Porsin S, Simonet P, Montagne N, Boucher CA, Schapiro JM and Dellamonica P (1999) Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet*, **353**(9171), 2195–2199.

Evans WE and Relling MV (1999) Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science*, **286**, 487–491.

Exner DV, Dries DL, Domanski MJ and Cohn JN (2001) Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction. *The New England Journal of Medicine*, **344**, 1351–1357.

Farlow MR, Lahiri DK, Poirier J, Davignon J and Hui S (1996) Apolipoprotein E genotype and gender influence response to tacrine therapy. *Annals of the New York Academy of Sciences*, **802**, 101–110.

Flowers C and Veenstra DL (2000) Will pharmacogenomics in oncology be cost-effective? *Oncology Economics*, **1**, 26–33.

Friend SH (1999) How DNA microarrays and expression profiling will affect clinical practice. *British Medical Journal*, **319**, 1–2.

Fry RN, Avey SG and Sullivan SD (2003) The Academy of Managed Care Pharmacy Format for Formulary Submissions: an evolving standard – a Foundation for Managed Care Pharmacy Task Force report. *Value Health*, **6**(5), 505–521.

Garber AM and Phelps CE (1997) Economic foundations of cost-effectiveness analysis. *Journal of Health Economics*, **16**(1), 1–31.

Hisashige A (1997) Healthcare technology assessment and the challenge to pharmacoeconomics in Japan. *Pharmacoeconomics*, **11**(4), 319–333.

Hirsch MS, Brun-Vezinet F, D'Aquila RT, Hammer SM, Johnson VA, Kuritzkes DR, Loveday C, Mellors JW, Clotet B, Conway B, *et al*. (2000) Antiretroviral drug resistance testing in adult

HIV-1 infection: recommendations of an International AIDS Society-USA Panel. *The Journal of the American Medical Association*, **283**, 2417–2426.
Hodgson J and Marshall A (1998) Pharmacogenomics: Will the regulators approve? *Nature Biotechnology*, **16**, 243–246.
Holtzman NA and Marteau TM (2000) Will genetics revolutionize medicine? *The New England Journal of Medicine*, **343**, 141–144.
Israel E, Drazen JM, Ligget SB, Boushey HA, Cherniack RM, Chinchilli VM, Cooper DM, *et al*. (2000) The effect of polymorphisms of the B2-adrenergic receptor on the response to regular use of albuterol in asthma. *American Journal of Respiratory and Critical Care Medicine*, **162**(1), 75–80.
Iyer L and Ratain MJ (1998) Pharmacogenetics and cancer chemotherapy. *European Journal of Cancer*, **34**, 1493–1499.
Jonsson EG, Nothen MM, Gustavsson JP, Neidt H, Bunzel R, Propping P and Sedvall GC (1998) Polymorphisms in the dopamine, serotonin, and norepinephrine transporter genes and their relationships to monoamine metabolite concentrations in CSF of healthy volunteers. *Psychiatry Research*, **79**, 1–9.
Kleyn PW and Vesell ES (1998) Genetic variation as a guide to drug development. *Science*, **281**, 1820–1821.
Krynetski EY and Evans WE (1999) Pharmacogenetics as a molecular basis for individualized drug therapy: The thiopurine S-methyltransferase paradigm. *Pharmaceutical Research*, **16**(3), 342–349.
Kuivenhoven JA, Jukema JW, Zwinderman AH, de Knijff P, McPherson R, Bruschke AV, Lie KI and Kastelein JJ (1998) The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. *The New England Journal of Medicine*, **338**, 86–93.
Langley PC (1999) Formulary submission guidelines for Blue Cross and Blue Shield of Colorado and Nevada. Structure, application and manufacturer responsibilities. *Pharmacoeconomics*. **16**(3), 211–224.
Lerman C, Narod S, Schulman K, Hughes C, Gomez-Caminero A, Bonney G, Gold K, Trock B, Main D, Lynch J, *et al*. (1996) BRCA1 testing in families with hereditary breast-ovarian cancer. A prospective study of patient decision making and outcomes. *JAMA*, **275**(24), 1885–1892.
Lichter JB and Kurth JH (1997) The impact of pharmacogenetics on the future of healthcare. *Current Opinion in Biotechnology*, **8**, 692–695.
Liggett SB (1998) Pharmacogenetics of relevant targets in asthma. *Clinical and Experimental Allergy*, **1**(Suppl. 28), 77–79.
Lindpainter K (1999) Genetics in drug discovery and development: Challenge and promise of individualizing treatment in common complex diseases. *British Medical Bulletin*, **55**, 471–491.
Marshall A (1997) Getting the right drug into the right patient. *Nature Biotechnology*, **15**, 1249–1252.
Mather DB, Sullivan SD, Augenstein D, Pete Fullerton DS and Atherly D (1999) Incorporating clinical outcomes and economic consequences into drug formulary decisions: A practical approach. *The American Journal of Managed Care*, **5**, 277–285.
McHutchinson JG, Gordon SC, Schiff ER, Shiffman ML, Lee WM, Rustigi VK, Goodman ZD, Ling M, Cort S and Albrecht JK (1998) Interferon Alfa–2b alone or in combination with Ribavirin as initial treatment for chronic hepatitis C. *The New England Journal of Medicine*, **339**, 1485–1492.
NICE. National Institute for Clinical Excellence (1999) *Appraisal of New and Existing Technologies: Interim Guidance for Manufacturers and Sponsors*, August Obtained at **http://www.nice.org.U.K./appraisals/apr_gide.htm**, accessed 12/20/99.
Pahor M, Psaty BM, Alderman MH, Applegate WB, Williamson JD, Cavazzini C and Furberg CD (2000) Health outcomes associated with calcium antagonists compared with other first-line antihypertensive therapies: A meta-analysis of randomised controlled trials. *Lancet*, **356**, 1949–1954.
Persidis A (1998) The business of pharmacogenomics. *Nature Biotechnology*, **16**, 209–210.

PBAC. Pharmaceutical Benefits Advisory Committee (1999) *Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the PBAC*, Obtained at **http://partners.health.gov.au/hfs/haf/docs/pharmpac/part1.htm**, accessed 12/20/99.

Phillips KA, Veenstra DL and Sadee W (2001a) Implication of the human genome project for health services research: Pharmacogenomics. *Health Services Research*, **35**(5), Part III, pp 128–140.

Phillips K, Veenstra DL, Oren E, Lee JK and Sadee WS (2001b) The potential role of pharmacogenomics in preventing adverse drug reactions: A systematic review. *The Journal of the American Medical Association*, **286**(18) 2270–2279.

Phillips KA, Veenstra D, Van Bebber S and Sakowski J (2003) An introduction to cost-effectiveness and cost-benefit analysis of pharmacogenomics. *Pharmacogenomics*, **4**(3), 231–239.

Poirier J (1999) Apolipoprotein E: A pharmacogenetic target for the treatment of Alzheimer's disease. *Molecular Diagnosis*, **4**, 335–341.

Regalado A (1999) Inventing the pharmacogenomics business. *American Journal of Health-system Pharmacy*, **56**, 40–50.

Rettie AE, Wienkers LC, Gonzalez FJ, Trager WF and Korzekwa KR (1994) Impaired (S)-warfarin metabolism catalysed by the R144C allelic variant of CYP2C9. *Pharmacogenetics*, **4**, 39–42.

Rioux PP (2000) Clinical trials in pharmacogenetics and pharmacogenomics: Methods and applications. *Journal of Health Economics*, **57**, 887–898.

Sadee W (1999) Pharmacogenomics. *British Medical Journal*, **319**, 1–4.

Taube J, Halsall D and Baglin T (2000) Influence of cytochrome P-450 CYP2C9 polymorphisms on warfarin sensitivity and risk of over-anticoagulation in patients on long-term treatment. *Blood*, **96**(5), 1816–1819.

Temple R (1999) Are surrogate markers adequate to assess cardiovascular disease drugs? *The journal of the American Medical Association*, **282**, 790–795.

Veenstra DL, Higashi MK and Phillips KA (2000a) *Assessing the Cost-effectiveness of Pharmacogenomics*, AAPS PharmSci 2: Article 29 (**http://www.aaps.org/journals/pharmsci/veenstra/manuscript.htm**).

Veenstra DL, Tran C, Lum B and Cheung R (2000b) The cost-effectiveness of genetic screening and combination therapy for hepatitis C. Abstract. *Drug Information Association 2nd Annual Workshop on Pharmaceutical Outcomes Research, Seattle, May 11–12, 2000*.

Weinstein MC, Goldie SJ, Losina E, Cohen CJ, Baxter JD, Zhang H, Kimmel AD and Freedberg KA (2001) Use of genotypic resistance testing to guide HIV therapy: Clinical impact and cost-effectiveness. *Annals of Internal Medicine*, **134**(6), 440–450.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS and Russell LB (1996) Recommendations of the panel on cost-effectiveness in health and medicine. *The Journal of the American Medical Association*, **276**, 1253–1258.

Weinstein MC and Staason WB (1977). Foundations of cost-effectiveness analysis for health and medical practices. *The New England Journal of Medicine* **296**(13), 716–721.

Younossi ZM, Singer ME, McHutchison JG and Shermock KM (1999) Cost effectiveness of interferon alpha2b combined with ribavirin for the treatment of chronic hepatitis C. *Hepatology*, **30**, 1318–1324.

Short Specialist Review

Molecular dysmorphology

Leslie G. Biesecker
National Institutes of Health, Bethesda, MD, USA

1. The origins of dysmorphology

Dysmorphology is defined as the study of abnormal form, typically in humans. It is the study of abnormal human embryonic development that can lead to major anomalies (those that impair health or require surgical correction) or minor anomalies (such as a low-set ear or mildly curved fifth finger). Clinical dysmorphology is a subdiscipline of clinical genetics. It is practiced by detailed clinical evaluation of patients to catalog their physical features such as cranial shape, eye position and shape, ear configuration, and dozens of other physical assessments, and these morphological data are integrated with functional assessments such as mental retardation, renal function, and so on, to form a global assessment of an individual. These attributes are then matched to clinical references (Cohen, 1997; Jones, 1997; Guest *et al*., 1999; Gorlin *et al*., 2001) to determine if the findings in the patient are sufficiently similar to a previously recognized malformation syndrome. The roots of the discipline arose around the beginning of the twentieth century in Germany, but with the collapse of German science after World War II, this discipline lay dormant until it was reactivated in the 1960s. This reactivation was pioneered by Robert Gorlin, John Opitz, David Smith, and Josef Warkany. During this phase, the primary activity was the clinical delineation of discrete pleiotropic disorders of human development, known as *birth defect* or *malformation syndromes*. Hundreds of these disorders were described that could be diagnosed clinically but only a handful were associated with etiologic factors such as chromosomal aberrations (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1, e.g., Pallister–Killian syndrome) or environmental exposures (e.g., fetal alcohol syndrome). A large number of syndromes were inherited as simple Mendelian traits, but the technology did not exist to identify the genes that were mutated in those disorders.

2. Coupling genomics to dysmorphology

One technologic advance that facilitated the rapid dissection of the molecular basis of single gene dysmorphic syndromes was the conceptual development of positional cloning and the resources of the Human Genome Project (*see* Article 23, **The technology tour de force of the Human Genome Project**, Volume 3).

The existence of familial Mendelian dysmorphic syndromes allowed the meiotic mapping of the loci for these disorders and positional cloning of mutations in genes (*see* Article 9, **Genome mapping overview**, Volume 3). The determination that these genes were mutated in humans with dysmorphic syndromes implicated the genes as being necessary for normal development and coupled human clinical dysmorphology to developmental biology. This fusion of disciplines has been termed *molecular dysmorphology*, and it has facilitated rapid advances in the understanding of the molecular basis of normal and abnormal development (Epstein *et al*., 2004).

3. Utility of molecular dysmorphology

The morphology and pathology of humans has been studied in more detail than any other organism (Arguably, *Caenorhabditis elegans* may exceed the human because the fate and origin of every cell in the organism has been determined. However, the functional consequences of abnormal development are not nearly so detailed.). A by-product of intense and sophisticated human medical care is the cataloging of enormous amounts of clinical data on functional and anatomic variants. In contrast to the situation in the human is the situation with the mouse. Although there are many mouse mutants and the genetics of mouse breeding and transgenesis are very powerful (*see* Article 38, **Mouse models**, Volume 3), the phenotypic data are generally much weaker than in the human. This problem is most evident in human disorders that cause mental retardation or other functional disorders of the CNS (central nervous system). In a number of cases, it has been very difficult to characterize the CNS dysfunction in mouse mutants or knockout models. For these reasons, human phenotypic and positional cloning efforts have been coupled to complementary studies in animals where the genetics and other experimental parameters can be manipulated and controlled (*see* Article 37, **Functional analysis of genes**, Volume 3).

4. Examples of the power of molecular dysmorphology

Vertebrate limb development is an attractive system because it is relatively simple compared to some other organ systems (e.g., the CNS) and accessible to study. Major limb anomalies include limb reduction defects and limb duplications such as polydactyly. There are more than 100 heritable disorders that include polydactyly, either isolated or syndromic forms (Biesecker, 2002). Although each of these disorders is rare, as a group, the affected patients are frequently encountered in clinics and wards. Complementing the need for improved understanding of the medical causes and care of patients with limb anomalies is the opportunity to discover basic mechanisms of limb development. This opportunity drives the research on the molecular etiology of birth defects and there are many examples of successes. We have studied a group of disorders with a rare form of polydactyly, central polydactyly (Winter *et al*., 1993). This form of polydactyly occurs in

at least five different syndromes but interestingly has never been reported as a nonsyndromic form of polydactyly. The implication of this is that the genes that are involved in central digit specification are all likely to be part of a coordinated genetic pathway and because they are pleiotropic, are also critical for the development of other organ systems. We began by studying Pallister–Hall syndrome, which includes central polydactyly, bifid epiglottis and cleft larynx, imperforate anus, pulmonary segmentation anomalies, and hamartoma of the hypothalamus (Biesecker and Graham, 1996). It is inherited as an autosomal dominant trait, and the causative mutations were positionally cloned in 1997. The causative gene is the *GLI3* transcription factor, and all patients with PHS have mutations that predict production of truncated forms of the protein (Kang *et al*., 1997). This demonstrated that PHS was allelic to the Greig cephalopolysyndactyly syndrome (GCPS), which manifests polydactyly of the thumb/great toe or fifth digits, wide-spaced eyes, and macrocephaly (Vortkamp *et al*., 1991). A wide variety *GLI3* mutations cause GCPS. By comparing *GLI3* to its *D. melanogaster* homolog, *cubitus interruptus (ci)*, a hypothesis was developed regarding the basic biology of the human gene. Intriguingly, *ci* was known to be proteolytically processed into two forms, an activator and repressor form, but this was only known to occur in *Drosophila* (Aza-Blanc *et al*., 1997). Because all PHS mutations predicted truncated proteins that were similar to the processed repressor form of the ci protein, the prediction was made that GLI3 was similarly processed in mammals (Biesecker, 1997). The processing of GLI3 was subsequently described in mice (Wang *et al*., 2000) and is essentially certain to also occur in humans, because of evolutionary conservation.

There are many other examples of disorders where the positional cloning of the human gene led to important biological insights. These have included the fibroblast growth factor receptors in skeletal dysplasias and craniosynsotoses (Cohen, 2002; Wilkie *et al*., 2002), filamins A and B in at least seven dysmorphologic syndromes and skeletal dysplasias (Robertson *et al*., 2003; Krakow *et al*., 2004), PAX6 in aniridia, Peter's anomaly, other eye anomalies, and CNS malformations (van Heyningen and Williamson, 2002), and there are many other examples. These successes are impressive but must be regarded only as a first step. After the identification of the causative mutations, critical follow-up studies include detailed phenotype–genotype studies to determine if the patterns and types of mutations reveal more about the biology of the gene. In addition, the human data are then used to develop new scientific questions that can then be tested either by additional human studies, or more commonly, by moving into appropriate animal model systems such as the mouse, zebrafish, chicken, fruit fly, and so on.

In summary, molecular dysmorphology is a scientific approach to the study of human malformations that is a piece of the large scientific mosaic that needs to be assembled to develop a full understanding of the biologic and medical aspects of malformation syndromes. It couples the power of genomics and positional cloning with detailed phenotypic assessments of variations in human form. When coupled with complementary studies in animal model systems, molecular dysmorphology can be a crucial tool to understand the molecular basis of vertebrate developmental biology.

Acknowledgments

This chapter is dedicated to the memory of Professor Robin M. Winter (1950–2004), a pioneer of dysmorphology and the objective analysis of human birth defects.

References

Aza-Blanc P, Ramírez-Weber F-A, Laget M-P, Schwartz C and Kornberg TB (1997) Proteolysis that is inhibited by hedgehog targets Cubitus interruptus protein to the nucleus and converts it to a repressor. *Cell*, **89**, 1043–1053.

Biesecker LG (1997) Strike three for *GLI3*. *Nature Genetics*, **17**, 259–260.

Biesecker LG (2002) Polydactyly: how many disorders and how many genes? *American Journal of Medical Genetics*, **112**, 279–283.

Biesecker LG and Graham JM, Jr (1996) Syndrome of the month: Pallister-Hall syndrome. *Journal of Medical Genetics*, **33**, 585–589.

Cohen MM, Jr (1997) *The Child with Multiple Birth Defects*, *Second Edition*, Oxford University Press: New York.

Cohen MM, Jr (2002) Malformations of the craniofacial region: evolutionary, embryonic, genetic, and clinical perspectives. *American Journal of Medical Genetics*, **115**, 245–268.

Epstein CJ, Erickson RP and Wynshaw-Boris A (2004) *Inborn Errors of Development*, Oxford University Press: New York.

Gorlin RJ, Cohen MM, Jr and Hennekam RCM (2001) *Syndromes of the Head and Neck*, *Fourth Edition*, Oxford University Press: Oxford.

Guest SS, Evans CD and Winter RM (1999) The online London Dysmorphology database. *Genetics in Medicine*, **1**, 207–212.

Jones KL (1997) *Smith's Recognizable Patterns of Human Malformation*, *Fifth Edition*, W B Saunders: Philadelphia.

Kang S, Graham JM, Olney AH and Biesecker LG, Jr (1997) *GLI3* frameshift mutations cause autosomal dominant Pallister-Hall syndrome. *Nature Genetics*, **15**, 266–268.

Krakow D, Robertson SP, King LM, Morgan T, Sebald ET, Bertolotto C, Wachsmann-Hogiu S, Acuna D, Shapiro SS, Takafuta T, *et al*. (2004) Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis. *Nature Genetics*, **36**, 405–410.

Robertson SP, Twigg SR, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, Kenwrick SJ, Kim CA, Morava E, Newbury-Ecob R, *et al*. (2003) Localized mutations in the gene encoding the cytoskeletal protein filamin a cause diverse malformations in humans. *Nature Genetics*, **33**, 487–491.

van Heyningen V and Williamson KA (2002) PAX6 in sensory development. *Human Molecular Genetics*, **11**, 1161–1167.

Vortkamp A, Gessler M and Grzeschik K-H (1991) *GLI3* zinc finger gene interrupted by translocations in Greig syndrome families. *Nature*, **352**, 539–540.

Wang B, Fallon J and Beachy P (2000) Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell*, **100**, 423–434.

Wilkie AO, Patey SJ, Kan SH, van den Ouweland AM and Hamel BC (2002) FGFs, their receptors, and human limb malformations: clinical and molecular correlations. *American Journal of Medical Genetics*, **112**, 266–278.

Winter RM, Schroer RJ and Meyer LC (1993) *In Human Malformations and Related Anomalies*, Vol. 2, Stevenson RE, Hall JG and Goodman RM (Eds.), Oxford University Press: New York, pp. 805–843.

Short Specialist Review

Changing paradigms of genetic counseling

Katherine A. Schneider
Dana-Farber Cancer Institute, Boston, MA, USA

Vickie Venne
Huntsman Cancer Institute, Salt Lake City, UT, USA

1. Introduction

Genetic counseling, as a medical specialty, has been around since the 1970s, and like most things that have been around that long, the profession has undergone major evolution. This article will provide an overview of genetic counseling providers and services, with a focus on the profession's current "face".

2. Genetic counseling providers

The first genetic counseling providers were physicians, mostly pediatricians who cared for babies with congenital anomalies. With the onset of prenatal testing in the 1970s and the increased number of molecular tests available in the 1990s, overwhelmed physicians came to rely on allied health professionals to provide genetic counseling. Today, the typical genetic counseling provider has a Master of Science (MS) degree from an accredited genetic counseling training program and has passed a certification exam by the American Board of Genetic Counseling (ABGC). Currently, there are over 30 genetic counseling training programs in North America, as well as programs in England and Australia. Two states (Utah and California) have passed legislation to issue state licenses to genetic counselors and licensure is actively being sought in at least 15 other states. Other health care professionals who provide genetic counseling include geneticists, with either MD or Ph.D. degrees, and nurses with special training in genetics.

Almost all practicing genetic counselors in the United States belong to The National Society of Genetic Counselors (NSGC), a membership organization of about 2000 members. Affiliate professional organizations include the American Society of Human Genetics, American College of Medical Genetics, Canadian Association of Genetic Counselors, and the International Society of Nurses in Genetics. Genetic counselors also have close relationships with consumer support organizations, especially the Alliance for Genetic Support Groups.

3. Genetic counseling settings

In the 1970s, fewer than 10 training programs graduated genetic counselors, who then went on to work primarily in prenatal and pediatric genetic clinics at a few select medical centers. Today, genetic counselors are working in all 50 states and Canadian provinces and can be found in clinical settings as well as in educational, policy, research, administrative, biotechnology, and laboratory settings. Genetic counseling services are offered at almost all major medical centers, in satellite clinics, and in some private physician practices.

Genetic counselor job opportunities have changed dramatically over the past 10 years. In addition to the growing number of positions that do not involve direct clinical care, there is also a shift from prenatal and pediatric jobs to multiple different specialty areas. Similar to other medical providers, clinical genetic counselors are becoming increasingly specialized in areas as diverse as teratology, assisted reproductive technology, behavioral disorders, and adult-onset disorders. The most rapidly expanding job market for clinical genetic counselors is in the area of adult-onset disorders, specifically neurology, heart disease, and cancer. For more specific information about cancer genetic counseling, please refer to Article 88, **Cancer genetics**, Volume 2.

4. Indications for genetic counseling referrals

In the 1970s, most genetic counseling referrals were clients with prenatal or pediatric concerns. Genetic counseling referrals now occur at any point of life, from preconception or prenatal counseling through pediatric and adolescent years to adult-onset disorders. Individuals are referred to genetic counselors because of concerns about their own future health or the health of their offspring. Referrals can be triggered by the recognition of features or symptoms in a client, a pattern of disease within the family tree (pedigree), or a prospective parent's age or ethnicity. Genetic counseling clients have often been referred by a physician, although clients are increasingly self-referred.

5. Components of genetic counseling sessions

In 1975, the American Society of Human Genetics published a formal definition of genetic counseling as a communication process dealing with issues of occurrence, or risk of occurrence of a genetic condition (American Society of Human Genetics Ad Hoc Committee on Genetic Counseling, 1975). Over time, this definition has been adapted to include: (1) identifying a genetic condition, (2) explaining the inherited component to the client, (3) evaluating the client's options for action, (4) helping the client make a decision, and (5) supporting the client before, after, and during the process.

Historically, genetic counselors followed a nondirective counseling style. In this model, a counselor does not make direct recommendations to clients, but rather

helps them make the decision by presenting the pros, cons, and logistics of various options. Many genetic counselors are currently advocating for a change to a "shared decision-making" or "clients as partners" model, which is a more accurate reflection of current genetic counseling practices.

Genetic counseling services are best conducted as face-to-face interactions, although adjunct services include the use of CD-ROMs and Internet-based resources, group counseling, and telephone-based interactions. As telemedicine and Web-based learning becomes incorporated into the genetic counseling repertoire, the challenges will be how to obtain reimbursement for services, avoid potential provider liability, and still maintain close rapport with clients.

Genetic counseling sessions vary greatly depending on the specialty area. General components include: collection of the personal, pregnancy, and family history; risk assessment; description of potential genetic syndromes or heritable component of the condition; discussion and arrangement of genetic testing or medical management options; and counseling to help adapt to the choices, condition, and other familial issues that arise.

Collecting family history: The goal of pedigree collection and analysis is to determine if there is a pattern of disease in the family consistent with an inherited condition. Depending on the reason for the referral, genetic counselors typically collect health information about relatives over three to four generations. The information is diagrammed in a pedigree (Figure 1) and includes information about current ages or ages at death, presence of major health problems or diagnoses, pregnancy histories, and ethnicity.

Risk assessment: On the basis of information obtained from the family history (and often confirmed by medical records), the genetic counselor will discuss the likelihood that the client has a hereditary condition or is at risk for having a child with a genetic disorder. Some families will receive reassuring information, while those that do have a high likelihood of an inherited disorder will then receive additional information and support.

Description of genetic syndrome: Clients are educated about the hereditary condition, including the inheritance pattern, associated health and mental findings, the natural course and variability of the condition, and medical management issues, including procedures that are likely to be needed as well as screening and prevention options.

Genetic testing: Genetic testing is currently available for hundreds of inherited conditions, and includes cytogenetic, biochemical, and molecular analysis. Genetic testing can occur in prenatal settings, as part of an evaluation of a child or adult with a genetic condition, or be offered to adults worried about their predisposition of developing a specific condition in their family. A discussion about genetic testing includes the possible test results, the implications of the test result to the patient and other family members, the pros and cons of testing, and logistics of how testing would be done, including the cost of the analysis and when results will be available. Certain genetic tests may be available only from research laboratories. Individual results cannot be given to clients without confirmation from a clinically certified (CLIA approved) laboratory.

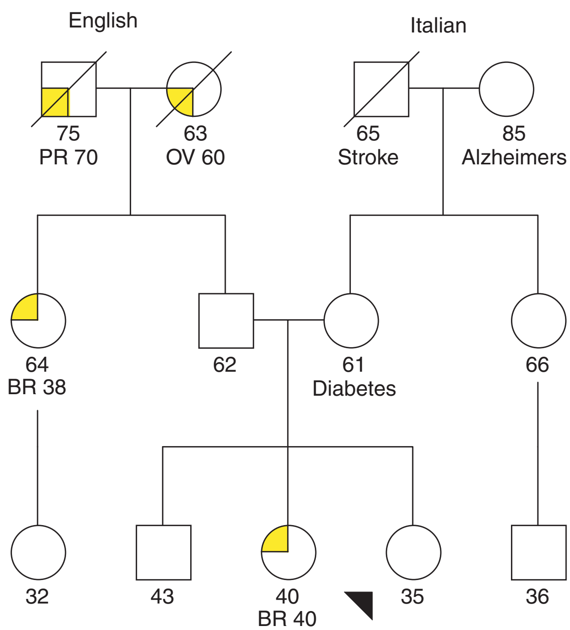


Figure 1 Example of a Genetic Counseling Pedigree. This is a 3-generation family tree depicting a family history of breast and ovarian cancer. Note that squares are men and circles are women. Current age or age at death is also noted. (Diagonal lines indicate people who are deceased.) The client (proband) is indicated by the arrow. Each cancer diagnosis and age at onset are indicated underneath the affected relative and the quadrants are shaded according to the general cancer type: upper left quadrant – breast cancer; lower left – genitourinary cancers; lower right – gastrointestinal cancers; and upper left – all other cancer types. In terms of this case example, the shaded circles and squares illustrate personal histories of breast cancer in the proband's sister and paternal aunt, ovarian cancer in the paternal grandmother, and prostate cancer in the paternal grandfather

Most clinical genetics programs require one to two additional counseling visits to arrange for testing and then to disclose and counsel about the result.

Counseling: During any consultation, genetic counselors will assess the psychological ability of the individual and/or family to cope with the information and management or testing choices that are being discussed. Family and social networks that can be used to support both the genetic counseling/test/diagnostic process and the natural history of the diagnosis are also identified. Genetic counselors use coping strategies, decision-making models, and other mechanisms to help clients adapt to the diagnosis or high-risk status.

6. Genetic counseling complexities and special issues

Most genetic conditions are rare, and family members become the medical experts regarding their condition. They may find themselves in the uneasy position of

having to describe the syndrome to their health care providers. Although this situation also occurs in other medical specialties, it is only one of the issues unique to or more intensified in the world of genetic counseling. This section illustrates some of the interesting family dilemmas and difficult ethical situations surrounding genetic counseling and testing.

Genetic conditions are a family affair: The identification of a genetic condition can send a ripple effect through the family as the awareness of risk spreads. Certain family members will be at risk while others will not, a fact that can seriously strain family relationships. Genetic test results may exacerbate communication problems in families that are already dysfunctional or who operate with secretive or deceptive communication styles. Situations may also arise in which one family member wants genetic testing information but other relatives do not. Or perhaps a person with a positive genetic test refuses to divulge risk information to his/her adult offspring despite potential consequences for their health. Clients may worry about their obligations to disclose genetic test results to other family members. Individuals need to balance the importance of warning relatives about the condition versus the anxiety and guilt that may be triggered by this information.

Genetic counseling raises sensitive issues: Meeting with a genetic counselor and discussing risks of birth defects or disease can be a frightening prospect. Even collecting information about the family history can evoke memories of relatives who have died as well as other family secrets. The discussions can trigger a number of coping strategies from denial to overintellectualization. Within a counseling session, clients may exhibit a range of emotions, including grief, fear, anger, and even humor. Genetic counselors are well versed in counseling techniques that allow them to be empathetic and supportive throughout the interactions. Since emotional distress can derail the capacity to learn, counselors must also continually gauge the client's emotional state during the encounter and determine whether it is important to pause or explore these emotions further.

Genetics is complicated and can outpace treatments: Another challenge is the enormous amount of complex information that needs to be understood by the client. A discussion of risk will include information about biology, math, and statistics – concepts often not familiar to the average person. The discussion can have intense emotional overtones, since the information is about reproductive fitness or the health of the client and/or close relatives. The cloning of the 34 000 genes through the Human Genome Project provides great hope for future therapies and prevention of genetic disorders. However, currently, there are very few genetic conditions that can be fixed, cured, or prevented. Genetic testing has been described by clients as looking into a crystal ball to determine their fate, but at the moment, it remains a cloudy view. In most genetic conditions, uncertainties remain as to the severity of symptoms or, in the cases of adult-onset conditions, if and when the problems will occur.

7. Genetic testing is changing the landscape of medicine

In the 1970s, genetic counseling was about providing clinical service to prenatal and pediatric clients. The field has expanded into areas that did not exist 30 years ago, and much of the basic information about genetics is now being provided by primary care providers. The onset of predictive genetic tests will continue the shift in medicine from urgent care to prevention. The potential number of genetic predictive tests is staggering, in areas as diverse as disease prediction (e.g., heart-attack risk) and pharmacogenomics (e.g., targeting drugs to people's genotypes). Everyone likely has a genetic predisposition to one or more disorders. Therefore, the entire health care profession will need to be prepared to deal with a population of people at various genetic risks. This will require increased knowledge about genetics and greater access to genetic services, especially in underserved areas. Families will need to be reassured that genetic test results will not lead to discrimination from third-party sources, such as insurers, employers, or colleges. Finally, the increased availability of personalized genetic screens and subsequent additional medical care by at-risk people will potentially place a burden on medical services. Given the complexities of the human genome, genetic counselors and geneticists will continue to play important, albeit different roles, in the "genomic era".

Reference

American Society of Human Genetics Ad Hoc Committee on Genetic Counseling (1975) Genetic counseling. *American Journal of Human Genetics*, **27**, 240–242.

Short Specialist Review

Ethical and legal issues in medical genetics

Mary Kay Pelias

GenELSI Consulting Inc., New Orleans, LA, USA

1. Introduction

Over the course of the twentieth century, the knowledge and technology of medical genetics expanded exponentially. From tenuous beginnings in descriptions of single-gene diseases, through the mid-century elucidation of the structure and functions of DNA, the practice of medical genetics has expanded to include a host of therapies, including specialized pharmaceuticals and the inception of gene transfer for controlling the ravages of deleterious genes. Concomitant with the growth of medical options, an array of ethical and legal questions has focused our attention on the interests and obligations of professionals, families, patients, and those who participate as subjects in medical research. Many of the ethical issues have received considerable debate and study. Some issues have generated resolution, while many continue to elude consensus among professionals in both clinical practice and scientific research arenas and among professionals and lay experts in bioethics.

2. Professional–patient and professional–subject relationships

Interactions between health care professionals and researchers, on the one hand, and their patients and subjects, on the other, evolved significantly over the twentieth century. The ancient Hippocratic admonition to "do no harm" acquired new meanings as the practice of medicine embraced such innovations as antisepsis and anesthesia. These advances and many others expanded the array of options that could be offered to patients and subjects, and these options resulted in significant shifts in the balance of relationships between health care providers and consumers. The physician–patient relationship traditionally followed the ethical principle of beneficence and was characterized by a fiduciary imbalance that is derived from the physician's specialized training and knowledge and the patient's relative lack of medical expertise. Indeed, the patient expected the knowledgeable physician to exercise his role of beneficence in making diagnostic and treatment decisions on behalf of the patient.

As medical knowledge and options continued to grow, the emphasis on professional beneficence shifted to an emphasis on the personal autonomy of patients and subjects, who became aware of multiple choices for decisions about health care. This new awareness was first articulated in civil litigation brought by plaintiff patients whose physicians had not acquired express permission for specific therapies or procedures. The foundation of respect for the patient's personal autonomy was articulated in a 1914 holding that "every human being of adult years and sound mind has a right to determine what shall be done with his own body" (*Schloendorff v. Society of New York Hospital*, 1914). This principle was applied in the world of biomedical research only after atrocities of human experimentation became public knowledge in the mid-twentieth century and the subject of protective legislation in the United States in the 1970s.

3. Informed consent

As the ethical principle of beneficence yielded to the ethical principle of personal autonomy, the courts and legislatures further defined the obligations of health care professionals in the principles of informed consent. These principles evolved along two pathways. In the civil litigation of medical negligence, health care professionals are now required to convey all "material" information to their patients, with "material information" defined as any information that could cause the patient to choose another option. In the legislative regulation of biomedical research, the research professional must provide the subject with information about the purpose and goals of the research, about possible risks and benefits of participation, and any alternatives, including the right to decline to participate and the right to withdraw without prejudice. Thus, the traditional fiduciary duty of professionals to make decisions for patients and subjects has yielded to a duty of the professional to inform the patient or subject fully so that he or she is equipped to make a rational choice among available options. This duty obligates the professional to be informed about new developments in medicine and research and to anticipate changes in the standard of care as the courts and legislatures become aware of new knowledge in basic science and new avenues of treatment (Pelias, 1991).

4. Genetic testing in children and parental autonomy

Exhaustive study during the 1990s of the risks and benefits of testing children for later-onset diseases resulted in guidelines for prudent decisions about generating genetic information on minors (ASHG/ACMG Report, 1995). Genetics professionals enjoy a consensus that the primary reason for testing children is to provide an immediate medical benefit. Beyond this justification, however, testing children for other reasons should be approached with caution. Such reasons might include, for example, plans for financial support, or trusts, for children, or plans to relocate near a center that specializes in medical care for particular disorders or diseases. Thorough and careful counseling should provide parents with information about both the risks and benefits of testing children and about how the information might

be disseminated to children judiciously, without causing psychological damage or creating negative stigmata. Once the genetics professional has provided such counseling, however, the professional should defer to the parents as the primary decision makers for their own families (Pelias and Markward, 2000).

5. Privacy and nondiscrimination

The concept of privacy in medicine and biomedical research can be viewed from three perspectives. Personal privacy describes privacy of the body and the sphere of personal space around the body that is shared only with consent. Decisional privacy describes personal autonomy in making decisions for oneself, as in the prescriptions for informed consent. Informational privacy describes the confidentiality of information that is shared in physician–patient and researcher–subject relationships.

While confidentiality is the cornerstone of the relationship of professionals to patients and subjects, a growing change is anticipated by arguments for a "duty to recontact" patients or subjects when significant new information becomes available. This issue has been litigated in at least three early medical malpractice cases not related to genetics, but these cases may eventually be resurrected for parallel arguments in the context of genetic testing. Further, the possibility of recontacting patients and subjects raises the question of establishing reciprocity in the professional relationship so that new information can be forwarded expediently to the benefit of both parties (Pelias, 1991). Similarly, genetics professionals are examining the possibility of a "duty to warn" at-risk relatives when genetic testing of a patient or subject indicates that relatives of the patient or subject might have an increased risk of bearing the same deleterious gene. Guidelines for discussing these exceptional circumstances with patients or subjects before testing and for respecting the rights of patients and relatives have been assembled in order to maximize medical benefits and reduce legal risks for both professionals as well as patients, subjects, and their relatives (Knoppers *et al*., 1998).

Confidentiality of genetic information has also become a focus of attention with respect to the insurance industry and protections in the workplace. Individuals whose private genotype information has been disclosed have lost health insurance coverage and, in some cases, have resorted to litigation to establish and protect their own interests. Some cases illustrate the poor understanding of insurance and workplace personnel in the science and implications of genetics (Billings *et al*., 1992). Indeed, egregious violations of personal privacy are documented in industrial settings when employers provide health care benefits and physical examinations that are stretched to include genotyping without the knowledge or consent of the employee. A significant outcome of the documentation of genetic discrimination and privacy violations is the plethora of state legislation efforts to prohibit such prejudicial and intrusive activities and to protect the privacy of the individual with respect to genes and genetic prospects (Roche, 2002). Finally, federal protections have been achieved under the Americans with Disabilities Act (ADA), in the Health Insurance Portability and Accountability Act (HIPAA), and in several recent congressional bills expressly aimed at protecting genetic privacy and preventing discrimination based on genotype (Hustead and Goldman, 2002).

6. Newborn screening and use of archived tissue samples

Genetic testing of newborn infants has conferred immense benefits to children and to the public for over four decades. Identification of some deleterious genotypes in the newborn period allows early initiation of treatments that prevent the accumulation of harmful metabolites that could otherwise cause retardation, serious illness, and even death. Because of the immense benefits that accrue from early detection and treatment, all but two states have legislated mandatory screening that circumvents the requirement of parental consent: the minor intrusion to obtain a few drops of blood is viewed as insignificant when weighed against the immense benefits of early detection and treatment.

New questions are arising, however, as the technology for testing expands to permit testing for genes that cause late-onset problems or are related to susceptibilities for complex disorders. The justification for mandatory screening for serious diseases of infancy and childhood may not be stretched to other genetic tests for which parental consent is appropriate, if not obligatory. Similar concerns arise over the recent thrust to obtain archived blood spots, with or without personal identifiers, for exploring new and future tests and technologies. Whether and how archived tissue samples may be used in future research remains an unsettled issue (Pelias and Markward, 2001).

7. Race, ethnicity, and pharmacogenetics

As the molecular structure of a growing number of genes is characterized, researchers in both academia and industry are focusing on designing pharmaceuticals that are tailored to interact with disease-related alleles or gene products. Such alleles are, however, often found to be concentrated in one or another racial or ethnic human group, unevenly distributed across the entire population. A major source of ethical concern derives from the possibility that smaller human groups will be characterized and stigmatized by their particular deleterious genes. Among suggestions for avoiding such negative social consequences of pharmacogenetics is a call for concentrated education programs directed to the general public. Such an effort is viewed as a massive undertaking but essential for public appreciation of human variance and the potential benefits of allele-specific pharmaceuticals (Rothstein and Epps, 2001).

8. Conclusions

The ethical and legal issues outlined above are past and present concerns in the practice of medical genetics, in genetic counseling, and in clinical and basic research. Some concerns, such as genetic testing in children and newborn screening for diseases of infancy, have attained satisfactory resolution. Others, however, such as newborn screening for late-onset diseases and susceptibilities and social questions in pharmacogenetics, remain unsettled. As yet, unanticipated ethical issues can be

expected to arise as we continue to explore the human genome and the nature of genetic health and disease. Perhaps, the most significant outcome of the discussions and studies of these issues is a new awareness among geneticists about ethical, legal, and social issues, and a new concern for fostering the understanding of these issues among the public at large and within smaller racial and ethnic groups in the human population.

References

ASHG/ACMG Report (1995) Points to consider: Ethical, legal, and psychosocial implications of genetic testing in children and adolescents. *American Journal of Human Genetics*, **57**, 1233–1241.

Billings PR, Kohn MA, deCuevas M, Beckwith J, Alper JS and Natowicz MR (1992) Discrimination as a consequence of genetic information. *American Journal of Human Genetics*, **50**, 476–482.

Hustead JL and Goldman J (2002) Genetics and privacy. *American Journal of Law & Medicine*, **28**, 285–307.

Knoppers BM, Strom C, Clayton EW, Murray T, Fibison W and Luther L (1998) Professional disclosure of familial genetic information. *American Journal of Human Genetics*, **62**, 474–483.

Pelias MZ (1991) Duty to disclose in medical genetics: A legal perspective. *American Journal of Medical Genetics*, **39**, 347–354.

Pelias MZ and Markward NJ (2000) The human genome project and public perception: Truth and consequences. *Emory Law Journal*, **49**, 837–858.

Pelias MZ and Markward NJ (2001) Newborn screening, informed consent, and future use of archived tissue samples. *Genetic Testing*, **5**, 179–185.

Roche PA (2002) The genetic revolution at work: Legislative efforts to protect employees. *American Journal of Law & Medicine*, **28**, 272–283.

Rothstein MA and Epps PG (2001) Pharmacogenomics and the (ir)relevance of race. *The Pharmacogenomics Journal*, **1**, 104–108.

Schloendorff v. Society of New York Hospital, 211 N.Y. 125, 105 N.E. 92 (1914).

Mechanisms of inheritance

Arthur S. Aylsworth
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

1. Introduction

Gregor Mendel grew up in an area of central Europe called Moravia, where there was great interest in plant and animal breeding. As a young man, Mendel was intrigued by the observation that for some traits, when two individuals with different phenotypes are crossed, all members of the first generation express the trait characteristic of only one of the parents. The trait is faithfully expressed and is not a combination of the traits from the two parents. But when members of that first generation are crossed, the trait that had been lost suddenly reappears, perfectly expressed, in the next generation in a predictable and reproducible fashion. This "lost" trait is seen in one-quarter of the individuals produced, while the remaining three-quarters continue to express the trait seen in the previous generation. While studying in Vienna, Mendel formulated his hypothesis that heredity is determined by *units* that occur in pairs. He returned to his monastery at Bruno, and after spending 8 years carefully selecting seven traits in *pisum* that follow this pattern, proceeded to test and demonstrate the truth of his hypothesis. Note that it took Mendel many years to select a few traits that follow this inheritance pattern. That is because most traits in peas and humans do not obey these deterministic rules of expression that we now refer to as "Mendelian inheritance".

Mendel's great contribution was the discovery of *unitary inheritance*, which was in dramatic contrast to the then existing theory of *inheritance by a blending of characters*. Modern interpretations are usually summarized as the principles of *segregation* and *independent assortment*. The principle of segregation states that alleles are paired and each gamete normally receives only one member of each pair. Exceptions to this principle include trisomy and uniparental disomy. Independent assortment refers to the fact that pairs of genes tend to segregate independently of each other. The exception to this principle is genetic linkage.

The concepts of *dominant* and *recessive* defined by Mendel refer to the *phenotypic expression* of alleles in heterozygotes, not to intrinsic characteristics of gene loci. Therefore, the older practice used by some early geneticists of referring to a locus as dominant or recessive should be avoided.

In diploid organisms, a *homozygote* (adj. *homozygous*), strictly speaking, is an individual that has two identical alleles at a gene locus, while a *heterozygote* (adj. *heterozygous*) has two different DNA sequences at a particular locus. Note that in

the fields of human and medical genetics, the term "homozygote" is also frequently used to refer to an individual with a homozygous phenotype, even though the genotype may consist of two different disease-causing mutations. Such individuals are *compound heterozygotes* or *allelic compounds*.

Inheritance patterns are determined by one or more genes. From a functional standpoint, genetic factors can be thought of as monogenic or multigenic and deterministic or nondeterministic. Medical geneticists traditionally subdivide inheritance patterns into (1) chromosome abnormalities (multigenic, deterministic), (2) single-gene disorders (monogenic and usually deterministic, but sometimes nondeterministic), and (3) multifactorial or complex conditions (multigenic, nondeterministic). One also must consider the roles of environment, epistasis, and chance in the production of phenotypic expression.

Exceptions and variations on the simple, basic themes of monogenic inheritance commonly involve issues of *penetrance* and *expression*.

1.1. Penetrance

Penetrance (P) refers to the proportion of gene carriers that express a particular trait. P ranges from 0 to 1 (or 0 to 100%). Traits or conditions with decreased or incomplete penetrance ($P < 1.0$ or $< 100\%$), therefore, fall into the category of monogenic nondeterministic inheritance.

1.2. Expression

Expression refers to the observable characteristics of the phenotypic trait in question such as quantity, quality, severity, age of onset, body part or organ system involved, and so on. Variable expression may occur within families (intrafamilial variability) or between/among different families (interfamilial variability). Intrafamilial variability may be due to epistasis, environment, chance, or mosaicism. Interfamilial variability may be due to epistasis, environment, chance, and allelic or locus heterogeneity.

Penetrance and expression may be affected by the gender of an individual, the age of an individual (i.e., time), transmission through families, environmental factors, the effects of other genes (epistasis), stochastic factors (chance), and the presence of mosaicism.

- Examples of sex-influenced expression include male pattern baldness, breast and ovarian cancer, and anomalies of internal and external genitalia.
- For all disorders of postnatal onset, penetrance will depend on age. It should be noted that the term "variable penetrance" is commonly misused for the concept of "variable expressivity". In fact, the term "variable penetrance" is only appropriate for conditions of postnatal onset, where there are different penetrance figures at different ages.
- Transmission through a family can be associated with a change in the size and, therefore, the deleterious effects of trinucleotide repeat expansion mutations, a process called *anticipation*. *Imprinting* causes expression to be determined by

the gender of the transmitting parent (*see* Article 36, **Variable expressivity and epigenetics**, Volume 1).

- Environmental factors may be required for phenotypic expression. For example, the autosomal recessive inborn error of metabolism phenylketonuria is characterized by mental retardation. But a genetic deficiency in the activity of the enzyme phenylalanine hydroxylase requires dietary phenylalanine to produce this deleterious phenotype.
- Epistasis is assumed to be a ubiquitous principle, based on developing knowledge about the complexities of genetically determined biomolecular interactions in cells and tissues and on data collected from plant and animal experiments. For most or all traits, expression of an allele probably depends at some level on interactions with other gene products.
- A factor traditionally ignored in discussions about the cause of genetic disease is chance. Kurnit *et al*. (1987) used computer modeling to show that some of the non-Mendelian familial clustering of anomalies usually attributed to concepts such as "reduced penetrance" and "multifactorial inheritance" may be accounted for by simple, random chance. Central to this concept is the idea that biological processes, like many other phenomena that we encounter daily, are error-prone, nonlinear systems. The complex processes of embryologic development and physiologic homeostasis may be very sensitive to small perturbations, which are by themselves "within normal limits", that subsequently, *by chance*, are amplified by the randomness that is inherent in all such systems.
- Mosaicism describes the situation in which an individual has cells with more than one genotype. For example, females are mosaic for X-linked gene expression, because some of their cells express the father's X and other cells express the mother's X. Individuals with sporadic (i.e., nonfamilial) cases of autosomal dominant disorders may have less severe manifestations than those who inherit the condition from a parent because in some sporadic cases, the mutation occurred in the postfertilization zygote and only a fraction of body cells carry the mutant allele. Postzygotic chromosome rearrangements cause a mosaic pattern to be found on karyotyping (*see* Article 18, **Mosaicism**, Volume 1). Postzygotically acquired mutations may be evenly distributed throughout the body (e.g., mild osteogenesis imperfecta caused by mosaicism for a severe lethal disease mutation), limited to specific parts or sides of the body (e.g., segmental neurofibromatosis type I), or limited to specific tissues of origin (e.g., tissue-specific mutations and chromosome rearrangements associated with malignancy).

2. Multigenic deterministic inheritance

This category includes abnormalities that involve large segments of the genome, traditionally called "chromosome abnormalities" in standard texts (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1, Article 12, **The visualization of chromosomes**, Volume 1, Article 13, **Meiosis and meiotic errors**, Volume 1, Article 16, **Nondisjunction**, Volume 1, Article 17, **Microdeletions**, Volume 1, Article 22, **FISH**, Volume 1, Article 67, **Approach to**

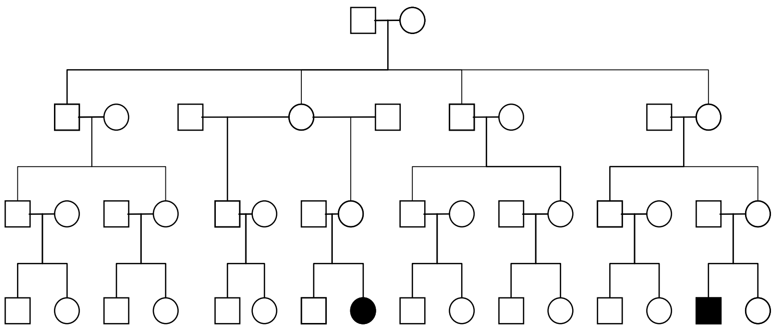


Figure 1 Pedigree with only two second cousins affected out of 30 relatives in a four-generation family. Such pedigrees are frequently assumed to be explained by complex or multifactorial inheritance. Note that this pedigree is also consistent with several other possible mechanisms of inheritance including the following: an unbalanced chromosomal rearrangement where intervening, unaffected relatives carry the balanced rearrangement; autosomal dominant with incomplete penetrance; X-linked inheritance with incomplete penetrance in heterozygous females; and mitochondrial inheritance where expression depends on an environmental exposure such as aminoglycoside-induced deafness (compare with Figure 2)

rare monogenic and chromosomal disorders, Volume 2, Article 71, **Advances in cytogenetic diagnosis**, Volume 2, and Article 87, **The microdeletion syndromes**, Volume 2). Examples include trisomies, partial trisomies, partial monosomies, and smaller duplications and deletions. Cytogenetically detectable abnormalities are found in a little over 0.5% of newborn babies, and in a much higher percentage of miscarriages and fetal deaths. Submicroscopic cytogenetic rearrangements include microduplications and microdeletions that cause "contiguous gene" syndromes.

Submicroscopic microdeletions and microduplications may be vertically transmitted from generation to generation as are the autosomal dominant disorders described below. Syndromes caused by chromosome abnormalities may be inherited through families in a pattern suggesting complex or multifactorial inheritance (Figure 1). For example, cousins may have unbalanced karyotypes and be affected, while intervening, phenotypically unaffected relatives are asymptomatic carriers of a balanced chromosomal rearrangement.

3. Monogenic deterministic inheritance

Frequently called "Mendelian" inheritance patterns, these are well covered in standard textbooks of genetics. They include the "single-gene disorders", autosomal and X-linked dominant and recessive inheritance, and mitochondrial patterns. The major features are summarized below.

3.1. *Autosomal dominant inheritance*

This occurs when one allele of a heterozygous pair causes a phenotypic trait. Many classic human genetic disorders are caused by this mechanism and follow this pattern. An organism heterozygous for the trait-causing allele has a 50% chance,

or one chance out of two, of passing that allele on to each offspring. Transmission in families, therefore, is from one generation to the next in a pattern that is sometimes called "vertical inheritance". Males and females are typically both affected and affected equally unless there is sex-limited expression. Male-to-male transmission occurs, distinguishing this from X-linked dominant inheritance.

Mendel's definition of dominant expression was that the heterozygote appeared identical to the mutant homozygote. In human genetics, autosomal dominant diseases are relatively rare (frequencies generally less than one in several thousand). Therefore, heterozygote matings are extremely rare, and the human homozygous phenotype is unknown for most human "dominant" disorders. A few are known. Huntington disease appears to be a true Mendelian dominant because homozygote and heterozygote expressions are similar. On the other hand, the condition known as classic achondroplasia, traditionally called an autosomal dominant, is not. Homozygous offspring from two heterozygous parents are typically much more severely affected, and usually die at birth or in the first few weeks of life. Heterozygous achondroplasia is compatible with long survival and reproduction, while homozygous achondroplasia is a genetic lethal. Other terms used to describe this phenomenon include "incomplete dominant" and "semidominant" inheritance.

A paternal age effect is observed for some autosomal dominant traits. This refers to the phenomenon where new mutations are associated with advanced paternal age because mutations due to unrepaired DNA copy errors increase proportionally with increasing paternal age.

3.2. Autosomal recessive inheritance

This occurs when two phenotypically normal parents carry mutant alleles at the same locus and produce a phenotypically affected offspring that inherits a mutant allele from both parents. Penetrance in heterozygotes is zero, because they appear identical to the wild-type homozygotes. Because affected sibs may be produced by heterozygous parents, this pedigree pattern has been referred to as "horizontal inheritance". Carrier parents have a 25% risk, or one chance out of four with each pregnancy, of having a child who is affected by virtue of being homozygous or an allelic compound. Unaffected sibs of affected patients each have a two-third chance of being a carrier.

Consanguinity is frequently a cause of both parents carrying the same rare recessive allele. Parental consanguinity is not significantly increased, however, for common recessives such as sickle-cell disease, cystic fibrosis, or hemochromatosis. Consanguinity can result in a phenomenon called "pseudodominant inheritance". This occurs when a homozygous (or compound heterozygous) affected individual mates with an unaffected heterozygote, producing an affected offspring. A parent and child are affected but the condition is autosomal recessive because the heterozygote is not affected. Pseudodominant inheritance of common recessive traits does not require consanguinity.

3.3. X-linked inheritance

This occurs when a phenotype is caused by an allele on the X chromosome. Because of the pedigree patterns produced, this type of inheritance is sometimes called

"diagonal inheritance". X-linked inheritance is characterized by the lack of male-to-male transmission and the fact that affected males pass their mutation-bearing X chromosomes on to all of their daughters but to none of their sons. Traditionally, human geneticists talk about X-linked recessive and dominant disorders. The former is characterized by no apparent manifestations in carrier females, but sometimes, subtle findings are apparent if a careful evaluation is performed. The latter term is used when both males and females are significantly affected, even though on close examination differences may be found.

3.4. Autosomal and X-linked intermediate expression

When carrier females have a phenotype much milder than that of affected males, a term such as X-linked intermediate may be used. In fact, because of random variation in Lyonization, a variable degree of phenotypic expression can be expected in females who carry alleles for X-linked disorders and traits.

Phenotypes may be dominant or recessive depending on how one defines them. For example, the disease sickle-cell anemia is autosomal recessive but the phenomenon of in vitro sickling is a dominant trait.

Some genes may produce a wide range of mutant phenotypes. Mutations at the locus for the tissue nonspecific alkaline phosphatase gene may be recessively expressed (i.e., no phenotype in the heterozygote but severe hypophosphatasia in the homozygote or allelic compound) or dominantly expressed as childhood or adult-onset hypophosphatasia. Heterozygous carriers of alleles that cause the recessive, multisystem disorder homocystinuria may have a degree of hyperhomocyst(e)inemia that places them at increased risk for adult vascular disease. It seems reasonable to predict that future research may find subtle phenotypic effects associated with heterozygosity for a number of other recessive disorders.

3.5. Y-linked inheritance

Holandric inheritance occurs when a phenotype is caused by an allele on the Y chromosome and is characterized by exclusive and obligatory male-to-male transmission. As of November 2004, the OMIM database (OMIM, 2000) listed 45 genes or phenotypes in the Y-linked catalog.

3.6. Mitochondrial inheritance

When a phenotype is caused by an allele in the mitochondrial genome, a pedigree phenomenon called "maternal inheritance" is produced, whereby all of the offspring of an affected woman are either affected or nonpenetrant carriers and none of the offspring of affected men are affected or carriers (Figure 2). Variable expression is due to the fact that within an organism's cells there is a mixture of normal and mutant mitochondria (heteroplasmy). Lower levels of mutant mtDNA are associated with milder or minimal symptoms, while higher levels cause more severe disease expression.

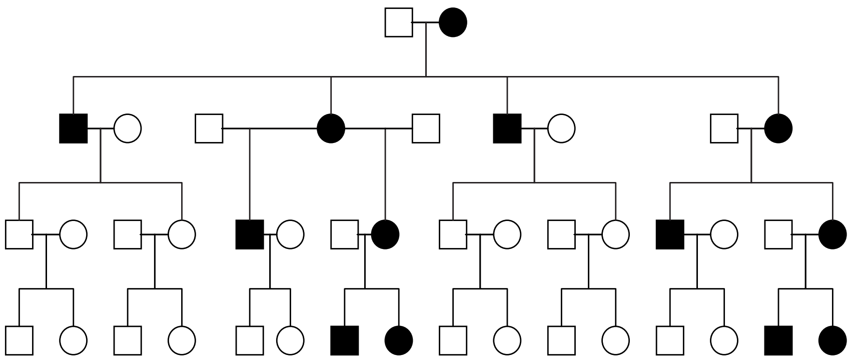


Figure 2 Pedigree showing maternal inheritance of a trait caused by a mutation in mitochondrial DNA. All offspring of affected women are also affected, while none of the offspring of affected males are affected

4. Monogenic nondeterministic inheritance

This category is frequently covered in standard texts as exceptions to Mendelian inheritance. When lifetime penetrance is incomplete, some individuals who carry disease-causing genes will be unaffected. Conditions with lifetime penetrance figures that range from 99% down to 10% or so usually are recognized and classified as monogenic disorders with incomplete penetrance. That is, they are recognized as having a major single genetic factor that causes the disorder. On the other hand, it is not clear how to classify conditions or traits caused by a single major genetic factors with penetrances of 1, 5, or even 10%. These can be thought of as *susceptibility genes* that predispose an organism to develop a particular trait or disease. Figure 1 shows a pedigree of a family segregating such a low penetrance, disease-susceptibility gene.

The less deterministic a single major gene effect, the more likely interactions with other gene products and environmental factors will significantly affect expression. Acute intermittent porphyria is an autosomal dominant condition with incomplete penetrance that depends on environmental exposures. Similarly, hemochromatosis is an autosomal recessive susceptibility genotype with low penetrance that interacts with environmental factors to produce disease. A classic example of an autosomal dominant structural anomaly with incomplete penetrance is split-hand split-foot malformation. In this case, there is some experimental evidence that phenotypic expression is affected by other genes (i.e., the individual's genetic background). Since the factors that influence penetrance and expressivity may be a mix of genes and environmental factors, the monogenic model, in practice, may actually be a complex "mixed" model of major single-gene effects on a polygenic background, where liability may be determined by more than one gene.

Many instances of mitochondrial inheritance fit in this category also. Variable expression and even incomplete penetrance may be due to heteroplasmy. Figure 1 illustrates how mitochondrial inheritance, using the same pedigree transmission as in Figure 2, can suggest complex or multifactorial inheritance. In this case,

the mitochondrial mutation required an environmental exposure for phenotypic expression. An example of this would be aminoglycoside-induced deafness.

5. Multigenic nondeterministic inheritance

The terms "complex inheritance" or "complex disease" imply that a single, causative, completely penetrant gene does not always produce the phenotype in question (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2, Article 58, **Concept of complex trait genetics**, Volume 2, and Article 59, **The common disease common variant concept**, Volume 2). Instead, a combination of effects from more than one gene, or one or more genes with environmental (nongenetic) factors may produce the phenotype. "Complex" phenotypes are *causally heterogeneous*. This usually means that over an extended population, the causes of a particular phenotype, trait, or disease may include both low frequency, high-penetrance causative (i.e., deterministic) alleles and more common, low-penetrance susceptibility alleles interacting with environmental factors. Most of the common disorders of children and adults are complex phenotypes. The common disorders of childhood include birth defects, mental retardation, and short stature. Common disorders in adults include cancer, diabetes, cardiovascular disease, hypertension, stroke, the psychoses, and so on. These conditions are frequently familial, but pedigrees usually do not suggest a straightforward pattern of Mendelian inheritance (Figure 1)

To explain these "non-Mendelian" patterns of familial recurrence, mathematical models were developed that assumed a continuous distribution of genetic "liability" to malformation or disease in the general population. "Polygenic" determination initially referred to a mathematical model in which a number of genes with small, additive effects provide an underlying genetic predisposition to malformation or disease. Some quantitative traits and clinical disorders in humans have been studied and found to be compatible with this mechanism of determination. "Multifactorial" describes models in which environmental factors interact with genetic predisposition, which is frequently thought of as polygenic. In the case of quantitative traits such as blood pressure, weight, and height, a normal curve would represent the distribution of measurements in a population. This model was adapted to account for discontinuous traits by the addition of a threshold, the point within that liability distribution beyond which individuals are affected.

5.1. Digenic causation

This designates a subset in which diseases or traits result from mutations in two genes at different loci. For example, if two unlinked genes encode polypeptide subunits of a functional protein complex, then mutations in either or both genes cause dysfunction and clinical phenotypes. While this may seem like a simple variation on the theme of monogenic inheritance, it is included under "multigenic nondeterministic inheritance" because its existence confounds standard linkage analysis. Examples include some forms of retinitis pigmentosa, nonsyndromic

hearing loss, junctional epidermolysis bullosa, and autosomal dominant polycystic kidney disease.

5.2. Oligogenic causation

"Oligogenic" designates a subset of "polygenic" causation in which diseases or traits result from the effects of relatively few genes, some of which may have rather large effects. Many current research efforts are directed toward identifying a small number of interacting genes that produce clinical phenotypes and diseases, without the necessity of identifying a quantitative, polygenic background of susceptibility. An example is Hirschsprung disease (HSCR), a disorder characterized by congenital absence of ganglion cells in the wall of the colon. Eight incompletely penetrant genes from three different but interacting signaling pathways are implicated in causing HSCR.

5.3. Environmental factors

These influence both embryonic development and adult disease. For example, the risk of developing emphysema is greater in individuals with both an environmental exposure such as smoking and a genetic predisposition such as alpha-1-antitrypsin deficiency than it is for individuals with only one of these risk factors or none at all. Severe respiratory distress in asthmatics is triggered by exposure to environmental triggers. Those with a familial susceptibility to skin cancer can decrease their risk of developing cancer by modifying their sun exposure. Obesity strongly affects morbidity in individuals with inherited susceptibilities to type II diabetes.

The relative merits of hypothetical models such as polygenic, multifactorial threshold, and major single gene with incomplete penetrance have been debated vigorously over the years. It seems reasonable to assume that complex phenotypes result from many different stochastic combinations of genetic and environmental factors, where the genetic contribution may be from one or more genes, and expression is merely a possibility, pending environmental exposure(s) and/or random fluctuations of normal processes in the embryo, fetus, or adult organism.

References

Kurnit DM, Layton WM and Matthysse S (1987) Genetics, chance, and morphogenesis. *American Journal of Human Genetics*, **41**, 979–995.

OMIM (2000) *Online Mendelian Inheritance in Man, OMIM*™, McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University: Baltimore, MD and National Center for Biotechnology Information, National Library of Medicine: Bethesda, MD.

Genetic family history, pedigree analysis, and risk assessment

Robin L. Bennett
Division of Medical Genetics, University of Washington, Seattle, WA, USA

1. Introduction

An array of high-tech genetic diagnostic tools have ushered in the era of genomic medicine (*see* Article 71, **Advances in cytogenetic diagnosis**, Volume 2 and Article 72, **Current approaches to molecular diagnosis**, Volume 2). Yet, obtaining a medical family history remains one of the most powerful but simple and cost-effective approaches to identifying individuals at risk for rare single-gene disorders as well as for common disorders with a genetic etiology (e.g., cancer, dementia, heart disease, stroke, diabetes, etc.). Health care professionals and genetic researchers should have basic competence in obtaining and interpreting a family medical history and pedigree (refer to NCHPEG genetics core competencies at **www.nchpeg.org**). A pedigree is an important method of establishing patient rapport, and serves as a visual demonstration for providing patient education (e.g., on variation in disease expression in a family), as well as identifying other relatives at risk for disease or who can be informative for inclusion in genetic research (Bennett, 1999).

2. Obtaining a pedigree

Typically a pedigree includes two generations of ascent from the consultand (the person requesting the medical or genetic information) or proband (the affected individual that brings the family to medical attention), and two generations of descent. Information is usually collected on first-degree relatives (children, siblings, and parents), second-degree relatives (half-siblings, aunts, uncles, nieces, nephews, grandparents, and grandchildren), and sometimes third-degree relatives (e.g., first cousins). For example, pedigree assessment for a 60-year-old man with a family history of colorectal cancer would include information about his parents, grandparents, aunts and uncles, and possibly his cousins. If a condition is suspected of being inherited, the pedigree is extended back as many generations as possible to include affected relatives.

Standardized pedigree symbols (Figures 1, 2, and 3) are used to record relationships and medical information in the form of a pedigree (Bennett *et al*., 1995). The square or circle representing an affected relative can be shaded. More than one condition can be shown on the pedigree by partitioning the circle or square into two or

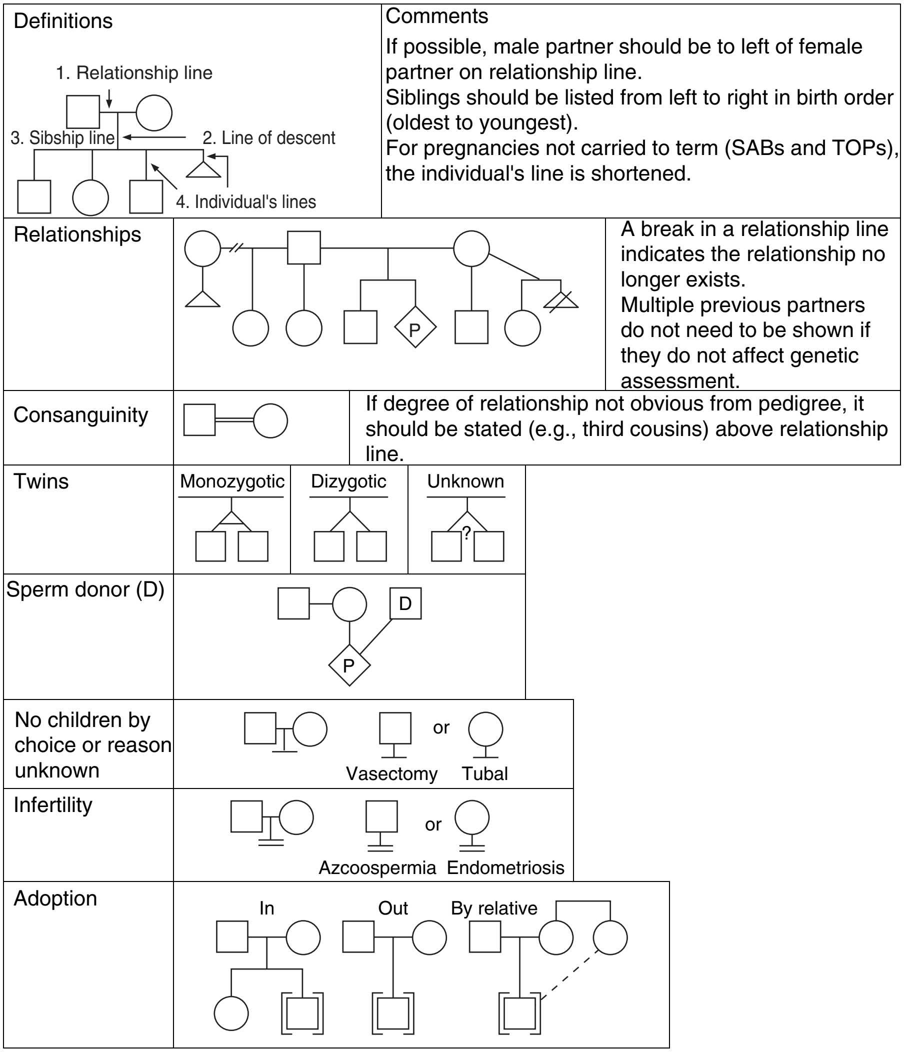


Figure 1 Common pedigree symbols (Reproduced from Bennett *et al*., 1995 by permission of University of Chicago Press)

four sectors and shading the appropriate sector or using different fill patterns (e.g., hatched). A pedigree key or legend is used to identify the medical conditions that are being traced in the pedigree and to explain any unusual abbreviations or less commonly used symbols (e.g., a gamete donor, adoption). Record on the pedigree: the age of relatives, their age at diagnosis of the condition in question, their age at death, and the cause of death. It is important to include both affected and unaffected relatives on the pedigree. The date the pedigree was obtained should be noted (for easy reference) as well as the name of the person who obtained the information.

Ideally, corroboration of the clinical diagnosis with medical records or death certificates provides the most accurate information for pedigree analysis and risk

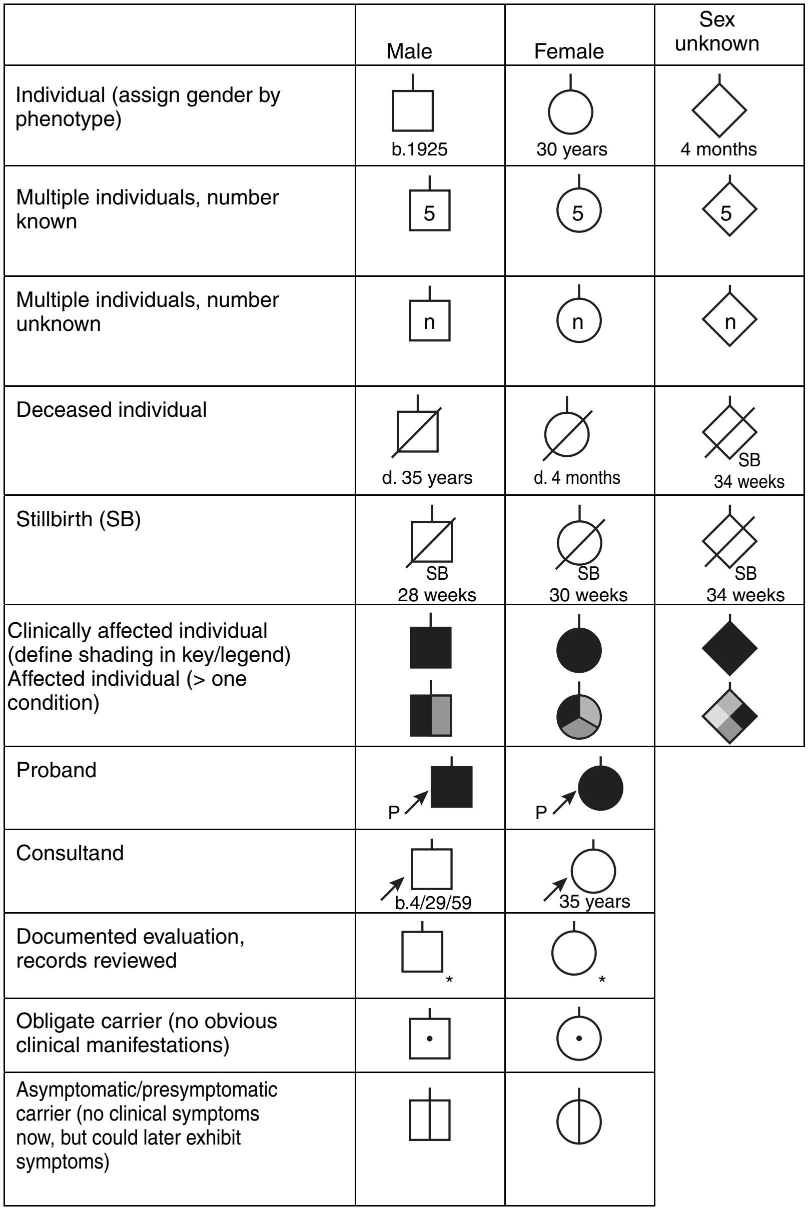


Figure 2 Standardized nomenclature for relationships in human pedigrees (Reproduced from Bennett *et al*., 1995 by permission of University of Chicago Press)

assessment. Even family photographs may be helpful when attempting to confirm a diagnosis. A client's recall of information about second- and third-degree relatives (e.g., aunts and uncles, grandparents, and cousins) is much less likely to be accurate than the information about more closely related relatives (e.g., siblings, parents, and children) (Bennett, 1999).

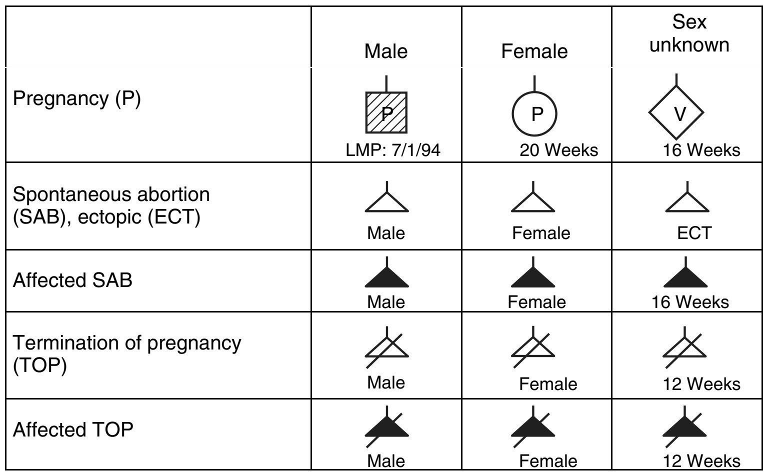


Figure 3 Standardized pedigree symbols and nomenclature for pregnancies and pregnancies not carried to term (Reproduced from Bennett *et al.*, 1995 by permission of University of Chicago Press)

The concept of race and ethnicity in genetic terms is evolving, but it is useful to note the countries of ancestral origin on a pedigree because the genetic risk assessment and the approach to genetic testing may vary on the basis of the client's ethnicity. For example, the likelihood of finding a BRCA1 or BRCA2 mutation in an individual with a family history of breast cancer who is of Ashkenazi ancestry is higher than for a person with a similar history of breast cancer who is Caucasian (Frank *et al.*, 2002; **www.myriadtests.com/provider/doc/mutprev.pdf**). Document the ancestral origins of both sets of grandparents (e.g., German, Japanese, Italian, Amish). Noting consanguinity (relationships between individuals with a common ancestor) is important for genetic risk assessment. Individuals with a common ancestor are more likely to share the same deleterious genes in common.

3. Pedigree tools

The pencil-and-paper approach to taking a family history remains a cheap and simple way of collecting family history. There are several software programs for use by professionals to record family histories that are particularly useful for large families involved in research protocols. Several on-line resources that can be downloaded to a personal computer have been developed to encourage clients to collect, organize, and maintain their medical family history, which can then be made available to their health care professionals (Guttmacher *et al.*, 2004). These include pedigree tools developed through the National Human Genome Resource Institute and the Centers for Disease Control (**www.hhs.gov/familyhistory**), the American Medical Association (**www.ama-ass.org/ama/pub/article/2380-2844.html**), and a joint collaboration of the National Society of Genetics Counselors, the Genetic Alliance, and the American Society of Human Genetics (**www.nsgc.org**), and the March of Dimes (**www.marchofdimes.com**).

4. Pedigree analysis and risk assessment

One of the major purposes of drawing a pedigree is to recognize patterns of inheritance. Table 1 reviews the basic patterns of inheritance and variables that can mask the recognition of these patterns.

Statistical risk assessment based on pedigree analysis, epidemiology (such as Hardy–Weinberg equilibrium), various risk models (such as Bayes theorem), and the sensitivity and specificity of various genetic tests is central to genetic counseling and risk assessment (Claus *et al*., 1994; Gail *et al*., 1989; Nussbaum *et al*., 2001; Parmigiani *et al*., 1998, Young, 1999). Genetic counseling involves explaining risks in multiple ways, such as distinguishing absolute from relative risks (e.g., a 10% absolute risk but a threefold increased risk), and using percentages to frame the magnitude of risks from different perspectives. For example, if autosomal recessive is the mode of inheritance, the chance of having an affected child would be framed in terms of a 25% or one in four chances of occurrence or recurrence, as well as a 75% chance of or three in four chances that a son or daughter would be unaffected. Comparing individual risk to population risks can help clients put their personal risk in perspective. Clinicians and researchers should attempt to remove their own biases about whether a risk is high or low from presentation of risk information.

A client's and family's reaction to genetic risk will depend on multiple variables including what are the client's preconceived notions about patterns of inheritance, chances of testing positive, or developing the family disorder? If the genetic risk information is divergent from a client's perception, the client may have difficulty incorporating the information or implementing any disease-management recommendations. What are the personal, family, religious, and cultural beliefs about disease causation? Reactions will also depend on the perceived burden of the condition – does the person have personal experience with the condition, and if so, have relatives survived the disease or has the disease been debilitating and/or resulted in early death? Other factors that are relevant to risk perception include level of cognitive functioning, temperament, and personality (e.g., pessimists or optimists, high-achievers who plan to "beat the risk"), external or internal locus of control, and coping style (e.g., information seekers, avoidance, dependency) (Bottoroff *et al*., 1998; Hallowell *et al*., 1997; McCarthy Veach *et al*., 2003).

5. Summary

A pedigree is an important instrument for genetic diagnosis and research. A pedigree can be used to establish the pattern of inheritance and diagnosis of a condition, and to demonstrate familial variation in disease expression (e.g., variable age of onset, variable disease severity). A pedigree can help a practitioner decide upon molecular testing strategies. It is also an excellent way to establish patient rapport and provide patient education. A pedigree can be recorded through a simple and easily available pen-and-paper method or incorporated into software tools that provide an easy mechanism to update the pedigree with new family history information, and to incorporate data analysis tools to help in the interpretation of

Table 1 Examples of clues for recognizing patterns of inheritance and variables that can mask the recognition of these patterns

Inheritance pattern	Mode of transmission	Pedigree clues	Confounding factors	Disease examples
Autosomal dominant	50% risk to each son/daughter	Condition in consecutive generations Male–male transmission Males/females affected Often variability in disease severity Homozygotes may be affected more severely Homozygous state may be lethal	Reduced penetrance Can miss diagnosis in relatives if there is a mild expression for disease New mutations may be mistaken for sporadic if family size is small	Neurofibromatosis 1 Breast–ovarian cancer syndrome Postaxial polydactyly Familial adenomatous polyposis Marfan syndrome Myotonic muscular dystrophy
Autosomal recessive	25% risk to each son/daughter Parents "healthy"	Usually one generation Males/females affected Often seen in newborn, infancy, childhood Often inborn errors of metabolism May be more common in certain ethnic groups (e.g., Tay–Sachs disease and Ashekenazism) Sometimes parental consanguinity	May be mistaken as sporadic if family size is small If carrier frequency is high, can look autosomal dominant (e.g., hemochromatosis)	Hemochromatosis Cystic fibrosis Sickle-cell anemia Phenylketonuria Tay–Sachs disease Beta-thalassemia Nonsyndromic neurosensory deafness Medium-chain-acyl-dehydrogenase deficiency (MCAD)
X-linked dominant	Heterozygous women affected with 50:50 risk to have affected daughter/50:50 chance for affected male (though lethal)	Females usually express condition but have milder symptoms than males No male-to-male transmission Often lethal in males so paucity of males in pedigree is seen	Small family size	Incontinentia pigmenti Orofacial digital syndrome 1 Rett syndrome

		May see multiple miscarriages (due to male fetal lethality) Females usually express condition but have milder symptoms than males		
X-linked	Women have 50% chance for affected son/50% chance for heterozygous daughter	No male-to-male transmission Males affected Females may be affected, but may be milder and/or with later onset than males	May be missed if paucity of females in family Lyonization	Duchenne muscular dystrophy Red-green color blindness Hemophilia A Fabry disease Fragile X syndrome
Chromosomal	Increased risk for trisomy observed with advanced maternal age Risk for affected fetus depends on specific chromosomal rearrangement (ranges from 1 to 15% or higher)	Suspect if the individual has 2+ major birth defects or 3+ minor birth defects Fetus with structural anomalies Unexplained MR (static) especially with dysmorphic features Unexplained psychomotor retardation Ambiguous genitalia Lymphedema or cystic hygroma in newborn Couples with three or more pregnancy losses Individuals with multiple congenital anomalies and family history of MRUnexplained infertility		Trisomy 21 (Down syndrome) Trisomy 18 Turner syndrome (45,X) Robertsonian translocations Reciprocal translocations Subtelomeric defects

(*continued overleaf*)

Table 1 (*continued*)

Inheritance pattern	Mode of transmission	Pedigree clues	Confounding factors	Disease examples
Mitochondrial	0–100%	No male transmission to offspring, only maternal transmission	Generally considered rare	Mitochondrial encephalopathy with ragged-red fibers (MERRF)
		Highly variable clinical expression		Mitochondrial encephalopathy, lactic acidosis, strokes (MELAS)
		Often central nervous disorders		Neropathy with ataxia and retinitis pigmentosa (NARP)
		Males and females affected, often in multiple generations		Neural tube defects
Multifactorial	Based on empiric risk tables	Males and females affected	May actually be single gene (e.g., missing diagnosis of velocardiofacial syndrome in family with cleft palate and congenital heart defect)	Neural tube defects
		No clear pattern		Scoliosis
		Skips generations		Cleft lip
		Few affected family members		Schizophrenia
				Bipolar disorder

MR = mental retardation. (Adapted from *Primary Care Clinics in Office Practice*, V1: 495–497, Acheson LS *et al*., The family medical history, Copyright 2004, with permission from Elsevier)

pedigree information. Its usefulness will continue to develop in tandem with the growing spectrum of molecular, bioinformatics, and proteomic tools.

Further reading

Bennett RL (2004) The family medical history. In *Primary Care Clinics in Office Practice 31*, Acheson LS and Wiesner GL, (Eds.), Saunders: Philadelphia, pp. 497–495.

References

Bennett RL (1999) *The Practical Guide to the Genetic Family History*, John Wiley & Sons: New York.

Bennett RL, Steinhaus KA, Uhrich SB, O'Sullivan CK, Resta RG, Lochner-Doyle D, Markel DS, Vincent V and Hamanishi J (1995) Recommendations for standardized human pedigree nomenclature. *American Journal of Human Genetics*, **56**, 745–752.

Bottoroff JL, Ratner PA, Johnson JL, Lovato CY and Joab SA (1998) Communicating cancer risk information: the challenge of uncertainty. *Patient Education and Counseling*, **33**, 67–81.

Claus EB, Risch N and Thompson WD (1994) Autosomal dominant inheritance of early-onset breast cancer: Implications for risk prediction. *Cancer*, **73**, 643–651.

Frank TS, Deffenbaug AM, Reid Je, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, School T, Tavtigian SV, Pruss DR, *et al*. (2002) Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: Analysis of 10,000 individuals. *Journal of Clinical Oncology*, **20**, 1480–1490.

Gail MH, Brinton LA, Byar DP, Corke DK, Green SB, Schairer C and Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, **81**, 1879–1886.

Guttmacher AE, Collins FS and Carmona RH (2004) The family history – more important than ever. *The New England Journal of Medicine*, **351**, 2333–2336.

Hallowell N, Statham H, Murton F, Green J and Richards M (1997) Talking about chance: The presentation of risk information during genetic counseling for breast and ovarian cancer. *Journal of Genetic Counseling*, **6**, 269–286.

McCarthy Veach P, LeRoy BS and Bartels DM (2003) *Facilitating the Genetic Counseling Process, A Practice Manual*, Springer: New York, pp. 122–149.

Nussbaum RL, McInnes RR, Willard HF and Thompson MW (2001) *Thompson and Thompson Genetics in Medicine*, *Sixth Edition*, WB Saunders: New York.

Parmigiani G, Berry DA and Aguilar O (1998) Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics*, **62**, 145–158.

Young ID (1999) *Introduction to Risk Calculation in Genetic Counseling*, *Second Edition*, Oxford University Press: Oxford.

The physical examination in clinical genetics

Marni J. Falk
Case Western Reserve University School of Medicine, University Hospitals of Cleveland, Cleveland, OH, USA

Nathaniel H. Robin
University of Alabama at Birmingham, Birmingham, AL, USA

1. Introduction

The physical examination is a valuable tool in medical practice that provides an objective supplement to historical information (*see* Article 78, **Genetic family history, pedigree analysis, and risk assessment**, Volume 2). To understand the special nature of the "genetic" physical exam, one must first recognize that the primary goal of a medical genetic evaluation is to identify a unifying etiology for seemingly unrelated birth defects, developmental problems, or other abnormal findings present in a fetus, child, or adult. Some may question the benefits of making a genetic diagnosis, as there are very few "cures" for such conditions. However, it is only by establishing a correct diagnosis that appropriate clinical management can be provided, along with accurate prognostic and recurrence risk counseling (*see* Article 75, **Changing paradigms of genetic counseling**, Volume 2). Understanding the pathogenesis of a patient's problems can further help families begin to cope with guilt they may feel about "why" their child has a particular problem and can direct families to contact appropriate support groups.

The genetic physical exam relies heavily on the art of dysmorphology, defined as the study of abnormal form (Aase, 1990). A detailed analysis of body structure is employed to detect potential embryologic deviations from normal development. While the actual physical examination is similar to the general medical examination, much closer attention is given to form, size, proportion, positioning, spacing, and symmetry (Hall, 1993). Indeed, precise observation and accurate description of all physical features is the cornerstone of the genetic dysmorphology exam. A genetic diagnosis may subsequently be reached through careful collation, interpretation, and categorization of any physical differences present (Hall, 1993; *see also* Article 74, **Molecular dysmorphology**, Volume 2).

2. Categorization of physical abnormalities

Physical exam findings that are indicative of a genetic diagnosis may be of varying clinical relevance. *Major anomalies* are structural alterations arising during embryologic development that have severe medical or cosmetic consequences and typically require therapeutic intervention. Examples include congenital heart defects, neural tube defects, and cleft lip or palate. *Minor anomalies*, in contrast, are medically and cosmetically insignificant departures from normal development that do not require significant surgical or medical treatment and have no risk for long-term sequelae. Minor external anomalies most commonly are found in areas where structures are most complex and variable, such as in the face, auricles, hands, and feet. Examples include wide-set eyes, low-set ears, or brachydactyly (short fingers or toes). A third category of physical findings is *minor*, or *normal*, *variants*. While these findings may also represent medically insignificant departures from normal development, they occur at a low frequency among the normal general population. Examples include café-au-lait macules, single palmar creases, and fifth finger clinodactyly. It is important to appreciate that minor anomalies and variants are significant only when taken in context. For example, while preauricular ear pits are common in the general population, they may lead to the diagnosis of branchio-oto-renal syndrome when identified in a patient with sensorineural hearing loss and renal anomalies.

Although major anomalies are more obvious and tend to draw attention to the patient more quickly, it is important to recognize that they are not more important than minor anomalies or normal variants as clues to a diagnosis (Hall, 1993). Rather, it is often the most subtle findings, discernable only through a genetics physical exam, that lead to the diagnosis of a genetic syndrome. Learning to differentiate a minor anomaly from normal variation is vitally important, as it is through these subtle anomalies that most genetic diagnoses are made. In particular, the rarer physical features may be the most useful in arriving at the correct differential diagnosis. Descriptions of normal features and variations can be found in excellent texts such as Diagnostic Dysmorphology (Aase, 1990) or Smith's Recognizable Patterns of Human Malformation (Jones, 1997).

3. Guidelines to performing the genetic physical exam

The exam itself has two interrelated components: descriptive observation and careful measurement. Each physical feature should be evaluated, then normal and variant features should be accurately described using accepted standard terminology. While impressions are important (e.g., "the palpebral fissures look small"), they are insufficient alone. "No clinical judgment should ever be made on a measurable parameter without actually having measured and compared it to standard references" (Hall, 1993). For example, hypertelorism (widely spaced eyes) noted on casual observation may be an illusion caused by a widened nasal bridge or epicanthal folds; measuring interpupillary distance may reveal that eye spacing actually falls within the normal range. Distinguishing such differences may be of considerable importance when trying to determine if minor anomalies are present and consistent with a particular syndrome.

In some instances, a genetic syndrome can be diagnosed by the overall appearance, or "gestalt" of the facial appearance. This ability to instantaneously recognize a syndrome is powerful and impressive to one's colleagues. However, even an experienced dysmorphologist can be wrong if one is too quick to make a diagnosis, as many findings are common to multiple disorders. Therefore, a careful and complete exam is warranted even in apparently obvious cases. Starting with the "big picture" is a useful approach, however, for examining each body region if followed by sequential evaluation of smaller component subunits within that region (Hall, 1993).

4. Components of the physical exam

When examining the patient's physical characteristics, it is important to be systematic and thorough. Growth parameters (i.e., head circumference, height, and weight) should be recorded on all patients, including adults. The general appearance and behavior of a patient should also be carefully observed. Many genetic disorders have characteristic behaviors that may have as much diagnostic importance as any physical finding. Examples include spontaneous laughter in patients with Angelman syndrome, self-hugging in patients with Smith–Magenis syndrome, and excessive anxiety with a friendly personality in patients with Williams syndrome (Cassidy and Morris, 2002).

A useful regional approach is to start with examination of the head and proceed downward. Greatest attention is given to examination of the head and face, as this is where the greatest variability of human features is seen. Assessing symmetry, placement, and proportions both of overall appearance and of all paired structures (e.g., eyes and ears) may detect differences not otherwise appreciated. The skull should be observed for symmetry and contour, then palpated for ridging of sutures and fontanel sizes. The forehead should be evaluated for contour and breadth. The placement, rotation, size, and configuration of the ears should be noted, as should the presence of preauricular pits or tags. The eye distances should be measured. The external eye should be evaluated for unusual position or structural abnormalities of the lids, irises, or pupils. Ophthalmologic evaluation for cataracts or retinal abnormalities should be attempted. The nose should be evaluated for the configuration of the nasal bridge and tip, as well as symmetry of the nasal septum. The configuration of the philtrum should be noted, along with the size of the mouth and any unusual features of the lips. The arch of the palate should be observed, along with any cleft of the palate or uvula. Teeth should be evaluated for size, placement, and abnormalities of enamel or configuration. In addition, the profile should be assessed for prominence or recession of the forehead, eyes, midface, and chin. Any redundancy of nuchal skin should be noted. Finally, a description of hair, eyebrows, and eyelashes should include assessment of texture, color, thickness, and length.

Evaluation of the trunk includes measuring relative proportions of upper and lower segment lengths (as divided at the symphysis pubis), sternal length, chest circumference, and internipple distance. The chest, clavicles, ribcage, and spine should be examined for deformity or abnormalities of contour. The positioning and form of the nipples and sternum should be observed. Any heart murmur, abdominal

wall defect, or organomegaly should be noted. Sacral spinal defects such as hair tufts or pits should be noted. Examination of the genitalia should include assessment of proper proportions and positioning of structures, as well as Tanner stage.

Careful examination of the extremities can also yield useful diagnostic clues. The relative proportions and symmetry of the various segments of each limb should be noted, as specific abnormalities may indicate a particular skeletal dysplasia. All joints should be assessed for contractures, abnormal angulation, or hypermobility (preferably using the Beighton score) (Beighton *et al*., 1997, 1998). Examination of the hands and feet require careful consideration, as they offer a wide range of insight into early fetal development. Measurements should be made of total hand, palm, middle finger, and foot lengths. Fingers and toes should be assessed for placement, contractures, webbing, and spacing. The contour of the soles of the feet should be noted. Careful observation should also be given to the texture and configuration of fingernails and toenails. For example, subungual fibromas may be the presenting sign of Tuberous Sclerosis in a patient with mental retardation and seizures.

The skin should be carefully examined in an effort to detect and describe any birthmarks or abnormalities in pigment, texture, elasticity, or wound healing. This may require using ultraviolet light in fair-skinned individuals. Observing dermatoglyphic patterns and skin creases is also important. Some conditions have characteristic dermatoglyphic patterns, as exemplified by a predominance of fingertip whorls in Smith–Lemli–Opitz syndrome. More generally, abnormal skin creases reflect an abnormality in early fetal movement before 18 weeks' gestation.

5. Placing the physical exam in context

Information obtained from a detailed medical and family history will often guide the physical exam, allowing one to focus extra attention on specific findings (*see* Article 78, **Genetic family history, pedigree analysis, and risk assessment**, Volume 2). For example, a postpubertal male referred for evaluation of mental retardation should be assessed for large ear size and macroorchidism, features that would raise suspicion for fragile X syndrome. Similarly, using background knowledge will allow one to direct attention to features that may have been missed. For example, the examination of an infant with ambiguous genitalia should include careful analysis of the toes, as cutaneous syndactyly (webbing) between the second and third toes in this context would raise suspicion for Smith–Lemli–Opitz syndrome.

Examining a patient's family members and/or reviewing their photographs can establish the background features that may assist the examiner in understanding where their patient's features might have originated. In particular, any parameter thought to be abnormal in the patient should be examined in their relatives.

A final point to consider is that physical findings change with time in many genetic syndromes. In Prader–Willi syndrome, for example, affected newborns are hypotonic and manifest failure to thrive, but within several years they become hyperphagic and obese (*see* Article 29, **Imprinting in Prader–Willi and Angelman syndromes**, Volume 1). Periodic evaluations are important in cases in

which a diagnosis cannot be reached, as physical examination findings may have changed into a recognizable pattern consistent with a known diagnosis. Similarly, reviewing photographs of a patient at various ages may detect phenotypic features classic for a particular condition that are no longer recognizable.

Related article

Article 86, **Uses of databases**, Volume 2

Further reading

Winters R and Baraitser M (1998) *Oxford Medical Database*, Oxford University Press Electronic Publishing, Polyhedron Software Limited.

Hall JG, Froster-Iskenius UG and Allanson JE (1989) *Handbook of Normal Physical Measurements*, Oxford University Press: Oxford.

References

Aase JM (1990) *Diagnostic Dysmorphology*, Plenum Medical Book Co.: New York.

Beighton P, De Paepe A, Steinmann B, Tsipouras P and Wenstrup RJ (1997, 1998) Ehlers-Danlos syndromes: revised nosology, Villefranche. *American Journal of Medical Genetics*, **77**, 31–37.

Cassidy SB and Morris CA (2002) Behavioral phenotypes in genetic syndromes: genetic clues to human behavior. *Advances in Pediatrics*, **49**, 59–86.

Hall BD (1993) The state of the art of dysmorphology. *American Journal of Diseases of Children*, **147**, 1184–1189.

Jones KL (1997) *Smith's Recognizable Patterns of Human Malformation*, *Fifth Edition*, W.B. Saunders Company: Philadelphia.

Genetic testing and genotype–phenotype correlations

Elfride De Baere
Ghent University, Ghent, Belgium

Ludwine Messiaen
University of Alabama at Birmingham, Birmingham, AL, USA

1. Introduction

In the wake of the Human Genome Project, thousands of genes involved in monogenetic Mendelian disorders have been cloned, allowing identification of the germ-line mutation (genotype) underlying the specific features in an affected person (phenotype). It was expected that knowledge of mutations would lead to consistent genotype–phenotype correlations, clarifying why a given genetic change results in a particular clinical phenotype. Elucidation of genotype–phenotype correlations would move clinical genetics toward predictive medicine, allowing a prognosis and better selection of therapeutic strategies for any individual patient. However, it has become clear that the correlation between genotype and phenotype is often incomplete. For most "single-gene" disorders, a significant number of mutant alleles do not correlate absolutely with a clinical phenotype because of the effect of additional independently inherited genetic variations and/or environmental influences (Dipple and McCabe, 2000).

Intrafamilial variability can result from a combination of the effects of other unlinked genes (modifier genes, quantitative trait loci), epigenetic factors (skewed X-inactivation, imprinting, methylation, etc.), and environmental factors (nutritional, toxins, drugs, chance events). Interfamilial variability can be enhanced by allelic heterogeneity when many different mutations within a given gene can be seen in patients presenting with a given disease.

To identify genotype–phenotype correlations, the type of mutation must be related to the severity of the disorder and/or the spectrum of features. First, the spectrum of mutations underlying the disorder must be fully understood. The coding region of the gene needs to be analyzed in a large number of patients with a given genetic disease. Mutations can lead to loss-of-function (i.e., decrease or absence of normal gene product, leading to a hypomorphic and null alleles respectively; loss of normal gene function plus antagonistic effect of the mutant gene product with the normal gene product, called *dominant negative effect* or *antimorphic allele*) or gain-of-function (increased or novel function of the protein).

Examples of diseases and disease genes where genotype–phenotype correlations have been observed to some extent are described below.

2. Loss-of-function and gain-of-function mutations in the same gene

In some genes, mutations can lead either to loss-of-function or gain-of-function, resulting in distinct clinical phenotypes.

The *RET* proto-oncogene (MIM 164761) on chromosome 10q11.2 encodes a cell surface tyrosine kinase, consisting of extracellular, transmembrane, and intracellular tyrosine kinase domains activating signal transduction. Mutations that cause loss-of-function (total gene and small intragenic deletions, nonsense mutations, and splicing mutations resulting in a truncated protein) result in Hirschsprung disease (MIM 142623), a congenital disorder characterized by the absence of enteric ganglia along a variable length of intestine. In contrast, mutations that cause gain-of-function effect result in either multiple endocrine neoplasia type 2A or 2B (MEN 2A/2B) (MIM 171400/162300), characterized by a high incidence of thyroid carcinoma and phaeochromocytoma. MEN2A patients have, in addition, an increased risk for parathyroid adenoma, and MEN2B patients develop mucosal neuromas, ganglioneuromatosis of the gastrointestinal tract, and a "Marfanoid" habitus. The activating mutations causing MEN 2A are clustered in five cysteine residues in the extracellular domain, while MEN 2B is caused by a unique mutation (M918T) in the tyrosine kinase domain (Edery *et al*., 1997) (Figure 1).

Fibroblast growth factors (FGF) play key roles in embryogenesis. The transduction of extracellular *FGF* signals is mediated by a family of four transmembrane tyrosine kinase receptors: the fibroblast growth factor receptors (FGFR), each of which contains an extracellular region with three immunoglobulin-like domains, a transmembrane segment, and two intracellular tyrosine kinase domains (Figure 2).

A dominant gain-of-function mutation of *FGFR1* (MIM 136350) underlies a form of craniosynostosis (Pfeiffer syndrome, MIM 101600) (Muenke and Schell,

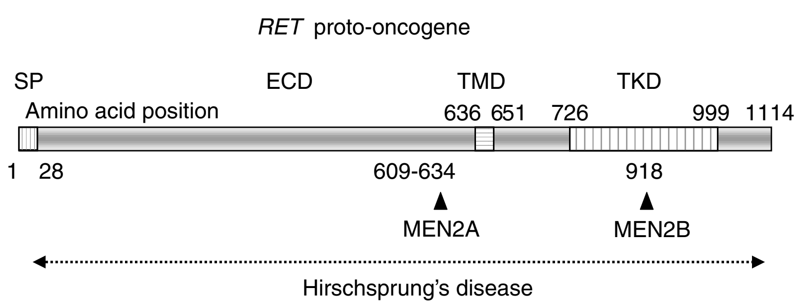


Figure 1 The *RET* proto-oncogene and its genotype–phenotype correlations. On top, the structure of the RET proteins is shown. Numbers represent amino acid residues. The most common mutation sites in different clinical entities are shown by arrows. Abbreviations used: SP: signal peptide; ECD: extracellular domain; TMD: transmembrane domain; TKD: tyrosine kinase domain; MEN2A and 2B: multiple endocrine neoplasia type 2A and 2B

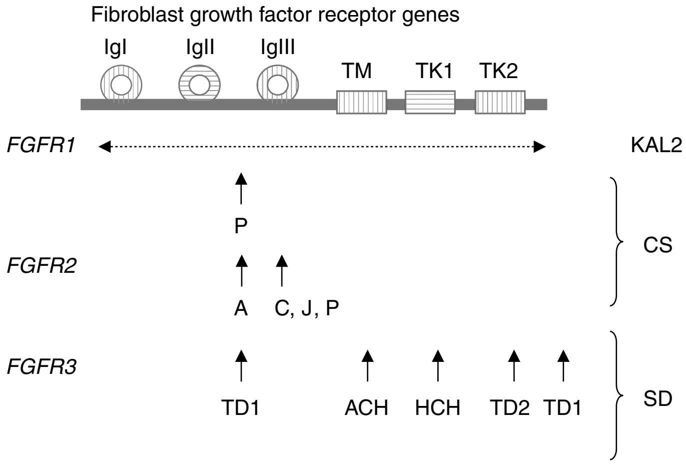


Figure 2 Fibroblast growth factor receptors (FGFR) and their genotype–phenotype correlations. On top, the structure of the FGFR proteins is shown. Arrows indicate the location of mutations with respect to the developmental disorders they cause (craniosynostosis syndromes, skeletal dysplasias, and autosomal dominant Kallmann syndrome). Abbreviations used: KAL2: autosomal dominant Kallmann syndrome; CS: craniosynostosis group; SD: skeletal dysplasia group; P: Pfeiffer; A: Apert; C: Crouzon; J: Jackson–Weiss; TD: thanatophoric dysplasia; ACH: achondroplasia; HCH: hypochondroplasia

1995). In contrast, a dominant form of Kallmann syndrome (KAL2, MIM 147950), characterized by hypogonadotropic hypogonadism and anosmia, results from loss-of-function mutations in *FGFR1* (Dode *et al*., 2003) (Figure 2).

Apert syndrome (MIM 101200), characterized by craniosynostosis and typical syndactyly, is caused by a mutation in one of the adjacent *FGFR2* (MIM 176943) residues that link the second and third immunoglobulin loops. In contrast, mutations in the third immunoglobulin loop cause either Crouzon syndrome (MIM 123500), in which the limbs are normal, or Pfeiffer syndrome (MIM 101600) (Muenke and Schell, 1995) (Figure 2).

In achondroplasia (MIM 100800), the limbs show proximal shortening and the head is enlarged. About 98% of achondroplasia patients carry the mutation 1138G>A (Gly380Arg) and 1% 1138G>C (Gly380Arg) in the transmembrane domain of *FGFR3* (Figure 2). Hypochondroplasia, a milder form of skeletal dysplasia, is caused by mutations in the proximal tyrosine kinase domain of *FGFR3*. Finally, the lethal thanatophoric dysplasia is caused by mutations in either the residues linking the second and third immunoglobulin domain of *FGFR3*, or the distal *FGFR3* tyrosine kinase domain (Muenke and Schell, 1995) (Figure 2).

3. Trinucleotide repeat expansion

A classical example of genotype–phenotype correlations is found in myotonic dystrophy (DM) (MIM 160900), a disorder affecting skeletal and smooth muscle, the eye, heart, endocrine system, and central nervous system. It presents in three

overlapping clinical phenotypes (mild, classical, congenital). The diagnosis is confirmed by detection of a CTG trinucleotide repeat expansion in the 3′ UTR of the *DMPK* gene on 19q13.2-q13.3 (MIM 605377). The number of CTG repeats ranges from 5 to 37 on normal alleles. Individuals with 38–49 CTG repeats do not have disease symptoms, but their children are at risk to have inherited a larger repeat size (anticipation). Persons with a number of CTG repeats >50 to ∼150 are frequently mildly affected, developing cataract and mild myotonia. Patients with repeats between 100 and ∼1000/1500 have classical MD with an age of onset between 10 and 30 years, presenting with weakness, myotonia, cataracts, balding, and cardiac arrhythmia. The average age of death is between 48 and 55 years. In the severe congenital form, neonats present with infantile hypotonia, respiratory deficits, and mental retardation. These patients typically have ∼1000 to >2000 CTG repeats. The average age of death is 45 years (Tapscott and Thornton, 2001). Expansion of the CTG repeat may affect processing of the primary transcript, or may affect expression of a whole series of genes by altering chromatin structure in this gene-rich chromosomal region (Klesert *et al*., 1997).

4. Loss-of-function

4.1. X-linked recessive conditions

Duchenne and Becker muscular dystrophy (DMD/BMD) (MIM 310200/300376) are progressive severe skeletal muscle disorders, caused by mutations in the *DMD* gene located at Xp21.2 (MIM 300377). DMD usually presents in early childhood and is rapidly progressive, with affected children becoming wheelchair-bound by age 12 years. Few patients survive after the third decade. The milder BMD is characterized by later-onset skeletal muscle weakness. The average age of death is after 40. The phenotypes are best correlated with the degree of expression of dystrophin. In general, DMD is caused by frameshift or nonsense mutations leading to a severely truncated dystrophin protein. Frame-neutral mutations, resulting in a shorter than normal dystrophin protein molecule and a residual production of dystrophin, generally cause BMD. More than 70% of disease-causing alleles consist of the deletion or duplication of one or more exons, but also total gene deletions and small intragenic mutations occur (Muntoni *et al*., 2003).

4.2. Haploinsufficient conditions and lack of genotype–phenotype correlations

Loss-of-function mutations leading to a dominant condition are referred to as *haploinsufficient*, prevalent in dosage sensitive genes (e.g., genes coding for transcription factors). Haploinsufficiency conditions often show variable expression and lack of genotype–phenotype correlation, as the changes in gene dosage depend on interactions that are subject to modifications elsewhere in the genome.

Neurofibromatosis type 1 (NF1, MIM 162200) is a progressive autosomal dominant neurocutaneous disorder notorious for its phenotypical intra- and interfamilial

variability. It is caused by mutations in the *NF1* gene on chromosome 17. The extreme allelic heterogeneity, added to the size and complexity of the gene, further complicates the search for genotype–phenotype correlations in this disorder. So far, the only genotype–phenotype correlation identified concerns the whole-gene deletion phenotype, being more severe with more neurofibromas at an earlier age, a lower average IQ, facial dysmorphisms, and an increased risk for the development of malignant peripheral nerve sheath tumors. These findings have led to speculation on modifiers of the haploinsufficient state of the NF1 gene product neurofibromin (Easton *et al*., 1993; Viskochil, 2002).

5. Concluding remarks

Now that many genes causing Mendelian disorders have been identified, the study of genotype–phenotype correlations has moved to center stage. However, with the collection of mutation data in "single-gene" disorders, geneticists have observed that the correlation between genotype and phenotype is often inconsistent or incomplete. The frequent lack of correlations supports the notion that a mutant gene product is part of a complex system in which tissue-specific alternative splicing, intragenic SNPs, epigenetic changes, protein–gene and protein–protein interactions, modifying genes, and environmental factors play a role.

The insight that "simple" Mendelian traits are in fact complex traits has consequences for families and their physicians, and is a challenge for the scientific community. Hence, there is a strong clinical and scientific motivation to identify factors playing a role in genotype–phenotype correlations.

References

Dipple KM and McCabe ER (2000) Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *American Journal of Human Genetics*, **66**, 1729–1735.

Dode C, Levilliers J, Dupont JM, De Paepe A, Le Du N, Soussi-Yanicostas N, Coimbra RS, Delmaghani S, Compain-Nouaille S, Baverel F, *et al*. (2003) Loss-of-function mutations in FGFR1 cause autosomal dominant Kallmann syndrome. *Nature Genetics*, **33**, 463–465.

Easton DF, Ponder MA, Huson SM and Ponder BA (1993) An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): evidence for modifying genes. *American Journal of Human Genetics*, **53**, 305–313.

Edery P, Eng C, Munnich A and Lyonnet S (1997) RET in human development and oncogenesis. *Bioessays*, **19**, 389–395.

Klesert TR, Otten AD, Bird TD and Tapscott SJ (1997) Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of DMAHP. *Nature Genetics*, **16**, 402–406.

Muenke M and Schell U. (1995) Fibroblast-growth-factor receptor mutations in human skeletal disorders. *Trends in Genetics*, **11**, 308–313.

Muntoni F, Torelli S and Ferlini A (2003) Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet Neurology*, **2**, 731–740.

Tapscott SJ and Thornton CA (2001) Biomedicine. Reconstructing myotonic dystrophy. *Science*, **293**, 816–817.

Viskochil D (2002) Genetics of neurofibromatosis 1 and the NF1 gene. *Journal of Child Neurology*, **17**, 562–570; discussion 571–572, 646–651.

Genetic counseling process

Gretchen H. Schneider
Harvard Medical School-Partners Healthcare Center for Genetics and Genomics, Boston, MA, USA

1. Introduction

Genetic counseling was first defined in 1975 when an ad hoc committee of the American Society of Human Genetics described it as a "communication process which deals with the human problems associated with the occurrence or risk of occurrence of a genetic disorder in a family". The committee went on further to explain that genetic counseling includes helping a patient or family (American Society of Human Genetics, 1975):

1. understand a diagnosis, its likely course of action, and treatment possibilities;
2. appreciate the hereditary nature and the risks to family members;
3. be aware of the options for dealing with risks involved;
4. choose an appropriate course of action;
5. make the best adjustment to the disorder or risk to family members.

This original explanation has remained a valid description and is still widely disseminated. It is important to recognize, however, that the genetic counseling process often begins before a diagnosis or precise risk is determined, and thus also encompasses the steps involved in the collection and assessment of information relevant to the questions being addressed in the genetic counseling session.

Genetic counseling is often aimed at answering a question such as: What is wrong with me (my child)? How did this happen? Is my pregnancy at risk for a specific disease? Should I have genetic testing? Am I at risk for cancer due to my family history?

2. Information collection

The first step in answering such a question is gathering information so that precise and accurate predictions can be made. This should include a complete pregnancy, medical, and family history (*see* Article 78, **Genetic family history, pedigree analysis, and risk assessment**, Volume 2) on the patient(s). Information from medical records on the patient or family members may be necessary to clarify

or verify what is obtained during the session. A physical examination by a clinical geneticist (*see* Article 79, **The physical examination in clinical genetics**, Volume 2) may also be an important means of assessing whether evidence of a genetic disease is present. In some cases, review of medical literature in databases, textbooks, or journals will allow comparison of findings in the patient to what has been described in other individuals. Finally, additional testing or evaluations by other specialists might also be warranted and provide additional data pertinent to the patient's assessment.

3. Assessment

Synthesis of information gathered as part of the genetic evaluation will hopefully allow for either a diagnosis to be made or a calculation of risk for a patient or family. This will, in turn, facilitate the discussion of its implications as part of the genetic counseling process. The accuracy of any assessment is dependent on many factors. In the case of a diagnosis, those made based on clinical findings (for example, a child who has features suggestive of Bardet Biedl syndrome) are more dependent on clinical interpretation than one confirmed by a genetic test (for example, the identification of two *CFTR* mutations in someone with clinical features of cystic fibrosis). In situations in which the assessment entails providing an estimate of risk, it is crucial that the risk is derived using precise information that may include, but not be limited to, accurate history, confirmation of diagnoses in family members, or solid data from the literature. This will allow for provision of numbers that truly reflect the risk to the patient or family. If an assessment is not based on complete or correct information, then this may result in decisions being made based on erroneous data.

4. Discussion

As stated in the formal definition of genetic counseling above, a detailed explanation of what is determined in the assessment is a critical part of the genetic counseling process. When a diagnosis has been made, this should include: what the disease is, how the diagnosis was established, how a disorder manifests over a period of time, what interventions are available for affected individuals, what the underlying hereditary nature is, and whether other family members are at risk of having the same disorder. If a genetic risk is being discussed, the information provided should encompass: what the risk is, how this risk was determined, whether there are limitations to the risk estimate (for example, is it an empiric number from the literature that represents an average), and what types of options are available for reducing the risk (such as prenatal diagnosis in future pregnancies).

Given both the complexities and the potentially sensitive nature of genetic information, it is important that discussions be done in a respectful manner. Explanations should be in a language that the patients or families can understand, and efforts should be made to determine the extent to which they comprehend what is discussed. Care should be taken to give patients time to absorb the information and ask questions. If possible and when appropriate, children should be excluded from

the discussion so that parents can focus on what is being said and ask questions openly without worry of information being misunderstood by children. Finally, written documentation should be sent to the patient or family after the genetic counseling visit for them to have as a summary.

5. Decision making

In many cases, particularly in the prenatal setting, genetic counseling for patients or families culminates in their having information on which to base decisions. Should we pursue carrier testing? How can we reduce the recurrence risk for future children? Do we want prenatal diagnosis? Therefore, an important component of a genetic counseling session includes discussion of what choices a patient or family has based on the information that has been provided to them. In some situations, for example, a clinical diagnosis for which there is no genetic testing and more than one inheritance pattern, testing other pregnancies is not an option. For others, however, the possibility of reliable carrier testing, accurate prenatal diagnosis, or alternatives such as egg or sperm donation, adoption, or preimplantation genetic diagnosis may be available and should be presented to families as ways of dealing with the risks quoted to them.

It is crucial to recognize that the decision-making process in a genetic counseling session can be different for different persons. One couple in which the woman is of advanced maternal age may clearly want an amniocentesis, whereas for others in the same situation, the decision may not be as straightforward. One couple may choose to have carrier testing for cystic fibrosis based on ethnicity, while others would decline. In any situation, the discussion should be unbiased and in-depth to include a review of all possible options (testing vs. not testing, and the limitations of both), testing outcomes (abnormal, normal, inconclusive) and subsequent actions (testing options for future pregnancies or termination or continuation, in the case of an ongoing pregnancy). For those having difficulty in choosing a course of action, it can be helpful to have patients go through the different scenarios so that they can sort out their feelings and best select their next step. Even having done this, though, some couples are left having difficulty in making a decision.

Genetic counseling has traditionally been defined as nondirective, where the genetic counselor provides information so the patients can make the best decision for their own particular situation. The goal is not to tell the patients what to do, but to encourage patient autonomy so that he or she is able to select an appropriate course of action. Recent discussions in the field, however, now question nondirectiveness as the central feature of genetic counseling, claiming nondirectiveness is ill-suited in many situations (Biesecker, 2003; Kessler, 1997; Weil, 2003). Regardless, the objectives in any genetic counseling remain unchanged: providing information, facilitating decision making, and offering support in a way that is customized to every individual situation.

6. Support

Genetic counseling would be inadequate without psychological assessment and support for patients or families going through the process. While the extent to which

this is needed varies by situation and by family, most situations call for some type of intervention at some point during the patient's or family's coming to terms with a genetic disease or risk. This may include ongoing assessment of their ability to cope with their situation and acknowledgment and validation of the feelings that they are experiencing. For those who are continually struggling, the genetic counselor may be called on to identify additional resources for therapeutic interventions that lay outside the scope of their expertise. Other families may be in need of information on a disease-specific foundation or support group or be interested in speaking with other families struggling with the same issues. Provision of these services is part of the ongoing nature of the genetic counseling process. Finally, merely remaining available as a source of information or to address questions or concerns can be of great help to patients and families.

7. Conclusion

As research on the human genome continues to unravel the molecular basis of genetic disease, genetic counseling will only grow in importance as well as become more complicated. Advanced technologies will translate to better testing options than those that now exist. Knowledge of previously poorly defined diseases will allow us to better diagnose, manage, and assess risk for other family members. Perhaps most importantly, increased understanding of the underlying genetic basis for common disorders will likely result in a more widespread use of genetic testing to determine predisposition to many diseases. This promises to not only change the field of genetics but also revolutionize all areas of medicine.

References

American Society of Human Genetics, Ad Hoc Committee on Genetic Counseling (1975) Genetic counseling. *American Journal of Human Genetics*, **27**, 240–242.

Biesecker BB (2003) Back to the future of genetic counseling: commentary on "psychosocial genetic counseling in the post-nondirective ear". *Journal of Genetic Counseling*, **12**, 213–217.

Kessler S (1997) Psychological aspects of genetic counseling. XI. Nondirectiveness revisited. *American Journal of Medical Genetics*, **72**, 164–171.

Weil J (2003) Psychosocial genetic counseling in the post-nondirective era: a point of view. *Journal of Genetic Counseling*, **12**, 199–211.

Treatment of monogenic disorders

Maria Descartes and Joseph R. Biggio Jr
University of Alabama at Birmingham, Birmingham, AL, USA

1. Introduction

Although molecular genetic technology has yet to realize its potential for the large-scale cure or treatment of disease, it has provided important new avenues of therapy for single gene disorders by replacing or restoring the function of defective proteins or by minimizing the consequences of the protein deficiency (Nussbaum *et al*., 2001). According to a 1985 study of inborn metabolic errors, the monogenic disorders most frequently targeted for such therapy, only 12% of patients demonstrated a complete response to such treatment, with partial benefits seen in 40% and no response at all in 48% (Hayes *et al*., 1985). In a 1997 study, the rate of complete response remained stable at 12% but the rates for partial or nonresponsiveness changed to 54% and 34% respectively (Treacy *et al*., 2001). Though the stagnant complete response rate signals that complete cure of monogenic disorders has remained an elusive goal, the higher proportion of partial responses marks real progress in controlling and reducing the symptoms associated with them. (Data from Journal of Medicine, **www.wiley.co.uk/genmed/clinical**, updated January 2004.)

Two main treatment options exist for monogenic disorders. First, symptoms may be treated as they occur, with alterations to the milieu to minimize future symptom occurrence. Second, the functioning of the defective protein can be enhanced or the defective protein can be bypassed or replaced altogether.

Changes intended to reduce the incidence of symptoms may be accomplished pharmacologically, surgically, or environmentally. For example, certain medications may be avoided that precipitate porphyric crises, displaced extremities in those with distal arthrogryposis may be splinted or surgically corrected, and smoke-filled environments may be avoided by those with cystic fibrosis.

2. Dietary modification

For inborn metabolic errors, the archetype for monogenic disorders, treatment of this type is targeted at restricting dietary intake, increasing excretion of problematic substances, providing deficient substances, or altering the primary metabolic rate

(Scriver and Treacy, 1999). For example, mental retardation in phenyketonuria patients can be avoided or made less severe by restricting the dietary intake of phenylalanine, an essential amino acid. Dietary restriction has also been used to successfully treat urea cycle disorders, several organic acidemias, and maple syrup urine disease. However, the limited understanding of the metabolic pathways involved in disease progression still poses a barrier to effective treatment, as can be seen in the partial response to treatment by patients with galactosemia; despite a reduction in the occurrence of cataracts and mental retardation, females with this disorder invariably experience premature ovarian failure (Guerrero, 2000).

3. Alternative elimination

For disorders characterized by accumulation of a toxic precursor or by-product, excretion of the offending substance is the preferred therapeutic method. Alternative pathways can be activated, allowing the harmful substance to be converted to a benign form that can be excreted. Pharmacologic agents such as sodium benzoate, phenylbutyrate, or phenylacetate can be used in patients with urea cycle disorders to promote nitrogen elimination and avoid the toxic accumulation of ammonium ion. Additional clearance mechanisms have been employed in disorders characterized by a failure to excrete excess amounts of substrate. Chelating therapy with penicillamine has been used to increase excretion of copper in Wilson disease, and serial phlebotomy has been preferred over chelation to remove the excess iron that characterizes hemochromatosis.

4. Metabolite replacement

Replacement of a deficient metabolic product can also be an effective therapeutic tool in certain disorders. For example, the number of hypoglycemic episodes suffered by patients with Type I and Type III glycogen storage disorders can be reduced by encouraging them to eat frequently and to ingest cornstarch, a slowly digested glucose polymer that provides a sustained glucose source. Cholesterol supplementation has provided at least some benefit to patients with Smith–Lemli–Opitz syndrome, a disorder due to impaired cholesterol biosynthesis (Irons, 1997). Therapeutic measures of this type have also been applied to many endocrine disorders. The defects responsible for congenital adrenal hyperplasia are the prototype of this type of therapy; patients deficient in 21-hydroxylase, as well as other enzymes in this pathway, do not produce sufficient cortisol to trigger feedback inhibition, leading to increased levels of adrenocorticotrophic hormone (ACTH). Because of the metabolic block, cortisol precursors are shunted to the androgen production pathway and result in masculinization. This overproduction of sex steroids and the resulting masculinization have been successfully prevented both pre- and postnatally by replacing the deficient glucocorticoid using pharmacologic agents such as dexamethasone (Pang *et al*., 1990).

5. Enzymatic blockade

In some inborn metabolic disorders, the detrimental effects are due not to the defect in the primary metabolic pathway that results in the lack of a needed product but to a metabolite produced by an alternative pathway. In such cases, therapy has focused on inhibiting the normal pathway at an early stage to prevent the enhancement of the alternative pathways. For example, in Type I tyrosinemia, the lack of the enzyme fumarylacetoacetate hydrolase results in the accumulation of fumarylacetoacetate and maleylacetoacetate, which are then metabolized via an alternative pathway to succinylacetone, a toxic metabolite responsible for many of the neurologic symptoms of this disorder. Treatment with NTBC (2-(2-nitro-4-trifluoromethylbenzoyl)-1,3-cyclohexanedione) results in the inhibition of hydroxy-phenyl pyruvate dioxygenase and the blockade of the metabolic pathway at an early stage, leading to dramatic decreases in succinylacetone levels. Because this therapy leads to an accumulation of tyrosine, dietary restriction of tyrosine and phenylalanine are typically instituted jointly with NTBC therapy (Lock, 1998). Allopurinol therapy has similarly been used for the prevention of uric acid accumulation in patients with Lesch–Nyhan syndrome. Normal processes of feedback inhibition can also be exploited to turn off enzyme systems that result in accumulation of toxic substances. In acute intermittent porphyria, hematin therapy during an acute porphyric crisis decreases the activity of δ-aminolevulinic acid synthetase and thereby reduces porphyrin production (Watson *et al*., 1978).

6. Enzyme potentiation

As an alternative to the milieu-altering treatments designed to reduce symptoms outlined above, the functioning of the defective protein can be enhanced by facilitating its binding to a vitamin cofactor. Such binding is known to be a key step in the activation and regulation of many metabolic pathways, including those involved in congenital lactic acidosis, organic acidemias, and homocystinuria. For example, the activity of the enzyme pyruvate dehydrogenase is enhanced by the cofactor thiamine. Indeed, an altered form of the enzyme, which shows a decreased affinity for thiamine, is observed in conditions characterized by a deficiency of the enzyme, such as congenital lactic acidosis. In some patients suffering from this disorder, supra-pharmacologic doses of thiamine have increased the enzymatic activity by forcing thiamine binding (Naito *et al*., 1994). Similarly, excess biotin has been used successfully for holocarboxylase deficiency, excess thiamine for homocystinuria due to cystathionine β-synthetase deficiency, and excess vitamin B12 for some types of methylmalonic acidemia (Morrone, 2002).

7. Enzymatic bypass

Alternatively, the defective protein can be bypassed. Because an adequate supply of cofactors for further enzymatic reactions is dependent upon recycling reactions,

enzymatic defects leading to inadequate recycling can deplete cofactor stores and disrupt metabolism. Some of these defective enzymes do not play a significant role in other metabolic pathways and so cofactors can simply be replaced under certain conditions. For example, administration of pharmacologic doses of biotin can virtually reverse the depletion of biotin seen when biotinidase, the enzyme responsible for the removal of the cofactor biotin from the active form of a carboxylase enzyme, is deficient. Likewise, high doses of folic acid can bypass the insufficient recycling caused by defects in folic acid metabolism and prevent the hypomethioninemia and hyperhomocystinemia resulting from depleted folate stores and decreased methyl donors.

8. Protein replacement

Finally, the entire defective protein can be replaced. Protein replacement can be accomplished, at least theoretically, via three different strategies: (1) actual administration of the wild-type/normally functioning protein, (2) transplantation of cells or tissues that produce the normal protein, or (3) gene therapy in which only the affected gene is introduced via a vector.

In the first strategy, a normal protein is administered to replace the defective one. Such protein replacement has thus far been applied successfully only to a small number of conditions because the administered protein must reach its site of normal activity in order to achieve normal function. That requirement may be formidable when proteins need to cross the blood–brain barrier to function in the central nervous system. Thus, excellent clinical responses to enzyme replacement therapy have been reported for some lysosomal storage disorders such as nonneuronopathic Gaucher disease and Fabry disease, but only limited responses have been seen in neuronopathic Gaucher disease and Hurler syndrome, because the accumulated toxic metabolites within the nervous system remain relatively inaccessible to the infused protein (Brady *et al*., 2001). In addition to problems of delivery to the normal physiologic compartment, the stability, pharmacokinetics, and antigenicity of the administered medications must be considered (Barranger and O'Rourke, 2001). In many disorders, because the mutant protein may be missing key antigenic sites or have an altered conformation, the normal protein, when introduced, may be recognized by the immune system as "nonself" or foreign and trigger an immune response, thereby not only increasing the risk of infusion reactions but also decreasing the stability of the protein and necessitating an escalation of dose in order to maintain therapeutic results (Kakkis *et al*., 2001).

The second strategy for protein replacement seeks to bypass the repeated administrations required by protein infusion therapy by exploiting the more long-lasting benefits provided by tissue and organ transplantation. Both solid organ transplants, such as liver transplants for ornithine transcarbamylase deficiency, and bone marrow transplants, such as those for Hurler syndrome, have been performed to provide a sustained source of the previously deficient protein. While both have been able to provide sufficient quantities of protein to ameliorate many symptoms, the protein, like the endogenously produced protein, must still penetrate all the physiologic

compartments necessary to arrive at its locale of normal activity. Because the transit of the blood–brain barrier can be problematic for large proteins and because transplantation of tissue directly into the central nervous system is not feasible, the utility of bone marrow and organ transplantation in the treatment of lysosomal storage disorders associated with the accumulation of toxic metabolites in the central nervous system remains uncertain. For example, while bone marrow transplants have resulted in an improvement in hepatosplenomegaly and cardiac function in patients with Hurler syndrome, the neurologic symptoms often remain unabated (Peters *et al*., 1998).

One potential tool for the treatment of monogenic disorders is a modality that seeks to replace the defective protein by reactivation of silent or poorly expressed genes. The end result of this "transcriptional therapy" is an increase in the number of mRNA signals produced from the target genes. In vitro and in vivo studies have demonstrated reactivation of several genes by "transcriptional therapy", including ALDPL1, SMN2, and fetal hemoglobin, but, with the exception of the induction of gamma globin chain expression by hydroxyurea, these techniques have yet to be applied clinically (Chiurazzi and Neri, 2003; Bunn, 1997).

Third, gene therapy has been investigated as a means of providing a renewable protein source. Disorders resulting from a simple deficiency of a specific gene product are generally the most amenable to treatment because even low-level expression of an introduced normal allele should be sufficient to overcome the deficiency. While, in principle, this approach seems straightforward, it has encountered the by now familiar obstacle of protein accessibility to diseased tissue as well as problems in the regulation of gene expression. In addition, the efficacy of such treatment requires not only a wild-type form of the protein but also the presence in any transfected tissue of all cofactors and enzymes necessary for the production of mature functional protein.

Gene therapy has been used for monogenic disorders only in the absence of other therapeutic options or, as is the case with severe combined immunodeficiency syndrome, when the disorder is severe enough to warrant the risks. The development of effective gene therapies for protein replacement has been hampered by the attendant risks, which include an adverse reaction to the vector or transferred gene, the potential induction of a mutation in the patient's germ line, and the potential integration of the transferred gene into the patient's DNA, resulting in activation of a proto-oncogene, disruption of a tumor-suppressor gene, or the disruption of an otherwise normal, essential gene.

In summary, the tremendous progress made over the last quarter of a century in understanding the pathophysiologic basis of many monogenic disorders has facilitated the development of more effective palliative therapies, but the quest for the ultimate therapy utilizing molecular techniques to replace or repair the defective gene remains ongoing.

References

Barranger JA and O'Rourke E (2001) Lessons learned from the development of enzyme therapy for Gaucher disease. *Journal of Inherited Metabolic Disease*, **24**(Suppl 2), 89–96.

Brady RO, Murray GJ, Moore DF and Schiffman R (2001) Enzyme replacement therapy for Fabry disease. *Journal of Inherited Metabolic Disease*, **24**(Suppl 2), 18–24.

Bunn HF (1997) Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine*, **337**(11), 762–769.

Chiurazzi P and Neri G (2003) Reactivation of silenced genes and transcriptional therapy. *Cytogenetic and Genome Research*, **100**(1-4), 56–64.

Gene Therapy Clinical Trials World Wide. Provided by the *Journal of Gene Medicine*, **www.wiley.co.uk/genmed/clinical**, John Wiley & Sons: New Jersey. Updated January 31, 2004.

Guerrero NV, Singh RH, Manatunga A, Berry GT, Steiner RD and Elsas LJ (2000) Risk factors for premature ovarian failure in females with galactosemia. *The Journal of Pediatrics*, **137**(6), 833–841.

Hayes A, Costa T, Scriver CR and Childs B (1985) The effect of mendelian disease on human health II: Response to treatment. *American Journal of Medical Genetics*, **21**, 243–255.

Irons M, Elias ER, Abuelo D, Bull MJ, Greene CL, Johnson VP, Kepper L, Schanen C, Tint GS and Salen G (1997) Treatment of Smith-Lemli-Opitz syndrome: results of a multicenter trial. *American Journal of Medical Genetics*, **68**(3), 311–314.

Kakkis ED, Muenzer J, Tiller GE, Waber L, Belmont J, Passage M, Izykowski B, Phillips J, Doroshow R, Walot I, *et al*. (2001) Enzyme-replacement therapy in mucopolyssachoaridosis I. *New England Journal of Medicine*, **344**(3), 182–188.

Lock EA, Ellis MK, Gaskin P, Robinson M, Auton TR, Provan WM, Smith LL, Prisbylla MP, Mutter LC and Lee DL (1998) From toxicological problem to therapeutic use: the discovery of the mode of action of 2-(2-nitro-4-trifluoromethylbenzoyl)-1, 3-cyclohexanedione (NTBC), its toxicology and development as a dry. *Journal of Inherited Metabolic Disease*, **21**(5), 498–506.

Morrone A, Malvagia S, Donati MA, Funghini S, Ciani F, Pela I, Boneh A, Peter H, Pasquini E and Zammarchi E (2002) Clinical findings and biochemical and molecular analysis of four patients with holocarboxylase synthetase deficiency. *American Journal of Medical Genetics*, **111**(1), 10–18.

Naito E, Ito M, Takeda E, Yokota I, Yoshijima S and Kuroda Y (1994) Molecular analysis of abnormal pyruvate dehydrogenase in a patient with thiamine-responsive lactic acidemia. *Pediatric Research*, **36**(3), 340–346.

Nussbaum RL, McInnes RR and Willard HF (2001) *Thompson & Thompson Genetics in Medicine*, Sixth Edition, W.B. Saunders Company: Philadelphia.

Pang SY, Pollack MS, Marshall RN and Immken L (1990) Prenatal treatment of congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *New England Journal of Medicine*, **322**(2), 111–115.

Peters C, Shapiro EG, Anderson J, Henslee-Downey PJ, Klemperer MR, Cowan MJ, Saunders EF, de Alarcon PA, Twist C, Machman JB, *et al*., The Storage Disease Collaborative Study Group (1998) Hurler syndrome:II. Outcome of HLA-genotypically identical sibling and HLA-haploidentical related donor bone marrow transplantation in fifty-four children. *Blood*, **91**(7), 2601–2608.

Scriver CR and Treacy EP (1999) Is there treatment for "genetic" disease? *Molecular Genetics and Metabolism*, **68**, 93–102.

Treacy EP, Valle D and Scriver CR (2001) Treatment of genetic disease. In *The Metabolic and Molecular Bases of Inherited Disease*, Eighth Edition, Scriver CR, Beaudet AL, Sly WS and Valle D (Eds.), McGraw-Hill: New York.

Watson CJ, Pierach CA, Bossenmaier I and Cardinal R (1978) Use of hematin in the acute attack of the "inducible" hepatic porphyrias. *Advances in Internal Medicine*, **23**, 265–286.

Carrier screening: a tutorial

Wayne W. Grody
David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Jean A. Amos
Molecular Genetics Specialty Laboratories, Valencia, CA, USA

1. Obstetrical visit and rationale for carrier testing

John and Judy Brown are seeing their obstetrician for a routine prenatal visit; Judy is now 11 weeks pregnant. In taking the family history (Figure 1), the obstetrician notes that John had a brother who died of cystic fibrosis (CF) many years ago in childhood. His condition was characterized by failure to thrive, short stature, malabsorption, and chronic obstructive lung disease. John and Judy are both non-Jewish Caucasians of European descent. Unlike John, Judy has no family history of this disease.

The obstetrician explains that current practice dictates that all pregnant couples be offered CF carrier screening, and that this is even more important when there is a positive family history. The objective of this program is to identify couples at risk so that they can then be offered prenatal diagnosis.

Although it would be helpful to first identify the familial mutations in the affected index case, this is not possible because John's brother is long deceased, and there are no stored tissue specimens; moreover, John initially declined to ask his parents, both of whom are obligate carriers of a familial mutation, to participate in his carrier study. The reason this approach is advantageous is because of the large number of mutations in the *CFTR* gene (over 1300 reported to date). Identification of the mutations in John's brother would allow for more targeted testing of John, as opposed to a generic screening panel. If John were found not to carry either of his brother's mutations, then his carrier risk would be reduced to near zero (assuming proper paternity), and no further testing would be indicated. In contrast, a negative result of screening for a limited subset of mutations in John in the absence of any knowledge of his brother's mutations would still leave him with an appreciable risk of being a carrier (see below) simply because of his family history and our inability to test for all possible mutations in the gene.

Since the familial mutations are not known a priori in this case, both John and Judy are offered screening using the standard panel of 23 mutations as recommended by the American College of Medical Genetics and the American College of Obstetricians and Gynecologists (Grody *et al*., 2001; Watson *et al*., 2004). They

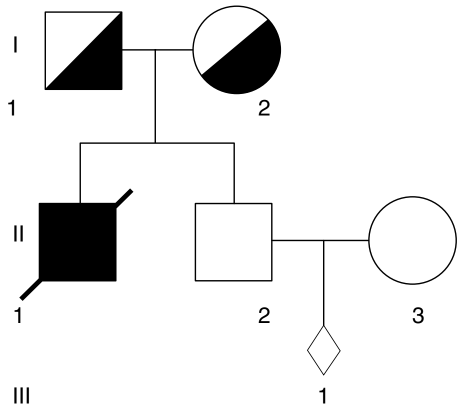


Figure 1 Pedigree of the Brown family indicating John's deceased brother (II-1) who was affected with cystic fibrosis, and Judy's pregnancy (III-1)

consent to proceed, and blood specimens are drawn from both and sent to the Molecular Genetics Laboratory for testing. In two weeks, the couple returns to the clinic to receive their test results and undergo further counseling.

2. Mutation analysis

The prior risk that John is a CF carrier, based on his positive family history, is 2/3. Judy's prior risk is based on the frequency of CF carriers in the Caucasian population, 1/25, since she has no family history of CF. The risk that their fetus has inherited CF is $2/3 \times 1/25 \times 1/4 = 1/150$.

CF mutation analysis is negative for both members of this couple, but the interpretation of these results are very different for each of them, based on their different family histories. The revised CF carrier risk after a negative mutation analysis is based on the prior risk of the individual tested and the rate of detection of carriers in their ethnic group. For both John and Judy, both non-Jewish Caucasians of European descent, this detection rate is 90%. Using Bayesian risk analysis, the testing laboratory reports to the obstetrician that the revised carrier risks for John and Judy are 1/6 and 1/241, respectively, and that the risk to their fetus is 1/5784. The laboratory also encourages the obstetrician to refer this couple for genetic counseling and to request testing of John's parents to identify whether the familial CF mutation(s) are included on the testing panel. On the basis of this recommendation, the obstetrician refers John and Judy to a genetic counselor at the teaching hospital in their city.

3. Genetic counseling and additional mutation analysis

The genetic counselor explains to John and Judy that their negative mutation analyses have reduced the risk of CF to their pregnancy but that identification of the familial mutation might further reduce John's carrier risk and the subsequent risk

of CF to their fetus. John arranges for his parents to submit a blood sample each, and the testing laboratory determines that both are carriers of the most common CF mutation, ΔF508. The laboratory then states that, based on knowledge of the familial mutation, John's revised carrier risk is essentially zero, as is the revised risk to their fetus. John and Judy are reassured by these results and his parents feel gratified that they have contributed to the knowledge that their grandchild's risk to inherit the disease that devastated their own son is extremely low.

4. Comment

In some cases, a couple in which one member has a positive family history elects only to have the partner with the negative family history screened for CF mutations. For the Brown family, this approach would have led to the 1 in 241 carrier risk revision for Judy, and a fetal risk of $2/3 \times 1/241 \times 1/4 = 1$ in 1440, a reduction of almost fivefold from the prior risk. In the authors' experience, many couples find this risk reduction acceptable and do not seek further testing. However, we have seen two couples who have used this approach and have had affected children. Subsequent testing revealed that these affected babies inherited a common mutation from the parent with the affected relative and a rare mutation not included in any US clinical testing panel from the negative family history parent.

Further reading

Richards CS, Bradley LA, Amos J, Allitto B, Grody WW, Maddalena A, McGinnis MJ, Prior TW, Popovich BW, Watson MS, *et al*. (2002) Standards and guidelines for CFTR mutation testing. *Genetics in Medicine*, **4**, 379–391. Erratum in: *Genetics in Medicine*, **4**, 471 (2002).

Richards CS and Grody WW (2004) Prenatal screening for cystic fibrosis: past, present and future. *Expert Review of Molecular Diagnostics*, **4**, 49–62.

Watson MS, Desnick RJ, Grody WW, Mennuti MT, Popovich BW and Richards CS (2002) Cystic fibrosis carrier screening: issues in implementation. *Genetics in Medicine*, **4**, 407–409.

References

Grody W, Cutting G, Klinger K, Richards CS, Watson M and Desnick R (2001) Laboratory standards and guidelines for population-based cystic fibrosis carrier screening. *Genetics in Medicine*, **3**, 149–154.

Watson MS, Cutting GR, Desnick RJ, Driscoll DA, Klinger K, Mennuti M, Palomaki GE, Popovich BW, Pratt VM, Rohlfs E, *et al*. (2004) Cystic fibrosis carrier screening: 2004 revision of the American College of Medical Genetics mutation panel. *Genetics in Medicine*, **6**, 387–391.

Prenatal aneuploidy screening

Katharine D. Wenstrom
University of Alabama at Birmingham School of Medicine, Birmingham, AL, USA

1. Case

Ms FG, a 28-year old white woman currently pregnant at 18 weeks' gestation, is referred for evaluation because an ultrasound exam of her fetus indicated the presence of intracranial cysts. She denied any family history of birth defects, heritable diseases, recurrent pregnancy loss, or infertility, and reported that her pregnancy has been otherwise uncomplicated.

2. Background

Prior to the 1980s, the screening test for fetal aneuploidy was a question: "Will you be age 35 or older when your baby is born?" The basis for this question was the fact that the risk of fetal trisomy, such as Down syndrome or trisomy 18, increases along with maternal age. At age 35, the risk of aneuploidy roughly equals the risk of procedure-related rupture of the membranes leading to pregnancy loss, thus justifying invasive fetal testing (American College of Obstetricians and Gynecologists, 1987). Women who were aged 35 or older were offered definitive fetal diagnosis via either traditional amniocentesis (performed between 14 and 22 weeks' gestation) or chorionic villous sampling (performed between 10 and 13 6/7 weeks). The goal of fetal diagnosis is to give the woman and her partner autonomy; some parents would terminate an affected pregnancy, whereas others would use fetal diagnosis to plan their delivery. For example, foreknowledge that the fetus has trisomy 18, a lethal disorder, would allow the woman to avoid a needless Cesarean delivery for fetal stress in labor (which is common in trisomy 18 pregnancies), since it would not change the neonatal outcome. Foreknowledge of fetal Down syndrome with a cardiac defect would allow delivery conditions to be optimized, for example, by planning delivery at a tertiary center with a pediatric cardiologist available.

A major advance in aneuploidy screening occurred in the 1980s, when it was discovered that the levels of certain maternal serum analytes are altered when the fetus has Down syndrome or trisomy 18 (Haddow *et al*., 1992; Haddow *et al*., 1994). A variety of analytes have been investigated, but the best performing analytes in current use are first-trimester levels of PAPP-A, estriol, AFP, and free β hCG, or second-trimester levels of hCG, estriol, AFP, and inhibin (Wald *et al*.,

1998; Cuckle, 2001). These analytes can be combined with the maternal age-related Down syndrome risk to create either a first- or second-trimester *multiple marker screening test*. For each analyte, the level measured in the woman's serum is compared to the levels found in both normal and Down syndrome pregnancies at the same gestational age. This comparison allows determination of the *relative risk* of fetal Down syndrome associated with each woman's unique analyte level. The individual relative risks associated with each analyte are then combined to create a composite relative risk, and the composite risk is used to modify the woman's maternal age-related Down syndrome risk. Women whose final estimated risk of Down syndrome is above a predetermined cut off – usually 1:200 or 1:270 – are offered definitive (invasive) fetal testing. If all screen-positive women undergo definitive fetal testing, most of the first- and second-trimester multiple marker tests in current use have a 70–75% Down syndrome detection rate at a 5% false-positive rate (Wald *et al*., 1998; Cuckle, 2001). Using a similar protocol, 75–80% of trisomy 18 fetuses are also detected, at a screen-positive rate of 2% (Palomaki *et al*., 1995).

Obstetrical ultrasound equipment has also improved significantly since the 1980s, and currently available machines allow fetal structures to be seen in great detail. Both first- and second-trimester ultrasound exams can now be performed as *sonographic* screening tests for fetal Down syndrome, to identify women at sufficiently increased risk to justify invasive fetal testing. An experienced sonographer can identify anatomic features indicating an increased risk of aneuploidy in up to 87% of Down syndrome fetuses, with a false-positive rate of 13% (Vintzileos *et al*., 1997; Vintzileos *et al*., 1996). Ultrasound also allows detection of 75% of trisomy 18 fetuses; 25% have isolated choroid plexus cysts, and 50% have choroid plexus cysts along with other dysmorphisms or structural anomalies. The false-positive rate for isolated choroid plexus cysts is approximately 0.5% (Gupta *et al*., 1995). Because the majority of fetuses evaluated sonographically are euploid, the negative predictive value of ultrasound for Down syndrome detection is 99.7% (Vintzileos *et al*., 1996). On the other hand, because a proportion of Down syndrome fetuses appear structurally normal on ultrasound, the positive predictive value of ultrasound as a screening test is only 19.4% (Vintzileos *et al*., 1996).

The "genetic ultrasound exam" includes a complete survey of all major anatomic structures as well as evaluation for the presence of dysmorphisms that have been associated with fetal aneuploidy. Such dysmorphisms include choroid plexus cysts, increased nuchal translucency (first trimester) or nuchal thickness (second trimester), presence and size of the nasal bone, an echogenic cardiac focus, renal pelviectasis, echogenic bowel, slightly shortened femur or humerus, hypoplastic middle phalanx of the little finger, and sandal foot (a space between the first and second toes). One aspect of the fetal exam, the measurement of the nuchal translucency or nuchal thickness, requires additional specialized training as well as ongoing monitoring of measurement technique to insure correct assessment (D'Alton *et al*., 2003). The presence of a major structural malformation increases the risk of fetal aneuploidy sufficiently to justify invasive fetal testing; depending on the maternal age, the presence of one or more dysmorphisms may also sufficiently increase the risk. Table 1 lists some common anomalies along with their associated risks of aneuploidy, and Table 2 lists the dysmorphisms most frequently associated with fetal Down syndrome.

Table 1 Malformations associated with fetal aneuploidy

Structural defect	Aneuploidy risk (%)	Most common aneuploidy (trisomy)
Cystic hygroma	60–75	45X (80%); 21,18,13,XXY
Hydrops	30–80	13,21,18,45X
Hydrocephalus	3–8	13,18, triploidy
Holoprosencephaly	40–60	13,18,18p-
Cardiac defects	5–30	21,18,13,22,8,9
Diaphragmatic hernia	20–25	13,18,21,45X
Omphalocele	30–40	13,18
Duodenal atresia	20–30	21
Bladder outlet obstruction	20–25	13,18
Facial cleft	1	13,18, deletions
Limb reduction	8	18
Club foot	6	18,13,4p-,18q-

Reprinted from the American College of Obstetricians and Gynecologists Technical Bulletein. © American College of Obstetricians and Gynecolgists.

Table 2 Sonographic dysmorphisms associated with fetal Down syndrome[c]

Ultrasound marker	Incidence ($N = 420$)	Positive predictive value
Structural anomalies (including cardiac)	17 (4%)	7[a]/17 (41.1%)
Short femur	18 (4.3%)	4[a]/18 (22.2%)
Short humerus	16 (3.8%)	7[a]/16 (44%)
Pyelactasis	20 (4.7%)	4[b]/20 (20%)
Nuchal fold thickening $\geq$ 6 mm	15 (3.6%)	9[a]/15 (60%)
Echogenic bowel	4 (1%)	0/4
Choroid plexus cysts	14 (3.3%)	0/14
Hypoplastic/absent mid-phalanx 5th digit	13 (3.1%)	2[a]/13 (15.4%)
Wide space, 1st–2nd toes	4 (1%)	1[a]/4 (25%)
2-vessel umbilical cord	3 (0.7%)	0/3

[a] All cases were associated with additional ultrasound markers.
[b] Three of four cases were associated with additional ultrasound markers.
[c] Reprinted from American Journal of Obstetrics and Gynecology, V87, Vintzileos AM *et al*. The use of second-trimester genetic sonogram in guiding clinical management of patients at increased risk for fetal trisomy 21, pp. 948–952. ©1996, with permission from Elsevier.

Finally, a first-trimester screening test that combines maternal serum analytes with an ultrasound marker has been developed. When evaluated according to a specified protocol, the first-trimester nuchal translucency measurement can be converted to a multiple of the median and used to derive a relative risk; this risk can then be combined with the relative risks associated with PAPP-A and free β hCG, and the ultimate composite risk used to modify the woman's age-related risk. According to the largest multicenter trial, which included more than 33 000 women, this first-trimester test has a Down syndrome detection rate of 80% when the screen-positive rate is held at 3.4%, or 94% when the screen-positive rate is allowed to increase to 10.8% (Malone *et al*., 2003).

3. Approach to this case

A targeted ultrasound exam should be performed, and the patient should be offered the multiple marker screening test; a positive multiple marker screening test should be followed by counseling and consideration of a definitive fetal diagnostic test (Figure 1). In this case, the multiple marker test indicated no increased risk, and

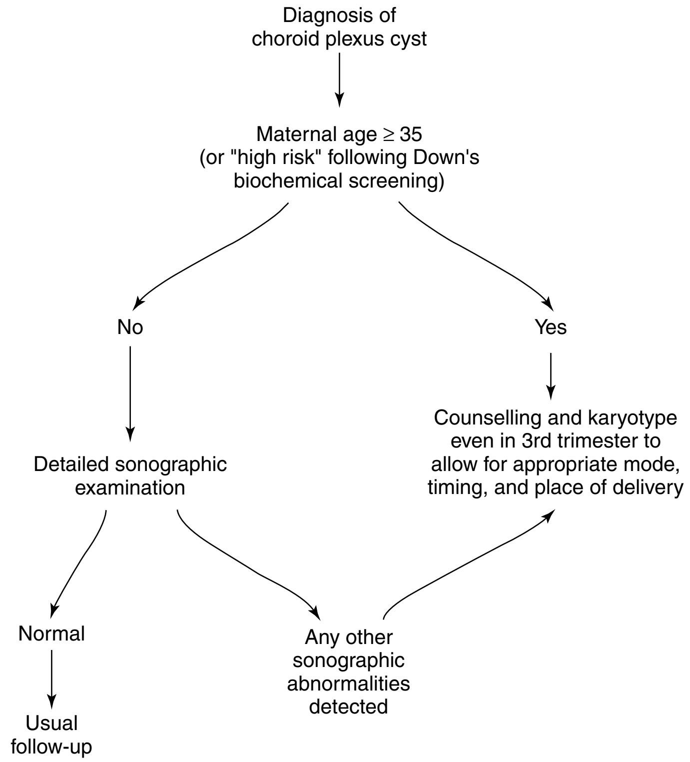


Figure 1 Management plan for suspected choroid plexus cysts (adapted from Gupta JK, Cave M, Lilford RJ, Farrell TA, Irving HC, Mason G, Hau CM (1995) Clinical significance of fetal choroid plexus cysts. *Lancet*, **345**, 724–29)

the targeted ultrasound exam confirmed the presence of choroid plexus cysts. These cysts typically form within the choroid plexus of the lateral ventricles sometime during the first trimester. They usually do not deform or disrupt associated structures, and in most cases resolve completely before delivery. They are of interest mainly because they are found in 4.3% of fetuses with trisomy 18 (Gupta and Lilford, 1997). However, they are also present in 0.47% of euploid fetuses (Gupta and Lilford, 1997). The fetus should therefore be carefully examined to determine if there are any other sonographic features of trisomy 18, such as a ventricular septal defect or another cardiac abnormality, a renal anomaly, clenched fists with the second and fifth fingers overlapping the third and fourth, micrognathia, omphalocele, or rocker bottom feet. In this case, the rest of the fetal exam was normal.

The ultrasound data can then be used to further refine the patient's risk of trisomy 18, using Bayes theorem (Gupta and Lilford, 1997):

$$\text{Posterior risk} = \text{likelihood ratio} \times \text{prior risk} \tag{1}$$

In this case, the likelihood ratio is the relative risk of trisomy 18 associated with isolated choroid plexus cysts. These cysts are found in 4.3% of fetuses with trisomy 18 and in 0.47% of euploid fetuses; the likelihood ratio is thus 4.3/0.47, or 9.04. The prior risk is the maternal age–related risk of trisomy 18 at age 28, or 1:3351. The posterior risk associated with isolated fetal choroid plexus cysts is thus 1:3351 × 9 = 1:371 (The presence of additional fetal abnormalities would have increased her risk by a factor of 1800, to approximately 1:2). Since the most frequently quoted post-procedure loss rate after amniocentesis is 1:200 (American College of Obstetricians and Gynecologists, 1987), the patient's final risk of 1:371 indicates that a procedure-related problem is more likely than fetal trisomy 18. Most patients would decline amniocentesis after considering these relative risks. However, each patient also assigns her own "relative risk", reflecting her personal values, to procedure-related pregnancy loss as well as the birth of a child with trisomy 18. As a result, some women may request amniocentesis even after counseling regarding the relative risk indicated by the ultrasound exam.

References

American College of Obstetricians and Gynecologists (1987) *Antenatal Diagnosis of Genetic Disorders*, ACOG Technical Bulletin #108, American College of Obstetricians and Gynecologists: Washington, pp. 1–8.

Cuckle HS (2001) Time for total shift to first-trimester screening for Down's syndrome. *Lancet*, **358**, 1658–1659.

D'Alton ME, Malone FD, Lambert-Messerlian G, Ball RH, Nyberg DA, Comstock CH, *et al*. (2003) Maintaining quality assurance for nuchal translucency sonography in a prospective multicenter study: Results from the FASTER Trial. *American Journal of Obstetrics and Gynecology*, **189**, S79.

Gupta JK, Cave M, Lilford RJ, Farrell TA, Irving HC, Mason G and Hau CM (1995) Clinical significance of fetal choroid plexus cysts. *Lancet*, **345**, 724–729.

Gupta JK and Lilford RJ (1997) Management of fetal choroid plexus cysts. *British Journal of Obstetrics and Gynaecology*, **104**, 881–886.

Haddow JE, Palomaki GE, Knight GJ, Cunningham GC, Lustig LS and Boyd PA (1994) Reducing the need for amniocentesis in women 35 years of age or older with serum markers for screening. *The New England Journal of Medicine*, **330**, 1114–1118.

Haddow JE, Palomaki GE, Knight GJ, Williams J, Pulkkinen A, Canick JA, Saller DN Jr. and Barsel Bowers G (1992) Prenatal screening for Down syndrome with use of maternal serum markers. *The New England Journal of Medicine*, **3**, 588–593.

Malone FD, Wald NJ, Canick JA, Ball RH, Nyberg DA, Comstock CH, *et al*. (2003) First and second-trimester evaluation of risk (FASTER) trial: principal results of the NICHD multicenter Down syndrome screening study. *American Journal of Obstetrics and Gynecology*, **189**, S56.

Marchese CA, Carozzi F and Mosso R (1985) Fetal karyotype in malformations detected by ultrasound. *American Journal of Human Genetics*, **37**, A223.

Palomaki GE, Haddow JE, Knight GJ, Wald JN, Kennard A, Canick JA, Saller DN, Blitzer MG, Dickerman LH, Fisher R, *et al*. (1995) Risk-based prenatal screening for trisomy 18 using alpha-fetoprotein, unconjugated oestriol and human chorionic gonadotropin. *Prenatal Diagnosis*, **15**, 713–723.

Vintzileos AM, Campbell WA, Guzman ER, Smulian JC, McLean DA and Anath CV (1997) Second-trimester ultrasound markers for detection of trisomy 21: Which markers are best? *Obstetrics and Gynecology*, **89**, 941–944.

Vintzileos AM, Campbell WA, Rodis JF, Guzman ER, Smulian JC and Knuppel RA (1996) The use of second-trimester genetic sonogram in guiding clinical management of patients at increased risk for fetal trisomy 21. *Obstetrics and Gynecology*, **87**, 948–952.

Wald NJ, Kennard A, Hackshaw A and McGuire A (1998) Antenatal screening for Down's syndrome. *Health Technology Assessment*, **2**(1), 1–112.

Waldimiroff JW, Sachs ES and Reuss A (1998) Prenatal diagnosis of chromosome abnormalities in the presence of fetal structural defects. *American Journal of Medical Genetics*, **28**, 289.

Williamson RA, Weiner CP, Patil S, *et al*. (1987) Abnormal pregnancy sonogram: selective indication for fetal karyotype. *Obstetrics and Gynecology*, **69**, 15.

Gene identification in common disorders: a tutorial

Mark O. Goodarzi and Jerome I. Rotter
Cedars-Sinai Medical Center, Los Angeles, CA, USA

Elucidation of the genetic determinants of common disorders has proven much more challenging than the genetics of rare conditions (King *et al*., 2002). Whereas a rare genetic disorder is typically caused by a loss- or gain-of-function mutation in one gene that has a dramatic effect on phenotype (for example, the *HUNTINGTIN* gene in Huntington's Disease), common disorders (such as diabetes mellitus, coronary artery disease, or migraine headache) are thought to arise from the complex interaction of variants in several genes and environmental factors (*see* Article 58, **Concept of complex trait genetics**, Volume 2). Each of the several genes contributing to a complex disorder is thought to have only a mild effect on phenotype; it is the combined effect of predisposing genes that contributes to disease. This presents great challenges in terms of identifying these genes; large study populations may be needed to provide enough power to detect genes with a moderate effect on phenotype (*see* Article 68, **Approach to common chronic disorders of adulthood**, Volume 2). In this tutorial, we describe general principles related to gene identification in common disorders.

The first stage is the selection of a disorder that appears to be genetically determined. This is not always straightforward, as common environmental exposures may be misinterpreted as genetic underpinnings. Examination of relatives of probands with a particular disorder is often used to determine whether that disorder has a genetic component; if the disorder occurs at higher frequency in family members than in the general population (familial aggregation), this is evidence for genetic determinants.

The next challenge is to recruit a large number of subjects, on the basis of the specific study design (Figure 1). A case-control study requires a significant number of subjects with the condition and a number of subjects without the condition. Controls must be matched as closely to the cases as possible, differing only in the absence of the condition under study. If the controls differ significantly from the cases in other aspects, for example, ethnic makeup or body composition, then genetic differences between the two may be related to these other factors, a phenomenon referred to as *population stratification*. Family-based studies enroll families based on a proband with the condition of interest. Population stratification is less of an issue when healthy relatives are studied as controls. Often, multicenter collaborations are necessary to successfully recruit a large number of subjects for study.

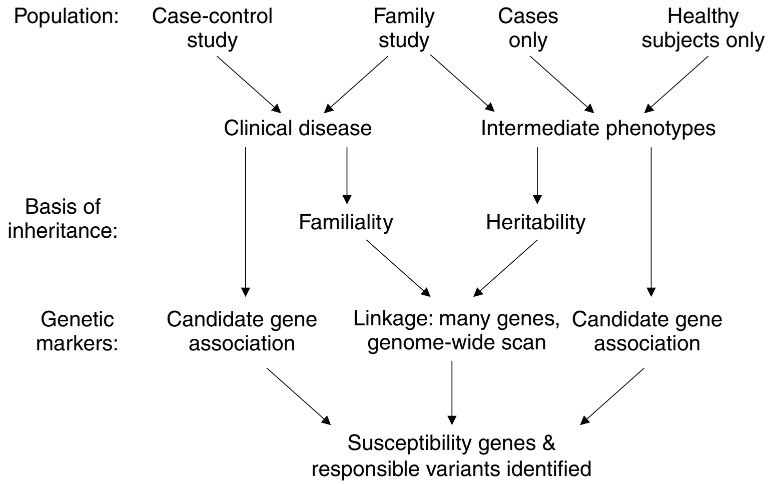


Figure 1 Flowchart illustrating steps in the identification of genes for common disorders

If one has access to only a large group of subjects with a common disorder, genetic studies can be carried out that examine the inheritance of phenotypes that characterize the disorder. For example, diabetes mellitus is characterized by abnormal levels of insulin. If one has only a large number of subjects with diabetes, but no control group, then insulin, rather than the presence or absence of diabetes, can be used as the phenotype in a genetic study. Phenotypes that characterize a disorder are known as component or intermediate phenotypes (also subclinical markers). Intermediate phenotypes can also be studied in subjects who have no overt evidence of disease. In this instance, intermediate phenotypes are thought to be closer to predisposing genes as they may represent the earliest stages of the condition.

Characterization of the traits of the study subjects, that is, phenotyping, is an important feature of genetic studies. At times, one seeks to ask whether a disorder is present or absent. In this case, a precise, reproducible definition of the disorder must be used. Some disorders are difficult to define; in such cases, a definition that is not strict may allow the inclusion of subjects with conditions that resemble each other but do not share genetic determinants. For example, genetic studies of migraine headaches must be careful to exclude other headache disorders. When intermediate phenotypes are to be studied, a balance must be found between the cost of phenotyping and the sophistication of the phenotyping. For example, insulin resistance, a metabolic condition in which the hormone insulin is ineffective in removing glucose from the bloodstream, can be quantified in a number of different ways. Fasting glucose may be used for this purpose and has the advantage of low cost and ease of measurement. On the other hand, insulin resistance may also be quantified by physiologic studies (such as the euglycemic clamp) that involve several hours in a research laboratory for each subject. These are time-consuming, expensive, and technically difficult; however, they yield very reliable measurements. In fact, insulin resistance quantified by physiologic study was found to have a higher heritability than insulin resistance measured by simpler measures (Bergman *et al*.,

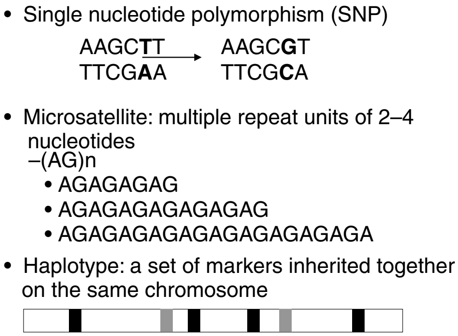


Figure 2 Genetic markers used in studies of common disorders

2003). Thus, the greater difficulty of conducting detailed physiologic phenotyping may yield great benefit in terms of traits that are more genetically determined. This was shown in a study of the lipoprotein lipase (*LPL*) gene, wherein variation in the gene was associated with insulin resistance quantified by a detailed physiologic study, but not with simpler indices of insulin resistance (Goodarzi *et al*., 2004).

Given that we often start with no knowledge of genetic variants that lead to a disease, we must take advantage of markers in the genome. Markers, such as microsatellites and single nucleotide polymorphisms (SNPs, Figure 2), are polymorphic variants interspersed throughout the genome. They are used as tags to track disease-causing variants or mutations. The underlying principle is that markers that are close to disease-causing variants tend to be inherited on the same chromosomes (*see* Article 59, **The common disease common variant concept**, Volume 2). Linkage refers to the situation wherein markers in a region of the genome are inherited in a nonrandom fashion in relation to a particular phenotype. Association refers to the situation wherein a particular allele of a marker is found with greater frequency in those with a particular phenotype. Linkage scans of the entire genome are often carried out using a panel of microsatellites that cover the whole genome. Whole-genome association studies are being contemplated, as genotyping technology advances make such an effort feasible.

A relatively new tool being utilized in the study of common disorders is the haplotype, which is a collection of marker alleles inherited together on the same chromosome. Haplotypes span large regions on the genome, on an average of 10 to 20 kb (Gabriel *et al*., 2002). Haplotypes reflect the global gene structure, encompassing chromosomal blocks that have remained unbroken by recombination during the population history of the gene. The use of haplotypes may be more likely to identify disease-variation associations than is the use of a random single marker. The identification of a haplotype associated with increased or decreased disease risk should facilitate the identification of the actual functional variant that affects disease risk, because this variant should lie on chromosomes identified by that haplotype (Figure 3). A study that examined SNPs within the *LPL* gene did not find any association with coronary artery disease; however, when these SNPs were organized into haplotypes, association with coronary artery disease was evident (Goodarzi *et al*., 2003).

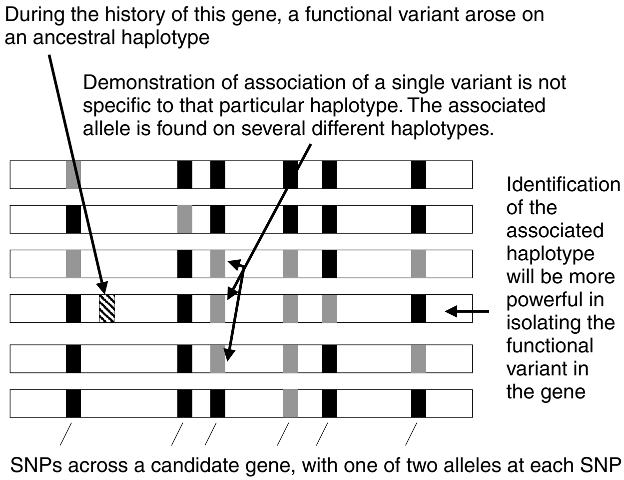


Figure 3 Benefits of using haplotypes in the isolation of genes for common disorders

Related articles

Article 58, **Concept of complex trait genetics**, Volume 2; Article 59, **The common disease common variant concept**, Volume 2; Article 68, **Approach to common chronic disorders of adulthood**, Volume 2

References

Bergman RN, Zaccaro DJ, Watanabe RM, Haffner SM, Saad MF, Norris JM, Wagenknecht LE, Hokanson JE, Rotter JI and Rich SS (2003) Minimal model-based insulin sensitivity has greater heritability and a different genetic basis than homeostasis model assessment or fasting insulin. *Diabetes*, **52**, 2168–2174.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Goodarzi MO, Guo X, Taylor KD, Quinones MJ, Saad MF, Yang H, Hsueh WA and Rotter JI (2004) Lipoprotein lipase is a gene for insulin resistance in Mexican Americans. *Diabetes*, **53**, 214–220.

Goodarzi MO, Guo X, Taylor KD, Quiñones MJ, Samayoa C, Yang H, Saad MF, Palotie A, Krauss RM, Hsueh WA, *et al*. (2003) Determination and use of haplotypes: ethnic comparison and association of the lipoprotein lipase gene and coronary artery disease in Mexican-Americans. *Genetics in Medicine*, **5**, 322–327.

King RA, Rotter JI and Motulsky AG (2002) *The Genetic Basis of Common Diseases*, Oxford University Press: New York.

Basic Techniques and Approaches

Uses of databases

Roberta A. Pagon
University of Washington, Seattle, WA, USA

1. Introduction

Databases are used by clinicians practicing genetic medicine and clinical genetics to determine the molecular basis of inherited disorders, establish diagnoses, identify sources and uses of molecular genetic testing, assess and manage inherited cancer risk, provide consumer-health-oriented information to patients, and identify clinical genetic services for geographically dispersed family members. In this chapter, only those databases that are freely available on the Internet and, hence at the point of care, are discussed. Physician use of Medline (**www.ncbi.nlm.nih.gov/pubmed**), a widely used database in the practice of medicine, is assumed.

2. Molecular basis of inherited disorders

OMIM (Online Mendelian Inheritance in Man) (**www.ncbi.nlm.nih.gov/omim**) is funded by the NIH, authored and edited by Dr. Victor A. McKusick and an editorial team at The John Hopkins University, and distributed by the National Library of Medicine (Hamosh, 2002; Maglott *et al*., 2002). OMIM is a timely, authoritative compendium of bibliographic material and observations on inherited disorders, human genes, and gene loci. In May 2004, OMIM had over 15 000 records, including information on over 9000 loci. OMIM is continuously updated with information abstracted from the medical literature; closely integrated with Medline and the NCBI genomic databases; and has extensive links to other genomic databases and resources. OMIM comprises the following:

- *MIM entries:* Descriptive, full-text entries that include approved gene name and symbol, alternative names and symbols in common use, and a loosely structured text description of the disease or gene.
- *OMIM Gene Map:* A tabular listing of the genes and loci represented in MIM ordered pter to qter from chromosome 1-22, X, and Y. The information in the map includes the cytogenetic location, symbol, title, MIM number, method of mapping, comments, and associated disorders and their MIM numbers.
- *OMIM Morbid Map:* Alphabetical listing of disease genes in the OMIM Gene Map organized disease.

- *Clinical synopses:* List of clinical findings in a disease MIM entry by organ system.
- *Mini-MIM:* A text synopsis of an MIM entry.

3. Diagnostic tools

OMIM can be used in a limited way for syndrome identification because of its sophisticated search capabilities by word or phrase that may be a disease name, symptom, or clinical finding. OMIM search results provide a list of entries ranked by the closest match.

4. Sources and uses of molecular genetic testing

GeneTests (**www.genetests.org**), funded by the NIH and developed at the University of Washington, Seattle (Pagon *et al*., 2002), is composed of:

- *Laboratory Directory:* Listing of contact information and test methods for international clinical and research medical genetics laboratories testing for inherited disorders using molecular genetic tests, specialized cytogenetic tests, or biochemical genetic tests. In May 2004, ~600 laboratories testing for ~1050 diseases (~700 clinical; ~350 research only) were listed. Listings are revised as needed and updated every 2 years.
- *GeneReviews:* Disease descriptions that focus on the use of currently available genetic testing in patient diagnosis, management, and genetic counseling authored and peer-reviewed by experts from around the world. Each entry is highly structured and navigable by a Table of Contents. In May 2004, 245 *GeneReviews* were posted; one new one is added each week. Entries are revised as needed and formally updated every 2 years.

The GeneTests hierarchical naming system is used to clarify the evolving understanding of relationships between genes and phenotypes. GeneTests "Disease" search results are displayed as a parent–child hierarchy that relates disease names that refer to a change in a gene ("gene-related" name) to names that refer to a phenotype ("phenotype-related" name). When the parent is a gene-related name, all children must be phenotype-related names; conversely, when the parent is a phenotype-related name, all children must be a gene-related name. This naming

Parent (gene-related name): Severe chronic Neutropenia [Testing] [*Reviews*]
  Child (phenotype-related name): Congenital Neutropenia
  Child (phenotype-related name): Cyclic Neutropenia

Figure 1 Alteration in one gene associated with two phenotypes

Parent (phenotype-related name): Hypohidrotic Ectodermal Dysplasia [*Reviews*]
Child (gene-related name): Hypohidrotic Ectodermal Dysplasia, Autosomal [Testing]
Child (gene-related name): Hypohidrotic Ectodermal Dysplasia, X-linked [Testing]

Figure 2 One phenotype associated with alteration in one of the two genes

system allows clinically available testing ("Testing" button) to be associated with gene-related names only. Examples of search results are given in Figures 1 and 2.

5. Cancer risk assessment and management

PDQ Cancer Genetics section (**www.cancer.gov/cancerinfo/pdq/genetics**) of the Cancer Information portion of the National Cancer Institute website provides a Cancer Genetics Overview, a discussion of the Elements of Cancer Genetics Risk Assessment and Counseling, and summaries of evidence-based information about the genetic basis of breast and ovarian cancer, colorectal cancer, medullary thyroid cancer, and prostate cancer.

6. Consumer-health-oriented information

1. The Genetic and Rare Conditions site, University of Kansas Medical Center (**www.kumc.edu/gec/support/**), maintained by Debra Collins, Mississippi. Over 550 pages with links to lay advocacy and support groups, information on over 200 genetic conditions/birth defects for professionals, educators, and individuals, links to sites for children and teens.
2. Family Village: A Global Community of Disability-Related Resources (**www.familyvillage.wisc.edu/index.html**), provides both disease-specific information as well as general information directed at parents of individuals who have disabilities. This site has general resources such as communication, adaptive products and technology, recreational activities, education, worship, and disability-related media and literature.
3. Genetic Home Reference (**http://ghr.nlm.nih.gov**), developed by the National Library of Medicine of the NIH, contains descriptions of ~100 genetic conditions and the genes responsible for those conditions, links to resources and patient support information for those disorders, and general educational materials and a glossary.
4. GeneTests (**www.genetests.org**) provides disease-specific consumer-health-oriented resources accessed within a *GeneReview* or the "Resources" button displayed in a disease search result.
5. Genetic Alliance (**www.geneticalliance.org**), a consumer-health-oriented organization of more than 600 advocacy, research, and health care organizations that maintains a directory of support groups that can be searched by genetic condition, organization, and services offered.

7. Clinical genetics service providers (Table 1)

Table 1 Clinical genetics service providers

Organization	Directory	Search parameters
GeneTests: Clinic Directory **www.genetests.org**	>1000 US genetics clinics	Zip code, state, city, services, specialty clinics
American Board of Genetic Counseling, American Board of Medical Genetics, American College of Medical Genetics, American Society of Human Genetics	Clinical geneticists, laboratory directors, genetic counselors, nurses in genetics, researchers	Alphabetical, geographical
National Society of Genetic Counselors **www.nsgc.org/resourcelink.org**	Genetic counselors	Location, name, specialty, zip code

Related articles

Article 69, **Current approaches to prenatal screening and diagnosis**, Volume 2; Article 72, **Current approaches to molecular diagnosis**, Volume 2; Article 74, **Molecular dysmorphology**, Volume 2; Article 80, **Genetic testing and genotype–phenotype correlations**, Volume 2; Article 81, **Genetic counseling process**, Volume 2; Article 82, **Treatment of monogenic disorders**, Volume 2; Article 83, **Carrier screening: a tutorial**, Volume 2; Article 84, **Prenatal aneuploidy screening**, Volume 2; Article 87, **The microdeletion syndromes**, Volume 2; Article 88, **Cancer genetics**, Volume 2; Article 89, **Familial adenomatous polyposis**, Volume 2

Further reading

Guttmacher AE (2001) Human genetics on the web. *Annual Review of Genomics and Human Genetics*, **2**, 213–233.

References

Hamosh A (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **30**, 52–55.

Maglott D, Amberger JS and Hamosh A (2002) Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders. **www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/chtd1.pdf**.

Pagon RA, Tarczy-Hornoch P, Covington ML, Baskin PK, Edwards JE, Espeseth M, Beahler C, Bird TD, Popovich B, Nesbitt C, *et al*. (2002) Genetests and geneclinics: genetic testing information for a growing audience. *Human Mutation*, **19**, 501–509.

The microdeletion syndromes

Jodi D. Hoffman and Elaine H. Zackai
The Children's Hospital of Philadelphia, Philadelphia, PA, USA

1. Case presentation

A 3-year-old white girl presents to your clinic for evaluation of multiple problems. She recently moved from another state, and her new pediatrician wonders if there might be a genetic etiology for her medical issues.

She was the second child to her 31-year-old gravida 2 para 2 mother after a pregnancy complicated by the finding of an interrupted aortic arch on fetal ultrasound. After the discovery of the cardiac defect, an amniocentesis was performed that showed a normal 46,XX karyotype. The child was born at 38 weeks via spontaneous vaginal delivery at a university hospital with a weight in the 25th percentile and height and head circumference in the 10th percentile. She was stabilized in the NICU prior to surgery. On chest X ray, she was noted to have thymic aplasia. Her cardiac defect was repaired without complication, other than multiple episodes of hypocalcemia, which resolved with medical management. She was discharged at 4 weeks of age. Since her discharge, she has had multiple ear infections, gastroesophageal reflux with weight loss requiring treatment with a proton pump inhibitor, has been noted to have short stature, and has a delayed speech compared to her older sister at similar age. Her parents are slightly concerned about her gross motor milestones and enrolled her in a local early intervention program.

On physical examination, she is noted to be a shy, petite child who does not strongly resemble either parent. She says several words and you note a hypernasal tone to her voice. Her weight is 10th percentile, height just below the 5th percentile, and head is 10th percentile. Her face appears slightly long. Her eyes are deeply set and hooded. Her nose has a bulbous tip. Her ears are slightly low set and the helices are overfolded. She has a chest scar from her cardiac surgery. Her chest and abdominal exam are unremarkable other than the presence of a small umbilical hernia. Her fingers are long-appearing and tapered.

In summary, you have a short 3-year-old female with a history of repaired interrupted aortic arch, resolved hypocalcemia, multiple ear infections, gastroesophageal reflux, delayed hypernasal speech, dysmorphic features, who had a previously normal karyotype. How should you approach this case?

2. Discussion

This child has multiple system involvement (cardiac, endocrine, immunologic, gastrointestinal) and dysmorphia. What should one consider in children who have

differences in several apparently unrelated systems? In taking a birth and pregnancy history, it is important to ask about medication and drug exposures (alcohol, valproate, warfarin, or other teratogens) that could lead to multisystem involvement. If there are no obvious exposures, next it is important to assess whether any in utero infections (CMV, parvovirus, rubella, etc.) could be the etiology of the child's differences. If the infectious work-up is also unrevealing, chromosomal abnormalities should be considered.

The constellation of multiple congenital abnormalities and dysmorphic features that occur together repeatedly in unrelated patients is referred to as a syndrome. Syndromes can be due to a multitude of etiologies including chromosomal deletions, duplications, and rearrangements, uniparental disomy, gene deletions, single gene mutations (autosomal recessive, autosomal dominant, X-linked), triplet repeat expansions, and imprinting.

In the case described above, the child had an amniocentesis in utero due to the finding of a cardiac defect. The karyotype was a normal (46,XX), decreasing the chance that the cause of this child's abnormalities was a large chromosomal deletion, duplication, or rearrangement. Because the study was performed on amniocytes, which cannot be analyzed at as detailed a level as lymphocytes, there is still a small possibility that a cytogenetically detectable abnormality was missed.

In order to understand this child's findings, the next question to ask is: Has this combination of findings ever been seen before? The patient has a cardiac defect, thymic aplasia, and hypocalcemia; cardinal features of classic DiGeorge syndrome (DiGeorge, 1965). A small number of patients were found to have translocations in a common area of the long arm of chromosome 22 (de La Chapelle *et al*., 1981; Kelley *et al*., 1982). Later, 25% of patients were found to have cytogenetically visible interstitial deletions in the 22q11 region.

In 1991, a method now referred to as FISH (fluorescence in situ hybridization) became available to look for noncytogenetically apparent chromosomal changes such as microdeletions (Lichter *et al*., 1991; Trask, 1991). With FISH, a fluorescent probe is specifically designed for a given chromosomal region. The probe is then used to look for the presence or absence of two intact copies of a critical region known to be deleted in those with a syndrome. In 1992, it was found that the vast majority of patients with DiGeorge syndrome who did not have a cytogenetically visible abnormality had a microdeletion of approximately 3 million base pairs on the long arm of chromosome 22 (Driscoll, 1992). Microdeletions are described as "contiguous-gene" deletions (Emanuel, 1988), which may lead to syndromes with multiple systems involvement.

This patient has hypernasal speech and tapered fingers, features described in the velocardial facial syndrome (VCFS) (Shprintzen *et al*., 1978). This syndrome as well as the conotruncal anomaly face syndrome (CFS) and some cases of Opitz G/BBB syndrome were all eventually found to have the same etiology – a microdeletion of chromosome region 22q11.2. Owing to the ability to identify a unifying etiology for these overlapping phenotypes, many people now group DiGeorge syndrome, VCFS, and CFS together as the 22q11.2 deletion syndrome.

Although in 10% of cases the 22q11.2 deletion syndrome is transmitted in a familial fashion, it is most often seen sporadically, with a high prevalence, approximately 1 in 3000 live births. Recently, it has been found that chromosomal

Table 1 Microdeletion syndromes

Syndrome	Deletion	Classic features
DiGeorge/VCFS/CFS	22q11.2	Congenital heart defects, dysmorphic features, immune abnormalities, hypocalcemia, developmental delay
Williams Syndrome	7q11.23	Supravalvular aortic stenosis, hoarse voice, prominent lips, mental retardation, hypercalcemia
Smith Magenis Syndrome	17p11.2	Self destructive behavior, speech delay, mental retardation, brachycephaly, midfacial hypoplasia
Langer–Giedion Syndrome	8q24.11-13	Multiple exostoses, redundant skin, bulbous nose, prominent ears, mental retardation
Prader Willi syndrome	15q11-13	Hypotonia, severe obesity, small hands and feet, mental retardation
Angelman syndrome	15q11-13	Severe mental retardation, absence of speech, paroxysms of laughter, ataxic gait, seizures, characteristic facies
Miller–Dieker Syndrome	17p13.3	Lissencephaly, microcephaly, vertical forehead ridging, mental retardation, initial hypotonia, failure to thrive, seizures
WAGR Syndrome	11p13	Wilms tumor, aniridia, genitourinary abnormalities, mental retardation

regions that are repeatedly deleted in multiple patients are often flanked by low-copy repeats that make the DNA in the area unstable and more susceptible to deletions, duplications, and rearrangements (Emanuel and Shaihk, 2001).

The 22q11.2 syndrome as well as other microdeletion syndromes (Table 1) are most often not visible cytogenetically. As in this case, the signs and symptoms in an individual patient may lead to a specific syndrome diagnosis for which there is now the availability of FISH on a clinical diagnostic basis. These FISH tests must be ordered specifically.

Further reading

Jones KL (1997) *Smith's Recognizable Patterns of Human Malformation*, W.B. Saunders.

References

de La Chapelle A, Herva R, Koivisto M and Aulo O (1981) A deletion in chromosome 22 can cause Digeorge syndrome. *Human Genetics*, **57**, 253–256.

DiGeorge A (1965) Discussion on a new concept of the cellular basis of immunology. *The Journal of Pediatrics*, **67**, 907.

Driscoll DA, Budarf ML and Emanuel BS (1992) A genetic etiology for DiGeorge syndrome: consistent deletions and microdeletions of 22q11. *American Journal of Human Genetics*, **50**, 924–933.

Emanuel BS (1988) Molecular cytogenetics: toward dissection of the contiguous gene syndromes. *American Journal of Human Genetics*, **43**, 575–578.

Emanuel BS and Shaihk TH (2001) Segmental duplications: an expanding role in genomic stability and disease. *Nature Reviews Genetics*, **2**, 791–800.

Kelley RI, Zackai EH, Emanuel BS, Kistenmacher M, Greenberg F and Punnett HH (1982) The association of the DiGeorge anomalad with partial monosomy of chromosome 22. *The Journal of Pediatrics*, **101**, 197–200.

Lichter P, Boyle A, Cremer T and Ward D (1991) Analysis of genes and chromosomes by nonisotopic in situ hydbridization. *Genetic Analysis Techniques and Applications*, **8**, 24–35.

Shprintzen RJ, Goldberg RB, Lewin ML, Sidoti EJ, Berkman MD, Argamaso RV and Young D (1978) A new syndrome involving cleft-palate, cardiac anomalies, typical facies, and learning disabilities: velo-cardio-facial syndrome. *The Cleft Palate Journal*, **15**, 56.

Trask BJ (1991) Fluorescence in situ hybridization: Applications in cytogenetics and gene mapping. *Trends Genetic*, **7**, 149–154.

Cancer genetics

Katherine A. Schneider, Kelly J. Branda and Anu B. Chittenden
Dana-Farber Cancer Institute, Boston, MA, USA

Kristen M. Shannon
Massachusetts General Hospital, Boston, MA, USA

1. Introduction

Over the past 10 years, clinical cancer genetics has become established as an important medical specialty that bridges oncology and genetics. This overview article will provide basic information about cancer susceptibility genes, describe features of 12 hereditary cancer syndromes, and detail the main components of cancer genetic counseling sessions.

2. Basic cancer genetics

At a cellular level, the process of growth and division must occur properly in order to maintain a state of balance in the body. The cell cycle is controlled by complex processes and is highly regulated. Underlying the development of all cancers is the accumulation of mutations in the genes that control the cell cycle. While the protein products of thousands of genes play roles in cell cycle control at different times and in different tissues during a person's life, there are three classes of genes that are highly important in the development of cancers. These gene groups are tumor suppressor genes, DNA repair genes, and proto-oncogenes. Germline mutations in any of the genes in these three groups can lead to hereditary susceptibility to cancer. However, it is important to note that hereditary cancer syndromes only account for a small percentage (5–10%) of the 1.3 million cancers that will be diagnosed in the United States this year (American Cancer Society, 2004).

3. Tumor suppressor genes

Tumor suppressor genes serve as the security branch of the cell and are the most common cause of hereditary cancer. The protein products of these genes are

involved in recognizing replication errors, aiding in DNA repair and facilitating cell suicide (apoptosis) if necessary (Fearon, 1998). Since almost all genes are inherited in pairs, most cells have two copies of a necessary tumor suppressor gene in place to provide a backup system if one copy (allele) stops working because of gene mutation.

One of the best-known tumor suppressor genes is the *Rb* gene. After studying children with sporadic and hereditary forms of retinoblastoma (a childhood eye tumor), Dr. Alfred Knudson proposed his now-famous hypothesis regarding the development of cancer (Knudson, 1993). His "two-hit" model states that two mutagenic events are needed, which knock out both copies of the *Rb* susceptibility gene. Individuals with sporadic retinoblastoma develop eye tumors, because two "hits" (mutations in each *Rb* gene) were acquired somatically in a single retinal cell. However, individuals with the hereditary form of retinoblastoma only required one somatic hit to develop a tumor, because their initial hit was an inherited mutation in one copy of the *Rb* gene. This explained why children with an inherited form of retinoblastoma tended to develop bilateral disease at younger than usual ages.

4. DNA repair genes

The second most common type of cancer susceptibility gene is the DNA repair gene. These genes are involved in identifying and correcting mistakes that occur during the replication process. Mutations in DNA repair genes can lead to the accumulation of mutations in other genes involved in cell cycle control, and thereby lead to overgrowth and cancer (Pearson *et al*., 1998). The most common hereditary form of colorectal cancer, termed hereditary nonpolyposis colorectal cancer (HNPCC), is caused by mutations in one of several DNA mismatch repair genes (see Table 1) (Lucci-Cordisco *et al*., 2003).

5. Proto-oncogenes

Although this occurs less commonly, proto-oncogenes can also be involved in inherited susceptibility to cancer. Proto-oncogenes are genes whose protein products typically activate some process in the cell that switches the cell cycle into the "on" position, triggering cellular growth. Inherited mutations in these genes can lead to the inability of a cell to turn off the growth process (Park, 1998). The most well known example of proto-oncogene mutation leading to hereditary susceptibility to cancer is the *RET* proto-oncogene. Germline *RET* mutations lead to one of three endocrine cancer syndromes, all of which involve hereditary susceptibility to medullary thyroid carcinoma (MTC) (Eng, 1999). It is now considered standard-of-care in medicine to test anyone with a diagnosis of MTC for mutations in the *RET* proto-oncogene. If a mutation is identified, predictive genetic testing of family members can reveal which family members are at risk to develop MTC. Early screening and preventative removal of the thyroid can be life-saving in this situation (Eng, 1999).

Table 1 Clinical features and associated genes of selected hereditary cancer syndromes

Syndrome	Gene (chromosomal location)	Gene type	Inheritance pattern	Clinical features
Cowden Syndrome (CS)	*PTEN* (10q23)	Tumor suppressor	AD	Cancers: thyroid, breast, endometrial, renal cell, melanoma, glioblastoma Other: trichilemmomas, acral keratoses, papillomatous papules, mucosal lesions, macrocephaly, Lhermittte–Duclos disease, thyroid disease, mental retardation, gastrointestinal hamartomas, fibrocystic breasts, lipomas or fibromas, genitourinary tumors, genitourinary malformations
Familial Adenomatous Polyposis (FAP)	*APC* (5q21)	Tumor Suppressor	AD	Cancers: colon, thyroid, ampulla of Vater, bile duct, small intestine, hepatoblastoma, brain tumors Other: desmoid tumors, dental abnormalities, CHRPE, osteomas
Familial Medullary Thyroid Cancer (FMTC)	*RET* (10q11.2)	Proto-oncogene	AD	Cancers: medullary thyroid cancer Other: none
Familial Melanoma (FAMM)	*CMM1* (1p36) *CDKN2A/TP16* (9p21) *CDK4* (12q14)	Tumor suppressor	AD	Cancers: melanoma, astrocytomas, pancreas
Familial Retinoblastoma (RB)	*RB1* (13q14.1)	Tumor suppressor	AD	Cancers: retinoblastoma, osteosarcoma, Ewing sarcoma, leukemia, lymphoma, melanoma, lung cancer, bladder cancer Other: short stature, microcephaly, mental retardation, genital malformations, ear abnormalities
Hereditary Breast/Ovarian Cancer (HBOC)	*BRCA1* (17q21) *BRCA2* (13q12)	Tumor suppressor	AD	Cancers: female breast, male breast, ovary, pancreas, prostate
Hereditary Nonpolyposis Colorectal Cancer (HNPCC)	*MLH1* (3p21.3) *MSH2* (2p22-p21) *MSH6* (2p16) *PMS1* (2q31) *PMS2* (7p22)	Mismatch repair	AD	Cancers: colon, endometrial, stomach, small intestine, ureter, kidney, brain tumors
Li Fraumeni Syndrome (LFS)	*TP53* (17p13.1)	Tumor suppressor	AD	Cancers: osteosarcomas, soft tissue sarcomas, breast cancer, brain tumors, adrenocortical carcinomas, acute leukemias, stomach, colon, lung, nueroblastomas, melanoma

(*continued overleaf*)

Table 1 (*continued*)

Syndrome	Gene (chromosomal location)	Gene type	Inheritance pattern	Clinical features
Multiple Endocrine Neoplasia, type 2A (MEN2A)	*RET* (10q11.2)	Proto-oncogene	AD	Cancers: medullary thyroid Other: pheochromocytomas, parathyroid tumors
Multiple Endocrine Neoplasia, type 2B (MEN2B)	*RET* (10q11.2)	Proto-oncogene	AD	Cancers: medullary thyroid cancer Other: pheochromocytomas, parathyroid tumors, developmental delay, enlarged lips, mucosal neuromas, ganglioneuromatosis of the intestine, Marfanoid habitus
Von Hippel Lindau Disease (VHL)	*VHL* (3p25)	Tumor suppressor	AD	Cancer: renal cell (clear cell type) Other: retinal angiomas, hemangioblastomas of cerebellum and spine, cysts and adenomas of kidney and pancreas, pheochromocytomas
Xeroderma Pigmentosum (XP)	*XPA* (9q34.1) *XPB* (2q21) *XPC* (3p25.1) *XPD* (19q13.2) *XPE* (11p12-p11) *XPF* (16p13.2-p13.1) *XPG* (13q32-q33)		AR	Cancers: basal cell and squamous cell, melanoma, sarcomas, ocular melanoma, brain tumors, lung, gastric, leukemia Other: mental subnormality, microcephaly, sensorineural deafness, hyporeflexia or areflexia, spacitity/ataxia, abnormal EEG

References: Offit, 1998; Schneider, 2002; Genetests, 1993–2004: **http://www.genetests.org** and OMIM, 2000: **http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM**.

6. Hereditary cancer syndromes

Over 200 cancer syndromes have been identified, but most are exceedingly rare. Table 1 lists the major clinical features and underlying genes for 12 hereditary cancer syndromes.

Two of the best-defined inherited cancer syndromes are hereditary breast and ovarian cancer syndrome and hereditary nonpolyposis colorectal cancer. These two syndromes are detailed below.

6.1. Hereditary breast and ovarian cancer syndrome

The majority of cases of hereditary breast and ovarian cancer syndrome (HBOCS) are caused by germline mutations in the *BRCA1* or *BRCA2* gene. These genes are inherited in an autosomal dominant manner, meaning that they can be inherited from the maternal or paternal lineage and that they are passed on to daughters and sons with equal frequency. The incidence of inherited *BRCA1* and *BRCA2* mutations in the general population is estimated to be 1/500 to 1/1000 (Szabo and King, 1997). Founder mutations have been identified in several ethnic groups at greater frequencies. For example, 1 of 40 people of Ashkenazi Jewish ancestry will have one of three founder mutations: 187delAG and 5385 insC mutations in *BRCA1* and 6174delT mutation in *BRCA2* (Struewing *et al*., 1997).

Women with HBOCS have a 50–85% lifetime risk of breast cancer. The risk of developing a second primary cancer in the contralateral breast is thought to be 50–60%. In general, breast cancers due to *BRCA2* mutations occur at older ages than those associated with *BRCA1* mutations. The risk of ovarian cancer is 20–40% for women with *BRCA1* mutations and 15–20% for women with *BRCA2* mutations. Risks of peritoneal cancer and fallopian tube cancer are also increased. Breast cancer risk in men is 6–10% with *BRCA2* mutations and is also thought to be increased, but less so, for men with *BRCA1* mutations. Men with HBOCS also have increased risks for prostate cancer. Individuals with *BRCA2* mutations may also have small, increased risks of melanoma (of the skin and eye) and cancers of the pancreas, bile duct, and stomach (Petrucelli *et al*., 2004).

6.2. Hereditary nonpolyposis colon cancer

Hereditary nonpolyposis colon cancer (HNPCC) is an inherited cancer syndrome that accounts for about 5% of colon cancer diagnoses (Offit, 1998). People with HNPCC have up to an 80% risk of developing colorectal cancer over the course of their lifetime, although this risk is largely preventable with routine screening. These tumors often arise in the right-sided region of the colon, and are typically derived from an adenomatous polyp, although frank polyposis is absent. Women with HNPCC have up to a 40% risk of endometrial (uterine) cancer and a 10–12% risk of ovarian cancer. Other HNPCC malignancies include cancers of the stomach, pancreas, and small intestine, as well as transitional cell cancers of the kidney and bladder. Additional features are present in subtypes of HNPCC. These include sebaceous gland tumors and skin keratocanthomas found in Muir–Torre syndrome and brain tumors (primarily glioblastomas) seen in Turcot syndrome (Kohlmann and Gruber, 2004).

HNPCC is clinically diagnosed using Amsterdam I or II criteria. Amsterdam I criteria states that families must have at least three cases of colorectal cancer in at least two successive generations, two of these cases must occur in first-degree relatives and one of the colon cancers must have been diagnosed at age 50 or younger (Vasen *et al*., 1991). Amsterdam II criteria broadens the definition to include other HNPCC malignancies (Vasen *et al*., 1999).

HNPCC can be caused by germline mutations in several different mismatch repair genes, but about 40–80% of families that meet classic Amsterdam I criteria will have a detectable *MLH1* or *MSH2* mutation (Syngal *et al*., 2000). It is thought that families with variant forms of HNPCC may have specific gene mutations leading to these phenotypes. For example, several families with Turcot syndrome have specific *PMS2* mutations (Hamilton *et al*., 1995).

Nearly 90% of the colon tumors that develop in individuals with HNPCC show a feature known as microsatellite instability (MSI). In contrast, MSI is only seen in about 15% of sporadic colon tumors (Wahlberg *et al*., 2002). Microsatellites are a family of repetitive DNA, which is typically repeated a fixed number of times at any given chromosomal locus. The difference in MSI rates allows this to be a useful screening test to determine the likelihood that a family has HNPCC. The Bethesda criteria can be useful in determining whether individuals should undergo MSI tumor analysis (Rodriguez-Bigas *et al*., 1997; Boland *et al*., 1998).

7. Cancer genetic counseling

Cancer genetic counseling is a communication process concerning an individual's risks of developing specific inherited forms of cancer. This risk may be higher than or similar to the general population risks of cancer. Counseling can, but does not always, lead to genetic testing. For additional information about genetic counseling services and providers, please refer to Article 75, **Changing paradigms of genetic counseling**, Volume 2.

8. Genetic counseling sessions

Patients are often referred to a cancer genetic counselor by their oncologists or other health care providers. However, an increasing number of patients are self-referred. Genetic counseling sessions involve collecting the family history, providing an assessment of the risk of certain cancers and the likelihood of having an inherited susceptibility to cancer, describing options for medical follow-up, and arranging genetic tests as appropriate (Schneider, 2002).

8.1. Collecting family histories

The first and most important step in assessing a patient's hereditary cancer risk is the collection of a careful and detailed family history. Pedigrees typically span 3–4 generations and include both sides of the family. For each affected member of the family, information is obtained about the cancer diagnosis including the exact location of the tumor, the stage at diagnosis and how the cancer was treated. Precancerous conditions (e.g., colonic polyps) are also ascertained, as should benign, but possibly related, conditions (e.g., nevi). Since the risk assessment and subsequent recommendations are based on the pattern of cancerous and

precancerous tumors in the family, it is important that the information be accurate. Thus, counselors typically request written documentation of the relatives' cancer diagnoses.

8.2. Providing risk assessments and follow-up options

Families can be categorized as having a high, moderate, or low likelihood of having a hereditary predisposition to cancer. High-risk families have patterns of cancer that are suggestive of a hereditary cancer syndrome and affected members are assumed to carry an inherited gene mutation. In this instance, the patient's cancer risks can be estimated by his/her position in the pedigree, current age, absence or presence of associated features, and the age-specific penetrance of the genes in question. At-risk individuals are generally recommended to undergo additional cancer surveillance at more frequent intervals than in the general population. They may also have the option of risk-reducing strategies, such as chemoprevention or prophylactic surgery.

In moderate-risk families, there may be a few features suggestive of an inherited cancer syndrome, but not enough to make a diagnosis. Estimating a patient's cancer risk in this instance may involve the use of empiric data or a range of risk. Recommendation about medical follow-up may be based on the specific cancer(s) seen in the family rather than all of the malignancies associated with the syndrome in question.

Members of low-risk families can be reassured about the low likelihood of an inherited risk factor in the family and can be given standard information about cancer surveillance and risk avoidance.

8.3. Arranging genetic testing

Genetic testing is currently available for over 50 hereditary cancer syndromes (**www.genetests.org**). Discussions about genetic testing include testing eligibility (including the person in the family who should be tested initially), possible test results, limitations of test results, implications of a positive test result for the patient and other family members, pros and cons of testing, and logistics. Logistical issues include whether testing will be done in a clinical or research laboratory, the cost of the analysis, the number of counseling visits in the testing program, when results are likely to become available, how results will be given, and whether results will be placed in the patient's medical record. Special counseling issues include patient concerns about discrimination and stigma, emotional distress engendered by the family's experiences with cancer or test results, strain on family relationships, and ethical dilemmas that range from decisions about duty to warn other family members about cancer risks to balancing family rights.

References

American Cancer Society (2004) *Cancer Facts & Figures*, American Cancer Society: Atlanta.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, *et al*. (1998) A national cancer institute

workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Research*, **58**, 5248–5257.

Eng C (1999) RET proto-oncogene in the development of human cancer. *Journal of Clinical Oncology*, **17**, 380–393.

Fearon ER (1998) Tumor suppressor genes. In *The Genetic Basis of Human Cancer*, Vogelstein B and Kinzler K (Eds.), McGraw-Hill: New York, pp. 229–230.

GeneTests: Medical Genetics Information Resource (Database Online), Copyright, University of Washington Press: Seattle, (1993–2004), **http://www.genetests.org**.

Hamilton SR, Liu B, Parsons RE, Papadopoulos N, Jen J, Powell SM, Krush AJ, Berk T, Cohen Z, Tetu B, *et al*. (1995) The molecular basis of Turcot's syndrome. *The New England Journal of Medicine*, **332**, 839–847.

Knudson AG (1993) Antioncogenes and human cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 10914.

Kohlmann W and Gruber SB (2004) Hereditary non-polyposis colon cancer. *GeneReviews*. **http://www.genetests.org**

Lucci-Cordisco E, Zito I, Gensini F and Genuardi M (2003) Hereditary nonpolyposis colorectal cancer and related conditions. *American Journal of Medical Genetics*, **122A**, 325–334.

Offit K (1998) The common hereditary cancers. *Clinical Cancer Genetics: Risk, Counseling, and Management*, Wiley-Liss pubs: New York, pp. 66–156.

Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore) and National Center for Biotechnology Information, National Library of Medicine (Bethesda), (2000), World Wide Web URL: **http://www.ncbi.nlm.nih.gov/omim/**

Park M (1998) Oncogenes. In *The Genetic Basis of Human Cancer*, Vogelstein B and Kinzler K (Eds.), McGraw-Hill: New York, pp. 205–207.

Pearson PL and Van Der Luijt RB (1998) The genetic analysis of Cancer. *Journal of Internal Medicine*, **243**, 413–417.

Petrucelli N, Daly MB, Burke W, Culver JOB, Hull JL, Levy-Lahad E and Feldman GL (2004) BRCA1 and BRCA2 Hereditary breast/ovarian cancer. *GeneReviews*. **http://www.genetests.org**

Rodriguez-Bigas MA, Boland CR, Hamilton SR, Henson DE, Jass JR, Khan PM, Lynch H, Perucho M, Smyrk T, Sobin L, *et al*. (1997) A national cancer institute workshop on hereditary nonpolyposis colorectal cancer syndrome: meeting highlights and Bethesda guidelines. *Journal of the National Cancer Institute*, **89**, 1758–1762.

Schneider K (2002) Predisposition testing and counseling. *Counseling About Cancer: Strategies for Genetic Counseling*, Second Edition, Wiley-Liss: New York, pp. 249–290.

Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Brody LC and Tucker MA (1997) The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine*, **336**, 1401–1408.

Syngal S, Fox EA, Eng C, Kolodner RD and Garber JE (2000) Sensitivity and specificity of clinical criteria for HNPCC associated mutations in MSH2 and MLH1. *Journal of Medical Genetics*, **37**, 641–645.

Szabo CI and King MC (1997) Population genetics of BRCA1 and BRCA2 [editorial; comment]. *American Journal of Human Genetics*, **60**, 1013–1020.

Vasen HF, Mecklin JP, Khan PM and Lynch HT (1991) The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Diseases of the Colon and Rectum*, **34**, 424–425.

Vasen HF, Watson P, Mecklin JP and Lynch HT (1999) New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology*, **116**, 1453–1456.

Wahlberg SS, Schmeits J, Thomas G, Loda M, Garber J, Syngal S, Kolodner RD and Fox E (2002) Evaluation of microsatellite instability and immunohistochemistry for the prediction of germ-line MSH2 and MLH1 mutations in hereditary nonpolyposis colon cancer families. *Cancer Research*, **62**, 3485–3492.

Familial adenomatous polyposis

Madhuri R. Hegde
Baylor College of Medicine, Houston, TX, USA

C. Sue Richards
Oregon Health & Science University, Portland, OR, USA

1. Case studies

1.1. Case 1

DK (III-1), a 17-year-old Caucasian male, presented to his GP with irregular bowel and bleeding. DK was referred for evaluation to a GI specialist because of a family history of colon cancer in three generations (Figure 1). GI evaluation showed the presence of thousands of adenomatous and hyperplastic polyps. Subsequent eye examination by an ophthalmologist with a genetics subspecialty showed congenital hypertrophy of the retinal pigment epithelium (CHRPE). On the basis of the family history and clinical presentation, a diagnosis of familial adenomatous polyposis (FAP) was given and genetic testing was ordered to confirm the diagnosis.

1.2. Case 2

GS (II-2), a 47-year-old Caucasian male, presented to his GP with inflammatory bowel disease. GS was referred for evaluation to a GI specialist for surveillance colonoscopy due to his increased risk for colorectal cancer (CRC) associated with these conditions and because of his family history of colon cancer (Figure 2). GI evaluation showed the presence of approximately 100 adenomatous polyps. GS was given a probable diagnosis of Attenuated FAP (AFAP) and referred for genetic counseling based on his family history. Molecular testing for *APC* gene mutations was recommended to confirm his diagnosis and to characterize the mutation.

1.3. Case 3

NJ (III-1), a newborn male of mixed Caucasian/Hispanic heritage, presented to his pediatrician with voluminous osteoma in the frontal areas associated with diffuse subcutaneous lipomas. Clinical examination revealed skin cysts and diffuse

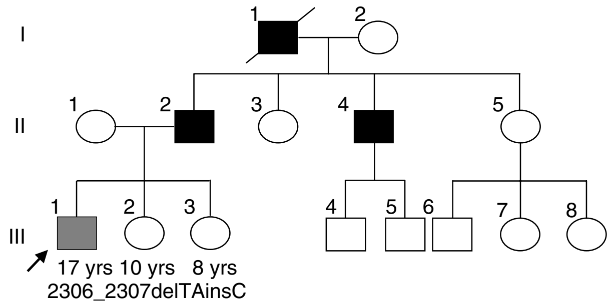


Figure 1 Pedigree for case 1. The 2306_2307delTAinsC mutation was identified in the proband. Black symbols = CRC; gray symbol = symptomatic proband

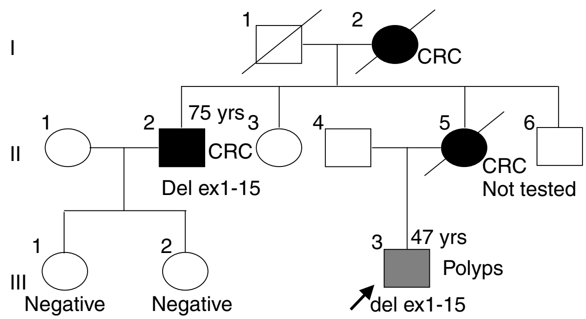


Figure 2 Pedigree for case 2, an AFAP family. A deletion of one copy of exons 1-15 was identified in the proband and his affected maternal uncle. Black symbols = CRC; gray symbol = symptomatic proband

increased pigmentation of the skin. NJ was referred to a GI specialist for further evaluation with a possible diagnosis of Gardner syndrome. GI evaluation revealed the presence of a carpet of colonic polyps. NJ has no family history of FAP (Figure 3). Molecular testing for *APC* gene mutations was ordered to confirm his diagnosis.

2. Background

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and cause of cancer deaths in the United States, with an estimated 147 500 newly diagnosed cases and 57 100 deaths in 2003 (Lynch and de la Chapelle, 2003; *see* Article 65, **Complexity of cancer as a genetic disease**, Volume 2). Most of the CRC is sporadic, but some of the genes responsible for the hereditary CRC syndromes have been identified, and routine genetic testing is available. The most prevalent of inherited CRCs include hereditary nonpolyposis colon cancer (HNPCC) and familial adenomatous polyposis (FAP), and more recently MYH associated polyposis (MAP) (Sieber *et al*., 2003). Individuals who inherit a germline mutation in one of the mismatch repair genes have an approximately 80% lifetime risk of developing

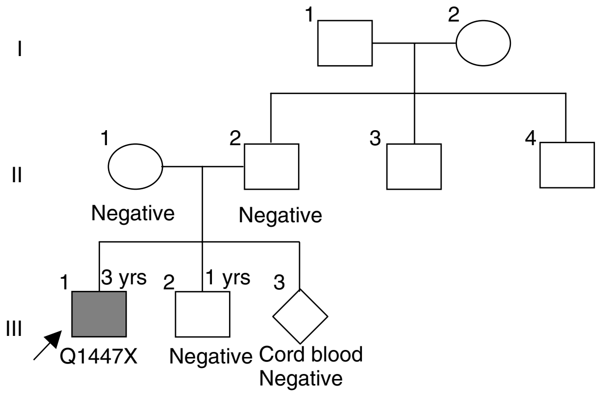


Figure 3 Pedigree for case 3. The novel Q1447X mutation was identified in this sporadic case of Gardner syndrome. Both parents and two unaffected siblings tested negative. Solid symbol = affected

CRC (Lynch and de la Chapelle, 2003). That risk is even greater (>95%) for individuals who inherit an *APC* mutation resulting in the FAP phenotype. For these families, screening surveillance is more intense and begins at an earlier age. Surveillance includes annual colonoscopy from adolescence, and if polyps are detected, prophylactic colectomy to eliminate the risk of CRC. If untreated, the majority of classical FAP patients will develop CRC, generally by the fourth decade of life.

FAP is inherited as an autosomal dominant and accounts for up to 1% of all colorectal cancers (Cama *et al*., 1997). The incidence of FAP is estimated at approximately 1 in 5000 individuals in the United States. Classic FAP is characterized by the presence of hundreds to thousands of adenomatous colorectal polyps, which usually first appear in the distal colon during adolescence and progress to CRC if prophylactic total colectomy is not performed. Extra-colonic features may include gastric polyps, osteomas, congenital hypertrophy of the retinal pigment epithelium (CHRPE), soft-tissue tumors, thyroid carcinoma, and hepatoblastoma. Other known variants of FAP include attenuated FAP (AFAP) (Friedl *et al*., 1996), which is characterized by late onset and lesser number of polyps, Gardner syndrome, which typically is associated with a number of extra-colonic features, and Turcot syndrome associated with brain tumors (medulloblastoma), in addition to colon polyps and CRC.

2.1. *APC gene*

In 1991, the understanding of the molecular pathogenesis of CRC was profoundly altered by the identification of the causative gene for FAP, the adenomatous polyposis coli *(APC)* tumor suppressor gene, located on chromosome 5q21-22 (Fearnhead *et al*., 2001; Bodmer *et al*., 1987). The *APC* gene consists of 8535 base pairs, encoding a 2843 amino acid protein (Figure 4) (Fearnhead *et al*., 2001). While the gene is composed of 15 exons, exon 15 accounts for >75% (6.5 kb) of the

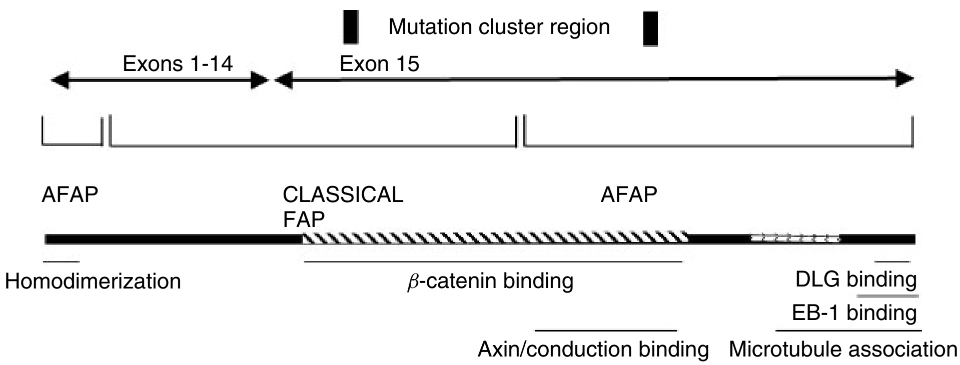


Figure 4 A schematic representation of the *APC* gene: exons 1-14 and exon 15 are shown with mutation cluster region, functional domains, and disease phenotype for each region. Functional domains are shown under the bar: homodimerization, β- catenin binding, DLG binding domain (lethal (1) discs large-1), EB-1 protein binding domain, axin/conduction binding domain, and microtubule association region

coding sequence and contains the vast majority of mutations, which are truncating mutations (~90%) and include small deletions, small insertions, and nonsense mutations. The remainder consists of missense mutations (~4%) and gross alterations (~5%) (Fearnhead *et al*., 2001). The mutation spectrum is scattered across the gene with a number of recurrent mutations, particularly codon 1309 and 1068 in exon 15. Exon 15 (~1200–1400) contains a mutation cluster region (MCR) in which sporadic mutations occurring in CRC tumors as second events in Knudson's hypothesis are often found.

The majority of *APC* mutations in FAP are found to cause premature truncation of the protein product, which results in abnormal protein function. These germline truncating mutations in *APC* are detectable in over 80% of patients with classic FAP. Most commonly, the protein truncation test (PTT) (van der Luijt *et al*., 1994) has been used for the detection of truncating mutations in the *APC* gene, but this technique depends on the availability of cDNA, and accuracy of mutation detection is based on the position and type of mutation. Recently, several new techniques have been described and include mutation scanning using conformational sequencing gel electrophoresis (CSGE) (Arancha *et al*., 2002) and denaturing high-performance liquid chromatography (dHPLC) (Wu *et al*., 2001), followed by sequencing to characterize the mutation. More recent studies have also demonstrated the presence of large rearrangements/deletions or mutations within the *APC* promoter that affect gene expression. These large gene rearrangements may contribute to the remaining 20% of *APC* mutations (Su *et al*., 2000).

2.2. *APC gene function*

The *APC* gene product functions as a tumor suppressor with a gatekeeper role at the level of initiation of tumorigenesis (Munemitsu *et al*., 1995). The interactions of the various domains of the *APC* gene product have led to a greater understanding of the many roles for this protein. Structurally, the APC protein can be divided (from

amino to carboxyl terminus) into the following domains: a homodimeric-binding region; β-catenin binding region; axin-binding region; and microtubule-binding region (Figure 4). Perhaps the most important role for the APC protein is a negative regulator in the wnt-signaling pathway. In this role, APC binds β-catenin and also binds to axin, which increases the binding to β-catenin, and facilitates the phosphorylation of β-catenin by GSK-3β, leading to the degradation of β-catenin in the proteosome (Rubinfeld *et al*., 1996). If β-catenin is not regulated by APC, its concentration increases within the cell and it diffuses into the nucleus where it interacts with Tcf and other transcription factors to increase the expression of downstream genes, including myc and cyclin D, thus leading to unregulated growth control. Thus, mutations occurring in the β-catenin-binding domain generally lead to more classical phenotypes.

2.3. *Treatment*

After the diagnosis of FAP is made, if the patient is left untreated, the risk of colon cancer is near 100%, and median life expectancy is approximately 42 years. Disease management consists of lifetime endoscopic surveillance, initial colon resection followed by complete resection, if required, and may involve repeated surgeries. Given the serious consequences of FAP in terms of cancer risk and the need for repeated major surgical interventions, there has been interest in developing a systemic treatment with low toxicity that could reduce polyp burden as an adjunct to surgery. One potential pharmacologic target in impeding growth of adenomatous tissue is the cyclo-oxygenase (COX) enzyme (Phillips *et al*., 2002). The recent development of selective COX-2 inhibitors has provided the opportunity to more safely test the hypothesis that selective inhibition of COX-2 might be useful in the prevention or treatment of adenomatous polyps, and clinical trials are under way.

2.4. *Genetic counseling*

Genetic counseling is strongly recommended for FAP families. Unlike other genetic disorders with adult onset, FAP in the classical form is manifested in children. Thus, it is important to counsel parents about the clinical symptoms, course of disease, surveillance measures, and treatment options, as well as provide an avenue for genetic testing. Genetic testing of the proband in the family is important first to identify the *APC* mutation segregating within the family and to confirm the diagnosis. Among confirmed gene mutation carriers, preventive health options include frequent screening, and for those individuals who develop many colorectal polyps, surgical removal of polyps via colectomy. Identification of the mutation in the proband then allows genetic testing for at-risk family members. In such cases, there are two possible outcomes. If the at-risk family member is found to carry the same *APC* gene mutation as in the proband and is currently asymptomatic, enrollment in drug trials aimed at delay or prevention of tumor development may be an option along with frequent surveillance monitoring. For those family members

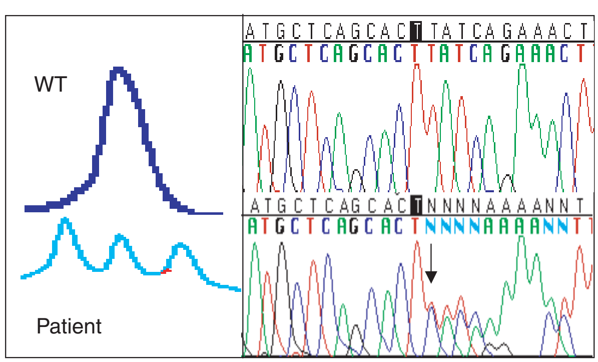
who receive genetic testing and are found to be negative for the familial mutation, frequent, costly and unpleasant colonoscopies can be avoided. It is important to remember that negative results do not eliminate the risk of CRC; thus, the same general population guidelines for colon screening at a later age would apply.

3. Case study approaches

3.1. Approach to case 1

A blood specimen from DK was sent to the clinical laboratory for *APC* mutation analysis. dHPLC analysis was performed and a mutation was identified in exon 15 in one copy of the *APC* gene (Figure 5). Sequence analysis revealed a 2-bp deletion and 1-bp insertion mutation at nucleotide position 2306_2307delTAinsC (codon 769). This indel mutation results in a frameshift and a stop codon at nucleotide 2326 (codon 776), producing a truncated protein. This mutation has not been previously reported for the *APC* gene in the HGMD (Human Genome Mutation Database), although it is interpreted as a disease-causing mutation. This mutation is predicted to act as a dominant negative mutation, by which, in its presence, the abnormal product from the allele inactivates the product of the normal allele. This mutation lies within the β-catenin binding domain, and thus is predicted to be important for normal function of the APC protein. On the basis of the results of the molecular testing in DK, his diagnosis is confirmed, and testing is available for at-risk family members.

Genetic counseling is strongly encouraged for this family. DK (III-1) has a strong paternal family history of CRC, with two at-risk siblings (III-2 and III-3) and two at-risk paternal first cousins (III-4 and III-5). Genetic counseling issues should be discussed with DK and his family, including the implications of his mutation-positive status for his health care and for other family members, with a



(a) (b)

Figure 5 (a) dHPLC profile for case 1. A dHPLC profile showing the patient (lower) has a different profile from the wild type (upper). (b) Sequence analysis was performed to characterize the 2306_2307delTAinsC mutation (lower) in comparison to wild-type sequence (upper)

recommendation for molecular testing of the affected parent for mutation confirmation. The family should be advised that presymptomatic testing is available for at-risk family members and that additional genetic testing will target the familial mutation identified in DK. The family should also be informed that prenatal testing is available for future pregnancies of mutation carriers. The medical implications discussed with this family should include virtual certainty of development of colorectal cancer for mutation-positive carriers, necessity of continued screening, and options for surgical intervention. The risk for each successive generation based on inheritance of an autosomal dominant disorder should be presented to the family, along with options for genetic testing.

3.2. Approach to case 2

A blood specimen from GS (II-3) was sent to the clinical laboratory for *APC* mutation analysis. After full sequence analysis of the coding region of *APC*, no mutation was detected. However, GS was found to be homozygous for the eight common single nucleotide polymorphisms (SNPs) in the *APC* gene, suggesting a large gene deletion on one allele. Additional testing was performed for gross gene alterations using a real-time quantitative PCR-based assay, and a deletion of the entire *APC* gene was detected (Figure 6a). Confirmation of this finding was done using multiplex ligation probe amplification (MLPA) (Bunyan *et al*., 2004) (Figure 6b) and fluorescence in situ hybridization (FISH) (Rogan *et al*., 2001) analysis (Figure 6c). This exon 1-15 deletion mutation is expected to result in haploinsufficiency of the APC protein, resulting in inefficient downregulation of β-catenin. Thus, these molecular results confirm the diagnosis of AFAP for this patient and allow testing for at-risk family members.

GS has a family history of CRC (Figure 2) with his mother (II-5), maternal grandmother (I-2), and maternal uncle (II-2) affected with CRC. Subsequent targeted genetic testing for the familial *APC* mutation was performed on the affected 75-year-old maternal uncle, identifying the same exon 1-15 deletion mutation and confirming these results. Subsequent genetic testing performed on at-risk individuals (III-1 and III-2) produced negative results. These findings indicate that III-1 and III-2 are at no greater risk for developing CRC than the general population and thus do not require the frequent monitoring by colonoscopy as do mutation-positive carriers.

3.3. Approach to case 3

NJ's (III-1) blood sample was sent to the clinical laboratory for mutation analysis of the *APC* gene. dHPLC and sequencing identified a mutation in exon 15 in one copy of the *APC* gene (Figure 7). A nonsense mutation was detected at nucleotide position 4339C > T, leading to a change of glutamic acid to a stop at codon 1447 (Q1447X) and producing a truncated protein. This novel mutation has not previously been reported in the HGMD, although it is the type expected to cause FAP.

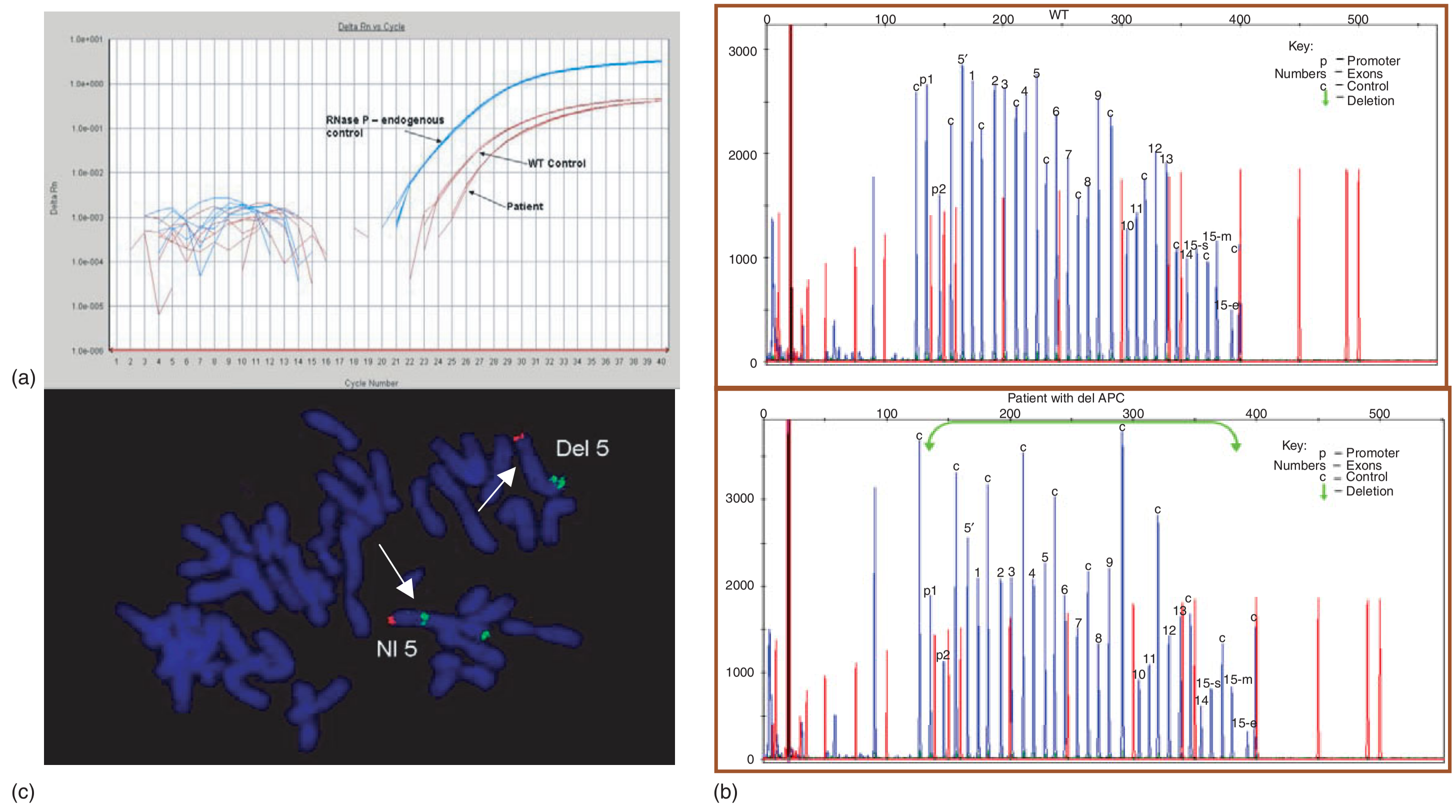


Figure 6 (a) Real-time PCR analysis for the detection of suspected deletions in case 2. Patient sample shows a large deletion spanning the entire coding region of the *APC* gene. The shift showing a difference between the wild-type product and that of the patient indicates that the patient sample has a deletion in the region spanning the exon. (b) MLPA confirmed that the patient has the entire *APC* gene deleted on one allele. The height of the numbered peaks corresponding to each exon is compared to that of the control peaks (c) and the wild-type profile (top panel) is compared with the patient (lower). (c) Confirmation of the gene deletion using FISH. FISH was performed using a biotin labeled BAC RP11-107C15 containing the *APC* gene as a probe (green signal). The 5p telomere (green signals) and the 5q telomere (red signals) were used as control probes. Note that the arrow points to *APC* gene, which is missing in the upper portion of the figure

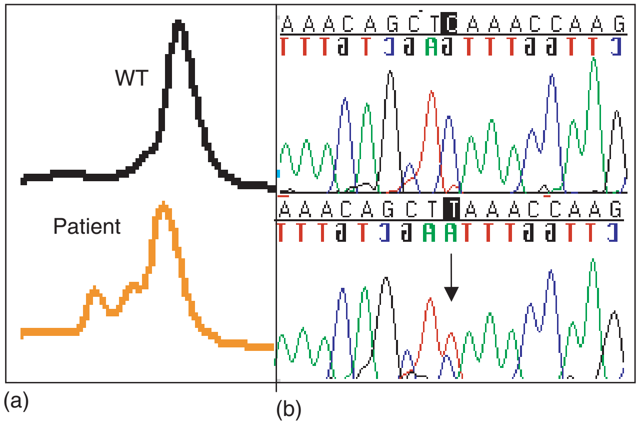


Figure 7 (a) dHPLC and sequence analysis for case 3. The dHPLC profile shows the patient (lower panel) has a different profile from that of the wild-type sample (upper panel). (b) Sequence analysis was performed to characterize the Q1447X mutation. The patient sequence (lower panel) is compared to wild-type sequence (upper panel)

On the basis of the results of the molecular testing for NJ, genetic counseling was strongly recommended. NJ has no family history of Gardner syndrome, and his mother is pregnant (Figure 3). Both parents were tested and were found to be negative for the Q1447X mutation, indicating that their blood cells do not carry the Q1447X mutation detected in their son. However, these results do not rule out the possibility for gonadal mosaicism as the de novo event. Approximately 25% of germline *APC* mutations are de novo events arising either in the germline of the affected offspring or somatically in the germ cells of the parent. Clinical laboratory testing does not distinguish between these two possibilities. Genetic counseling issues were discussed with NJ's family, including the possibility of gonadal mosaicism and the implications for the current and future pregnancies. A sibling (III-2) and the fetus of the current pregnancy (III-3) were subsequently examined to ascertain their carrier status by targeted mutation analysis for the familial mutation, and both were found to be negative. The parents were informed about the recommended surveillance regimen, the availability of COX-2 inhibitors pediatric clinical trials, treatment options, and the expected clinical progression of disease for their affected child. They were reassured that their other two children were at a decreased risk due to their negative test results and would not require frequent lifetime screening procedures.

References

Arancha C, Ruiz-Llorente S, Cascon A, Osorio A, Martinez-Delgado B, Benitez J and Robledo M (2002) A rapid and easy method for multiple endocrine neoplasia type 1 mutation detection using conformation-sensitive gel electrophoresis. *Journal of Human Genetics*, **47**(4), 190–195.

Bodmer WF, Bailey CJ, Bodmer J, Bussey HJ, Ellis A, Gorman P, Lucibello FC, Murday VA, Rider SH, Scambler P, *et al*. (1987) Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature*, **328**, 614–616.

Bunyan DJ, Eccles DM, Sillibourne J, Wilkins E, Thomas NS, Shea-Simonds J, Duncan PJ, Curtis CE, Robinson DO, Harvey JF, *et al*. (2004) Dosage analysis of cancer predisposition genes by multiplex ligation-dependent probe amplification. *British Journal of Cancer*, **91**(6), 1155–1159.

Cama A, Guanti G, Mareni C, Radice P, Saglio G, Varesco L and Viel A (1997) Recommendations for the molecular diagnosis of familial adenomatous polyposis. *Tumori*, **83**, 795–799.

Fearnhead NS, Britton MP and Bodmer WF (2001) The ABC of APC. *Human Molecular Genetics*, **10**, 721–733.

Friedl W, Meuschel S, Caspari R, Lamberti C, Krieger S, Sengteller M and Propping P (1996) Attenuated familial adenomatous polyposis due to a mutation in the 3′ part of the APC gene. A clue for understanding the function of the APC protein. *Human Genetics*, **97**, 579–584.

Lynch HT and de la Chapelle A (2003) Genomic medicine: hereditary colorectal cancer. *The New England Journal of Medicine*, **6348**, 919–932.

Munemitsu S, Albert I, Souza B, Rubinfeld B and Polakis P (1995) Regulation of intracellular beta-catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 3046–3050.

Phillips RK, Wallace MH, Lynch PM, Hawk E, Gordon GB, Saunders BP, Wakabayashi N, Shen Y, Zimmerman S, Godio L, *et al*. (2002) FAP Study Group. A randomised, double blind, placebo controlled study of celecoxib, a selective cyclooxygenase 2 inhibitor on duodenal polyposis in familial adenomatous polyposis. *Gut*, **50**(6), 857–860.

Rogan PK, Cazcarro PM and Knoll JH (2001) Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization. *Genome Research*, **11**, 1086–1094.

Rubinfeld B, Albert I, Porfiri E, Fiol C, Munemitsu S and Polakis P (1996) Binding of GSK3beta to the APC-beta-catenin complex and regulation of complex assembly. *Science*, **272**(5264), 1023–1026.

Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RK, Bisgaard ML, Orntoft TF, Aaltonen LA, Hodgson SV, *et al*. (2003) Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *The New England Journal of Medicine*, **348**(9), 791–799.

Su LK, Steinbach G, Sawyer JC, Hindi M, Ward PA and Lynch PM (2000) Genomic rearrangements of the APC tumor-suppressor gene in familial adenomatous polyposis. *Human Genetics*, **106**, 101–107.

van der Luijt R, Khan PM, Vasen H, van Leeuwen C, Tops C, Roest P, den Dunnen J and Fodde R (1994) Rapid detection of translation-terminating mutations at the adenomatous polyposis coli (APC) gene by direct protein truncation test. *Genomics*, **20**, 1–4.

Wu G, Wu W, Hegde M, Fawkner M, Chong B, Love D, Su LK, Lynch P, Snow K and Richards CS (2001) Detection of sequence variations in the adenomatous polyposis coli (APC) gene using denaturing high-performance liquid chromatography. *Genetic Testing*, **5**, 281–290.

Introductory Review

Gene therapy I: principles and clinical applications

J. Wesley Ulm
Harvard Medical School, Brookline, MA, USA

1. Introduction

Gene therapy is a novel pharmacological approach in which the drug is supplied in the form of a nucleic acid – DNA, RNA, or some modification or combination thereof. The origin of gene therapy lies in the watershed advancements of two fields, particularly in the 1970s and 1980s: recombinant DNA technology, which enables the precise manipulation and amplification of defined DNA segments and virology, in which progress in the basic science of viral life cycles and replication has allowed for the exploitation of viral particles as vectors to ferry therapeutic DNA or RNA into cells. Although many obstacles remain to be overcome before gene therapy can enter the mainstream of treatment and prevention of human disease, the technology holds great promise. Effective and safe gene therapy would bring pharmacology to the core of the central dogma in biology, treating disease at the level of the purine and pyrimidine nucleic acid polymers that control the fundamental processes of normal physiology and pathophysiology.

2. Why gene therapy?

From a pharmacological perspective, gene therapy presents an attractive option since its very nature is conducive to a high therapeutic index, at least at the cellular level. A DNA or RNA sequence prefabricated to target a particular disease locus constitutes, in principle, a highly specific and minimally toxic therapeutic (as discussed below, the viral vectors currently used in most gene therapy protocols unfortunately introduce a degree of nonspecificity and toxicity, often in the form of an immune response). Gene therapy also affords the prospect of long-term effectiveness from a small number of interventions, enabling physicians to treat efficiently many classes of disease that would otherwise be recalcitrant to therapy. Heritable loss-of-function genetic diseases such as cystic fibrosis (CF), the hemophilias, sickle-cell anemia, Duchenne muscular dystrophy, and many others represent the classic targets for gene therapy. A gene therapy protocol providing a stable supply of a missing or damaged protein could obviate the need for the chronic, expensive,

and often only partially effective (e.g., hemophilia, the thalassemias) or palliative (e.g., CF, Tay–Sachs disease, Lesch–Nyhan syndrome) treatments that currently exist. In practice, most protocols recently have focused on cancer, heart failure, and angiogenesis, but recessive genetic diseases still represent the canonical target for which gene therapy was originally conceived.

Finally, gene therapy provides a flexible treatment strategy that could be readily adapted to treat novel pathogens and even overcome resistance. Modern medicine, particularly the antimicrobial and anticancer chemotherapeutic drugs introduced in the wake of World War II, has in some respects become a victim of its own success. Such treatment regimens frequently become case studies in cellular microevolution, as the drugs are ultimately selected for targets that are resistant to their mode of action. This challenge has repeatedly foiled many a promising new cancer treatment: a novel drug seems to clear the tumor from a patient's body, only to have a relapse occur months or years later because of the outgrowth of a few cells with a multidrug-resistance membrane pump or an overamplified or mutated protein target. It is here that a gene therapy scheme could demonstrate great benefit. If a drug target (e.g., a pathogenic bacterium or a relapsing tumor) were to acquire a resistance phenotype against a small molecule therapeutic, it would be a long and time-consuming process to generate and screen new small molecules to adapt to the change, even with modern improvements in computer modeling and robotic technologies. A linear DNA sequence, in contrast, could be rapidly and reliably altered to target the new gain-of-resistance mutation. In a novel paradigm of treatment, gene therapy remedies could be tailored to each patient and dynamically adapted to respond to resistance as it arises.

3. Prospects for clinical application

As will be discussed in specific cases below, many gene therapy modalities have made the leap from bench to bedside in the form of clinical trials. While gene-replacement strategies represent the canonical application of gene therapy, in practice, most trials have focused on cancer and blood vessel disease, two areas with a large pool of patients and a high index of clinical applicability for the therapeutic vectors of interest. Although many novel modalities have attracted substantial interest and shown promise in the laboratory, to date most clinical trials with gene therapy have had disappointing results. The first such clinical trial was undertaken by French Anderson and colleagues in the early 1990s to treat pediatric patients with adenosine deaminase (ADA) deficiency, a disorder that prevents proper development of lymphocytes and renders patients immunodeficient. These studies involved the ex vivo introduction of retroviral vectors (discussed below) bearing an ADA expression cassette into hematopoietic cells, followed by their injection into patients' bone marrow. The results of the work were confounded, however, by the concurrent administration to the trial participants of PEG-ADA, a traditional pharmaceutical regimen for ADA patients that has made the interpretation of the trial data rather difficult.

Hundreds of other trials have since been launched, occasionally with hints of efficacy but in general failing to exceed the threshold that would allow gene therapies to surpass more established pharmaceuticals in clinical practice. The one exception to this in recent years has been the work of Alain Fischer and colleagues in France, who in 2000 and 2002 reported the successful use of a retroviral-based modality to reconstitute the immune systems of patients suffering from an X-linked congenital form of severe combined immune deficiency (discussed below). The results of the trial have been partly marred by the subsequent development of a T-cell leukemia (albeit one highly amenable to chemotherapeutic treatment) in 3 out of the 10 patients in the study; yet, the French trial nonetheless represents the first demonstration of a sustainable, long-term correction of an inherited disorder using a gene therapy approach (see Cavazzana-Calvo *et al*., 2000; Hacein-Bey-Abina *et al*., 2002).

The difficulties and frustrations encountered in most clinical trials so far should not lead to the premature conclusion that gene therapy is unworkable in a clinical setting. Most such trials encounter the same subset of obstacles – usually related to vector delivery, efficiency of transduction and intracellular expression, and long-term maintenance of the transgene (see review by Mulligan, 1993) – for which greater understanding is obtained with each new attempt. The benchtop studies and clinical trials cross-pollinate each other, since the basic science investigations are used to design the initial therapeutic protocols and, in turn, themselves benefit from the experience and knowledge gleaned from the trials. The work of Fischer and colleagues in France in particular constitutes a watershed not only for inherited diseases but also for the field in general, and demonstrates that gene therapy can indeed provide a viable approach to treat patients for whom no other effective treatment is available. Furthermore, strategies like oncolytic viruses and neoplastic apoptosis induction offer the prospect of anticancer treatments far more selective than most available today, while improved applications of RNA interference techniques especially suggest a practical methodology to design new and effective drugs far more rapidly than is required for most small molecules. Gene therapy is thus built on well-established phenomena and proven technologies that can potentially fine-tune medicine in a manner that is elusive for most pharmacological modalities, and the early hurdles of clinical trials constitute the difficult yet inevitable learning curve that must be surmounted to attain clinical practicality.

The next sections will discuss the various gene therapy strategies, organized according to classes of disease targets, requirements for safe and effective treatment, and viral gene therapy vectors currently under development.

4. Classes of disease targets

To which areas of human disease could a viable gene therapy protocol be applied to the greatest advantage? While this answer can only be conjectured at the present, current proposals for clinical trials and topic reviews tend to focus on the following (summarized in Table 1).

Table 1 Examples of diseases for which gene therapy has been used or is being considered

Class	Exemplary diseases	Mutated gene product	Result
Mendelian, homozygous recessive loss-of-function	Cystic fibrosis	CFTR, a cell-surface chloride channel	Deficient luminal water transport, salty sweat, pulmonary infections, male sterility, exocrine pancreas insufficiency, intestinal obstruction
	Familial hyper-cholesterolemia	Low-density lipoprotein (LDL) receptor	Extremely elevated plasma LDL, cholesterol, and triglycerides; xanthomas, xanthelasma; angina, early-onset acute myocardial infarction
	Tay–Sachs disease	Hexosaminidase A	Accretion of ganglioside GM2 in CNS neurons; apathy, retardation, blindness, convulsions, paralysis, macular degeneration, death by age 3
Protein aggregation	Alzheimer's disease	*Tau* protein, amyloid precursor protein (APP)	Anterograde and retrograde amnesia; damage to cholinergic neurons in frontal and temporal lobes (especially hippocampus and amygdala)
Neoplastic	Malignant melanoma	B-Raf, cyclin dependent kinases and cdk inhibitors	Development of dysplastic cutaneous nevi, neoplastic transformation, eventual metastasis; target of "cancer vaccine"
Cardiovascular	NIDDM-induced small vessel disease	Polygenic disorder	Diabetic vasculitides, secondary to chronically elevated blood glucose, lead to paresthesias and neuritides; retinal angiogenesis leads to retinopathy

4.1. Diseases attributable to simple recessive Mendelian, X-linked, triplet-repeat, or mitochondrial inheritance

Many human diseases can be traced to mono- or diallelic mutations in a single gene that is crucial for homeostasis. A large class of disorders is caused by homozygous recessive mutations that cause a loss of function in a critical cellular protein, often associated with signal transduction pathways that regulate hematological function, endocrine and metabolic systems, or ion transport in epithelia or nerves. Such anomalies are notoriously difficult to treat with conventional remedies. While a deficiency in a circulating molecule such as insulin (diabetes mellitus type I) or factor IX (hemophilia B) can be, at least partially, ameliorated by exogenous administration of the deficient protein, a loss of function in a cognate receptor, intracellular enzyme, or structural protein is typically unresponsive to such therapy. Unless the mutation can be circumvented, this loss of function may present an

intractable danger to the patient's health. The ΔF508 mutation in CF – pursuant to the loss of a single phenylalanine codon in the *cftr* gene on both chromosomes – renders an otherwise functional chloride transporter incapable of reaching the plasma membrane of epithelial cells, instead occasioning its sequestration in the endoplasmic reticulum and its subsequent degradation by proteasomes. This homozygous deletion (and others like it) affects critical cellular functions such as water transport, and ultimately leads to the fatal sequelae of CF.

Similarly, familial hypercholesterolemia (FHC) is caused by mutations in both genes coding for the cell surface low-density lipoprotein (LDL) receptor; decreased or absent receptor expression prevents cellular uptake of LDL and provokes the deleteriously high levels of plasma LDL that cause accelerated atherosclerosis. A litany of enzyme deficiencies – for example, Tay–Sachs disease (hexosaminidase A); Gaucher's disease (glucocerebrosidase); Fabry disease (alpha-galactosidase A); ataxia-telangiectasia, Bloom's syndrome, and Fanconi's anemia (DNA-repair enzymes); immune system disorders such as ADA deficiency and chronic granulomatous disease (NADPH oxidase) – result in usually fatal illnesses in affected patients. The insufficiency can be partly remedied in some cases (e.g., Gaucher's, ADA deficiency, and Fabry) by continual administration of the missing enzyme as a drug; nevertheless, cures for these diseases (as through bone marrow transplants) remain exceptional and elusive. Mutations in intracellular structural proteins (muscular dystrophy), membrane components (myelin in Type Ia Charcot-Marie-Tooth syndrome), and specialized functional proteins (hemoglobin in sickle-cell anemia and thalassemia) also pose detrimental or even grave threats to the health of affected individuals.

Most of the above are examples of recessive or X-linked diseases. Some disorders transmit in families according to autosomal dominant patterns. Neurofibromatosis 1 (resulting from germline monoallelic loss of neurofibromin, a GTPase-activating protein) and Li–Fraumeni syndrome (consequent to germline monoallelic loss of one p53 allele) are both recessive at the *cellular* level, since they require two hits against a tumor-suppressor protein to manifest in a given cell. However, if one allele is mutated in the germline, then stochastic events will inevitably give rise to the loss of heterozygosity at an early age, indicative of dominant transmission. Other autosomal dominant diseases manifest through dosage effects or dominant-negative actions, thus precipitating deleterious clinical sequelae. Hereditary triplet-repeat disorders – such as Huntington's disease, fragile X syndrome, and myotonic dystrophy – are caused by the expansion of a repeated 3-base sequence, which can occur in the regulatory or coding region of the affected gene, that is amplified in the germline from one generation to the next. This amplification fosters an increasingly severe phenotype in each successive generation (so-called anticipation). Finally, mitochondrial DNA disorders – such as Leber's optic neuropathy – are passed on maternally.

The common thread in these genetic diseases is the loss of function of a critical cellular component or the accumulation of a toxic gene product (see Table 1). Because these diseases are hereditary, the molecular and cellular defect is both widespread and lifelong. Traditional drug therapy is generally ineffective against this panel of disorders. Thus, these diseases represent the canonical challenge for

gene therapy strategies, which seek to replenish or counteract the defective function in the cells of interest.

4.2. Protein aggregation diseases

The converse of the heritable loss-of-function diseases includes dominantly inherited disorders in which there is a generation of an altered form of a particular gene product that has a dominant negative effect on the normal function or a toxic gain of function. In the latter case, for example, the abnormal protein may accrue in tissues, with toxic effects. A classic case is Alzheimer's disease. Though the precise etiology is still controversial, autopsies of Alzheimer's brains show both intracellular neurofibrillary tangles, or NFTs, of the microtubule-associated *tau* protein, and extracellular deposits of beta-amyloid protein (BAP). These gene products, as well as auxiliary proteins (e.g., apolipoprotein E4 and the presenilins), are often altered in Alzheimer's disease patients, and their dysregulation is believed to contribute to NFT formation and amyloid deposition. A similar pattern is observed in other amyloid accumulation disorders (e.g., amyloid kidney disease, diabetes mellitus type II, and systemic amyloidosis), and analogous processes are suspected in other adult-onset neurodegenerative illnesses such as amyotrophic lateral sclerosis (ALS). (In ALS, a somatically mutated superoxide dismutase I (SOD-I) gene may induce the pathology.) By reducing or abrogating the production of the abnormal protein, gene therapy strategies such as antisense RNA or gene conversion could contribute enormously to the management of these disorders.

4.3. Neoplastic diseases

Cancer develops when a clonal population of somatic or germ cells possesses a combination of genetic and chromosomal aberrations that cause inappropriate cell cycling, inefficient DNA repair, loss of contact inhibition and apoptosis, and karyotypic instability. Because cancers originate in gain-of-function (oncogenic) and loss-of-function (tumor-suppressor) mutations in the genome, this class of diseases provides another potential target for gene therapy. Putative treatment strategies could focus on downregulating oncogenes or augmenting tumor-suppressor genes, or on selectively killing tumor cells directly or via enhanced T-cell-mediated cytolytic attack. Some modalities unite gene therapy with more traditional small molecule-based approaches. For example, a vector expressing an enzyme that converts a prodrug into a toxic molecule can be transduced into tissues in the vicinity of a tumor, with the result that the neoplastic cells are selectively killed. Other strategies attempt to deliver apoptotic factors selectively to cancer cells, to differentiate the neoplastic cells, or to thwart the angiogenesis that is essential for continued growth or metastasis of a tumor (Tandle *et al*., 2004) (yet another strategy, utilizing engineered adenoviruses as "oncolytic viruses" is discussed below). Cancer is an especially popular subject of gene therapy clinical trials, and many of the above methodologies have been or are currently being investigated in patients.

4.4. Cardiovascular diseases

Diseases of the heart and vascular system may be especially amenable to gene therapy strategies. Many disorders, such as the increasingly prevalent noninsulin-dependent (type II) diabetes mellitus (NIDDM), entail the attrition of blood delivery in general, and the loss of small-blood vessel capacity in particular, because of inflammation, chemical complexation, and immune-complex damage at the vessel walls – the proximate cause of the peripheral neuropathies, renal disease, susceptibility to infection, and retinopathy of NIDDM. Patients with Type II diabetes could therefore benefit from gene therapy modalities that engender targeted revascularization, as could those with myocardial ischemia or infarction from sclerosed or obstructed coronary arteries. In addition to promoting the establishment of a collateral circulation, gene therapy can facilitate the elaboration of modulatory factors that can help reduce complications like intimal proliferation and lumenal restenosis of grafted coronary vessels. New vessels produced by such protocols often suffer from reduced structural integrity and stability, but revascularization remains a promising area of research for gene therapy modalities.

4.5. Infectious diseases

The most common application of gene therapy in the realm of infectious diseases is in the production of vaccines. Inoculation against microbial infection can be achieved by the delivery of naked DNA that expresses epitopes specific for the pathogen. Viral vectors have also found utility. The same methods used to engineer viral vectors to express a missing gene can be applied to engender an immune response, by generating pathogen-specific proteins that stimulate humoral and cellular arms of the immune system. Some viruses (such as HIV) against which vaccination is desirable may be too dangerous to be administered as attenuated particles in healthy people, and strategies are therefore sought to express immunogenic epitopes in the context of other viruses. One possible frontier for gene therapy strategies may lie in confronting the increasing peril of microbial resistance to antibiotics and antiviral drugs. Persistent microbial evolution, high replication rates, and the paucity of pharmacological targets in viruses and some bacteria have combined to generate a formidable medical challenge for the coming century. Rapidly adaptable gene therapy modalities could contribute to more effective infectious disease management in the face of such challenges.

5. Requirements for safe and effective gene therapy

In the classical case of gene replacement, the objective of gene therapy is to modulate expression of a target gene, often by restoring the function of a mutated protein but sometimes by downregulating or controlling the expression of such a protein. Other applications (such as cancer therapy or angiogenesis) involve the delivery of a novel gene to kill target cells selectively, or production of a factor

to effect a change in the physiological state of a damaged tissue. These modalities must meet several demands in order to be safe and effective:

1. *Transgene persistence and long-term stable expression*
   (a) Expression of the novel locus (supplied by the gene therapy vector in targeted cells) should be both steady and long-term, with the transgene escaping degradation and epigenetic silencing.
   (b) Cells expressing the transduced gene should be protected from immune clearance.
   (It should be noted that the above two conditions apply primarily to the classical modality of gene replacement. In some applications, such as neovascularization, it may be desirable to curtail gene expression deliberately over time, while in cancer therapy a vigorous immune response may be instrumental to the treatment strategy itself.)
2. *Specific expression*
   Vectors should be directed to the target cell population of interest, and transduction of bystander tissues should be minimized.
3. *Regulated expression*
   If the dosage of a restored normal gene must be precisely controlled (e.g., a hemoglobin gene in thalassemia), the gene therapy system should include a mechanism to manipulate expression levels reliably so as to approximate the physiological state.
4. *Reliable delivery*
   Parenterally administered in vivo vectors should be sufficiently stable to reach the nucleus of target cells intact, evading plasma proteases and nucleases as well as lysosomal degradation and interferon-mediated suppression.
5. *Minimal toxicity*
   Gene therapy vectors should be minimalistic, so as to minimize both immunogenicity and direct toxicity from the vector's components.

There are nearly 100 trillion cells in an adult human body, and shuttling a bolus of the nucleic acid "drug" to its targets is a formidable challenge. To circumvent obstacles to delivery, some gene therapy efforts that address heritable genetic diseases have focused on ex vivo transduction of hematopoietic and tissue stem cells, which are then reimplanted and utilized to restore the normal protein to the affected patient. In vivo gene therapy, in which the nucleic acid is administered parenterally much like many small molecule drugs, requires the vector to home to the targeted tissues via the blood and lymphatic circulation or to be injected directly into the tissue. For both ex vivo and in vivo approaches, proper tissue targeting is essential for success. Moreover, since the immune system is designed to react to foreign proteins such as those supplied in the gene therapy vector, drug delivery strategies must be designed either to evade immune mechanisms or to achieve the effective delivery over a minimum number of treatment sessions before immune memory sets in. Finally, the pharmacodynamics and pharmacokinetics of in vivo gene therapy vectors are typically more intricate than those of small molecule drugs. Along with the traditional concerns of route of administration, plasma stability, and hepatorenal clearance, vectors must confront *immune* and *intracellular* clearance,

and one must consider the body's handling of not only the DNA or RNA "drug" itself but also the vehicle that ferries the drug into cells.

Because of their genomic modularity and innate ability to deliver genetic material into cells, viruses have become the vectors of choice for gene therapy. Recombinant DNA technology allows disparate fragments of viral genomes to be maintained on separate DNA plasmids, permitting safe and convenient manipulation that supplants viral expression cassettes with the therapeutic genes of interest. The next chapter summarizes the salient features of viral vectors that have found application in gene therapy.

References

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, *et al*. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.

Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay JP, Thrasher AJ, Wulffraat N, Sorensen R, Dupuis-Girod S, *et al*. (2002) Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *The New England Journal of Medicine*, **346**, 1185–1193.

Mulligan RC (1993) The basic science of gene therapy. *Science*, **260**, 926–932.

Tandle A, Blazer DG III and Libutti SK (2004) Antiangiogenic gene therapy of cancer: recent developments. *Journal of Translation Medicine*, **2**, 22.

Introductory Review

Gene therapy II: viral vectors and treatment modalities

J. Wesley Ulm
Harvard Medical School, Brookline, MA, USA

1. Introduction

In its classical manifestation, gene therapy exploits the capabilities of mammalian viruses – including, remarkably, a number of infectious agents notorious for their pathogenic potential in human hosts – as vehicles to deliver therapeutic nucleic acids to cells and tissues. The previous chapter introduced some basic requirements for the vectors used in gene therapy modalities, and this chapter will introduce the viral systems most commonly employed in clinical protocols. Each choice of vector exhibits a specific set of properties with regard to the maximum size of a therapeutic insert, efficiency of production in culture, tissue tropism within the host, capacity to integrate in a target cell, long-term persistence in the host, and overall toxicity. Investigators designing gene therapy protocols must weigh the benefits and shortcomings specific to each viral type in deciding which vector to use in the treatment of a particular disease. Viral vectors are in general rendered nonreplicating, which has necessitated the development of specialized tissue culture systems to generate viral particles in high titers. Furthermore, the expansion of treatment modalities available to gene therapy researchers – including RNA-based techniques such as antisense and RNA interference – has led, in recent years, to additional production systems specially tailored to maximize the therapeutic potential of these options in the setting of a clinical application.

2. Viral gene therapy vectors: applications and limitations

Only a small subset of the immense variety of animal viruses have been tested as bearers of therapeutic genes. These have been chosen primarily from among the common human pathogens with broad tissue tropism, for which detailed information on genes and sequences is available. Viruses vary considerably in size, life cycle, titer, ease of laboratory manipulation, safety, immunogenicity, and composition (and they may possess RNA or DNA genomes). Currently, nearly all gene therapy protocols use retroviruses, adenoviruses, adeno-associated viruses (AAVs), herpes viruses, or Epstein–Barr viruses (EBVs) as the gene delivery agent. The

general features, advantages, and disadvantages of each virus family are summarized in Table 1. Detailed discussions of retroviral, adenoviral, and AAV vectors are provided in other chapters; therefore, these vector systems will be only briefly summarized here.

2.1. Retroviruses

Retroviruses (family *Retroviridae*) were so named because of their "retrograde" flow of genetic information from RNA in the virion to DNA in the host cells. While many other RNA viruses exist, retroviruses stand out in their use of a DNA intermediate and in their possession of an RNA-dependent DNA polymerase called *reverse transcriptase* (RT), which generates a double-stranded complementary DNA (cDNA) from the diploid retroviral genomic template, bounded by long terminal repeats (LTRs), in the cytoplasm of an infected cell. For purposes of gene therapy, retroviruses possess another salient feature that has made them a vector of choice for gene replacement applications: As part of its replication cycle, a retrovirus's cDNA – flanked by the LTRs – integrates stably into the genome of a host cell (becoming an *integrated provirus*), a feature that greatly assists long-term retention and expression of a therapeutic transgene.

Most retroviruses can infect only actively dividing cells, a limitation that can nonetheless be an asset in anticancer modalities. However, one subgroup of the retrovirus family, the lentiviruses – which includes HIV – can productively infect resting, noncycling cells. For reasons of both safety and efficiency, retroviral vectors are generally rendered nonreplicating, with essential viral components maintained on separate plasmids lacking LTRs, or supplied in trans via complementation by *packaging cells*, which express viral gene products off genomic loci. (Lentiviral vectors are endowed with additional safety features because of their high pathogenicity in the wild (reviewed in Delenda *et al*., 2002.) Retroviral vectors have been utilized in several notable clinical trials over the past decade. In a promising recent study, Alain Fischer and colleagues in Paris used the Moloney murine leukemia virus (Mo-MLV) to transduce dividing hematopoietic progenitors of patients with a rare X-linked immune disorder, SCID-X1, caused by an inherited deficiency in the γ-c subunit of the interleukin receptor that is critical for lymphocyte maturation (Cavazzana-Calvo *et al*., 2000; Hacein-Bey-Abina *et al*., 2002) (summarized in Figure 1). A T-cell acute lymphoblastic leukemia emerged in three of the patients in the study, owing to aberrant activation of the LMO-2 protooncogenic locus consequent to proviral integration; however, these leukemias have been highly responsive to therapy, and further adjustments in the retroviral systems should help circumvent this complication. Retroviruses have also been utilized as immunotherapeutic "cancer vaccines". Melanoma cells from affected patients are isolated and transduced with a retroviral vector to produce GM-CSF, a cytokine that activates dendritic cells – a potent form of antigen-presenting cell (APC) – then reimplanted into patients to stimulate a strong and specific anti-melanoma immune response.

Table 1 Viral gene therapy vectors

Viral vector	Taxonomy and structure	Advantages	Disadvantages
Retrovirus	– Retroviridae – ssRNA viruses – Reverse-transcribe viral genome into dsDNA after entry into cells – ~7–10-kb genome – Flanked by LTRs, which mark boundary between transgenic sequence and host genome	– Reverse-transcribed DNA is integrated into host cell genome, providing the potential for long-term expression off retroviral U3 promoter or other viral or cellular promoters – Highly modular, can express *gag, pol, env* genes off separate plasmids and modify as desired – Replication-incompetent viral particles can be produced in packaging cells – No recombination steps required in construction	– Transgene integration is random, creating a danger of interfering with crucial gene expression or activating protooncogenes (recently observed in clinical trial) – Some difficulties with long-term expression – Can infect only actively dividing cells after breakdown of the nuclear envelope (except lentiviruses like HIV, which can infect resting cells)
Adenovirus	– Adenoviridae – Linear dsDNA – 30–40-kb genome – Sequential gene expression – Can infect many tissues – A cause of the common cold	– High titers and efficiency of infection – E1a, E1b, E3, E4 regions can all be substituted with therapeutic genes; multiple inserts are therefore possible – High tropism for liver (good for hepatic diseases) and various epithelia – Vectors can be complemented and packaged conveniently in 293 cells – Episomal residence and gene expression can be a safety feature	– Construction requires recombination steps owing to large size – Residual viral gene products (and transgene, to a lesser degree) elicit strong immune response against cells infected with the vector – Substantial difficulty achieving long-term expression off adenoviral episome, due to immune clearance – Potential toxicity in infected cells, especially in liver – Danger of contamination with replication-competent helper virus in preparations
Adeno-associated virus (AAV)	– Parvoviridae – Linear ssDNA – 4.5–5.5-kb genome – Flanked by ITRs	– Can transduce nondividing cells – Good transduction of CNS, muscle – Low immunogenicity – Availability of distinct serotypes that can infect different primary cell types	– Small capacity limits size of packaged insert – Difficulty in achieving large-scale production
Herpesvirus	– Herpesviridae – Linear dsDNA – Very large (150 kb) genome; latency	– Large size enables packaging of sizable inserts, for example, potentially useful for treatment of DMD (an enormous gene) – Tropic for neural tissue	– Some cellular damage observed even in replication-defective mutants – Silencing of promoters in CNS
Epstein-Barr virus (EBV)	– A herpesvirus – Circular DNA episome	– Persistence of episome makes it useful in hybrid vectors	– Low infection titers by itself, so not useful as a stand-alone gene therapy vector

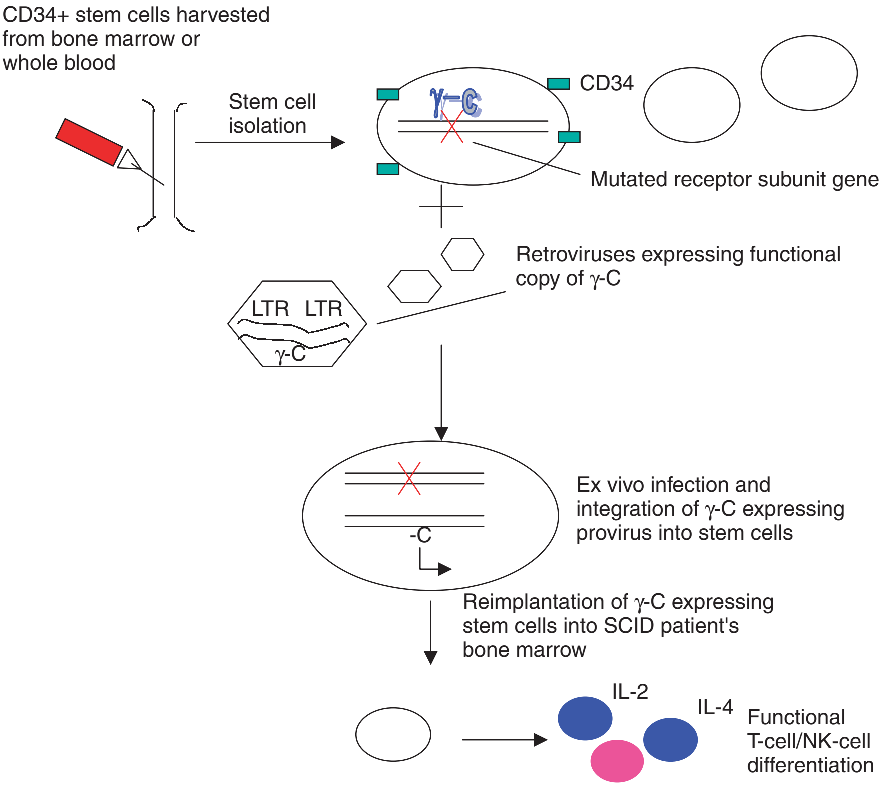


Figure 1 Stem cell-based gene therapy: Ex vivo retroviral transduction of functional γ-C receptor subunit-expressing gene into stem cells of a SCID-X1 patient. Use of the patient's own stem cells helps to obviate immune rejection of the transplanted cells, although the new gene is ectopic to its original location. It is hoped to use similar methodology for a wide array of diseases, such as Duchenne muscular dystrophy, by differentiating stem cells into other tissues. Silencing of the retroviral transgene and toxicity associated with ectopic expression of the transgene are concerns. Based on work of Alain Fischer *et al*

2.2. Adenoviruses and adeno-associated viruses

Members of the family *Adenoviridae* are large (30–40 kb), linear dsDNA viruses that are frequently implicated in human illness, causing eye and bladder disorders as well as common cold. The family is quite diverse and there is frequent recombination among different serotypes, but the general organization is the same from virus to virus. *Adenoviridae* display a complex temporal regulation of gene expression divided into early (E), delayed early (IX and Iva2), and late (L) genes, with viral genomes flanked on each side by an *inverted terminal repeat* (ITR). Unlike retroviruses, adenoviruses are double-stranded DNA viruses that do not integrate into the host genome, but remain as linear episomes in the nucleus, separate from the host's genomic DNA. Furthermore, adenoviruses are capable of infecting both resting and dividing cells.

When adenoviruses are used as gene therapy vectors, foreign DNA (up to about 7 kb) can be introduced into E1, E2, and parts of the E4 region. Adenoviruses have a large tissue host range, with an especially marked tropism for the liver. They can be produced at high titers in 293 cells, a human complementation line that supplies E1a and E1b gene products in trans; as a result, adenoviruses are convenient for the in vivo expression of an essential gene in significant quantities.

Adenoviral vectors are often used in vaccine protocols to express immunostimulatory epitopes so as to activate both humoral and cellular immunity (Imler, 1995), as well as in numerous anticancer regimens (reviewed by Zhang, 1999). For example, E1a and E1b-deleted adenoviral variants can serve as *oncolytic viruses* that selectively infect p53- and Rb-deficient neoplastic cells (tumor suppressors that are complexed by E1a and E1b in the wild). Adenoviral vectors, unfortunately, are notorious for stimulating a vigorous immune response against infected cells, and this and other factors heighten their toxicity and reduce their long-term persistence. Moreover, special care must be taken to ensure the absence of replication-competent "helper" virus in adenoviral vector preparations.

AAVs are a family of small, 4.5–5.5 kb single-stranded DNA parvoviruses flanked by ITRs. They derive their name from the fact that they require the adenoviral E1a, E1b, E2a, and E4 gene products in trans to propagate themselves, and therefore are naturally infectious only in tandem with a concurrent adenoviral infection of a target cell. In practice, since replication deficiency is desirable in a gene therapy vector, adenoviral complementation is furnished only in the cells in which the AAV vectors are initially generated. AAVs are available in multiple serotypes, which facilitates their application for targeting specific tissues, such as muscle and brain, including nondividing cells.

AAVs can integrate into target DNA like retroviruses, and are even capable of doing so in a site-specific manner. The AAV *rep* protein recognizes a cognate sequence that results in enhanced integration into a region of the long arm of chromosome 19 called the AAVS1 site. In practice, however, AAV vectors do not express *rep* upon infection of target cells, and therefore integrate randomly (like retroviruses), either as single proviruses or as head-to-tail concatemers.

2.3. Herpesviruses, Epstein-Barr virus (EBV), and nonviral systems

Herpesviruses are enormous (>150 kb) dsDNA viruses, among the largest found in nature. They are ubiquitous in human disease, herpes simplex virus (HSV)-1 causing cold sores and fever blisters, HSV-2 inducing genital pruritus and blisters, varicella-zoster virus (VZV) causing both chicken pox and shingles, and human herpes virus (HHV)-8 inducing AIDS-associated Kaposi's sarcoma. Herpesvirus recombinant vectors can accommodate up to 30 kb of inserted sequence, and so-called amplicons (which are analogous to gutless adenoviruses) can incorporate up to 150 kb of foreign DNA sequence. Thus, herpesviruses provide one of the only viable means to package very large gene cassettes, such as a cDNA for the dystrophin gene that is mutated in Duchenne muscular dystrophy. Herpesviruses are also tropic for the nervous system, which they access in nature via retrograde axonal flow from infection at a mucosal surface to residence at

a dorsal root ganglion. This property makes herpesviruses valuable for treatment modalities that must surmount the blood-brain barrier (Jacobs *et al*., 1999; Lam and Breakefield, 2000). Furthermore, their propensity to enter a state of latency following infection helps reduce immune provocation and facilitate long-term establishment in target cells. Like other vectors, herpesviruses suffer from the potential for gene silencing in transduced cells (including silencing of the "latency" promoters). Moreover, use of even nonreplicating herpesviruses can have some toxicity.

One interesting herpesvirus variant is the EBV, the pathogen that causes infectious mononucleosis. As a virus with transforming potential that is tropic for lymphocytes, EBV is also implicated in Burkitt's lymphoma and nasopharyngeal carcinoma. EBV is unusual in that it persists in infected cells as a circularized episome, and it partitions – along with the host cell genome – into dividing cells following replication. Its titers are generally low and, while it is impractical as a gene therapy vector by itself, EBV is popular as a component of so-called hybrid vectors that utilize two or more viral backbones to generate the vector. An adenoviral-EBV vector, for example, combines the high efficiency and wide host range of the adenoviruses with the episomal persistence of EBV, to forge a "virus within a virus". EBV elements are packaged within an adenoviral "shell", then excised following infection by a lox-CRE-mediated system to liberate the episome. Adenovirus- or herpesvirus-retrovirus hybrids pursue a similar objective, introducing retroviral integration elements into the genome of the adenovirus or herpesvirus to engender an integrating component.

One additional area of research involves nonviral gene therapy systems. The simplest example of such methods is the DNA vaccine, in which naked DNA is injected locally into a tissue, with subsequent cellular uptake engendering the production of relevant epitopes against a pathogenic agent (reviewed in Wiethoff and Middaugh, 2003; Parker *et al*., 2003). In general, however, nonviral systems employ more sophisticated modalities such as DNA-coated circulating microparticles to deliver a therapeutic nucleic acid bolus to affected tissues. Elements of nonviral systems generally mimic essential properties of viral vectors, which have made them useful as vehicles for therapeutic genes, such as their capacity to efficiently endocytose through plasma membranes and escape lysosomal degradation to enter cellular nuclei. Progress in nonviral vector modalities is technologically driven, strongly correlating with advances in the capacity to mass-produce sophisticated, microfabricated lipid or polymeric particles in cell-free systems. These methods are currently inefficient, but technical advances in biochemistry and microparticle engineering may soon help nonviral systems to become viable alternatives to viral-based modalities, with a greater safety margin than most viral protocols.

2.4. Antisense strategies, dsRNA, RNAi

A relatively new and promising application for gene therapy consists of a set of strategies that employ RNA molecules to inhibit specifically a gene of interest, for example, a gene whose product is toxic. Toxic proteins or peptides are thought

to be involved in degenerative diseases of the nervous system such as Huntington's disease, amyotrophic lateral sclerosis, Kreutzfeld–Jakob disease, retinitis pigmentosa, and (probably) Alzheimer's disease. In such cases, the protein or peptide accumulates (often by aggregation) at levels that are toxic to nearby neurons, and therapeutic approaches directed at minimizing the deleterious accumulation may be beneficial. Cancer is another salient target of RNA inhibitory approaches, because neoplasia inevitably involves the inappropriate and constitutive expression of growth-enhancing and cell-cycling oncogenes, whose repression by specific downregulating RNA molecules could furnish a means to selectively target and kill neoplastic cells. In addition, RNA inhibitory strategies against cancer may be more adaptable than other modalities in treating cancer relapses due to new mutations in tumor cells, because an RNA molecule that targets its novel (mutated) complement is far easier to generate than a small molecule or peptide that targets a protein. Finally, RNA inhibition could potentially yield a rich harvest of new antimicrobial drugs that are targeted against the wealth of unique gene products present in foreign pathogens.

From a physiological perspective, "RNA inhibition" consists of a series of conserved mechanisms, in both prokaryotes and eukaryotes, which regulate gene expression and combat viral infection by modulating mRNA amounts and controlling the levels of mRNA translation (see reviews by Hannon, 2002; Inouye, 1988; Brantl, 2002). RNA inhibition entails two overlapping yet distinct phenomena – (1) antisense RNA and (2) RNA interference or RNAi. (The action of ribozymes, in which catalytic RNA cleaves an RNA target in cis or in trans, is also related to these phenomena.) There is occasional confusion between the two, in part because of their similarities, but antisense RNA is not the same as RNAi and, in fact, the latter was discovered as a consequence of work on the former.

The principle of antisense RNA is elementary, relying on the ability of RNA, like DNA, to form duplexes. Thus, an RNA transcript elaborated from the antiparallel strand of an open reading frame in the genome will hybridize to the mRNA produced upon activation of transcription at that gene. Watson–Crick pairing of a capped and polyadenylated message with its complement can then interfere with mRNA processes, ribosome interactions, and/or the docking of translation initiation factors, effectively rendering the mRNA incompetent for peptide production. Interaction of a transcript with its antisense sequence can also promote RNAse H–mediated degradation of the message (Figure 2). The principle of antisense inhibition was established in the early 1980s, and it found utility as a laboratory tool for selective blocking of gene expression in a wide variety of organisms (Dolnick, 1997). Eukaryotic antisense transcripts have been implicated in the phenomenon of genomic imprinting, which involves the selective, epigenetic downregulation of alleles specific to the paternal or maternal chromosome (Ward and Dutton, 1998). Antisense RNA approaches have been recently applied in the clinic; fomivirsen is a recently approved antisense RNA used to treat retinitis caused by cytomegalovirus in patients immunocompromised by HIV infection.

Like antisense RNA, RNAi requires the formation of a dsRNA structure. Otherwise, however, the two mechanisms are remarkably different. A dsRNA substrate, several hundred base pairs in length, can be processed by a dsRNA-specific endonuclease, called *Dicer*, into short fragments approximately 20-bp long. These

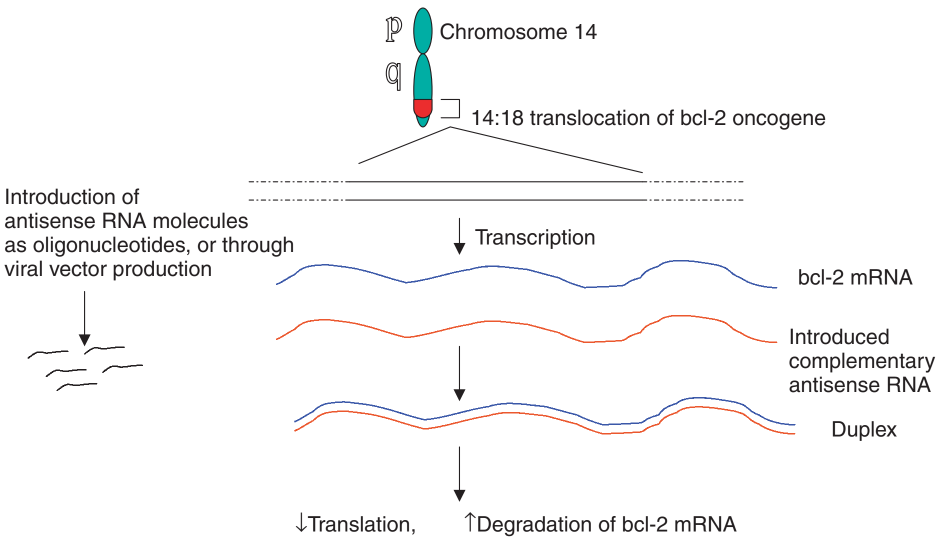


Figure 2 Use of an exogenous antisense RNA molecule to block production of an oncogenic gene product

short dsRNAs, called siRNAs (for small interfering RNA particles), are the active species in RNAi. The siRNAs form a triplex RNA structure with homologous mRNAs, and these complexes draw the attention of the RISC (RNA-induced silencing complex) array, which digests the target mRNA. In this way, mRNAs identical or homologous to the siRNAs are selectively degraded. RNAi has been shown to be operational across many species, including plants, *Caenorhabditis elegans*, *Drosophila*, and mammals.

RNA inhibition strategies represent a potentially exciting frontier because of their power and selectivity in downregulating target genes. One drawback, of course, is that they act primarily at the level of elaborated RNA transcripts rather than the genome itself. Thus, continuous administration would be required to inhibit a targeted gene chronically. RNA inhibition also represses gene expression incompletely, and as with all gene therapy strategies, successful delivery is a serious challenge. Nonetheless, the RNAi approach embodies many of the facets that make gene therapy so potentially valuable – in particular, the ability to modify the drug rapidly and adapt it to changes (such as mutations) in disease targets. RNAi has already demonstrated promise in cell culture, thwarting the infection of cells by HIV and polio virus (Coburn and Cullen, 2002; Gitlin *et al*., 2002; Hu *et al*., 2002). Moreover, the Dicer-directed endonuclease-processing step can be bypassed to enable direct synthesis and utilization of the active siRNA, and methodologies to express siRNA species directly off polIII or phage promoters may permit continuous expression of these inhibitory molecules in cells of interest.

2.5. *In situ gene correction*

An additional frontier for the gene therapy field involves the use of specialized vectors to engender the repair of nonfunctional gene targets in vivo and in situ. In

most gene replacement strategies – envisaged and implemented for diseases such as cystic fibrosis, hemophilia, and sickle cell anemia – the viral vector expresses the missing or defective gene *ectopically*, outside the native genomic location of the gene in normal cells. This can occur either as an integrated provirus (as for retroviruses) randomly integrated into a chromosome or as an intranuclear, nonintegrated episome for vectors such as adenoviruses. Such ectopic expression, however, can be problematic since the physical location of a gene in normal cells is often germane to its expression levels, regulation, and integration into signal transduction networks. Epigenetic mechanisms like chromatin remodeling, methylation, and histone acetylation are often critical factors in the precise modulation of gene expression that is seen in vivo, and such controls may not be operative if a gene integrates into a different chromosomal site or resides on an episome. Careful regulation of expression is especially crucial in many hematological and endocrine disorders, whose sine qua non is the loss or mutation of a protein whose dosage levels are normally maintained within narrow limits to ensure proper physiological function and response to stresses.

To achieve gene correction in situ, a gene therapy vector would not carry an intact expression cassette to be integrated or episomally expressed; rather, it would furnish a *repair template* oligonucleotide containing a section of the gene that is mutated or deleted, possibly in conjunction with specialized proteins involved in homologous recombination, mismatch repair, or other cellular processes that participate in DNA modification or double-strand break (DSB) formation and repair. The human equivalents of the Rad52 epistasis group proteins, found in yeast, are central players in the repair of cellular DSBs and have thus attracted interest as potential facilitators of gene correction.

One possible methodology to effect such correction is afforded through the use of AAV vectors, which have been used successfully to perform gene targeting in mammalian cells (Russell and Hirata, 1998). While targeting efficiency is still low (approximately 1%), further knowledge of the mechanisms involved in such AAV-mediated repair may enable far higher rates of gene conversion. Indeed, recent work has suggested that AAV-driven correction is greatly enhanced in the presence of induced DSBs (Miller *et al*., 2004), and techniques to induce site-specific DSBs in the genome – through the use of "meganucleases" such as the I-SceI endonuclease, which recognize an 18-bp sequence that can be unique even within the large human genome – can potentially augment gene correction levels even further. The precise mechanism through which a single- or double-stranded DNA-repair template finds its target within the genome is unclear, though it may ensue from transient Watson–Crick base pairing between the oligonucleotide and the complementary strand of a genomic duplex, which has temporarily unwound due to stochastic "breathing" and opening of the two strands.

Another proposed yet highly controversial scheme to effect gene correction involves a technique dubbed *chimeraplasty* (Cole-Strauss *et al*., 1996). This modality uses a unique oligonucleotide consisting of a double-stranded RNA-DNA hybrid, in which the corrective bases in the DNA of one strand are flanked by RNA bases – often chemically modified – and paired with a DNA complement. The modification was claimed to both bolster stability of the repair template (protecting it from exonuclease digestion) as well as to augment its capacity

to base pair with a mutated target. The chimeraplast, paired with its genomic target, would then constitute a substrate to attract DNA-repair machinery (perhaps mismatch repair proteins), which would then correct the mutated base pair to result in the proper sequence. Early reports had suggested gene conversion levels of up to 50% in cells with the sickle-cell anemia mutation, yet the chimeraplasty field has been surrounded by controversy since later studies have shown mixed results in substantiating the original findings (see review by Taubes, 2002). The chimeraplastic oligos are difficult to synthesize and apply, and many laboratories have reported a degree of gene correction but at much lower levels. A recent study in plants has suggested that at least a portion of the correction originally attributed to the chimeraplasts may in fact ensue from spontaneous mutation (Ruiter *et al*., 2003).

The prospect of in situ gene correction remains a promising, yet thus far unrealized approach to specific and efficacious gene therapy. As with many other gene therapy modalities, the intriguing conceptual basis of this methodology is hampered in practice by mundane issues of delivery and efficiency within the cell. Many of the processes that underlie gene correction remain poorly defined, just as in more established therapies that involve retroviral transgene or adenoviral episome persistence. The most fruitful avenues to advance these modalities to clinical applicability may therefore reside in basic science investigations that better elucidate the underlying mechanisms, enabling them to be more effectively exploited in specific treatments.

3. Conclusion

Modern medicine stands at the precipice of an entirely new pharmacology. For the first time in history, physicians seek to manage disease at the level of the genetic information that underlies all biology. The promise of this technology resides in its potentially magnificent specificity and high therapeutic index, as well as in the potential for rapid adaptation of therapies on an individualized basis. This latter property may be especially valuable in confronting cancer and a wide array of infectious diseases, for, as has become evident over the past several decades, natural selection is as operative in microbial and cellular pathology as it is in animal biology. The success of new chemotherapeutic drugs is too often attenuated or overcome altogether by the rise of resistant clones; the long-term solution to this challenge may be the capacity on the part of the clinician to respond just as rapidly. Currently, gene therapy faces significant hurdles in terms of vector delivery, stable transgene expression, and safety. But as work to overcome these obstacles progresses, the potential for both ex vivo and in vivo gene therapy comes closer to realization and clinical application.

Further reading

Bischoff JR, Kirn DH, Williams A, Heise C, Horn S, Muna M, Ng L, Nye JA, Sampson-Johannes A, Fattaey A, *et al*. (1996) An adenovirus mutant that replicates selectively in p53-deficient human tumor cells. *Science*, **274**, 373–376.

Dranoff G, Jaffee E, Lazenby A, Golumbek P, Levitsky H, Brose K, Jackson V, Hamada H, Pardoll D and Mulligan RC (1993) Vaccination with irradiated tumor cells engineered to secrete murine granulocyte-macrophage colony-stimulating factor stimulates potent, specific, and long-lasting anti-tumor immunity. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 3539–3543.

Porteus MH and Baltimore D (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300**, 763.

Wirth T, Zender L, Schulte B, Mundt B, Plentz R, Rudolph KL, Manns M, Kubicka S and Kuhnel F (2003) A telomerase-dependent conditionally replicating adenovirus for selective treatment of cancer. *Cancer Research*, **63**, 3181–3188.

References

Brantl S (2002) Antisense-RNA regulation and RNA interference. *Biochimica et Biophysica Acta*, **1575**, 15–25.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, *et al*. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.

Coburn GA and Cullen BR (2002) Potent and specific inhibition of human immunodeficiency virus type 1 replication by RNA interference. *Journal of Virology*, **76**, 9225–9231.

Cole-Strauss A, Yoon K, Xiang Y, Byrne BC, Rice MC, Gryn J, Holloman WK and Kmiec EB (1996) Correction of the mutation responsible for sickle cell anemia by an RNA-DNA oligonucleotide. *Science*, **273**, 1386–1389.

Delenda C, Audit M and Danos O (2002) Biosafety issues in lentivector production. *Current Topics in Microbiology and Immunology*, **261**, 123–141.

Dolnick BJ (1997) Naturally occurring antisense RNA. *Pharmacology and Therapeutics*, **75**, 179–184.

Gitlin L, Karelsky S and Andino R (2002) Short interfering RNA confers intracellular antiviral immunity in human cells. *Nature*, **418**, 430–434.

Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay JP, Thrasher AJ, Wulffraat N, Sorensen R, Dupuis-Girod S, *et al*. (2002) Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *The New England Journal of Medicine*, **346**, 1185–1193.

Hannon GJ (2002) RNA interference. *Nature*, **418**, 244–251.

Hu W, Myers C, Kilzer J, Pfaff S and Bushman F (2002) Inhibition of retroviral pathogenesis by RNA interference. *Current Biology*, **12**, 1301.

Imler JL (1995) Adenovirus vectors as recombinant viral vaccines. *Vaccine*, **13**, 1143–1151.

Inouye M (1988) Antisense RNA: its functions and applications in gene regulation – a review. *Gene*, **72**, 25–34.

Jacobs A, Breakefield XO and Fraefel C (1999) HSV-1-based vectors for gene therapy of neurological diseases and brain tumors: part II. Vector systems and applications. *Neoplasia*, **1**, 402–416.

Lam PY and Breakefield XO (2000) Hybrid vector designs to control the delivery, fate and expression of transgenes. *Journal of Genetics Medicine*, **2**, 395–408.

Miller DG, Petek LM and Russell DW (2004) Adeno-associated virus vectors integrate at chromosome breakage sites. *Nature Genetics*, **36**, 767–773.

Parker AL, Newman C, Briggs S, Seymour L and Sheridan PJ (2003) Nonviral gene delivery: techniques and implications for molecular medicine. *Expert Reviews in Molecular Medicine*, Sep 3, 1–15.

Ruiter R, van den Brande I, Stals E, Delaure S, Cornelissen M and D'Halluin K (2003) Spontaneous mutation frequency in plants obscures the effect of chimeraplasty. *Plant Molecular Biology*, **53**, 675–689.

Russell DW and Hirata RK (1998) Human gene targeting by viral vectors. *Nature Genetics*, **18**, 325–330.

Taubes G (2002) The strange case of chimeraplasty. *Science*, **298**, 2116–2120.

Ward A and Dutton JR (1998) Regulation of the Wilms' tumour suppressor (WT1) gene by an antisense RNA: a link with genomic imprinting? *The Journal of Pathology*, **185**, 342–344.

Wiethoff CM and Middaugh CR (2003) Barriers to nonviral gene delivery. *Journal of Pharmaceutical Sciences*, **92**, 203–217.

Zhang WW (1999) Development and application of adenoviral vectors for gene therapy of cancer. *Cancer Gene Therapy*, **6**, 113–138.

Specialist Review

Hematopoietic stem cell gene therapy

Adrian J. Thrasher

Institute of Child Health, London, UK

Fabio Candotti

National Institutes of Health, Bethesda, MD, USA

1. Introduction

Hematopoietic stem cell (HSC) transplantation is an established curative procedure for a variety of inherited disorders including hemoglobinopathies, immunodeficiencies, lysosomal storage diseases, and bone marrow failure syndromes. However, the high incidence of adverse immunologic effects associated with transplantation of allogeneic cells, including graft rejection and graft versus host disease, remains problematic. Since the early development of viral vectors as efficient tools for gene transfer into mammalian cells, autologous HSCs have been attractive targets for the development of alternative strategies based on gene correction or augmentation. Targeting human HSCs has been, and continues to be, a challenging task for several reasons. The identification of human HSCs, which themselves are heterogeneous, is difficult, and stringent purification methods that are easily transferable to the clinical setting are currently not available. Most strategies utilize cell populations that are enriched for HSC by selection of cells expressing the CD34 cell surface antigen by magnetic bead sorting. Even so, the large number of CD34+ cells that is required to achieve successful engraftment complicates the transduction process, and has implications for the development of toxicity arising from insertional mutagenesis (due to the large number of unique integration events in each graft). The identification of improved selection strategies for HSCs is therefore of significant importance.

Effective gene transfer to HSC and their progeny requires stability, which currently can only be achieved efficiently using integrating vectors, most commonly based on mammalian retroviruses (*see* Article 98, **Retro/lentiviral vectors**, Volume 2). Though effective, this leads to difficulties associated with variegation of transgene expression (dependent on the local chromatin environment), and potential for harmful mutagenesis (Baum *et al*., 2004; Challita and Kohn, 1994; Klug *et al*., 2000; Yao *et al*., 2004). Efficient transduction of a variety of primary hematopoietic cells can be achieved with vectors based on gammaretroviruses, lentiviruses, and

foamy viruses. At the time of writing, almost all clinical studies have employed murine gammaretroviruses, which are dependent on active proliferation of the target cell population for effective gene transfer because the nuclear membrane must be disrupted for entry of the preintegration complex. As HSCs are mostly quiescent, extensive studies have been dedicated to the identification of optimal ex vivo culture conditions that stimulate HSC proliferation, without inducing differentiation and loss of long-term repopulating ability. Though far from perfect, most protocols in current clinical use employ a combination of cytokines such as interleukin (IL)-3, IL-6, stem cell factor (SCF), thrombopoietin, and Flt-3 ligand, increasingly with defined serum-free culture conditions. The manipulation of cells with alternative molecules such as the homeodomain-containing transcription factor HOXB4 and bone morphogenetic protein (BMP)-4 has shown some promise, and may permit amplification of HSC populations (Bhatia *et al*., 1999; Sorrentino, 2004). Vectors based on lentiviruses and foamy viruses are under intense investigation as alternatives to murine gammaretroviruses as they are less dependent on cell division for effective gene transfer, and are highly efficient (Ailles *et al*., 2002; Case *et al*., 1999; Demaison *et al*., 2002; Josephson *et al*., 2002; Piacibello *et al*., 2002; Vassilopoulos *et al*., 2001). Clinical trials employing HIV-1-based lentiviral vectors for transduction of HSC are imminent.

2. Primary immunodeficiencies as models for HSC gene therapy

Primary immunodeficiencies (PID) are a heterogeneous group of disorders in which inherited genetic defects compromise host immunity (Fischer, 2004). The most severe forms of PID are known as severe combined immunodeficiency (SCID), in which T-lymphocyte development is invariably compromised, and associated with diverse disorders of development and functionality of B lymphocytes, and natural killer (NK) cells (Buckley, 2004). Although clinically severe, bone marrow transplantation is usually highly successful if a genotypically matched family donor or unrelated donor is available (Antoine *et al*., 2003; Buckley *et al*., 1999). However, for the majority of individuals, this is not the case, and survival from mismatched family (usually parental donor) transplants is substantially lower, and associated with predictable toxicity arising from the administration of chemotherapeutic agents to ensure adequate HSC engraftment. SCID is a particularly attractive target for gene therapy as a profound growth and survival advantage is conferred to corrected cells (though this may be variable between different molecular types). In other words, owing to the huge proliferative capacity of the hematopoietic system (and particularly the lymphoid compartment), effective gene transfer to a small proportion of bone marrow precursor cells can result in substantial correction of the immunological deficit. This is most clearly witnessed by the renewed development of lymphocytes in patients with rare somatic gene reversion events (Hirschhorn, 2003). Other primary immune deficiencies may be immediately less severe, but are associated with significant accumulative morbidity and mortality. The difficulties associated with conventional HSC transplantation

have, therefore, driven the development of novel gene therapy strategies that have recently produced remarkable clinical effects.

2.1. Adenosine deaminase deficiency (ADA-D)

Deficiency of the purine salvage enzyme adenosine deaminase (ADA) accounts for approximately 10–20% of all SCID. ADA catalyzes the deamination of deoxyadenosine (dAdo), and adenosine to deoxyinosine and inosine respectively. Deficiency of ADA leads to the accumulation of the metabolites deoxyATP (dATP) and dAdo, which have profound effects on lymphocyte development and function through a number of cellular mechanisms. There is variation in the severity of the condition but most ADA patients have very low numbers of T and B lymphocytes. An alternative modality of treatment to that of HSC transplantation is regular exogenous enzyme replacement with polyethylene glycol-conjugated bovine ADA. This can result in correction of metabolic and immunological abnormalities albeit only partially in a significant number of cases (Hershfield, 2004). The first human gene therapy studies were conducted on patients with ADA deficiency in the early 1990s (Blaese *et al*., 1995; Bordignon *et al*., 1995; Hoogerbrugge *et al*., 1996; Kohn *et al*., 1998). Peripheral blood lymphocytes and/or HSCs were used as targets for gammaretroviral vector-mediated gene transfer, and though overall results were disappointing, long-term persistence of transduced cells was clearly demonstrated (Muul *et al*., 2003; Schmidt *et al*., 2003). The most important reason for the limited success of these studies was probably the continued administration of PEG-ADA. This compromised the efficient engraftment and development of transduced cells by removing their selective growth and survival advantage over nontransduced counterparts. This suggestion is supported by evidence of increasing levels of transduced peripheral blood T lymphocytes in two patients in whom PEG-ADA was discontinued sometime after gene therapy (Aiuti *et al*., 2002b; Kohn *et al*., 1998). Furthermore, results from a new study have clearly demonstrated that ADA-SCID can be successfully treated by gammaretroviral vector-mediated gene therapy in the absence of concomitant enzyme replacement (Aiuti *et al*., 2002a).

In this recent successful study, two important changes were incorporated into the protocol that may have influenced the positive outcome. First, patients were not commenced on PEG-ADA (or PEG-ADA was discontinued), and second, patients received low-intensity myelosuppressive bone marrow conditioning using an alkylating agent to facilitate engraftment of transduced cells (4 mg kg^{-1} of Busulphan as 2 mg kg^{-1} on two successive days). All patients treated to date have demonstrated an impressive recovery of immunological function, directly attributable to the development of new immunologically competent cells from transduced precursors (the absolute requirement for a conditioning procedure has not been rigorously tested in clinical trials. However, two additional recent studies in which transduced cells were administered concomitantly or after withdrawal of PEG-ADA, in the absence of prior chemotherapy, have not resulted in significant clinical recovery). Some variability in the levels of recovery may reflect the effectiveness of bone marrow suppression, or the dose of transduced cells in individual patients. However, in all cases, there has been impressive correction of

the metabolic defects, comparable to that achieved following successful allogeneic bone marrow transplantation. Importantly, there is also clear evidence for stable transduction of multipotent HSCs. This has positive implications for the treatment of other hematopoietic disorders where similar levels of engraftment would be predicted to result in therapeutic benefit. A second similar study using a gibbon-ape leukemia virus (GALV)-pseudotyped gammaretrovirus incorporating the spleen focus forming virus (SFFV) LTR, and the Woodchuck posttranscriptional regulatory element (WPRE) in order to force maximal expression of ADA in transduced cells (and therefore to achieve maximal systemic detoxification) has also reported good clinical effects. In this case, an alternative conditioning regimen using melphalan (140 mg m^{-2} as a single dose) was employed, although it is too early to draw meaningful comparisons between the two studies.

2.2. *X-linked severe combined immunodeficiency (SCID-X1)*

X-linked severe combined immunodeficiency (SCID-X1) accounts for approximately 50–60% of all SCID, and is caused by mutations in the gene encoding the common cytokine receptor gamma chain (γc). This is a subunit of the cytokine receptor complex for interleukins (IL) 2, 4, 7, 9, 15, and 21. In the absence of γc signaling, many aspects of immune cell development and function are compromised. The classical immunophenotype of SCID-X1 is the absence of T and NK cells, and persistence of dysfunctional B cells (T-B+NK-SCID). If a genotypically matched donor is available, bone marrow transplantation is a highly successful procedure with a long-term survival rate of over 90%. The high survival rates are partly due to the fact that the absence of T and NK cells in SCID-X1 patients allows engraftment in the absence of myelosuppressive conditioning.

Many incremental advances in gene transfer technology have contributed to the successful application of gene therapy for SCID-X1 and ADA-D (including the optimization of cell culture and gene transfer conditions ex vivo), which complement the intrinsic profound selective growth advantage imparted to successfully transduced cells (for SCID-X1, this is probably even more potent than that observed following restoration of ADA activity). The first dramatic demonstration of effective somatic gene therapy in human disease was derived from a study on patients with SCID-X1 (Cavazzana-Calvo *et al*., 2000; Hacein-Bey-Abina *et al*., 2002). Here, an amphotropic gammaretroviral vector encoding a γc cDNA (regulated by Moloney murine leukemia virus long terminal repeat sequences) was used to transduce autologous $CD34^+$ cells ex vivo, which were reinfused into the patients in the absence of preconditioning (see Figure 1). In nearly all patients, NK cells appeared between 2 and 4 weeks after infusion of cells, followed by new thymic T-lymphocyte emigrants at 10–12 weeks. With some variation, the number and distribution of these T cells normalized rapidly (more rapidly than observed following haploidentical transplantation). They also appeared to function normally in terms of proliferative response to mitogens, T-cell receptor (TCR), and specific antigen stimulation, and to have a complex phenotypic and molecular diversity of TCR. Functionality of the humoral system was also restored, maybe not quite as effectively, but to a sufficient degree that discontinuation of immunoglobulin therapy

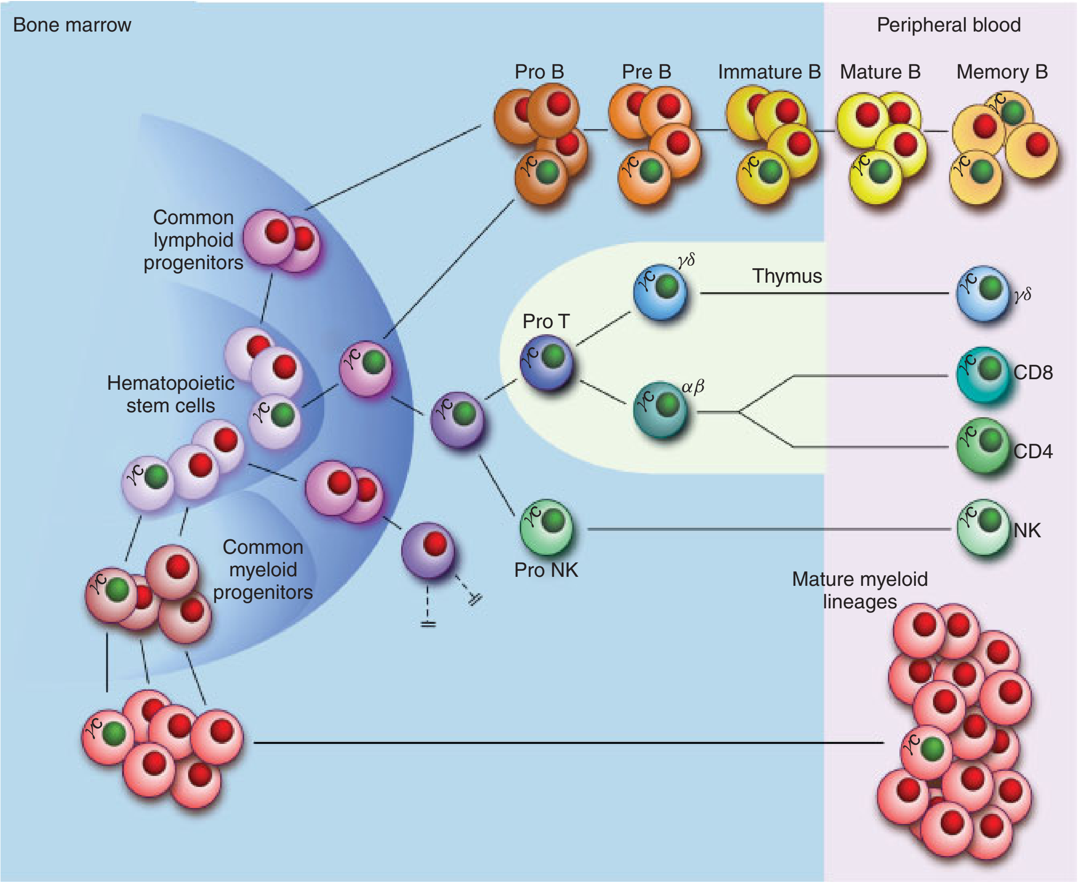


Figure 1 Schematic representation of the effects of gene therapy for SCID-X1 as demonstrated in recent clinical trials. In the absence of γc, T lymphocyte and natural killer (NK) cell differentiation is compromised, but when functionally corrected by transduction of some HSCs, the profound growth and survival advantage imparted to progenitors allows the differentiation and development of large numbers of mature circulating T and NK cells expressing the transgene. In human patients, B-lymphocyte differentiation occurs relatively normally in the absence of γc expression, although the cells have intrinsic functional defects. Gene-corrected B-cell lineages therefore do not appear to have clear selective advantage over mutated counterparts until the stage of "memory B cell" where some accumulation of γc-expressing cells has been reported. Restoration of γc expression to myeloid cells does not confer any selective advantage, and they differentiate normally in its absence. In this population, gene-corrected cells are, therefore, found at low levels (particularly because no preconditioning has been administered to patients to enhance HSC engraftment). Expression of functional γc is marked on gene-corrected cells (green nuclei). Nontransduced host cells are also shown (red nuclei)

was possible in most patients. A second study using a GALV-pseudotyped gammaretroviral vector and serum-free culture conditions has produced similar results (Gaspar *et al*., 2004). Persistent long-term marking in myeloid cells, albeit at low level, suggests that long-lived stem or progenitor cells have also been successfully transduced. The contribution to the initial burst of thymopoiesis from relatively late T-cell precursors in the original transduced CD34+ cell population, versus that from cells earlier in the hematopoietic differentiation hierarchy (or true HSCs)

that have engrafted in the bone marrow, has not yet been determined. This may have important implications for the durability of immunological reconstitution, and for sustained production of new T cells. Ultimately, the longevity of functional reconstitution can only be determined by clinical monitoring, but it may also be feasible to repeat gene therapy on multiple occasions. Definition of the effective window within which gene therapy will be effective is vitally important, as true for other more conventional therapeutic modalities. This has been clearly demonstrated by the failure of immunological reconstitution in two older patients following effective gene transfer to bone marrow CD34+ cells (Thrasher *et al*., 2005). At least for SCID, it is likely that there are host-related restrictions to efficacy, for example, due to the inability to reinitiate an exhausted or failed program of thymopoiesis.

2.3. Other forms of SCID as targets for gene therapy

The molecular basis of autosomal recessive T-B+NK-SCID is mutation of the receptor tyrosine kinase gene *JAK-3*. The dependence of γc on signaling through JAK-3 is responsible for a clinical and immunological phenotype identical to that of SCID-X1, and the rationale for gene therapy is, therefore, similar. Correction of a murine model of JAK-3 deficient SCID has been achieved using both myelosuppresssive, and more relevant to clinical studies, conditioning-free protocols (Bunting *et al*., 2000). Patients with mutations of the recombinase-activating genes *RAG-1* and *RAG-2* characteristically present with absence of both B and T cells. Murine gammaretroviral vectors have been shown to effectively reconstitute RAG-2 deficient mice in the absence of detectable toxicity, even though gene expression was not tightly regulated (Yates *et al*., 2002). One way to obviate toxicity arising from dysregulated gene expression in any condition, and to achieve physiological activity, is to correct genetic mutation by gene repair or homologous recombination. It has recently been shown that RAG2−/− mutant murine embryonic stem (ES) cells, repaired by standard homologous recombination technology, can be grown in vitro to provide sufficient hematopoietic progenitors for engraftment and correction of RAG-2 mutant mice (Rideout *et al*., 2002). This is the first example of gene therapy combined with a therapeutic cloning strategy, and clearly has important implications for future treatment of many genetic disorders.

2.4. Other primary immunodeficiencies as targets for gene therapy

Chronic granulomatous disease (CGD) is caused by mutations in genes encoding components of the phagocyte NADPH-oxidase complex, which is responsible for mediating efficient killing and digestion of many bacteria and fungi. In many ways, this has become a model disorder for testing HSC gene therapy strategies, as there is no growth or survival advantage for corrected cells (Barese *et al*., 2004; Malech, 1999). Gene expression is also only important in relatively short-lived terminally differentiated effector cells such as neutrophils and macrophages, meaning that long-term efficacy is entirely dependent on efficient stable transduction of HSCs. Important information can be obtained from the study of variant patients, who

retain partial NADPH-oxidase activity, and carriers of the X-linked form of the disease. From these, it can be predicted that over 10% correction in terms of cell numbers will be therapeutically effective, but that levels of correction per cell (in other words, the levels of enzyme activity) probably needs to be more than 30%. Therefore, the challenges are to achieve sufficient engraftment of transduced HSC, and efficient gene expression in terminally differentiated cells. Several clinical studies have been performed using standard gammaretroviral vectors, but in the absence of bone marrow preconditioning, only transient low-level correction has been achieved (Malech *et al*., 1997). More recently, studies have been initiated using low-intensity conditioning (busulphan or melphalan as for ADA gene therapy studies) to create space for incoming transduced HSC. These have provided good evidence for substantial correction associated with genuine therapeutic effect (clearance of infections), although prolonged follow-up is necessary to determine durability (communication from Dr. Manuel Grez, European Society for Gene Therapy meeting, 2004). It is likely that ongoing improvements in vector type and design will facilitate reliable high-level correction of this disease. Progress in gene transfer strategies for CGD will likely be directly translated to similar approaches for diseases such as Leukocyte Adhesion Deficiency type 1 (LAD-1), an inherited disorder of leukocyte function caused by defective expression of the common β_2-integrin subunit (CD18). As in CGD, no survival advantage is expected in gene-corrected cells and the critical cell population to be targeted is terminally differentiated neutrophils. As for CGD, one previous attempt at treating this disease by gammaretrovirus vector-mediated transfer into HSC, and engraftment into nonmyelosuppressed recipient patients resulted in only minimal and transient correction of myeloid cells (Bauer and Hickstein, 2000; Malech, 1999). The Wiskott–Aldrich syndrome (WAS) is also classed as a primary immunodeficiency, although an invariable feature is nonimmune microthrombocytopenia. The WAS protein (WASp) is expressed in all hematopoietic cell types, and is responsible for regulated organization of the actin cytoskeleton. For this disease, efficient expression of transgene must be achieved in multiple hematopoietic lineages, at levels that correct (at least partially) the cytoskeletal defect, without toxicity that may arise from overexpression. Regulation of near-physiological gene expression is, therefore, of key importance, and has implications for selection of vector and vector design. Following several studies demonstrating correction of cellular defects in WASp-deficient mice, clinical trials for treatment of patients with this disorder are planned for the near future (Charrier *et al*., 2005; Dupre *et al*., 2004; Klein *et al*., 2003; Strom *et al*., 2003).

3. Other applications of gene transfer to HSCs

Though rare, PID offer good models on which to study gene therapy strategies for other more common diseases such as beta thalassemia and sickle-cell disease (Persons and Tisdale, 2004; Sadelain, 2002). Each disease has its own unique requirement for achieving adequate levels of correction, and for correct regulation of gene expression. Considerable thought therefore needs to be invested in the design of safe but effective patient conditioning protocols, as most applications

will not benefit from preferential outgrowth of transduced cells. It may also be possible to select pharmacologically for transduced HSC in vivo as has been shown in animal studies using drug resistance genes such as methylguanine methyltransferase (MGMT) (Milsom and Fairbairn, 2004). MGMT encodes a DNA-repair enzyme that confers resistance to the combination of the MGMT inhibitor O(6)-benzylguanine (O(6)BG) and nitrosourea drugs such as carmustine, and methylating agents such as temozolomide (Neff *et al*., 2005). This strategy may therefore have utility for chemoprotection of bone marrow in cancer studies, as well as for the forced selection in vivo of cells cotransduced with a therapeutic transgene. Alternatively, the level of gene-corrected cells can be amplified in vivo without the need for myelosuppression using selective amplifier genes (SAGs). These are chimeric molecules that allow specific pharmacological agents to trigger selective proliferation of gene-modified cells. Preliminary experiments in mice and nonhuman primates have indicated the feasibility of such an approach, and the applicability to HSC diseases such as CGD (Hara *et al*., 2004; Neff and Blau, 2001; Ueda *et al*., 2004). Some genes will, in addition, require complex regulation, which may create profound challenges for vector design (*see* Article 104, **Control of transgene expression in mammalian cells**, Volume 2). For beta-thalassemia, significant advances have been made using lentiviral vectors to obtain high-level expression of complex globin gene cassettes. Therapeutic correction in murine models of both beta-thalassemia and sickle-cell anemia has been achieved using this approach (Hanawa *et al*., 2004; Imren *et al*., 2004; May *et al*., 2000; Puthenveetil *et al*., 2004). HSC and their progeny can be used as vehicles for disseminating therapeutic gene products throughout tissues such as the central nervous system. Such a strategy is under investigation for treatment of lysosomal storage disorders such as metachromatic leukodystrophy (MLD) and Gaucher disease type 1, and the peroxisomal disorder, adrenoleukodystrophy (ALD) (Benhamida *et al*., 2003; Biffi *et al*., 2004). Similarly, HSCs are considered good targets for augmentation with genes that will promote resistance of progeny to infection by pathogenic viruses such as HIV (Fanning *et al*., 2003).

4. Insertional mutagenesis and risks of HSC gene therapy

The dependence of retroviruses on chromosomal integration for stability of transduction brings with it the risk of insertional mutagenesis. On the basis of numerous animal studies and over 300 clinical trials in which patients have received gammaretroviral vectors, the risk of clinically manifesting insertional mutagenesis has been judged to be low. However, reproducible leukemogenesis and oncogenesis have now clearly been demonstrated in preclinical models, and may be directly associated with vector dose or cell copy number. Cooperating effects from expression of the transgene, or from other elements within the vector backbone may also be important, and are likely to be context dependent (Baum *et al*., 2004). In human clinical trials, three patients with SCID-X1 having initially achieved successful immunological reconstitution developed T-cell lymphoproliferative disease approximately 3 years after the gene therapy procedure (Hacein-Bey-Abina *et al*., 2003). In at least two of these patients, retroviral vector insertion into or near the *LMO-2*

proto-oncogene resulted in high-level expression of LMO-2 in the clones, as a result of retroviral enhancer-mediated activation of transcription. Activation of LMO-2 is known to participate in human leukemogenesis by chromosomal translocation, and results in the development of T-cell lymphoproliferation and leukemia in mice, albeit with a long latency. It is therefore likely that other contributing factors are required for these events to manifest. One consideration is a contribution from the activity of the γc transgene, although there is currently no evidence of dysregulated expression in lymphoid cells. Interestingly, at least one tumor derived in susceptible mice following infection with replication gammaretroviruses has been shown to harbor separate but coincident integrations at the γc and LMO-2 gene loci, suggesting that there may be a significant synergistic interaction (Dave *et al*., 2004). Cells with high proliferative potential such as HSC and thymocytes are also likely to be more susceptible to transformation following an insertional event than quiescent cells if they acquire additional adverse mutations unrelated to the gene therapy itself.

5. Future prospects for HSC gene therapy

Much can be done to improve efficiency and safety of current protocols. The design of vectors used for gene delivery is clearly important and modifications may be possible that will limit the risks of mutagenesis, for example, by incorporation of DNA and RNA insulator sequences in integrating vectors, by the use of self-inactivating vectors (in which the powerful viral LTR enhancer sequences are deleted), or by targeting safe regions in the genome. The detailed molecular analysis of insertion events in patients undergoing HSC gene therapy will greatly assist in the delineation of favored integration points within the genome, but is unlikely to be able to predict potential for mutagenesis unless recurrent hotspots associated with clinical disease become evident. Patterns of integration into host chromosomes are also to some degree vector dependent and could thereby contribute to the likelihood of inadvertent gene activation. For example, gammaretroviruses have been shown to integrate preferentially around the transcriptional start site of genes. Methods to minimize the number of integration events per cell, and to limit the number of engrafting clones by more stringent purification of HSC populations, may therefore be beneficial. Elimination of powerful viral enhancer sequences that can dysregulate gene expression over large chromatin domains, and replacement with more physiological and tissue specific regulatory elements may be feasible, and is under investigation for several applications. Lentiviral vectors in particular, provide greater capacity for incorporation of more complex and physiological regulatory sequences. The relative risk for each type of vector modification needs to be determined in clinically relevant animal-model systems and the effectiveness of these models to predict side effects in humans will have to be validated. The development of homologous recombination or gene repair to correct mutations, or the construction of mitotically stable extrachromosomal vectors would obviate many of these problems, but current technologies are inefficient. The applicability of any novel therapy, including HSC gene therapy, ultimately depends on the balance of risks against those of alternative treatments. The accurate characterization of adverse events, the utilization of protocols to test toxicity in a rigorous way, and

the development of methods to minimize risks yet retaining efficacy are therefore essential.

References

Ailles L, Schmidt M, Santoni de Sio FR, Glimm H, Cavalieri S, Bruno S, Piacibello W, von Kalle C and Naldini L (2002) Molecular evidence of lentiviral vector-mediated gene transfer into human self-renewing, multi-potent, long-term NOD/SCID repopulating hematopoietic cells. *Molecular Therapy*, **6**(5), 615–626.

Aiuti A, Slavin S, Aker M, Ficara F, Deola S, Mortellaro A, Morecki S, Andolfi G, Tabucchi A, Carlucci F, *et al*. (2002a) Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science*, **296**(5577), 2410–2413.

Aiuti A, Vai S, Mortellaro A, Casorati G, Ficara F, Andolfi G, Ferrari G, Tabucchi A, Carlucci F, Ochs HD, *et al*. (2002b) Immune reconstitution in ADA-SCID after PBL gene therapy and discontinuation of enzyme replacement. *Nature Medicine*, **8**(5), 423–425.

Antoine C, Muller S, Cant A, Cavazzana-Calvo M, Veys P, Vossen J, Fasth A, Heilmann C, Wulffraat N, Seger R, *et al*. (2003) Long-term survival and transplantation of haemopoietic stem cells for immunodeficiencies: report of the European experience 1968-99. *Lancet*, **361**(9357), 553–560.

Barese CN, Goebel WS and Dinauer MC (2004) Gene therapy for chronic granulomatous disease. *Expert Opinion on Biological Therapy*, **4**(9), 1423–1434.

Bauer TR Jr and Hickstein DD (2000) Gene therapy for leukocyte adhesion deficiency. *Current Opinion in Molecular Therapeutics*, **2**(4), 383–388.

Baum C, von Kalle C, Staal FJ, Li Z, Fehse B, Schmidt M, Weerkamp F, Karlsson S, Wagemaker G and Williams DA (2004) Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Molecular Therapy*, **9**(1), 5–13.

Benhamida S, Pflumio F, Dubart-Kupperschmitt A, Zhao-Emonet JC, Cavazzana-Calvo M, Rocchiccioli F, Fichelson S, Aubourg P, Charneau P and Cartier N (2003) Transduced CD34+ cells from adrenoleukodystrophy patients with HIV-derived vector mediate long-term engraftment of NOD/SCID mice. *Molecular Therapy*, **7**(3), 317–324.

Bhatia M, Bonnet D, Wu D, Murdoch B, Wrana J, Gallacher L and Dick JE (1999) Bone morphogenetic proteins regulate the developmental program of human hematopoietic stem cells. *Journal of Experimental Medicine*, **189**(7), 1139–1148.

Biffi A, De Palma M, Quattrini A, Del Carro U, Amadio S, Visigalli I, Sessa M, Fasano S, Brambilla R, Marchesini S, *et al*. (2004) Correction of metachromatic leukodystrophy in the mouse model by transplantation of genetically modified hematopoietic stem cells. *The Journal of Clinical Investigation*, **113**(8), 1118–1129.

Blaese RM, Culver KW, Miller AD, Carter CS, Fleisher T, Clerici M, Shearer G, Chang L, Chiang Y and Tolstoshev P (1995) T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. *Science*, **270**(5235), 475–480.

Bordignon C, Notarangelo LD, Nobili N, Ferrari G, Casorati G, Panina P, Mazzolari E, Maggioni D, Rossi C and Servida P (1995) Gene therapy in peripheral blood lymphocytes and bone marrow for ADA- immunodeficient patients. *Science*, **270**(5235), 470–475.

Buckley RH (2004) Molecular defects in human severe combined immunodeficiency and approaches to immune reconstitution. *Annual Review of Immunology*, **22**, 625–655.

Buckley RH, Schiff SE, Schiff RI, Markert L, Williams LW, Roberts JL, Myers LA and Ward FE (1999) Hematopoietic stem-cell transplantation for the treatment of severe combined immunodeficiency. *The New England Journal of Medicine*, **340**(7), 508–516.

Bunting KD, Lu T, Kelly PF and Sorrentino BP (2000) Self-selection by genetically modified committed lymphocyte precursors reverses the phenotype of JAK3-deficient mice without myeloablation. *Human Gene Therapy*, **11**(17), 2353–2364.

Case SS, Price MA, Jordan CT, Yu XJ, Wang L, Bauer G, Haas DL, Xu D, Stripecke R, Naldini L, *et al*. (1999) Stable transduction of quiescent CD34(+)CD38(−) human hematopoietic cells

by HIV-1-based lentiviral vectors. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(6), 2988–2993.

Cavazzana-Calvo M, Hacein-Bey S, de Saint BG, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, *et al*. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**(5466), 669–672.

Challita PM and Kohn DB (1994) Lack of expression from a retroviral vector after transduction of murine hematopoietic stem cells is associated with methylation in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **91**(7), 2567–2571.

Charrier S, Stockholm D, Seye K, Opolon P, Taveau M, Gross DA, Bucher-Laurent S, Delenda C, Vainchenker W, Danos O, *et al*. (2005) A lentiviral vector encoding the human Wiskott-Aldrich syndrome protein corrects immune and cytoskeletal defects in WASP knockout mice. *Gene Therapy*, **12**, 597–606.

Dave UP, Jenkins NA and Copeland NG (2004) Gene therapy insertional mutagenesis insights. *Science*, **303**(5656), 333.

Demaison C, Parsley K, Brouns G, Scherr M, Battmer K, Kinnon C, Grez M and Thrasher AJ (2002) High-level transduction and gene expression in hematopoietic repopulating cells using a human immunodeficiency [correction of imunodeficiency] virus type 1-based lentiviral vector containing an internal spleen focus forming virus promoter. *Human Gene Therapy*, **13**(7), 803–813.

Dupre L, Trifari S, Follenzi A, Marangoni F, Lain dL, Bernad A, Martino S, Tsuchiya S, Bordignon C, Naldini L, *et al*. (2004) Lentiviral vector-mediated gene transfer in T cells from Wiskott-Aldrich syndrome patients leads to functional correction. *Molecular Therapy*, **10**(5), 903–915.

Fanning G, Amado R and Symonds G (2003) Gene therapy for HIV/AIDS: the potential for a new therapeutic regimen. *The Journal of Gene Medicine*, **5**(8), 645–653.

Fischer A (2004) Human primary immunodeficiency diseases: a perspective. *Nature Immunology*, **5**(1), 23–30.

Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, Sinclair J, Brouns G, Schmidt M, von Kalle C, Barington T, *et al*. (2004) Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet*, **364**(9452), 2181–2187.

Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay JP, Thrasher AJ, Wulffraat N, Sorensen R, Dupuis-Girod S, *et al*. (2002) Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *New England Journal of Medicine*, **346**(16), 1185–1193.

Hacein-Bey-Abina S, von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E, *et al*. (2003) LM02-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**(5644), 415–419.

Hanawa H, Hargrove PW, Kepes S, Srivastava DK, Nienhuis AW and Persons DA (2004) Extended beta-globin locus control region elements promote consistent therapeutic expression of a gamma-globin lentiviral vector in murine beta-thalassemia. *Blood*, **104**(8), 2281–2290.

Hara T, Kume A, Hanazono Y, Mizukami H, Okada T, Tsurumi H, Moriwaki H, Ueda Y, Hasegawa M and Ozawa K (2004) Expansion of genetically corrected neutrophils in chronic granulomatous disease mice by cotransferring a therapeutic gene and a selective amplifier gene. *Gene Therapy*, **11**(18), 1370–1377.

Hershfield MS (2004) Combined immune deficiencies due to purine enzymes defects. In *Immunologic Disorders in Infants and Children*, *Fifth Edition*, Stiehm ER, Ochs HD and Winkelstein JA (Eds.), Elsevier-Saunders: Philadelphia, pp. 480–504.

Hirschhorn R (2003) In vivo reversion to normal of inherited mutations in humans. *Journal of Medical Genetics*, **40**(10), 721–728.

Hoogerbrugge PM, van Beusechem VW, Fischer A, Debree M, Le Deist F, Perignon JL, Morgan G, Gaspar B, Fairbanks LD, Skeoch CH, *et al*. (1996) Bone marrow gene transfer in three patients with adenosine deaminase deficiency. *Gene Therapy*, **3**(2), 179–183.

Imren S, Fabry ME, Westerman KA, Pawliuk R, Tang P, Rosten PM, Nagel RL, Leboulch P, Eaves CJ and Humphries RK (2004) High-level beta-globin expression and preferred intragenic

integration after lentiviral transduction of human cord blood stem cells. *The Journal of Clinical Investigation*, **114**(7), 953–962.

Josephson NC, Vassilopoulos G, Trobridge GD, Priestley GV, Wood BL, Papayannopoulou T and Russell DW (2002) Transduction of human NOD/SCID-repopulating cells with both lymphoid and myeloid potential by foamy virus vectors. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(12), 8295–8300.

Klein C, Nguyen D, Liu CH, Mizoguchi A, Bhan AK, Miki H, Takenawa T, Rosen FS, Alt FW, Mulligan RC, *et al*. (2003) Gene therapy for Wiskott-Aldrich syndrome: rescue of T-cell signaling and amelioration of colitis upon transplantation of retrovirally transduced hematopoietic stem cells in mice. *Blood*, **101**(6), 2159–2166.

Klug CA, Cheshier S and Weissman IL (2000) Inactivation of a GFP retrovirus occurs at multiple levels in long-term repopulating stem cells and their differentiated progeny. *Blood*, **96**(3), 894–901.

Kohn DB, Hershfield MS, Carbonaro D, Shigeoka A, Brooks J, Smogorzewska EM, Barsky LW, Chan R, Burotto F, Annett G, *et al*. (1998) T lymphocytes with a normal ADA gene accumulate after transplantation of transduced autologous umbilical cord blood CD34+ cells in ADA- deficient SCID neonates. *Nature Medicine*, **4**(7), 775–780.

Malech HL (1999) Progress in gene therapy for chronic granulomatous disease. *Journal of Infectious Diseases*, **179**(Suppl 2), S318–S325.

Malech HL, Maples PB, Whiting-Theobald N, Linton GF, Sekhsaria S, Vowells SJ, Li F, Miller JA, DeCarlo E, Holland SM, *et al*. (1997) Prolonged production of NADPH oxidase-corrected granulocytes after gene therapy of chronic granulomatous disease. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(22), 12133–12138.

May C, Rivella S, Callegari J, Heller G, Gaensler KM, Luzzatto L and Sadelain M (2000) Therapeutic haemoglobin synthesis in beta-thalassaemic mice expressing lentivirus-encoded human beta-globin. *Nature*, **406**(6791), 82–86.

Milsom MD and Fairbairn LJ (2004) Protection and selection for gene therapy in the hematopoietic system. *The Journal of Gene Medicine*, **6**(2), 133–146.

Muul LM, Tuschong LM, Soenen SL, Jagadeesh GJ, Ramsey WJ, Long Z, Carter CS, Garabedian EK, Alleyne M, Brown M, *et al*. (2003) Persistence and expression of the adenosine deaminase gene for 12 years and immune reaction to gene transfer components: long-term results of the first clinical gene therapy trial. *Blood*, **101**(7), 2563–2569.

Neff T, Beard BC, Peterson LJ, Anandakumar P, Thompson J and Kiem HP (2005) Polyclonal chemoprotection against temozolomide in a large-animal model of drug resistance gene therapy. *Blood*, **105**(3), 997–1002.

Neff T and Blau CA (2001) Pharmacologically regulated cell therapy. *Blood*, **97**(9), 2535–2540.

Persons DA and Tisdale JF (2004) Gene therapy for the hemoglobin disorders. *Seminars in Hematology*, **41**(4), 279–286.

Piacibello W, Bruno S, Sanavio F, Droetto S, Gunetti M, Ailles L, Santoni dS, Viale A, Gammaitoni L, Lombardo A, *et al*. (2002) Lentiviral gene transfer and ex vivo expansion of human primitive stem cells capable of primary, secondary, and tertiary multilineage repopulation in NOD/SCID mice. Nonobese diabetic/severe combined immunodeficient. *Blood*, **100**(13), 4391–4400.

Puthenveetil G, Scholes J, Carbonell D, Qureshi N, Xia P, Zeng L, Li S, Yu Y, Hiti AL, Yee JK, *et al*. (2004) Successful correction of the human beta-thalassemia major phenotype using a lentiviral vector. *Blood*, **104**(12), 3445–3453.

Rideout WM III, Hochedlinger K, Kyba M, Daley GQ and Jaenisch R (2002) Correction of a genetic defect by nuclear transplantation and combined cell and gene therapy. *Cell*, **109**(1), 17–27.

Sadelain M (2002) Globin gene transfer for the treatment of severe hemoglobinopathies: a paradigm for stem cell-based gene therapy. *The Journal of Gene Medicine*, **4**(2), 113–121.

Schmidt M, Carbonaro DA, Speckmann C, Wissler M, Bohnsack J, Elder M, Aronow BJ, Nolta JA, Kohn DB and von Kalle C (2003) Clonality analysis after retroviral-mediated gene transfer

to CD34+ cells from the cord blood of ADA-deficient SCID neonates. *Nature Medicine*, **9**(4), 463–468.

Sorrentino BP (2004) Clinical strategies for expansion of haematopoietic stem cells. *Nature Reviews. Immunology*, **4**(11), 878–888.

Strom TS, Turner SJ, Andreansky S, Liu H, Doherty PC, Srivastava DK, Cunningham JM and Nienhuis AW (2003) Defects in T-cell-mediated immunity to influenza virus in murine Wiskott-Aldrich syndrome are corrected by oncoretroviral vector-mediated gene transfer into repopulating hematopoietic cells. *Blood*, **102**(9), 3108–3116.

Thrasher AJ, Hacein-Bey-Abina S, Gaspar HB, Blanche S, Davies EG, Parsley K, Gilmour K, King D, Howe S, Sinclair J, *et al*. (2005) Failure of SCID-X1 gene therapy in older patients. *Blood*.

Ueda K, Hanazono Y, Shibata H, Ageyama N, Ueda Y, Ogata S, Tabata T, Nagashima T, Takatoku M, Kume A, *et al*. (2004) High-level in vivo gene marking after gene-modified autologous hematopoietic stem cell transplantation without marrow conditioning in nonhuman primates. *Molecular Therapy*, **10**(3), 469–477.

Vassilopoulos G, Trobridge G, Josephson NC and Russell DW (2001) Gene transfer into murine hematopoietic stem cells with helper-free foamy virus vectors. *Blood*, **98**(3), 604–609.

Yao S, Sukonnik T, Kean T, Bharadwaj RR, Pasceri P and Ellis J (2004) Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications. *Molecular Therapy*, **10**(1), 27–36.

Yates F, Malassis-Seris M, Stockholm D, Bouneaud C, Larousserie F, Noguiez-Hellin P, Danos O, Kohn DB, Fischer A, De Villartay JP, *et al*. (2002) Gene therapy of RAG-2-/- mice: sustained correction of the immunodeficiency. *Blood*, **100**(12), 3942–3949.

Specialist Review

Gene therapy in the central nervous system

Shyam Goverdhana, Maria G. Castro and Pedro R. Lowenstein

Cedars-Sinai Medical Center, University of California, Los Angeles, CA, USA

1. Obstacles in gene delivery to the CNS: structural and innate barriers

The brain is separated from the circulating bloodstream by the blood brain barrier (BBB). The BBB is a selective physical barrier formed by endothelial cells that display tight junctions of high electrical resistance providing an effective barrier against the entry of large and highly polar compounds. Accessing the brain requires crossing through the BBB or via the cerebrospinal fluid that circulates within the CNS (central nervous system) or through direct injection into the brain.

Approaches to cross the BBB to deliver viral vectors or chemotherapeutic agents to the brain parenchyma are osmotic disruption of the BBB, administration of vasomodulator compounds, and targeted vector administration via specific receptor binding. Osmotic BBB disruption approach has been employed in animal studies (Neuwelt *et al*., 1999) and also used in the clinic to promote enhanced chemotherapeutic agent delivery. Osmotic disruption facilitates transient BBB disruption by opening of the BBB to water-soluble drugs and macromolecules in vivo, and is achieved by infusing a hypertonic solution of arabinose or mannitol into the carotid artery (Neuwelt *et al*., 1980; Kaya *et al*., 2004; Sato *et al*., 1998; Liu *et al*., 2001). BBB opening involves widening of tight junctions between endothelial cells of the cerebralvasculature and is mediated by endothelial cell shrinkage and vascular dilatation. Only transient opening of the BBB is achieved, and there is restricted delivery in the vicinity of administration. In contrast to targeted delivery, systemic delivery provides only minimal benefit as this approach lacks specificity and can induce systemic toxicity via the immune system. The viral vector cannot be selectively delivered to the target brain area and is lost and eliminated via circulation. This can result in potential toxicity from activation of macrophages and the complement system leading to significant inhibition of viral vector entry into the brain (Ikeda *et al*., 1999). The current and efficient mode of viral vector delivery in animal models and human patients is attained by direct stereotaxic injection, which normally involves the use of small amounts of the viral vector to achieve localized expression. The extent of distribution of transgene expression from the

injected site is as important in achieving an optimum therapeutic outcome. The distribution is dependent on the type of viral vector (replicating or nonreplicating), vector dose, the type of therapeutic/cytotoxic transgene encoded, the target site and cells, and principally, the mode of viral vector administration. Studies with lentiviral and AAV vectors have shown predominant infection within neurons (Consiglio *et al*., 2001; Davidson *et al*., 2000), whereas adenoviruses have shown to infect astrocytes and neurons (Smith-Arica *et al*., 2000). Upon viral vector injection, the virus binds to receptors specific to viral capsid proteins. The molecular mechanisms by which viral vectors infect cell types within the brain remains unclear, and further studies on viral uptake and infection will provide a greater insight into the mechanisms of viral vector spread and infection in the brain. Certain elements within transgene sequences have been speculated to have an effect on viral vector spread (Lowenstein and Castro, 2002). Adenoviral vector–mediated expression of β-galactosidase compared to vectors expressing thymidine kinase (tk) gene, derived from the herpes-simplex virus, demonstrated a larger dispersion of tk gene expression in the brain (Dewey *et al*., 1999). Further analysis of the above factors is crucial to study viral vector–based gene expression spread within the brain.

2. Gene therapy approaches to treat CNS diseases: tumors of the CNS

CNS tumors are one of the most serious and frequent forms of cancer affecting both children and adults. Examples of major types of tumors include glioblastoma multiforme (GBM), acoustic neuroma, oligodendroglioma, and pituitary adenomas. The conventional methods of treating brain tumors have been chemotherapy, radiation therapy, and surgical tumor resection. As an alternative approach for brain tumor treatment, gene therapy has progressed into an important therapeutic candidate. The employment of animal brain tumor models, transgenic animals, and established and emerging tumor cell lines all have facilitated to further understand tumor biology, assess, and improve gene therapy and anticancer approaches. Commonly employed animal glioma models are the rat CNS1 model C6 glioma and 9L gliosarcoma, as well as the VM Dk murine astrocytoma and the GL261 murine glioma. The use of transgenic animals is still in its primitive phase (Holland *et al*., 2000; Reilly *et al*., 2000). Major strategies for brain tumor treatment include tumor suppressor gene therapy (Gomez-Manzano *et al*., 1996; Fueyo *et al*., 1998), inactivation of selective oncogenes (Cheney *et al*., 1998), repression of angiogenesis (Kirsch *et al*., 1998), cytokine-mediated enhancement of immune responses (Iwadate *et al*., 2001; Natsume *et al*., 2000), modified dendritic cell–based immunotherapy (Yu *et al*., 2001; Okada *et al*., 2001), antibody-directed prodrug therapy (Napier *et al*., 2000), and virus-mediated enzyme prodrug therapy. An effective tumor treatment strategy is suicide gene therapy in which a cytotoxic gene is selectively expressed in tumor cells, resulting in their death without affecting neighboring normal cells. One example of an enzyme prodrug therapy approach is the Herpes simplex virus (HSV) type 1 thymidine kinase (TK)/glanciclovir (HSV-TK) (Kraiselburd, 1976). Other suicide gene therapy approaches that involve

enzyme-drug complexes include cytosine deaminase/-fluorocytosine (Huber *et al*., 1994) and Carboxypeptidase G2 (Springer *et al*., 1991).

Amalgamation of conventional approaches with gene therapy have successfully been attempted and tested in human clinical trials using retroviral, replication-competent herpes simplex, and adenoviral vectors (Klatzmann *et al*., 1998; Packer *et al*., 2000; Rampling *et al*., 2000; Shand *et al*., 1999; Chiocca *et al*., 2004; Immonen *et al*., 2004). The use of high-capacity adenoviral vectors to selectively deliver and tightly regulate suicide or tumor gene expression would generate a further effective gene therapeutic strategy to the management of brain tumors (Thomas *et al*., 2000).

3. Degenerative diseases of the CNS

3.1. Parkinson's disease gene therapy

Parkinson's disease (PD), though a well-known neurodegenerative disease, its etiology remains unidentified. The disease is not a consequence of a single cause as there is proof of both genetic and environmental components contributing to the development of the disease. The pathological hallmarks are progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta of the basal ganglia. Distinctive cardinal signs of PD include recurring rest tremors, abnormal gait patterns, and unbalanced posture. Gene therapy for PD has made enormous advances in treatment strategies that regenerate and sustain dopaminergic neuronal circuitry and function, and curtailment of cardinal clinical manifestations of the disease.

PD gene therapy studies include prevention of neurodegeneration using lentiviral vector–mediated delivery of glial derived neurotropic factor (GDNF) in MPTP-induced animal models of PD (Kordower *et al*., 2000). Adenoviral vector–mediated delivery of GDNF also produced prolonged protection from dopaminergic neuron degeneration with reduced behavioral symptoms in 6-OHDA-induced animal models of PD (Do Thi *et al*., 2004). Improved HSV-1 amplicons and HSV-1 "disabled" vectors have also demonstrated their effectiveness in repairing this neuronal pathway. HSV-1 amplicon-mediated delivery of tyrosine hydroylase (TH) into the rat striatum reduced behavioral symptoms in 6-OHDA-induced animal models (During *et al*., 1994). Disabled HSV-1-mediated delivery of Bcl-2 to the substantia nigra prevented neuronal degeneration induced in 6-OHDA animals models (Yamada *et al*., 1999). HSV-1 vectors have also demonstrated their capacity to be carried in a retrograde process within the neuronal circuitry from the injection site, providing a suitable approach for achieving targeted delivery of therapeutic genes, which otherwise would be surgically complex (Lilley *et al*., 2001). This has important implications for treatments of neurodegenerative diseases such as PD, as one can precisely deliver therapeutic transgenes by means of neuronal retrograde transport to reach the affected substantia nigra.

Key issues that must be addressed before implementation of neurotrophic factor gene therapy for PD for human clinical trials are (1) the safety of the engineered viral vector and the extent of side effects from the vector dosage given, as

determined from preclinical studies; (2) capacity to control stable and long-term expression of the trophic factor, and (3) the ability to generate negligible levels of immune responses to the elements of the transgene and the viral vector. Regulation of therapeutic transgene expression at the transcriptional level is essential, and can be achieved by switching transgene expression "on" and "off" appropriately.

4. Inherited metabolic and dominant diseases

Common inherited metabolic diseases include phenylketonuria (PKU), cystinuria, and albinism. Examples of dominant inherited diseases include Huntington disease (HD), amyotrophic lateral sclerosis (ALS), and Kennedy's disease. Common aspects of these diseases are that they are adult-onset and progressive, have complex and poorly understood pathophysiology, and induce neuron degeneration and subsequent impairment of motor function. Defined mechanisms by which mutated proteins are leading to neuronal injury in dominant diseases are unclear. Some studies demonstrate possible self-activation mechanisms resulting in the pathogenic process (Zoghbi *et al*., 2000). HD, an autosomal dominant disorder, is a consequence of a trinucleotide CAG expansion at the IT15 locus of chromosome 4 within the Huntington protein whose role is yet unclear. The molecular mechanisms and pathogenesis of the trinucleotide expansion leading to polyglutamine-containing aggregates and subsequent neuronal death are still not clearly known, thus limiting the development of effective therapeutic strategies. The excessive CAG repeats result in degeneration of striatal and cortical neurons in Huntington's and spinal cord in ALS and Kennedy's causing abnormalities in cognitive, motor functions and progressive dementia. There is currently no cure or an effective treatment available for inherited CNS diseases. Gene therapy offers potential hope in providing effective treatment modalities for HD and other inherited CNS disorders, principally involving the use of therapeutic growth factors that are known to enhance neuronal survival neurons. Studies on neurotrophic factor gene therapy using different viral vectors demonstrated their potential in achieving neuronal protection in animal models with HD. AAV vector–mediated delivery of GDNF has been shown to successfully deliver the protein within the striatum providing neuronal and behavioral protection in animal models with HD (McBride *et al*., 2003). First-generation adenoviral vector–mediated delivery of ciliary neurotropic factor (CNTF) and brain derived neurotropic factor (BDNF) also generated widespread neuroprotection in the striatum (Mittoux *et al*., 2002; Bemelmans *et al*., 1999). Using a dual vector approach, lentiviral-mediated tetracycline-regulated delivery of CNTF produced regulated CNTF expression in the striatum using the tet-off regulatory system (Regulier *et al*., 2002). Amyotrophic Lateral Sclerosis (ALS), another neurodegenrative disease, involves advancing degeneration of motor neurons in the spinal cord and the brain, resulting in the patient being paralyzed. A number of hypotheses on the underlying pathogenic mechanisms involved in ALS have been put forward, principally, autoimmunity and excitotoxicity and its potential treatment with I6F1 (Kaspar *et al*., 2003). Further studies at the molecular level will shed light into isolating prominent players in the onset and progression of these diseases.

4.1. *RNA interference strategies to treat inherited dominant diseases*

A promising approach under extensive research for the treatment of dominant inherited diseases is the use of small interfering (si) RNAs. Because dominant diseases commonly arise as a consequence of disease-causing mutated gene sequences, siRNA-based approaches can silence mutated genes responsible for causing the disease. Yet, several dominant disease genes may also encode essential proteins and it will be crucial to inactivate specifically mutant alleles while preserving expression of the nonmutated gene. SiRNA-based inactivation of dominant disease genes emerged from a number of in vitro studies. A pathological hallmark of the inherited dominant familial disease ALS is the occurrence of a point mutation of the *SOD1* gene, which eventually causes symptoms of progressive muscle weakness and degeneration. SiRNA silencing of a mutated version of superoxide dismutase (SOD1) gene resulted in the prevention of cell death from cyclosporin-A toxin insult, in which efficacy and selectivity of allele-specific silencing was achieved using wild type and mutated forms of the HSOD gene (Maxwell *et al*., 2004).

In polyglutamine neurodegenerative disorders such as Machado–Joseph disease and spinocerebellar ataxia type 3 diseases, the characteristic feature is the presence of polyglutamine (polyQ) degeneration, the consequence of CAG-repeat expansions in a particular disease-causing mutated gene. As the normal form of the gene also contains a CAG-repeat, selective targeting to isolate the disease-causing CAG repeat within the expansion chain is difficult. To resolve this problem, allele-specific silencing of disease genes was achieved by using an associated linked single-nucleotide polymorphism (SNP) to generate siRNA that solely inactivated the mutant allele, (Miller *et al*., 2003). The report demonstrated the feasibility of silencing diseased genes differing by a single nucleotide. Other successful siRNA-based therapies include allele-specific silencing of the mutant Torsin A gene for inherited DYT1 dystonia, in which there is a three-nucleotide deletion in the *TOR1A* gene (Gonzalez-Alegre *et al*., 2003), and siRNA-based inhibition specific for mutant SOD1 with single-nucleotide alteration in familial ALS dominant disease (Yokota *et al*., 2004).

Like siRNA, short hairpin RNAs (shRNAs) under the control of RNA polymerase III (pol III) promoters have been shown to induce degradation of messenger RNAs (mRNAs) and subsequently inhibit targeted gene expression (McManus and Sharp, 2002; Brummelkamp *et al*., 2002). Pol III-based shRNA sequences engineered into viral vectors successfully generated RNAi (Brummelkamp *et al*., 2002; Rubinson *et al*., 2003). Using an engineered recombinant AAV vector, Xia *et al*. (2004) reported RNAi suppression of polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia (SCA1). SCA1 is a dominant polyglutamine expansion disease resulting in a progressive, untreatable neurodegeneration. In a mouse model of SCA1 caused by mutant ataxin-1, intracerebellar injection of AAV vector encoding short hairpin RNAs considerably enhanced motor coordination, sustained cerebellar morphology, and determined characteristic ataxin-1 inclusions in Purkinje cells (Xia *et al*., 2004).

5. CNS gene therapy in the near future

Gene therapy has now been proposed for the treatment of the major groups of brain diseases, including inherited metabolic brain diseases, brain tumors, neurodegenerations, and dominantly inherited diseases. While clinical trials are being pursued, the challenges of sustained and regulatable transgene expression, cytotoxicity, and immune responses remain. It is likely that within a few years gene therapies will become alternative available treatments for brain diseases, which will be offered to patients in addition to classical pharmacological or surgical approaches.

References

Bemelmans AP, Horellou P, Pradier L, Brunet I, Colin P and Mallet J (1999) Brain-derived neurotrophic factor-mediated protection of striatal neurons in an excitotoxic rat model of Huntington's disease, as demonstrated by adenoviral gene transfer. *Human Gene Therapy*, **10**, 2987–2997.

Brummelkamp TR, Bernards R and Agami R (2002) A system for stable expression of short interfering RNAs in mammalian cells. *Science*, **296**, 550–553.

Cheney IW, Johnson DE, Vaillancourt MT, Avanzini J, Morimoto A, Demers GW, Wills KN, Shabram PW, Bolen JB, Tavtigian SV, *et al*. (1998) Suppression of tumorigenicity of glioblastoma cells by adenovirus-mediated MMAC1/PTEN gene transfer. *Cancer Research*, **58**, 2331–2334.

Chiocca EA, Abbed KM, Tatter S, Louis DN, Hochberg FH, Barker F, Kracher J, Grossman SA, Fisher JD, Carson K, *et al*. (2004) A phase I open-label, dose-escalation, multi-institutional trial of injection with an E1B-Attenuated adenovirus, ONYX-015, into the peritumoral region of recurrent malignant gliomas, in the adjuvant setting. *Molecular Therapy*, **10**, 958–966.

Consiglio A, Quattrini A, Martino S, Bensadoun JC, Dolcetta D, Trojani A, Benaglia G, Marchesini S, Cestari V, Oliverio A, *et al*. (2001) In vivo gene therapy of metachromatic leukodystrophy by lentiviral vectors: correction of neuropathology and protection against learning impairments in affected mice. *Nature Medicine*, **7**, 310–316.

Davidson BL, Stein CS, Heth JA, Martins I, Kotin RM, Derksen TA, Zabner J, Ghodsi A and Chiorini JA (2000) Recombinant adeno-associated virus type 2, 4, and 5 vectors: transduction of variant cell types and regions in the mammalian central nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3428–3432.

Dewey RA, Morrissey G, Cowsill CM, Stone D, Bolognani F, Dodd NJ, Southgate TD, Klatzmann D, Lassmann H, Castro MG, *et al*. (1999) Chronic brain inflammation and persistent herpes simplex virus 1 thymidine kinase expression in survivors of syngeneic glioma treated by adenovirus-mediated gene therapy: implications for clinical trials. *Nature Medicine*, **5**, 1256–1263.

Do Thi NA, Saillour P, Ferrero L, Dedieu JF, Mallet J and Paunio T (2004) Delivery of GDNF by an E1,E3/E4 deleted adenoviral vector and driven by a GFAP promoter prevents dopaminergic neuron degeneration in a rat model of Parkinson's disease. *Gene Therapy*, **11**, 746–756.

During MJ, Naegele JR, O'Malley KL and Geller AI (1994) Long-term behavioral recovery in Parkinsonian rats by an HSV vector expressing tyrosine hydroxylase. *Science*, **266**, 1399–1403.

Fueyo J, Gomez-Manzano C, Yung WK, Liu TJ, Alemany R, Bruner JM, Chintala SK, Rao JS, Levin VA and Kyritsis AP (1998) Suppression of human glioma growth by adenovirus-mediated Rb gene transfer. *Neurology*, **50**, 1307–1315.

Gomez-Manzano C, Fueyo J, Kyritsis AP, Steck PA, Roth JA, McDonnell TJ, Steck KD, Levin VA and Yung WK (1996) Adenovirus-mediated transfer of the p53 gene produces rapid and generalized death of human glioma cells via apoptosis. *Cancer Research*, **56**, 694–699.

Gonzalez-Alegre P, Miller VM, Davidson BL and Paulson HL (2003) Toward therapy for DYT1 dystonia: allele-specific silencing of mutant TorsinA. *Annals of Neurology*, **53**, 781–787.

Holland EC, Celestino J, Dai C, Schaefer L, Sawaya RE and Fuller GN (2000) Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice. *Nature Genetics*, **25**, 55–57.

Huber BE, Austin EA, Richards CA, Davis ST and Good SS (1994) Metabolism of 5-fluorocytosine to 5-fluorouracil in human colorectal tumor cells transduced with the cytosine deaminase gene: significant antitumor effects when only a small percentage of tumor cells express cytosine deaminase. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 8302–8306.

Ikeda K, Ichikawa T, Wakimoto H, Silver JS, Deisboeck TS, Finkelstein D, Harsh GRt, Louis DN, Bartus RT, Hochberg FH, *et al*. (1999) Oncolytic virus therapy of multiple tumors in the brain requires suppression of innate and elicited antiviral responses. *Nature Medicine*, **5**, 881–887.

Immonen A, Vapalahti M, Tyynela K, Hurskainen H, Sandmair A, Vanninen R, Langford G, Murray N and Yla-Herttuala S (2004) AdvHSV-tk gene therapy with intravenous ganciclovir improves survival in human malignant glioma: a randomised, controlled study. *Molecular Therapy*, **10**, 967–972.

Iwadate Y, Yamaura A, Sato Y, Sakiyama S and Tagawa M (2001) Induction of immunity in peripheral tissues combined with intracerebral transplantation of interleukin-2-producing cells eliminates established brain tumors. *Cancer Research*, **61**, 8769–8774.

Kaspar BK, Llado J, Sherkat N, Rothstein JD and Gage FH (2003) Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science*, **301**(5634), 839–842.

Kaya M, Gulturk S, Elmas I, Kalayci R, Arican N, Kocyildiz ZC, Kucuk M, Yorulmaz H and Sivas A (2004) The effects of magnesium sulfate on blood-brain barrier disruption caused by intracarotid injection of hyperosmolar mannitol in rats. *Life Sciences*, **76**, 201–212.

Kirsch M, Strasser J, Allende R, Bello L, Zhang J and Black PM (1998) Angiostatin suppresses malignant glioma growth in vivo. *Cancer Research*, **58**, 4654–4659.

Klatzmann D, Valery CA, Bensimon G, Marro B, Boyer O, Mokhtari K, Diquet B, Salzmann JL and Philippon J (1998) A phase I/II study of herpes simplex virus type 1 thymidine kinase "suicide" gene therapy for recurrent glioblastoma. Study group on gene therapy for glioblastoma. *Human Gene Therapy*, **9**, 2595–2604.

Kordower JH, Emborg ME, Bloch J, Ma SY, Chu Y, Leventhal L, McBride J, Chen EY, Palfi S, Roitberg BZ, *et al*. (2000) Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. *Science*, **290**, 767–773.

Kraiselburd E (1976) Thymidine kinase gene transfer by herpes simplex virus. *Bulletin du Cancer*, **63**, 393–398.

Lilley CE, Groutsi F, Han Z, Palmer JA, Anderson PN, Latchman DS and Coffin RS (2001) Multiple immediate-early gene-deficient herpes simplex virus vectors allowing efficient gene delivery to neurons in culture and widespread gene delivery to the central nervous system in vivo. *Journal of Virology*, **75**, 4343–4356.

Liu Y, Hashizume K, Samoto K, Sugita M, Ningaraj N, Asotra K and Black KL (2001) Repeated, short-term ischemia augments bradykinin-mediated opening of the blood-tumor barrier in rats with RG2 glioma. *Neurological Research*, **23**, 631–640.

Lowenstein PR and Castro MG (2002) Progress and challenges in viral vector-mediated gene transfer to the brain. *Current Opinion in Molecular Therapeutics*, **4**, 359–371.

Maxwell MM, Pasinelli P, Kazantsev AG and Brown Jr RH (2004) RNA interference-mediated silencing of mutant superoxide dismutase rescues cyclosporin A-induced death in cultured neuroblastoma cells. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(9), 3178–3183.

McBride JL, During MJ, Wuu J, Chen EY, Leurgans SE and Kordower JH (2003) Structural and functional neuroprotection in a rat model of Huntington's disease by viral gene transfer of GDNF. *Experimental Neurology*, **181**, 213–223.

McManus MT and Sharp PA (2002) Gene silencing in mammals by small interfering RNAs. *Nature Reviews. Genetics*, **3**, 737–747.

Miller VM, Xia H, Marrs GL, Gouvion CM, Lee G, Davidson BL and Paulson HL (2003) Allele-specific silencing of dominant disease genes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7195–7200.

Mittoux V, Ouary S, Monville C, Lisovoski F, Poyot T, Conde F, Escartin C, Robichon R, Brouillet E, Peschanski M, *et al*. (2002) Corticostriatopallidal neuroprotection by adenovirus-mediated ciliary neurotrophic factor gene transfer in a rat model of progressive striatal degeneration. *The Journal of Neuroscience*, **22**, 4478–4486.

Napier MP, Sharma SK, Springer CJ, Bagshawe KD, Green AJ, Martin J, Stribbling SM, Cushen N, O'Malley D and Begent RH (2000) Antibody-directed enzyme prodrug therapy: efficacy and mechanism of action in colorectal carcinoma. *Clinical Cancer Research*, **6**, 765–772.

Natsume A, Tsujimura K, Mizuno M, Takahashi T and Yoshida J (2000) IFN-beta gene therapy induces systemic antitumor immunity against malignant glioma. *Journal of Neuro-oncology*, **47**, 117–124.

Neuwelt EA, Abbott NJ, Drewes L, Smith QR, Couraud PO, Chiocca EA, Audus KL, Greig NH and Doolittle ND (1999) Cerebrovascular Biology and the various neural barriers: challenges and future directions. *Neurosurgery*, **44**, 604–608, discussion 608–609.

Neuwelt EA, Frenkel EP, Rapoport S and Barnett P (1980) Effect of osmotic blood-brain barrier disruption on methotrexate pharmacokinetics in the dog. *Neurosurgery*, **7**, 36–43.

Okada H, Pollack IF, Lieberman F, Lunsford LD, Kondziolka D, Schiff D, Attanucci J, Edington H, Chambers W, Kalinski P, *et al*. (2001) Gene therapy of malignant gliomas: a pilot study of vaccination with irradiated autologous glioma and dendritic cells admixed with IL-4 transduced fibroblasts to elicit an immune response. *Human Gene Therapy*, **12**, 575–595.

Packer RJ, Raffel C, Villablanca JG, Tonn JC, Burdach SE, Burger K, LaFond D, McComb JG, Cogen PH, Vezina G, *et al*. (2000) Treatment of progressive or recurrent pediatric malignant supratentorial brain tumors with herpes simplex virus thymidine kinase gene vector-producer cells followed by intravenous ganciclovir administration. *Journal of Neurosurgery*, **92**, 249–254.

Rampling R, Cruickshank G, Papanastassiou V, Nicoll J, Hadley D, Brennan D, Petty R, MacLean A, Harland J, McKie E, *et al*. (2000) Toxicity evaluation of replication-competent herpes simplex virus (ICP 34.5 null mutant 1716) in patients with recurrent malignant glioma. *Gene Therapy*, **7**, 859–866.

Regulier E, Pereira de Almeida L, Sommer B, Aebischer P and Deglon N (2002) Dose-dependent neuroprotective effect of ciliary neurotrophic factor delivered via tetracycline-regulated lentiviral vectors in the quinolinic acid rat model of Huntington's disease. *Human Gene Therapy*, **13**, 1981–1990.

Reilly KM, Loisel DA, Bronson RT, McLaughlin ME and Jacks T (2000) Nf1;Trp53 mutant mice develop glioblastoma with evidence of strain-specific effects. *Nature Genetics*, **26**, 109–113.

Rubinson DA, Dillon CP, Kwiatkowski AV, Sievers C, Yang L, Kopinja J, Rooney DL, Ihrig MM, McManus MT, Gertler FB, *et al*. (2003) A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nature Genetics*, **33**, 401–406.

Sato S, Kawase T, Harada S, Takayama H and Suga S (1998) Effect of hyperosmotic solutions on human brain tumour vasculature. *Acta Neurochirurgica (Wien)*, **140**, 1135–1141, discussion 1141–1132.

Shand N, Weber F, Mariani L, Bernstein M, Gianella-Borradori A, Long Z, Sorensen AG and Barbier N (1999) A phase 1-2 clinical trial of gene therapy for recurrent glioblastoma multiforme by tumor transduction with the herpes simplex thymidine kinase gene followed by ganciclovir. GLI328 European-Canadian Study Group. *Human Gene Therapy*, **10**, 2325–2335.

Smith-Arica JR, Morelli AE, Larregina AT, Smith J, Lowenstein PR and Castro MG (2000) Cell-type-specific and regulatable transgenesis in the adult brain: adenovirus-encoded combined transcriptional targeting and inducible transgene expression. *Molecular Therapy*, **2**, 579–587.

Springer CJ, Bagshawe KD, Sharma SK, Searle F, Boden JA, Antoniw P, Burke PJ, Rogers GT, Sherwood RF and Melton RG (1991) Ablation of human choriocarcinoma xenografts in nude mice by antibody-directed enzyme prodrug therapy (ADEPT) with three novel compounds. *European Journal of Cancer*, **27**, 1361–1366.

Thomas CE, Schiedner G, Kochanek S, Castro MG and Lowenstein PR (2000) Peripheral infection with adenovirus causes unexpected long-term brain inflammation in animals injected intracranially with first-generation, but not with high-capacity, adenovirus vectors: toward

realistic long-term neurological gene therapy for chronic diseases. *Proceedings of The National Academic Science of the United States of America*, **97**(13), 7482–7487.

Xia H, Mao Q, Eliason SL, Harper SQ, Martins IH, Orr HT, Paulson HL, Yang L, Kotin RM and Davidson BL (2004) RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Medicine*, **10**, 816–820.

Yamada M, Oligino T, Mata M, Goss JR, Glorioso JC and Fink DJ (1999) Herpes simplex virus vector-mediated expression of Bcl-2 prevents 6-hydroxydopamine-induced degeneration of neurons in the substantia nigra in vivo. *Proceedings of The National Academic Science of the United States of America*, **96**, 4078–4083.

Yokota T, Miyagishi M, Hino T, Matsumura R, Tasinato A, Urushitani M, Rao RV, Takahashi R, Bredesen DE, Taira K, *et al*. (2004) siRNA-based inhibition specific for mutant SOD1 with single nucleotide alternation in familial ALS, compared with ribozyme and DNA enzyme. *Biochemical and Biophysical Research Communications*, **314**, 283–291.

Yu JS, Wheeler CJ, Zeltzer PM, Ying H, Finger DN, Lee PK, Yong WH, Incardona F, Thompson RC, Riedinger MS, *et al*. (2001) Vaccination of malignant glioma patients with peptide-pulsed dendritic cells elicits systemic cytotoxicity and intracranial T-cell infiltration. *Cancer Research*, **61**, 842–847.

Zoghbi HY, Gage FH and Choi DW (2000) Neurobiology of disease. *Current Opinion in Neurobiology*, **10**, 655–660.

Specialist Review

Cardiovascular gene therapy

Shalini Bhardwaj, Himadri Roy and Seppo Ylä-Herttuala
University of Kuopio, Kuopio, Finland

1. Introduction

Recent advances in gene transfer technologies and better understanding of molecular and genetic bases of cardiovascular disease have made gene therapy an emerging alternative treatment strategy. Promising results have been obtained in animal models of restenosis and vein graft thickening, and limb and cardiac ischemia. Gene therapy for the induction of angiogenesis is based on the concept that myocardial and peripheral ischemia can be improved by stimulating neovessel formation and collateral development from the existing vasculature. Therapeutic vascular growth includes stimulation of angiogenesis, arteriogenesis, and lymphangiogenesis (Ylä-Herttuala and Martin, 2000). Prevention of restenosis can be achieved by inhibiting smooth muscle cell proliferation, migration, matrix synthesis, remodeling, and thrombosis.

Gene therapy has the potential advantage of enabling gene expression for a sufficiently long period, at an adequate concentration to stimulate effective therapeutic response from a single administration. However, full potential of this therapy can be achieved only after further development of gene transfer technology and selection of effective treatment genes.

2. Choice of angiogenic factors

Angiogenic signals are mediated by a number of growth factors and cytokines. The endothelial specific growth factors include members of the vascular endothelial growth factor (VEGF) family of proteins and the angiopoietin (Ang) family. Different members of the VEGF family act as key regulators of endothelial cell function, controlling vasculogenesis, angiogenesis, vascular permeability, and endothelial cell survival (Ferrara and Davis-Smyth, 1997). Other factors that promote angiogenesis are fibroblast growth factors (FGFs), hepatocyte growth factor (HGF), epherins, platelet derived growth factors (PDGFs), and hypoxia-inducible factor-1 (HIF-1). Vessel survival is dependent on VEGF and other exogenous survival factors. Angs act during remodeling of vascular plexus and a combination therapy with VEGF and Ang-1 may produce more stable vessels.

Another important aspect is induction of angiogenesis by angiogenic master switch genes, such as HIF-1α and HGF, which stimulate multiple neovascularization cascades.

2.1. Vascular endothelial growth factors (VEGFs)

VEGF family comprises of six members, VEGF-A, -B, -C, -D, -E, and placental growth factor (PIGF), which differ in their molecular mass and biological properties. The mechanism of action is through tyrosine kinase receptors, VEGFR-1, VEGFR-2, and VEGFR-3. VEGF-A is known to play a crucial role in angiogenesis and vasculogenesis and is a ligand for VEGFR-1 and VEGFR-2 (Ferrara, 2001). VEGF-A promotes increased microvascular permeability and fibrin deposition that may be responsible for enhanced migration of endothelial cells in extracellular matrix. It supports the survival of endothelial cells by expressing antiapoptotic proteins in these cells. In phase I and II clinical trials, VEGF-A plasmid/liposome or adenovirus vector has been used for coronary artery disease (CAD), in-stent restenosis, and peripheral artery occlusive disease. Vascular endothelial growth factor B (VEGF-B) is structurally related to VEGF-A and binds only to VEGFR-1 (Olofsson *et al*., 1998). VEGF-B is a very weak mitogen when tested in mammalian cells. The receptors for VEGF-B are located on endothelial cells; thereby, it is more likely to act in a paracrine manner.

Expression of VEGF-C occurs during early embryonic development before the emergence of lymphatics, which is suggestive of its role in vasculogenesis and angiogenesis. VEGF-D has angiogenic and lymphangiogenic potentials (Bhardwaj *et al*., 2003; Rissanen *et al*., 2003). Both VEGF-C and VEGF-D act through VEGFR-2 and VEGFR-3 (Hamada *et al*., 2000; Achen, 1998). Biological activity of VEGF-C and -D depends on proteolytic cleavage. VEGF-E was discovered in the genome of the Orf virus, and signaling through VEGFR-2 and neuropilin receptor (NRP-1) causes endothelial cell mitogenesis. PIGF-1 binds specifically to VEGFR-1 and is a nonheparin binding protein. PIGF-2 is a heparin binding protein that binds to VEGFR-1 and NRP-1. VEGF and PIGF form heterodimers have been found to bind to VEGFR-2. High concentration of PIGF saturates VEGFR-1 binding sites and augments the action of VEGF, which then acts through VEGFR-2. PIGF is chemotactic to endothelial cells and monocytes.

2.2. Angiopoietins (Angs)

Angs are a group of growth factors that affect the growth of endothelial cells. They bind to the receptor Tie2 (tyrosine kinase with immunoglobulin and epidermal growth factor homology domain) (Sato *et al*., 1995). Ang1/Tie-2 along with VEGF/VEGFR-2 is critical for mobilization and recruitment of hemopoietic stem cells and the circulation of endothelial cell precursors. Property of Ang-1 to reduce vascular leakage and inflammation might prove beneficial in vascular gene therapy. Ang-2 is an antagonist for Ang-1 and is probably needed for vascular remodeling.

2.3. Hypoxia-inducible factor-1 (HIF-1)

HIF-1 is a transcription factor that acts as a regulator of oxygen homeostasis. It acts as a transcriptional activator of VEGF gene. A cellular enzyme HIF-1α prolyl hydrolase (HIF-PH) probably serves as a cellular oxygen sensor. HIF-1α administered via gene transfer induces expression of VEGF, which leads to therapeutic neovascularization of ischemic tissues. This property of HIF-1 has been used for promoting therapeutic angiogenesis.

2.4. Platelet-derived growth factors (PDGFs)

The family of PDGFs currently compromises of four members, PDGF-A, -B, -C, and -D, which bind to receptors PDGFR-α and PDGFR-β. They are major mitogens for fibroblasts, smooth muscle cells, and several other cell types. PDGF-A and PDGF-B form homo and heterodimers with their tyrosine kinase receptors, whereas PDGF-C and PDGF-D form apparently only homodimers. Increased PDGF activity has been implicated in several pathological conditions in adults, including atherosclerosis, restenosis, fibrosis, and tumorigenesis. PDGF receptor (PDGFR) inhibition is known to reduce restenosis in experimental animals.

2.5. Fibroblast growth factors (FGFs)

FGFs are known to stimulate cell migration and cell mitosis and affect cellular senescence. FGF signaling contributes to multiple, distinct steps in vessel formation. These steps include proliferation and differentiation of Flk1-positive hemangioblastic precursor cells from mesoderm, assembly of endothelial cells during vasculogenesis, and sprouting angiogenesis. FGFs can regulate vascular morphogenesis by acting either directly through FGFRs or indirectly by inducing other angiogenic factors like VEGFs. FGF is produced by angiogenic tissue, and it can be released to stimulate endothelial cells, smooth muscle cells, and pericytes. Thus, FGFs might be responsible for the maturation of blood vessels. Adenoviral mediated FGF-4 gene delivery has been used in Phase II clinical trials for peripheral artery occlusive disease and coronary artery disease.

2.6. Hepatocyte growth factor (HGF)

Hepatocyte growth factor stimulates proliferation and migration of endothelial cells through *c-Met* (a transmembrane tyrosine kinase) receptor present on endothelial cells and some other cell types including smooth muscle cells and pericytes. Overexpression of HGF in the skin increases granulation tissue formation, angiogenesis, and VEGF levels. It has been used to promote therapeutic angiogenesis in animal models and in a clinical trial.

2.7. Ephrins (Eph)

The Eph family of receptor tyrosine kinases is the largest known family of receptor tyrosine kinases (RTKs) identified so far. Expression of ephrin ligands may be induced by growth factors and cytokines in various cell types. Ligands of EphB family induce capillary sprouting in vitro. Expression of ephrin-B2 and its cognate EphB receptors in mesenchymal cells adjacent to vascular endothelial cells suggests an EphB/ephrin-B2 interaction at endothelial-mesenchymal contact zones. Ephrin-A1 is expressed at sites of vascular development. The Eph receptor family represents a new class of receptor tyrosine kinases, and their role in angiogenesis is yet to be defined.

2.8. Risks associated with angiogenic gene therapy

Certain risks are associated with therapeutic angiogenesis and include formation of hemangiomas or vascularization of tumors, neovascularization in atherosclerotic lesions leading to plaque rupture, development of nonfunctional vessels, and edema. Increasing the tissue specificity of the gene constructs and promoters and regulating the transgene expression should minimize these risks.

3. Targeted gene delivery systems

The efficacy and safety of gene therapy also depends on targeting genes to particular cells and effectively controlling their expression. Developing vectors with defined cell-type trophism or using cell-specific promoters and regulatory elements can produce better targeting (Harris and Lemoine, 1996).

Receptor-mediated targeting is based on receptor ligand interaction. Modified vectors have been prepared that target binding to alternative attachment receptors, improving vector specificity. Adenoviruses are widely used vectors for gene transfer to dividing and nondividing cells (*see* Article 96, **Adenovirus vectors**, Volume 2). However, they have broad cell tropism and transgene expression is often detected in various ectopic organs. Novel adenoviruses targeted to vascular wall have been developed. They include matrix metalloproteinase-2 and -9 (MMP-2 and -9) targeted TIMP-1 encoding adenoviruses, $\alpha\nu$ integrin targeted human interleukin-2 encoding adenoviruses, and endothelial cell targeted adenoviruses. Additionally, blocking of (Coxsackie and adenovirus) CAR receptors may lead to targeted expression by adenoviral vectors (Kibbe *et al*., 2000).

In transcriptional targeting, tissue or cell type–specific promoters and regulatory elements are used to restrict expression in nontarget tissues. Endothelial specific promoters include *fms-like* tyrosine kinase-1 (FLT-1), intercellular adhesion molecule-2 (ICAM-2), von Willebrand factor, and Tie promoters. The SM22alfa promoter restricts transgene expression exclusively to smooth muscle cells after adenovirus-mediated gene transfer to arterial wall. Along with the use of viral promoters in cardiovascular gene transfer, vectors containing inducible promoters are now being used to regulate gene expression and to optimize therapeutic

effect. An example of inducible promoters is *Escherichia coli* tetracycline responsive element *tet*, which activates transcription of the transgene only in the presence of tetracycline. Another strategy is the use of endogenous stimuli to regulate transgene expression. Examples of this approach are vectors containing transcription regulatory elements sensitive to hypoxia, which can be effectively used for the regulation of transgene expression in ischemic tissues. The hypoxia response element (HRE) is introduced into an expression cassette and gene expression is activated by HIF-1 under ischemic conditions (Dachs *et al*., 1997).

4. Potential therapeutic targets

4.1. Atherosclerotic vascular disease and thrombosis

Atherosclerosis is characterized by deposition of atheromas or plaques in the inner layers of arteries. These plaques can ultimately occlude an artery, or an unstable plaque can result in thromboembolic episodes. Complex etiology of atherosclerosis makes the use of a single or local gene transfer for its prevention or treatment a controversial issue. But several genetic disorders with a single gene defect, which predispose to the development of atherosclerosis, can be treated with gene therapy. In cases of low-density lipoprotein (LDL) receptor deficiency, LDL receptor and very low density lipoprotein (VLDL) receptor gene transfers to liver may prove beneficial. Lecithin cholesterol acyl transferase (LCAT) or lipid transfer protein gene transfer can be used to treat certain dyslipoproteinemias. It is possible to inhibit the elevated levels of atherogenic apolipoprotein (apo) B100 by apobec-1 gene transfer, which is a catalytic subunit of apoB editing enzyme. ApoA1 gene transfer that promotes reverse cholesterol transport might be used to treat apoA1 deficient patients. ApoE gene transfer might be useful for decreasing lipoprotein levels in the treatment of Type III Hyperlipoproteinemia. Lipoprotein lipase and hepatic lipase gene transfers could benefit patients having deficiency of these enzymes. Class A soluble scavenger receptor gene transfer could decrease lipid accumulation in macrophages and class B soluble scavenger receptors can alter high-density lipoprotein (HDL) levels. Decreased nitric oxide (NO) bioavailability probably results in endothelial dysfunction, occurring in early atherosclerosis. It could be corrected by using endothelial nitric oxide synthase (eNOS) and VEGF genes. In advanced cases of atherosclerosis, however, an increased NO production may not be useful. Rho family GTPases participates in the regulation of actin cytoskeleton and cell adhesion. Inhibiting Rho kinase (RhoK) by dominant negative RhoK gene transfer decreases atherosclerosis. Overexpression of antioxidant enzymes like superoxide dysmutase (SOD) also helps in decreasing atherosclerosis. It has been seen that Interleukin-10 (IL10) and platelet activating factor acetyl hydrolase (PAF-AH) gene transfers decrease atherosclerosis probably via their antiinflammatory effects. Rupture of an unstable plaque and subsequent thrombosis in an atherosclerotic artery might precipitate an acute ischemic episode. TIMP gene transfer may prove useful to stabilize unstable plaques. Other gene transfers used to decrease thrombotic episodes in animal models include cyclooxygenase, hirudin, thrombomodulin, tissue plasminogen activator, and tissue factor pathway inhibitor.

4.2. Coronary artery disease (CAD) and peripheral artery disease (PAD)

Coronary artery occlusion due to atherosclerosis can result in myocardial ischemia. Angiogenic gene therapy is aimed at promoting new blood vessel formation in ischemic myocardium, thereby improving cardiac perfusion, exercise tolerance, and quality of life. Therapeutic angiogenesis using VEGF, FGF, HGF, and HIF-1α has proved beneficial in many animal models. Improvement in exercise tolerance was reported after adenovirus-mediated VEGF gene transfer to ischemic myocardium. Targeted delivery of angiogenic growth factors using sophisticated delivery systems like NOGA catheters might further improve chances of rescuing ischemic myocardium.

In PAD, there is decreased blood supply to the limbs because of arterial obstruction and vasoconstriction. Many of these patients suffer from disabling symptoms like severe ischemic rest pain, and amputation is often required to alleviate suffering. Therapeutic angiogenesis using angiogenic growth factors has been recently used to treat critical limb ischemia. VEGF, FGF, and HGF gene transfers have been used to promote development of collateral blood vessels in animal models and clinical trials. Angiopoietins can also possibly enhance the maturity of new vessels formed after VEGF gene therapy. Other cytokines like Monocyte chemotactic protein-1 (MCP-1) and PDGFs can also promote angiogenesis indirectly.

4.3. Arterial restenosis and vein graft disease

Maladaptive response to injury can result in occlusion of an artery as seen after balloon angioplasty, stenting, or in bypass vein graft. Restenosis is defined as a diameter stenosis of 50% at follow up. Restenosis occurs in 10–30% of patients after balloon angioplasty and stenting. Multiple factors including smooth muscle cell proliferation, matrix accumulation, remodeling, thrombosis, and platelets and leukocyte adhesion are involved in the development of arterial restenosis after angioplasty, stenting, and in vein graft disease. Various gene therapy strategies have been employed to decrease cellular proliferation. These include antisense oligonucleotides and ribozymes against *c-myb, c-myc, cdc-2, cdk-2, ras, bcl-x*, or decoy constructs against transcription factors such as E2F and NFkB. Cell cycle inhibitors like nonphosphorylated retinoblastoma gene, p21, p27, p53, and *gax* can decrease cellular proliferation and neointima formation. Similarly, (Herpes Simplex Virus-Thymidine Kinase) HSV-TK, cytosine deaminase, preprocecropine A, and *fas* ligand gene transfers have been shown to decrease cellular proliferation and smooth muscle cell migration in the blood vessels. Transfer of VEGF and HGF genes to vessel wall has been shown to decrease neointima formation, possibly by enhancing endothelial repair. It has been hypothesized that rapid regeneration of endothelial cells results in secretion of antiproliferative substances like nitric oxide, C-type natriuretic peptide, and prostacyclin I_2. Gene transfer of TIMP-1, nitric oxide synthtase, and dominant negative Rho kinase has resulted in decreased neointima formation in animal models. Inhibition of thrombosis by recombinant hirudin (inhibitor of thrombin) gene transfer resulted in decreased neointima formation

in animal models. Only a limited number of gene therapy clinical trials have been conducted for restenosis and vein graft disease. Ex vivo gene transfer of E2F decoy in vein grafts has been successful in decreasing graft failure rate in human trials. Other clinical trials for gene therapy in restenosis and vein graft disease have so far been inconclusive.

4.4. Systemic hypertension

Essential hypertension is a progressive disease characterized by chronically elevated blood pressure of unknown etiology (*see* Article 63, **Hypertension genetics: under pressure**, Volume 2). Multifactor and intricate etiology of systemic hypertension has led to the question of feasibility of gene therapy in hypertension. But it has been shown that altering certain mediators by gene therapy can result in effective lowering of systemic blood pressure. Argument in favor of gene therapy for hypertension has been that a single gene transfer might be able to control systemic hypertension for a long term, thereby improving patient compliance. One approach has been to transfer genes, which increase vasodilator proteins like tissue kallikrein, atrial natriuretic peptide (ANP), adrenomedullin, and eNOS. Another approach has been to decrease the vasoconstrictor proteins. Antisense oligonucleotides and DNA have been used against β-adrenoreceptors, angiotensin converting enzyme (ACE), angiotensin type-1 receptors, angiotensin gene activating element, thyrotropin releasing hormone (TRH) and TRH receptor, carboxypeptidase y, *c-fos,* and CYP4A1. Although promising results have been obtained in animal models for hypertension, no clinical trial for gene therapy in systemic hypertension has yet taken place.

4.5. Pulmonary hypertension

Pulmonary hypertension is characterized by progressively increasing pulmonary artery pressure. Primary pulmonary hypertension (PPH) is a disease of unknown etiology, while secondary pulmonary hypertension results from diseases like collagen vascular disease, congenital heart disease, chronic thrombotic/ and or embolic disease, chronic obstructive pulmonary disease, chronic hypoxia, and certain drugs. Mutations in BMPR-II (encoding bone morphogenetic protein receptor II) have been reported in many cases of sporadic cases of PPH. It is usually a progressive and fatal disease. Gene therapy to decrease cellular proliferation and vasospasm in pulmonary vessels has so far been limited to animal studies.

MCP-1, Prepro-calcitonin gene related peptide, atrial natriuretic peptide, eNOS, prostacyclin synthtase and VEGF gene transfers have been used with varying degree of success in animal models for pulmonary hypertension.

4.6. Congestive cardiac failure and cardiomyopathies

Alterations in myocardial β-adrenergic receptor system and intracellular calcium signaling play a crucial role in the pathophysiology of heart failure. Ability to genetically manipulate beta-adrenergic receptor system and calcium signaling might prove beneficial in the management of chronic congestive cardiac failure of primary

Table 1 Therapeutic genes and their disease targets

Treatment target	Therapeutic genes
Atherosclerosis	VLDL receptor, LDL receptor, apoE, apoA-1, lipoprotein lipase, hepatic lipase, LCAT, apoB, lipid transfer proteins, Lp(a) inhibition, soluble scavenger-receptor decoy, soluble VCAM or ICAM, SOD, IL-10, PAF-AH
Thrombosis	Hirudin, tPA, thrombomodulin, TFPI, COX
Unstable plaque	TIMPs, COX, soluble VCAM
Restenosis and vein graft failure	VEGF-A, C, eNOS, iNOS, TIMPs, TK, COX, gax, CyA, p53, Rb, sdi-1, fas ligand, p16, p21, p27, NFkB and E2F decoys, cdk-2, cdc-2, c-myb, c-myc, ras, bcl, Gbg, PCNA antisense oligonucleotides, ribozimes, cecropine A, blocking PDGF or TGF-b expression or their receptors
Systemic hypertension	Tissue kallikrein, ANP, adrenomedullin, eNOS, Adrenoreceptor, ACE, angiotensin II type-1 receptor, angiotensin gene activating element, TRH receptor, TRH, carboxypeptidase y, c-fos, CYP4A1 antisense oligonucleotides
Therapeutic angiogenesis	VEGF-A, -B, -C, -D, -E, PlGF-1, FGF-1, -2, -4, -5, Angiopoetin-1, -2, HGF, MCP-1, PDGF, eNOS, iNOS
Pulmonary hypertension	Prepro-calcitonin gene related peptide, ANP, eNOS, prostacyclin syntase, VEGF-A

VEGF: Vascular endothelial growth factor; PIGF: Placental growth factor; HGF: Hepatocyte growth factor; FGF: Fibroblast growth factor; MCP-1: Monocyte chemotactic protein-1; PDGF: Platelet-derived growth factor; NOS: Nitricoxide synthase; ANP: Atrial natriuretic peptide; TIMP: Tissue inhibitor of metalloproteinase; TK: Thymidine kinase; CyA: Cytosine deaminase; Rb: Retinoblastoma gene; sdi-1: Senescent cell-derived inhibitor-1; PCNA: Proliferating cell nuclear antigen; TGF- β: Transforming growth factor β; VCAM: Vascular-cell adhesion molecule; ICAM: Intercellular adhesion molecule; LCAT: Lecithin cholesterol acyl transferase; COX: Cyclooxygenase; NFkB: Nuclear factor kappa B; ACE: Angiotensin converting enzyme; TRH: Thyrotropin releasing hormone; SOD: Super oxide dysmutase; IL-10: Interleukin-10; PAF-AH: Platelet activating factor acetyl hydrolase; TFPI: Tissue factor pathway inhibitor; tPA: Tissue plasminogen activator.

myocardial origin, where conventional drug therapy is often inadequate. HGF gene transfer that has an antifibrosis and antiapoptosis action in the myocardium was beneficial in an animal model of cardiomyopathy with heart failure. Recent genetic studies have revealed that mutations in genes for cardiac sarcomere components lead to dilated cardiomyopathy. Mutations in the Z-line region of titin were found along with decreased binding affinities of titin to Z-line proteins. Gene therapy directed at correcting defective sarcomeric proteins may prove beneficial in cases of familial cardiomyopathy.

Various therapeutic genes and their disease targets are listed in Table 1.

5. Delivery systems for cardiovascular gene transfer

Selection of an appropriate system results in efficient expression of the therapeutic gene. Gene delivery to the cardiovascular system could be local, regional, and systemic. Local gene transfer is required in cases of atherosclerotic lesions, vein graft, ischemic conditions of myocardium or skeletal muscle, regional as in the case of pulmonary hypertension or systemic as in atherosclerosis and hypertension. Perivascular collars or sheaths, needle injection catheters, and biodegradable gels can be used for delivering the vectors to adventitia (Laitinen *et al*., 1997; Laitinen and Ylä-Herttuala, 1998). When specific physical or biological targeting methods are available, they usually improve transgene expression (Ylä-Herttuala and Alitalo,

2003). Ultrasound, microbubbles and electronic pulses can be used to enhance gene transfer efficiency. Local delivery to small arterioles and capillaries can also be achieved by using coated biodegradable microspheres.

Successful transfection of smooth muscle cells in media could be achieved by high-pressure intra luminal gene delivery. Disruption of intima and internal elastic lamina by balloon angioplasty and subsequent delivery of the vector by infusion catheter results in transfection of medial SMC. Coated stents and hydrogel coated balloon catheters are also useful for delivering vectors to endothelial cells or medial smooth muscle cells (Riessen *et al*., 1993). Different types of catheters have been developed for the delivery of vectors to blood vessels, such as double, gel coated, porous channel balloon and dispatch catheters.

Vector delivery to ischemic myocardium for CAD has been achieved using intramyocardial injections via thoracotomy and intracoronary injections (Hedman *et al*., 2003). A recent introduction to myocardial gene transfer is nonfluoroscopic catheter-based electromechanical mapping system (NOGA, Biosense Webster, Inc). NOGA system assesses the reduction in electrical voltage and mechanical activity in ischemic myocardium, thereby differentiating it from healthy tissue. Pericardial delivery of the vectors can also be useful for transferring genes to myocardium and coronary arteries. Gene delivery for the treatment of peripheral arterial disease (PAD) can be done by the use of direct intramuscular injections (Isner *et al*., 1998), infusion–perfusion catheters, and hydrogel coated balloon catheters.

6. Issues in clinical trial design

The field of therapeutic angiogenesis for CAD is fast growing from basic and preclinical investigations to clinical trials although many new issues need to be addressed. These include a deep understanding of the biology of angiogenesis, selection of appropriate patient populations for clinical trials, choice of therapeutic end points for curative or palliative purposes, choice of therapeutic strategy, route of administration, and the side effect profile (Isner *et al*., 2001).

The induction of arteriogenesis is clinically more relevant in maintaining the cardiac function and patient survival than angiogenesis as demonstrated by the fact that myocardial infarction patients are less likely to develop ventricular aneurysms and show improved survival, if they have collateral arteries. Patient selection is another important feature, both with respect to age and genetic background. There is age dependent impairment of VEGF expression at least in part caused by impaired induction of HIF-1 activity in response to ischemia or hypoxia. Combination therapies using different growth factors according to their role in the initiation of growth and maintenance of blood vessels are needed to ensure long-term viability of these vessels. An alternative to this is the use of “angiogenic master switch” genes like HIF-1α that is capable of initiating multiple neovascularization cascades.

Presently, most gene transfer studies in cardiovascular gene therapy are preclinical, although a number of high profile vascular gene therapy clinical trials are in progress. The most important targets for cardiovascular gene therapy are CAD and PAD. The clinical trials for cardiovascular gene therapy are summarized in Table 2A and 2B.

Table 2 Clinical trials in cardiovascular gene therapy

(A) Trials for therapeutic angiogenesis

Investigator	Location	Disease	Delivery route	No. of patients	Treatment	Phase	Vector
Isner *et al*.	St. Elizabeth's Medical Center, Boston, MA, USA	Peripheral artery occlusive disease	Intramuscular injection	6	VEGF-A	I	Plasmid
Baumgartner *et al*.	St. Elizabeth's Medical center, Boston, MA, USA	Peripheral artery occlusive disease	Intramuscular injection	9	VEGF-A	I	Plasmid
Berlex Schering AG	Multicenter trial, Europe	Peripheral artery occlusive disease	Intramuscular injection	[a]	FGF-4	II	Adenovirus
Symes *et al*. (1999)	St. Elizabeth's Medical Center, Boston, MA, USA	Severe coronary artery disease	Intramyocardial injection via thoracotomy	20	VEGF-A	I	Plasmid
Rosengart *et al*.	Cornell Medical Center, New York, USA	Coronary artery disease	Intramyocardial injection during bypass surgery or minithoracotomy	21	VEGF-A	I	Adenovirus
Grines *et al*. (2002) Berlex Laboratories	Multicenter trial, USA	Coronary artery disease	Intracoronary injection	131	FGF-4	II	Adenovirus
Kastrup *et al*.	Multicenter trial, Europe	Coronary artery disease	Intramyocardial injections NOGA	48[a]	VEGF-A	II	Plasmid
Hedman *et al*.	Kuopio University Central Hospital, Kuopio, Finland	Coronary artery disease, in-stent restenosis	Infusion–perfusion catheter after angioplasty and stenting	103	VEGF-A	II	Plasmid/ liposome or adenovirus
Mäkinen *et al*. (2002)	Kuopio University Central Hospital, Kuopio, Finland	II Peripheral artery occlusive disease	Infusion–perfusion catheter after angioplasty	54	VEGF-A	II	Plasmid/liposome or adenovirus

[a]Ongoing recruitment.

(B) Trials for restenosis and vein graft failure

Investigator	Location	Disease	Delivery route	No. of patients	Treatment	Phase	Vector
Laitinen *et al*. (2000)	Kuopio University Central Hospital, Kuopio, Finland	Coronanry artery disease, restenosis	Infusion–perfusion catheter after angioplasty	10	VEGF-A	I	Plasmid/liposome
Mann *et al*.	Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA	Vein graft stenosis, infrainguinal bypass surgery	Pressure mediated ex vivo delivery	41	E2F Decoy	II	Oligonucleotide
Grube *et al*.	Multicenter trial	Coronary vein graft stenosis	Pressure mediated ex vivo delivery	200	E2F Decoy	II	Oligonucleotide
Kutryk *et al*.	University Hospital Dijkzigt, Rotterdam, Netherlands	Coronary artery disease, in-stent restenosis	Catheter after stenting	85	c-myc antisense	II	Oligonucleotide
Cardion AG	Multicenter trial	Coronary artery disease, in-stent restenosis	Infiltrator catheter af ter stenting	[a]	iNOS	I	Plasmid/lipoplex

[a] Ongoing recruitment

References

Achen M, Jeltsch M, Kukk E, Makinen T, Vitali A, Wilks A, Alitalo K and Stacker SA (1998) Vascular endothelial growth factor D (VEGF-D) is a ligand for the tyrosine kinases VEGF receptor 2 (Flk1) and VEGF receptor 3 (Flt4). *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 548–553.

Bhardwaj S, Roy H, Gruchala M, Viita H, Kholova I, Kokina I, Achen MG, Stacker SA, Hedman M, Alitalo K, *et al*. (2003) Angiogenic responses of vascular endothelial growth factors in periadventitial tissue. *Human Gene Therapy*, **14**, 1451–1462.

Dachs GU, Patterson AV, Firth JD, Ratcliffe PJ, Townsend KM, Stratford IJ and Harris AL (1997) Targeting gene expression to hypoxic tumor cells. *Nature Medicine*, **3**, 515–520.

Ferrara N (2001) Role of vascular endothelial growth factor in regulation of physiological angiogenesis. *American Journal of Physiology. Cell Physiology*, **280**, C1358–C1366.

Ferrara N and Davis-Smyth T (1997) The biology of vascular endothelial growth factor. *Endocrine Reviews*, **18**, 4–25.

Grines CL, Watkins MW, Helmer G, Penny W, Brinker J, Marmur JD, West A, Rade JJ, Marrott P, Hammond HK, *et al*. (2002) Angiogenic Gene Therapy (AGENT) trial in patients with stable angina pectoris. *Circulation*, **105**, 1291–1297.

Hamada K, Oike Y, Takakura N, Ito Y, Jussila L, Dumont DJ, Alitalo K and Suda T (2000) VEGF-C signaling pathways through VEGFR-2 and VEGFR-3 in vasculoangiogenesis and hematopoiesis. *Blood*, **96**, 3793–3800.

Harris JD and Lemoine NR (1996) Strategies for targeted gene therapy. *Trends in Genetics*, **12**, 400–405.

Hedman M, Hartikainen J, Syvanne M, Stjernvall J, Hedman A, Kivela A, Vanninen E, Mussalo H, Kauppila E, Simula S, *et al*. (2003) Safety and feasibility of catheter-based local intracoronary vascular endothelial growth factor gene transfer in the prevention of postangioplasty and in-stent restenosis and in the treatment of chronic myocardial ischemia: phase II results of the Kuopio Angiogenesis Trial (KAT). *Circulation*, **107**, 2677–2683.

Isner JM, Baumgartner I, Rauh G, Schainfeld R, Blair R, Manor O, Razvi S and Symes JF (1998) Treatment of thromboangiitis obliterans (Buerger's disease) by intramuscular gene transfer of vascular endothelial growth factor: preliminary clinical results. *Journal of Vascular Surgery*, **28**, 964–973.

Isner JM, Vale PR, Symes JF and Losordo DW (2001) Assessment of risks associated with cardiovascular gene therapy in human subjects. *Circulation Research*, **89**, 389–400.

Kibbe MR, Murdock A, Wickham T, Lizonova A, Kovesdi I, Nie S, Shears L, Billiar TR and Tzeng E (2000) Optimizing cardiovascular gene therapy: increased vascular gene transfer with modified adenoviral vectors. *Archives of Surgery*, **135**, 191–197.

Laitinen M, Hartikainen J, Hiltunen MO, Eranen J, Kiviniemi M, Narvanen O, Makinen K, Manninen H, Syvanne M, Martin JF, *et al*. (2000) Catheter-mediated VEGF gene transfer to human coronary arteries after angioplasty. *Human Gene Therapy*, **110**, 263–270.

Laitinen M, Pakkanen T, Donetti E, Baetta R, Luoma J, Lehtolainen P, Viita H, Agrawal R, Miyanohara A, Friedmann T, *et al*. (1997) Gene transfer into the carotid artery using an adventitial collar: comparison of the effectiveness of the plasmid-liposome complexes, retroviruses, pseudotyped retroviruses, and adenoviruses. *Human Gene Therapy*, **8**, 1645–1650.

Laitinen M and Ylä-Herttuala S (1998) Adventitial gene transfer to arterial wall. *Pharmacological Research*, **37**, 251–254.

Mäkinen K, Manninen H, Hedman M, Matsi P, Mussalo H, Alhava E and Ylä-Herttuala S (2002) Increased vascularity detected by digital subtraction angiography after VEGF gene transfer to human lower limb artery: a randomized, placebo-controlled, double-blinded phase II study. *Molecular Therapy*, **6**, 127–133.

Olofsson B, Korpelainen E, Pepper MS, Mandriota SJ, Aase K, Kumar V, Gunji Y, Jeltsch MM, Shibuya M, Alitalo K, *et al*. (1998) Vascular endothelial growth factor B (VEGF-B) binds to VEGF receptor-1 and regulates plasminogen activator activity in endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 11709–11714.

Riessen R, Rahimizadeh H, Blessing E, Takeshita S, Barry JJ and Isner JM (1993) Arterial gene transfer using pure DNA applied directly to a hydrogel-coated angioplasty balloon. *Human Gene Therapy*, **4**, 749–758.

Rissanen TT, Markkanen JE, Gruchala M, Heikura T, Puranen A, Kettunen MI, Kholova I, Kauppinen RA, Achen MG, Stacker SA, *et al*. (2003) VEGF-D is the strongest angiogenic and lymphangiogenic effector among VEGFs delivered into skeletal muscle via adenoviruses. *Circulation Research*, **92**, 1098–1106.

Sato TN, Tozawa Y, Deutsch U, Wolburg-Buchholz K, Fujiwara Y, Gendron-Maguire M, Gridley T, Wolburg H, Risau W and Qin Y (1995) Distinct roles of the receptor tyrosine kinases Tie-1 and Tie-2 in blood vessel formation. *Nature*, **376**, 70–74.

Symes JF, Losordo DW, Vale PR, Lathi KG, Esakof DD, Mayskiy M and Isner JM (1999) Gene therapy with vascular endothelial growth factor for inoperable coronary artery disease. *The Annals of Thoracic Surgery*, **68**, 30–836.

Ylä-Herttuala S and Alitalo K (2003) Gene transfer as a tool to induce therapeutic vascular growth. *Nature Medicine*, **9**, 694–701.

Ylä-Herttuala S and Martin JF (2000) Cardiovascular gene therapy. *Lancet*, **355**, 213–222.

Short Specialist Review

Artificial self-assembling systems for gene therapy

Pierre Lehn
Hôpital Robert Debré, Paris, France

1. Introduction

Nonviral gene delivery systems are nowadays widely investigated as an alternative approach to recombinant viruses for gene therapy studies. Indeed, although nonviral vectors still suffer from a limited efficiency, they are free from a number of inconveniences associated with the use of viruses. One may consider nonviral systems to include all physical and chemical methods for gene transfer. Although combining physical techniques (such as electroporation, gene gun, hydrodynamic pressure, and ultrasound) has recently improved gene transfection using naked DNA, we will herein focus on chemical gene transfer systems that can be viewed as virus-like systems, as all steps involved in virus-mediated gene transduction are of chemical nature. Thus, we will describe the main characteristics of the current chemical vectors and outline the cellular and molecular barriers still limiting their transfection activity. In this forward-looking chapter, we will also discuss strategies to improve in vivo gene delivery, in particular, the development of sophisticated modular systems constituting true "artificial viruses".

2. Nonviral vectors versus recombinant viruses

The use of an efficient and safe gene delivery system is an obvious prerequisite for successful gene therapy. Viral vectors have been shown to be particularly efficient for gene transfer. Indeed, viruses are smart nucleic acid–containing supramolecular assemblies that have been tailored by evolution for transferring their genes from one cell to another. Accordingly, several viral vector systems have been developed (retroviruses, adenoviruses . . .) over the last decade. All of these viral systems rely on the generation of replication-defective recombinant viral particles by genetically engineered "producer" cells. It is therefore no surprise that viral vectors suffer from drawbacks resulting from their biological nature, in particular, safety concerns, immunogenicity issues, and practical issues relating to large-scale production and quality control. These inherent limitations of viral vectors have led to the development of an alternative de novo approach in which synthetic organic

molecules are used as DNA carriers (Lehn *et al*., 1998). Such artificial vectors (termed nonviral vectors) are indeed free of the infection risks of recombinant viruses as they are well-characterized compounds obtained in the test tube via chemical synthesis. Other advantages of nonviral vectors include their probable low immunogenicity, ease of large-scale production, and cost-effectiveness. Further, in contrast to viral vectors in which capsids have a predetermined size, there is no (upper, lower) size limit for the DNA to be transferred by chemical vectors, as the vector/DNA complexes are formed via a self-assembling process in the test tube. Thus, synthetic vectors allow the transfer of not only eukaryotic expression cassettes for cDNAs but also large genomic constructs as well as short nucleic acid sequences (Aissaoui *et al*., 2002).

3. Current vectors: DNA condensing agents

The current nonviral vectors belong to one of two main categories: cationic liposomes/micelles or cationic polymers (Kabanov *et al*., 1998; Huang *et al*., 1999). Spontaneous formation of self-assembled nanometric vector/DNA complexes is in both cases due to electrostatic interactions between the positively charged chemical vector and the negatively charged DNA. Lipoplex is the name given to the complexes formed between cationic lipids and DNA, whereas the complexes formed by cationic polymers are termed polyplexes. The DNA entrapped in these complexes is then protected from degradation by nucleases. It is generally agreed that use of an excess of cationic vector yields DNA complexes with a net positive charge whose binding to negative cell surface residues (such as proteoglycans) leads to nonspecific, electrostatic-driven endocytosis. Once inside the cell, the DNA has, however, still to perform several steps before being transcribed in the nucleus: escape from the endosome into the cytoplasm (to avoid degradation in the lysosome), trafficking to the perinuclear region, and finally passage across the nuclear membrane (Zabner *et al*., 1995).

Cationic lipids are especially attractive as they can be prepared with relative ease and extensively characterized. All cationic lipids are positively charged amphiphiles containing three functional domains: (1) a polar hydrophilic head group, which is positively charged, generally via protonation of one (monovalent lipid) or several (multivalent reagents) amino groups; (2) a linker whose nature and length impact on the stability and biodegradability (and consequently on the toxicity) of the vector; (3) a hydrophobic moiety composed of either two alkyl chains (saturated or unsaturated) or cholesterol. The first use of a monovalent cationic lipid for in vitro gene transfection into cultured cells was reported in 1987 (Felgner *et al*., 1987). A multivalent lipopolyamine soon followed, whereas use of cholesterol as the hydrophobic anchor was subsequently validated (Miller, 1998). Since this initial “proof of principle” period, many other cationic lipids have been synthesized (often on a basis of trial and error) in order to develop better vectors (Miller, 1998; Martin *et al*., 2003). Modifications have been made in the design of each of the fundamental constituent parts of a cationic lipid. First, the choice of head group has expended into the use of natural architectures and functional groups with recognized DNA binding modes. For example, the transfection efficiency of

cationic cholesterol derivatives characterized by a head group with guanidinium functions or natural aminoglycoside structures has been reported by us (Vigneron *et al*., 1996; Belmont *et al*., 2002). Second, modifications of the hydrophobic portion revealed that optimal vector design was also highly dependent on this moiety. Here, because of its rigidity, cholesterol has been used when lipoplexes with a high degree of stability were required as for aerosol delivery. Finally, stable linking of the hydrophilic and hydrophobic portions is commonly achieved using a variety of chemical bonds (carbamate, amide, ester, ether). Of note, cationic lipids are often formulated as liposomes with the neutral colipid DOPE (dioleoyl phosphatidylethanolamine), as DOPE is thought to have fusogenic properties that may enhance endosomal escape of the DNA to the cytoplasm (Miller, 1998; Martin *et al*., 2003).

Several cationic polymers have been reported to form complexes with DNA and promote gene transfection. The linear polymer poly-L-Lysine (pLL) was the first cationic polymer to be used for gene delivery (Wagner, 1998). Its efficiency was, however, found to be low in the absence of additional agents facilitating cellular uptake or endosomal release. Cellular uptake could be improved by conjugation to the carrier of a variety of ligands (such as transferrin, antibodies, RGD tripeptide motifs ...) specific for receptors on the target cells. Endosomal escape was enhanced via the use of lysosomotropic agents (chloroquine), defective virus particles, and fusogenic peptides. Next, polyamidoamine (PAMAM) dendrimers are spherical macromolecules bearing a large number of amine groups on their surface which were also tested for DNA delivery. In contrast to intact dendrimers, degraded or fractured PAMAM dendrimers mediated very efficient gene transfection in vitro (Tang *et al*., 1996). Polyethylenimine (PEI) is a recent but impressive addition to the list (Boussif *et al*., 1995). PEI has a very high charge density potential, as every third atom is an amino nitrogen atom. However, at physiological pH, only a fraction of these amino groups are protonated. Thus, it has been proposed that the high gene transfer efficiency of PEI was due to its strong buffering capacity. Indeed, the capacity of PEI to capture protons at the acidic pH of the endosome may cause osmotic swelling and subsequent endosome disruption with release of the DNA ("proton sponge" mechanism) (Kichler *et al*., 1999). Finally, it is noteworthy that efficient gene transfection has also been mediated by several other cationic polymers and block copolymers and, in the particular case of the muscle, even by nonionic polymers and copolymers (Kabanov *et al*., 1998; Kabanov *et al*., 2002).

4. Clinical gene therapy trials

Encouraging data from a variety of animal studies have provided a reasonable basis for subsequent clinical trials in man (Kabanov *et al*., 1998; Huang *et al*., 1999). For example, local intratumoral or regional (intracavitary) gene therapy with various types of lipoplexes showed a significant antitumor effect. Other studies demonstrated that nonviral vectors mediated efficient transfection of the airway epithelial cells. Accordingly, in the first clinical gene therapy trials with cationic lipids, the lipoplexes were applied via in situ administration such as instillation

into the airways (of cystic fibrosis patients) or direct intratumoral injection (Hersh and Stopeck, 1998; Davies *et al*., 2001; see also the interactive database of The Journal of Gene Medicine clinical trials website at **http://www.wiley.co.uk/genmed/clinical**). These clinical trials basically showed encouraging safety and biological activity data, but they also clearly emphasized that more efficient systems were required.

5. Future directions: toward multifunctional systems

It is thus nowadays agreed that more efficacious nonviral vectors need to be developed before gene therapy can be regarded as a viable therapy. As such improved systems need to be capable of overcoming the multiple barriers encountered in vivo, most research efforts focus at present on providing the vector/DNA complexes with the various functions required for surmounting these barriers: (1) stabilization of the lipoplexes/polyplexes in the extracellular medium (via "stealth technology" involving pegylation); (2) equipment of the DNA complexes with receptor-specific ligands (sugar residues, folate ...) for targeted transfection; (3) enhancement of endosomal escape and "triggered" decomplexation of the DNA via incorporation of functional groups sensitive to cellular stimuli, such as the decrease in pH along the endosomal/lysosomal pathway or the cytoplasmic redox potential; (4) facilitated trafficking to the perinuclear region (probably along the microtubule network); and (5) inclusion of ligands (such as nuclear localization signals, steroids ...) for nuclear uptake (Miller, 1998; Yonemitsu *et al*., 1998; Martin *et al*., 2003; Kichler, 2004; Wagner, 2004). The future will require the difficult task of incorporating all the functions described above into a single system. In addition, the different functional components need to be capable of working in a chronological order to avoid unwanted interreactivity of the individual functions (Lehn *et al*., 1998). Finally, present research also focuses on the development of improved gene constructs. Indeed, nonviral vectors are nowadays generally used to transfer plasmids (containing eukaryotic expression cassettes) that lead to transient expression of the transgene. Here, research efforts aim, in particular, at decreasing the immune response to the unmethylated CpG motifs in the plasmid DNA and designing self-replicating or integrating plasmid expression systems to increase the duration of transgene expression (Aissaoui *et al*., 2002).

6. Conclusions

Over the last decade, a variety of nonviral gene delivery systems have been developed and shown to mediate efficient gene transfection in vitro. Their in vivo efficiency was, however, found to be much less satisfactory. The goal of current research is thus to develop improved vectors for efficient and safe in vivo gene delivery. The design of such improved vectors can nowadays be based on a better understanding of the various barriers encountered by the DNA complexes while trafficking from the extracellular medium to the nucleus of the target cells. Accordingly, improved vectors may consist of sophisticated modular systems where

each functional component enables the DNA complex to overcome a critical cellular barrier. Such true "artificial viruses" may then constitute a serious alternative to recombinant viruses for gene therapy applications.

References

Aissaoui A, Oudrhiri N, Petit L, Hauchecorne M, Kan E, Sainlos M, Julia S, Navarro J, Vigneron JP, Lehn JM, *et al*. (2002) Progress in gene delivery by cationic lipids: guanidinium-cholesterol-based systems as an example. *Current Drug Targets*, **3**, 1–16.

Belmont P, Aissaoui A, Hauchecorne M, Oudrhiri N, Petit L, Vigneron JP, Lehn JM and Lehn P (2002) Aminoglycoside-derived cationic lipids as efficient vectors for gene transfection in vitro and in vivo. *Journal of Gene Medicine*, **4**, 517–526.

Boussif O, Lezoualch F, Zanta MA, Mergny MD, Scherman D, Demeinex B and Behr JP (1995) A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo – Polyethylenimine. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 7297–7301.

Davies JC, Geddes DM and Alton EWFW (2001) Gene therapy for cystic fibrosis. *Journal of Gene Medicine*, **3**, 409–417.

Felgner PL, Gadek TR, Holm M, Roman R, Chan HW, Wenz M, Northrop JP, Ringlod GM and Danielsen M (1987) Lipofection: a highly efficient, lipid-mediated DNA transfection procedure. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 7413–7417.

Hersh EM and Stopeck AT (1998) Cancer gene therapy using nonviral vectors: preclinical and clinical observations. In *Self-assembling Complexes for Gene Delivery*, Kabanov AV, Felgner PL and Seymour LW (Eds.), John Wiley & Sons: Chichester, pp. 421–436.

Huang L, Hung MC and Wagner E (1999) *Nonviral Vectors for Gene Therapy*, Academic Press: San Diego.

Kabanov AV, Felgner PL and Seymour LW (1998) *Self-assembling Complexes for Gene Delivery*, John Wiley & Sons: Chichester.

Kabanov AV, Lemieux P, Vinogradov S and Alakhov V (2002) Pluronic block copolymers: novel functional molecules for gene therapy. *Advanced Drug Delivery Reviews*, **54**, 223–233.

Kichler A (2004) Gene transfer with modified polyethylenimines. *Journal of Gene Medicine*, **6**, S3–S10.

Kichler A, Behr JP and Erbacher P (1999) Polyethylenimines: a family of potent polymers for nucleic acid delivery. In *Nonviral Vectors for Gene Therapy*, Huang L, Hung MC and Wagner E (Eds.), Academic Press: San Diego, pp. 191–206.

Lehn P, Fabrega S, Oudrhiri N and Navarro J (1998) Gene delivery systems: dridging the gap between recombinant viruses and artificial vectors. *Advanced Drug Delivery Reviews*, **30**, 5–11.

Martin B, Aissaoui A, Sainlos M, Oudrhiri N, Hauchecorne M, Vigneron JP, Lehn JM and Lehn P (2003) Advances in cationic lipid-mediated gene delivery. *Gene Therapy and Molecular Biology*, **7**, 273–289.

Miller AD (1998) Cationic liposomes for gene therapy. *Angewandte Chemie International Edition*, **37**, 1769–1785.

Tang MX, Redemann CT and Szoka FC (1996) In vitro gene delivery by degraded polyamidoamine dendrimers. *Bioconjugate Chemistry*, **7**, 703–714.

Vigneron JP, Oudrhiri N, Fauquet M, Vergely L, Bradley JC, Basseville M, Lehn P and Lehn JM (1996) Guanidinium-cholesterol cationic lipids: efficient vectors for the transfection of eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 9682–9686.

Wagner E (1998) Polylysine-conjugate based DNA delivery. In *Self-assembling Complexes for Gene Delivery*, Kabanov AV, Felgner PL, and Seymour LW (Eds.), John Wiley & Sons: Chichester, pp. 309–322.

Wagner E (2004) Strategies to improve DNA polyplexes for in vivo gene transfer: will "artificial viruses" be the answer? *Pharmaceutical Research*, **21**, 8–14.

Yonemitsu Y, Alton EWFW, Komori K, Yoshimizu T, Sugimachi K and Kaneda Y (1998) HVJ (Sendai virus) liposome-mediated gene transfer: current status and future perspectives. *International Journal of Oncology*, **12**, 1277–1285.

Zabner J, Fasbender AJ, Moninger T, Poellinger KA and Welsh MJ (1995) Cellular and molecular barriers to gene transfer by a cationic lipid. *Journal of Biological Chemistry*, **270**, 18997–19007.

Short Specialist Review

Adenovirus vectors

Monika Lusky
Transgene SA, Strasbourg, France

1. Adenovirus biology

For a better understanding of Ad vectors, the life cycle of adenovirus, extensively reviewed in Shenk (1996), is summarized below. The adenovirus is a nonenveloped, icosahedral virus of 60–90 nm in diameter. The genome of the most commonly used human adenovirus (group C, serotypes 2, 5) consists of a linear 36 kb double-stranded DNA molecule. The major viral capsid proteins consist of 240 hexon, 12 penton capsomeres, and 12 homotrimeric units of the fiber protein as the spike components of the penton capsomeres. The viral protein IX (pIX) forms 80 homotrimeric units and acts as cement protein for the viral capsid. A dominant pathway of cell entry is dictated by the interaction of the fiber protein's globular knob domain with the cellular coxsackie and adenovirus receptor (CAR). During the internalization process, a sequential disassembly of the virion occurs. The arrival of the viral DNA in the nucleus triggers the onset of early viral transcription.

Transcription of the viral genome occurs on both strands, and viral gene expression is coordinated through a precisely temporirally regulated splicing program of almost all the transcripts. Early transcription units (E1, E2, E3, E4) are differentiated from late ones (L), depending on the expression pattern relative to the onset of viral DNA synthesis (Figure 1). The first viral gene to be expressed is the major viral transcriptional transactivator encoded by the *E1A* gene, activating viral early transcription through the interaction with multiple transcription factors. As the E2 gene products (E2A: DNA binding protein, DBP; E2B: preterminal protein and DNA polymerase) accumulate, viral DNA replication can commence. The inverted terminal repeats (ITRs) of the viral chromosome serve as the replication origins. DNA synthesis occurs through protein priming mediated by the preterminal protein (pTP).

The adenoviral late genes, encoding mostly viral capsid proteins, begin to be expressed efficiently at the onset of viral DNA replication. The adenovirus late coding regions are organized into a single large transcription unit whose primary transcript is approximately 29 000 nt in length (Figure 1). This major late transcription unit (MLTU), controlled by the major late promoter (MLP), is processed by differential polyA site utilization and splicing to generate at least 18 distinct mRNAs, grouped into five distinct families, L1 to L5. With the accumulation of excess quantities of the viral capsid proteins, virus assembly begins with the formation of

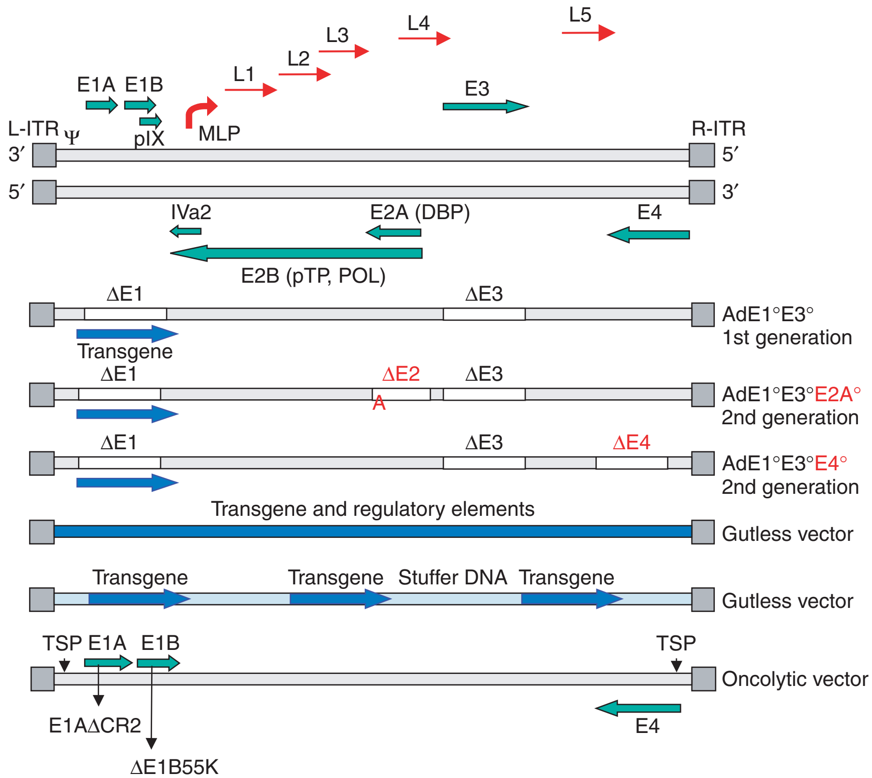


Figure 1 Schematic representation of the Ad5 genome organization and different species of recombinant Ad vectors. The direction of transcription of early (E) and late (L) mRNAs and of transgenes in the recombinant vectors is indicated by arrows. Deletions of viral regions and insertion of transgenes are depicted for first- and second-generation vectors as well as for gutless vectors. Typical genome modifications and the insertion of tumor-specific promoters (TSP) are indicated for an oncolytic vector

an empty capsid and, subsequently, a viral DNA molecule enters the capsid. The DNA-capsid recognition event is mediated by the packaging sequence (ψ), a cis-acting DNA sequence at about 260 nt from the left end of the viral chromosome. A single infectious cycle can lead to the production of 10^4 to 10^5 progeny particles.

2. Replication-defective Ad vectors

2.1. E1-deleted Ad vectors

Replication-deficient Ad vectors with critical viral functions deleted serve two purposes: (1) inhibition of viral spreading into the environment renders such viruses safe; (2) by deletion of viral genes, space is provided, which allows for the insertion of foreign transgenes. The earliest, first-generation recombinant Ad vectors have the E1 region deleted (E1). In addition, in most AdE1° vectors, the viral E3 region

is also deleted, as the E3 functions are not required for the viral life cycle in vitro (Shenk, 1996). In most cases, a heterologous expression cassette with a transgene is inserted in place of the E1 region (Figure 1). Such AdE1° and AdE1°E3° vectors can be propagated to high yields in permissive E1 complementation cell lines, such as 293 cells that provide the E1 functions in trans (Graham *et al*., 1977). 293 cells were generated by transformation of human embryonic kidney cells with sheared adenovirus DNA and carry the sequences between 1 and 4137 nt integrated into chromosomal DNA (Louis *et al*., 1997). Most AdE1° vectors carry an E1 deletion between approximately 400 and 3500 nt of the genome. Thus, an extensive sequence overlap exists between the E1 sequences present in the cell line and the sequences present in the vector. Owing to double crossover recombination events, the E1 sequences will be incorporated into the vector, leading to a replication competent adenovirus vector (RCA; Lochmüller *et al*., 1994). The strategy to prevent RCA was to eliminate any sequence overlap between vector sequences and viral sequences in the cell. This concept resulted in the generation of new E1 complementation cells carrying essentially only E1 coding sequences; these cells are either based on human embryonic retina cells (Per.C6; Fallaux *et al*., 1998) or human amniotic cells (Schiedner *et al*., 2000).

2.2. Ad vectors with multiple deletions

Disadvantages of AdE1° vectors including a high level of tissue toxicity and inflammation interfering with persistent transgene expression in many preclinical and clinical studies have stimulated further manipulation of the viral genome. Second-generation vectors with simultaneous deletions of several regulatory regions, AdE1°E3°E2A° or AdE1°E3°E4°, and the respective complementation cell lines (Lusky *et al*., 1998 and references therein) were generated (Figure 1). The additional deletions improve the cloning capacity to approximately 11 kb. Importantly, AdE1°E3°E4-modified vectors carrying the E4 ORF3 or E4ORF3 + ORF4 functions were able to allow persistent transgene expression in vivo, in selected animal models, in the absence of vector-induced toxicity and inflammation (Armentano *et al*., 1999; Christ *et al*., 2000). Multiple deleted Ad vectors might also be useful in cancer therapy (Senzer *et al*., 2004).

2.3. Gutless Ad vectors

Gutless, high-capacity Ad vectors lack all viral genes (gutless) and contain only the cis-acting sequences required for viral replication (ITRs) and the packaging signal (Schiedner *et al*., 2002). These vectors can accommodate up to 36 kb of nonviral DNA (high capacity), allowing the insertion of multiple expression cassettes, large genes, and their regulatory sequences (Figure 1). The toxicity and immunogenicity of these vectors in vivo is significantly reduced because of lack of viral gene expression, making long-term transgene expression possible. The production of a gutless Ad vector requires the presence of a helper virus, providing all missing viral functions in trans. In order to enrich for the gutless vector during production

and to reduce the contamination of helper virus in the purified gutless vector product, most systems are based on the genetic inactivation of the helper virus through recombinase-mediated excision of the packaging signal of the helper virus (Schiedner *et al*., 2002). This results in preparations of gutless vectors with a contamination of helper vector below 1%.

3. Replicative Ad vectors

Tumor-selective, replication-competent oncolytic viruses are attractive candidates as cancer therapeutics because they are intended to replicate and spread exclusively in tumor cells, leading to their destruction, while not affecting normal cells (Dobbelstein, 2004). Thus, the antitumor effect is not delivered with a transgene but by virus-mediated oncolysis and subsequent spreading of the virus in the tumor tissue. Two major approaches are pursued to achieve tumor selectivity of replication (Figure 1). One is to eliminate viral genes that are dispensable in tumor cells and the second is to replace viral promoters with tumor-specific promoters to express viral genes required for replication. The first approach exploits the fact that many tumors lack functional tumor suppressor genes such as p53 or the retinoblastoma tumor suppressor gene pRB (Johnson *et al*., 2002). Therefore, viral genes interacting with these host cell proteins should not be required for the replication of the virus in tumor cells. This concept has been explored with Ad vectors deficient in selected E1 genes. The prototype of such vector dl1520 (ONYX015) is deficient in p53 interaction due to a deletion in the E1B 55 kDa gene. This was the first therapeutic oncolytic virus shown to preferentially propagate in cancer cells based upon their p53 functional status (Bischoff *et al*., 1996).

Similarly, oncolytic Ad vectors with a mutation (CR2) in the pRB binding domain of the viral *E1A* gene have been designed toward selective replication in tumor cells with a disrupted retinoblastoma tumor suppressor protein (pRB) pathway. Since the pRB pathway is disrupted in nearly all human tumors, the oncolytic activity of a virus with the E1A-CR2 mutation should not be limited to tumors of a particular tissue type (Johnson *et al*., 2002). For such type of vectors, it was shown that its pRB selectivity could further be enhanced by replacement of the viral E1 and E4 promoter with the human E2F promoter (Johnson *et al*., 2002). The second approach for developing tumor cell selectivity has been the use of tumor-selective or tumor-specific promoters to control the expression of the adenoviral early genes E1, E2, or E4 and to restrict the oncolytic activity to the specific tumor targeted (Fernandez and Lemoine, 2004).

Both types of oncolytic vectors have been tested extensively in clinical trials, which have demonstrated a high level of safety of these vectors and very low toxicity levels. However, therapeutic efficacy of these oncolytic monotherapies has been limited so far (Kirn, 2002). Therefore, strategies are developed to increase the efficiency of oncolytic virus therapy. Genetic strategies include the incorporation of a therapeutic transgene into the oncolytic virus to generate an "armed therapeutic virus" (Hermiston and Kuhn, 2002). Several oncolytic vectors have been generated expressing prodrug-activating enzymes, and promising clinical results have emerged recently with an oncolytic Ad vector carrying a gene fusion

of the cytosine deaminase (CD) and the thymidine kinase (TK) gene (Hermiston and Kuhn, 2002).

4. Modification of Ad tropism

Owing to the broad tropism of adenovirus, significant efforts have been undertaken to de-target the virus from its natural tropism (CAR) and to retarget the virus to new receptors, with the goal of restricting transduction to the organ or tumor tissue of interest and to limit inflammatory responses. Another rationale for retargeting Ad vectors is that the CAR receptor, while expressed in most healthy tissues throughout the human body, is present in low levels in some target cell types and lack or downregulation of CAR has been reported for various tumor types (Anders *et al*., 2003). Owing to a detailed structure –function understanding of the fiber protein, residues in the fiber knob domain involved in CAR interaction as well as the HI-loop and the C-terminus in the knob for the suitable incorporation of ligands could be identified (Legrand *et al*., 2002). Viruses deficient in CAR interaction were de-targeted for the CAR entry pathway in vitro, but not in vivo, suggesting the existence of additional entry pathways in vivo. In this context, the ubiquitously expressed heparin sulfate proteoglycan molecules were described as an alternative Ad5 receptor (Dechecchi *et al*., 2001). Recent studies have shown that specific mutations in the fibershaft almost completely abolish the natural tropism of the virus in mice and nonhuman primates; however, the role of the fibershaft for the different entry pathways is not known (Smith *et al*., 2003). Several model ligands such as the integrin binding motif RGD or a heparansulfate proteoglycan binding polylysine peptide inserted into the fiber protein have been described to prove the concept of retargeting in vitro and have showed enhancement of tumor transduction in vivo (Wickham, 2002; Curiel, 2002). Fiber pseudotyping presents an alternative approach toward vector retargeting (Havenga *et al*., 2002). More recently, another capsid protein, pIX, was proposed as a candidate for the insertion of ligands. Incorporation of RGD or polylysine motifs at the extruding C-terminus of this protein showed enhanced transduction rates in vitro (Dimitriev *et al*., 2002; Velinga *et al*., 2004). Together, although it is now possible to generate fiber-modified viruses that show a specifically redirected binding and infection in vitro, the impact of these modifications is less clear in vivo.

An alternative or additional targeting strategy exploits the possibility of differential and tissue-/tumor-selective transcription through the use of tissue- or tumor-specific promoters (TSP). Thus transcriptional targeting aims to genetically limit the expression of the introduced gene or the replication of the virus to specific tissues or to the tumor mass through the use of promoter sequences whose activity is upregulated in the target tissues (Fernandez and Lemoine, 2004).

5. Toxicity and immunogenicity

Ad vector particles cause an acute immediate toxicity following systemic administration, resulting in a strong and rapid activation of chemokine expression by Ad-transduced macrophages. These initial events are followed by the cellular immune

response against vector-encoded MHC-presented proteins (St George, 2003). In addition, neutralizing antibodies raised against Ad vector particles in the host will prevent systemic readministration of the vector (O'Riordan, 2002). Approaches to overcome this problem include the sequential pseudotyping of Ad vectors (Havenga *et al*., 2004).

6. Ad vector construction

Initially, recombinant Ad vectors were generated in eukaryotic cells, such as in 293 cells, using ligation methods and homologous recombination (reviewed in Lusky *et al*., 2002).

Recently, several novel methods based on bacterial systems have been developed for the generation of Ad. Three basic methods have evolved to enable the manipulation of the full-length adenoviral genome as a stable plasmid and facilitate the efficient construction of precisely tailored and infectious Ad in *Escherichia coli*. Two powerful methods in use are based on homologous recombination (Chartier *et al*., 1996) and direct ligation (Mizuguchi and Kay, 1998) technologies. These methods combine major advantages over traditional approaches: (1) Manipulation of the viral genome at any point is possible. (2) Recombinant viral DNA is purified from individual bacterial clones and therefore generates homogenous virus preparations, obviating the need for tedious plaque screening and purification. (3) Importantly, and in contrast to the traditional in vivo approaches, these methods entirely separate viral vector construction from virus production. The first step is performed in bacteria and the second step takes place in the mammalian complementation cell line. The methods in use are simple, highly efficient, and can generate recombinant Ad vectors in a very short period of time (Lusky *et al*., 2002).

7. Conclusion and future direction

Cancer gene therapy is aiming at the destruction of malignant cells, whereas gene therapy for monogenetic disorders aims to restore a long-term function in target cells. Therefore, the requirements for viruses to be used for these applications are fundamentally different. In most clinical trials today, first- and second-generation Ad vectors are used in cancer gene therapy or in vaccination for tumor therapy or infectious diseases. For these applications, short-term and robust gene expression is often sufficient to show therapeutic effects. In contrast, gutless Ad vectors will find their applications in genetic diseases requiring long-term gene expression and potential readministration.

While much knowledge has been gained in the years of Ad vector development, many challenges remain. These include the development of strategies to (1) reduce vector-induced toxicity, (2) to enable readministration, (3) to specifically direct the vector to the targeted tissue in vivo specifically after systemic administration, and (4) to increase the therapeutic index of oncolytic vectors in vivo.

Further reading

Armentano D, Zabner J, Sacks C, Sookdeo CC, Smith MP, St. George JA, Wadsworth SC, Smith AE and Gregory RJ (1997) Effect of the E4 region on the persistence of transgene expression from adenovirus vectors. *Journal of Virology*, **71**, 2408–2416.

Fernandez M, Lemoine N (2002) Tumor/tissue selective promoters. In *Vector targeting for therapeutic gene delivery*, Curiel D, Douglas, J (Eds.), John Wiley & Sons: Chichester, pp. 459–480

References

Anders M, Christian C, McMahon M, McCormick F and Korn WM (2003) Inhibition of the Raf/MEK/ERK pathway up-regulates expression of the coxsackievirus and adenovirus receptor in cancer cells. *Cancer Research*, **63**, 2088–2095.

Bischoff JR, Kirn DH, Williams A, Heise C, Horn S, Muna M, Ng L, Nye JA, Sampspon-Johannes A, Fattaey A, *et al*. (1996) An adenovirus that replicates selectively in p53-deficient human tumor cells. *Science*, **274**, 373–376.

Chartier C, Degryse E, Gantzer M, Dieterle A, Pavirani A and Mehtali M (1996) Efficient generation of recombinant adenovirus vectors by homologous recombination in *Escherichia coli*. *Journal of Virology*, **70**, 4805–4810.

Christ M, Louis B, Stoeckel F, Dieterle A, Grave L, Dreyer D, Kintz J, AliHadji D, Lusky M and Mehtali M (2000) Modulation of the inflammatory properties and hepatotoxicity of recombinant adenovirus vectors by the viral E4 gene products. *Human Gene Therapy*, **1**, 415–427.

Curiel DT (2002) Strategies to alter the tropism of adenoviral vectors via genetic capsid modifications. In *Vector Targeting for Therapeutic Gene Delivery*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 171–201.

Dechecci MC, Melotti P, Bonizatto A, Santacaterina M, Chilosi M and Cabrini G (2001) Heparin sulfate glycosaminoglycans are receptors sufficient to mediate the initial binding of adenovirus types 2 and 5. *Journal of Virology*, **75**, 8772–8780.

Dimitriev IP, Kashentseva EA and Curiel DT (2002) Engineering of adenovirus vectors containing heterologous peptide sequences in the C-terminus of capsid protein IX. *Journal of Virology*, **76**, 6893–6899.

Dobbelstein M (2004). Replicating adenoviruses in cancer therapy. In *Current Topics in Microbiology and Immunology*, Vol. 273, Doerfler W, Böhm (Ed.), Springer Verlag, Berlin, pp. 291–334.

Fallaux FJ, Bout A, van der Velde I, van den Wollenberg DJ, Hehir KM, Keegan J, Auger C, Cramer SJ, van Ormondt H, van der Eb AJ, *et al*. (1998) New helper cells and matched early region 1-deleted adenovirus vectors prevent generation of replication-competent adenoviruses. *Human Gene Therapy*, **9**, 1909–1917.

Graham FL, Smiley J, Russell WC and Nairn R (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *The Journal of General Virology*, **36**, 59–74.

Havenga MJE, Vogels R, Bout A Mehtali M (2002) Pseudotyping of adenoviral vectors. In *Vector targeting for therapeutic gene delivery*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 89–122.

Hermiston TW and Kuhn I (2002) Armed therapeutic viruses: Strategies and challenges to arming oncolytic viruses with therapeutic genes. *Cancer Gene Therapy*, **9**, 1022–1035.

Johnson L, Shen A, Boyle L, Kunich J, Pandey K, Lemmon M, Hermiston T, Giedlin M, McCormick F and Fattaey A (2002) Selectively replicating adenoviruses targeting deregulated E2F activity are potent, systemic antitumor agents. *Cancer Cell*, **1**, 325–337.

Kirn D (2002) Replication-selective oncolytic adenovirus E1-region mutants: virotherapy for cancer. In *Adenoviral Vectors for Gene Therapy*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 329–374.

Legrand V, Leissner P, Winter A, Mehtali M and Lusky M (2002) Transductional targeting with recombinant adenovirus vectors. *Current Gene Therapy*, **3**, 323–339.

Lochmüller H, Jani A, Huard J, Prescott S, Simoneau M, Massie B, Karpati G and Acsadi G (1994) Emergence of early region 1-containing replication competent adenovirus in stocks of replication-defective adenovirus recombinants (ÐE1 + ÐE3) during multiple passages in 293 cells. *Human Gene Therapy*, **5**, 1485–1492.

Louis N, Evelegh C and Graham FL (1997) Cloning and sequencing of the cellular viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology*, **233**, 423–429.

Lusky M, Christ M, Rittner K, Dieterlé A, Dreyer D, Mourot B, Schultz H, Stoeckel F, Pavirani A and Mehtali M (1998) *In vitro* and *in vivo* biology of recombinant adenovirus vectors with E1, E1/E2A, or E1/E4 deleted. *Journal of Virology*, **72**, 2022–2032.

Lusky M, Degryse E, Mehtali M Chartier C (2002). Adenoviral vector construction II: bacterial systems. In *Adenoviral Vectors for Gene Therapy*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 105–128.

Mizuguchi H and Kay MA (1998) Efficient construction of a recombinant adenovirus vector by an improved in vitro ligation method. *Human Gene Therapy*, **9**, 2577–2583.

O'Riordan C (2002) Humoral immune response. In *Adenoviral Vectors for Gene Therapy*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 375–408.

Schiedner G, Clemens PR, Volpers C, Kochanek S (2002) High-capacity gutless adenoviral vectors: technical aspects and applications. In *Adenoviral Vectors for Gene Therapy*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 429–446.

Schiedner G, Hertel S and Kochanek S (2000) Efficient transformation of primary human amniocytes by E1 functions of Ad5: generation of new cell lines for adenoviral vector production. *Human Gene Therapy*, **11**, 2105–2116.

Senzer N, Mani S, Rosemurgy A, Nemunaitis J, Cunningham C, Guh Bayol N, Gillen M, Chu K, Rasmussen C, Rasmussen H, *et al*. (2004) TNFerade biologic, an adeno vector with radiation-inducible promoter, Carrying the human tumor necrosis factor alpha gene in phase I study in patients with solid tumors. *Journal of Clinical Oncology*, **22**, 577–579.

Shenk T (1996) Adenoviridae: the viruses and their replication. In *Fundamental Virology*, Fields BN, Knipe DM, Howley PM (Eds.), Raven Press, Philadelphia, PA, pp. 976–1016.

Smith TA, Idamakanti N, Marshall-Neff J, Rollence ML, Wright P, Kaloss M, King L, Mech C, Dinges L, Iverson WO, *et al*. (2003) Receptor interactions involved in adenoviral-mediated gene delivery after systemic administration in non-human primates. *Human Gene Therapy*, **14**, 1595–1604.

St. George JA (2003) Gene therapy progress and prospects: adenoviral vectors. *Gene Therapy*, **10**, 1135–1141.

Velinga J, Rabelink MJWE, Cramer SJ, van den Wollenberg DJM, Van der Meulen H, Leppard KN, Fallaux FJ and Hoeben RC (2004) Spacers increase the accessibility of peptide ligands linked to the CARboxyl terminus of adenovirus minor capsid protein IX. *Journal of Virology*, **78**, 3470–3479.

Wickham TJ (2002) Genetic targeting of adenoviral vectors. In *Vector Targeting for Therapeutic Gene Delivery*, Curiel D, Douglas J (Eds.), John Wiley & Sons: Chichester, pp. 143–170.

Short Specialist Review

Adeno-associated viral vectors: depend(o)ble stability

Richard O. Snyder
University of Florida, Gainesville, FL, USA

The adeno-associated viruses (AAV) are members of the family *Parvoviridae*, genus *Dependovirus,* and as the genus name implies, AAVs are dependent upon coinfection with a helper virus, such as an adenovirus, for efficient replication. The latent phase of the life cycle of these nonpathogenic viruses has made them attractive for development as gene transfer vectors (Muzyczka, 1992). As AAV vector technology has improved, the safety and efficacy of AAV-mediated gene transfer in animal models has been demonstrated (reviewed in Snyder, 1999), and the vectors have been shown to persist and are safe in humans (Wagner *et al*., 1999; Flotte *et al*., 2003; Stedman *et al*., 2000; Manno *et al*., 2003; Janson *et al*., 2002; During *et al*., 2001; Flotte *et al*., 2004).

To date, more than 100 serotypes of AAV have been identified (Gao *et al*., 2004). Of the vectors that have been characterized, some are capable of infecting different types of cells from several species, and the cellular receptors that determine tropism are being investigated. For AAV2, three cellular receptors have been identified: heparin sulfate proteoglycan, the fibroblast growth factor receptor (FGFR1), and $\alpha V\beta 5$ integrin. Serotypes other than AAV2 interact with different cell surface molecules (Di Pasquale *et al*., 2003). The difference in transduction efficiency of these vector serotypes in different tissue targets can be quite significant, and matching vector serotype to organ targets is an area of intense investigation.

AAV virions contain a single-stranded DNA genome of 4.7 kb and both plus and minus strands are packaged equally. Infectious clones have facilitated the study of the genetics of the virus, defining two major open reading frames (ORFs). The *cap* ORF encodes the viral capsid proteins VP-1, VP-2, and VP-3 that assemble into particles with $T = 1$ icosahedral symmetry (Xie *et al*., 2002). The *rep* ORF encodes the four nonstructural Rep proteins that have been shown to possess functions required to replicate the genome, modulate transcription from AAV and heterologous promoters, mediate site-specific integration into the human genome, and encapsidate single-stranded genomes. The palindromic terminal repeats are 145 nucleotides that form T-shaped structures and encode sequences required for packaging, integration, and rescue, and also serve as the origins of DNA replication.

In tissue culture cells, there are preferred sites within the human genome where wild-type AAV2 establishes a latent infection that can be stable for years and many passages. The highly preferred site located on chromosome 19q13.3-qter (called

AAVS1) was initially characterized by Kotin *et al*. (1992), and subsequent analyses have shown that as little as 33 bp of AAVS1 is required to direct integration. DNA binding experiments demonstrated the ability of the Rep 78 and 68 proteins to bind to AAVS1 and mediate the juxtaposition with the AAV2 ITR. A search of the May 2004 freeze of the human genome using the UC Santa Cruz genome browser (**http://genome.ucsc.edu**) did not reveal the presence of any significant regions of AAV DNA homology (M. Nickerson and R. Snyder, unpublished observation). Gao *et al*. (2004) were able to isolate AAV genomes from a variety of human tissues. This raises two possibilities: (1) that AAV genomic sequences may not be integrated in the human genome but episomal forms were coisolated during the total DNA isolation procedure used by Gao *et al*. or (2) the compilation of the consensus human genome sequence may exclude differences between sequenced DNA samples such that unique sequences, including integrated AAV, may have been "filtered out".

The long-term and stable gene expression in normal, mature, and quiescent tissue has been achieved in a variety of animal models and organ targets using rAAV vectors encoding therapeutic and marker proteins (*see* Article 92, **Hematopoietic stem cell gene therapy**, Volume 2, Article 93, **Gene therapy in the central nervous system**, Volume 2, Article 94, **Cardiovascular gene therapy**, Volume 2, Article 101, **Gene transfer to skeletal muscle**, Volume 2, Article 102, **Gene transfer to the liver**, Volume 2, and Article 103, **Gene transfer to the skin**, Volume 2). These studies demonstrate the versatility of rAAV vectors and their applicability to treating a broad spectrum of human diseases. Following rAAV transduction, vector dose-dependent and sustained protein expression has been achieved utilizing constitutive promoters from cellular and viral sources, and exogenously regulated promoters based on the tetracycline (Rendahl *et al*., 1998) or rapamycin (Rivera *et al*., 2004) systems (*see* Article 104, **Control of transgene expression in mammalian cells**, Volume 2). Acute toxicity has not been observed by histological and serum biochemical analyses following transduction of a variety of tissues. Safety studies to support clinical trials require sensitive, reproducible, and validated assays carried out under Good Laboratory Practices (GLP) regulations.

Long-term studies in animals and humans indicate that few serious adverse events (such as tumorigenesis and germline transmission) arise following rAAV-mediated gene delivery (Manno *et al*., 2003; Wagner *et al*., 1999; Tenenbaum *et al*., 2003). In a study carried out by Donsante *et al*. (2001), data after 1 year in a small number of MPS-VII mice showed that liver tumors developed following IV neonatal transduction. This result seems to be an exception because similar long-term liver transduction of normal animals using AAV vectors have been carried out at higher input doses, and these have not resulted in liver tumors (Snyder *et al*., 1997; Mount *et al*., 2002). Integration of rAAV viral vectors (lacking the *rep* gene) into the genome of human tissue culture cells appears random. In animal models, sites homologous to AAVS1 are not found and studies to determine the status of the rAAV vector following in vivo transduction indicate that the vector genome may integrate inefficiently at random sites in head-to-tail concatameric arrays with a preference for transcribed genes (Nakai *et al*., 2003). Conversely, AAV has been detected in episomal form in lung tissue, skeletal muscle, and liver, and an episomal

intermediate may exist before integration. Targeted integration of wild-type AAV has been demonstrated in vivo in transgenic rats and mice harboring AAVS1.

Humoral immunity to multiple AAV serotypes preexists in 80–90% of humans; of this population, approximately 20–50% carry neutralizing antibodies to the capsid. Administration of AAV vectors to animals elicits a humoral response to the AAV capsid proteins, except when the primary transduction event is accompanied by immunosuppressive therapy. The repertoire of vectors derived from different AAV serotypes could allow repeat dosing or primary dosing in patients with preexisting antibodies (*see* Article 99, **Immunity and tolerance induction in gene therapy**, Volume 2).

Sustained expression of certain foreign proteins in immunocompetent animals has been achieved using rAAV vectors. Transient infiltrates following transduction of murine skeletal muscle have been observed between 2 and 6 weeks, but transduced cells were not eliminated. It was shown that rAAV vectors transduce dendritic cells very poorly, and thus induction of an immune response to transgene products may not be sufficiently primed (Wang *et al*., 2004), but this can be route and dose dependent. Additionally, the slow rise in gene expression may help in the development of tolerance toward a foreign gene product. Steady state protein expression has a characteristic 2–4-week delay following transduction in vivo that can be explained in part by the need to create a transcriptionally active template that is concatamerized and/or integrated into the cellular genome. The conversion of the viral single-stranded genome to a double-stranded molecule that is a template for RNA polymerase is a rate-limiting step for AAV transduction.

During the lytic cycle, greater than 10^6 genomes, 10^7 preformed capsids, 10^6 total virions, and 10^4 infectious virions are generated per cell. The adenoviral (Ad) gene products that play a role in the AAV lytic cycle include E1A, E1B, E2A, E4, and VA, and these act throughout the AAV replicative cycle to promote AAV production. The herpesviruses can also supply helper functions but the *HSV* genes responsible for helping AAV have distinct functions from the Ad genes and include a subset of genes required for HSV DNA replication.

For vectors, the AAV inverted terminal repeats (ITRs) supply the cis-acting sequences for production and transduction (Samulski *et al*., 1989), although sequences located inside of the ITRs possess some of these activities (Nony *et al*., 2003). A transgene can replace the AAV coding sequences because both the *rep* and *cap* gene products can be supplied in trans to make infectious rAAV virions (Samulski *et al*., 1989). The primary disadvantage in the use of rAAV vectors is the size limitation of approximately 5000 nucleotides for a recombinant genome that is capable of being packaged.

In standard methods used to produce rAAV vectors, helper and vector plasmids are cotransfected into tissue culture cells along with the Ad genes required for AAV vector production, and no infectious adenovirus is generated by this method. Approximately 100–200 infectious particles (5×10^3–5×10^4 packaged vector genomes (vg)) are produced per cell and suitable quantities of vector can be produced for early human clinical trials using this technology. For increased scale, advances such as the creation of stable packaging cell lines or helperviruses harboring the AAV genes have been developed. AAV vector sequences have been incorporated into the E1A region of Ad, and in the presence of *rep* and *cap* (e.g.,

supplied in a stable cell line), rAAV vectors can be produced at titers similar to traditional means (~100–200 infectious units per cell). Incorporating the AAV genes into Ad has been very difficult and is likely due to Rep toxicity toward adenovirus that may be of similar origin to the toxicity seen in tissue culture cells. Herpesviruses harboring the AAV genes have been constructed for AAV production and can achieve outputs of 5000 infectious units per cell (nearly equal to wild-type yields). Production of rAAV using cultured insect cells and three recombinant baculoviruses encoding AAV2 Rep proteins, capsid proteins, and a recombinant AAV vector cassette has been described recently with yields approaching 5×10^4 vg per Sf9 cell.

The size and physiochemical stability of parvoviral virions allows harvest and purification by traditional methods utilized in the industrial manufacture of protein therapeutics. The virions are very small (18–26 nm) and stable to a wide range of temperatures (−80°C to 56°C), pH 3–9, sonication, microfluidization, solvents ($CHCl_3$), detergents, proteases, and nucleases. The virions are also stable during precipitation using $(NH_4)_2SO_4$, PEG, or $CaCl_2$, and during column chromatographic procedures. Purifying rAAV vectors using automated column chromatography is economical, scalable, and reproducible, and can be validated in compliance with current Good Manufacturing Practice (cGMP) regulations. In addition, the stability of the virions is suitable for long-term storage, and direct in vivo administration.

Characterization of the vector stocks utilizing robust, rugged, and reproducible testing methods needs to be performed to develop and monitor the clinical manufacturing process, evaluate product, and compare animal studies using different vector lots (Table 1). The level of protein contaminants and the proper ratio of the AAV capsid proteins can be assessed on stained denaturing protein gels, and by immunoblotting using antibodies to possible contaminants (such as adenovirus proteins and bovine serum albumin) and AAV capsid proteins. Total and infectious particle titers need to be accurately determined to evaluate dosing and the total particle:infectious particle ratio. Ideally, the titering incorporates the use of an accepted reference standard, so doses that are prepared by different laboratories can be compared. In addition to vector characterization, product safety testing ensures patient protection. Levels of infectious adventitious viruses can be evaluated using plaque assays or cytopathic effect (CPE) assays. Contaminating cellular and plasmid DNA can be assessed by PCR or hybridization. The degree of replication competent

Table 1 AAV clinical batch testing

Safety	Characterization
Sterility including bacteriostasis/fungistasis	Purity: Silver staining/Coomassie blue
Mycoplasma	Potency: Infectious unit titer
Endotoxin	Strength: vector genome titer
In vitro adventitious viral contaminants	Strength: particle titer
General safety	Identity
rcAAV	Appearance
Residual host cell DNA	

AAV (rcAAV) generated through homologous or nonhomologous mechanisms can be determined using replication center assays, PCR, or Southern blot analyses.

Several parallel efforts are advancing the use of rAAV vectors for treating human diseases. For several serotypes, large-scale manufacturing technology is being developed that will generate vector batches that are safe, pure, potent, and stable. The collection of AAV serotypes is being matched to specific organs and cell types to maximize uptake efficiency, and these vectors will be tested in the clinic soon. Improvements in vector expression cassettes have achieved tissue specificity and regulation, and increased vector potency that provides therapeutic benefit at lower vector doses and less cost. Well-designed acute and long-term toxicology studies are being conducted to better understand the safety profile in animals, and experience in humans is accumulating. Finally, with the sequencing of the human genome, matching therapeutic genes to specific diseases has the potential for different molecular approaches for treating each disease.

Notice

RS is an inventor on patents related to recombinant AAV technology and owns equity in a gene therapy company that is commercializing AAV for gene therapy applications. To the extent that the work in this manuscript increases the value of these commercial holdings, RS has a conflict of interest.

References

Di Pasquale G, Davidson BL, Stein CS, Martins I, Scudiero D, Monks A and Chiorini JA (2003) Identification of PDGFR as a receptor for AAV-5 transduction. *Nature Medicine*, **9**, 1306–1312.

Donsante A, Vogler C, Muzyczka N, Crawford JM, Barker J, Flotte T, Campbell-Thompson M, Daly T and Sands MS (2001) Observed incidence of tumorigenesis in long-term rodent studies of rAAV vectors. *Gene Therapy*, **8**, 1343–1346.

During MJ, Kaplitt MG, Stern MB and Eidelberg D (2001) Subthalamic GAD gene transfer in Parkinson disease patients who are candidates for deep brain stimulation. *Human Gene Therapy*, **12**, 1589–1591.

Flotte TR, Brantly ML, Spencer LT, Byrne BJ, Spencer CT, Baker DJ and Humphries M (2004) Phase I trial of intramuscular injection of a recombinant adeno-associated virus alpha 1-antitrypsin (rAAV2-CB-hAAT) gene vector to AAT-deficient adults. *Human Gene Therapy*, **15**, 93–128.

Flotte TR, Zeitlin PL, Reynolds TC, Heald AE, Pedersen P, Beck S, Conrad CK, Brass-Ernst L, Humphries M, Sullivan K, *et al*. (2003) Phase I trial of intranasal and endobronchial administration of a recombinant adeno-associated virus serotype 2 (rAAV2)-CFTR vector in adult cystic fibrosis patients: a two-part clinical study. *Human Gene Therapy*, **14**, 1079–1088.

Gao G, Vandenberghe LH, Alvira MR, Lu Y, Calcedo R, Zhou X and Wilson JM (2004) Clades of Adeno-associated viruses are widely disseminated in human tissues. *Journal of Virology*, **78**, 6381–6388.

Janson C, McPhee S, Bilaniuk L, Haselgrove J, Testaiuti M, Freese A, Wang DJ, Shera D, Hurh P, Rupin J, *et al*. (2002) Clinical protocol. Gene therapy of Canavan disease: AAV-2 vector for neurosurgical delivery of aspartoacylase gene (ASPA) to the human brain. *Human Gene Therapy*, **13**, 1391–1412.

Kotin RM, Linden RM and Berns KI (1992) Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *The EMBO Journal*, **11**, 5071–5078.

Manno CS, Chew AJ, Hutchison S, Larson PJ, Herzog RW, Arruda VR, Tai SJ, Ragni MV, Thompson A, Ozelo M, *et al*. (2003) AAV-mediated factor IX gene transfer to skeletal muscle in patients with severe hemophilia B. *Blood*, **101**, 2963–2972.

Mount JD, Herzog RW, Tillson DM, Goodman SA, Robinson N, McCleland ML, Bellinger D, Nichols TC, Arruda VR, Lothrop CD Jr, *et al*. (2002) Sustained phenotypic correction of hemophilia B dogs with a factor IX null mutation by liver-directed gene therapy. *Blood*, **99**, 2670–2676.

Muzyczka N (1992) Use of adeno-associated virus as a general transduction vector for mammalian cells. *Current Topics in Microbiology and Immunology*, **158**, 97–129.

Nakai H, Montini E, Fuess S, Storm TA, Grompe M and Kay MA (2003) AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nature Genetics*, **34**, 297–302.

Nony P, Chadeuf G, Tessier J, Moullier P and Salvetti A (2003) Evidence for packaging of rep-cap sequences into adeno-associated virus (AAV) type 2 capsids in the absence of inverted terminal repeats: a model for generation of rep-positive AAV particles. *Journal of Virology*, **77**, 776–781.

Rendahl KG, Leff SE, Otten GR, Spratt SK, Bohl D, Roey MV, Donahue BA, Cohen LK, Mandel RJ, Danos O, *et al*. (1998) Regulation of gene expression in vivo following transduction by two separate rAAV vectors [In Process Citation]. *Nature Biotechnology*, **16**, 757–761.

Rivera VM, Gao GP, Grant RL, Schnell MA, Zoltick PJ, Rozamus LW, Clackson T and Wilson JM (2004) Long-term pharmacologically regulated expression of erythropoietin in primates following AAV-mediated gene transfer. *Blood*, **105**(4), 1424–1430.

Samulski RJ, Chang LS and Shenk T (1989) Helper-free stocks of recombinant adeno-associated viruses: Normal integration does not require viral gene expression. *Journal of Virology*, **63**, 3822–3828.

Snyder RO (1999) Adeno-associated virus-mediated gene delivery. *The Journal of Gene Medicine*, **1**, 166–175.

Snyder RO, Miao CH, Patijn GA, Spratt SK, Danos O, Nagy D, Gown AM, Winther B, Meuse L, Cohen LK, *et al*. (1997) Persistent and therapeutic concentrations of human factor IX in mice after hepatic gene transfer of recombinant AAV vectors. *Nature Genetics*, **16**, 270–276.

Stedman H, Wilson JM, Finke R, Kleckner AL and Mendell J (2000) Phase I clinical trial utilizing gene therapy for limb girdle muscular dystrophy: Alpha-, beta-, gamma-, or delta-sarcoglycan gene delivered with intramuscular instillations of adeno-associated vectors. *Human Gene Therapy*, **11**, 777–790.

Tenenbaum L, Lehtonen E and Monahan PE (2003) Evaluation of risks related to the use of adeno-associated virus-based vectors. *Current Gene Therapy*, **3**, 545–565.

Wagner JA, Messner AH, Moran ML, Daifuku R, Kouyama K, Desch JK, Manley S, Norbash AM, Conrad CK, Friborg S, *et al*. (1999) Safety and biological efficacy of an adeno-associated virus vector-cystic fibrosis transmembrane regulator (AAV-CFTR) in the cystic fibrosis maxillary sinus. *The Laryngoscope*, **109**, 266–274.

Wang CH, Liu DW, Tsao YP, Xiao X and Chen SL (2004) Can genes transduced by adeno-associated virus vectors elicit or evade an immune response? *Archives of Virology*, **149**, 1–15.

Xie Q, Bu W, Bhatia S, Hare J, Somasundaram T, Azzi A and Chapman MS (2002) The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10405–10410.

Short Specialist Review

Retro/lentiviral vectors

Douglas E. Brown and Andrew M. L. Lever
University of Cambridge, Cambridge, UK

Until the discovery of retroviruses and the elucidation of their life cycle, it was believed that DNA was the primary genetic store for most organisms and that this was transcribed into RNA, but that the reverse could not occur. Some viruses used RNA only, copying their RNA genome into mRNAs and new genomes. The identification of a family of viruses in which the virus particle contained an RNA genome, which acted as a template for a DNA copy that integrates into the DNA of the infected cell, altered this dogma. Identification of the enzyme responsible, reverse transcriptase, gave the viruses their name and earned the Nobel prize for Temin and Baltimore. Very soon it was realized that if one could subvert this system it could be used to deliver and integrate heterologous genes into the DNA of a cell for research and even therapeutic purposes.

Research on the known animal (mainly mouse) and avian retroviruses would have undoubtedly continued at a vigorous pace, but the emergence of a pathogenic human retrovirus HIV, which turned out to be a member of the relatively obscure lentivirus group of retroviruses, revolutionized the field. Unprecedented amounts of money and research effort were invested in understanding this virus, and it is now probably the most intensively studied genetic sequence ever. Lentiviruses are retroviruses but with a more complex array of regulatory genes and with the critical difference of being able to enter into and integrate into cells that are not undergoing mitosis. Thus, for gene delivery to cells of the nervous system (*see* Article 93, **Gene therapy in the central nervous system**, Volume 2), muscle (*see* Article 101, **Gene transfer to skeletal muscle**, Volume 2), and liver (*see* Article 102, **Gene transfer to the liver**, Volume 2), the lentivirus subfamily has opened up a new vista of possibilities over and above those available using earlier "simple" murine and avian viruses (Lever, 2003; Quinonez and Sutton, 2002; Trono, 2001).

A simplified retroviral life cycle is shown in Figure 1. To generate new retroviral vectors, it is necessary to understand the process from the time point at which the viral RNA in the cell is first transcribed through to the budding of the new viruses (5–10). The period (1–5) from binding of the virus to its target cell through to the transcriptionally active integrated DNA provirus must be understood to use the vectors to infect or "transduce" cells efficiently (Delenda, 2004).

To generate vectors, it was necessary to find out how the virus captures its own RNA from amongst the many species present in the living cell. Regions of the genomic RNA were identified, which, when deleted, did not affect protein

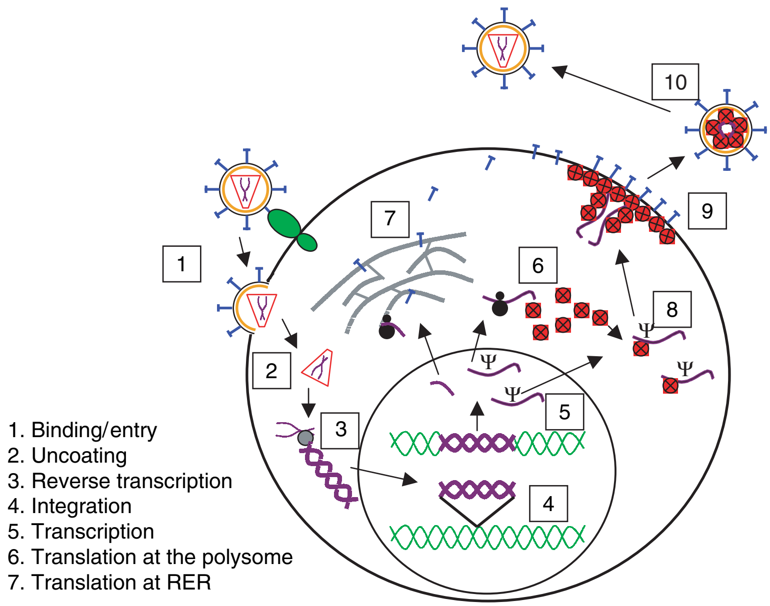


Figure 1 Simplified retroviral life cycle

coding but led to the production of virus particles devoid of genomes. Thus, "packaging signals" (Ψ) were discovered. Eliminating these from a construct encoding the viral proteins generated "empty" virus particles (Jewell and Mansky, 2000). Incorporating this same Ψ region into a heterologous RNA and expressing it in the same cell as the virus proteins led to the new chimeric RNA being captured, packaged, and incorporated into the budding virus (Lever, 2000; Linial and Miller, 1990). If the viral/vector particle infects a cell, this RNA is reverse-transcribed and integrated into that cell using the enzymes carried in the virus.

Retroviruses have another unique property amongst viruses, that of encapsidating a pair of RNA genomes – they are diploid (Haddrick *et al*., 1996). This is thought to be advantageous in repairing the reverse transcript if one copy is defective, and it also contributes to genomic diversity since two slightly different RNAs from different proviruses can, in theory, be packaged in the same particle and, by template hopping during reverse transcription, recombination can occur.

For vectors, this poses a potential danger since even if a packaging signal is deleted, the leakiness of the system means that there may still be a risk of encapsidating a small number of the mRNA molecules that code for the viral proteins along with the gene delivery vector containing the packaging signal. Recombination could recreate a full-size infectious virus. This is very undesirable for murine virus systems since the high risk of insertional mutagenesis when a replication competent vector is produced is well established (Kohn *et al*., 2003). Recombination to recreate an intact HIV would also clearly be disastrous. Eliminating this risk has been the focus of much effort, and considerable progress has also been made. First, the virus protein-coding construct is separated out into cassettes encoding either the structural (Gag) and enzymatic (Pol) proteins or the

envelope (Env) gene-coding region. In lentiviruses, the regulatory protein Rev is sometimes also required and is provided on a third construct (Gasmi *et al*., 1999). These two or three DNAs are then introduced into cells together with the gene vector with its intact packaging signal. Transient transfection or stable expression of the constructs have both been attempted for coexpression of these genes. For simple viruses, stable expression of some or all of these constructs in packaging or producer cell lines has been readily achieved (Cosset *et al*., 1995). Lentiviruses, probably because of cellular incompatibilities, have proven to be difficult to express stably although newer inducible cell lines have been developed (Farson *et al*., 2001; Kuate *et al*., 2002; Sparacio *et al*., 2001). In lentiviruses, building on many years of receptor research, the viral envelope is commonly substituted by a heterologous envelope, commonly from the Rhabdovirus, Vesicular Stomatitis Virus (VSV). The G-protein for VSV coats (pseudotypes) the particles well, makes them more stable, more easily purified and concentrated, and confers considerably wider tropism (Burns *et al*., 1993).

For all retroviral vectors, efforts have been made to eliminate the risk of recombinational reconstruction of a native virus. This has involved careful removal of all nonessential sequence from all the constructs to minimize regions of homology (Zufferey *et al*., 1997; Dull *et al*., 1998). Codon optimization reduces sequence homologies and has been shown to enhance expression (Fuller and Anson, 2001). In lentiviruses, the Rev/Rev responsive element (RRE) nuclear export system has been supplemented by export sequences from other viruses including the Mason–Pfizer Monkey virus constitutive transport element (CTE) (Srinivasakumar and Schuening, 1999) and the Woodchuck hepatitis posttranscriptional response element (WPTRE) (Zufferey *et al*., 1999). All of these enhance export of the transgene (*see* Article 104, **Control of transgene expression in mammalian cells**, Volume 2) in the target cell, but in the producer cell, despite equivalent nuclear export of RNA, the Rev protein, in concert with its response element appears to have extra effects on enhancing vector titer, suggesting it has additional roles in virus assembly and RNA encapsidation (Anson and Fuller, 2003). Vectors now commonly contain both the Rev/RRE system and a second nuclear export signal.

Compared to some viral vectors like those based on Adenovirus (*see* Article 96, **Adenovirus vectors**, Volume 2), retro/lentiviral vectors are of lower titer. 10^7 infectious units per milliliter is readily achievable and using concentration procedures, this can be raised further. Of the cells that are transduced by vectors, the proportion in which the transgene is transcriptionally active is uncertain. Simple retroviruses have LTR regions, which may be subject to transcriptional silencing, as are some heterologous promoters when used in vivo long term (Pannell and Ellis, 2001). Lentivirus vectors, if transcriptionally active on transfer, appear to maintain gene expression. More work is needed on this aspect of their use.

Murine retroviral vectors are one of the most widely used vectors in the clinic (*see* Article 100, **Gene transfer vectors as medicinal products: risks and benefits**, Volume 2). The clear definition of their packaging signals and the strictness of RNA encapsidation (Linial and Miller, 1990; Linial *et al*., 1978) together with the ease of production of packaging cell lines (Miller, 1990) has been influential in their popularity. Perhaps the most stunning success has been the transfer of the γ-C receptor gene to blood stem cells of children with severe combined

immunodeficiency (Cavazzana-Calvo *et al*., 2000). Although two children have suffered a vector insertion–related lymphoproliferative disorder (Hacein-Bey-Abina *et al*., 2003a,b), both have now been successfully treated and are in remission, bringing the number of successfully treated and clinically well children given this gene therapy to over 15. Some of these would undoubtedly be ill or dead by now and the long-term outlook for them all in the absence of this treatment was very poor. Despite the logistic barriers to producing clinically useful quantities of vector and the use of transient transfection processes in manufacture, the first lentiviral vector based on HIV-1 is now in clinical trial (Dropulic, 2001). It can only be a matter of time before this pioneering study is joined by others using the specific advantages of lentiviruses.

Retroviral vectors have a niche for long-term integrated gene expression. Their requirement for cell division can be used to advantage to transduce toxic genes into tumor cells in the background of nondividing normal cells such as in treatment of tumors in the central nervous system. They are useful for targeted treatment and ex vivo delivery to selected cell populations. Lentiviruses also have a growing potential. They seem to maintain gene expression, possibly linked to their propensity to integrate into active chromatin, while long-term gene delivery to and expression in mitotically inactive cells is a real advantage (Naldini *et al*., 1996). Their close rival, Adeno associated viruses (AAV) (*see* Article 97, **Adeno-associated viral vectors: depend(o)ble stability**, Volume 2) have the drawback of a much smaller gene-carrying capacity. Lentiviral vectors do not, however, penetrate very well into tissue and require a relatively long exposure to the target tissue for effective transduction. This may be solved by some of the newer polymer delivery systems (Dishart *et al*., 2003). The use of a heterologous envelope obviates the risk of recombinational formation of HIV; however, the later risk of gene mobilization by an exogenous HIV infection needs to be borne in mind. Insertional mutagenesis remains a specter but as long as they are being used for severe illnesses for which there is no alternative better treatment and for which the prognosis is very poor, this is a risk worth running. New strategies strengthening polyadenylation signals, using insulator sequences or locus control regions and tissue-specific promoters will reduce the risk of read-through activation of adjacent genes (*see* Article 104, **Control of transgene expression in mammalian cells**, Volume 2).

In the long term, all of these vectors will provide the building blocks for synthetic vectors, which incorporate useful individual proteins and cis-acting sequences, to create truly nonviral vectors that are safe and efficient. As yet, however, the efficiency with which viruses enter cells and deliver their genetic material dwarfs anything we have managed with synthetic vectors – but they have had many more years of practice.

References

Anson DS and Fuller M (2003) Rational development of a HIV-1 gene therapy vector. *The Journal of Gene Medicine*, **5**, 829–838.

Burns JC, Friedmann T, Driever W, Burrascano M and Yee JK (1993) Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: Concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 8033–8037.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, *et al*. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.

Cosset FL, Takeuchi Y, Battini JL, Weiss RA and Collins MK (1995) High-titer packaging cells producing recombinant retroviruses resistant to human serum. *Journal of Virology*, **69**, 7430–7436.

Delenda C (2004) Lentiviral vectors: Optimization of packaging, transduction and gene expression. *The Journal of Gene Medicine*, **6**, S125–S138.

Dishart KL, Denby L, George SJ, Nicklin SA, Yendluri S, Tuerk MJ, Kelley MP, Donahue BA, Newby AC, Harding T, *et al*. (2003) Third-generation lentivirus vectors efficiently transduce and phenotypically modify vascular cells: Implications for gene therapy. *Journal of Molecular and Cellular Cardiology*, **35**, 739–748.

Dropulic B (2001) Lentivirus in the clinic. *Molecular Therapy*, **4**, 511–512.

Dull T, Zufferey R, Kelly M, Mandel RJ, Nguyen M, Trono D and Naldini L (1998) A third generation lentivirus vector with a conditional packaging system. *Journal of Virology*, **72**, 8463–8471.

Farson D, Witt R, McGuinness R, Dull T, Kelly M, Song J, Radeke R, Bukovsky A, Consiglio A and Naldini L (2001) A new-generation stable inducible packaging cell line for lentiviral vectors. *Human Gene Therapy*, **12**, 981–997.

Fuller M and Anson DS (2001) Helper plasmids for production of HIV-1-derived vectors. *Human Gene Therapy*, **12**, 2081–2093.

Gasmi M, Glynn J, Jin MJ, Jolly DJ, Yee JK and Chen ST (1999) Requirements for efficient production and transduction of human immunodeficiency virus type 1-based vectors. *Journal of Virology*, **73**, 1828–1834.

Hacein-Bey-Abina S, von Kalle C, Schmidt M, Le Deist F, Wulffraat N, McIntyre E, Radford I, Villeval JL, Fraser CC, Cavazzana-Calvo M, *et al*. (2003a) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *The New England Journal of Medicine*, **348**, 255–256.

Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E, *et al*. (2003b) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.

Haddrick M, Lear AL, Cann AJ and Heaphy S (1996) Evidence that a kissing loop structure facilitates genomic RNA dimerisation in HIV-1. *Journal of Molecular Biology*, **259**, 58–68.

Jewell NA and Mansky LM (2000) In the beginning: Genome recognition, RNA encapsidation and the initiation of complex retrovirus assembly. *The Journal of General Virology*, **81**, 1889–1899.

Kohn DB, Sadelain M, Dunbar C, Bodine D, Kiem HP, Candotti F, Tisdale J, Riviere I, Blau CA, Richard RE, *et al*. (2003) American Society of Gene Therapy (ASGT) ad hoc subcommittee on retroviral-mediated gene transfer to hematopoietic stem cells. *Molecular Therapy*, **8**, 180–187.

Kuate S, Wagner R and Uberla K (2002) Development and characterization of a minimal inducible packaging cell line for simian immunodeficiency virus-based lentiviral vectors. *The Journal of Gene Medicine*, **4**, 347–355.

Lever AML (2000) HIV RNA packaging and lentivirus-based vectors. *Advances in Pharmacology*, **48**, 1–28.

Lever AML (2003) Lentiviral vectors in gene therapy. In *Nature Encyclopedia of the Human Genome*, Cooper D (Ed.), Macmillan Publishers: Basingstoke, (In press).

Linial M, Medeiros E and Hayward WS (1978) An avian oncovirus mutant (SE 21Q1b) deficient in genomic RNA: Biological and biochemical characterization. *Cell*, **15**, 1371–1381.

Linial ML and Miller AD (1990) Retroviral RNA packaging: Sequence requirements and implications. *Current Topics in Microbiology and Immunology*, **157**, 125–152.

Miller AD (1990) Retrovirus packaging cells. *Human Gene Therapy*, **1**, 5–14.

Naldini L, Blomer U, Gallay P, Ory D, Mulligan P, Gage FH, Verma IM and Trono D (1996) *In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science*, **272**, 263–267.

Pannell D and Ellis J (2001) Silencing of gene expression: Implications for design of retrovirus vectors. *Reviews in Medical Virology*, **11**, 205–217.

Quinonez R and Sutton RE (2002) Lentiviral vectors for gene delivery into cells. *DNA and Cell Biology*, **21**, 937–951.

Sparacio S, Pfeiffer T, Schaal H and Bosch V (2001) Generation of a flexible cell line with regulatable, high-level expression of HIV Gag/Pol particles capable of packaging HIV-derived vectors. *Molecular Therapy*, **3**, 602–612.

Srinivasakumar N and Schuening FG (1999) A lentivirus packaging system based on alternative RNA transport mechanisms to express helper and gene transfer vector RNAs and its use to study the requirement of accessory proteins for particle formation and gene delivery. *Journal of Virology*, **73**, 9589–9598.

Trono D (2001) Lentiviral Vectors. *Lentiviral Vectors (Current Topics in Microbiology and Immunology)*, Springer-Verlag: Berlin, Heidelberg.

Zufferey R, Donello JE, Trono D and Hope TJ (1999) Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. *Journal of Virology*, **73**, 2886–2892.

Zufferey R, Nagy D, Mandel RJ, Naldini L and Trono D (1997) Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nature Biotechnology*, **15**, 871–875.

Short Specialist Review

Immunity and tolerance induction in gene therapy

Jean Davoust, Nicolas Bertho, Carole Masurier and David Gross
Généthon, Evry, France

1. Introduction

The goal of gene therapist facing the three-headed Cerberus guardian of the immune system composed of innate, cellular, and humoral effectors is to understand and control immune responses induced after gene transfer in order to design safe and efficient gene therapy protocols (Second Cabo gene therapy working group, 2002). This is of central importance in order to enable survival, growth or persistence of the corrected cells, which express a foreign antigen absent during the initial development of the immune system. As dendritic cells (DCs) maintain and/or initiate peripheral tolerance against self antigens and can eventually be tailored to induce tolerance against foreign antigens expressed after gene transfer, they are central to this question and careful attention should be devoted to DC/gene therapy vector interactions both in animal models and in humans in vitro cells.

Immune responses can either be immediate in case of hypersensitivity reactions to the vector, or adaptive and linked to the development of humoral and cellular responses to vectors and transgene products. Preexisting immunity to viral vectors can notoriously obviate gene engraftment, and the fine dissection of such responses must be pursued to identify the effector arms leading to rejection of transduced cells (Kafri *et al*., 1998; Brown and Lillicrap, 2002; Wang *et al*., 2005). Preexisting responses against the vector may not only hamper gene delivery, but may behave as an adjuvant eliciting adaptive responses to the transgene itself. This risk can, in principle, be evaluated in animal model, when the transgene product bears nonself–amino acid sequences. This is usually not straightforward, as one needs to design appropriate animal models with precise mutations in the target gene, and to deliver animal-matched gene sequences to correct the disease.

Several approaches can be used to counteract the deleterious effects of immune responses on transduced cells, in order to maintain long-lasting corrections in treated tissues. Nonspecific immunosuppressive treatments improve the duration of the gene transfer (Jooss *et al*., 1996; Jooss *et al*., 1998) but may compromise the vital functions of the immune system of the host. Antigen delivery to privileged tissues such as liver can induce a state of tolerance (Dobrzynski *et al*., 2004;

Mingozzi *et al*., 2003), and a general antigen-specific cotreatment has recently been proposed on the basis of adoptive transfer of antigen-educated regulatory CD4+ CD25+ T lymphocytes to inhibit the responses against recombinant adenovirus-associated vector (rAAV) transgene expression in the host (Jooss *et al*., 2001).

Several factors influencing the presentation of non-self antigens expressed from recombinant vectors should be taken into account, including the nature of the vector itself and the ability of the transgene to transduce directly the dendritic cells or to be cross-presented by DCs after the capture of transduced cell bodies (Sarukhan *et al*., 2001a,b; Berard *et al*., 2000; Guermonprez *et al*., 2003). DCs are the first line of immune cells encountering vector preparations and transgene products in vivo, and as such represent an important research field to be explored to evaluate the orientation of immune responses in each therapeutic setting (Pulendran, 2004). DCs also represent highly valuable tools to identify the epitopes being processed and presented by MHC class I and MHC class II molecules after antigen capture in the form of either viral particles or cell fragments containing transgene products (Guermonprez *et al*., 2002).

2. Deciphering dendritic cells/viral vector interactions

DCs play an important role in the distinction of self/non-self antigens, both in the thymus where they present self antigen to developing T lymphocytes and delete autoreactive T cells, and in the periphery where they are thought to activate immunoregulatory T cells. Although of major importance, central tolerance mechanisms are ineffective for a variety of self antigens either uniquely present in the periphery, expressed at late stages of the formation of the T cell repertoire, or for Ag that engage promiscuous T cell receptors that cross-react with self determinants. As recently proposed, DCs that reside in the periphery may migrate to secondary lymphoid organs at the steady state, inducing tolerance through cognate interactions with regulatory T cells (Steinman and Nussenzweig, 2002; Steinman *et al*., 2003). Conversely, DCs that have captured antigens in the periphery and/or are activated by a series of inflammatory cytokines, adjuvant components, and cell–cell contacts, may initiate immunostimulatory responses in lymph nodes leading to the expansion of B and T cell effectors.

The underlying role of DCs, and probably of other cell types expressing class II MHC molecules, emerges therefore as an important element in the initiation of immune responses to the vector and/or transgene products and in the induction of peripheral tolerance to specific antigens. Previous anatomo-pathology studies carried out in human lymph node draining inflammatory lesions have allowed to determine that immature DCs of the Langerhans type (LC) can be found in enflamed lymph nodes and that partially matured LCs acquired migration responsiveness to lymph nodes chemokines in vitro. These observations indicated that LCs as well as other DC subtypes can transport antigens to secondary lymphoid organs to induce either immunity or peripheral tolerance depending on their maturation stages. This transport and the local production of IL-10 could play a central role in the tolerization process of CD4+ T cells, and consequently of CD8+ T cells recruited in the lymph node microenvironment because of specific recognition

of transgene epitopes. Two lines of research need, therefore, to be conducted to tackle subtle interactions of DCs and gene therapy vectors, namely, the in vivo evaluation of DC activation by currently used gene transfer protocols, and the in vitro evaluation of human DC interactions with gene therapy vectors to qualify vector preparations for immune innocuousness and assess the tropism of transgene promoters.

2.1. Interactions between gene therapy vectors and antigens presenting cells

With the promising use of various vectors of clinical interest, such as AAV (adeno-associated virus vectors) and LV (lentivirus vectors), it is now important to assess the transduction efficacy according to the promoter selected for each application, and the ensuing activation of human DCs (activation of endocytosis, migration, and maturation properties). This knowledge will be helpful to minimize the immunogenicity of viral vectors preparations. For this purpose, one can differentiate different subtypes of DC (interstitial type, Langerhans, or plasmacytoid DCs) derived either from circulating monocytes or from CD34+ progenitors, to explore their transduction efficacy with nonreplicative viral vectors, their antigen presentation capacity after transduction, and the ensuing initiation of T cell responses.

2.2. Prospects for in vivo manipulation of dendritic cells and tolerance induction to gene therapy

In the long run, gene therapists wish to understand how DCs initiate peripheral tolerance against self antigens and can be tailored to induce tolerance against foreign antigens, which are expressed in cells corrected by gene therapy. DCs are cornerstones of immunity being effective antigen-presenting cells that play prominent roles in infectious diseases, immune disorders, tolerance, and vaccination (Banchereau and Steinman, 1998; Banchereau *et al*., 2000). At present, it is commonly believed that in the steady state, nonactivated DCs induce tolerance to peripheral self antigens by presenting these antigens but failing to engage T cell costimulation (Dhodapkar and Steinman, 2002). However, upon signal reception from pathogens, inflammatory or immune mediators, maturing DCs present antigens to costimulation-dependent, naive T cells and initiate immune responses. The mechanisms of T cell tolerance include anergy, changes in T cell development, induction of regulatory T cells, and T cell deletion. It is commonly accepted that targeting the foreign antigens to DC subpopulations in an immature state, or preventing DC maturation by direct interference on activation factors, will lead to the establishment of antigen-specific immune regulation.

DC-based tolerance induction implies, therefore, that one could target immature DCs and activate regulatory T cells in the steady state, to exert a suppressive role on specific CD4+ and CD8t T lymphocytes. The cell-to-cell contact dependence and the antigen specificity of the induction of regulatory T cells are now actively

explored (Tang *et al*., 2004; Tarbell *et al*., 2004) and can be summarized by a three-step model (Steinbrink *et al*., 2002; Mahnke *et al*., 2002; Jonuleit *et al*., 2002; Jonuleit *et al*., 2001) in humans: (1) APCs, probably resting DCs, activate CD4+ T regulatory cells, (2) they also interact with conventional CD4+ T cells resulting in the development of additional suppressor T cell populations, (3) these induced CD4+ T regulatory cells suppress the proliferation of freshly isolated conventional CD4+T cells via a process partially mediated by soluble TGF-ß, and suppress also CD8+ T cells (Dhodapkar and Steinman, 2002; Piccirillo and Shevach, 2001). Immature DCs can, in addition, induce regulatory CD8+ T cells (Dhodapkar and Steinman, 2002), and activated CD4+ CD25-regulatory T cells secrete huge amounts of IL-10, which spread the response and maintain DCs and monocytes in an immature or inactive stage in inflammatory sites. This model focuses our attention on MHC class II antigen presentation as a key element in inducing antigen-specific CD4+ T cell help and Ag-specific CD4+ tolerance. One may conceive that from an initial decisive interaction with DCs, which presumably occurs in draining lymph nodes, the CD4 T cell compartment generates either T helper or T regulatory responses, which determine the fate of B cells and cytotoxic CD8+ T cells effectors.

In future prospects, it seems therefore crucial to delineate epitopes presented by MHC class II molecules using transgenic strains for HLA-DR MHC class II molecules, gearing regulatory T cell responses.

3. Immunological design of transgene-specific tolerance with CD4+ CD25+ regulatory T cells

Successful gene transfer requires immunosuppression to arrest the immune response and achieve expression of the engrafted gene. However, long-lasting and life-threatening immunosuppression is undesirable in most therapeutic applications. Alternative antigen-specific immunosuppressive protocols can now be designed with regulatory T cells. The role of CD4+ CD25+ regulatory T cells (Tregs) and of IL-10 producing T cells during the control of immune responses has now been established in several situations: (1) transplantation (Waldmann and Cobbold, 2004; Walsh *et al*., 2004), after which the graft versus host disease may be alleviated by the donor's CD4+ CD25+ regulatory T lymphocytes (Taylor *et al*., 2004; Trenado *et al*., 2003); (2) autoimmune pathologies and chronic inflammations, where regulatory T lymphocytes can attenuate the immune responses; (3) tolerance induction by persisting pathogens, and tolerance induction in the airway and intestinal tracks (Shevach, 2002; Sakaguchi, 2004; O'Garra and Vieira, 2004). Although their mode of action is still a matter of debate, the CD4+ regulatory T cell subsets can probably interact with antigen-presenting cells (APC) loaded with fragments from foreign proteins to exert their suppressive role. The recruitment of antigen-specific regulatory T cells at the site of gene transfer and/or in draining lymphoid organs should be a mean to nullify the initiation of immune responses by APCs carrying transgene products.

The potency of antigen-specific Tregs to counteract immune responses after gene transfer can be evaluated using regulatory T lymphocytes bearing a transgenic T

cell receptor that recognizes a single epitope derived from a model transgene, such as the hemagglutinin (HA) protein of Influenza virus. Indeed, HA-restricted Tregs transferred into unconditioned animals could exert their suppressive function on both humoral and cellular immune responses and blocked the rejection of muscle fibers transduced with the HA model transgene (Gross *et al*., 2003). To gain insight in the suppressive mechanism, it is now of importance to explore the type of DCs involved as well as the robustness of the phenomenon of immune tolerance. As discussed above, it is likely that in this system, DCs either capture foreign Ag and present them through a cross-presentation pathway (Guermonprez *et al*., 2003) or express the transgene directly to prime regulatory T cells.

4. Monitoring immune responses through identification of epitopes present on transgenes

As cellular immune responses may compromise long-term expression in various gene therapy strategies, studying the cytotoxic responses generated after gene transfer in diseased animal models remains of central importance. To explore humanized HLA-restricted responses in a preclinical setting, it is therefore desirable to construct models in which the mutated phenotype is present in mice carrying human HLA class I molecules such as the most represented HLA-A*0201 in Caucasian populations (Firat *et al*., 1999). Choosing as a model mdx mice lacking full dystrophin sequence expression due to a stop codon on exon 23, the HLA-A*0201 character would be able to educate a "humanized" T cell repertoire, which contains reactivities to nonself dystrophin sequences, allowing therefore the identification of HLA-A*0201 dystrophin rejection epitopes. In such mice, cellular immune responses may block long-term restoration of wild-type dystrophin provided by gene therapy treatments, and cytotoxic T cell reactivities are to be analyzed by combining epitope prediction with in vivo cytotoxic assays after gene delivery using full-length dystrophin plasmid. These studies led to the identification of a human specific HLA-A*0201 epitope hDys1281 (Ginhoux *et al*., 2003) and may lead to the identification of new epitopes revealed only on the mdx/HLA-A*0201 background.

Recent results indicated that in contrast to plasmid delivery, rescue of muscle dystrophin expression by exon-skipping after administration of either an oligonucleotide (Lu *et al*., 2003), or an AAV vector expressing antisense sequences linked to a modified U7 small nuclear RNA (Goyenvalle *et al*., 2004) restored long-term expression of dystrophin within muscle fibers. Both of these restorations failed to trigger immune rejection of muscle fibers expressing the corrected dystrophin sequence. The above-mentioned mdx/HLA-A*0201 animals may, therefore, represent appealing models to compare the induction of CTL activity against dystrophin depending on the gene repair protocol being used. Preliminary results indicated that the dystrophin exon skipping procedure, which restores nearly full-length dystrophin messenger mRNA (Lu *et al*., 2003; Goyenvalle *et al*., 2004), did not trigger cytotoxic immune responses on mdx/HLA-A*0201 background, indicating that immune responses initiated in dystrophin defective mice depend on the gene transfer protocol.

The above-mentioned example proves that appropriate diseased HLA-A*0201 animal models could provide interesting clues on possible states of immune ignorance, susceptible to breakage after a secondary inflammatory challenges. To assay the occurrence of deleterious cellular responses in restored muscles, inflammatory agents, and/or antigenic peptide are to be injected, as ultimate tests for immunological safety. As in other gene therapy settings in which foreign sequences are transferred in immunocompetent hosts, the HLA-A*0201-restricted human dystrophin epitopes unraveled here (Ginhoux *et al*., 2003) are important tools to monitor the occurrence of immune responses in humans.

Finally, to circumvent the advent of unavoidable cellular responses, attempts have been conducted to shield vectors and transgenes from the immune response. The overall success of this strategy may depend on gene transfer modalities and on subtle combinations of purity, transduction efficiencies, and cell target specificities of the viral vectors used. Along this line, we initiated a study with a model transgene vectorized either with plasmid or AAV viral particles and showed that epitope engineering designed to mask an immunodominant sequence did generated in fact a set secondary rejection epitopes (Ginhoux *et al*., 2004). A Sisyphean task awaits therefore the gene therapists wishing to shield nonself epitopes recognized by the T cell repertoire in the host.

5. Conclusion

Humanized animal models are now highly desirable to explore the type of immune responses encountered and to define molecular determinants, which gear adaptive immune responses to transgene products. However, a lack of response in small animal models does not necessarily mean a lack of immunogenicity in humans. Nonspecific inflammatory conditions, which may differ from mice to humans and preexisting immune responses to vectors themselves may favor the initiation of immune responses to the transgene, much like the conversion of autoimmune T cell reactivity into overt autoimmune disease (Lang *et al*., 2005). So studying both the factors, which activate initial innate immune responses, and the molecular determinants present on transgenes will give important clues in the outcome of immune responses to gene transfer.

We hope *in fine* to induce long-lasting immune tolerance against foreign antigens expressed after a gene transfer. This can be tentatively achieved with regulatory T cells, which can blunt the initiation of immune responses and probably control inflammatory processes. In ideal circumstances, the establishment of mixed hematopoietic chimerism (Sykes, 2001) using the transgene of interest as a minor antigens should provide a strong therapeutic tolerance and obviate immune responses, by subduing all three heads of the Cerberus guardian.

Acknowledgments

We wish to thank Terry Partridge, Carole Masurier, David Gross, Nicolas Bertho, and Olivier Danos for critical reading of the manuscript, Florent Ginhoux and

Sabrina Turbant for privileged communications, and Huseyin Firat, Marylène Leboeuf, and François Lemonnier for constant support in the construction of HLA-A*0201 mouse models. This work was supported by the Association Française contre les Myopathies, by ATIGE grant from the GIP Genopole Evry, France, by the Fondation pour la Recherche Médicale and the EC Inherinet program (QLK3-CT-2001-00427).

References

Banchereau J, Briere F, Caux C, Davoust J, Lebecque S, Liu YJ, Pulendran B and Palucka K (2000) Immunobiology of dendritic cells. *Annual Review of Immunology*, **18**, 767–811.

Banchereau J and Steinman RM (1998) Dendritic cells and the control of immunity. *Nature*, **392**, 245–252.

Berard F, Blanco P, Davoust J, Neidhart-Berard EM, Nouri-Shirazi M, Taquet N, Rimoldi D, Cerottini JC, Banchereau J and Palucka AK (2000) Cross-priming of naive CD8 T cells against melanoma antigens using dendritic cells loaded with killed allogeneic melanoma cells. *The Journal of Experimental Medicine*, **192**, 1535–1544.

Brown BD and Lillicrap D (2002) Dangerous liaisons: the role of "danger" signals in the immune response to gene therapy. *Blood*, **100**, 1133–1140.

Dhodapkar MV and Steinman RM (2002) Antigen-bearing immature dendritic cells induce peptide-specific CD8(+) regulatory T cells in vivo in humans. *Blood*, **100**, 174–177.

Dobrzynski E, Mingozzi F, Liu YL, Bendo E, Cao O, Wang L and Herzog RW (2004) Induction of antigen-specific CD4+ T-cell anergy and deletion by in vivo viral gene transfer. *Blood*, **104**, 969–977.

Firat H, Garcia-Pons F, Tourdot S, Pascolo S, Scardino A, Garcia Z, Michel ML, Jack RW, Jung G, Kosmatopoulos K, *et al*. (1999) H-2 class I knockout, HLA-A2.1-transgenic mice: a versatile animal model for preclinical evaluation of antitumor immunotherapeutic strategies. *European Journal of Immunology*, **29**, 3112–3121.

Ginhoux F, Doucet C, Leboeuf M, Lemonnier FA, Danos O, Davoust J and Firat H (2003) Identification of an HLA-A*0201-restricted epitopic peptide from human dystrophin: application in Duchenne muscular dystrophy gene therapy. *Molecular Therapy*, **8**, 274–283.

Ginhoux F, Turbant S, Gross DA, Poupiot J, Marais T, Lone Y, Lemonnier FA, Firat H, Perez N, Danos O, *et al*. (2004) HLA-A*0201-restricted cytolytic responses to the rtTA transactivator dominant and cryptic epitopes compromise transgene expression induced by the tetracycline on system. *Molecular Therapy*, **10**, 279–289.

Goyenvalle A, Vulin A, Fougerousse F, Leturcq F, Kaplan JC, Garcia L and Danos O (2004) Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping. *Science*, **306**, 1796–1799.

Gross DA, Leboeuf M, Gjata B, Danos O and Davoust J (2003) CD4+CD25+ regulatory T cells inhibit immune-mediated transgene rejection. *Blood*, **102**, 4326–4328.

Guermonprez P, Saveanu L, Kleijmeer M, Davoust J, Van Endert P and Amigorena S (2003) ER-phagosome fusion defines an MHC class I cross-presentation compartment in dendritic cells. *Nature*, **425**, 397–402.

Guermonprez P, Valladeau J, Zitvogel L, Thery C and Amigorena S (2002) Antigen presentation and T cell stimulation by dendritic cells. *Annual Review of Immunology*, **20**, 621–667.

Jonuleit H, Schmitt E, Kakirman H, Stassen M, Knop J and Enk AH (2002) Infectious tolerance: human CD25(+) regulatory T cells convey suppressor activity to conventional CD4(+) T helper cells. *The Journal of Experimental Medicine*, **196**, 255–260.

Jonuleit H, Schmitt E, Stassen M, Tuettenberg A, Knop J and Enk AH (2001) Identification and functional characterization of human CD4+CD25+ T cells with regulatory properties isolated from peripheral blood. *The Journal of Experimental Medicine*, **193**, 1285–1294.

Jooss K, Ertl HC and Wilson JM (1998) Cytotoxic T-lymphocyte target proteins and their major histocompatibility complex class I restriction in response to adenovirus vectors delivered to mouse liver. *Journal of Virology*, **72**, 2945–2954.

Jooss K, Gjata B, Danos O, von Boehmer H and Sarukhan A (2001) Regulatory function of in vivo anergized CD4(+) T cells. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8738–8743.

Jooss K, Yang Y and Wilson JM (1996) Cyclophosphamide diminishes inflammation and prolongs transgene expression following delivery of adenoviral vectors to mouse liver and lung. *Human Gene Therapy*, **7**, 1555–1566.

Kafri T, Morgan D, Krahl T, Sarvetnick N, Sherman L and Verma I (1998) Cellular immune response to adenoviral vector infected cells does not require de novo viral gene expression: implications for gene therapy. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 11377–11382.

Lang KS, Recher M, Junt T, Navarini AA, Harris NL, Freigang S, Odermatt B, Conrad C, Ittner LM, Bauer S, *et al*. (2005) Toll-like receptor engagement converts T-cell autoreactivity into overt autoimmune disease. *Nature Medicine*, **11**, 138–145.

Lu QL, Mann CJ, Lou F, Bou-Gharios G, Morris GE, Xue SA, Fletcher S, Partridge TA and Wilton SD (2003) Functional amounts of dystrophin produced by skipping the mutated exon in the mdx dystrophic mouse. *Nature Medicine*, **9**, 1009–1014.

Mahnke K, Schmitt E, Bonifaz L, Enk AH and Jonuleit H (2002) Immature, but not inactive: the tolerogenic function of immature dendritic cells. *Immunology and Cell Biology*, **80**, 477–483.

Mingozzi F, Liu YL, Dobrzynski E, Kaufhold A, Liu JH, Wang Y, Arruda VR, High KA and Herzog RW (2003) Induction of immune tolerance to coagulation factor IX antigen by in vivo hepatic gene transfer. *The Journal of Clinical Investigation*, **111**, 1347–1356.

O'Garra A and Vieira P (2004) Regulatory T cells and mechanisms of immune system control. *Nature Medicine*, **10**, 801–805.

Piccirillo CA and Shevach EM (2001) Cutting edge: control of CD8+ T cell activation by CD4+CD25+ immunoregulatory cells. *Journal of Immunology*, **167**, 1137–1140.

Pulendran B (2004) Modulating vaccine responses with dendritic cells and Toll-like receptors. *Immunological Reviews*, **199**, 227–250.

Sakaguchi S (2004) Naturally arising CD4+ regulatory t cells for immunologic self-tolerance and negative control of immune responses. *Annual Review of Immunology*, **22**, 531–562.

Sarukhan A, Camugli S, Gjata B, von Boehmer H, Danos O and Jooss K (2001a) Successful interference with cellular immune responses to immunogenic proteins encoded by recombinant viral vectors. *Journal of Virology*, **75**, 269–277.

Sarukhan A, Soudais C, Danos O and Jooss K (2001b) Factors influencing cross-presentation of non-self antigens expressed from recombinant adeno-associated virus vectors. *The Journal of Gene Medicine*, **3**, 260–270.

Second Cabo gene therapy working group (2002) Cabo II: immunology and gene therapy. *Molecular Therapy*, **5**, 486–491.

Shevach EM (2002) CD4+ CD25+ suppressor T cells: more questions than answers. *Nature Reviews. Immunology*, **2**, 389–400.

Steinbrink K, Graulich E, Kubsch S, Knop J and Enk AH (2002) CD4(+) and CD8(+) anergic T cells induced by interleukin-10-treated human dendritic cells display antigen-specific suppressor activity. *Blood*, **99**, 2468–2476.

Steinman RM, Hawiger D and Nussenzweig MC (2003) Tolerogenic dendritic cells. *Annual Review of Immunology*, **21**, 685–711.

Steinman RM and Nussenzweig MC (2002) Avoiding horror autotoxicus: the importance of dendritic cells in peripheral T cell tolerance. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 351–358.

Sykes M (2001) Mixed chimerism and transplant tolerance. *Immunity*, **14**, 417–424.

Tang Q, Henriksen KJ, Bi M, Finger EB, Szot G, Ye J, Masteller EL, McDevitt H, Bonyhadi M and Bluestone JA (2004) In vitro-expanded antigen-specific regulatory T cells suppress autoimmune diabetes. *The Journal of Experimental Medicine*, **199**, 1455–1465.

Tarbell KV, Yamazaki S, Olson K, Toy P and Steinman RM (2004) CD25+ CD4+ T cells, expanded with dendritic cells presenting a single autoantigenic peptide, suppress autoimmune diabetes. *The Journal of Experimental Medicine*, **199**, 1467–1477.

Taylor PA, Panoskaltsis-Mortari A, Swedin JM, Lucas PJ, Gress RE, Levine BL, June CH, Serody JS and Blazar BR (2004) L-Selectin(hi) but not the L-selectin(lo) CD4+25+ T-regulatory cells are potent inhibitors of GVHD and BM graft rejection. *Blood*, **104**, 3804–3812.

Trenado A, Charlotte F, Fisson S, Yagello M, Klatzmann D, Salomon BL and Cohen JL (2003) Recipient-type specific CD4+CD25+ regulatory T cells favor immune reconstitution and control graft-versus-host disease while maintaining graft-versus-leukemia. *The Journal of Clinical Investigation*, **112**, 1688–1696.

Waldmann H and Cobbold S (2004) Exploiting tolerance processes in transplantation. *Science*, **305**, 209–212.

Walsh PT, Taylor DK and Turka LA (2004) Tregs and transplantation tolerance. *The Journal of Clinical Investigation*, **114**, 1398–1403.

Wang L, Dobrzynski E, Schlachterman A, Cao O and Herzog RW (2005) Systemic protein delivery by muscle gene transfer is limited by a local immune response. *Blood*, **105**, 4226–4234.

Short Specialist Review

Gene transfer vectors as medicinal products: risks and benefits

Christian J. Buchholz and Klaus Cichutek
Paul-Ehrlich-Institut, Langen, Germany

The potential applications of gene therapy are plentifold, reaching from hereditary disease to acquired multifactorial disorders. The vector systems used for the transfer of the therapeutic gene into the patient are equally diverse including, for example, naked plasmid DNA as well as engineered viruses or genetically modified cells. Regarded as medicinal product when used in vivo, gene transfer medicinal products (GT-MPs) are unique in their extent of diversity and complexity not only relative to conventional chemical drugs but also relative to biological pharmaceuticals like recombinant proteins. They, therefore, equally challenge research and regulation.

Medicinal products, also termed drugs or medicines, are used for the treatment, diagnosis, or prevention of diseases in or on human subjects. Gene therapy and somatic cell therapy using genetically modified cells are best summarized under the term clinical gene transfer and involve the treatment of human subjects with GT-MPs. GT-MPs belong to the group of advanced biotechnology products for which testing provisions for marketing authorization have been specified in a legally binding document for the European Union. Marketing authorization for a GT-MP was recently granted for the first time worldwide in China for an adenoviral vector transferring the human p53 gene (Pearson *et al*., 2004). This product is being used for cancer treatment to restore the apoptosis pathway in p53-deficient tumor cells. Data showing clinical efficacy of this GT-MP have not been published.

GT-MPs have been used worldwide by the end of 2004 in almost 1000 clinical trials for in vivo diagnostics, prevention, or therapy. Most clinical gene transfer protocols are aimed at the treatment or prevention of cancer, cardio-vascular disease, infectious disease such as AIDS or monogeneic disorders, or at the prevention of infectious disease by vaccination. The vectors most often used ex vivo are MLV-derived gamma-retroviral vectors, whereas adenoviral and poxviral vectors have mostly been used in vivo. An increasing number of studies involves the use of nonviral vectors or naked DNA (refer to Table in (*see* Article 95, **Artificial self-assembling systems for gene therapy**, Volume 2).

More than 6000 human subjects have been treated with GT-MPs worldwide, most of them in the United States. Within Europe, probably the highest number of clinical gene transfer studies have been carried out in the United Kingdom and

Table 1 Clinical progress in gene therapy

• Adenosine deaminase deficiency	Adenosine deaminase gene (ada)	Blood stem cells/retroviral vector	2 patients cured	Aiuti *et al*. (2002)
• SCID-X1	Gamma-c-chain (IL-2 R)	Blood stem cells//retroviral vector	10 of 11 babies cured[a]	Cavazzana-Calvo *et al*. (2000), Gaspar *et al*. (2004)
• Peripheral artery occlusive disease	Vascular endothelial growth factor (VEGF)	i.m./plasmid DNA	Improved vascularization	Gruchala *et al*. (2004)
• Head and neck tumors	Conditionally replicating adenovirus, no transgene	Tumor cells	Local tumor regression	Post (2002)
• Leukemia, graft versus host treatment	Herpes simplex virus thymidine kinase (HSV-tk)	T cells/retroviral vector	Succesful graft versus host treatment	Bonini *et al*. (1997)
• Hemophilia B	Factor IX	i.m./AAV vector[b]	Improved plasma levels	Couto and Pierce (2003)

[a]Two patients developed lymphoproliferative disease due to vector integration.
[b]AAV: adenovirus associated virus.

in Germany. A few clinical trials have resulted so far in benefit for the patients involved (Table 1). It is becoming clear that each disease needs the development of a particular gene transfer method and regiment.

A particular case was the SCID-X1 trial performed by Alain Fischer's group at the Hospital Necker in Paris. This clinical trial drew most attention, reaching beyond the gene therapy field. From the same study, spectacular therapeutic effects were reported; however, novel, until then unexpected serious adverse reactions were also reported. SCID-X1 is caused by an inherited disorder, which occurs in the γ subunit of cytokine receptors encoded by the γc gene (Fischer *et al*., 2002). The γc polypeptide is used as a joint subunit in a number of interleukin receptors relevant for hematopoesis, including the receptors for IL-2, IL-4, IL-7, or IL-9. In children with SCID-X1, no functionally active cytokine receptor is found on the surface of lymphocyte precursor cells. This results in a complete failure of the cells to differentiate into T lymphocytes and natural killer (NK) cells. Thus, newborns with this congenital disorder lack the functional immune response to infectious diseases and are therefore forced to live under germ-free conditions. Conventional treatment of this disease requires allogeneic bone marrow transfer, if a suitable donor is available.

The gene therapy approach uses $CD34^+$ cells isolated from the patient's bone marrow, activated with a cytokine mixture, and then transduced with the γc gene by means of a conventional replication-incompetent retroviral vector derived from murine leukemia virus (Figure 1). The transduced cells are then reinfused into the patients. Until October 2002, when owing to the observed adverse events the trial was put on hold, 10 SCID-X1 patients had been treated. In nine patients, the gene therapy approach was able to provide a fully functional immune system during an

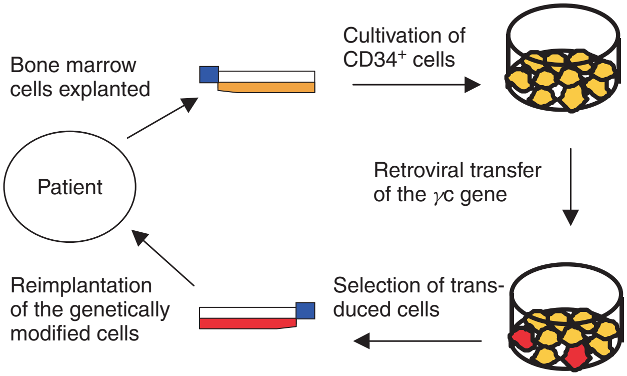


Figure 1 Gene therapy of SCID-X1. The first step is to purify, stimulate, and culture CD34$^+$ cells from the patient's bone marrow. The cells are then transduced with the γc gene using a retroviral vector, before they are reinfused into the patient

observation period of 3 years, partly even longer. For the affected children and their parents, this meant leading a normal life. The children could even tolerate vaccination against common infectious diseases.

However, about 30 months after treatment, two of the cured patients developed a T cell proliferative syndrome reminiscent of adult T cell leukemia showing a strong increase in the number of T lymphocytes, accompanied by splenomegaly and anemia. Owing to conventional chemotherapy, both patients recovered; one of them is currently in good health but the other patient died because of leukemia (Hacein-Bey-Abina *et al*., 2003). It is meanwhile commonly accepted that this disease was a result of chromosomal integration of the retroviral vector into the locus of the *Imo-2* gene, a known proto-oncogene, resulting in strong overexpression. The *Imo-2* gene encodes a transcription factor (rhombotin-2), which is upregulated during hematogenesis but usually not expressed in mature T lymphocytes. Expression of *Imo-2*, however, can lead to acute T cell leukemia (Herblot *et al*., 2000). Cofactors besides lmo-2 overexpression that contributed to the T cell proliferation might include the proliferative signals mediated through the transferred γc gene product, as well as familial predisposition to cancer or a chickenpox infection in one of the patients.

Thus, for the first time, insertional oncogenesis had manifested itself in a patient treated with retroviral vectors in the SCID-X1 clinical trial. Cancer development after retroviral gene transfer could therefore no longer be considered exclusively as a theoretical risk. As outlined below, it had immediate consequences for this and also for other clinical gene transfer studies using retroviral vectors.

In the European Union and also in the United States, national authorities are responsible for the registration of clinical trials. In the European Union, Directive 2001/20/EC lays down the rules for good clinical practice (GCP). According to this directive, which had to be transformed into national laws by May 2004, clinical trials using GT-MPs require approval by the competent national authority as well as a positive appraisal by an ethics committee. While the ethics committees, which are usually supported by advice from central gene therapy committees with expert

members that cover the different fields of gene therapy, focus on the ethical and medical acceptability of the study, the competent authorities evaluate the acceptability according to the current standards of science. For phase I clinical trials, which is by far the majority of all ongoing clinical gene therapy trials, this includes evaluation of preclinical data on quality, pharmacology, toxicology, and potency of the investigational drug.

The surveillance of clinical trials is another important task carried out by the national authorities. In the event of a suspected risk for enrolled patients or subjects, the competent authority can coordinate and exert suitable measures. Immediately when the reports about the severe adverse reactions in the SCID-X1 trials carried out in France became available, expert hearings took place not only in France but also in other countries such as Germany, the United Kingdom, and the United States to reconsider the safety of retroviral gene transfer. As an initial action (provision), the enrollment of further patients was put on hold at least for all studies using retroviral vectors for blood stem cell modification, in some countries for all clinical gene therapy studies using live retrovirally modified cells. The principal investigators, however, were given the opportunity to review the risk/benefit analysis on the basis of the new results, and to make the appropriate changes in the patient informed consent and the inclusion and exclusion criteria. All studies are currently being continued.

Especially, the risk/benefit analysis is of particular importance in this situation. For life-threatening disease where no alternative treatment is available, even the risk of leukemia may be justified. For example, treatment of patients suffering from chronic granulomatosis, who have a shortened life expectancy due to their congenital immune deficiency disorder, may be continued after changes in the protocol. On the other hand, pure marker gene studies using retroviral vectors without any direct benefit for the patient to be treated currently have a low likelihood of being authorized.

References

Aiuti A, Slavin S, Aker M, Ficara F, Deola S, Mortellaro A, Morecki S, Andolfi G, Tabucchi A, Carlucci F, *et al*. (2002) Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science*, **296**, 2410–2413.

Bonini C, Ferrari G, Verzeletti S, Servida P, Zappone E, Ruggieri L, Ponzoni M, Rossini S, Mavilio F, Traversari C, *et al*. (1997) HSV-TK gene transfer into donor lymphocytes for control of allogeneic graft-versus-leukemia. *Science*, **276**, 1719–1724.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, *et al*. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.

Couto LB and Pierce GF (2003) AAV-mediated gene therapy for hemophilia. *Current Opinion in Molecular Therapeutics*, **5**, 517–523.

Fischer A, Hacein-Bey S and Cavazzana-Calvo M (2002) Gene therapy of severe combined immunodeficiencies. *Nature Reviews Immunology*, **2**, 615–621.

Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, Sinclair J, Brouns G, Schmidt M, Von Kalle C, Barington T, *et al*. (2004) Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet*, **364**, 2181–2187.

Gruchala M, Roy H, Bhardwaj S and Ylä-Herttuala S (2004) Gene therapy for cardiovascular diseases. *Current Pharmaceutical Design*, **10**, 407–423.

Hacein-Bey-Abina S, von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E, *et al*. (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.

Herblot S, Steff AM, Hugo P, Aplan PD and Hoang T (2000) SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression. *Natural Immunity*, **1**, 138–144.

Pearson S, Jia H and Kandachi K (2004) China approves first gene therapy. *Nature Biotechnology*, **22**, 3–4.

Post LE (2002) Selectively replicating adenoviruses for cancer therapy: an update on clinical development. *Current Opinion in Investigational Drugs*, **3**, 1768–1772.

Gene transfer to skeletal muscle

Jeffrey S. Chamberlain
University of Washington School of Medicine, Seattle, WA, USA

1. Gene vectors for muscle tissues

Gene transfer into muscles can take a variety of forms, from delivery of whole genes, mini-gene expression cassettes, to oligonucleotides. Each of these types of genetic elements can, in theory, be delivered either as a simple DNA sequence, in the context of a plasmid vector, or embedded within a whole or modified viral genome. Oligonucleotides represent the simplest type of DNA to deliver, as they are easily synthesized and generally nontoxic. These sequences are typically short (15–100 bases) and do not encode an actual gene, but are intended to modify gene expression or directly target genetic mutations. Plasmids are also a simple method to transfer genes, and have the capacity to carry oligonucleotides, entire genes (or fragments thereof), or synthetic expression cassettes. Many applications of muscle gene therapy focus on the use of virally derived vectors to transfer genes or oligonucleotides. Viral vectors display the most efficient ability to target and enter muscle cells, so the majority of research on muscle gene therapy has used these vectors. However, viral vectors deliver not only the therapeutic nucleic acid sequence but also portions of the viral genome and/or virally encoded proteins. Consequently, a variety of specific modifications to the viral genomes involving deletion of some or all viral genes are typically used to increase the safety and to modulate the immunogenicity of these vectors. The viral vectors commonly used for direct muscle gene transfer include recombinant vectors derived from adenovirus and adeno-associated viruses.

2. Oligonucleotide transfer to muscle

Delivery of short oligonucleotides composed of DNA or DNA/RNA chimeras have been explored to modulate muscle gene expression or direct specific genomic DNA sequence modifications (mutagenesis) in muscle cells. The major application of this technology has been with the delivery of short DNA oligonucleotides that can bind to splice donor or acceptor sequences in muscle genes to trigger alternative splicing of RNA transcripts. This approach is a potentially powerful means to treat Duchenne muscular dystrophy (DMD), which typically arises from either a premature stop codon or by frameshifting mutations resulting from deletions or

splice site mutations that lead to exclusion of a particular exon or exon(s) from the encoded dystrophin mRNA (Lu *et al*., 2003b; van Deutekom *et al*., 2001). These oligonucleotides are delivered to myogenic cells in culture by conventional DNA transfection methods, or to intact muscles of animal models by direct injection. Oligonucleotides are generally nontoxic, do not elicit a cellular or humoral immune response, and they have a relatively short half-life, enabling discontinuation of an experiment or therapeutic intervention if unwarranted side effects develop. They can also be mass-produced using synthetic technologies. The disadvantages are that oligonucleotides display a relatively short half-life in vivo, on the order of weeks, which would require repetitive dosing to treat a chronic disorder. The other main disadvantage is that cellular uptake of oligonucleotides is inefficient, such that huge quantities are required, and a relatively low percentage of the musculature is targeted. To date, oligonucleotides have only been applied for localized delivery to individual muscles, although technological improvements, including delivery using viral vectors (below) may lead to methods for widespread delivery.

The second major application of oligonucleotides in muscle has been for the correction of genetic mutations. Several classes of oligonucleotides, including long (~50–200 bp) DNA oligonucleotides and chimeric DNA/RNA oligonucleotides (chimeroplasts) can lead to targeting of specific genomic sequences and subsequent repair or conversion of a mutant or mismatched base to the corresponding sequence of the oligonucleotide (Bertoni and Rando, 2002). Through mechanisms that are largely undefined, base pairing between the mutant genomic DNA sequence and the oligonucleotide can lead to correction of the mutation in the genomic DNA. The theoretical advantage of this approach is that it combines the safety factor of nonviral vectors with a method to permanently correct a genetic mutation. The disadvantage is that to date this approach has proved to be inefficient, with an ability to correct at most ~1% of the mutant genomes near the injection site.

3. Plasmids for gene transfer to muscle

Plasmids were the first vectors used to deliver genes to muscles and are still in wide use. In theory, any gene or expression cassette can be inserted into a plasmid for delivery, so these vectors can accommodate large genes. However, the efficiency of gene transfer seems to drop with increasing plasmid size. Moderate levels of transduction can be obtained by direct intramuscular injection of plasmids, although efficiencies vary with the age of the animal, the species, and the method of injection (Zhang *et al*., 2004a). Plasmid expression vectors enable one to choose a particular promoter, such as a muscle-specific promoter, to drive gene expressions (Ferrer *et al*., 2000). Although intramuscular injections do not result in widespread muscle gene transfer, intravenous and intravascular methods have been developed using high pressure and volume injections that enable wider gene transfer to numerous muscles in a limb (Zhang *et al*., 2004b). Increased muscle transduction can be achieved by combining direct injection with the use of electroporation, and injection of block polymers (pluronics), or microbubbles in combination with ultrasound (Lu *et al*., 2003a). To date the best transfer efficiencies are between 10 and 50% of the myofibers in limb muscles (Zhang *et al*., 2004b). This promising method has

already been used in a phase 1 human clinical trial for DMD without adverse events, although little gene expression was detected with the doses used (Romero *et al*., 2002).

4. Adenoviral vectors

Vectors derived from adenoviruses were the first viral vectors used for muscle gene transfer (*see* Article 97, **Adeno-associated viral vectors: depend(o)ble stability**, Volume 2). Adenoviral (Ad) vectors are deleted for the E1 region of the viral genome, rendering them largely replication-defective. These conventional Ad vector have a cloning capacity of about 7 kb and are relatively efficient at muscle gene transfer, although the efficiency drops with increasing muscle maturation due to loss of the Ad receptor CAR (Nalbantoglu *et al*., 2001). Nonetheless, dystrophic and regenerating muscles of adult animals can still be efficiently targeted (DelloRusso *et al*., 2002). The major drawback of Ad vectors is their propensity to elicit potent cellular and innate immune responses, due primarily to leaky viral gene expression and effects of the capsid proteins, respectively. Improved vectors with reduced viral gene expression and lower replication capacity have been developed by deletion of additional viral genes, primarily the *E2b* and the *E4* genes (Barjot *et al*., 2002). The best Ad vectors for muscle gene transfer are fully deleted, or gutted Ad vectors, that lack all viral genes (DelloRusso *et al*., 2002; Chen *et al*., 1997). These vectors are more difficult to grow, and require a helper Ad vector that is removed from the vector preparations by Cre-recombinase-mediated destruction of the helper virus followed by CsCl density gradient centrifugation (Parks *et al*., 1996). Gutted Ad vectors can carry genes up to 30 kb in size, and have been used to efficiently deliver the full-length 14-kb dystrophin cDNA to adult *mdx* mice, a model for DMD (DelloRusso *et al*., 2002; Chen *et al*., 1997). Ad vector delivery has been limited to intramuscular injection, as they appear too large to traverse the vascular endothelial barrier lining blood vessels (Gregorevic *et al*., 2004a; Greelish *et al*., 1999). However, the residual ability of these vectors to elicit an innate immune response coupled with their difficult production requirements have greatly limited their application in human clinical trials. These vectors are still a good choice for gene transfer in culture, and are still being developed owing to their large cloning capacity.

5. Adeno-associated viral (AAV) vectors

The most promising vectors for muscle gene transfer currently are derived from adeno-associated viruses (rAAV) (Gregorevic *et al*., 2004b; *see also* Article 96, **Adenovirus vectors**, Volume 2). AAV-derived vectors typically do not elicit significant innate or cellular immune responses, at least as assayed in normal mice, and have been shown to support gene expression for at least 4 years following delivery to muscles of mice, dogs, and nonhuman primates (Xiao *et al*., 1996; Manno *et al*., 2003). In dystrophic muscle, however, some vectors expressing foreign transgenes elicit a cellular immune response when the transgene is expressed

from a ubiquitously active promoter, such as CMV, but not when expressed from a muscle-specific promoter, such as MCK (Yuasa *et al*., 2002). The major limitation of rAAV vectors is their small, 5 kb cloning capacity. For large genes, such as dystrophin, this limited cloning capacity has been partially addressed by the development of truncated, yet highly functional micro-dystrophin cDNAs expressed from small muscle-specific promoters (Wang *et al*., 2000; Harper *et al*., 2002). Delivery of micro-dystrophins to dystrophic *mdx* mice has been shown to prevent and partially reverse dystrophic pathologies in mice. Nine different serotypes of AAV have been tested for the ability to target a variety of different tissues, and vectors derived from AAV1, 5, and 6 have shown the best ability to transduce muscle (Gao *et al*., 2004; Blankinship *et al*., 2004). These vectors are easily delivered by direct intramuscular injection. Unlike Ad vectors, which begin expressing within 24 h of injection, rAAV vectors take 2–5 weeks to achieve maximal transgene transcription (Malik *et al*., 2000; Blankinship *et al*., 2004). The small size of rAAV vectors has also proven to be a great advantage in the development of methods for whole-body targeting of muscle tissues. Vectors based on AAV6 have been shown capable of transducing all the striated muscles in adult mice with various transgenes, including micro-dystrophins, when injected directly into the tail vein (Gregorevic *et al*., 2004a). Other methods have been used for targeting muscles of individual limbs (Greelish *et al*., 1999). These approaches represent potentially powerful methods for gene therapy of a variety of muscle disorders. Vectors based on rAAV may also be useful for delivering other expression cassettes, including antisense oligonucleotides, which was recently shown to be a promising approach for DMD (Goyenvalle *et al*., 2004). rAAV vectors are easily grown in cell cultures by a 2-plasmid cotransfection system (Blankinship *et al*., 2004). Unlike Ad vectors, rAAV vectors typically do not work well for gene transfer in cell culture systems.

6. Conclusions

Numerous vectors are available for gene transfer to muscle. The choice of vectors and delivery route is dependent on the specific application. AAV vectors offer numerous advantages over other vector systems, as they are easily prepared at high titer, they evoke minimal immune reactions, and they can transduce muscles body-wide. However, AAV vectors have a limited cloning capacity and do not work well in cell culture. Oligonucleotides and plasmids work well in cell culture, are not limited by a vector cloning capacity and seem to have the safest potential in avoiding immune reactions. Ad vectors have the greatest cloning capacity and work well in cell culture, but are much more cumbersome to prepare in forms that minimize immune reactions. For basic research studies and for preclinical studies, these various vectors offer a wide choice in transferring genes to myogenic cells. Nonetheless, further work is needed to adapt any of these vector systems toward a therapy for human neuromuscular disorders, and other vector systems as yet poorly studied may someday prove to be better than the current vectors. Currently, plasmids, oligonucleotides, and rAAV vectors expressing micro-dystrophins are in various stages of being tested in human clinical trials for DMD. It is hoped that at least one of these systems will have an important clinical impact on this and other

muscle disorders (*see* Article 100, **Gene transfer vectors as medicinal products: risks and benefits**, Volume 2).

Acknowledgments

This work was supported by grant AG015434 from the National Institutes of Health.

References

Barjot C, Hartigan-O'Connor D, Salvatori G, Scott JM and Chamberlain JS (2002) Gutted adenoviral vector growth using E1/E2b/E3-deleted helper viruses. *Journal of Gene Medicine*, **4**, 480–489.

Bertoni C and Rando TA (2002) Dystrophin Gene Repair in mdx Muscle Precursor Cells In Vitro and In Vivo Mediated by RNA-DNA Chimeric Oligonucleotides. *Human Gene Therapy*, **13**, 707–718.

Blankinship M, Gregorevic P, Allen JM, Harper SQ, Harper HA, Halbert CL, Miller AD and Chamberlain JS (2004) Vectors based on adeno-associated virus serotype 6 efficiently transduce skeletal muscle. *Molecular Therapy*, **10**, 671–678.

Chen HH, Mack LM, Kelly R, Ontell M, Kochanek S and Clemens PR (1997) Persistence in muscle of an adenoviral vector that lacks all viral genes. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 1645–1650.

DelloRusso C, Scott J, Hartigan-O'Connor D, Salvatori G, Barjot C, Robinson AS, Crawford RW, Brooks SV and Chamberlain JS (2002) Functional correction of adult mdx mouse muscle using gutted adenoviral vectors expressing full-length dystrophin. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12979–12984.

Ferrer A, Wells K and Wells D (2000) Immune responses to dystrophin: implications for gene therapy of Duchenne muscular dystrophy. *Gene Therapy*, **7**, 1439–1446.

Gao G, Vandenberghe LH, Alvira MR, Lu Y, Calcedo R, Zhou X and Wilson JM (2004) Clades of Adeno-associated viruses are widely disseminated in human tissues. *Journal of Virology*, **78**, 6381–6388.

Goyenvalle A, Vulin A, Fougerousse F, Leturcq F, Kaplan JC, Garcia L and Danos O (2004) Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping. *Science*, **306**, 1796–1799.

Greelish JP, Su LT, Lankford EB, Burkman JM, Chen H, Konig SK, Mercier IM, Desjardins PR, Mitchell MA, Zheng XG, *et al*. (1999) Stable restoration of the sarcoglycan complex in dystrophic muscle perfused with histamine and a recombinant adeno-associated viral vector. *Nature Medicine*, **5**, 439–443.

Gregorevic P, Blankinship M, Allen J, Crawford RW, Meuse L, Miller DG, Russell DW and Chamberlain JS (2004a) Systemic delivery of genes to striated muscles using adeno-associated viral vectors. *Nature Medicine*, **10**, 828–835.

Gregorevic P, Blankinship MJ and Chamberlain JS (2004b) Viral vectors for gene transfer to striated muscle. *Current Opinion in Molecular Therapeutics*, **6**, 491–498.

Harper SQ, Hauser MA, DelloRusso C, Duan D, Crawford RW, Phelps SF, Harper HA, Robinson AS, Engelhardt JF, Brooks SV, *et al*. (2002) Modular flexibility of dystrophin: Implications for gene therapy of Duchenne muscular dystrophy. *Nature Medicine*, **8**, 253–261.

Lu QL, Bou-Gharios G and Partridge TA (2003a) Non-viral gene delivery in skeletal muscle: a protein factory. *Gene Therapy*, **10**, 131–142.

Lu QL, Mann CJ, Lou F, Bou-Gharios G, Morris GE, Xue SA, Fletcher S, Partridge TA and Wilton SD (2003b) Functional amounts of dystrophin produced by skipping the mutated exon in the mdx dystrophic mouse. *Nature Medicine*, **9**, 1009–1014.

Malik AK, Monahan PE, Allen DL, Chen BG, Samulski RJ and Kurachi K (2000) Kinetics of recombinant adeno-associated virus-mediated gene transfer. *Journal of Virology*, **74**, 3555–3565.

Manno CS, Chew AJ, Hutchison S, Larson PJ, Herzog RW, Arruda VR, Tai SJ, Ragni MV, Thompson A, Ozelo M, *et al*. (2003) AAV-mediated factor IX gene transfer to skeletal muscle in patients with severe hemophilia B. *Blood*, **101**, 2963–2972.

Nalbantoglu J, Larochelle N, Wolf E, Karpati G, Lochmuller H and Holland PC (2001) Muscle-specific overexpression of the adenovirus primary receptor CAR overcomes low efficiency of gene transfer to mature skeletal muscle. *Journal of Virology*, **75**, 4276–4282.

Parks RJ, Chen L, Anton M, Sankar U, Rudnicki MA and Graham FL (1996) A helper-dependent adenovirus vector system: removal of helper virus by Cre-mediated excision of the viral packaging signal. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13565–13570.

Romero NB, Benveniste O, Payan C, Braun S, Squiban P, Herson S and Fardeau M (2002) Current protocol of a research phase I clinical trial of full-length dystrophin plasmid DNA in Duchenne/Becker muscular dystrophies. Part II: clinical protocol. *Neuromuscular Disorders*, **12**(Suppl 1), S45–S48.

van Deutekom JC, Bremmer-Bout M, Janson AA, Ginjaar IB, Baas F, den Dunnen JT and van Ommen GJ (2001) Antisense-induced exon skipping restores dystrophin expression in DMD patient derived muscle cells. *Human Molecular Genetics*, **10**, 1547–1554.

Wang B, Li J and Xiao X (2000) Adeno-associated virus vector carrying human minidystrophin genes effectively ameliorates muscular dystrophy in mdx mouse model. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 13714–13719.

Xiao X, Li J and Samulski RJ (1996) Efficient long-term gene transfer into muscle tissue of immunocompetent mice by adeno-associated virus vector. *Journal of Virology*, **70**, 8098–8108.

Yuasa K, Sakamoto M, Miyagoe-Suzuki Y, Tanouchi A, Yamamoto H, Li J, Chamberlain JS, Xiao X and Takeda S (2002) Adeno-associated virus vector-mediated gene transfer into dystrophin-deficient skeletal muscles evokes enhanced immune response against the transgene product. *Gene Therapy*, **9**, 1576–1588.

Zhang G, Budker VG, Ludtke JJ and Wolff JA (2004a) Naked DNA gene transfer in mammalian cells. *Methods in Molecular Biology*, **245**, 251–264.

Zhang G, Ludtke JJ, Thioudellet C, Kleinpeter P, Antoniou M, Herweijer H, Braun S and Wolff JA (2004b) Intraarterial delivery of naked plasmid DNA expressing full-length mouse dystrophin in the mdx mouse model of duchenne muscular dystrophy. *Human Gene Therapy*, **15**, 770–782.

Basic Techniques and Approaches

Gene transfer to the liver

Katherine Parker Ponder
University of Washington School of Medicine, St. Louis, MO, USA

The liver is an important target for gene therapy. First, many genetic diseases could be corrected with liver-directed gene therapy. These include deficiencies in metabolic enzymes such as phenylketonuria, deficiencies in blood proteins such as hemophilia, lysosomal storage diseases, infectious diseases such as hepatitis B or C, or liver tumors. Second, hepatocytes have direct contact with blood, allowing the vector to reach most cells with a simple portal vein, hepatic artery, or intravenous injection, which is the mode of delivery currently used in most studies. Third, hepatocytes are long-lived, and are what replicate in response to normal growth or liver injury. This makes it possible to obtain long-term correction after the transfer into fetal or newborn animals if the vector integrates into the chromosome.

Substantial success has been achieved in the liver using retroviral, adenovirus-associated virus (AAV), adenoviral, and plasmid vectors for gene therapy (Kren *et al*., 2002; Thomas *et al*., 2003; VandenDriessche *et al*., 2003). The best long-term expression has been achieved using vectors that utilize a promoter that is normally expressed in the liver, such as the human α1-antitrypsin promoter. In contrast, most viral promoters such as the cytomegalovirus (CMV) promoter have attenuated substantially over time. Gamma-retroviral vectors (Kren *et al*., 2002; Thomas *et al*., 2003; VandenDriessche *et al*., 2003), which only transduce dividing cells, have transferred genes into small and large animals in the newborn period, when hepatocytes are replicating because of the rapid growth. This approach has reduced clinical manifestations of hemophilia A and B, and mucopolysaccharidosis VII in mouse and canine models. This neonatal approach has additional advantages, as it usually induces tolerance to otherwise immunogenic proteins, and corrects the genetic defect earlier. In adults, delivery of gamma-retroviral vectors has required induction of hepatocyte replication with a hepatic growth factor or liver damage. Although effective in mice with hemophilia or thrombophilia, this approach has not yet been applied in large-animal models. Lenti-retroviral vectors can transduce nondividing cells, and have delivered genes to adult rodents without induction of hepatocyte replication. Lenti-retroviral vectors have reduced clinical manifestations in mice with hemophilia, but have not yet been applied in large animals. A major concern of retroviral vectors is the risk of insertional mutagenesis, in which adjacent genes might be activated by enhancer elements of the vector. Indeed, insertional mutagenesis contributed to the leukemias that developed after bone marrow–directed gene therapy for severe combined immunodeficiency, although

other factors almost certainly contributed. To date, no tumors have developed in mice or dogs after liver-directed gene therapy with retroviral vectors. Thus, retroviral vectors show substantial promise for gene therapy, although further safety evaluation needs to be performed.

AAV vectors can transduce nonreplicating hepatocytes of small and large animals in adults. They are primarily maintained in an episomal form although a small fraction of the sequences integrate into the chromosome (Grimm and Kay, 2003). AAV vectors have resulted in amelioration of a number of genetic diseases in mice, and have reduced bleeding in dogs with hemophilia B. One problem with AAV vectors is their small capacity of only ~5 kb of total sequence. This has made it difficult to achieve high levels of expression of large genes such as the Factor VIII gene that is deficient in hemophilia A. AAV vectors have also transduced hepatocytes of newborn mice with mucopolysaccharidosis VII. The substantial fall in expression that occurred during normal growth was likely due to the lack of integration of most copies. There is currently a clinical trial using an AAV2 vector to treat humans with hemophilia B. Although patients who received the highest dose had expression for about a month, this fell to undetectable levels thereafter, which was likely due to cytotoxic T lymphocyte responses against cells that contain Factor IX or the AAV capsid proteins. If the immune response was against the AAV2 capsid proteins, this might be avoided by generating vectors with capsid proteins from other serotypes, some of which do not normally infect humans but are efficient at transferring genes into the livers of animals. Thus, AAV vectors show substantial promise for gene therapy, although existing immunity to AAV2 capsid proteins may be problematic, and they have some risk of insertional mutagenesis.

Adenoviral vectors have a large capacity and can transduce nondividing hepatocytes extremely efficiently (Kren *et al*., 2002; Thomas *et al*., 2003; VandenDriessche *et al*., 2003). They have resulted in long-term expression of several therapeutic proteins in rodents. However, stable expression has not been achieved in most large-animal studies, which may be due to the failure to integrate. An additional problem with adenoviral vectors is the induction of inflammatory responses, which was likely responsible for the death of a patient with a urea cycle disorder. Development of adenoviral vectors that do not contain any viral genes has not eliminated this problem, although reduction of the dose with a strong liver-specific promoter has been effective. The role of adenoviral vectors in gene therapy for nonmalignant disorders is uncertain, as the instability and the inflammatory responses evoked are major hurdles.

Plasmid vectors have transferred genes into the liver and achieved therapeutic levels of expression (Kren *et al*., 2002). Stable expression has been achieved in rodents by cotransfer of a transposase that results in the integration of the therapeutic gene into the chromosome of hepatocytes, or by using plasmids that are maintained long-term extrachromosomally. The major problem with plasmid-based systems is that it has been difficult to achieve high-level transfer into hepatocytes. The most efficient method is to use a high-pressure injection, which has substantial toxicity. Use of simple plasmid vectors for gene therapy will require identification of an efficient and nontoxic method for getting them into hepatocytes. Alternatively, replication-deficient SV40 vectors have transferred circular DNA molecules into

the liver and maintained expression long-term, although these vectors still contain some viral genes that might provoke cytotoxic T lymphocyte responses against transduced cells in some species. Thus, plasmid-based vectors show promise for gene therapy, although delivery to the liver is a problem, and vectors that result in random integration will also have a risk of insertional mutagenesis.

References

Grimm D and Kay MA (2003) From virus evolution to vector revolution: Use of naturally occurring serotypes of adeno-associated virus (AAV) as novel vectors for human gene therapy. *Current Gene Therapy*, **3**, 281–304.

Kren BT, Chowdhury NR, Chowdhury JR and Steer CJ (2002) Gene therapy as an alternative to liver transplantation. *Liver Transplantation*, **8**, 1089–1108.

Thomas CE, Ehrhardt A and Kay MA (2003) Progress and problems with the use of viral vectors for gene therapy. *Nature Reviews. Genetics*, **4**, 346–358.

VandenDriessche T, Collen D and Chuah MK (2003) Gene therapy for the hemophilias. *Journal of Thrombosis and Haemostasis: JTH*, **1**, 1550–1558.

Gene transfer to the skin

Laurent Gagnoux-Palacios, Flavia Spirito and Guerrino Meneguzzi
University of Nice-Sophia Antipolis, Nice, France

1. Introduction

The skin is the physical barrier that preserves the body from environmental aggressions. The outer compartment of this organ, the epidermis, is a continuously renewing tissue produced by a proliferative basal layer of keratinocytes generated by transiently amplifying cells, which derive from epithelial stem cells. Differentiating basal keratinocytes migrate into the suprabasal cell layer and enter a stepwise process of keratinization and lipogenesis that defines the distinct layers of the epidermis: the spinous, granular, and corneum stratum. The stratum corneum is made of dead, terminally differentiated keratinocytes, and assumes defense functions including barrier to water permeability, penetration of xenobiotics, and parasitic infection. The epidermis is anchored to the dermis, the connective tissue component of the skin, which in addition to the large populations of fibroblasts that produce and organize the extracellular matrix, contains epithelial, vascular, neural, and haematopoietic cells. In association with the antigen-presenting cells (Langerhans and T-cells lymphocytes) of the epidermis, the macrophages, neutrophils, and lymphocytes of the dermis provide high levels of local immune surveillance. Thus, gene transfer to the skin is faced to an outer barrier, the stratum corneum, which protects the target living cells of the inner epidermis and dermis.

The skin is an optimal target organ for gene transfer because of its accessibility to manipulations and the ease with which the epidermal and dermal cells are biopsied, expanded in culture, genetically manipulated, and grafted back to the patients, following well-established procedures. Besides applications in the correction of genodermatosis, genetic manipulations of the skin have potential use in vaccination against cancer and infectious diseases. Engineered keratinocytes are also potential bioreactors for the production of locally and systemically active polypeptides with a therapeutic interest in the treatment of hormonal and metabolic disorders affecting other organs. An additional advantage over the other organs is that any adverse effect of genetic manipulations on the skin is easily monitored and the involved tissue readily removed.

A range of techniques for the introduction of recombinant DNA into the skin has been developed for *in vivo* and *ex vivo* gene transfer. In the *in vivo* approach, the DNA is directly administered to the skin, while *ex vivo* gene transfer requires

introduction of the desired gene into cells harvested from skin biopsies and generation of self-renewing tissue transplantable back to the donor. Technical achievements in gene transfer combined with the sophisticated methods of skin grafting make the *ex vivo* approach most attractive, despite its limited application to the few cell types for which culture conditions and transplantation techniques are well defined and the fact that the viability of the grafted cells, the need for multiple surgical procedures, and the risk of scarring may hamper the clinical applicability. Alternative *in vivo* systems require much less biotechnological expertise and effort. However, the efficiency of *in vivo* gene transfer is generally low and associated with a transient expression of the therapeutic gene. Most of the *ex vivo* and *in vivo* strategies rely on the use of either viral or nonviral vectors designed by specific therapeutic needs: transient transgene expression is adapted in tissue repair, vaccination, or anticancer treatments, while permanent gene expression, possibly regulated by inducible promoters, is required for the permanent correction of inherited and acquired conditions.

1.1. *Viral vectors*

Recombinant viruses have been the most successful methods for gene transfer and clinical applications because the manipulation of the viral genomes is relatively easy and controllable. Various viral systems have been adapted to develop high efficiency gene transfer to the skin. These include replication-defective retroviruses (*see* Article 98, **Retro/lentiviral vectors**, Volume 2) and adenovirus (*see* Article 96, **Adenovirus vectors**, Volume 2) and recombinant adeno-associated viruses (AAV) (*see* Article 97, **Adeno-associated viral vectors: depend(o)ble stability**, Volume 2).

Retroviral vectors based on murine leukaemia virus (MLV) are currently used to achieve long-term expression of a foreign gene in the skin. The major advantages of these vectors include their capacity of transferring genes at high efficiency, and to permanently integrate them into the host cell chromosomes, which assures a stable and efficient expression. Since integration of MLV-based vectors requires host cell division, these vectors are conveniently used both to transfect cultured keratinocytes and to target proliferative cells in regenerating tissue *in vivo*. Disadvantages of retroviral gene transfer are represented by the limited size of the transferred DNA, the complexity of the methods for virus production, the relatively low titers of the viral suspensions, the limited types of proliferating skin cells that can be efficiently targeted, and the lack of control on integration into the host genome. Permanent *ex vivo* transduction of keratinocyte stem cells with a low number of integration events is achieved by integrating the producer cells into the lethally irradiated feeder for primary keratinocytes (Dellambra *et al*., 1998). The development of retroviral MSC vectors has allowed highly efficient *ex vivo* transduction of epithelial stem cells by direct transduction with purified virus. The recent construction of retroviral vectors based on lentiviruses, which yield high titers of infectious virions and integrate into the genome of nonproliferating cells, has overcome the major constraints in MLV retroviral gene transfer to the skin. Lentiviral vectors have been successfully used to efficiently transfer relatively

large cDNAs in cultured keratinocyte with sustained and stable expression of the transgene after transplantation of the engineered cells *in vivo* (Chen *et al*., 2002). Long-term transduction of mouse keratinocytes by a retroviral vector has also been demonstrated *in vivo* after mechanical abrasion to remove the interfollicular epithelium and induce re-epithelialization from proliferating stem cells located in the hair follicles, and by intradermal injections of viral particles (Hengge *et al*., 1995; Ghazizadeh *et al*., 1999). However, administration of the virus by injections between the scab and the regenerating epithelium hampers the quantification of the material delivered to the cells and it is technically too complex for clinical applications. Intradermal injection results in transduction of all the cell populations close to the inoculation site, including endothelial and dendritic cells, which causes undesired side effects. The possible activation of proto-oncogenes consequent to integration into the host genome implies a risk of cancerization of the transduced cells that must be considered. Therefore, the choice of the transduction protocols and retroviral vectors should favor procedures that limit the number of the genomic integrations of the therapeutic gene.

Recombinant adenoviral vectors provide transient and highly efficient gene transfer, and because of their inability to integrate into the host genome, they represent a useful alternative to retroviral gene transfer when the permanent expression of the transgene in the skin is not required. Their ability to infect nondividing cells makes adenoviruses particularly adapted for *in vivo* transduction of slowly dividing or differentiated cells such as melanocytes, Langherans cells, and suprabasal keratinocytes. However, the interest of adenoviral vectors for skin gene therapy is limited because the most efficient constructs induce a transient and dose-dependent inflammatory reaction to the viral proteins that is enhanced when repeated treatments are required.

Safety, an elevated physical stability, and lack of immunogenicity are the interesting properties of recombinant AAV vectors. However, AAV vectors transduce dermal fibroblasts inefficiently and transduction of keratinocytes is poorly documented. In our hands, primary keratinocyte cell cultures enriched in stem cells are resistant to the AAV infection, whereas efficient transduction with persistent expression of the transgene was observed with keratinocytes either senescent or immortalized. Exploitability of AAV vectors in skin gene therapy is therefore questionable.

1.2. Nonviral gene transfer

This is based on technologies involving easy production and direct application of the genetic material *in vivo* and *in vitro* (*see* Article 95, **Artificial self-assembling systems for gene therapy**, Volume 2). Most of nonviral vectors are synthetic means of encapsulating transgenic DNA until it reaches the cellular target. These vectors are safe to prepare, suitable for large-scale manufacturing procedures, and the risk of pathological and immunological complications is negligible. This approach, however, suffers from inefficient transfer and transitory expression of the transgene that does not integrate into the host genome. For *in vivo* delivery, the therapeutic DNA must overcome the epidermal barrier and reach the underlying target cells. To facilitate penetration into the skin, various techniques (prolonged occlusion,

electrically assisted delivery by iontophoresis or electroporation, sonography) based on the disruption or alteration of the stratum corneum have been devised but they remain unapplicable for clinical use. Lipofection has been used *in vivo* and in clinical trials, but the advantage of liposomes in terms of no constraint on the size of the transgene, the nontoxicity, and the possibility of multiple applications is counterbalanced by the low frequency of stable transfection, the stability, and the cost of the reagents. The delivery of the transgene is efficient to hair follicles, which may constitute the ideal target for this application. Direct injection of DNA has low potential for biohazardous risk but remains inefficient. Introduction of naked plasmid DNA into the superficial dermis by single inoculation leads to transient transgene expression concentrated in the layers of the epidermis overlying the injection site and low expression in the injected connective tissue. Efficiency can be enhanced by multiple microinoculations (microseeding) into the skin or by particle bombardment (ballistic gene transfer). The ballistic microprojectiles are made of small gold particles coated with DNA that are projected at high speed by a gene gun. This technique has been successfully used in various animal models, but the transgene expression remains transient (5 to 30 days) and efficiency is elevated only when the basal keratinocytes of the epidermis are targeted.

Because keratinocytes and fibroblasts are easily expanded in tissue culture, relatively efficient gene transfer can be achieved *ex vivo* using systems that involve insertion of the desired gene into nonviral episomal expression vectors containing specific regulatory and selection elements, and following the procedures currently used to transfect the mammalian cells. These include DNA-mediated transfection by methods such as calcium phosphate precipitation, DEAE-dextran transfection, polybrene-dimethyl sulfoxide shock, direct microinjection, jet injection or electroporation of pure plasmid DNA or DNA complexed with surface-receptors ligands, cationic liposomes, or ballistic microprojectiles. Cell toxicity or induction of keratinocyte differentiation may constitute a hindering side effect in a number of these methods. All these nonviral transfer systems permit only transient gene expression, because efficient host integration does not occur, except in transposon-based plasmids. For instance, the phage ϕC31 integrase mediates integration in the human cell DNA at endogenous attachment sites of extrachromosomal vectors carrying a therapeutic cDNA (Ortiz-Urda *et al*., 2002). However, the integration frequency is low, just above the background of random integration, so that observation of the curative effect requires the enrichment of the recipient keratinocytes using a selectable gene. Thus, the ideal gene transfer system that reliably permits high efficiency, safe, and prolonged expression of a therapeutic gene is still missing. At the current state, the nonviral technology is unsuitable for *in vivo* gene therapy.

Direct DNA injection into the skin yields a protective immune response. The nature of the genetic immunization depends on the technique used (injection of naked DNA or bombardment of DNA-coated microprojectiles), the amount of DNA, and the site of immunization. DNA vaccines remain an attractive mode of treatment and investigation in gene expression technology and in enhancing various aspects of adjuvant effects. The immune system of the skin is stimulated even by low expression of exogenous gene products directly introduced in Langerhans cells and dermal dendritic cells; therefore, expression of an exogenous transgene is expected to elicit an immune response against the genetically modified

cells. Immune responses can also be elicited by antigens generated by the secretion or processing of recombinant polypeptides synthesized by *ex vivo* transduced keratinocytes, with destruction of the genetically manipulated cells. Because any transgene product containing novel epitopes is a potential target of the host immune response, deep investigations are necessary to evaluate the host tolerance to each gene product delivered to the skin and assess the potential setbacks in setting up human clinical trials (Ghazizadeh *et al.*, 2003).

Further reading

Hengge UR and Volc-Platzer B (Eds.) (2000) *The Skin and Gene Therapy*, Springer-Verlag: Berlin.

References

Chen M, Kasahara N, Keene DR, Chan L, Hoeffler WK, Finlay D, Barcova M, Cannon PM, Mazurek C and Woodley DT (2002) Restoration of type VII collagen expression and function in dystrophic epidermolysis bullosa. *Nature Genetics*, **32**, 670–675.

Dellambra E, Vailly J, Pellegrini G, Bondanza S, Golisano O, Macchia C, Zambruno G, Meneguzzi G and De Luca M (1998) Corrective transduction of human epidermal stem cells in laminin-5-dependent junctional epidermolysis bullosa. *Human Gene Therapy*, **9**, 1359–1370.

Ghazizadeh S, Harrington R and Taichman L (1999) In vivo transduction of mouse epidermis with recombinant retroviral vectors: implications for cutaneous gene therapy. *Gene Therapy*, **6**, 1267–1275.

Ghazizadeh S, Kalish RS and Taichman LB (2003) Immune-mediated loss of transgene expression in cutaneous epithelium: Implications for cutaneous gene theraphy. *Molecular Therapy*, **7**, 296–303.

Hengge UR, Chan EF, Foster RA, Walker PS and Vogel JC (1995) Cytokine gene expression in epidermis with biological effects following injection of naked DNA. *Nature Genetics*, **10**, 161–166.

Ortiz-Urda S, Thyagarajan B, Keene DR, Lin Q, Fang M, Calos MP and Khavari PA (2002) Stable nonviral genetic correction of inherited human skin disease. *Nature Medicine*, **8**, 1166–1170.

Basic Techniques and Approaches

Control of transgene expression in mammalian cells

Beat P. Kramer and Martin Fussenegger
Institute of Chemical and Bio-Engineering, Zurich, Switzerland

1. Introduction

Since most genetic disorders result from deregulated transcription or mutated genes, current gene therapy initiatives focus on complementation of genetic defects using adjustable expression of therapeutic transgenes to ensure precise titration into the therapeutic window, coordination with patient-specific disease dynamics, and termination of therapeutic interventions. Ideal conditional therapeutic transcription interventions require complex systems that (1) are of heterologous origin to ensure an interference-free operation, (2) enable seamless integration into regulatory networks of target cells, (3) exhibit low immunogenicity, (4) provide high-level expression under induced and low basal expression under repressed conditions, (5) show adjustability to intermediate levels over a wide range of inducer concentrations, (6) are responsive to a bioavailable inducer including clinically licensed small-molecule drugs, other inert molecules, or any specific well-tolerated physical condition, (7) exhibit high compatibility with approved viral and nonviral gene transfer technologies, (8) support configurations to restrict interventions to specific tissues or disease foci, and (9) are amenable to compact genetic design to limit pleiotropic effects associated with repeated molecular intervention on the patient's chromosome.

Generic design concepts of most advanced transcription control systems include artificial transcription-modulating (fusion) proteins that bind or assemble at specific target promoters in a small molecule-dependent manner or at specific physical conditions.

2. Antibiotic-responsive gene regulation systems

Antibiotic-responsive gene regulation systems are derived from prokaryotic antibiotic response regulators evolved to coordinate resistance to specific classes of antibiotics. A protein represses a resistance gene until its release from the target promoter following binding of a specific antibiotic. Such antibiotic-responsive protein–DNA interactions have successfully been assembled in three different configurations for conditional transgene transcription in mammalian cells: (1) Fusion of

the bacterial antibiotic resistance gene repressor to a generic transactivation domain creates an antibiotic-dependent transactivator, which, in the absence of regulating antibiotics, binds and activates chimeric promoters containing transactivator-specific (tandem) operator modules 5′ of a minimal eukaryotic promoter (Gossen and Bujard, 1992; Fussenegger *et al*., 2000a; Weber *et al*., 2002). (2) The aforementioned transactivator can be mutated to enable reverse antibiotic-dependent binding, resulting in dose-dependent transgene induction in the presence of regulating antibiotics (Gossen *et al*., 1995). (3) The bacterial antibiotic resistance gene repressor bound to its cognate operator may repress transcription from 5′-located mammalian promoters. The antibiotic-induced repressor release then correlates with increased transgene transcription (Fussenegger *et al*., 2000a; Yao *et al*., 1998; Weber *et al*., 2002). Three different transcription control systems responsive to tetracyclines (Gossen and Bujard, 1992), streptogramins (Fussenegger *et al*., 2000b), and macrolides (Weber *et al*., 2002) have been developed. Antibiotic-responsive gene regulation systems comply with ideal systems at a high standard. Yet, potential immunogenicity of bacterial components, tissue-specific accumulation, and/or promotion of antibiotic resistance will remain ongoing challenges for clinical implementation (Darteil *et al*., 2002).

3. Hormone-inducible gene expression

Lipophilic hormones are key players of intercellular communication in higher eukaryotes. They freely cross cell membranes, bind intracellular receptors, and modulate target gene expression following nuclear translocation. The generic design concept for hormone-inducible gene regulation systems consists of fusing the hormone-binding domain of a hormone receptor to a heterologous DNA-binding module and optionally to a transactivation/transsilencing moiety. This chimeric hormone receptor will initiate/repress transcription from target promoters equipped with DNA-binding domain-specific operator modules in the presence of regulating hormones. The use of human hormone receptor mutants specific for endogenous hormone agonists is expected to enable increased immunocompatibility without interfering with endogenous hormone regulatory networks. Yet, some hormone agonists exhibit a major clinical impact. Three hormone-inducible transcription control systems responsive to estrogen (Braselmann *et al*., 1993), the progesterone antagonist mifepristone (Wang *et al*., 1994), and the insect moulting hormone ecdysone (No *et al*., 1996) have been constructed, and are continuously improved for human compatibility.

4. Dimerizer-regulated gene expression

Chemically induced dimerization (CID) is a phenomenon by which two proteins (hetero)dimerize in the presence of a molecule. The most prominent heterodimerizer is the immunosuppressive agent rapamycin, which unites FKBP (FK506-binding protein) and FRAP (FKBP rapamycin-associated protein), thus impairing cell cycle regulatory networks involved in T cell expansion. Rapamycin-inducible

heterodimerization of FKBP fused to artificial DNA-binding domains and FRAP to a transactivation domain reconstitutes a chimeric transactivator that initiates transcription from target promoters containing specific operator modules. In order to alleviate cosuppression of the immune system when rapamycin-based CID transcription control is in action, protein engineering initiatives to alter FKBP and FRAB's specificities for a nontoxic, clinically inert molecule are promising and may secure a clinical future of this technology (Pollock and Clackson, 2002).

5. Systems of the future

Despite quantitative differences in regulation performance, most aforementioned transcription control systems qualify for clinical implementation. However, improvement of immunocompatibility and tolerability of inducers remain future challenges. Current systems based on prokaryotic antibiotic resistance operons promise specific regulation by clinically licensed molecules, but are compromised by the use of bacterial epitopes and long-term accumulation of antibiotics in various tissues of the body. Hormone- and CID-responsive transgene control systems can be optimally humanized, yet remain limited by pleiotropic and other side effects of their inducing agents. Even when using artificial molecule-protein interactions, immunocompatibility and side effects of designer molecules remain imminent concerns. On the way toward ideal transgene regulation modalities, a variety of strategies have been designed, which promise important improvements over existing technologies: (1) construction of transcription modulators responsive to clinically inert compounds, temperature, light, and dynamic electromagnetic fields, (2) systems responsive to specific nucleic acids, and (3) epigenetic gene networks.

Prokaryotes manage inter- and intrapopulation communication by quorum-sensing molecules, which bind to receptors in target cells and initiate specific regulon switches by modulating the receptors affinity to cognate promoters (Bassler, 2002). Systems derived from bacterial, quorum-sensing regulatory networks are expected to be of particular interest since many regulating molecules of commensal prokaryotic populations have a long history of human-prokaryotic coevolution. Following the generic design principle of antibiotic-adjustable transcription control modalities, bacterial cross talk systems responsive to butyrolactones have been successfully validated in mice without signs suggestive of inducer-related side effects (Weber *et al*., 2003b; Neddermann *et al*., 2003).

Throughout the development of eukaryotes, production of ribosomal proteins (rp) is regulated at the translational level. Translation control is mediated by a terminal oligopyrimidine element (TOP) present in the 5′ untranslated region of rp-encoding proteins. TOP elements adopt a translation-prohibitive secondary structure, which is resolved upon binding of a specific cellular nucleic acid–binding protein (CNBP). TOP-complementary nucleic acids interfere with the translation-promoting TOP–CNBP interaction and so repress top-tagged mRNAs in a dose-dependent manner (Schlatter and Fussenegger, 2003).

Two low-temperature-inducible mammalian gene regulation systems have been designed capitalizing on (1) a heat-labile alphaviral replicase that transcribes target genes driven by subgenomic promoters only at permissive temperatures (Boorsma

et al., 2000) and (2) a thermosensor managing heat-shock response in *Streptomyces albus* (Weber *et al*., 2003a). Owing to dominant environmental conditions, clinical implementation of temperature-controlled expression technology would require local temperature control by a Pelletier element.

Light-inducible gene regulation may represent an alternative to temperature-based control. A recently discovered photosensitive plant protein heterodimerizes with a partner protein following exposure to the cofactor phynocynanobilin and transient red and far-red light pulses. Assembly of these heterodimerizing proteins analogous to CID systems enabled light-inducible transgene expression in yeast (Shimizu-Sato *et al*., 2002). Owing to limited tissue penetration, nonvisible electromagnetic fields have come into the limelight of the gene control community, but these systems are at best in the discovery stage: (1) several electromagnetic field response elements (EMRE) have been discovered in humans (Lin *et al*., 2001; Rubenstrunk *et al*., 2003) and (2) radio-frequency magnetic fields were shown to remotely control hybridization of oligonucleotides linked to gold nanoparticles (Hamad-Schifferli *et al*., 2002).

Throughout multicellular systems, cell identity is maintained by epigenetic regulation circuits, which imprint transient morphogen gradients during early embryogenesis by locking the transcriptome of adult cells in a cell phenotype-specific manner. By combining two repressors, which control each other's expression, an epigenetic circuitry able to switch between two stable expression states by transient addition of two inducers has been pioneered (Kramer *et al*., 2004). Owing to transient administration of regulating agents, long-term side effects will be eliminated.

6. Conclusions

Much like drug dosing is the key parameter in modern molecular medicine since Paracelsus' statement that "the dose makes the poison", gene expression dosing will be of central importance for next-generation gene therapy and tissue engineering initiatives. Capitalizing on achievements accumulated for over a decade, conditional transcription control of therapeutic transgenes stands now on the threshold to a clinical reality. With the first transcription control units being assembled into regulatory gene networks and prototype epigenetic gene switches being designed for mammalian cells, the future for multigene-based therapeutic interventions in regulatory networks of patients' cells has just begun.

References

Bassler BL (2002) Small talk. Cell-to-cell communication in bacteria. *Cell*, **109**, 421–424.

Boorsma M, Nieba L, Koller D, Bachmann MF, Bailey JE and Renner WA (2000) A temperature-regulated replicon-based DNA expression system. *Nature Biotechnology*, **18**, 429–432.

Braselmann S, Graninger P and Busslinger M (1993) A selective transcriptional induction system for mammalian cells based on Gal4-estrogen receptor fusion proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 1657–1661.

Darteil R, Wang M, Latta-Mahieu M, Caron A, Mahfoudi A, Staels B and Thuillier V (2002) Efficient gene regulation by PPAR gamma and thiazolidinediones in skeletal muscle and heart. *Molecular Therapy*, **6**, 265–271.

Fussenegger M, Morris RP, Fux C, Rimann M, von Stockar B, Thompson CJ and Bailey JE (2000a) Streptogramin-based gene regulation systems for mammalian cells. *Nature Biotechnology*, **18**, 1203–1208.

Fussenegger M, Morris RP, Fux C, Rimann M, von Stockar B, Thompson CJ and Bailey JE (2000b) Streptogramin-based gene regulation systems for mammalian cells. *Nature Biotechnology*, **18**, 1203–1208.

Gossen M and Bujard H (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 5547–5551.

Gossen M, Freundlieb S, Bender G, Muller G, Hillen W and Bujard H (1995) Transcriptional activation by tetracyclines in mammalian cells. *Science*, **268**, 1766–1769.

Hamad-Schifferli K, Schwartz JJ, Santos AT, Zhang S and Jacobson JM (2002) Remote electronic control of DNA hybridization through inductive coupling to an attached metal nanocrystal antenna. *Nature*, **415**, 152–155.

Kramer BP, Usseglio Viretta A, Daoud-El Baba M, Aubel D, Weber W and Fussenegger M (2004) An engineered epigenetic transgene switch in mammalian cells. *Nature Biotechnology*, **22**, 867–870.

Lin H, Blank M, Rossol-Haseroth K and Goodman R (2001) Regulating genes with electromagnetic response elements. *Journal of Cellular Biochemistry*, **81**, 143–148.

Neddermann P, Gargioli C, Muraglia E, Sambucini S, Bonelli F, De Francesco R and Cortese R (2003) A novel, inducible, eukaryotic gene expression system based on the quorum-sensing transcription factor TraR. *EMBO Reports*, **4**, 159–165.

No D, Yao TP and Evans RM (1996) Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 3346–3351.

Pollock R and Clackson T (2002) Dimerizer-regulated gene expression. *Current Opinion in Biotechnology*, **13**, 459–467.

Rubenstrunk A, Orsini C, Mahfoudi A and Scherman D (2003) Transcriptional activation of the metallothionein I gene by electric pulses in vivo: basis for the development of a new gene switch system. *Journal of Gene Medicine*, **5**, 773–783.

Schlatter S and Fussenegger M (2003) Novel CNBP- and La-based translation control systems for mammalian cells. *Biotechnology and Bioengineering*, **81**, 1–12.

Shimizu-Sato S, Huq E, Tepperman JM and Quail PH (2002) A light-switchable gene promoter system. *Nature Biotechnology*, **20**, 1041–1044.

Wang Y, O'Malley BW Jr, Tsai SY and O'Malley BW (1994) A regulatory system for use in gene transfer. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 8180–8184.

Weber W, Fux C, Daoud-El Baba M, Keller B, Weber CC, Kramer BP, Heinzen C, Aubel D, Bailey JE and Fussenegger M (2002) Macrolide-based transgene control in mammalian cells and mice. *Nature Biotechnology*, **20**, 901–907.

Weber W, Marty RR, Link N, Ehrbar M, Keller B, Weber CC, Zisch AH, Heinzen C, Djonov V and Fussenegger M (2003a) Conditional human VEGF-mediated vascularization in chicken embryos using a novel temperature-inducible gene regulation (TIGR) system. *Nucleic Acids Research*, **31**, E69.

Weber W, Schoenmakers R, Spielmann M, Daoud-El Baba M, Folcher M, Keller B, Weber CC, Link N, van de Wetering P, Heinzen C, *et al*. (2003b) *Streptomyces*-derived quorum-sensing systems engineered for adjustable transgene expression in mammalian cells and mice. *Nucleic Acids Research*, **31**, E71.

Yao F, Svensjo T, Winkler T, Lu M, Eriksson C and Eriksson E (1998) Tetracycline repressor, tetR, rather than the tetR-mammalian cell transcription factor fusion derivatives, regulates inducible gene expression in mammalian cells. *Human Gene Therapy*, **9**, 1939–1950.

Introductory Review

Eukaryotic genomics

Mark D. Adams
Case Western Reserve University, Cleveland, OH, USA

1. Introduction

In the early 1990s, a plethora of strategies for genome sequencing were proposed as part of the initial phases of the Human Genome Project (HGP). Each strategy relied to varying extents on three types of maps: (1) genetic and radiation hybrid maps consist of sequence tagged site (STS) markers of known order throughout the genome that can be used as landmarks (*see* Article 14, **The construction and use of radiation hybrid maps in genomic research**, Volume 3 and Article 15, **Linkage mapping**, Volume 3), (2) physical maps are composed of overlapping cloned regions of the genome that are tied to the landmark maps and can be used as the physical source of DNA for sequencing a segment of the genome (*see* Article 9, **Genome mapping overview**, Volume 3 and Article 18, **Fingerprint mapping**, Volume 3), and (3) the sequence map, which is the genome sequence itself. So many factors underlie these interconnected maps that the initial plan for the HGP chose to defer decision making on many aspects of the latter two maps until they could be developed together as technology for collecting DNA sequence improved and lessons had been learned from model organism sequencing projects.

By the end of the 1990s, genome sequences were in hand for *Saccharomyces cerevisiae* (Goffeau *et al*., 1996; Goffeau *et al*., 1997), *Caenorhabditis elegans* (*C. elegans* Consortium, 1998), and several bacteria and archaea. Sequencing technology had improved dramatically and a few large laboratories were running scores of automated DNA sequencers and producing thousands of high-quality DNA sequence lanes per day. With the introduction of 96-channel capillary sequencers in 1998, the rush was on to capitalize on even greater sequencing capacity to get the human genome sequence done. Two strategies were chosen by Celera Genomics and the International Human Genome Sequencing consortium. These two strategies and several blends between them continue to be used for the sequencing of other eukaryotic genomes. A brief analysis of the “whole-genome shotgun” and “hierarchical shotgun” method will provide an introduction to this section on Genome Sequencing.

2. Sequencing and assembly approaches

2.1. Hierarchical shotgun

Traditional genome-sequencing methods (Blattner, *et al*., 1997, Goffeau, *et al*., 1997, *C. elegans* Consortium, 1998) have relied on making a carefully constructed map of genome subclones, sequencing each subclone, and then reassembling the complete genome by piecing together the subclone sequences. The maps are generally constructed by a combination of marker-driven methods (probing a subclone library with short sequences such as STSs) and fingerprint methods (restriction digest patterns of clones are compared to one another to identify overlapping clones). A set of subclones is then chosen for sequencing on the basis of selecting the smallest number of clones that reliably covers the genome. The advantage of this approach, termed "hierarchical shotgun sequencing", is that there are several opportunities for checking the quality of the map as it progresses (i.e., do the fingerprint and marker order data agree?). A second advantage is that the map itself has value since the clones can be used for follow-up study. The disadvantage is that it is laborious and difficult to automate. It is also highly dependent on the nature of the subclones that are used. Early map-building efforts suffered from clones that were unstable at both ends of the size spectrum: yeast artificial chromosomes (YACs, 0.5–2 Mb) and cosmids (30 kb). Bacterial artificial chromosomes (BACs, ~150 kb), which are carried as single copy in *Escherichia coli*, have proven to stably clone most segments of the human genome, and the hierarchical shotgun strategy relied to a large extent on BACs for sequencing the human genome. Each BAC was then subcloned into small (~2 kb) fragments that were sequenced. Assembly of each BAC was followed by a second assembly step that joined all of the BAC sequences together on the basis of information from the map, thus the term "hierarchical" shotgun assembly. The BAC clones have also served as a distributed platform for finishing the human genome sequence to high quality, with different clones completed and checked at dozens of laboratories throughout the world (IHGSC, 2004).

2.2. Whole-genome shotgun

With the development of the Applied Biosystems 3700 Automated DNA Analyzer, the speed and accuracy with which raw DNA sequence data could be obtained increased dramatically. This forced a shift in thinking away from the map-based approaches toward a whole-genome shotgun strategy that would take maximum advantage of the increased output of raw sequence data. The whole-genome strategy relies on computational algorithms rather than extensive map-building to reassemble the genome sequence from the raw data (Weber and Myers, 1997). In the whole-genome shotgun strategy, the entire genome is sheared into small to medium fragments (~2 kb or ~10 kb); these are sequenced directly. By sequencing both ends of each subcloned fragment, the two sequences are constrained to be adjacent to one another in the genome; these clone end sequences are called mate pairs. A sufficient number of fragments are sequenced to represent the genome 5–10 times. This 5X to 10X coverage means that most DNA bases have been sequenced many

times, but a small fraction is missing because the coverage is random. The whole-genome strategy poses two problems: the shear number of fragments (about 30 million for 5X coverage of the human genome) and the presence of repetitive DNA. The first problem is largely computational – data management, data structures, and assembly algorithms have been developed to effectively organize and handle the quantity of data (Myers, *et al*., 2000, Batzoglou, *et al*., 2002). The second problem is more complicated and has implications for sequencing all eukaryotic genomes.

If the 2.9 billion base-pair sequence of the human genome were composed of a random distribution of the four DNA bases, any given sequence of at least 12 bases (4^{12}) would be highly likely to be unique in the genome. The 500 to 700 bases in a typical sequence fragment from an automated sequencer have more than enough information content to be unique in the human genome. The problem then is not the size of the genome but the presence of highly similar sequences at more than one location in the genome. There are several categories of these repeated sequences ranging from the ~300-bp Alu element (100 000 copies) to duplications of up to several megabases that are frequent around the centromeres of the chromosomes (*see* Article 26, **Segmental duplications and the human genome**, Volume 3). The length and identity of the repetitive elements determines the level of difficulty that they add to the assembly process. Repeats that are longer than the typical sequence read length and more similar than about 98.5% along their entire length are difficult to assign to their correct chromosomal location.

By obtaining mate-pair sequence from clones of several insert sizes, it is often possible to identify unique sequence that jumps across or spans repeats that are shorter than the average clone length. This approach can resolve the most common types of repetitive elements in the human genome. The mate pairs also serve to anchor together adjacent sequence contigs, resulting in long chains of correctly ordered sequence, with the gaps between contigs spanned by subclones. Additional computational techniques have been developed that attempt to improve on assembly in repeat-rich regions, especially relying on detection and classification of repeats, use of error-correction, and use of signature differences to separate repeat copies appropriately. Long tandem arrays of nearly identical repeats at the centromeres and telomeres of chromosomes cannot be sequenced with existing technology and approaches.

3. Prospects for the future

Sequencing of the human genome was a landmark event in the history of science. While a great deal was learned about the structure and content of the genome through an initial evaluation of the sequence, it has become increasingly clear that much more can be learned by comparing the genomes of multiple individuals and comparing the human genome to that of other primates, mammals, and other animals. The genomes of yeast, *C. elegans* (nematode worm), and *Drosophila melanogaster* (fruit fly) were obtained as part of the preparation for sequencing the human genome. Following human, the mouse, rat, and chimpanzee genomes have been completed to a "draft" stage. A "draft" genome sequence generally means that about 95% of the genome is covered in reasonably accurate sequence (less than

Table 1 Eukaryotic genome-sequencing projects

Species	Status	Genome size (Mb)	Sequencing strategy[a]	Publications
Published				
Human	Finished	2900	HS, WGS	(Venter *et al*., 2001; IHGSC, 2004)
Mouse	Draft+Finished	2600	WGS	(IMGSC, 2002)
Rat	Draft	2700	Hybrid	Gibbs, *et al*., 2004
C. elegans	Finished	100	HS	(*C. elegans* Consortium, 1998)
C. briggsae	Draft	105	WGS	(Stein *et al*., 2003)
D. melanogaster	Finished	120	WGS	(Adams *et al*., 2000; Myers *et al*., 2000; Rubin *et al*., 2000; Celniker *et al*., 2002)
A. gambiae	Draft	280	WGS	(Holt *et al*., 2002)
P. falciparum	Finished	23	SCS	(Gardner *et al*., 2002)
S. cerevisiae	Finished	16	HS	(Goffeau *et al*., 1997; Goffeau *et al*., 1996)
S. pombe	Finished	12.5	HS	(Wood *et al*., 2002)
Dog	Light Draft	2700	WGS	(Kirkness *et al*., 2003)
Arabidopsis	Finished	125	HS	(Arabidopsis Initiative, 2000)
Rice	Draft	400	Hybrid	(Yu *et al*., 2002)
Neurospora	Draft	40	WGS	(Galagan *et al*., 2003)
Fugu rubripes	Draft	365	WGS	(Aparicio *et al*., 2002)
C. intestinalis	Draft	117	WGS	(Dehal *et al*., 2002)
Unpublished/Ongoing				
Danio rerio (Zebrafish)	Draft+Finished	1600	Hybrid	**http://www.sanger.ac.uk/Projects/D_rerio/**
Cow	Ongoing	2900	Hybrid	**http://hgsc.bcm.tmc.edu/projects/bovine/**
Honeybee	Ongoing	200	Hybrid	**http://hgsc.bcm.tmc.edu/projects/honeybee/**
D. pseudoobscura	Draft	140	WGS	**http://www.hgsc.bcm.tmc.edu/projects/drosophila/**
Chimpanzee	Draft	2900	WGS	**http://genome.wustl.edu/projects/chimp/**
Macaca mulatta	Ongoing	2900	Hybrid	**http://hgsc.bcm.tmc.edu/projects/rmacaque/**
C. albicans	Draft	16	WGS	**http://www-sequence.stanford.edu/group/candida/**
Fusareum graminearum	Draft	36	WGS	**http://www.broad.mit.edu/annotation/fungi/fusarium/**

Ustilago maydis	Draft	19	WGS	**http://www.broad.mit.edu/annotation/fungi/ustilago_maydis/**
Ciona savignyi	Draft	180	WGS	**http://www.broad.mit.edu/annotation/ciona/index.html**
Aspergillus nidulans	Draft	30	WGS	**http://www.broad.mit.edu/annotation/fungi/aspergillus/**
Aspergillus fumigatus	Draft	35	WGS	**http://www.sanger.ac.uk/Projects/A_fumigatus/**, **http://www.tigr.org/tdb/e2k1/afu1/**
Magnaporthe grisea	Draft	39	WGS	**http://www.broad.mit.edu/annotation/fungi/magnaporthe/**
Coprinus cinereus	Draft	36	WGS	**http://www.broad.mit.edu/annotation/fungi/coprinus_cinereus/**
Cryptococcus neoformans serotype A	Draft	19	WGS	**http://www.broad.mit.edu/annotation/fungi/cryptococcus_neoformans/**
Dictyostelium discoideum	Draft	34	SCS	**http://dictygenome.bcm.tmc.edu/**
Entamoeba histolytica	Draft	20	WGS	**http://www.sanger.ac.uk/Projects/E_histolytica/**, **http://www.tigr.org/tdb/e2k1/eha1/**
Tetrahymena thermophila	Draft	100	WGS	**http://www.tigr.org/tdb/e2k1/ttg/index.shtml**
Theileria parva	Draft+Finished	9	WGS	**http://www.tigr.org/tdb/e2k1/tpa1/**
Brugia malayi	Ongoing	110	WGS	**http://www.tigr.org/tdb/e2k1/bma1/**
Plasmodium vivax	Draft	30	WGS	**http://www.tigr.org/tdb/e2k1/pva1/**
Plasmodium yoelli	Draft	30	WGS	**http://www.tigr.org/tdb/e2k1/pya1/**
Pneumocystis carinii	Ongoing	7	WGS	**http://pneumocystis.cchmc.org/**
Schistosoma mansoni	Draft	270	WGS	**http://www.sanger.ac.uk/Projects/S_mansoni/**, **http://www.tigr.org/tdb/e2k1/sma1/**
Toxoplasma gondii	Ongoing	80	WGS	**http://www.tigr.org/tdb/e2k1/tga1/**
Trichomonas vaginalis	Ongoing	60	WGS	**http://www.tigr.org/tdb/e2k1/tvg/**
Trypanosoma brucei	Ongoing	30	Hybrid	**http://www.tigr.org/tdb/e2k1/tba1/**, **http://www.sanger.ac.uk/Projects/T_brucei/**
Trypanosoma cruzi	Ongoing	40	Hybrid	**http://www.tigr.org/tdb/e2k1/tca1/**
Sea urchin	Draft	800	Hybrid	**http://hgsc.bcm.tmc.edu/projects/seaurchin/**
Tetraodon nigroviridis	Ongoing	400	WGS	**http://www.broad.mit.edu/annotation/tetraodon/index.html**
Kangaroo	Ongoing		WGS	**http://kangaroo.genome.org.au/**
Chicken	Ongoing	1200	Hybrid	**http://genome.wustl.edu/projects/chicken/**

[a]HS: hierarchical shotgun, WGS: whole-genome shotgun, Hybrid: both whole-genome shotgun and map-based clone sequencing used together, SCS: single chromosome shotgun.

one error in 5000 bases) that is well ordered and mapped to chromosomes. Many additional eukaryotic genome-sequencing projects are either completed, underway, or planned (see Table 1).

Given what has been learned so far, what is the best strategy for sequencing additional large eukaryotic genomes? The choice of sequencing strategy for these organisms will depend on the goals of the sequencing project and on the answers to three primary questions: (1) How closely related is the genome to the genome of another organism that has been sequenced? (2) What is the nature of the repetitive elements in the genome? (3) Will the genome eventually be finished to very high quality? Each of these issues will be addressed in the following paragraphs.

3.1. Comparative sequencing

Increasingly, phylogenetic relatives are being sequenced to assist in the analysis and interpretation of a reference genome sequence (*see* Article 48, **Comparative sequencing of vertebrate genomes**, Volume 3). *Drosophila pseudoobscura*, *C. briggsae,* and chimpanzee were all selected for sequencing not only on their own merits but by what a comparison of their sequence might reveal about better-studied close relatives. In the case of chimpanzee, the nucleotide identity is so high that virtually every sequence read from the chimp genome can be assigned to a unique corresponding region of the human genome sequence, with the exception of sequences that are chimpanzee-specific. *Drosophila pseudoobscura* and *C. briggsae* are more distantly related to their respective references (*D. melanogaster* and *C. elegans*) – about the same phylogenetic distance apart as human and mouse. At this distance, most of the nonfunctional sequence is no longer conserved, facilitating identification of genes and conserved regulatory elements. The primary goal of these projects is to identify matching regions in a reference genome, and secondarily to identify the sequence unique to each genome, rather than to construct a high-quality finished sequence. In this case, a whole-genome shotgun strategy is clearly the most efficient way to generate high-quality draft sequence for comparison.

3.2. Repetitive elements

When long, nearly identical repeats are present, and when it is important to correctly resolve those repeat structures, such as for the study of chromosome evolution, a hierarchical or hybrid approach is likely to be the most effective. Whole-genome shotgun data can indicate the presence of repeated sequences (based on excess sequence coverage at those locations), but physically separating each copy of each repeat in BAC clones is the best way of correctly assembling each copy of long identical repeats. In repeat-rich genomes, construction of a BAC map by use of restriction fingerprint patterns can also be quite challenging, necessitating additional laboratory work to confirm both map and sequence.

3.3. Gap closure

Genome finishing – the process of filling gaps and confirming the quality of the entire sequence – is a quite different task from collecting the initial sequence data for a project, regardless of whether a hierarchical or whole-genome strategy is used. Plasmid subclones from each BAC or from whole-genome library used in the initial sequencing phase are selected for additional sequencing if they span a gap in a contig or a low-quality region. Additional finishing techniques involve sequencing of PCR-amplified segments of the genome and direct sequencing of BAC clones. One of the most difficult challenges of genome finishing is in closing gaps where no cloned DNA is present. These so-called physical gaps (because they are not physically present in any of the clone libraries) often result from portions of the genome that are not clonable in the standard cloning vectors that propagate in *E. coli*. For small genomes, a combination of sequencing subclones from whole-genome shotgun libraries, direct BAC or genomic sequencing, and PCR have been very successful at achieving high-quality genomic sequence. For larger metazoan genomes, where the whole-genome libraries contain millions of clones, finishing has primarily been performed on a BAC-by-BAC basis.

The cost per basepair for genome finishing is easily 50 times the cost of producing the first ∼95% of the sequence in draft form. The high cost and technical complexity of producing a finished genome sequence means that there will be many more draft than completely finished genome sequences for the foreseeable future. Methods such as comparative gene-finding programs (Parra *et al*., 2003; Flicek *et al*., 2003) that take best advantage of the incomplete information present in draft genome sequences continue to evolve.

4. Conclusion

As the cost of DNA sequencing continues to decline and analytical methods for assembling, annotating, and interpreting genome sequence improve, it is clear that more eukaryotic genomes will be sequenced. In fact, more than three dozen projects are already well along (Table 1) and many more are planned. The wealth of genome sequence data that will result will prove quite powerful for assisting in understanding the evolution of metazoan species, the structure of chromosomes, the sets of functional genes, and the sequences that control their expression.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.

The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP and Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research*, **12**, 177–189.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, *et al*. (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol*, **3**, Research0079.

The *C. elegans* Consortium (1998) Genome sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science*, **282**, 2012–2018.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, *et al*. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.

Flicek P, Keibler E, Hu P, Korf I and Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Research*, **13**, 46–54.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, *et al*. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, *et al*. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D, *et al*. (1997) The Yeast Genome Directory. *Nature*, **387**, S1–S105.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, *et al*. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.

International Human Genome Sequencing Consortium (IHGSC) (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

International Mouse Genome Sequencing Consortium (IMGSC) (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, *et al*. (2003) The dog genome: survey sequencing and comparative analysis. *Science*, **301**, 1898–1903.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al*. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW and Guigo R (2003) Comparative gene prediction in human and mouse. *Genome Research*, **13**, 108–117.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al*. (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al*. (2003) The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. *PLoS Biology*, **1**, e45.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Weber JL and Myers EW, *et al*. (1997) Human whole-genome shotgun sequencing. *Genome Research*, **7**, 401–409.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, *et al*. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.

Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al*. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Introductory Review

Genome sequencing of microbial species

Jacques Ravel and Claire M. Fraser
The Institute for Genomic Research, Rockville, MD, USA

1. Whole-genome shotgun cloning: a revolution in the microbial field

Microbes account for most of life on earth and are critical to its ecological balance. However, researchers have only scratched the surface of the tremendous biodiversity that these organisms display. Less than 1% has been cultured – only a minute proportion of the microbial diversity present in the environment. If this diversity is an indicator of the physiological, metabolic, and adaptation ability of the uncultured microorganisms, one can barely start to imagine the enormous diversity that can be discovered among the microbes on earth. No other field of research has embraced and applied genomic technology more than the field of microbiology, and genomic science has provided information that cannot be obtained by any other means. Microbial genomics has a broad range of applications, from understanding basic biological processes, host–pathogen interactions, and protein–protein interactions to discovering DNA variations that can be used in genotyping or forensic analyses. In addition, genomic data is being applied to unravel gene expression patterns through the development and analysis of DNA microarray data.

In 1995, The Institute for Genomic Research led by J. Craig Venter sparked a revolution in genomics by using whole-genome shotgun sequencing (Fleischmann *et al*., 1995) (Figure 1) to determine the first complete genome sequence of a free-living organism, the bacterium *Haemophilus influenzae*. Since that first report, more than 220 microbial genomes have been sequenced and at least another 650 are in progress (February 2005; **http://www.genomesonline.org/**) (Figure 2). This global effort has focused primarily on pathogens, which to date account for the majority of all genome projects (Figure 2), and has generated a large amount of raw material for in silico analysis. Additionally, in recent years, multiple in recent years, multiple strains of the same species, or multiple species of the same genus, have been the targets of sequencing projects, opening the possibility of comparing closely related genomes (*see* Article 61, **Comparative genomics of the ε-proteobacterial human pathogens *Helicobacter pylori* and *Campylobacter jejuni***, Volume 4). This will improve our understanding of microbial biology, pathogenicity, and evolution. However, the major challenge in the postgenomic era is to fully exploit and decipher this new accumulating wealth of information.

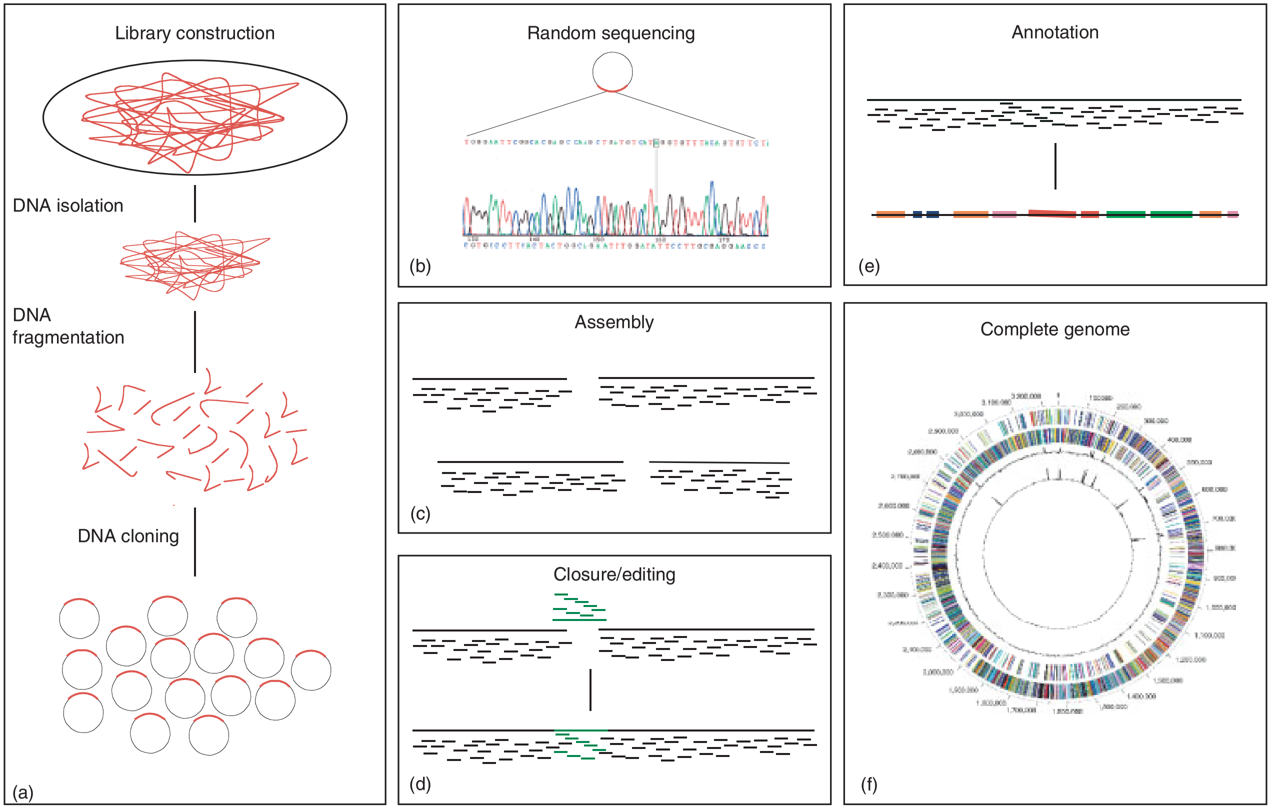


Figure 1 The steps involved in the whole-genome shotgun sequencing procedure. (a) Library construction. Total genomic DNA is extracted and mechanically sheared to smaller fragments. Each fragment is ligated into a cloning vector. (b) Random sequencing. About 6000 random clones per megabasepair are sequenced from both end of the insert to achieve 8X coverage. (c) The small sequences (~ 800 bp) are assembled into larger contigs using computational algorithms, such as the Celera Assembler. (d) The contigs are linked to each other during the closure phase, where the sequence is also manually edited. (e) Annotation. Using programs such as Glimmer, open reading frames (ORFs) are marked. The predicted protein sequences from these putative open reading frames are searched against nonredundant protein databases. (f) A Complete genome is obtained after manual curation of the annotation

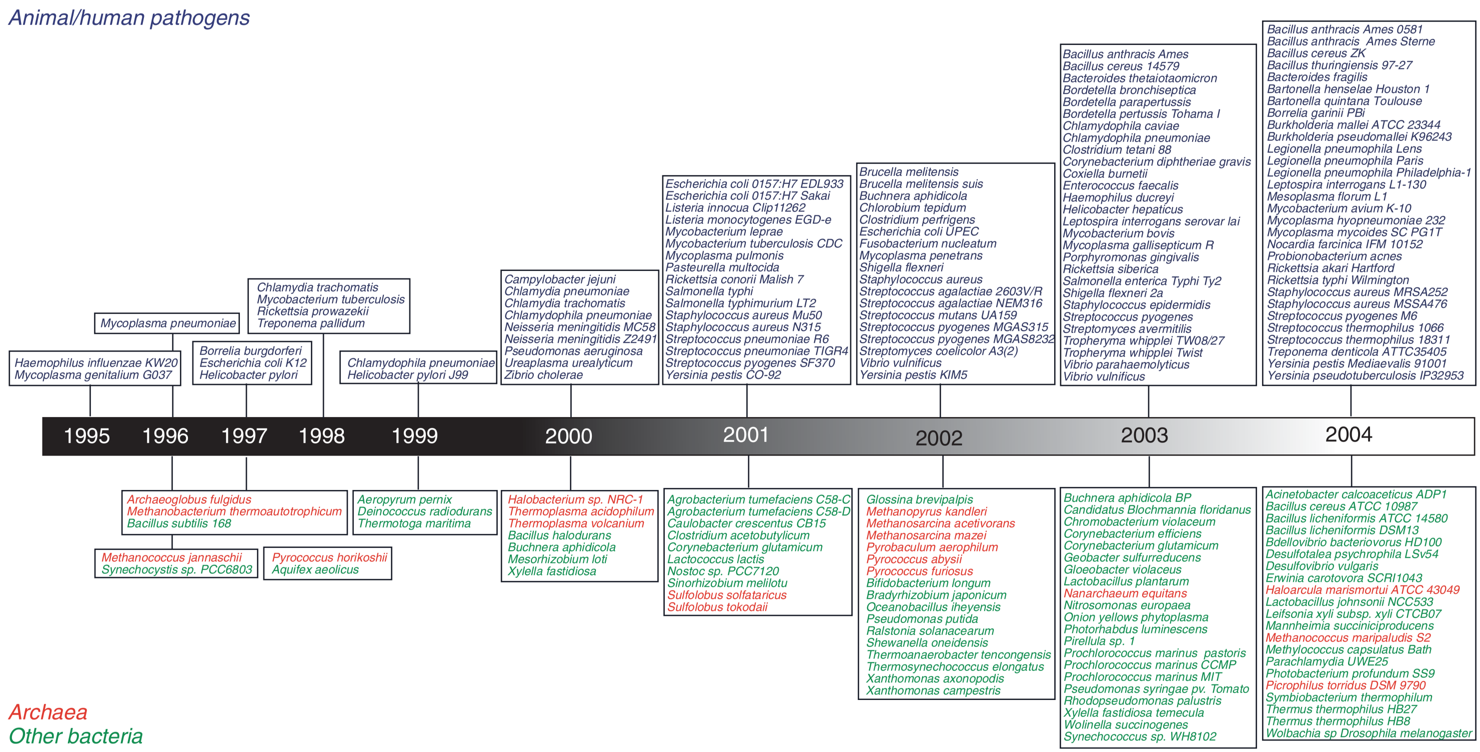


Figure 2 Completed genome sequencing project timeline

The whole-genome shotgun sequencing strategy does not require an initial mapping step to create a set of overlapping clones, and instead relies on computational methods (TIGR Assembler (Sutton *et al*., 1995), the Celera Assembler (Myers *et al*., 2000), and Phrap (**http://www.phrap.org**)) to correctly assemble tens of thousands of random DNA sequences 300–900-bp long. In some cases, the algorithms underlying the assembly software have also been shown to be powerful enough to successfully assemble larger eukaryotic genomes including the human genome (Venter *et al*., 2001; *see also* Article 25, **Genome assembly**, Volume 3). Given the current state of sequencing technologies, whole-genome shotgun sequencing remains the industry standard.

2. Genome annotation

The first step in the analysis of a completed and fully assembled genome is to determine the precise location and assign a putative function to all the protein coding regions, through a process known as annotation (*see* Article 29, ***In silico* approaches to functional analysis of proteins**, Volume 7). A wide variety of bioinformatics methods that have been developed to analyze sequence data have made annotation an increasingly sophisticated process. Computational gene finders (*see* Article 13, **Prokaryotic gene identification *in silico***, Volume 7) using Interpolated Markov modeling algorithms, such as Glimmer (Delcher *et al*., 1999), are routinely capable of finding more than 99% of all genes in a microbial genome. The predicted protein sequences from these putative open reading frames (ORFs) are searched against nonredundant protein databases and well-curated protein families, such as the PFAM (Bateman *et al*., 2002) and TIGRFAM (Haft *et al*., 2003) collections, that have been created using hidden Markov models (HMMs). HMMs are powerful statistical representations of groups of proteins that share sequence, and consequently, functional similarity. HMMs can represent very specific enzymatic functions or a superfamily of related functions. The use of HMMs has helped refine the annotation process. In addition, searches for PROSITE motifs (Sigrist *et al*., 2002), lipoproteins, signal peptides, and membrane-spanning regions are performed. On the basis of the evidence gathered, a two-stage annotation protocol is carried out whereby an initial automated annotation is followed by manual curation of each gene assignment by an expert biologist to ensure accuracy and consistency of the putative function of each predicted coding region. Proteins whose specific function cannot be confidently determined are designated "putative" or given a less specific family name. Proteins without any significant matches in any of the searches performed are annotated as hypothetical proteins. Consistent description and annotation of genes in different databases is critical to facilitate uniform queries across independent databases. This problem is being addressed by the development of controlled vocabularies (ontologies), such as the Gene Ontology (GO) project (The Gene Ontology Consortium, 2004; *see also* Article 82, **The Gene Ontology project**, Volume 8), where gene products are described in terms of their associated biological process, cellular components, and molecular functions in a species-independent manner.

3. What have we learned so far?

High-throughput genome sequencing technologies have only been around for less than 10 years, but the impact of these technologies has been profound. Genome sequence data have been obtained from representative species of all three domains of life (Figure 2); however, because of their relatively small size, bacterial and archaeal genomes have dominated the field (Figure 2). Taken together, comparative genome analysis has revealed interesting patterns pertaining to microbial species; for example, gene density in microbes is very consistent with about one gene per kilobase of DNA. Although we are able to identify microbial genes with a high degree of success, we cannot assign a function to about a quarter of all the ORFs in each species sequenced so far. This observation demonstrates how little is known about the biology and biochemistry of microbial species, and supports the idea of an incredible microbial diversity. These sets of genes that encode hypothetical proteins represent exciting opportunities for the research community and are not only potential sources of biological resources to be explored for future use, but also clearly indicate the need for further extensive genetic, enzymatic, and physiological analyses, before genomic data can be fully exploited.

Analysis of more than 150 microbial genome sequences has revealed an unexpected diversity and variability in genome size and structure, even in species previously thought to be identical. Many microbes possess diverse chromosome architectures that are quite different from the classical single circular chromosome. For example, the genome sequence of the human pathogen, *Vibrio cholerae*, unexpectedly revealed the presence of two circular chromosomes (Heidelberg *et al*., 2000), whereas the genome of *Borrelia burgdorferi* (Casjens *et al*., 2000; Fraser *et al*., 1997), the causative agent of Lyme disease, contained a relatively small (910 kb) linear chromosome and an unprecedented number of 21 linear and circular plasmids. On the other hand, the *Streptomyces coelicolor* linear chromosome is more than 9-Mb long (Bentley *et al*., 2002). In addition to differences in genome structure, microbial genomes vary largely in their GC content ranging from 24% to more than 70%. The effect of this disparity in GC content is reflected in the wide range of codon usage and the amino acid composition of proteins among various species.

As noted earlier, the study of bacterial pathogens has dominated and influenced the microbial genomic arena. This has resulted from the potential for developing a better understanding of virulence as well as identifying putative targets for vaccine and antimicrobial drugs. Access to the genomes of a variety of pathogens has allowed scientists to broaden their knowledge of pathogenicity through comparative genome analysis.

Organisms that belong to the same genus can differ in gene content by as much as 25% as it was found when the genome of *Escherichia coli* K-12 was compared to *E. coli* 0157:H7 (Hayashi *et al*., 2001; *see also* Article 51, **Genomics of enterobacteriaceae**, Volume 4). Insertion and deletion events appear to have played a major role and account for most of the differences observed. Pathogenicity islands, which are large blocks of self-mobile DNA that carry genes enabling an organism to act as a pathogen, have the ability to transfer from one organism and integrate into a new host. Other pathogens show little variation in chromosomal gene content, as demonstrated by the comparison of the genomes of two isolates of *Yersinia*

pestis (Deng *et al*., 2002; Parkhill *et al*., 2001), the etiologic agent of plague (*see* Article 58, ***Yersinia***, Volume 4). Remarkable differences in the chromosome structures, dominated by genome rearrangements, accounted for most of the variation observed between these two closely related strains. The differences appear to result from multiple inversions of genome segments at insertion sequences. *Y. pestis* sp. carry most of their virulence determinant on plasmids, which are absent in its ancestor, *Yersinia pseudotuberculosis*. A remarkable number of pseudogenes (degenerated and inactive genes) have been found on the genomes of *Y. pestis*, an indication of a recent and still evolving genome.

Often, differences between a pathogen and a nonpathogen cannot be explained solely by looking at gene presence or absence, but by subtle single nucleotide changes. These changes can have disproportionately large consequences. Important virulence genes have been shown to be completely inactivated by such changes. Virulence or survival can also be modulated by hypervariable short homopolymeric sequences, which vary in size during replication, and can result in frameshifts and inactivation or activation of important virulence genes, as seen in the human pathogens *Helicobacter pylori* and *Campylobacter jejuni* (Parkhill *et al*., 2000; *see also* Article 61, **Comparative genomics of the ε-proteobacterial human pathogens *Helicobacter pylori* and *Campylobacter jejuni***, Volume 4).

Genomic information can also be used to design novel vaccines and drugs (*see* Article 55, **Reverse vaccinology: a critical analysis**, Volume 4). In a pioneering study, Pizza *et al*. (2000) have exploited the genome sequence of *Neisseria meningitidis* to identify two highly conserved vaccine candidates within a set of cell-surface expressed or secreted proteins. There is no doubt that genomics has contributed enormously to a better understanding of bacterial pathogenicity, however, one genome is not enough. There is much that is still unknown and comparative genomics of close relatives of both pathogens and nonpathogens will be critical to unravel the secrets of microbial pathogenicity and continue the search for better and innovative vaccines or drugs.

The initial focus on pathogenic microbial species has shifted to include non-pathogenic environmental microbes. Understanding and accessing the tremendous microbial biochemical diversity that exists in the environment could have an important impact on industrial processes and help in resolving environmental issues, such as the bioremediation of human pollution.

Many archaea are considered extremophiles, as they often thrive under "extreme" conditions, such as high or low temperatures, high pressures or high salt concentrations among others. The novel enzymes encoded in these genomes (Figure 2) offer clear potential for biotechnological applications. In addition, genome analysis of the hyperthermophilic bacteria, *Thermotoga maritima* (Nelson *et al*., 1999) revealed that 20–25% of the genes in this species were more similar to genes from archaea than from bacteria, leading to a renewed interest in the process of lateral gene transfer and the role that it plays in microbial evolution and diversity (*see* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4).

Among the bacteria, the genome sequence of *Deinococcus radiodurans* (White *et al*., 1999), the most radiation-resistant organism on earth, and *Geobacter sulfurreducens* (Methe *et al*., 2003), which can clean up uranium and organic waste

contamination, will allow scientists to develop and optimize practical applications, such as the bioremediation of radioactive metals and harvesting electricity from waste organic matter. The genome of *Photorhabdus luminescens*, an insect pathogen living in symbiosis with a nematode has been fully sequenced (Duchaud *et al*., 2003). The analysis uncovered a variety of genes coding for entomopathogenic toxins, potentially useful in the fight against insect pests. Moreover, *P. luminescens* carries a large number of genes coding for the biosynthesis of antibiotics and fungicides, which could have potential applications for the treatment of infectious diseases. The genomes of *Streptomyces coelicolor* (Bentley *et al*., 2002) and *Streptomyces avermitilis* (Ikeda *et al*., 2003; Omura *et al*., 2001), both known to produce a wide variety of natural products, will assist in genome engineering to make novel and more efficient antimicrobial agents.

Researchers have only scratched the surface of microbial biodiversity. In order to harvest this enormous potential, genome shotgun sequencing is being applied to the environment. In a landmark study, the microbial populations from water samples collected in the Sargasso Sea were sequenced (Venter *et al*., 2004). An estimated 1.2 million new genes have been identified from at least 1800 genomic species. Similar techniques were applied to a community of microbes from a biofilm growing at pH 0.83 on the surface of acid mine drainage (Tyson *et al*., 2004). In this study, the low diversity genomic community was entirely reconstructed – the subsequent examination of the metabolic capabilities of this community gave valuable information on how each organism participates to the ecology of the biofilm. These types of microbial studies will help us define the entire repertoire of organisms in specialized niches and ultimately the mechanisms by which they interact in the biosphere.

With the technical advances of genome sequencing and analysis, genomics has also found an application in the field of microbial forensics. After the bioterror events of October 2001, where letters containing spores of *Bacillus anthracis*, the causative agent of anthrax, were sent through the mail, the genome of the *B. anthracis* isolate responsible for the death of a Florida man was rapidly sequenced and single nucleotide polymorphisms were found that could help identifying the origin of the samples used in this attack (Read *et al*., 2002).

4. Conclusions

Scientists in a number of different fields have employed the tools of genomics – no field has embraced and applied these technologies as quickly and effectively as the field of microbiology. Genomics will continue to improve the quality of human life well into the future as scientists continue to unravel the enormous amount of data that is being accumulated. More genome sequences are needed, new annotation tools must be developed and applied, and the databases that archive genomic data must be improved for better cross communication and up-to-date data. There is no question that genome-sequencing technologies are rapidly improving and that the data are going to accumulate at a faster pace in a future. The genomics community needs to be prepared to analyze and make use of this forthcoming deluge of

information. However, because genome sequence should not be considered an end-point and is only the first step in understanding biological processes, the microbial scientific community at large needs also to be trained and ready to make better use of this incredible resource.

References

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, *et al*. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.

Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ, *et al*. (2000) A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology*, **35**, 490–516.

The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261.

Delcher AL, Harmon D, Kasif S, White O and Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**, 4636–4641.

Deng W, Burland V, Plunkett G III, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, *et al*. (2002) Genome sequence of *Yersinia pestis* KIM. *Journal of Bacteriology*, **184**, 4601–4611.

Duchaud E, Rusniok C, Frangeul L, Buchrieser C, Givaudan A, Taourit S, Bocs S, Boursaux-Eude C, Chandler M, Charles JF, *et al*. (2003) The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*. *Nature Biotechnology*, **21**, 1307–1313.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, *et al*. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, *et al*. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research*, **8**, 11–22.

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, *et al*. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M and Omura S (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature Biotechnology*, **21**, 526–531.

Methe BA, Nelson KE, Eisen JA, Paulsen IT, Nelson W, Heidelberg JF, Wu D, Wu M, Ward N, Beanan MJ, *et al*. (2003) Genome of *Geobacter sulfurreducens*: Metal reduction in subsurface environments. *Science*, **302**, 1967–1969.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al*. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson LD, Nelson WC, Ketchum KA, *et al*. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.

Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, *et al*. (2001) Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12215–12220.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, *et al*. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, *et al*. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.

Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, *et al*. (2000) Identification of vaccine candidates against serogroup *B. meningococcus* by whole-genome sequencing. *Science*, **287**, 1816–1820.

Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang LX, Holtzapple E, Busch JD, Smith KL, Schupp JM, *et al*. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033.

Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A and Bucher P (2002) PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief in Bioinformatics*, **3**, 265–274.

Sutton G, White O, Adams M and Kerlavage AR (1995) TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9–19.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, *et al*. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, **286**, 1571–1577.

Introductory Review

Hierarchical, ordered mapped large insert clone shotgun sequencing

Bruce A. Roe

University of Oklahoma, Norman, OK, USA

The underlying concept of employing dideoxynucleotides as chain terminators in the DNA sequencing reaction, to create a replicated nested fragment set that is size fractionated and detected, has changed little since it was first reported by Sanger *et al*. in 1970 (Sanger *et al*., 1977). In contrast, the detailed methods implemented in the laboratory to create, resolve, and detect the actual dideoxynucleotide sequence data have improved greatly owing to the discovery and use of improved DNA polymerases (Chien *et al*., 1976; Saiki *et al*., 1988), the development of automated electrophoresis instrumentation (Ansorge and Barker, 1984; Smith *et al*., 1986; Karger *et al*., 1991), and the availability of highly sensitive fluorescently labeled nucleic acid derivatives that can be automatically detected efficiently after laser excitation (Bauer, 1990). Although all of these methodological improvements were important, it was the introduction of commercially available automated DNA sequencing instruments, and a significant influx of massive public and private sector funding (Choudhuri, 2003) over the last decade, that paved the way for the yearly almost log-scale increases in the amount of DNA sequenced data collected and a parallel significant reduction in DNA sequencing cost. As a result, a paradigm shift evolved that increased the emphasis on approaches and methods to generate and assemble large target DNA sequences, rather than the actual DNA sequencing data collection process. Clearly, the latter still remains important as improvements continue to be made through the introduction of newer DNA sequencing instruments, several of which are described in this section, as well as significant improvements in DNA sequence assembly programs, as described in other chapters. However, as with almost all science, changes evolve slowly over time. This was the case for DNA sequencing, which began by directed sequencing of restriction digested and subcloned short target sequences and that subsequently evolved into a hierarchical map-based approach to sequence larger genomic regions and then into the shotgun sequencing the underlying minimal tiling path ordered large insert clones combined with more directed closure and finishing phases. These methods now have evolved into widespread implementation of whole-genome shotgun sequencing and assembly to order and orient contiguous

but gapped sequences without much attention to the closure and finishing of the entire genome.

Initially, as the DNA sequencing data collection technologies evolved over the past decade, several groups also focused on developing strategies for obtaining the target DNA that subsequently was subjected to the sequencing process. Here the genomic DNA either was cleaved by enzymatic or physical methods and shotgun libraries were produced using various host/vector systems. Cosmid and yeast artificial chromosome (YAC) vectors (McCormick *et al*., 1987; Burke and Olson, 1991) initially were employed for this purpose, and hybridization methods were used to determine which cosmid or YAC clone(s) encoded the target region of interest (Feinberg and Vogelstein, 1983). When multiple, adjacent probes were used, either created as PCR products amplified off end-sequenced or fully sequenced cosmids (termed over-goes) or from sequencing fragmented YACs, it also was possible to overlap the cosmids and/or YACs and to generate a tiling path covering regions of genomic DNA several orders of magnitude larger than that covered by the initial cosmid or YAC. Although a valuable approach, using these hybridization approaches to completely sequence a large genomic region through making a tiling path of a large number of target clones was both time consuming and often prone to errors that could be traced to the specificity of the hybridization probe used. In addition, both YAC and cosmid vector systems had the tendency to either lose portions of the inserted DNA or otherwise rearrange it since there was little selective pressure to accurately maintain the originally cloned genomic DNA fragment. Therefore, more stable host/vector systems were developed including, namely, bacterial artificial chromosome (BAC)-based clones that could contain between 100 000 and 200 000 bp of genomic DNA insert (Shizuya *et al*., 1992) and fosmid-based cloning vectors that typically contained ∼40 000 bp of inserted genomic DNA (Kim *et al*., 1992). Since both types of clone libraries were engineered so that they were much less prone to deletions or rearrangements, improved methods to generate tiling paths for large segments of genomic DNA were now possible.

Thus, the hierarchical map-based approach needed to complete the sequence of large reference genomes, for example, flies, worms, humans, and mice, necessitated the development of BAC fingerprinting methods to create a tiling path of overlapping individual clones that then could be used to generate a minimal smaller set of BAC clones for eventual sequencing. Initially, these physical maps were constructed using high-throughput polyacrylamide gel electrophoresis to separate the restriction enzyme–digested BAC clone DNA followed by visualization using a fluorimager, followed by normalizing the band values and gel traces by editing the digitized images (Sulston *et al*., 1989). More recently, capillary electrophoresis of fluorescent-labeled DNA restriction digests has resulted in a more automated process by which thousands of BACs from a library can be rapidly fingerprinted (Ding *et al*., 1999; Ding *et al*., 2001). In either case, the resulting visualized and normalized restriction digestion patterns then are compared and overlapped via computer-based methods such as FPC (Marra *et al*., 1997) in which the clones are ordered into tiling paths on the basis of the occurrence of shared bands.

Once a minimum tiling path is obtained, the DNA from the underlying BAC clones is isolated and subjected to shotgun sequencing. This process entails breaking a large target DNA randomly into smaller fragments that then are cloned

into a vector. Initially, m13 phage vectors (Messing *et al*., 1977) were used for this purpose, but today double-stranded pUC-based plasmid vectors (Vieira and Messing, 1982) are used almost exclusively as both ends of the cloned insert can be more easily sequenced from the plasmid than from the single-stranded phage vector. After end sequencing, overlapping identical sequences are assembled to recreate the sequence of the original sequence of the BAC-cloned insert. This process is analogous to reconstituting the front page of a daily newspaper by putting thousands of copies of it through a shredder and then overlaying the pieces with similar words and pictures to give a single copy of the initial page.

The initial description of shotgun cloning was given by Steve Anderson in 1981, when he described the cloning of the products of a partial DNAse 1 digestion of a 4257-bp target fragment of the bovine mitochondrial genome into M13 vectors (Anderson, 1981) followed by randomly picking subclones and obtaining the end sequences of each of them. The resulting overlapping sequences then could be assembled into a final, contiguous, consensus sequence representing that of the initial DNA target fragment. This shotgun technique took several years to become widely accepted because the high number of DNA sequencing reactions and subsequent polyacrylamide gel–generated sequences that had to be manually read were both too expensive and too highly labor intensive. It was not until almost a decade later that two independent laboratories, Lee Hood's group at Cal Tech (Smith *et al*., 1986) and Ansorge's group at the EMBL laboratory in Heidelberg, Germany (Voss *et al*., 1990), introduced automated DNA sequence data collection methods that resulted in the first commercially available fluorescent-based DNA sequencers that were produced by Applied Biosystems and Pharmacia, respectively, in the early 1990s. The major advantage of these fluorescent-based DNA sequencing instruments was that the data collection process was automated. However, since the fluorescent-labeled reactions produced weaker fluorescent signal than the radioactive-labeled reactions, they required higher amounts of single-stranded DNA templates and fluorescent-labeled primers to produce the required signal strength during a constant temperature incubation. The later introduction of thermostable DNA polymerases allowing reaction temperature cycling, termed "cycle sequencing" (Murray, 1989), and fluorescent-labeled dideoxynucleotide terminators, eventually made it possible to use much less DNA template in a single reaction. This, when coupled with the automated data collection on slab gel–equipped instruments, ensured that the shotgun sequencing approach truly became widely accepted. More recently, the introduction of capillary-based DNA sequence data collection instruments, by Applied Biosystems, Molecular Dynamics, and Beckman, that have shorter runtimes and automated sample loading than previous slab gel–based machines, resulted in the elimination of the labor-intensive sequencing reaction pipetting and data collection steps.

This chapter includes descriptions of the work of several groups that have resulted in sequencing large numbers of DNAs from both higher eukaryotes (*see* Article 1, **Eukaryotic genomics**, Volume 3) and microbial genomes (*see* Article 2, **Genome sequencing of microbial species**, Volume 3), as well as a discussion of sequencing template preparation methods (*see* Article 4, **Sequencing templates – shotgun clone isolation versus amplification approaches**, Volume 3) and a description of robotics and automation techniques (*see* Article 5, **Robotics and**

automation, Volume 3). These articles are followed by contributions from three of the leading groups in developing the next generation of high throughput DNA sequencing methods that include microelectrophoresis devices for DNA sequencing (*see* Article 6, **Microelectrophoresis devices for DNA sequencing**, Volume 3), single molecule array-based sequencing (*see* Article 7, **Single molecule array-based sequencing**, Volume 3), and real-time DNA sequencing (*see* Article 8, **Real-time DNA sequencing**, Volume 3).

References

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, **9**(13), 3015–3027.

Ansorge W and Barker R (1984) System for DNA sequencing with resolution of up to 600 base pairs. *Journal of Biochemical and Biophysical Methods*, **9**(1), 33–47.

Bauer GJ (1990) RNA sequencing using fluorescent-labeled dideoxynucleotides and automated fluorescence detection. *Nucleic Acids Research*, **18**(4), 879–884.

Burke DT and Olson MV (1991) Preparation of clone libraries in yeast artificial-chromosome vectors. *Methods in Enzymology*, **194**, 251–270.

Chien A, Edgar DB and Trela JM (1976) Deoxyribonucleic acid polymerase from the extreme thermophile thermus aquaticus. *Journal of Bacteriology*, **127**(3), 1550–1557.

Choudhuri S (2003) The path from nuclein to human genome: a brief history of DNA with a note on human genome sequencing and its impact on future research in biology. *Bulletin of Science, Technology & Society*, **23**, 360–367.

Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, Lam JY, Kim YM and Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics*, **74**(2), 142–154.

Ding Y, Johnson MD, Colayco R, Chen YJ, Melnyk J, Schmitt H and Shizuya H (1999) Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. *Genomics*, **56**(3), 237–246.

Feinberg AP and Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochemistry*, **132**(1), 6–13.

Karger AE, Harris JM and Gesteland RF (1991) Multiwavelength fluorescence detection for DNA sequencing using capillary electrophoresis. *Nucleic Acids Research*, **19**(18), 4955–4962.

Kim UJ, Shizuya H, de Jong PJ, Birren B and Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Research*, **20**(5), 1083–1085.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**(11), 1072–1084.

McCormick M, Gottesman ME, Gaitanaris GA and Howard BH (1987) Cosmid vector systems for genomic DNA cloning. *Methods in Enzymology*, **151**, 397–405.

Messing J, Gronenborn B, Muller-Hill B and Hans Hopschneider P (1977) Filamentous coliphage M13 as a cloning vehicle: insertion of a HindII fragment of the lac regulatory region in M13 replicative form in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(9), 3642–3646.

Murray V (1989) Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Research*, **17**(21), 8889.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB and Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**(4839), 487–491.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(18), 8794–8797.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB and Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**(6071), 674–679.

Sulston J, Mallett F, Durbin R and Horsnell T (1989) Image analysis of restriction enzyme fingerprint autoradiograms. *Computer Applications in the Biosciences*, **5**(2), 101–106.

Vieira J and Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene*, **19**(3), 259–268.

Voss H, Zimmermann J, Schwager C, Erfle H, Stegemann J, Stucky K and Ansorge W (1990) Automated fluorescent sequencing of cosmid DNA. *Nucleic Acids Research*, **18**(4), 1066.

Sequencing templates – shotgun clone isolation versus amplification approaches

Rebecca Deadman and Carl W. Fuller
GE Healthcare, Piscataway, NJ, USA

1. Introduction

High-quality DNA is essential to obtaining the greatest success in DNA sequencing. Sequencing quality DNA should contain the lowest possible level of contaminating host DNA, have consistent yields from sample to sample, and be in sufficient quantities to perform multiple sequence reactions. Many thousands of clones are required whether a whole-genome shotgun or clone-by-clone approach is taken so the DNA isolation method must be amenable to automation for a high-throughput sequencing process.

To produce sequencing templates, random fragments of DNA are inserted into cloning vectors and propagated in *Escherichia coli*. Individual clones are cultured in 0.5–2 ml volumes and the episome containing the clone DNA is extracted and purified. Each method differs in the quality of DNA produced, the length of time and cost to perform, and its suitability to automation. Variable DNA concentrations from sample to sample make it difficult to optimize sequencing reactions. DNA isolation may be one of the most basic and routine molecular biology techniques but it is one of the most important for success.

Traditional extraction methods involve chemical isolation of the subclone DNA, taking advantage of the mass difference between the plasmid and chromosomal DNA. Some techniques further involve immobilization of the plasmid to a solid structure. DNA amplification methods circumvent the cell growth step, allowing DNA isolation to proceed directly from bacterial colonies or glycerol stocks. Table 1 shows a comparison of three common methods of DNA preparation for plasmid clones – alkaline lysis (Birnboim and Doly, 1979), solid-phase reversible immobilization (SPRI) (Hawkins *et al*., 1994), and rolling circle amplification (RCA) (Dean *et al*., 2001).

2. Cloning and sequencing vectors

Bacterial artificial chromosomes (BACs) are probably the most common clone type used today for genomic library construction. Libraries in BAC vectors such as

Table 1 Comparison of some popular DNA preparation methods

	DNA isolation method		
	Alkaline lysis	SPRI	RCA
Time to prepare 96 templates	18 h, including overnight culture.	16.5 h, including overnight culture.	4.5 h
Yield	7 µg from 1.5-ml culture	4 µg from 1.5-ml culture	1.5 µg from single colony
Quality of DNA	Good	High	High
Sample-to-sample variability	High	Low	No variability
Number of liquid-handling steps	8	7	2
Plastic ware	Culture plate Microwell plate	Culture plate Microwell plate	Microwell plate
Equipment required	Incubator Centrifuge Vortex	Incubator Magnet	Water bath
Ease of automation	Not easily automated	Easily automated	Easily automated
Key reagents	Bacterial growth media GTE NaOH SDS KOAc Ethanol	Bacterial growth media SprintPrep buffer Isopropanol Ethanol	Denature buffer TempliPhi Premix

pBACe3.6 were used for many of the model organism sequencing projects including human and mouse (Osoegawa *et al*., 2000). The BAC cloning vectors are based on the naturally occurring F-factor found in *E. coli* and they are maintained as supercoiled circular episomes within the bacteria, usually with a single copy per cell. The vector to insert ratio for BACs is very good. Inserts up to 300 kilobases (kb) can be stably introduced into the approximately 8-kb vector. Fosmid vectors (Kim *et al*., 1992) also contain the *E. coli* F-factor, but only 40 kb can be stably maintained in these vectors.

The most common sequencing vectors by far are the double-stranded plasmids, usually the high copy number pUC-based plasmids (Yanisch-Perron *et al*., 1985). With double-stranded vectors, sequence data from both forward and reverse strands can aid assembly of the genome or clone in question (Roach *et al*., 1995). The previously favored single-stranded bacteriophage M13-based vectors are now usually used only for regions that are not stably maintained in pUC plasmids (Chissoe *et al*., 1997). Increased sequence readlength from improvements in sequencing chemistry and instrumentation has allowed an increase in the typical subclone insert size. Sequence readlengths of 700–800 base pairs (bp) are not uncommon and so an average insert size of 2–4 kb is now routinely used for shotgun libraries in plasmid vectors.

3. Traditional plasmid DNA isolation techniques

When Wilson *et al*. (1992) described the methods involved in sequencing a 95-kb section of the mouse genome, the processing of 24 M13-based subclones took one

individual almost a day. With current levels of automation, thousands of subclones can be prepared per day, with human involvement reduced to loading and unloading of microwell plates and reagents.

The most common method for extraction of plasmid DNA from *E. coli* cells is still alkaline lysis. This method takes advantage of the mass differences between plasmid and chromosomal DNA. Bacteria are lysed by treatment with a solution containing sodium dodecyl sulfate (SDS) (CAS # 151-21-3) that denatures the proteins, and sodium hydroxide (NaOH) (CAS # 1310-73-2) that denatures chromosomal DNA. The mixture is neutralized with potassium acetate (KOAc) (CAS # 127-08-2) and the supercoiled, plasmid DNA reanneals rapidly due to its secondary structure and smaller size. The chromosomal DNA and proteins form a solid precipitate with the insoluble potassium salt and SDS and pellet under centrifugation. The plasmid is further purified from the supernatant by alcohol precipitation and washing.

An alternative method to alkaline lysis is the boiling miniprep (Holmes and Quigley, 1981). The cells are lysed by treatment with lysozyme (CAS # 12650-88-3) and heating in the presence of Triton X-100 (CAS # 9002-93-1) and sucrose (CAS # 57-50-1). This procedure releases the plasmid DNA but not the chromosomal DNA from the cell. Centrifugation pellets the cell debris including most of the chromosomal DNA, leaving the plasmid DNA in the supernatant, which is further purified by alcohol precipitation. This method is quicker than alkaline lysis, but the quality of the DNA is lower, having higher chromosomal DNA contamination and more variable yield.

Variability in yield can have a dramatic effect on sequence DNA quality. It is difficult to optimize sequencing reactions when the DNA templates vary widely in concentration. In addition, capillaries in DNA analysis systems can be adversely affected by excessive amounts of DNA in the samples. Sequencing capillaries vary in the range of DNA that they can tolerate, and the type of sequencing instrument should be a consideration when deciding which isolation method to use.

One of the main advantages of both the alkaline lysis and boiling methods is cost. The reagents are inexpensive and easily obtainable and no special equipment is needed, beyond a centrifuge. Once the overnight cell growth is complete, the procedures are fairly quick; two 96-well plates of cultures can be processed in a few hours by a single technician. DNA quality is usable, but probably the lowest of the methods that will be discussed – a chromosomal DNA contamination level of 5–10 % can be expected. As these methods involve centrifugation, they are difficult to automate which is vital for either a cost effective or high-throughput operation.

4. Filter-based purification methods

Most of the commercially available plasmid purification products, such as R.E.A.L™ (rapid extraction alkaline lysis) Prep 96 Plasmid Kit (Qiagen Inc.), begin with the alkaline lysis procedure but differ in the purification step. Following cell resuspension, lysis, and neutralization, the lysate is passed through a membrane that binds the plasmid DNA. The plasmid DNA is washed and then eluted with

water or Tris-EDTA (TE) buffer (CAS # 77-86-1 and 139-33-3). These so-called bind-wash-elute products usually use glass fiber membranes or glass beads that bind DNA in the presence of a chaotropic salt such as guanidine hydrochloride (CAS # 50-01-1). The lysate is usually drawn through the membrane using a vacuum manifold.

These methods eliminate some of the centrifugation steps in the alkaline lysis protocol, making them more amenable to automation and are available in single, 96-well and 384-well formats. Without the alcohol precipitation step, the methods are generally quicker than standard alkaline lysis methods. A 96-well plate of minipreps can be prepared from grown cultures in 45 min. DNA purity with these products is usually higher than with the standard alkaline lysis procedure but the overall cost is also increased owing to the additional filter plates required.

5. Alternative plasmid purification methods

5.1. *Solid phase reverse immobilization (SPRI)*

Technologies that use physical isolation of DNA instead of chemical isolation are commercially available. One such method, SPRI, is used in the SprintPrep™ and CosMCPrep™ DNA purification kits (Agencourt Biosciences Corp.). Carboxyl-coated magnetic beads in the presence of high polyethylene glycol (PEG), alcohol, and salts bind plasmid DNA from lysed bacterial cultures (Figure 1). Cell pelleting and resuspension steps are eliminated by using magnetic separation. Beads with absorbed DNA are washed with ethanol (CAS # 64-17-5) to remove contaminants, then the plasmid DNA is eluted from the beads with water. As this method requires neither centrifuge nor vacuum manifold, it can easily be automated. This method is the quickest of the ones discussed here. A 96-well plate of bacterial cultures can be processed in about 20 min.

5.2. *Rolling circle amplification*

All of the methods discussed so far employ overnight cell growth to propagate plasmid-containing cells and thus amplify the cloned DNA. These methods are effective when high copy-number vectors are used. An alternative strategy is to use multiply primed RCA. This method uses a highly processive, strand-displacing DNA polymerase to amplify the plasmid DNA directly from bacterial colonies, eliminating the need for overnight culture. TempliPhi™ DNA Sequencing Template Amplification kits (GE Healthcare) exploit this technology. Over 10 000-fold amplification can be achieved in as little as 4 h using random hexamer primers that initiate multiple replication forks (Figure 2).

The key to the technology is the DNA polymerase from bacteriophage Phi29. This DNA polymerase is highly processive, incorporating more than 70 000 nucleotides in a single binding event (Blanco *et al*., 1989). RCA is an isothermal reaction and does not require cycling to denature the DNA strands for the next round of amplification as in polymerase chain reaction (PCR). When the enzyme

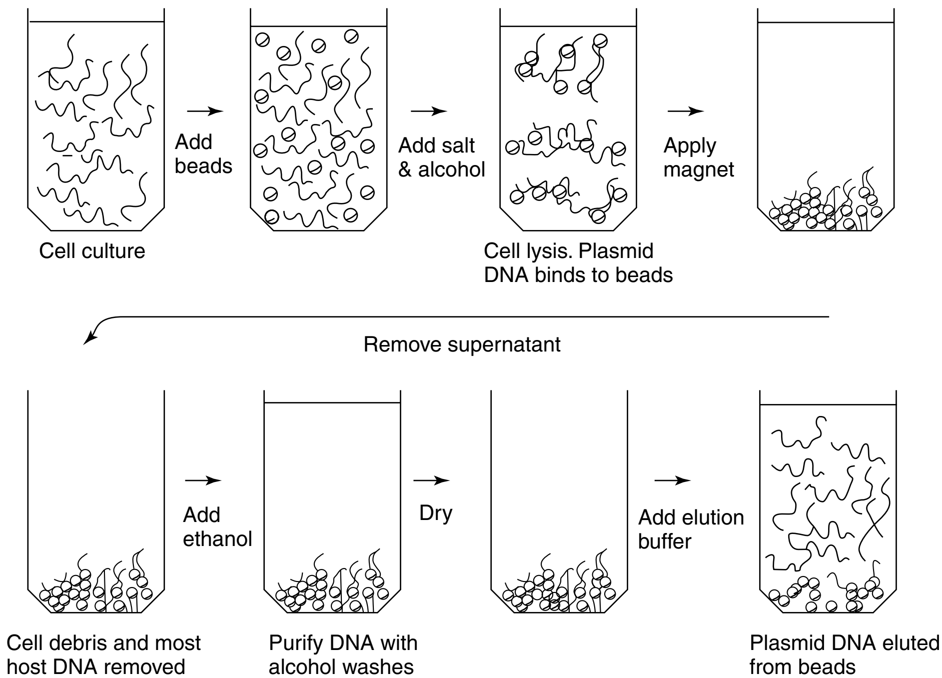


Figure 1 Schematic of SPRI plasmid isolation procedure. Paramagnetic beads are added to bacterial culture. Cells are lysed, and plasmid DNA binds to paramagnetic beads in the presence of isopropanol and salts. Immobilized plasmid DNA is further purified by ethanol washes. DNA is eluted from the beads with water

encounters a nontemplate strand, it simply displaces it, generating single-stranded DNA available for further primer annealing. This leads to exponential amplification of both strands. Phi29 DNA polymerase has a $3'-5'$ exonuclease activity, giving it an error rate of only 1 in 10^6-10^7 (Esteban *et al*., 1993), approximately 100 times lower than *Taq* DNA polymerase (Dunning *et al*., 1988). The product of the RCA process is double-stranded concatamers of the input DNA sequence (Figure 3). Approximately 80% of the product can be digested with restriction endonucleases generating unit-length DNA fragments (Dean *et al*., 2001).

There are a number of advantages to this technique. Speed is an obvious one; sequence ready DNA can be prepared in under 5 h, directly from colonies, with only 20 min of hands-on time for a 384-well plate of templates. Another is consistency of yield. Properly formulated, RCA is an exponential reaction, terminating only when all the nucleotides in the reaction mixture have been exhausted. Every reaction yields the same mass of DNA product, making optimization of downstream sequencing processes much simpler and more reliable than with other methods. In addition, the amplification product can be used directly in sequencing reactions without any further purification. It is not necessary to remove the excess hexamers prior to sequencing as they will not participate in the sequencing reaction owing to their lower melting temperature compared to sequencing primers. The one major disadvantage may be cost, the reagents being more expensive than those used in the alkaline lysis procedure but on par with other commercial plasmid purification

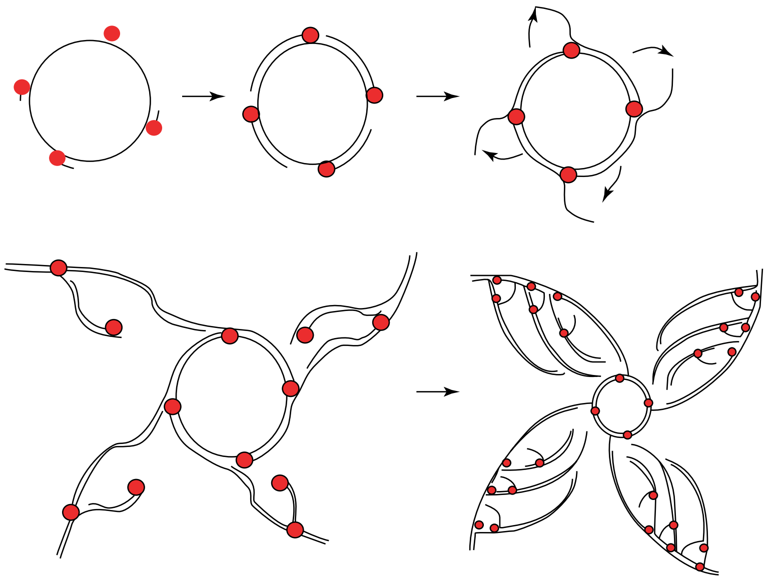


Figure 2 Schematic of rolling circle amplification. Random hexamers bind to the circular template, generating multiple replication forks. Phi29 DNA polymerase displaces the nontemplate strand, making them available for further primer binding. The amplification product is double-stranded tandem copies of the starting circle

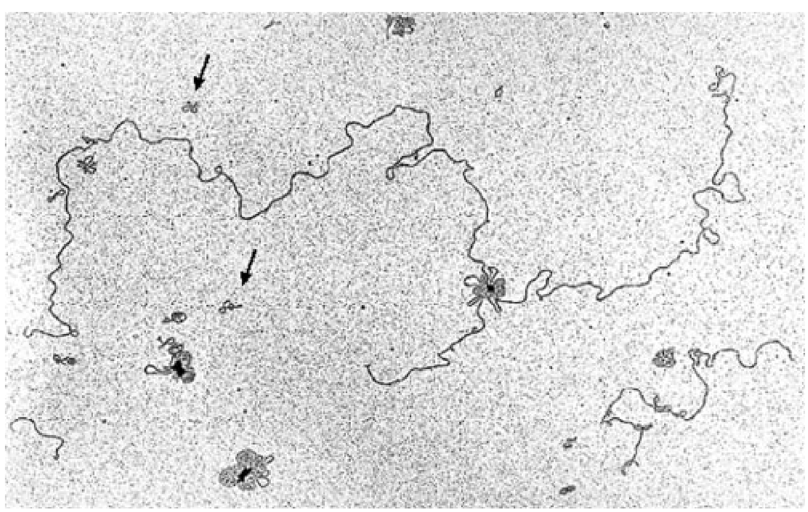


Figure 3 Electromicrograph of plasmid DNA amplified by rolling circle amplification. Image shows RCA products after 5 min amplification. Arrows indicate unit-length (nonamplified) plasmid molecules

methods. The increase in reagent cost may be offset by savings in the time, labor, and space, which can be achieved by the elimination of the bacterial growth and many liquid handling steps.

5.3. Colony PCR

Another method that bypasses culture growth is colony PCR (Gussow *et al.*, 1989). A colony is simply picked into a PCR cocktail containing primers in the flanking vector sequence designed to specifically amplify the entire insert. It is often necessary to purify the PCR product from the primers and excess nucleotides to prevent them from interfering in the sequencing reaction. Kits such as ExoSAP-IT™ (usb Corp.) contain *E. coli* exonuclease I and shrimp alkaline phosphatase to remove the single-stranded primers and free nucleotides. Colony PCR is a quick and simple method but has not been extensively used because of the amplification errors that can be introduced by the PCR process. The guidelines set out by the major sequencing centers on finishing DNA sequence (G16 Finishing Standards for the Human Genome Project – Version September 7, 2001 **http://www.genome.wustl.edu/Overview/finrulesname.php?G16=1**) limit the amount of the genome that can have sequence coverage only from PCR products, and any sequence derived from PCR products must be annotated. Despite the error rate, this method can be useful for quick colony screening.

6. Factors affecting plasmid yield

Plasmid yield is dependent on many factors including type of plasmid, (high or low copy number), size of plasmid, and *E. coli* host strain. For instance, copy number can vary from approximately 1000 for pUC vectors down to less than 10 for vectors with functional copy-control. Plasmid size should be taken into account when choosing an isolation procedure. Methods such as the boiling miniprep that rely on plasmid DNA being released from the cell when lysed are not suitable for large insert plasmids (>10 kb) as the plasmids get withheld along with the chromosomal DNA. This should also be a consideration for RCA where the harsher lysis conditions required to release the plasmid may release host DNA, which will be amplified in addition to the plasmid DNA. This is especially true for large vectors such as fosmids (see below). Optimization of PCR conditions may be required for colony PCR of plasmids with inserts larger than about 2 kb. Different PCR conditions may be required for vectors with different sized insert.

For alkaline lysis–based methods, including the filter-based methods, ideal yield from a 1.3-ml culture of a high-copy plasmid such as pUC is approximately 7 μg, although there is considerable sample to sample variability. With RCA, the yield is dependent on the amount of nucleotide in the reaction. Currently, two DNA amplification kits are available commercially from GE Healthcare that produce either 1.5 μg in about 4 h or 3.5 μg in 18 h. These reactions can be scaled up or down if more or less DNA is needed. With the SprintPrep DNA purification kit from Agencourt Biosciences, 150 μl of culture yields about 400 ng of plasmid DNA.

7. Preparation of M13-based vector DNA

Purification of M13-vectors is much simpler than for plasmids because the M13 phage particles are released into the growth media. Cells are pelleted by centrifugation, and the phage particles precipitated from the supernatant using PEG (CAS # 25322-68-3) and salt. The M13 DNA is released from the coat protein during the denaturation steps of cycle sequencing so no further purification is necessary for sequencing quality M13 DNA. If ultra pure M13 DNA is required, the DNA can be further purified by alcohol precipitation and washing. RCA and PCR can be used to amplify M13 templates directly from plaques. The product of both methods is double stranded and can immediately be sequenced from both the forward and reverse strands.

8. Preparation of fosmid and BAC DNA

The difficulty with purification of large vector constructs is twofold. First, they are usually present in only one or two copies per cell (although high copy-number vectors are recently available) and second, they are much larger than subclones, making purification based on size more difficult. There are many different protocols available for the isolation of BAC DNA, depending on the purity of DNA required. For sequencing purposes, some chromosomal DNA contamination is acceptable, but if the same DNA is to be used for fingerprinting, then a method that gives higher purity DNA may be required.

Alkaline lysis is used to purify BACs and fosmids as they are maintained in *E. coli* as supercoiled episomes. Owing to the increased size of BAC and fosmid constructs compared to plasmid subclones, some BAC DNA inevitably complexes with the SDS, protein, and chromosomal DNA, resulting in low yields. Some protocols allow the samples to stand for 30 min after the addition of the potassium acetate, presumably to allow time for the large construct DNA to reanneal. Depending on the level of purity required, either alcohol precipitation or cesium chloride gradient centrifugation can be performed following neutralization to improve DNA quality.

Filter-based methods are also available for purification of large constructs. As for plasmid purification kits, they are based on the alkaline lysis method followed by membrane purification in place of alcohol precipitation. Many employ the same glass fiber membranes used for plasmid isolation, while others such as the Montage BAC_{96} Miniprep Kit (Millipore) use size exclusion membranes. A 96-well plate of cultures can be processed in approximately 60 min with these vacuum or centrifuge-based filtration systems. Typical yields range from 0.5 to 1 μg from a 1-ml culture.

RCA can be used to amplify large constructs giving a much higher yield of DNA than alkaline lysis–based methods. Approximately 5 μg of DNA can be obtained with TempliPhi Large Construct DNA Amplification kit (GE Healthcare) in 18 h from 1 ng of starting DNA. Random hexamers in the kit will amplify any DNA in the reaction so that higher levels of chromosomal DNA are often present in the amplification product if purified BAC DNA is not used as the starting material. As a result, and because of the large size of BAC clones, a higher concentration

of DNA may have to be used in the sequencing reaction, and the RCA product is not ideal for library construction. The advantage of the method is that virtually any form of DNA can be the starting material, such as glycerol stocks or colonies, eliminating the need for culture growth.

9. Summary

The quality of sequencing template DNA directly affects the quality of sequence data obtainable. There are many template preparation methods available but no single method is perfect for all choices of vector, host strain, and sequencing application. Alkaline lysis remains the most popular method for isolating plasmid DNA but this and other inexpensive methods tend to be more time consuming, are difficult to automate, and suffer from low and variable yields. Consistent yield is the main concern when sequencing large numbers of templates, or when using a wide variety of vectors or hosts. It is difficult to establish a high-throughput sequencing pipeline when template yields are inconsistent. The column- or filter-based DNA purification methods offer higher yields and a higher purity product but still require many time-consuming steps and are subject to the same sample variability issues. These products are more expensive than simple miniprep methods. The SPRI technology eliminates many of the laborious steps of the traditional methods and as such is one of the quickest DNA purification methods available. Amplification technologies currently offer the most consistent yield and greatest flexibility for sequencing template preparation. RCA may be the method of choice when reliability, despite variation in vector and host strain, is of paramount importance. The RCA method eliminates culturing and purification steps, which can make it an attractive alternative despite higher initial costs. As DNA quality, quantity, and consistency vary between methods and with differences in host strain and vector, the choice of method has to be carefully considered, and more than one method may be required to meet all the sequencing template preparation needs.

Further reading

Elkin JE, Richardson PM, Fourcade HM, Hammon NM, Pollard MJ, Predki PF, Glavina T and Hawkins TL (2001) High-throughput plasmid purification for capillary sequencing. *Genome Research*, **11**, 1269–1274.

Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ and de Jong PJ (2001) A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Research*, **11**(3), 483–496.

References

Birnboim HC and Doly J (1979) A rapid alkaline extraction method for screening recombinant plasmid DNA. *Nucleic Acid Research*, **7**, 1513–1523.

Blanco L, Bernard A, Lazaro JM, Martin G, Garmendia C and Salas M (1989) Highly efficient DNA synthesis by phage Phi29 DNA polymerase. Symmetrical mode of DNA replication. *The Journal of Biological Chemistry*, **264**, 8935–8940.

Chissoe SL, Marra MA, Hillier L, Brinkman R, Wilson RK and Waterstone RH (1997) Representation of cloned genomic sequenced in two sequencing vectors: Correlation of DNA sequence and subclone distribution. *Nucleic Acids Research*, **25**, 2960–2966.

Dean FB, Nelson JR, Giesler TL and Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*, **11**, 1095–1099.

Dunning AM, Talmud P and Humphries SE (1988) Errors in the polymerase chain reaction. *Nucleic Acids Research*, **16**(21), 10393.

Esteban JA, Salas M and Blanco L (1993) Fidelity pf phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *The Journal of Biological Chemistry* **268**(4), 2719–2726.

Gussow D and Clackson T (1989) Direct clone characterization from plaques and colonies by the polymerase chain reaction. *Nucleic Acids Research*, **17**, 4000.

Hawkins TL, O'Connor-Morin T, Roy A and Santillian C (1994) DNA purification and isolation using solid phase. *Nucleic Acids Research*, **22**(21), 4543–4544.

Holmes DS and Quigley M (1981) A rapid boiling method for the preparation of bacterial plasmids. *Analytical Biochemistry*, **114**, 193–197.

Kim UJ, Shizuya H, de Jong PJ, Birren B and Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Research*, **20**(5), 1083–1085.

Osoegawa K, Tateno M, Woon PY, Frengen E, Mammoser AG, Catanese JJ, Hayashizaki Y and de Jong PJ (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Research*, **10**(1), 116–128.

Roach JC, Boysen C, Wang K and Hood L (1995) Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, **26**, 345–353.

Wilson RK, Koop BF, Chen C, Halloran N, Sciammis R and Hood L (1992) Nucleotide sequence analysis of 95 kb near the 3′ end of the murine T-cell receptor α/β chain locus: Strategy and methodology. *Genomics*, **13**, 1198–1208.

Yanisch-Perron C, Vierira J and Messing J (1985) Improved M13 phage cloning vectors and host strains: Nucleotide sequencing of M13mp18 and pUC19 vectors. *Gene*, **33**(1), 103–119.

Specialist Review

Robotics and automation

Elaine R. Mardis
Washington University School of Medicine, St. Louis, MO, USA

1. Introductory remarks

Reflecting on the past 15 years of the history of large-scale DNA sequencing, it is incredible to realize just how far this discipline has progressed in a relatively short period of time. The efforts, hopes, fears, and failures of many molecular biologists, engineers, and others have been played out in the arena of academic and industrial labs – all pursuing the advancement of the field. Nowhere have their efforts been more critical to these achievements than in the incorporation of robotics and automation to render tasks once entrusted to skilled technicians into the routine, programmed movements of robotic systems. This overview aims to trace the early origins of those efforts and their metamorphosis over time, as well as to present a state-of-the-art picture of high-throughput DNA sequence production for the interested reader.

2. Overview: critical components that rendered DNA sequence an "automatable" process

Before one can begin to chart the history of robotics and automation in large-scale DNA sequencing efforts, however, it is important to point out many of the factors that played a role in rendering DNA sequencing an "automatable" process. First and foremost, sequencing developed, over time, into a routine process that utilized the same series of steps and the same components at each iteration. One enormous contributor to the routine nature of sequencing reactions was encompassed by the development of cycled sequencing – an offshoot of PCR, in which a single primer is extended by a thermostable DNA polymerase on a template in the presence of dNTPs and ddNTPs, producing a great excess of dideoxy terminated fragments (in comparison to the input template amount) and eliminating stepwise addition of DNA sequencing reagents as in the past (McBride *et al*., 1989; Craxton, 1993). Later, a series of substantial improvements to DNA sequencing methods (enzymology and DNA fragment labeling in particular) and to associated methods such as DNA template preparation further enhanced the ability to automate reaction assembly by robotic means. Several of these key developments are highlighted below. Another important contribution to the routine nature of DNA sequencing

was that the detection hardware, DNA sequencing instruments, also improved significantly over time. A third development was the adaptation, borrowed from the clinical laboratory, of the multiwell microtiter plate format for sample handling, which provided a reproducible footprint and well-to-well spacing for access and plate handling by robotic devices. Along the way, industrial practices such as bulk solution manufacturing and quality control were developed and implemented by many high-throughput DNA sequencing facilities, contributing a reliability factor to the reagents used in these processes. The combination of these factors, along with the ever-increasing scale of DNA sequencing throughput became such that only robotic solutions were appropriate, especially in their elimination of manual errors, their ease of sample tracking (via bar code reading/entry), and the reproducible nature of their product.

3. Early days – the "do-it-yourself" era

When we and others first started thinking about automating DNA sequencing and its associated processes in mid-1980, suffice it to say that there were very few instrument manufacturers thinking about the same things. This situation improved slowly over time, but many first efforts at "automation" were fairly rudimentary, not truly "automated", and nearly always enlisted devices that were borrowed from other disciplines or were devised by genome center–associated engineering teams responding to demands for increased throughput. Predominantly, early efforts were aimed at applying robotics to individual processes in the DNA sequencing workflow rather than at the development of larger, integrated systems.

One of the first areas to be pinpointed for robotics application was that of liquid transfer, since reliable, reproducible manual liquid transfer of microliter volumes is a function of the quality of the pipettor and the skill of the person using it (not to mention adjunct factors of fatigue and interruption!). Most commercially available liquid-handling robots were designed for clinical applications that utilized the multiwell microtiter plate format. Hence, the adaptation of 96 well microtiter plates into DNA sequencing methods likely was a happenstance of the need for robotic liquid handling. Although many of the robots were standalone devices, they brought a reliability and reproducibility aspect to the sequencing process and enhanced throughput at very early stages. For example, our Center used the Robbins Hydra96™ pipettors to both prepare DNA templates (Mardis, 1994) and to pipette sequencing reactions for many years, and very early reports of sequencing reaction and DNA isolation methods used the Beckman Biomek 1000 robot (Wilson *et al*., 1988; Mardis and Roe, 1989; Koop *et al*., 1990).

Upstream of the DNA isolation and sequencing processes, harvesting recombinant clones from agar plates was another process that was targeted early on for automation, for quite obvious reasons – the process initially was done by technicians whose tools consisted of lightboxes (for imaging the plaques or colonies) and beakers of sterile wooden toothpicks (for harvesting the plaques or colonies). Several robots were designed and built, both in companies and academic labs (Panussis *et al*., 1996), for the purpose of picking M13 plaques and/or plasmid-containing colonies from agar plates. These robots typically combined a vision system for

imaging plates, an algorithmic selection of "pickable" plaques or colonies, a picking mechanism that contacted each plaque/colony to harvest it, and an axis/gantry system that precisely located the picking device over the plaque/colony to be harvested and over the microtiter plate well to be inoculated. The implementation of these devices into high-throughput sequencing labs provided an automated solution to a very tedious process and enabled a greatly enhanced scale of operations, not to mention making a significant dent in wooden toothpick sales worldwide.

The advent of large-scale cycle sequencing necessitated improvements to thermal cycler design – faster temperature transitions, smaller footprints, and multiple blocks were all developed to address the needs of high-throughput labs. Again, this was an area where both academic and commercial entities contributed their own designs and concepts, most of which again centered around the 96 well (and later 384) microtiter plate format. Hand in hand with thermal cycler development was the genesis of injection-molded microtiter plates that included the use of plastics with low thermal deformation and improved heat transfer characteristics, as well as design aspects that enabled both manual and robotic handling.

4. Impact of industrial/academic collaborations

As mentioned previously, there were a number of key discoveries or applications of existing technology that provided further ease of automation for DNA sequencing and related processes. Many of these resulted from efforts in academic labs funded to provide technology development for the Human Genome Project (HGP) by federal agencies such as NIH and DOE. These discoveries, in turn, were further developed into commercial products by companies that either became interested in, or were created to supply reagents, instrumentation, and develop technology for DNA sequencing.

Tabor and Richardson (1995) published a report of a mutated Taq polymerase with a single amino acid change in its nucleotide-binding site that mirrored the amino acid present at this position in the viral T7 polymerase. This single amino acid change effectively eliminated the polymerase's incorporation bias of dNTPs over ddNTPs, allowing a significant reduction in the ddNTP concentrations in sequencing reactions. Also in 1995, Ju and Mathies reported the use of fluorescence resonance energy transfer (FRET) technology in the design of base-specific dye primers (Ju *et al*., 1995a,b), effectively transforming a technique long used in protein structure determination into a fluorescent labeling strategy for modern day DNA sequencing applications. Energy Transfer ("ET") dye primers, and later "Big Dye" dye terminators (Lee *et al*., 1997), significantly enhanced DNA sequencing by dramatically reducing the amount of input template required to produce high quality sequence patterns, based on the enhanced quantum yield obtained from energy transfer from donor to acceptor dyes. Ultimately, the combination of FRET-based ("Big Dye™") dye terminators and the active site-modified thermostable polymerase yielded the most readily automated combination of technologies, and presently experiences widespread use in automated DNA sequencing. Namely, dye terminators enable one reaction per template (instead of four reactions when using

dye primers), the modified enzyme allows faster thermal cycling (fewer cycles, shorter extension times), and easier reaction cleanup to remove unincorporated terminators (since the ratio of fluorescent labeled ddNTP terminators to dNTPs is significantly reduced).

In addition to the commercialization of these key sequencing chemistry improvements, during the time period 1997–2001, many large-scale sequencing centers partnered with commercial robotics vendors to devise custom robotics solutions for specific high-throughput processes (Oefner *et al*., 1996; Marziali *et al*., 1999). Several of these robots, or scaled-down versions, subsequently transitioned into products for the sequencing market. In general, these robots either used a centrally positioned, articulated robotic arm or linear conveyor belts with gantry-mounted grippers to position microtiter plates onto various stations (liquid handling, plate sealing, mixing, lidding/delidding, etc.) on the workspace, and utilized scheduling software to keep all stations as occupied as possible in order to maximize throughput.

It is important to point out the critical contribution of sample tracking via barcodes and databases to these efforts. Without these tools, large numbers of samples would require manual logging as they progress through the DNA workflow, ultimately limiting the throughput.

Ultimately, DNA sequencing technology implementation is only as good as the instrument on which separation and excitation/detection occurs. As such, much of the ability to automate DNA sequencing was enabled by the evolution of DNA sequencing instrumentation, which has been considerable since commercial introduction of the Applied Biosystems 370 A in mid-1980 (Mardis and Roe, 1989). Early sequencing instruments, while automated in terms of laser scanning, fluorescence detection, and data analysis, required much user interaction including gel pouring, gel loading, and manual assignment of sample lanes ("tracking"). Developments that enhanced the automation of these instruments included replacing slab gels with capillaries (eliminating gel pouring and tracking), capillary illumination with a nonscanning laser beam (eliminating the time required for side-to-side scanning), and providing onboard robotics that scan barcodes and automate loading directly from a microtiter plate stacker (eliminating nearly all manual intervention) (Mardis, 1999).

5. Advent of high-throughput integrated DNA sequencing automation systems

The recent commercial introduction of DNA sequencing instruments with very high sensitivity has enabled the latest round of integrated robotics that approach full automation – that elusive ideal often referred to as *DNA in, sequence out*. In some of these efforts, the conventional, microtiter plate-based approach has been abandoned in favor of capillary tubes (an ironic twist of fate, since originally glass capillaries were used for DNA sequencing reaction vessels prior to the introduction of polypropylene microcentrifuge tubes). In other efforts, such as our own, 384 well microtiter plates are being used to prepare a sufficient amount of DNA for

one sequencing reaction in each well and then directly sequencing that DNA in the same plate. Regardless of the approach, the combination of enzymology, fluorescent labeling technology, submicroliter pipetting capability, rapid thermal cycling (due to enhanced thermal transfer and fewer cycles for small volumes), and detection sensitivity are making more fully integrated robotics approaches possible.

These efforts fall short of a fully automated approach since some preliminary work is required to grow the subclones in culture and since the sequencing products ultimately are loaded and detected on a separate instrument. As yet, the elusive goal of full automation continues to be the focus of many academic and commercial efforts to miniaturize the DNA preparation, sequencing, separation, and detection processes, by a variety of approaches. The most successful of these efforts should provide us with the next generation of automated DNA sequencing instrumentation systems, hopefully in the next few years.

References

Craxton M (1993) Cosmid sequencing. *Methods in Molecular Biology*, **23**, 149–167.

Ju J, Kheterpal I, Scherer JR, Ruan C, Fuller CW, Glazer AN and Mathies RA (1995a) Design and synthesis of fluorescence energy transfer dye-labeled primers and their application for DNA sequencing and analysis. *Analytical Biochemistry*, **231**(1), 131–140.

Ju J, Ruan C, Fuller CW, Glazer AN and Mathies RA (1995b) Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(10), 4347–4351.

Koop BF, Wilson RK, Chen C, Halloran N, Sciammis R, Hood L and Lindelien JW (1990) Sequencing reactions in microtiter plates. *Biotechniques*, **9**(1), 32, 34–37.

Lee LG, Spurgeon SL, Heiner CR, Benson SC, Rosenblum BB, Menchen SM, Graham RJ, Constantinescu A, Upadhya KG and Cassel JM (1997) New energy transfer dyes for DNA sequencing. *Nucleic Acids Research*, **25**(14), 2816–2822.

Mardis ER (1994) High-throughput detergent extraction of M13 subclones for fluorescent DNA sequencing. *Nucleic Acids Research*, **22**(11), 2173–2175.

Mardis ER (1999) Capillary electrophoresis platforms for DNA sequence analysis. *Journal of Biomolecular Techniques*, **10**, 137–147.

Mardis ER and Roe BA (1989) Automated methods for single-stranded DNA isolation and dideoxynucleotide DNA sequencing reactions on a robotic workstation. *Biotechniques*, **7**(8), 840–850.

Marziali A, Willis TD, Federspiel NA and Davis RW (1999) An automated sample preparation system for large-scale DNA sequencing. *Genome Research*, **9**(5), 457–462.

McBride LJ, Koepf SM, Gibbs RA, Salser W, Mayrand PE, Hunkapiller MW and Kronick MN (1989) Automated DNA sequencing methods involving polymerase chain reaction. *Clinical Chemistry*, **35**(11), 2196–2201.

Oefner PJ, Hunicke-Smith SP Chiang L, Dietrich F, Mulligan J and Davis RW (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Research*, **24**(20), 3879–3886.

Panussis DA, Stuebe ET, Weinstock LA, Wilson RK and Mardis ER (1996) Automated plaque picking and arraying on a robotic system equipped with a CCD camera and a sampling device using intramedic tubing. *Laboratory Robotics and Automation*, **8**, 195–203.

Tabor S and Richardson CC (1995) A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(14), 6339–6343.

Wilson RK, Yuen AS, Clark SM, Spence C, Arakelian P and Hood LE (1988) Automation of dideoxynucleotide DNA sequencing reactions using a robotic workstation. *Biotechniques*, **6**(8), 776–777, 781–787.

Short Specialist Review

Microelectrophoresis devices for DNA sequencing

Daniel J. Ehrlich, James H. Aborn, Sameh A. El-Difrawy, Elizabeth A. Gismondi, Roger Lam, Brian McKenna and Thomas O'Neil
Whitehead Institute, Cambridge, MA, USA

1. Introduction

Extensive applications in genomics and biomedicine remain inaccessible because of the current cost of DNA sequencing technology. Commercial capillary arrays enabled the Human Genome Project but currently define the, often, prohibitive cost structure of sequencing. Microelectromechanical systems (MEMS) have been proposed as one technical approach to a significant near-term advance.

The often-cited intrinsic advantages of the microdevice approach are the practical engineering of cross-channel sample injectors and very high density networks. The basic capability cross-injector is illustrated in Figure 1. Because the volume of sample that is caught in that area is very small, microfabricated devices have the ability to scale the sample consumed in a given experiment far below that of capillary devices.

The cost savings that would accrue from a practical implementation of nanoliter sample handling and injection would be an advance for biological applications. Much research is active in this area. The second main aspect of leverage for microdevice (electrophoretic) sequencing is massively parallel networks that would extend the productivity of systems beyond the, typically 96-lane, capability of commercial capillary arrays. Some substantial capability for MEMS electrophoresis systems was demonstrated in short devices by Paegel *et al*. (2002) (430-base reads in a compact 96-channel device) and Liu *et al*. (2000) (450-base reads in a 16-channel device). However, the cost model for de novo sequencing is very sensitive to read length, particularly when assembly costs are considered. Hence, for this application, the compromise in read length in short microdevices (when these devices are compared to capillary instruments) is not acceptable. Short microdevices might be considered for resequencing and genotyping applications. Koutny *et al*. (2000) demonstrated long read length (800-base reads in a single-channel device), however, in a very simple single-channel device.

Therefore, although some remarkable capability was demonstrated in the early systems, they did not attempt the combined parallelism, read length, and automation required to challenge the commercial capillary array machines. The current state

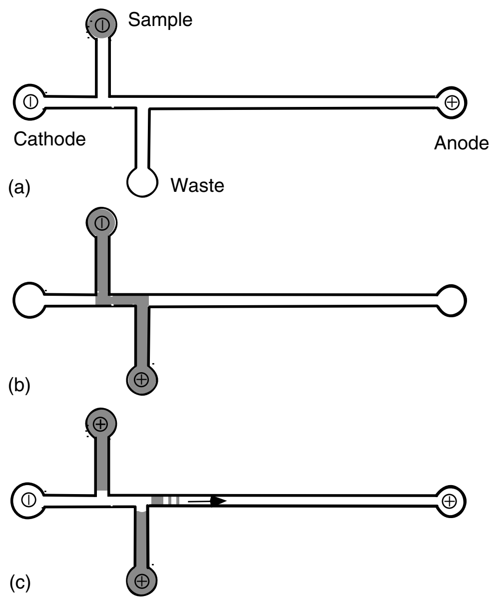


Figure 1 Schematic operation of a cross-injector used for DNA sample loading into a typical microdevice. The DNA sample is electrophoresed from the sample well to the waste well traveling, at least in part, through the separation channel. When the optimal loading time has been achieved, the voltages are switched, with the field now applied from one end of the separation channel to the other. This captures the DNA fragments that were in the separation channel at the time of the switch and begins the separation and detection process

of the art is now a 768-channel machine, which is in the final stages of testing and which is designed for exactly this purpose. The much more complex microdevice element is illustrated in Figure 2.

The factors that are most important in microdevice DNA sequencing are resolution and read length (Figure 3). The latter quantity includes consideration of the sample injection timing and the electrophoresis conditions. The ability to identify adjacent DNA fragments, particularly for large fragments, directly determines the performance and throughput characteristics of a given instrument. Because of the central importance of these parameters, each part of the instrument design is affected. The channel architecture cannot be permitted to add significant band broadening, and dense networks need to be robust to instabilities that occur in the network. Hence, the extension of single-channel performance to a dense network is nontrivial for long read applications.

2. Microdevice design and layout

A successful microdevice design must not only optimize conditions for fragment resolution and read length but must also take a number of other factors into account.The device design must have adequate space for the cross-injector elements, must be capable of interfacing with commercial sample introduction instruments (pipettors, capillary arrays, microtiter plates), must have channel separation lengths

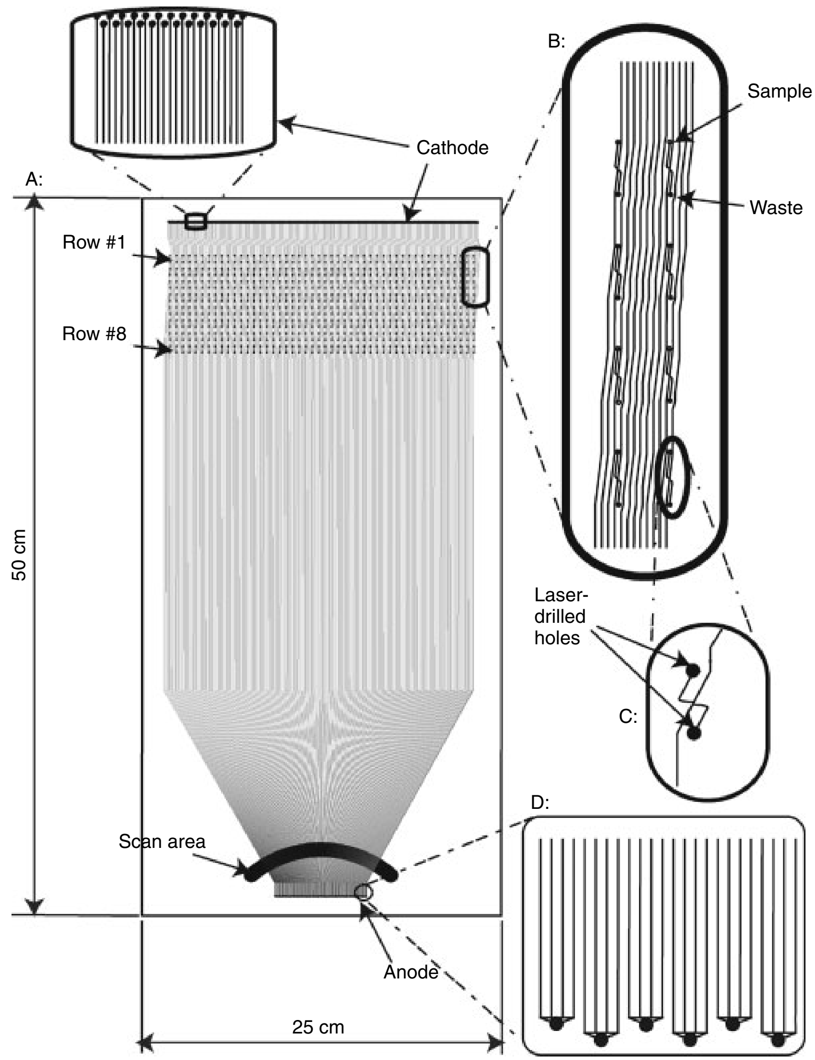


Figure 2 The plates contain 384 channels with double T cross-injectors and separation lengths varying from 37 to 45 cm. As the channels approach the detection window, they curve inward to achieve maximum channel density (~6 lanes/mm) during detection. This design provides full cross-injectors for all lanes, direct compatibility with commercial multitip pipettors, separation lengths optimized for long read performance, and channel geometry, which maximizes separation ability and lifetime

that allow for high quality, long read performance, and must minimize changes in channel direction, width, and/or cross section to maximize fragment resolution and device lifetime. Several devices have been constructed to this end.

A 16-channel device has been constructed that maintains channel length and minimizes the total size by allowing narrow turns in the device's separation channels. It has been reported that this device is capable of 450-base high-quality reads (Liu, 2000). Another device, a 96-lane microchip, does not have as extensive

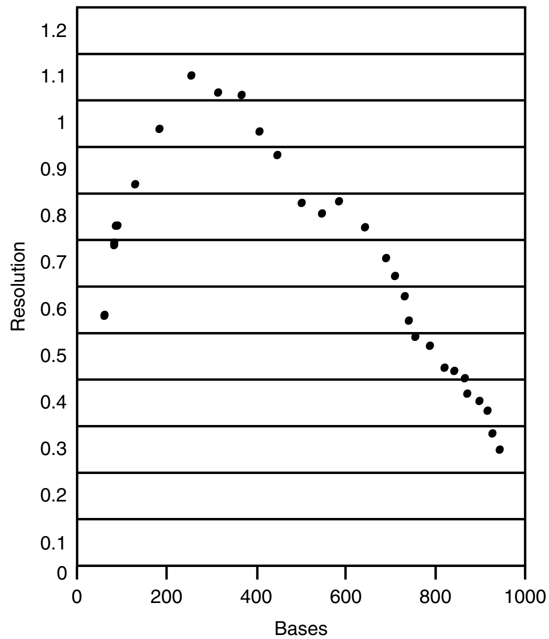


Figure 3 Resolution curve from a microfabricated single-channel sequencing device comparable to the high-quality performance of capillary machines. (Reprinted with permission from Koutny *et al*. (2000). ©2000 American Chemical Society)

separation lengths, but uses a radial design to maximize throughput capability and channel geometry consistency (Paegel, 2002). The 768-lane device mentioned earlier achieves long read length, >580 bases, and increased parallelism by using a large, 50 by 25 cm, substrate for device fabrication.

3. Microchip fabrication

These devices are created using standard and slightly modified microfabrication techniques on glass substrates described previously (Becker, 2000). Briefly, the glass substrates are coated with specialized photoresist and patterned via direct-write photolithography. Following this, the channels are etched into and access holes are drilled through the glass. Finally, two pieces of glass substrate are thermally fused.

4. Microchannel coating

Before the electrophoresis stage, the inner surfaces of the channels in the glass microdevice are usually chemically passivated to reduce or eliminate electroosmotic flow. Many different strategies have developed that often rely on the earlier work of Hjertén (1985) to form polymeric coatings for surface modification (Cobb, 1990; Fung, 1995; Madabhushi, 1998). These polymers can be attached to the channel wall either covalently or noncovalently. For glass microdevices, both covalent (Liu, 2000; Woolley, 1997; Shi, 1999) and noncovalent (Simpson, 2000; Backhouse, 2000) methods have been pursued extensively.

5. Separation media

The resolution and read length for a device are largely dependent on the media used for the fragment separation. Linear polyacrylamide (LPA) is used extensively as the separation matrix and is known for having high-quality sequencing ability (Koutny, 2000; Liu, 2000; Paegel, 2002). Several other media have been tested with varying degrees of success including polyethylene oxide (PEO) (Fung, 1995), poly(vinyl pyrrolidone)(PVP) (Song, 2001), hydroxycellulose (HEC) (Bashkin, 1996), and numerous mixed molecular weight and/or self-coating acrylamide copolymers (Sassi, 1996; Menchen, 1996; Dolnik, 2000).

6. Detection

Most microdevices for DNA sequencing use laser-induced fluorescence detection of tagged DNA fragments for detection. Many devices have a modified confocal scanner capable of exciting the fragments and collecting the emitted light (Dolnik, 2000; Woolley, 1997; Shi, 1999; Goedecke *et al*., 2004). Other detection systems use acousto-optical deflection-based laser beam scanners (Huang, 1999) and CCD cameras (Simpson, 2000).

Because of their inherent advantages, microdevices seem to be the next technological advance in high throughput DNA sequencing. Although there are certain drawbacks, such as the complex electrical networks being created, the sheer potential for parallelism and reduction in reagent costs are enormous.

References

Backhouse C, Caamano M, Oaks F, Nordman E, Carrillo A, Johnson B and Bay S (2000) DNA sequencing in a monolithic microchannel device. *Electrophoresis*, **21**, 150–156.

Bashkin J, Marsh M, Barker D and Johnston R (1996) DNA sequencing by capillary electrophoresis with a hydroxyethylcellulose sieving buffer. *Applied and Theoretical Electrophoresis*, **6**, 23–28.

Becker H and Gartner C (2000) Polymer microfabrication methods for microfluidic analytical applications. *Electrophoresis*, **21**, 12–26.

Cobb KA, Dolnik V and Novotny M (1990) Electrophoretic separations of proteins in capillaries with hydrolytically-stable surface structures. *Analytical Chemistry*, **62**, 2478–2483.

Dolnik V, Liu S and Jovanovich S (2000) Capillary electrophoresis on microchip. *Electrophoresis*, **21**, 41–54.

Fung EN and Yeung ES (1995) High-speed DNA sequencing by using mixed poly(ethylene oxide) solutions in uncoated capillary columns. *Analytical Chemistry*, **67**, 1913–1919.

Goedecke N, McKenna B, El-Difrawy S, Carey L, Matsudaira P and Ehrlich D (2004) A high-performance multilane microdevice system designed for the DNA forensics laboratory. *Electrophoresis*, **25**, 1678–1686.

Hjertén S (1985) High-performance electrophoresis: elimination of electroendosmosis and solute adsorption. *Journal of Chromatography*, **347**, 191–198.

Huang Z, Munro N, Huhmer AF and Landers JP (1999) Acousto-optical deflection-based laser beam scanning for fluorescence detection on multichannel electrophoretic microchips. *Analytical Chemistry*, **71**, 5309–5314.

Koutny L, Schmalzing D, Salas-Solano O, El-Difrawy S, Adourian A, Buonocore S, Abbey K, McEwan P, Matsudaira P and Ehrlich D (2000) Eight hundred-base sequencing in a microfabricated electrophoretic Device. *Analytical Chemistry*, **72**, 3388–3391.

Paegel B, Emrich C, Wedemayer G, Scherer J and Mathies RA (2002) High throughput DNA sequencing with a microfabricated 96-lane capillary array electrophoresis bioprocessor. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 574–579.

Liu S, Ren H, Gao Q, Roach DJ, Loder RTJ, Armstrong TM, Mao Q, Blaga I, Barker DL and Jovanovich SB (2000) Automated parallel DNA sequencing on multiple channel microchips. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5369–5374.

Madabhushi RS (1998) Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis*, **19**, 224–230.

Menchen S, Johnson B, Winnik MA and Xu B (1996) Flowable networks as DNA sequencing media in capillary columns. *Electrophoresis*, **17**, 1451–1459.

Sassi AP, Barron A, Alonso-Amigo MG, Hion DY, Yu JS, Soane DS and Hooper HH (1996) Electrophoresis of DNA in novel thermoreversible matrices. *Electrophoresis*, **17**, 1460–1469.

Shi Y, Simpson PC, Scherer JR, Wexler D, Skibola C, Smith MT and Mathies RA (1999) Radial capillary array electrophoresis microplate and scanner for high-performance nucleic acid analysis. *Analytical Chemistry*, **71**, 5354–5361.

Simpson JW, Ruiz-Martinez MC, Mulhern GT, Berka J, Latimer DR, Ball JA, Rothberg JM and Went GT (2000) A transmission imaging spectrograph and microfabricated channel system for DNA analysis. *Electrophoresis*, **21**, 135–149.

Song JM and Yeung ES (2001) Optimization of DNA electrophoretic behavior in poly(vinyl pyrrolidone) sieving matrix for DNA sequencing. *Electrophoresis*, **22**, 748–754.

Woolley AT, Sensabaugh GF and Mathies RA (1997) High-speed DNA genotyping using microfabricated capillary array electrophoresis chips. *Analytical Chemistry*, **69**, 2181–2186.

Short Specialist Review

Single molecule array-based sequencing

Simon T. Bennett and Tony J. Smith
Solexa Limited, Little Chesterford, UK

1. Introduction

Hidden within an individual's genomic sequence are the genetic instructions for the entire repertoire of cellular components that determine the complexities of biological systems. Unraveling genomic structure and characterizing the functional elements from within the code will allow connections to be made between the genetic blueprint, transcribed information, and the resulting systems biology, and will, in turn, accelerate the exploration of the biological sciences.

As pointed out in a recent review (Shendure *et al*., 2004), the vast majority of known DNA sequence data to date have been generated using the Sanger-based sequencing method. However, genotyping (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4) has been the tool most widely chosen for genetic exploration because the cost of sequencing individual genomes remains prohibitively expensive (recent estimates place the cost of sequencing a human genome in the region of tens of millions of US dollars). Technological advances in DNA resequencing, leveraging the availability of the consensus genome sequence for almost 200 species (**http://www.intlgenome.org/viewDatabase.cfm**), are transforming throughput and costs. An improvement in the region of four to five orders of magnitude over current sequencing costs is no longer an unrealistic prospect.

High-throughput sequence analysis, using capillary-electrophoretic separation and four-color fluorescent detection in instrument systems (such as the Applied Biosystems 3700/3730 and Amersham Biosciences MegaBACE 1000/4000), has been deployed successfully in factory-scale operations, largely within public-funded organizations, to sequence the human genome and that of many other species (*see* Article 5, **Robotics and automation**, Volume 3). Improvements in these systems continue to deliver incremental (maybe as high as 10-fold) increases in throughput and cost reductions. But these do not address the fundamental need for a transformation in cost-effectiveness that would be necessary for sequencing on a genome-wide scale to become a routine undertaking. Reagents are currently a highly significant cost element in current sequencing approaches and therefore a key target in cost-reduction approaches.

An initiative to address this key cost factor is being taken by the laboratory of Richard Mathies at U.C. Berkeley (Paegel *et al*., 2003). Mathies' lab is working

to achieve this goal by seeking to create an integrated microfabricated device that couples clone isolation, template amplification, Sanger extension, purification, and separation in a single microfluidic device. They highlight the development (in their lab and that of other workers) of highly parallel microfabricated capillary-array electrophoresis analyzers, nanoliter-scale DNA purification and amplification reactors, and microfluidic cell sorters and cytometers to support the feasibility of creating such an integrated microfabricated device. They calculate that the processing time could be reduced by 10-fold and reagent consumption by 100-fold, compared to the current state of the art. To go beyond this in cost reduction requires a fundamentally different strategy.

2. New sequencing approaches

There are several emerging sequencing technologies that aspire to ultralow cost, ultrahigh-throughput capabilities. Shendure *et al*. (2004) have classified these methods broadly into five different groups: microelectrophoretic methods (such as the work of Mathies and colleagues referred to above), hybridization, cyclic-array sequencing on amplified molecules, cyclic-array sequencing on single molecules, and real-time methods. While each of these approaches has potential to make the necessary breakthrough in technology, it is too early to predict whether expectations will be fulfilled. Yet, for some, and in particular the single molecule array-based approaches, recent developments have continued to stimulate community interest in sequencing technologies that have the capability to analyze entire genomes very quickly at an affordable price. This is particularly so for the human genome and the aspiration to achieve the so-called $1000 human genome concept (Zimmerman, 2004).

3. Single molecule–based approaches

Analysis at the single molecule level is challenging, yet it offers substantial advantages not only over conventional sequencing but also over other emerging technologies. Recent progress in the development of highly efficient strategies that dramatically reduce reagent consumption during analysis is bringing the routine analysis of whole-genome variation at the sequence level closer to reality. Methods under development, as reviewed by Smith (2004), fall into three main categories:

- *Single molecule separation:* Elongation of large fragments of genomic DNA that has been tagged with fluorescently labeled probes bound at specific sites, such as that being developed by OpGen (**http://www.opgen.com**) and US Genomics (**http://www.usgenomics.com**). The molecules can be analyzed at high speed, and this taken with the currently low (>1200 bp) resolution makes such techniques suited to mapping rather than sequencing.
- *Arrays of cloned single molecules:* Sequencing techniques related to high-density arrays of “colonies” of identical copies of template amplified from a single molecule, either immobilized on a solid surface (e.g., Solexa’s cluster array developed by former Swiss company Manteia) dispensed into a very high density

microtiter plate (e.g., 454 Life Sciences **http://www.454.com**), or amplified in a thin polyacrylamide gel matrix on a slide (e.g., George Church's group at Harvard; Mitra and Church, 1999).

- *Single molecule arrays:* Single molecule analysis in an array-based format to generate a massively parallel approach to sequencing, as pioneered by Solexa (**http://www.solexa.com**).

4. Single molecule array-based approaches

Single molecule array-based approaches are characterized by a number of distinct advantages over other technologies (Figure 1). In addition to minimal sample preparation and a novel sequencing chemistry, the rapid detection of single, fluorescently labeled dye molecules with very high signal-to-noise ratio is a critical feature of Solexa's Single Molecule Array™ (SMA) technology. The technology is massively parallel with an estimated 100 million single molecules of DNA sample template, dispersed randomly, per square centimeter of array. In the presence of a proprietary polymerase, specially designed nucleotides act as reversible terminators of sequencing so that, at each cycle, only a single base of DNA template is sequenced. Each of the four nucleotides is labeled with a distinguishable fluor and detected using a four-color detection system. Once the base has been identified, the block to further extension is relieved and the fluorescence removed so that the next cycle can be performed. Development of reversibly terminating nucleotides, by limiting each cycle to a single incorporation, overcomes the problem encountered by other approaches of having to decipher homopolymeric sequences and increases the accuracy of incorporation by the polymerase as all four nucleotides are present in the sequencing reaction.

The number of cycles of sequencing is dictated by the size of the genomic template that is under investigation. For example, with human resequencing, each template is sequenced to a length of 25 to 30 bases, derived from an analysis by Solexa of the human genome that revealed that unique alignment requires a read length of approximately 20 bases. Software aligns the n-mer reads against the reference sequence of the genome to identify a large part of the variation between

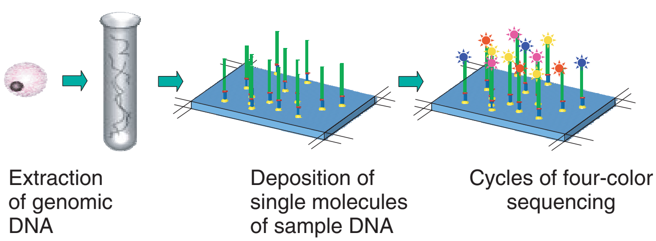


Figure 1 Single molecule array sequencing. Arrays of single molecules are created by binding randomly fragmented genomic DNA to a chip surface as primed templates. Addition of fluorescently labeled nucleotides and DNA polymerase allows sequence determination. (Reprinted from *Drug Discovery Today: Targets*, Volume **3**, Number 3, Smith T, Whole genome variation analysis using single molecule sequencing, 112–116, Copyright (2004), with permission from Elsevier)

the individual's DNA and the reference sequence. In this way, unknown SNPs as well as known SNPs can be detected and typed simultaneously at the same time as gathering data to determine haplotype structure and patterns of linkage disequilibrium. The approach is universally applicable to any organism for which a reference sequence is available, and shorter read lengths can be used where the genome, or a genome entity, such as a single chromosome, is less complex.

As the cost of sequencing per se is reduced, sample preparation will account for a significant proportion of total costs (*see* Article 4, **Sequencing templates – shotgun clone isolation versus amplification approaches**, Volume 3). Performed in a single reaction, the SMA approach does not require costly or time-consuming preparation, such as PCR amplification or cloning target DNA into bacteria. Another important advantage of SMA is the requirement only for very small quantities (picograms) of DNA starting material. This not only avoids averaging effects of using large samples, masking what is really happening in a biological system but also avoids representational bias by minimizing sample processing. These features, together with a dramatic reduction in reaction volume, combine to revolutionize the economic landscape of sequencing. Once viable economically at large scale, whole genome resequencing of each sample will enable true whole-genome association studies.

A critical consideration is that these new approaches will produce an unprecedented quantity of data, which will have to be processed, annotated, and applied. This will require an entirely new set of skills, systems, and databases, which, it is anticipated, will create an entirely new field of genomics. To this end, Solexa is working with the groups of Ewan Birney and Richard Durbin at the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, respectively, to extend and advance the *Ensembl* system (**http://www.ensembl.org**) to manage, query, and visualize multiple whole-genome sequence data sets. Furthermore, as a second strand to this project, statistical methods and tools are being developed with David Balding and colleagues at Imperial College London to allow epidemiological studies to exploit whole-genome data to localize gene effects involved in disease susceptibility and drug metabolism.

5. Arrays of cloned single molecules

There is a group of related techniques that seeks to overcome the high sensitivity required to analyze individual single molecules, by creating a high-density array of "colonies" of identical copies amplified from a single molecule (Figure 2). George Church's group at Harvard (Mitra and Church, 1999; Mitra *et al*., 2003) carry out PCR amplification in a thin polyacrylamide gel matrix on a slide to constrain lateral diffusion, thereby creating colonies of PCR products; they coined the term "polonies" to describe these. A related strategy, Manteia approach (Adessi *et al*., 2000), involved amplification of single-molecule templates immobilized on a solid surface. 454 Life Sciences (Leamon *et al*., 2003) have dispensed single molecules into a very high density microtiter plate, such that each 75-picoliter well contains no more than one molecule, and then carried out amplification. The sequences of several viral and bacterial genomes have been determined using this approach.

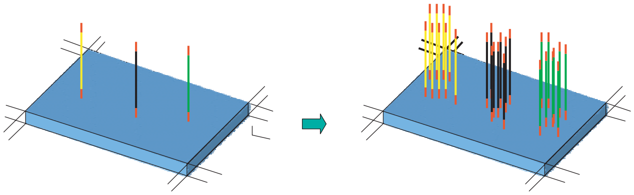


Figure 2 Arrays of cloned single molecules. Single molecules are amplified in a spatially defined way such that a large number of identical copies of each are generated in isolated "colonies". These colonies can then be subjected to sequencing in situ. (Reprinted from *Drug Discovery Today: Targets*, Volume **3**, Number 3, Smith T, Whole genome variation analysis using single molecule sequencing, 112–116, Copyright (2004), with permission from Elsevier)

These arrays of cloned single molecules are then subjected to sequencing using, for example, DNA polymerase–based incorporation of labeled nucleotides and fluorescence detection or pyrosequencing (**http://www.pyrosequencing.com**). The use of cloned single molecules facilitates detection by yielding a higher signal than an individual single molecule. In principle, this allows detection instrumentation that is relatively less sophisticated and less costly to be employed. A somewhat greater level of inefficiency in the sequencing biochemistry or loss of templates through the process can be tolerated, as the signal is derived from a large number of molecules. Balanced against these considerations, cloned single molecules can introduce problems owing to the individual molecules in a colony becoming out of phase with one another during the sequencing process and therefore creating high backgrounds and spurious signals. Other issues are the complexity and effort involved in generating the cloned array and the potential for the sequence representation of the sample not to be faithfully preserved.

6. Applications

By applying different simple methods of sample preparation and downstream analysis algorithms to the core technology, the range of capabilities of SMA technology is extended. SMA can be applied either to resequence whole genomes or to the same reproducible, specific genome sequence from several different individuals to provide a particular set number of SNP loci (e.g., a particular subset of, say, 1 million SNPs) for mapping traits (Bennett, 2004; *see also* Article 11, **Mapping complex disease phenotypes**, Volume 3 and Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3).

The primary application focus is on basic research, both in academia and in industry, where this breakthrough technology is anticipated to stimulate a new wave of research activity enabled by the newfound ability to measure variation comprehensively across whole genomes (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4). The technology will stimulate new methods of applying knowledge of individual variation with wide-ranging applications, such as is in functional/comparative genomics (*see* Article 48, **Comparative sequencing**

of vertebrate genomes, Volume 3), exploration of microbial diversity for the agricultural biology field, pathogen identification (*see* Article 49, **Bacterial pathogens of man**, Volume 4), transcriptome characterization and in particular of alternative splice variants, genotype–phenotype correlations, human and animal disease association (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2), pharmacogenomics, the development of new molecular diagnostics and drugs, and in personalized medicine. This process will begin largely in the major government-funded and not-for-profit-funded research institutes, leveraging the strong political will that exists to see real human health benefits from the large investment already made in genetics, and in particular in the Human Genome Project (*see* Article 24, **The Human Genome Project**, Volume 3) and its various ramifications.

7. Concluding remarks

Single molecule array-based sequencing technology has the potential to transform the economics of DNA sequencing by allowing the sequence of hundreds of millions of individual molecules to be determined rapidly in parallel. The approach drastically reduces, and at best obviates, the need for sorting, cloning, and amplification of genomic DNA samples with the consequential reduction in laboratory preparation and reagent overheads. Together, these facets of single molecule array-based sequencing will allow sequencing of large entities genomic, including whole genomes, at costs several orders of magnitude below current levels. For human genetics, next-generation technologies such as SMA offer the potential to achieve the much sought after $1000 human genome goal.

References

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P and Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*, **15**, e87.

Bennett S (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.

Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD, Berka J and Lohman KL (2003) A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **24**, 3769–3777.

Mitra RD and Church GM (1999) In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, **27**, e34.

Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE and Church GM (2003) Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences of the USA*, **100**, 5926–5931.

Paegel BM, Blazej RG and Mathies RA (2003) Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. *Current Opinions in Biotechnology*, **14**, 42–50.

Smith T (2004) Whole genome variation analysis using single molecule sequencing. *Drug Discovery Today Targets*, **3**, 112–116.

Shendure J, Mitra RD, Varma C and Church GM (2004) Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, **5**, 335–344.

Zimmerman Z (2004) *The $1000 Human Genome – Implications for Life Sciences, Healthcare, and IT*, IDC: Framingham.

Short Specialist Review

Real-time DNA sequencing

Susan H. Hardin
University of Houston, Houston, TX, USA

1. Introduction

DNA sequence information provides insights into a wide range of biological processes. The order of bases in DNA implies the order of bases in RNA and, consequently, the amino acid sequence of protein. DNA sequence specifies a molecular program that can lead to normal development or the manifestation of a genetic disease such as cancer. DNA sequence information also has the potential to instantly and conclusively identify a pathogen (or variation thereof), or uniquely identify and genetically characterize an individual.

The core elements required for DNA replication in a test tube include DNA polymerase, deoxynucleotide triphosphates, template, and primer in a buffer that promotes the activity. DNA synthesis occurs when the primer's 3′-end attacks the α-phosphate of the incoming nucleotide, which is complementary to the template strand. Of the three phosphates within the nucleotide, only the α-phosphate becomes part of the DNA strand. The β- and γ-phosphates (pyrophosphate, PP_i) are released into the solution.

Approximately 30 years ago, Sanger and colleagues developed a sequencing method that exploited the basic biochemistry of DNA replication (Sanger *et al*., 1977). Of particular importance for their method are the facts that DNA polymerase can incorporate a dideoxynucleotide triphosphate (ddNTP; a nucleotide analog lacking a 3′ OH), and that, once incorporated, additional nucleotide incorporation is not possible. Importantly, ddNTPs are incorporated by the polymerase using the same base incorporation rules that dictate incorporation of natural nucleotides. The reaction products are size-separated and examined to deduce the DNA sequence information.

The first human genome was sequenced using variations of Dr. Sanger's chemistry and important breakthroughs in instrumentation and process automation (Lander *et al*., 2001; Venter *et al*., 2001). This first human genome sequence has sparked a new era in genome analysis. Identifying differences in the genetic code that make each of us unique is the next challenge. Given current cost estimates of \$10–\$25 million to sequence a single human genome, it is unlikely that the large numbers of human genomes needed to identify these important differences will be completed using Sanger-based sequencing methods, and even less likely that this chemistry will be used to enable the promise of whole genome analysis for medical purposes (personalized medicine).

The desire to examine differences between genomes is so great that an industry directing analysis to regions previously associated with genetic variation has emerged. Single nucleotide polymorphism (SNP) analyses essentially involve skimming genomic information from predetermined regions owing to cost limitations and time constraints of current DNA sequencing methods. Ultrahigh throughput sequencing will enable a more comprehensive form of genetic variation detection that does not begin with assumptions. The fundamental importance of DNA sequence information drives researchers to continually strive to improve the efficiency and accuracy of sequencing methods.

2. A massively parallel, real-time sequencing strategy

We are developing a sequencing platform that will enable a more comprehensive form of genetic variation detection. Cutting-edge technologies, including single-molecule detection, fluorescent molecule chemistry, computational biochemistry, and biomolecule engineering and purification, are being combined to create this new platform. Our approach may make it easier to classify an organism or identify variations within an organism by sequencing the genome in question.

The basic biochemistry of DNA replication is being exploited in a new way to develop a radically different method to sequence DNA. DNA polymerase and nucleotides triphosphates are being engineered to act together as direct molecular sensors of DNA base identity at the single-molecule level. The general strategy involves monitoring real-time, single-pair fluorescence resonance energy transfer (spFRET) between a donor fluorophore attached to a polymerase and a color-coded acceptor fluorophore attached to the γ-phosphate of a dNTP (5′ fluorescently modified γ-phosphate) during nucleotide incorporation and pyrophosphate release (Figure 1). The purpose of the donor is to stimulate an acceptor to produce a characteristic fluorescent signal that indicates base identity (emission wavelength and intensity provide a unique signature of base identity). Equally important to our technology are the massively parallel arrays of nanomachines created to produce the unprecedented throughput of the sequencing system. Projected sequencing rates approach 1 million bases per second – rather than per day – per instrument; almost a 100 000-fold increase over current throughput.

The sequencing platform incorporates a laser that is tuned to excite the donor fluorophore. A spFRET-based strategy increases signal-to-noise by minimizing acceptor emission until the acceptor fluorophore is sufficiently close to a donor fluorophore to accept energy. Incorporating total internal reflectance fluorescence (TIRF) into the platform further increases signal-to-noise, since most of the labeled dNTPs in solution are not within the TIRF excitation volume and are, therefore, not directly excited by the incident light.

As an acceptor-labeled dNTP approaches the donor-labeled polymerase, it begins to emit its signature wavelength of light owing to energy transfer from the donor (they participate in spFRET), and the intensity of this fluorescence increases throughout the nucleotide's approach. The molecules are engineered to maximally

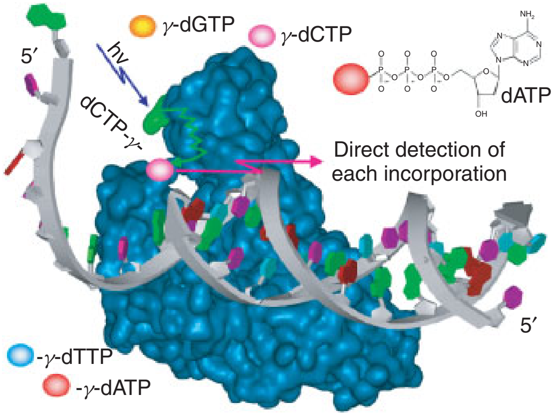


Figure 1 Real-time detection of dNTP incorporation. Components of the VisiGen Sequencing System include modified polymerase, color-coded nucleotides, primer, and template. Energy transfers from a donor fluorophore within polymerase to an acceptor fluorophore on the γ-phosphate of the incoming dNTP, stimulating acceptor emission, fluorescence detection, and incorporated nucleotide identification. Fluorescently tagged pyrophosphate leaves the complex, producing natural DNA. This nonserial approach enables rapid detection of subsequent incorporation events. Time-dependent fluorescent signals emitted from each complex are monitored in massively parallel arrays and analyzed to determine DNA sequence information. Courtesy of Dr. Tommie Lincecum, VisiGen Biotechnologies, Inc

FRET after the acceptor-dNTP docks at the active site of the polymerase (within the nucleotide binding pocket). During nucleotide insertion, the 3′ end of the primer attacks the α-phosphate within the dNTP, cleaving the bond between the α- and β-phosphates, and changing the spectral properties of the fluorophore (which remains attached to the PP_i). Donor fluorescence is also informative, as it undergoes anticorrelated intensity changes throughout the incorporation reaction. As the acceptor-tagged PP_i is released from the polymerase, the distance between it and the donor fluorophore increases, causing the intensity of the acceptor's fluorescence to decrease and that of the donor's to simultaneously increase. After an spFRET event, the donor's emission returns to its original state and is ready to undergo a similar intensity oscillation cycle with the next acceptor-tagged nucleotide. In this way, the donor fluorophore acts as a punctuation mark between incorporation events. The increase in donor fluorescence between incorporations is especially important during analysis of homopolymeric sequences.

Acknowledgments

Development of VisiGen's sequencing platform has been funded by DARPA and NIH in support of national defense (to identify pathogenic organisms or variations thereof) and to enable comprehensive genome analysis (basic science discovery and personalized medicine), respectively. Discussions with VisiGen development teams are gratefully acknowledged.

References

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA and Holt RA (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Introductory Review

Genome mapping overview

Wesley C. Warren and Michael Lovett
Washington University School of Medicine, St. Louis, MO, USA

1. Introduction

True physical maps, for example, fingerprint maps in which each clone is "fingerprinted" on the basis of the pattern of fragments generated by a restriction enzyme digest, of large-insert molecular clones have had a tremendous impact upon large-scale genomic DNA sequencing. The physical map of the human genome (McPherson *et al*., 2001) provided the necessary scaffolding for accurate final compilation of the human genomic DNA sequence. Computationally deconvoluting an entire shotgun sequence of small DNA fragments from large genomes is challenging, some more so than others. These final assembly steps are immeasurably helped by the availability of an ordered set of markers or molecular clones. Reference sets of ordered molecular clones have also impacted many other fields. For example, the use of mapped bacterial artificial clones (BACs) for array-based comparative genomic hybridization (CGH) (Ishkanian *et al*., 2004), and the use of mapped BACs to identify chromosomal rearrangement breakpoints (Volik *et al*., 2003) are just two of the many mapping spin-offs that have become widely used tools for disease analysis and gene discovery.

2. Genetic maps

Genetic linkage maps have provided a key framework for all of the mammalian genomes that currently have published draft DNA sequence assemblies. Genetic maps were first constructed in the early twentieth century (Morgan, 1910) and preceded all other map construction methods. These early maps started with the meiotic mapping of chromosome landmarks, as defined by phenotypes such as eye color in flies. They next moved to biochemical markers such as blood antigens or isozymes and finally evolved into our current use of multiple allelic polymorphic DNA markers. These markers are commonly referred to as simple sequence repeats (SSR) and single-nucleotide polymorphisms (SNP). Even today, the markers that initially anchored the first crude linkage maps sometimes still serve as the basis for defining quantitative trait loci (QTLs) in agriculturally relevant species and in organisms used as human disease models. The basic premise for building a genetic map from polymorphic DNA markers has not changed, and involves the

typing of a sufficient number of multiple allele markers in a defined population and use of standard quantitative genetic formulas to place markers into ordered linkage groups. The SSR-based linkage map has been the traditional choice for linkage analysis studies, whereas SNP maps are more commonly used in association studies (see review; Blangero, 2004). Linkage analysis attempts to define alleles within chromosomal regions that are identical by descent in a pedigree or set of pedigrees, using phenotypic data to infer this relationship. Association analysis, in contrast, does not necessarily rely on related individuals but instead tests the frequency of an observed genotype on the distribution of phenotype across a large population. Both methods rely on linkage disequilibrium between the observed alleles and the quantitative trait of interest. Several groups have successfully used a combination of linkage and association methods to link genotype to disease causation (Styrkarsdottir *et al*., 2003; Helms *et al*., 2003; Pajukanta *et al*., 2004).

In the past decade, the development of large-model organism-specific SNP databases has provided the impetus to design whole-genome association studies to infer the existence of heritable QTLs. These studies may suffer from large stochastic variances in linkage disequilibrium and may thus require large numbers of SNPs, estimates range from 200 000 to 1 million (Carlson *et al*., 2004). Although these are daunting numbers, if costs and throughput continue on the current trend, we anticipate that the use of large whole-genome SNP panels will become a routine method for the discovery of candidate QTL or disease-associated regions, especially for complex diseases. The availability of a human haplotype map (which will catalog the more common haplotype blocks in humans (The International HapMap Consortium, 2003) will further augment these association studies by enabling us to more efficiently design mapping studies on samples of different ethnic origin. It appears inevitable that the construction of whole-genome genetic linkage maps will decline in nonmodel and non-food-producing organisms, and they are certainly not absolutely required to derive a finished genomic DNA sequence. However, it seems likely that they will continue to be used in defining the etiology of heritable traits in many plants and animals. It is also likely that the successful use of SNPs in association studies of complex human disease will require improvements in the collection and quantitation of many phenotypic parameters for very large sample sets (Bell, 2004). In the end, even association studies can only pinpoint a region, an SNP or set of SNPs that is associated with increased disease risk. Improvements in the targeted resequencing of large genomic regions will be necessary to uncover all of the DNA variation that exists in affected individuals.

3. Physical maps

Physical maps are relatively new and have advanced rapidly in the last decade owing to advances in clone manipulation and high-throughput automation. Physical mapping (the positioning of molecular clones along a genome) has been a central technology in deriving a finished genomic DNA sequence for many genomes. Most notably, the high-quality DNA sequence of the human and many other model organism genomes has greatly accelerated our ability to map a phenotype of interest to a defined chromosomal region. A plethora of physical mapping techniques exist

and fall into three general categories: (1) cytogenetic characterization, in which fluorescent in situ hybridization (FISH) is the most used method, whereby markers are localized along a defined chromosomal axis (Heng *et al*., 1997); (2) radiation hybrid mapping, in which hybrid cell lines randomly segregate chromosome fragments. These fragments can then be ordered by the PCR amplification of specific markers across a panel of cell lines (Cox, 1992); and (3) restriction mapping, which relies on random distribution of restriction enzyme sites in a genome (Olson *et al*., 1986; Coulson *et al*., 1986; Marra *et al*., 1997).

Improvements in the resolution of FISH techniques have come from higher-resolution microscopy, more efficient dye labeling methods and improved template preparation, such as interphase nuclei FISH and fiber-FISH (Florijn *et al*., 1995). Molecular cytogenetics has experienced a resurgence in recent years with the modification of methods such as spectral karyotyping SKY (Weimer *et al*., 2001), and the advent of comparative genomic hybridization CGH (Kallioniemi *et al*., 1992). These have had a particular impact upon cancer (see review; Heng *et al*., 2004), but new methods such as array CGH (a fusion of CGH, large-insert genomic clones and microarray technologies; Pinkel *et al*., 1998) are providing insights into copy number changes in many other complex diseases. Radiation hybrid (RH) maps, as briefly mentioned above, evolved as a result of the need to improve marker resolution over genetic maps and to not rely on polymorphisms to order markers along a chromosomal axis. The use of RH maps for anchoring sequences along a chromosome and constructing synteny relationships among species is now well documented (Mouse Genome Sequencing Consortium, 2002; International Human Genome Sequencing Consortium, 2001). Another method related to RH mapping is HAPPY mapping, although to date its use has been relegated to chromosomes and not to whole-genome maps (Thangavelu *et al*., 2003). In this case, the segregation of markers is conducted entirely in vitro by randomly breaking DNAs and subpartitioning the DNA into smaller samples. The frequency of two markers being retained in any sample is proportional to their relative physical distance (as in RH maps). Thus, HAPPY mapping, like genetic maps and RH maps, relies upon statistical methods for marker ordering.

Maps have guided the genomic DNA sequence assemblies of many eukaryotic organisms. Sometimes, the phylogenetic relationship of a particular organism is so close to an already existing map from another species that organization of assembled sequences is relatively straightforward. One example of this would be the relationship of chimpanzee sequence assemblies to the sequenced human genome. However, even in this case, when utilizing known syntenic and karyotypic relationships between the two species, care was exercised so as to avoid "humanizing" the chimpanzee sequence. In general, some form of framework map is required as a scaffold on which to assemble the final consensus sequence, and this map becomes very important when duplications or gaps must be resolved. High-quality physical maps require the ordering of reference markers along a definable linear track. Frequently, this track has consisted of restriction enzyme cleavage points. Two predominant techniques have been employed in these types of maps; clone-based restriction enzyme fingerprinting methods and, to a lesser extent, optical mapping of restriction fragments. Fingerprint maps have aided the accurate sequence assembly of the human, mouse, rat, and chicken genomic DNA sequences

(Meyers *et al*., 2004; Wallis *et al*., 2004). In essence, they rely upon inferring large-insert clone overlaps and relationships by matching patterns in the lengths of various restriction enzyme digestion products. Optical maps, which depend upon stretching out DNAs and visualizing the length and order of various restriction enzyme digestion sites, have been helpful in quickly establishing genome order for relatively small genomes (Zhou *et al*., 2003) but suffer from the inability to actually archive a given stretch of DNA for further analysis. Whatever mapping method is employed, the lessons to date indicate that for most large genomes, once a draft genomic DNA sequence is achieved, it is necessary to validate the assembly by reconciling marker orders with some form of reference physical map.

4. Mapping informatics

For practitioners of map building and the use of maps in the genetic dissection of complex phenotypes, the main objective of informatics must be to present a great deal of information in a simple visualizable format. Typically, there are two definable stages associated with the positional cloning of candidate genes for a trait of interest. At early stages, a framework map of some type is routinely required to navigate even closer to the variant allele or haplotype block. Once an interval is roughly defined, additional genetic mapping accelerates progress toward higher interval resolution and potential candidate gene isolation. There are now numerous algorithms and software choices for how to accomplish this phase for quantitative trait loci in experimental model organism crosses, human pedigrees or human populations (see **http://linkage.rockefeller.edu/soft/list.html** for examples).

Comparative genomic analysis has become particularly important and powerful in defining genes and putative regulatory elements across species. There are now many such tools available for human and mouse genetics (many of which can be accessed through NCBI). One much favored entry point to these tools would be the UCSC genome browser and associated tools. However, many other software tools exist for the comparative evaluation of genetic and physical maps. For example, the package CMap (an extension of the genome model organism database (GMOD) project (**http://www.gmod.org**)) was originally written for map integration of various plant species, but it has pushed investigators toward common platforms, thus allowing experiments to quickly move from an in silico analysis to the bench or to the field. Likewise, the availability of FPC software has been pivotal in advancing the construction of fingerprint maps in numerous species (e.g., Soderlund *et al*., 1997). Improvements to FPC and additional independent mapping software have moved the physical mapping process from labor-intensive to highly automated in a short period of time (see **http://www.bcgsc.ca/bioinfo/software/** for software examples). For newly sequenced genomes, we expect that software development will continue to play a pivotal role in integrating map and sequence assemblies.

5. Conclusions

It is clear that maps have played a key part in assembling the linear order of sequenced genomes. They continue to provide the framework on which the genetic

dissection of complex phenotypes is based. However, just like geographic map projections of the world, they can contain distortions and ambiguities. Among these are the degree to which duplications, rearrangements, and deletions occur as common events throughout many genomes (including our own). Maps will continue to evolve and improve, in combination with overlayed information on transcription, regulation, and epigenetic levels of genomic control. It is hoped that the lessons of the past, in which even low-resolution framework maps played significant roles, will continue to be applied to future large-scale genomic DNA sequencing projects.

Further reading

Fuhrmann DR, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, Chinwalla A, Hillier LW, Waterston RH, McPherson JD, *et al*. (2003) Software for automated analysis of DNA fingerprinting gels. *Genome Research*, **13**, 940–953.

Steinmetz LM, Sinha H, Richards DR, Spielgelman JI, Oefner PJ, McCusker JH and Davis RW (2003) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326–330.

White R, Lalouel JM, Leppert M, Nakamura Y and O'Connell P (1989) Linkage maps of human chromosomes. *Genome*, **31**, 1066–1072.

References

Bell J (2004) Predicting disease using genomics. *Nature*, **429**, 453–456.

Blangero J (2004) Localization and identification of human quantitative trait loci: king harvest has surely come. *Current Opinion in Genetics & Development*, **14**, 233–240.

Carlson CS, Eberle MA, Kruglyak L and Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.

Coulson A, Sulston J, Brenner S and Karn J (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, **83**, 7821–7825.

Cox DR (1992) Radiation hybrid mapping. *Cytogenetics and Cell Genetics*, **59**, 80–81.

Florijn RJ, Bonden LA, Vrolijk H, Wiegant J, Vaandrager JW, Baas F, den Dunnen JT, Tanke HJ, van Ommen GJ and Raap AK (1995) High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. *Human Molecular Genetics*, **4**, 831–836.

Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JA, Robarge J, Ott J, *et al*. (2003) A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. *Nature Genetics*, **35**, 349–356.

Heng H, Spyropoulos B and Moens P (1997) FISH technology in chromosome and genome research. *BioEssays*, **19**, 75–84.

Heng H, Stevens JB, Liu G, Bremer SW and Ye SJ (2004) Imaging genome abnormalities in cancer research. *Cell & Chromosome*, **3**, 1–12.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterson RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**, 1072–1084.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Meyers BC, Scalabrin S and Morgante M (2004) Mapping and sequencing complex genomes: let's get physical! *Nature Genetics*, **5**, 578–588.

Morgan TH (1910) Sex-limited inheritance in Drosophila. *Science*, **32**, 120–122.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R and Frank T (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences*, **83**, 7826–7830.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, Duan XJ, Soro-Paavonen A, Naukkarinen J, Saarela J, *et al*. (2004) Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nature Genetics*, **36**, 371–376.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Soderlund C, Longden I and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences*, **13**, 523–535.

Styrkarsdottir U, Cazier JB, Kong A, Rolfsson O, Larsen H, Bjarnadottir E, Johannsdottir VD, Sigurdardottir MS, Bagger Y, Christiansen C, *et al*. (2003) Linkage of osteoporosis to chromosome 20p12 and association to BMP2. *PloS Biology*, **1**, 351–360.

Thangavelu M, James AB, Bankier A, Bryan GJ, Dear PH and Waugh R (2003) HAPPY mapping in a plant genome: reconstruction and analysis of a high-resolution physical map of a 1.9 Mbp region of *Arabidopsis thaliana* chromosome 4. *Plant Biotechnology Journal*, **1**, 23031.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo WL, *et al*. (2003) End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Sciences*, **100**, 7696–7701.

Wallis J, Aerts J, Groenen M, Crooijmans R, Layman D, Graves TA, Scheer D, Kremitzki C, Fedele M, Mudd N, *et al*. (2004) A physical map of the chicken genome. *Nature*, **432**, 761–764.

Weimer J, Koehler MR, Wiedemann U, Attermeyer P, Jacobsen A, Karow D, Kiechl M, Jonat W and Arnold N (2001) Highly comprehensive karyotype analysis by a combination of spectral karyotyping (SKY), microdissection and reverse painting (SKY-MD). *Chromosome Research*, **9**, 395–402.

Zhou S, Kvikstad E, Kile A, Severin J, Forrest D Runnheim R, Churas C, Hickman JW, Mackenzie C, Choudhary M, *et al*. (2003) Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides strain* 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome Research*, **13**, 2142–2151.

Introductory Review

Linking DNA to production: the mapping of quantitative trait loci in livestock

Stephen S. Moore, Christiane Hansen and Changxi Li
University of Alberta, Edmonton, AB, Canada

1. Introduction

The world's livestock populations provide a considerable resource with which functions can be assigned to genes. Unique populations have resulted from centuries of genetic selection focused on specific traits such as milk or meat production, fertility, and conformation, to name a few. Livestock production is of significant economic importance worldwide, and animal breeders and geneticists alike have historically been interested in linking the genetic makeup of an animal with its production. While the accuracy of traditional methods of selection is high, selection for certain traits can be limited by the fact that they can only be measured in one sex or are difficult to measure in a production setting. For these traits, molecular technologies can help improve the accuracy of their selection.

The majority of the traits of interest for the various livestock sectors are quantitative in nature in that they are controlled by many genes. Quantitative trait loci (QTL) are areas of the genome that affect a quantitative trait. A QTL may contain one or a number of genes affecting a specific trait. The size of the effect will vary depending on the gene and the trait involved. Our ability to identify QTL is a function of the size of the gene effect, the family or population structure under study, and the density of informative DNA marker information available. DNA marker information, compiled in the form of genetic linkage maps, is currently available for all of the major livestock species (Barendse *et al*., 1997; Kappes *et al*., 1997; Rohrer *et al*., 1996; De Gotari *et al*., 1998; Groenen *et al*., 2000).

DNA markers generally fall into two categories. The first category includes highly informative, multiallelic markers, of which *Simple Sequence Repeats* (SSRs or microsatellites) are the most widely used for genetic and QTL mapping in livestock. SSRs are hypermutable, which has resulted in multiple alleles segregating in any population. Studies using these markers have generally been restricted to within families, rather than populations, where the genetic phase, the specific allelic association with the trait, can be determined. This, in turn, has meant that in species such as cattle, QTL segregating in only a limited number of individuals, usually the sires or grandsires of families, have been studied.

The second type of marker is the *Single Nucleotide Polymorphism* (SNP). These markers are usually biallelic and much less mutable than SSRs. Studies using SNPs may be carried out in animal populations rather than families, allowing the analysis of many more individual genomes than is otherwise possible. SNPs promise to revolutionize the way in which QTL mapping is carried out in livestock species, promising both more rapid and sensitive approaches.

2. QTL mapping in livestock species

The majority of research in QTL identification has been carried out in cattle (dairy and beef) and swine. In both species, work has centered on traits that are of importance for production or product quality. For dairy cattle, the work has focused on milk production and quality-associated traits (e.g., Freyer *et al*., 2002; Olsen *et al*., 2002), while for beef cattle (see Table 1) and swine (e.g., Andersson-Eklund *et al*., 1998; Varona *et al*., 2002), the focus has been on growth and carcass-related traits. Work has also been done in both species to identify, among other things, QTL associated with reproductive, behavior and health-/disease-related traits (e.g., Desautes *et al*., 2003; Kuhn *et al*., 2003).

Table 1 Examples of QTL reported in the literature for several growth- and meat-quality traits in beef cattle

Trait	Chromosome	Region (cM)	Effect[a] (S.D.)	References
Birth weight	5	70–110	0.39–0.82	Davis *et al*. (1998)
		20–30	0.79	Li *et al*. (2002)
		60–70	n/a	Stone *et al*. (1999)
	6	20–70	0.39–0.82	Davis *et al*. (1998)
		25–60	3.8 (kg)	Casas *et al*. (2000)
	14	30–60	0.39–0.82	Davis *et al*. (1998)
	18	100–130	0.39–0.82	Davis *et al*. (1998)
	21	6–30	0.39–0.82	Davis *et al*. (1998)
Average daily gain on feed	5	0–20	0.62	Li *et al*. (2002)
		55–75	0.65	Li *et al*. (2002)
Preweaning average daily gain	5	0–20	0.68	Li *et al*. (2002)
		70–80	0.50	Li *et al*. (2002)
Backfat	5	65–70	0.67	Li *et al*. (2004)
	6	64–68	0.43	Li *et al*. (2004)
		81–83	0.42	Li *et al*. (2004)
	19	5–15	0.67	Li *et al*. (2004)
		39–46	1.33	Li *et al*. (2004)
		65–100	0.43	Li *et al*. (2004)
Marbling	2	25–45	n/a	Stone *et al*. (1999)
	17	10–60	32.77[b]	Casas *et al*. (2000)
	27	20–65	29.82[b]	Casas *et al*. (2000)

[a]Size of phenotypic variation attributable to the QTL.
[b]Actual effect.
n/a: Not available.

In chickens, the research focus has been primarily divided between the discovery of QTL underlying egg production-associated traits and growth/carcass/meat production-related traits (e.g., de Koning *et al*., 2003). Some research has also been carried out in identifying genes associated with disease, for example, Marek's disease (Yonash *et al*., 1999).

Experiments to discover QTL in sheep are more limited. For example, Beh *et al*. (2002) have performed a genome scan to identify QTL for resistance to the intestinal parasite *Trichostrongylus colubriformis* in sheep, and Rozen *et al*. (2003) have identified QTL associated with milk production traits.

3. Genes underlying QTL

While a significant amount of research has been done on QTL detection in livestock species, very few genes underlying the various QTL have actually been identified. Several examples are listed in Table 2. The limitation in identifying the genes has been due largely to the resolution of QTL mapping attainable using the current interval mapping approaches. Most regions identified as housing QTL have been in the order of 20–50 cM in length. This can approximate to as many as 50 million bases. The fact that such an interval can house thousands of genes, plus the fact that comparative maps with more well-studied genomes have until recently been rudimentary, has made it difficult to move from QTL regions to the genes themselves. More recently, Identity by Decent (IBD) methodology has been used to successfully narrow down QTL regions, and, in some cases, identify genes underlying QTL (Farnir *et al*., 2000; Moore *et al*., 2003; Li *et al*., 2002, 2004; *see also* Article 12, **Haplotype mapping**, Volume 3). This approach uses the historical recombination that has occurred over many generations within a given population, rather than the recombination observed within two or three generations in a family. The limitation has now become the reliance on SSR markers and the accompanying uncertainty regarding phase in wider populations. The use of the more stable SNP markers will greatly improve our ability to fine map genes.

Table 2 Examples of genes underlying or associated with QTL in livestock

Gene	Trait	References
AcylCoA: diacylglycerol acyltransferase 1 (DGAT1)	Milk composition (cattle)	Grisart *et al*. (2002)
Myostatin	Double-muscling (cattle)	Grobert *et al*. (1997)
Thyroglobulin	Marbling score (cattle)	Barendse (1999)
Ryanodine receptor	Stress susceptibility (pigs)	Fujii *et al*. (1991)
Leptin	Fatness (cattle)	Buchanan *et al*. (2002)
Estrogen receptor	Litter size (pigs)	Rothschild *et al*. (1996)

4. Future directions

The vast majority of the QTL work that has been done to date has relied on microsatellite markers coupled with interval mapping approaches. In order to fine-map many of the QTL that have been identified, SNP markers coupled with populationwide IBD methodologies will prove useful in the future. Linkage disequilibrium mapping (Jorde, 2000) also holds great promise in this regard. The completion of the bovine sequence and the analysis of the data generated will make application of these various techniques more feasible in the future.

References

Andersson-Eklund L, Marklund L, Lundstrom K, Haley CS, Andersson K, Hansson I, Moller M and Andersson L (1998) Mapping quantitative trait loci for carcass and meat quality traits in a wild boar x large white intercross. *Journal of Animal Science*, **76**, 694–700.

Barendse WJ (1999) Assessing lipid metabolism. International patent application PCT/AU98/00882, international patent publication WO 99/23248.

Barendse W, Vaiman D, Kemp SJ, Sugimoto Y, Armitage SM, Williams JL, Sun HS, Eggen A, Agaba M, Aleyasin SA, *et al*. (1997) A medium-density genetic linkage map of the bovine genome. *Mammalian Genome*, **8**, 21–28.

Beh KJ, Hulme DJ, Callaghan MJ, Leish Z, Lenane I, Windon RG and Maddox JF (2002) A genome scan for quantitative trait loci affecting resistance to trichostrongylus colubriformis in sheep. *Animal Genetics*, **33**, 97–106.

Buchanan FC, Fitzsimmons CJ, Van Kessel AG, Thue TD, Winkelman-Sim C and Schmutz SM (2002) Association of a missense mutation in the bovine leptin gene with carcass fat content and leptin mRNA levels. *Genetics, Selection and Evolution*, **34**, 105–116.

Casas E, Shackelford SD, Keele JW, Stone RT, Kappes SM and Koohmaraie M (2000) Quantitative trait loci affecting growth and carcass composition of cattle segregating alternative forms of myostatin. *Journal of Animal Science*, **78**, 560–569.

Davis GP, Hetzel DJS, Corbet NJ, Scacheri S, Lowden S, Renaud J, Mayne C, Stevenson R, Moore SS and Byrne K (1998) The mapping of quantitative trait loci for birth weight in a tropical beef herd. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, **26**, 441–444.

De Gotari MJ, Freking BA, Cuthbertson RP, Kappes SM, Keele JW, Stone RT, Leymaster KA, Dodds KG, Crawford AM and Beattie CW (1998) A second generation linkage map of the sheep genome. *Mammalian Genome*, **9**, 204–209.

de Koning DJ, Windsor D, Hocking PM, Burt DW, Law A, Haley CS, Morris A, Vincent J and Griffin H (2003) Quantitative trait locus detection in commercial broiler lines using candidate regions. *Journal of Animal Science*, **81**, 1158–1165.

Desautes C, Bidanel JP, Milan D, Iannuccelli N, Amigues Y, Bourgeois F, Caritez JC, Renard C, Chevalet C and Mormede P (2003) Genetic linkage mapping of quantitative trait loci for behavioral and neuroendocrine stress response traits in pigs. *Journal of Animal Science*, **80**, 2276–2285.

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, *et al*. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research*, **10**, 220–227.

Freyer G, Kuhn C, Weikard R, Zhang Q, Mayer M and Hoeschele I (2002) Multiple QTL on chromosome six in dairy cattle affecting yield and content traits. *Journal of Animal Breeding and Genetics*, **119**, 69–82.

Fujii J, Otsu K, Zorzato F, De Leon S, Khanna VK, Weiler JE, O'Brien PJ and MacLennan DH (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science*, **253**, 448–451.

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, *et al*. (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a

missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*, **12**, 222–231.

Grobert L, Royo Martin LJ, Poncelet D, Pirottin D, Brouwers B, Riquet J, Schoeberlein A, Dunner S, Menissier F, Massabanda J, *et al*. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nature Genetics*, **17**, 71–74.

Groenen MAM, Cheng HH, Bumstead N, Benkel BF, Briles WE, Burke T, Burt DW, Crittenden LB, Dodgson J, Hillel J, *et al*. (2000) A consensus linkage map of the chicken genome. *Genome Research*, **10**, 137–147.

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Research*, **10**, 1435–1444.

Kappes SM, Keele JW, Stone RT, McGraw RA, Sonstegard TS, Smith TPL, Lopez-Corrales NL and Beattie CW (1997) A second generation linkage map of the bovine genome. *Genome Research*, **7**, 235–249.

Kuhn C, Bennewitz J, Reinsch N, Xu N, Thomsen H, Looft C, Brockmann GA, Schwerin M, Weimann C and Hiendleder S (2003) Quantitative trait loci mapping of functional traits in the German holstein cattle population. *Journal of Dairy Science*, **86**, 360–368.

Li C, Basarab J, Snelling WM, Benkel B, Murdoch B and Moore SS (2002) The identification of common haplotypes on bovine chromosome 5 within commercial lines of *Bos taurus* and their associations with growth traits. *Journal of Animal Science*, **80**, 1187–1194.

Li C, Basarab J, Snelling WM, Benkel B, Kneeland J, Murdoch B, Hansen C and Moore SS (2004) Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21, and 23 in a commercial line of *Bos taurus*. *Journal of Animal Science*, **82**, 967–972.

Moore SS, Li C, Basarab J, Snelling WM, Kneeland J, Murdoch B, Hansen C and Benkel B (2003) Fine mapping of quantitative trait loci and assessment of positional candidate genes for backfat on bovine chromosome 14 in a commercial cross of *Bos taurus*. *Journal of Animal Science*, **81**, 1919–1925.

Olsen HG, Gomez-Raya L, Vage DI, Olsaker I, Klungland H, Svendsen M, Adnoy T, Sabry A, Klemetsdal G, Schulman N, *et al*. (2002) A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. *Journal of Dairy Science*, **85**, 3124–3130.

Rohrer GA, Alexander LJ, Hu Z, Smith TPL, Keele JW and Beattie CW (1996) A comprehensive map of the porcine genome. *Genome Research*, **6**, 371–391.

Rothschild M, Jacobson C, Vaske D, Tuggle C, Wang L, Shorts T, Eckardt G, Sasaki S, Vincent A, McLaren D, *et al*. (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Science USA*, **93**, 201–205.

Rozen FMB, Cappelletti CA, Arranz JJ, Diez TMC and San PF (2003) A search for quantitative trait loci for milk production traits on chromosome 9 in churra sheep. *Journal of Basic and Applied Genetics*, **15**, 11–17.

Stone R, Keele TJW, Shackelford SD, Kappes SM and Koohmaraie M (1999) A primary screen of the bovine genome for quantitative trait loci affecting carcass and growth traits. *Journal of Animal Science*, **77**, 1379–1384.

Varona L, Ovilo C, Clop A, Noguera JL, Perez-Enciso M, Coll A, Folch JM, Barragan C, Toro MA and Babot D (2002) QTL mapping for growth and carcass traits in an Iberian by landrace pig intercross: additive, dominant and epistatic effects. *Genetic Research*, **80**, 145–154.

Yonash N, Bacon LD, Witter RL and Cheng HH (1999) High resolution mapping and identification of new quantitative trait loci (QTL) affecting susceptibility to Marek's disease. *Animal Genetics*, **30**, 126–135.

Specialist Review

Mapping complex disease phenotypes

David A. Collier
The Institute of Psychiatry, London, UK

1. Introduction

Complex medical conditions such as obesity and diabetes, and psychiatric disorders such as depression and schizophrenia, are common disabling diseases. Their aetiology involves genetic factors, the environment, and their interaction, with genes typically explaining half or more of the variance. A complete model of causation would include all genetic and environmental factors, and their joint effect on risk of illness. Genetic mapping of complex disease phenotypes focuses on identifying just one component of a complex causal system, the role of genes (*see* Article 58, **Concept of complex trait genetics**, Volume 2).

The genetic factors involved in complex disease phenotypes are likely to be risk alleles that are relatively common in the population, but have a modest effect on risk, that is, they exert a weak genetic effect. This is in contrast to high-risk alleles, which have been found only rarely in complex disorders.

For example, more than 55 high-risk alleles for obesity have been identified in a total of six genes, but these are present in less than 2% of the obese population, and most are found in single families (Obesity gene map database; **http://obesitygene.pbrc.edu/**). Likewise, more than 150 rare high-risk alleles have been identified for Alzheimer's disease, but their population attributable fraction is only 5%. This is in contrast to the *APOE4* gene, which as a common, modest risk factor has a population attributable fraction of 20%. Rare high-risk alleles have been very valuable in unraveling the underlying aetiological pathways in diseases such as obesity and Alzheimer's disease, but for other diseases such as depression and schizophrenia, such rare high-risk alleles have been elusive.

Common, modest risk alleles are likely to be substantially more important than rare high-risk alleles in efforts to improve our understanding of aetiology and the identification of novel treatments. However, different methods are required for their identification as the tools used for identifying high-risk alleles have not been successfully applied (Cardon and Bell, 2001; Risch, 2000).

2. Linkage analysis

2.1. Linkage in complex phenotypes

The most successful method for identifying genes for human disorders has been linkage analysis, which uses families with the disorder to search for the cosegregation or sharing by affected members of genetic markers (*see* Article 15, **Linkage mapping**, Volume 3). It is reliant solely on the genetic analysis of the phenotype, avoiding the need for prior information on pathophysiology and the function of potential risk genes. Linkage analysis has been successful in the identification of genetic loci for many human genetic diseases, and in animal models of disease. It has substantial power for rare, highly penetrant risk alleles such as those in single-gene disorders, however, power is reduced for complex diseases, where risk alleles only have a moderate effect on risk (meaning that allele sharing between affected subjects will be much less evident), and which are influenced by environmental factors. In addition, classical parameters used for mapping Mendelian diseases relating to the mode of inheritance (dominant, recessive) cannot be readily applied as they are unknown for complex disorders (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1). To overcome this, two approaches are typically taken to linkage in complex disorders: the nonparametric approach whereby parameters are abandoned and the analysis focused on sharing of alleles between affected family members, usually sibling pairs, or the retention of the likelihood method, with the approximation of parameters in a flexible frame work (Sham, 1997).

Steps in a linkage study involve identifying the phenotype, which can be a dichotomous trait such as clinical diagnosis, or a quantitative trait such as body mass index (BMI) or neuroticism; identifying and collecting DNA from the family samples; selecting a genetic marker map and genotyping; and data cleaning and statistical analysis of the data to search for evidence of linkage.

Linkage analysis in complex disorders using dichotomous traits such as disease diagnoses is relatively straightforward. In its simplest form, it measures allele sharing by affected relatives using identity-by-state or identity-by-descent methods, and uses a test statistic to evaluate the significance of deviation from the null hypothesis. A typical linkage marker set consists of 500 or fewer genetic markers spaced throughout the genome, formed into a genetic map; the most commonly used map is the Marshfield map (**http://research.marshfieldclinic.org/genetics/Default.htm**), but many others are available.

Various study designs can be used to increase power and reduce effort in linkage analyses. These apply particularly to quantitative traits. For example, sibships with the greatest expected contributions to a true LOD peak can be selected for linkage analysis by selecting extreme discordant or concordant sib pairs for the trait under study (Purcell *et al*., 2001; Carey and Williamson, 1991; Risch and Zhang, 1995). This approach may be very powerful and efficient when mapping quantitative traits in a large population, as genotyping will only be needed in a subset (Nash *et al*., 2004).

2.2. Data cleaning

After genotyping, data cleaning is used to check the integrity of the data; this is particularly important for testing family structures. For example, sibling pairs may in fact be half-siblings or even unrelated, and the presence of unspecified monozygotic twins could inflate any linkage statistics; programs such as Graphical Relationship Representation (Abecasis *et al*., 2001) can be used to locate incorrect family relationships by a scatter plot of the mean against the variance of the number of alleles identity-by-state for the typed markers for all pairs of individuals in the sample. Programs such as PREST and ALTERTEST (Sun *et al*., 2002), which perform multiple tests of a number of possible relationships, can also be used to assess family relationships in linkage samples. Genotyping errors can be checked for by searching for Mendelian incompatibilities (e.g., with the program PEDSTATS) and double recombinants, which are indicative of unlikely genotypes. Genotyping error rates are typically 1% or less.

2.3. Statistical analysis

Many programs are available for complex disease linkage analysis (**http://linkage.rockefeller.edu/soft/**) and can be used for categorical or quantitative traits. Programs such as GeneHunter (Kruglyak *et al*., 1996) or Merlin (Abecasis *et al*., 2002) take a collection of categorical (GeneHunter) or quantitative (GeneHunter or Merlin) trait and genetic marker values, a pedigree, and a marker map. This data is used to perform single-point and multipoint linkage analyses of pedigree data, including parametric, identity by descent (IBD), and nonparametric and variance component linkage analyses. The general output generated is a plot of QTL (quantitative trait loci) location versus LOD score (for parametric analysis) or Z-score (for nonparametric analysis).

Levels of significance can be denoted as a genome-wide p-value, that is, the probability that the observed value will be exceeded anywhere in the genome, assuming the null hypothesis of no linkage. Criteria for linkage in complex disease are a little different from those for single-gene traits (Lander and Kruglyak, 1995; Sawcer *et al*., 1997), with an LOD score of about 3.3 being regarded as significant evidence of linkage. Lower LOD scores may still represent true positives, and an LOD score of 2 can be regarded as suggestive linkage. Linkage has been regarded as confirmed when a significant linkage observed in one study is confirmed by finding an LOD score or p-value that would be expected to occur 0.01 times by chance in a specific search of the candidate region. However, meta-analysis of linkage data is a more powerful approach for complex disease linkage analysis.

2.4. Meta-analysis

Linkage approaches have been partially successful for complex diseases, and a large body of linkage data has been built up for most complex phenotypes.

In schizophrenia for example, there have been at least 20 genome-wide scans (Lewis *et al*., 2003), and for BMI more than 30 (Obesity gene map database; **http://obesitygene.pbrc.edu/**), and these have led to the putative identification of genetic loci identified in more than one study. However, although statistically significant findings have sometimes been supported by subsequent studies, there is a general lack of consistency for most phenotypes in complex disease linkage analysis. Indeed, some genome scans fail to find linkage at all (Altmuller *et al*., 2001). This could be because susceptibility is conferred by alleles at combinations of loci, each with a small effect on risk, and that the loci of greatest effect vary considerably in their impact between samples, because of geographic variation. False-positive findings are also likely to have arisen, and this could occur more than once for particular loci given the large number of genome scans (Lander and Kruglyak, 1995).

One major issue with linkage is statistical power. Because susceptibility loci for complex diseases are expected to have a small population-wide effect on susceptibility, it is thus difficult to detect their presence consistently without very large samples. For small genetic effect sizes, genome scans with low statistical power tend to overestimate the effect of loci with the highest scores in the scan, that is, maximize the genetic parameters (Goring *et al*., 2001). Genome scans for complex diseases have typically been underpowered – up to 1000 sibling pairs would be required to reliably demonstrate locus-specific genetic effects causing an approximately 30% increase in risk to siblings (Hauser *et al*., 1996). Multiplicative genetic effects are even more difficult to detect (Rybicki and Elston, 2000).

One way to overcome the issue of power is to perform meta-analyses of genome scans, for which several strategies are available (Xu and Meyers, 1998; Gu *et al*., 2001; Zhang *et al*., 2001; Etzel and Guerra, 2002; Dempfle and Loesgen, 2004). The most robust approach would be to pool the raw data using the original genotypes for each study, construct a combined map of the markers, and perform new linkage analyses, which should find loci consistently (Guerra *et al*., 1999). In practice, this is not easily done as raw genotype data may not be readily available or restricted by commercial confidentiality. In the absence of raw data, combining significance or effect estimates can provide an overall, but more limited, assessment of different linkage studies. Genuine meta-analyses combine statistics from different studies, and can be divided into those that combine significance tests (e.g., p-values from across studies) and those that combine effect estimates and test the significance of the common effects (Dempfle and Loesgen, 2004).

Combining significance tests can be performed using Fisher's method for p-values (Guerra *et al*., 1999), and various modifications are available to combine p-values only below a certain threshold (Zaykin *et al*., 2002; Olkin and Saner, 2001) for avoiding bias when truncated LOD scores are used (Province, 2001; Wu *et al*., 2002), and for correcting for multiple testing on the basis of the size of the region-implicated multiple scan probability (MSP; Badner and Goldin, 1999).

Unlike the above methods, the Genome Scan Meta Analysis method (GSMA; Wise *et al*., 1999; Levinson *et al*., 2003) was specifically designed for meta-analysis of linkage data, and is a nonparametric rank method that relies on combining effect estimates and testing the significance of the common effects. It requires only placing markers within 30-cM bins and the rank ordering of each

bin within and then across studies, allowing the consideration of any linkage test statistic and avoiding the need for the same set of markers to be used. However, GSMA provides no formal test of genetic hetereogeneity, and the interpretation of genome-wide statistical significance currently relies on empirical grounds.

Other nonparametric meta-analyses combine IBD statistics, using the number of alleles common across relative pairs (Gu *et al*., 1998, Gu *et al*., 1999). Sample size can also be accounted for (Goldstein *et al*., 1999). Methods for pooling of quantitative trait data are also available (Zhang *et al*., 2001), such as combining Haseman–Elston regression coefficients in a random effects model (Etzel and Guerra, 2002).

2.5. QTL mapping complex phenotypes in the mouse

Genetic mapping of complex traits in animals is attractive because of the statistical power and simplicity of the genetics (Flint and Mott, 2001). Linkage analysis of a complex trait in a cross between two inbred strains of mice relies on the fact that there are only three genotypes at a given locus since the parents are homozygous, meaning that, in effect, a test of association is being performed. This is more powerful than the tests of linkage across human families as the variance has a common basis across all the animals tested, and sources of variance as low as 5% can be detected.

Mouse genetic models for complex diseases are very powerful when used for the mapping of QTLs, such as anxiety, hypertension, and adiposity (Abiola *et al*., 2003). Typically, two inbred strains are crossed to form the F1 progeny, which are then intercrossed to generate an F2 generation or backcrossed to one of the parental strains. Since each progeny chromosome has undergone one meiosis, it will contain about one recombinant per morgan on average, meaning only 3–4 markers per chromosome need be typed for mapping. The genotype at any locus in the F2 must be homozygous for either parental allele, or heterozygous. For each marker locus, the trait mean is examined for each grouped genotype from the F2 progeny, and tested for statistically significant difference. Markers close to a QTL have similar genotypes to those at the undetected QTL and, consequently, the test at such a marker will be almost equivalent to testing for differences at the QTL.

Using various breeding designs (Silver, 1999), such as F2 crosses, recombinant inbred strains (Plomin *et al*., 1991; Williams *et al*., 2001), congenic strains, and chromosome substitution strains (Nadeau *et al*., 2000; Singer *et al*., 2004), more than 100 QTL loci in mice have been detected, reflecting the large-scale simple genetic structure of QTL genetic effects (Flint and Mott, 2001). In theory, under such a simple architecture, fine-mapping in order to identify the underlying genetic variants should be possible, using large numbers of mice to generate recombinants around the QTL, despite the fact that the effect of each (∼5% of the variance) is weak. However, from these 100+ mouse QTLs, only one actual gene underlying a QTL effect has been isolated (Yalcin *et al*., 2004), and in plants, only genes for moderate to major QTLs have been identified despite the use of thousands of crosses for mapping. The reason for this is the hidden complexity of QTLs; each locus detectable by mapping may not map to a single gene, but a group of QTL “increaser” and “decreaser” alleles that lie within a cluster of genes covering a

large (up to 30 cM) region (Darvasi and Soller, 1997; Legare *et al*., 2000; Flint and Mott, 2001). Furthermore, loci can interact synergistically (epistasis), an effect that cannot easily be detected by QTL methods. As a result of these factors, methods such as recombination mapping and the use of congenic strains may fail to identify the underlying QTL.

Methods to overcome this have focused on intermating strategies to break down linkage and increase mapping resolution, particularly those that use outbred stocks of mice to create advanced intercrosses (Talbot *et al*., 1999; Mott *et al*., 2000). Thus, a mapping resolution for a QTL of less than 1 cM has been achieved with genetically heterogeneous HS mice, for which each chromosome is a fine-grained mosaic of the eight founder chromosomes that make up the stock.

With the optimum approach, it is possible to perform fine-mapping to identify at least a group of candidate genes; however, the final problem is the identification of which gene harbors the QTL. Mapping studies are aimed at identifying DNA polymorphisms that alter the trait of interest, and a functional polymorphism can lie anywhere within or near a gene; for example, enhancers tens of kilobases away from the coding part of a gene are known, so the location of the QTL allele may not necessarily implicate a particular gene. Furthermore, there may be hundreds of neutral polymorphisms within the region of interest, and it is currently difficult and laborious to use bioinformatic and functional genomic analysis to tell which is a QTL allele and which is not. However, methods such as transgene complementation can help identify which gene is involved, if not which polymorphism (Flint and Mott, 2001; De Luca *et al*., 2003; Yalcin *et al*., 2004).

3. Fine-mapping in humans

While it has been possible to map genes that have a large phenotypic effect and can thus be localized by the use of recombinants, the reduced penetrance in complex diseases means that recombination events cannot be used to reliably map the position of susceptibility alleles. Statistical approaches based on analysis of recombinants are also not reliable because of the small numbers that occur within each family. Thus, in almost all cases, it has not proved possible to identify complex disease genes by linkage mapping alone.

After the initial genome scan, a linked region will typically be refined with additional microsatellite markers to drain the region of any residual informativeness for linkage. Although this may increase the LOD score of the region, it may only sharpen the linkage peak a little. Further efforts at refining linkage peaks, such as ordered subset analysis, may be used (Hauser *et al*., 2004). However, linkage will leave a candidate locus of as much as 10 cM, which will contain on average 80 genes.

3.1. Positional candidate genes: a mapping shortcut

If the region is reasonably well defined and the pathophysiology of the disease fairly clear, this may allow the selection of candidate genes based on position and function, which can be directly evaluated for their contribution to the trait under test

by association analysis. While this is a strong approach for diseases with specific tissue or cellular localizations and characteristic pathology, such as eye disease or diabetes, this approach has been less successful for most diseases, including psychiatric diseases such as schizophrenia or depression, where information on pathophysiology is poor.

Many techniques can be used to identify strong candidate genes from within linked regions. These include data mining techniques that take advantage of the increasing level of knowledge on gene function. For example, Perez-Iratxeta *et al*. (2002), Perez *et al*. (2004), and others use systematic annotation of genes with controlled vocabularies to develop a scoring system for the possible functional relationships of human genes to genetically inherited diseases, including complex diseases. The Gene Ontology Annotation (GOA) database (**http://www.ebi.ac.uk/GOA**) (Camon *et al*., 2004) provides high-quality electronic and manual annotations to the UniProt Knowledgebase (Swiss-Prot, TrEMBL, and PIR-PSD) using the standardized vocabulary of the Gene Ontology (GO; *see* Article 82, **The Gene Ontology project**, Volume 8), allowing functional assessment of many genes. The goal of the GO project is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing (Ashburner *et al*., 2000; Lewis, 2005).

Other methods can also be used in the gene identification process, such as transcriptomics (reviewed in Farrall, 2004) and proteomics (Jaffe *et al*., 2004; *see also* Article 94, **Expression and localization of proteins in mammalian cells**, Volume 4); in disease mapping, these approaches can annotate gene databases with useful functional information and can be used to attempt to identify changes in protein or mRNA levels, distribution or function that can be used to implicate genes in the disease process. These approaches are largely unproven at present, as many of these changes can be secondary to the disease process or confounded by factors such as medication.

4. Positional cloning by linkage disequilibrium: fine-mapping

Fine-mapping strategies focus on systematically searching for genetic markers from within a linkage locus that are associated with the disease or trait in question, and can also be applied to genome-wide analysis (see below). In the human genome, there are more than 6 million common (minor allele >0.1) SNPs in about 3.2 billion bp (Kruglyak and Nickerson, 2001), plus more than 500 000 VNTRs (variable number of tandem repeats). This translates to about 1 SNP every 500 and 1 VNTR every 6000 or so base pairs, equivalent to tens of thousands of potential disease-susceptibility polymorphisms in any complex disease linkage locus.

This high density of markers provides a problem for fine-mapping studies, as there are a large number of potential susceptibility alleles within any given locus. However, the existence of linkage disequilibrium (LD; also known as allelic association) means that these markers are not independent of each other, and it is possible to infer the location of a disease-susceptibility allele without actually genotyping it (Weiss and Clark, 2002; *see also* Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3). Thus, if a particular genetic marker is in

LD with a disease or trait susceptibility allele, the marker will also be in LD with the disease or trait. LD in the human genome, at least in non-African populations, is higher than what had been expected, making the LD approach highly promising for mapping studies. However, intervals displaying association may be relatively wide and hence contain many genes, especially in admixed or isolated populations, a finding borne out by the analysis of QTLs in the mouse (Flint and Mott, 2001).

LD is present when recombination between alleles is rare, because they are physically close together on the same chromosome. Thus, instead of the alleles of two adjacent markers being randomly distributed with respect to each other, as they would be if they occurred on separate chromosomes (or indeed far apart on the same chromosome), their distribution becomes nonrandom and the alleles exhibit LD. This also means that there are a limited number of haplotypes in any given region, reducing genetic complexity (Boehnke, 2000).

4.1. Measurement of LD

LD is an unpredictable measure; unlike linkage, which is hierarchical, physical or genetic distance cannot be used to predict LD between markers. Markers only a few hundred base pairs apart may be in weak LD, whereas markers separated by hundreds of kilobases may be in very strong LD. Consequently, LD must be measured experimentally.

Measurements of LD typically capture the strength of association between pairs of biallelic sites (pairwise LD), and are usually measured using the statistics D' (Lewontin, 1964) or R^2 (sometimes denoted $\Delta 2$; Devlin and Risch, 1995) (see Wall and Pritchard, 2003 for review). Both are normalized statistics, that is, they measure the range from 0 (no LD) to 1 (complete LD), but their interpretation is different. D' is defined as equal to 1 if just two or three of the possible four haplotypes of a pair of biallelic markers are present. Intermediate values of D', where all four haplotypes are present, are variable and difficult to interpret (Hudson, 1985; Hudson, 2001; Pritchard and Przeworski, 2001). In contrast, the R^2 metric only reaches 1 if only two of the four haplotypes are present, that is, each allele is completely associated with just one other. It has a simpler inverse relationship with the sample size required to detect association with susceptibility loci. Thus, to detect a susceptibility allele using a nearby genetic marker in LD with it, the sample size needs to be increased by a factor of $1/R^2$ in comparison to examining the susceptibility polymorphism directly.

Other measures that examine LD across regions rather than just pairwise are possible, such as the measure ρ measures how much recombination would be required under a particular population model to generate the observed LD (Wall and Pritchard, 2003). Metric maps of LD in the human genome are also being created (Maniatis *et al*., 2002; Tapper *et al*., 2003) on the basis of LD units rather than positions in kilobases.

4.2. Fine-mapping strategies: subjects

Families used for linkage analysis are not likely to have sufficient power for fine mapping based on LD, and most investigators will collect a population sample of

cases and controls, or nuclear families, for the mapping process (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1 and Article 60, **Population selection in complex disease gene mapping**, Volume 2). There has been considerable debate on which are the most effective study samples for complex gene-mapping efforts (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1 and Article 60, **Population selection in complex disease gene mapping**, Volume 2).

Unrelated individuals have most often been used for LD mapping and association studies, mainly because of the simplicity of design and the ease of collecting samples, and the advantages of family-based analysis are in general not thought to be substantial (Morton and Collins, 1998). The main consideration for case–control association studies is population stratification (*see* Article 75, **Avoiding stratification in association studies**, Volume 4), as allele frequencies vary substantially between different human populations (a population is defined as stratified if it consists of more than two ethnic groups each with differing population allele frequencies). In effect, this results in poor case–control matching and false-positive (or sometimes false-negative) association results.

Using individual-specific inferred haplotypes as covariates in standard epidemiologic analyses (e.g., conditional logistic regression) is an attractive analysis strategy, as it allows adjustment for nongenetic covariates, provides haplotype-specific tests of association, and can estimate haplotype and haplotype × environment interaction effects (Kraft *et al*., 2005). Several methods, including the most likely haplotype assignment and the expectation substitution approach (Schaid, 2004; Zaykin *et al*., 2002; Stram *et al*., 2003) are available.

A variety of methods for the use of genomic controls to avoid stratification bias has been proposed (*see* Article 75, **Avoiding stratification in association studies**, Volume 4), which can detect and control for population stratification in genetic case–control studies (Devlin and Roeder, 1999; Devlin *et al*., 2001; Reich and Goldstein, 2001). A combined approach of careful ethnic assessment of study populations, because of the strong correspondence between genetic structure and self-reported race/ethnicity categories (Tang *et al*., 2005) combined with genomic control approaches may be the most efficient (Lee, 2004). Genotyping error can be minimized experimentally, for example, by using two different methods for genotyping the same sample; and using samples such as duplicates and identical twins to measure error rates. While this will increase costs, there may be significant enhancement in the ability to detect association, especially when the number and complexity of haplotypes is high (Kirk and Cardon, 2002).

Family-based association studies, such as the case-parents and the case-sibling designs (Risch and Merikangas, 1996), gained popularity for disease mapping since they avoid the problems of case–control matching though making marker comparisons are between members of the same family (Ewens and Spielman, 1995). However, theoretical and empirical study on the degree of population stratification bias in non-Hispanic European populations found the bias to be minimal (Wacholder *et al*., 2000). The use of nuclear families in association does not offer great advantages over case–control analysis for the detection of genotyping errors, particularly as there is no inheritance test for the nontransmitted alleles used as controls in family-based analysis. However, since phase information is available,

family-based haplotype tests may be particularly useful in mapping studies (Lange and Boehnke, 2004; Lin *et al*., 2004).

Certain populations may have advantages for genetic mapping; for example, LD intervals reach up to 1 Mb in general alleles of young subisolates, which may provide advantages for the initial locus positioning of complex traits (Varilo and Peltonen, 2004). Observations on LD parameters indicate that Eurasian populations (especially isolates with numerous cases) are efficient for genome scans, and populations of recent African origin (such as African-Americans) are efficient for identification of causal polymorphisms within a candidate sequence, since LD is lower (Lonjou *et al*., 2003). The main disadvantage of small isolates is statistical power; it may not be possible to obtain a large enough population for mapping studies and for the same reason opportunities for replication in the same population may be limited.

4.3. Fine-mapping strategies

Attempts to localize complex disease-susceptibility genes have focused on methods aimed at detecting LD between individual genetic markers, their haplotypes, and putative disease-susceptibility loci, and is already in use in complex disorders. The first applications were to major loci that could be assigned to haplotypes by family study (Kerem *et al*., 1989; Devlin and Risch, 1995; Terwilliger, 1995). These and other studies have provided the foundation for the application of LD mapping for positional cloning of common diseases in complex inheritance. A 10-cM region displaying linkage with a disease will contain about 20 000 SNPs, assuming its physical size is 10 megabases. To fine-map a 10-cM linked region with individual SNPs, about 3000 individual markers would be required, based on calculations used to estimate the number of SNPs required for a whole-genome scan (Carlson *et al*., 2003).

4.4. Mapping using haplotypes

Historically, association tests were limited to single variants, so that the allele was considered the basic unit for association testing. However, the use of haplotypes, or haploid genotypes, has become increasingly popular (*see* Article 12, **Haplotype mapping**, Volume 3). Many haplotype analysis methods require phase (i.e., family transmission) information inferred from genotype data. However, as the number of loci increases, the information loss due to haplotype ambiguity increases rapidly (Hoh and Hodge, 2000). Several strategies involving the expectation-maximization (EM) algorithm (Ott, 1977; Slatkin and Excoffier, 1996) have been proposed to overcome the problem of missing phase information for estimating haplotype frequencies (Excoffier and Slatkin, 1998; Hawley and Kidd, 1995; Chiano and Clayton, 1998). In general, EM estimation of haplotype frequencies for multiple genotypes is a better strategy for the recruitment of family members or intensive laboratory haplotyping for haplotype-based genetic studies. The availability of population-based haplotype databases will simplify this process further. However,

for most methods it is necessary either to discard families with ambiguous haplotypes or analyze the markers separately, resulting in potential loss of power (Cheng *et al*., 2003). For haplotype analysis, a frequency threshold for the inclusion of haplotypes (usually >3%) can be set to protect against misleading results due to rare alleles or haplotypes.

Methods for haplotype analysis of regions have focused on moving window analysis, in which a scan of sets of tightly linked SNPs is made across the region of interest, in order to identify the site of LD with the trait under test. This can require assigning a window width first and then analyzing multisite parental transmission data under this fixed width (Clayton, 1999; Zhao *et al*., 2000). Other procedures that can maximize LD with an appropriate window width of haplotype transmission data within a preset range have also been proposed (Cheng *et al*., 2003).

The analysis of LD in the human genome has led to the proposed use of haplotype-map-based LD approaches to mapping genes (Cardon and Abecasis, 2003; Wall and Pritchard, 2003). This arose from the observation that LD in the human genome appears to consist of "haplotype blocks", stretches of DNA where strong LD exists between markers, punctuated by areas of weak LD where recombination rates are much higher (Jeffreys *et al*., 2001; Daly *et al*., 2001; Patil *et al*., 2001). These blocks extend for <10 to more than 100 Kb, and have low haplotype diversity, meaning that, in theory, relatively few SNPs could be used to describe the haplotypes of a given region. Haplotype tagging is potentially an efficient way of mapping linkage loci, since it should be possible to use a small set of "htSNPs" to analyze each block in a linked region, reducing the effort by 75% or more (Johnson *et al*., 2001; Patil *et al*., 2001).

Under this model, mapping a linked locus of 10–20 cM would involve dividing the region into 100–200 blocks for analysis, each of which is tagged by a finite number of htSNPs (perhaps 5–10). Thus, a single large region could be fine-mapped by 500–1000 htSNPs. Various methods have been proposed for the selection of htSNPs (reviewed in Lin and Altman, 2004), including manual selection for small genomic regions (Daly *et al*., 2001; Johnson *et al*., 2001), systematic evaluation of subsets with a metric to evaluate each candidate set (Patil *et al*., 2001), analysis of all pairwise comparisons to select the htSNPs explaining the most haplotype diversity (Daly *et al*., 2001), entropy (Judson *et al*., 2002; Avi-Itzhak *et al*., 2003; Hampe *et al*., 2003), or minimization of the squared correlation between the estimated and the true number of copies of haplotypes (Chapman *et al*., 2003; Stram *et al*., 2003). Alternative approaches, which measure how well individual and sets of SNPs predict one another (Bafna *et al*., 2003), are based on set theory and recursive searches for the minimal set of SNPs from which the maximum number of the other SNPs in the data set can be derived (Sebastiani *et al*., 2003), or clustering SNPs by pairwise LD measures and then selecting one htSNP per cluster (Carlson *et al*., 2004). More recently, principal component analysis (PCA) has been proposed as an efficient method, and evidence suggests that it tends to select the smallest set of htSNPs to achieve a 90% reconstruction precision (Meng *et al*., 2003; Lin and Altman, 2004).

While good computational methods are available for efficient analysis of haplotype blocks, it is not yet clear how well defined they are in real populations and whether they are stable across diverse ethnic groups (van den Oord and

Neale, 2004). Some studies indicate that haplotype blocks are stable across diverse populations (Gabriel *et al*., 2002; Dawson *et al*., 2002), which would allow the generation of general human LD maps, analogous to the recombination maps created for linkage analysis (Maniatis *et al*., 2002). However, this model of the LD structure of the human genome may be excessively simplistic. For example, there is evidence that much of the genome may not be formed into haplotype blocks (Phillips *et al*., 2003), and that where blocks exist they are not necessarily discreet entities because of long range LD (Daly *et al*., 2001; Jeffreys *et al*., 2001; van den Oord and Neale, 2004).

The remaining challenge is then to refine the techniques for fine-mapping of the causal polymorphism(s) within regions of high LD. Obviously, if strong blocks exist, then it will require a combination of genetic and molecular biological methods to identify which of the SNPs within the block are causal. Methods based on single marker tests within a composite likelihood framework (Maniatis *et al*., 2005) can apply a model with evolutionary theory that incorporates a parameter for the location of the causal polymorphism.

4.5. Positional cloning by linkage disequilibrium – genome-wide approaches

The success of linkage analysis in identifying genes for single-gene disorders reflects the fact that it has substantial power to identify rare high-risk disease alleles, since IBD allele sharing for this type of genetic risk factor will be very high between affected individuals in pedigrees. However, for modest risk alleles, such as those operating in complex disorders, allele sharing between affected subjects will be much less evident (Carlson *et al*., 2004); for example, it has been estimated that samples of 600–1000 affected sibling pairs would be required to reliably demonstrate locus-specific genetic effects causing a 27–30% population-wide increase in risk to siblings (Hauser *et al*., 1996); typical sample sizes for complex disease linkage analysis are in the low hundreds.

Consequently, the most accepted route to mapping complex disease genes, linkage followed by LD mapping or positional candidate gene analysis, may miss loci that have modest effects on risk unless very large sample sizes or meta-analyses are employed. Association analysis, at least theoretically, has more power to detect common disease alleles that confer modest risk (Risch and Merikangas, 1996). It is also easier to attain high statistical power as larger numbers of case–control or family trio samples are available than multiply affected families, which can be elusive. However, because there are many more generations from the most recent common ancestor in an association sample, much higher marker densities are needed to detect association compared to linkage (Kruglyak, 1999; Gabriel *et al*., 2002).

Thus, whole-genome association has been proposed as an alternative approach to linkage followed by fine-mapping (*see* Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3), in which a dense genome-wide set of SNPs is tested for disease association under the assumption that if a risk polymorphism exists, it will either be genotyped directly or be in strong LD with

one of the genotyped SNPs. This is essentially a single-point approach, where markers are analyzed one by one, but other approaches such as "moving window" haplotype analysis or multipoint haplotype analysis are also possible (Morris *et al*., 2003). Family samples can also be used for genome-wide analysis. Lin *et al*. (2004) describe an algorithm and a statistical method that efficiently and exhaustively exploits haplotype information from sliding windows of all sizes to transmission disequilibrium tests, which can detect both common and rare disease variants of small effect.

Estimates of the number of SNPs required for whole-genome association vary, but a requirement for between 300 000 and 1.3 million SNPs has been suggested on the basis that strong LD operates over short distances, and empirical analysis of specific region (Kruglyak, 1999; Gabriel *et al*., 2002; Carlson *et al*., 2003). Both direct (gene-based) and indirect (neutral marker) mapping approaches have been suggested. In the indirect approach, a dense set of neutral SNPs is used in the hope of detecting LD with causative SNPs, whereas in the direct approach, each of the 25 000 or so human genes would be analyzed by a set of representative SNPs in each gene, with an attempt to focus on potentially functional SNPs (Kruglyak, 1999; Botstein and Risch, 2003; Neale and Sham, 2004). A focus on analyzing SNPs that alter coding regions (cSNPs) has been proposed, and there is some evidence that some complex disease causing polymorphisms will be cSNPs; however, it is at least as likely that complex disease risk alleles will lie in noncoding regulatory regions of genes, as seen for nonhuman QTLs (Yalcin *et al*., 2004; Flint and Mott, 2001) such as promoters, where it is difficult and laborious to assess the functionality (Buckland *et al*., 2004).

Neale and Sham (2004) have proposed a shift toward a gene-based approach in which all common variants within a gene are considered jointly, thus capturing all potential susceptibility alleles. This has advantages for the consideration of genetic differences in population, which are more readily resolved by use of a gene-based approach rather than either a neutral SNP-based or a haplotype-based approach. Thus, negative findings are subject only to the issue of power. Once all gene variants are characterized, the gene-based approach may become the natural end point for association analysis and will inform our search for functional variants relevant to disease aetiology.

Whole-genome approaches are dependent on assembling an adequate SNP marker map, for which difficulty in selecting SNPs comes principally from ethnic variation in LD and in SNP frequency. Carlson *et al*. (2003) estimate that even analyzing all 2.6 million SNPs known in 2003, only 80% of all common SNPs in European and 50% in African-origin populations would be detected; as many as a quarter of all SNPs seen in African populations are "private", that is, they do not exist elsewhere, meaning that SNP marker maps may need to be population specific and very dense. Improvements in efficiency may be achieved by forming SNPs into haplotypes and haplotype blocks that can be tagged with htSNPS, as proposed for locus-specific fine-mapping. However, this will require the completion of the human haplotype map (The International HapMap Consortium, 2003).

Microsatellite markers have also been proposed for whole-genome association. Ohashi and Tokunaga (2003) calculated a markedly higher power for microsatellite markers than for SNPs, even if more SNPs are analyzed, suggesting that the use of

microsatellite markers is preferable to the use of SNPs for genome-wide screening under certain assumptions. This method will be helpful for researchers who design genome-wide LD testing with microsatellite markers.

DNA pooling (Barcellos *et al*., 1997) has been proposed as a method to economize on the number of genotypes required for whole-genome association studies. Pooled genotyping is a powerful and efficient tool for high-throughput association analysis, both case–control and family-based. The use of pooling designs can significantly reduce the costs of a study. At the same time, since it is also extremely efficient with DNA resources, pooling can be an extremely effective method for conserving DNA (Sham *et al*., 2002; Norton *et al*., 2004). Sophisticated pooling designs are being developed that can take account of hidden population stratification, confounders, and interactions, and that allow the analysis of haplotypes (Hoh *et al*., 2003). Both microsatellites (Daniels *et al*., 1998; Breen *et al*., 1999) and SNPs (Breen *et al*., 2000) are amenable to the pooling approach. Pooling approaches that allow chip-based genotyping may be particularly cost-efficient and rapid.

Whole-genome analysis also raises the statistical problem of multiple testing and levels of significance when using hundreds of thousands of genetic markers (Carlson *et al*., 2004), requiring substantial Bonferroni correction. There are no obvious methods that can exclude false-positives while capturing true positives for genetic effects that are weak, aside from repeated replication and meta-analysis, which may be subject to bias and are not a substitute for an adequately powered primary study (Munafo and Flint, 2004). It may be possible to overcome this problem using very large samples sizes (e.g., 2000 cases and 2000 controls), and practical methods have been proposed to help identify important genetic factors more efficiently, such as ranking markers by proximity to candidate genes or by expected functional consequence. Single marker tests within a composite likelihood framework have also been proposed, and will avoid heavy Bonferroni correction (Maniatis *et al*., 2005).

Further reading

Morris AP, Whittaker JC and Balding DJ (2004) Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *American Journal of Human Genetics*, **74**(5), 945–953.

Service SK, Sandkuijl LA and Freimer NB (2003) Cost-effective designs for linkage disequilibrium mapping of complex traits. *American Journal of Human Genetics*, **72**(5), 1213–1220.

References

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2001) GRR: graphical representation of relationship errors. *Bioinformatics*, **17**(8), 742–743.

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B, Blankenhorn EP, Blizard DA, Bolivar V, Brockmann GA, *et al*. (2003) Complex Trait Consortium. The nature and identification of quantitative trait loci: a community's view. *Nature Reviews. Genetics*, **4**(11), 911–916.

Altmuller J, Palmer LJ, Fischer G, Scherb H and Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics*, **69**(5), 936–950; (2001) Erratum. *American Journal of Human Genetics*, **69**(6), 1413.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1), 25–29.

Avi-Itzhak HI, Su X and De La Vega FM (2003) Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. *Pacific Symposium on Biocomputing*, 466–477.

Badner JA and Goldin LR (1999) Meta-analysis of linkage studies. *Genetic Epidemiology*, **17**(Suppl 1), S485–S490.

Bafna V, Gusfield D, Lancia G and Yooseph S (2003) Haplotyping as perfect phylogeny: a direct approach. *Journal of Computational Biology*, **10**(3–4), 323–340.

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J and Thomson G (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *American Journal of Human Genetics*, **61**(3), 734–747.

Boehnke M (2000) A look at linkage disequilibrium. *Nature Genetics*, **25**(3), 246–247.

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl), 228–237.

Breen G, Harold D, Ralston S, Shaw D and St Clair D (2000) Determining SNP allele frequencies in DNA pools. *Biotechniques* **28**(3), 464–466, 468, 470.

Breen G, Sham P, Li T, Shaw D, Collier DA and St Clair D (1999) Accuracy and sensitivity of DNA pooling with microsatellite repeats using capillary electrophoresis. *Molecular and Cellular Probes*, **13**(5), 359–365.

Buckland PR, Hoogendoorn B, Guy CA, Coleman SL, Smith SK, Buxbaum JD, Haroutunian V and O'Donovan MC (2004) A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. *Biochimica Et Biophysica Acta*, **1690**(3), 238–249.

Camon E, Barrell D, Lee V, Dimmer E and Apweiler R (2004) The Gene Ontology Annotation (GOA) Database – An integrated resource of GO annotations to the UniProt Knowledgebase. *Silico Biology*, **4**(1), 5–6.

Cardon LR and Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, **19**(3), 135–140.

Cardon LR and Bell JI (2001) Association study designs for complex diseases. *Nature Reviews. Genetics*, **2**(2), 91–99.

Carlson CS, Eberle MA, Kruglyak L and Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**(6990), 446–452.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**(4), 518–521.

Carey G and Williamson J (1991) Linkage analysis of quantitative traits: increased power by using selected samples. *American Journal of Human Genetics*, **49**(4), 786–796.

Chapman JM, Cooper JD, Todd JA and Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity*, **56**(1–3), 18–31.

Cheng R, Ma JZ, Wright FA, Lin S, Gao X, Wang D, Elston RC and Li MD (2003) Nonparametric disequilibrium mapping of functional sites using haplotypes of multiple tightly linked single-nucleotide polymorphism markers. *Genetics*, **164**(3), 1175–1187.

Chiano MN and Clayton DG (1998) Fine genetic mapping using haplotype analysis and the missing data problem. *Annals of Human Genetics*, **62**(Pt 1), 55–60.

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics*, **65**(4), 1170–1177.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**(2), 229–232.

Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R and Owen MJ (1998) A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *American Journal of Human Genetics*, **62**(5), 1189–1197.

Darvasi A and Soller M (1997) A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics*, **27**(2), 125–132.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, *et al*. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**(6897), 544–548.

De Luca M, Roshina NV, Geiger-Thornsberry GL, Lyman RF, Pasyukova EG and Mackay TF (2003) Dopa decarboxylase (Ddc) affects variation in Drosophila longevity. *Nature Genetics*, **34**(4), 429–433.

Dempfle A and Loesgen S (2004) Meta-analysis of linkage studies for complex diseases: an overview of methods and a simulation study. *Annals of Human Genetics*, **68**(Pt 1), 69–83.

Devlin B and Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**(2), 311–322.

Devlin B, Roeder K and Bacanu SA (2001) Unbiased methods for population-based association studies. *Genetic Epidemiology*, **21**(4), 273–284.

Devlin B and Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

Ewens WJ and Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics*, **57**(2), 455–464.

Etzel CJ and Guerra R (2002) Meta-analysis of genetic-linkage analysis of quantitative-trait loci. *American Journal of Human Genetics*, **71**(1), 56–65.

Excoffier L and Slatkin M (1998) Incorporating genotypes of relatives into a test of linkage disequilibrium. *American Journal of Human Genetics*, **62**(1), 171–180.

Farrall M (2004) Quantitative genetic variation: a post-modern view. *Human Molecular Genetics*, **13**, Spec No 1, R1–R7.

Flint J and Mott R (2001) Finding the molecular basis of quantitative traits: successes and pitfalls. *Nature Reviews. Genetics*, **2**(6), 437–445.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**(5576), 2225–2229.

Goldstein DR, Sain SR, Guerra R and Etzel CJ (1999) Meta-analysis by combining parameter estimates: simulated linkage studies. *Genetic Epidemiology*, **17**(Suppl 1), S581–S586.

Gu C, Province M and Rao DC (1999) Meta-analysis of genetic linkage to quantitative trait loci with study-specific covariates: a mixed-effects model. *Genetic Epidemiology*, **17**(Suppl 1), S599–S604.

Gu C, Province MA and Rao DC (2001) Meta-analysis for model-free methods. *Advances in Genetics*, **42**, 255–272.

Gu C, Province M, Todorov A and Rao DC (1998) Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic homogeneity and identical markers. *Genetic Epidemiology*, **15**(6), 609–626.

Guerra R, Etzel CJ, Goldstein DR and Sain SR (1999) Meta-analysis by combining p-values: simulated linkage studies. *Genetic Epidemiology*, **17**(Suppl 1), S605–S609.

Goring HH, Terwilliger JD and Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, **69**(6), 1357–1369.

Hampe J, Schreiber S and Krawczak M (2003) Entropy-based SNP selection for genetic association studies. *Human Genetics*, **114**(1), 36–43.

Hawley ME and Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, **86**(5), 409–411.

Hauser ER, Boehnke M, Guo SW and Risch N (1996) Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. *Genet Epidemiol*, **13**(2), 117–137.

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD and Boehnke M (2004) Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic Epidemiology*, **27**(1), 53–63.

Hoh J and Hodge SE (2000) A measure of phase ambiguity in pairs of SNPs in the presence of linkage disequilibrium. *Human Heredity*, **50**(6), 359–364.

Hoh J, Matsuda F, Peng X, Markovic D, Lathrop MG and Ott J (2003) SNP haplotype tagging from DNA pools of two individuals. *BMC Bioinformatics*, **4**(1), 14.

Hudson RR (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, **109**(3), 611–631.

Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics*, **159**(4), 1805–1817.

Jaffe JD, Berg HC and Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, **4**(1), 59–77.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**(2), 217–222.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics*, **29**(2), 233–237.

Judson R, Salisbury B, Schneider J, Windemuth A and Stephens JC (2002) How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics*, **3**(3), 379–391.

Kerem BS, Buchanan JA, Durie P, Corey ML, Levison H, Rommens JM, Buchwald M and Tsui LC (1989) DNA marker haplotype association with pancreatic sufficiency in cystic fibrosis. *American Journal of Human Genetics*, **44**(6), 827–834.

Kirk KM and Cardon LR (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, **10**(10), 616–622.

Kraft P, Cox DG, Paynter RA, Hunter D and De Vivo I (2005) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genetic Epidemiology*, **28**(3), 261–272.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**(2), 139–144.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**(6), 1347–1363.

Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**(3), 234–236.

Lander E and Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**(3), 241–247.

Lange EM and Boehnke M (2004) The haplotype runs test: the parent-parent-affected offspring trio design. *Genetic Epidemiology*, **27**(2), 118–130.

Lee WC (2004) Case-control association studies with matching and genomic controlling. *Genetic Epidemiology*, **27**(1), 1–13.

Legare ME, Bartlett FS II and Frankel WN (2000) A major effect QTL determined by multiple genes in epileptic EL mice. *Genome Research*, **10**(1), 42–48.

Levinson DF, Levinson MD, Segurado R and Lewis CM (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis. *American Journal of Human Genetics*, **73**(1), 17–33.

Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I, Williams NM, Schwab SG, Pulver AE, Faraone SV, *et al*. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *American Journal of Human Genetics*, **73**(1), 34–48.

Lewis SE (2005) Gene Ontology: looking backwards and forwards. *Genome Biology*, **6**, 103.

Lewontin RC (1964) The interaction of selection and linkage. II. Optimum models. *Genetics*, **50**, 757–782.

Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N and Morton NE (2003) Linkage disequilibrium in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(10), 6069–6074.

Lin Z and Altman RB (2004) Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics*, **75**(5), 850–861.

Lin S, Chakravarti A and Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics*, **36**(11), 1181–1188.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X and Morton NE (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(4), 2228–2233.

Maniatis N, Morton NE, Gibson J, Xu CF, Hosking LK and Collins A (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Human Molecular Genetics*, **14**(1), 145–153.

Meng Z, Zaykin DV, Xu CF, Wagner M and Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *American Journal of Human Genetics*, **73**(1), 115–130.

Morris AP, Whittaker JC, Xu CF, Hosking LK and Balding DJ (2003) Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(23), 13442–13446.

Morton NE and Collins A (1998) Tests and estimates of allelic association in complex inheritance. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(19), 11389–11393.

Mott R, Talbot CJ, Turri MG, Collins AC and Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(23), 12649–12654.

Munafo MR and Flint J (2004) Meta-analysis of genetic association studies. *Trends in Genetics*, **20**(9), 439–444.

Nadeau JH, Singer JB, Matin A and Lander ES (2000) Analysing complex genetic traits with chromosome substitution strains. *Nature Genetics*, **24**(3), 221–225.

Nash MW, Huezo-Diaz P, Williamson RJ, Sterne A, Purcell S, Hoda F, Cherny SS, Abecasis GR, Prince M, Gray JA, *et al*. (2004) Genome-wide linkage analysis of a composite index of neuroticism and mood-related scales in extreme selected sibships. *Human Molecular Genetics*, **13**(19), 2173–2182.

Neale BM and Sham PC (2004) The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics*, **75**(3), 353–362.

Norton N, Williams NM, O'Donovan MC and Owen MJ (2004) DNA pooling as a tool for large-scale association studies in complex traits. *Annals of Medicine*, **36**(2), 146–152.

Ohashi J and Tokunaga K (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *Journal of Human Genetics*, **48**(9), 487–491.

Olkin I and Saner H (2001) Approximations for trimmed Fisher procedures in research synthesis. *Statistical Methods in Medical Research*, **10**(4), 267–276.

Ott J (1977) Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Annals of Human Genetics*, **40**(4), 443–454.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**(5547), 1719–1723.

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, *et al*. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics*, **33**(3), 382–387.

Perez-Iratxeta C, Bork P and Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, **31**(3), 316–319.

Perez AJ, Perez-Iratxeta C, Bork P, Thode G and Andrade MA (2004) Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, **20**(13), 2084–2091.

Plomin R, McClearn GE, Gora-Maslak G and Neiderhiser JM (1991) Use of recombinant inbred strains to detect quantitative trait loci associated with behavior. *Behavior Genetics*, **21**(2), 99–116. Review.

Pritchard JK and Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**(1), 1–14.

Province MA (2001) The significance of not finding a gene. *American Journal of Human Genetics*, **69**(3), 660–663.

Purcell S, Cherny SS, Hewitt JK and Sham PC (2001) Optimal sibship selection for genotyping in quantitative trait locus linkage analysis. *Human Heredity*, **52**(1), 1–13.

Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**(6788), 847–856.

Reich DE and Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, **20**(1), 4–16.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Risch N and Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, **268**(5217), 1584–1589.

Rybicki BA and Elston RC (2000) The relationship between the sibling recurrence-risk ratio and genotype relative risk. *American Journal of Human Genetics*, **66**(2), 593–604; (2000) Erratum. *American Journal of Human Genetics*, **67**(2): 541.

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN and Clayton D (1997) Empirical genomewide significance levels established by whole genome simulations. *Genetic Epidemiology*, **14**(3), 223–229.

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS and Ramoni MF (2003) Minimal haplotype tagging. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(17), 9900–9905.

Schaid DJ (2004) Linkage disequilibrium testing when linkage phase is unknown. *Genetics*, **166**(1), 505–512.

Sham PC (1997) *Statistics in Human Genetics*, Arnold: London, UK.

Sham P, Bader JS, Craig I, O'Donovan M and Owen M (2002) DNA Pooling: a tool for large-scale association studies. *Nature Reviews. Genetics*, **3**(11), 862–871.

Silver L (1999) *Mouse Genetics and Genomics*, Oxford University Press: Oxford.

Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, O'Brien WE, Conti DV, Witte JS, Lander ES, *et al*. (2004) Genetic dissection of complex traits with chromosome substitution strains of mice. *Science*, **304**(5669), 445–448.

Slatkin M and Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity*, **76**(Pt 4), 377–383.

Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE and Thomas DC (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity*, **55**(4), 179–190.

Sun L, Wilder K and McPeek MS (2002) Enhanced pedigree error detection. *Human Heredity*, **54**(2), 99–110.

Talbot CJ, Nicod A, Cherny SS, Fulker DW, Collins AC and Flint J (1999) High-resolution mapping of quantitative trait loci in outbred mice. *Nature Genetics*, **21**(3), 305–308.

Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, *et al*. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *American Journal of Human Genetics*, **76**(2), 268–275.

Tapper WJ, Maniatis N, Morton NE and Collins A (2003) A metric linkage disequilibrium map of a human chromosome. *Annals of Human Genetics*, **67**(Pt 6), 487–494.

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, **56**(3), 777–787.

The International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**(6968), 789–796.

van den Oord EJ and Neale BM (2004) Will haplotype maps be useful for finding genes? *Molecular Psychiatry*, **9**(3), 227–236.

Varilo T and Peltonen L (2004) Isolates and their potential use in complex gene mapping efforts. *Current Opinion in Genetics and Development*, **14**(3), 316–323.

Wacholder S, Rothman N and Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *Journal of the National Cancer Institute*, **92**(14), 1151–1158.

Wall JD and Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews. Genetics*, **4**(8), 587–597.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, **18**(1), 19–24.

Williams RW, Gu J, Qi S and Lu L (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biology*, **2**(11), RESEARCH0046.

Wise LH, Lanchbury JS and Lewis CM (1999) Meta-analysis of genome searches. *Annals of Human Genetics*, **63**(Pt 3), 263–272.

Wu X, Cooper RS, Borecki I, Hanis C, Bray M, Lewis CE, Zhu X, Kan D, Luke A and Curb D (2002) A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *American Journal of Human Genetics*, **70**(5), 1247–1256.

Xu J and Meyers DA (1998) Approaches to meta analysis in genetic disorders. *Clinical and Experimental Allergy*, **28**(Suppl 1), 106–107.

Yalcin B, Willis-Owen SA, Fullerton J, Meesaq A, Deacon RM, Rawlins JN, Copley RR, Morris AP, Flint J and Mott R (2004) Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nature Genetics*, **36**(11), 1197–1202.

Zaykin DV, Zhivotovsky LA, Westfall PH and Weir BS (2002) Truncated product method for combining P-values. *Genetic Epidemiology*, **22**(2), 170–185.

Zhang W, Collins A and Morton NE (2001) Combination of linkage evidence in complex inheritance. *Human Heredity*, **52**(3), 132–135.

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F and Kidd KK (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *American Journal of Human Genetics*, **67**(4), 936–946.

Specialist Review

Haplotype mapping

Alexandre Montpetit, Fanny Chagnon and Thomas J. Hudson

McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada

1. Introduction: linkage, association, and linkage disequilibrium (LD) mapping

Linkage mapping is based on coinheritance of large chromosomal stretches among members of a family sharing a common phenotype. This approach has been very successful for rare Mendelian diseases, but it has led to mixed results for common diseases (Lander and Kruglyak, 1995). Many factors, such as low penetrance, small sample size, different diagnostic or disease definitions, and uncertainty in the position of the linkage peak have contributed to the low success of family based genome scans. Association approaches can directly test functional polymorphisms or indirectly test mutations using genetic markers in linkage disequilibrium (LD) with the mutation. Association and LD mapping traditionally compare samples of cases and controls, but alternate strategies exist, such as the comparison of transmitted alleles inherited by an affected offspring with the untransmitted alleles of the parents. Association methods have greater power to detect small and moderate genetic effects than linkage analysis and, thus, are more suited for the identification of variants predisposing to common diseases (Risch and Merikangas, 1996). However, it is not currently feasible to perform association scans at the genome scale, and a mixed approach is typically used (see Table 1) (*see also* Article 11, **Mapping complex disease phenotypes**, Volume 3 and Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3 for more details on linkage and association mapping of common diseases).

While microsatellites were the markers of choice for linkage studies because they are highly informative, they cannot be used reliably for association studies because of their relatively high mutation rates. Single nucleotide polymorphisms (SNPs) are more stable and more abundant than microsatellite markers, enabling investigators to cover extensively any genomic region (Risch, 2000). There are estimates of 10 to 15 million common SNPs having a minimal minor allele frequency greater than 1% in the human population (Kruglyak and Nickerson, 2001). Using population simulations to estimate LD, Kruglyak (1999) hypothesized that the extent of useful levels of LD for association studies in the general outbred population is about 3 kb, implying that about 500 000 SNPs would be needed for a genome-wide association

Table 1 Examples of haplotype mapping studies

Disease	No. of cases/families (population)	No. of SNPs (density)	Gene/haplotype identified	Region	Haplotype frequency	Replication panel	References	Replication in independent study
Type 2 diabetes	110 cases (Mexican Americans)	114 (1/15 kb)	*CAPN10*	2q37	0.23/0.32	192 cases (Finnish)	Horikawa *et al*. 2000	Meta-analysis of 17 studies (Weedon *et al*., 2003)
Crohn disease	139 trios (Canadian)	301 (1/3.3 kb)	250 kb haplotype in the cytokine cluster	5q31	0.37	88 trios (Quebec)	Rioux *et al*. 2001	368 German trios (Giallourakis *et al*., 2003)
Asthma	130 cases (UK/US)	135 (1/18.5 kb)	*ADAM33*	20p13 (23 genes)	0.6	460 families (UK/US)	Van Eerdewegh *et al*. 2002	Werner *et al*. 2004
Schizophrenia	476 cases (Icelandic)	181 (1/8.3 kb)	*NRG1*	8p12-p21	0.075	-	Stefansson *et al*. 2002	609 Scottish cases (Stefansson *et al*., 2003)
Asthma	80 nuclear families (Australian)	54 (1/14 kb)	*PHF11*	13q14	0.08/0.4	237 cases (British)	Zhang *et al*. 2003	
Asthma	244 families (Australian/British)	82 (1/5.6 kb)	*DPP10*	2q14	0.3	270 cases (British)	Allen *et al*. 2003	
Susceptibility to leprosy	197 trios (Vietnamese)	81 (1/6 kb)	*PARK2/PACRG*	6q	0.31	587 cases (Brazilian)	Mira *et al*. 2004	
Myocardial infarction	779 cases (Icelandic)	48 (1/1.5 kb)	*ALOX5AP*	13q12-13	0.095	753 cases (British)	Helgadottir *et al*. 2004	

scan. With today's cheapest technology, it would be unfeasible to type them all in a typical case-control study. However, the rate of the LD decay is not uniform throughout the genome. Several studies observed that the local decay in LD is not continuous at all and can sometimes extend for very long distances up to 1-3 cM (Lonjou *et al*., 1999, Mohlke *et al*., 2001, Service *et al*., 2001). LD is known to extend for longer distances in isolated populations (such as populations from Iceland and Finland), or founder populations that experienced a rapid expansion (such as the Amish, or populations from Quebec and Newfoundland), and one might expect that this property should facilitate disease mapping (reviewed in Peltonen, 2000). While the use of isolated or founder populations helped in the identification of single mutations for rare diseases because it reduces allelic heterogeneity, their usefulness for common diseases is still debated (see Kruglyak, 1999).

2. Haplotype blocks

Recent studies using relatively high densities of SNPs enabled the analysis of the LD structure at smaller scales. It was observed that LD decays in sudden discrete drops to form blocklike structures that can extend from a few kilobase pairs to more than 100 kb (Daly *et al*., 2001, Reich *et al*., 2001, Patil *et al*., 2001). Moreover, these blocks seem to have very low diversity, pointing to the fact that most of this diversity comes from a small number of ancestral chromosomes. The specific set of alleles on a chromosomal segment is called a *haplotype*. New haplotypes can arise from new mutations or recombination. Each novel mutation is initially associated with alleles that happened to be present on a particular chromosome on which it arose (thus establishing a novel unique haplotype). These alleles will remain correlated until recombination events occur between them. The existence of recombination hot spots was suggested as an explanation for the presence of these large blocks of low haplotype diversity (or haplotype blocks) (Daly *et al*., 2001). Low haplotype diversity implies that only a few SNPs per haplotype block would need to be typed in order to extrapolate most of the information of the remaining SNPs, which could save considerable time and money.

For example, Rioux *et al*. (2001) applied an LD approach to the investigation of a Crohn disease susceptibility gene linked to chromosome 5q31. In a region spanning 250 kb encompassing several cytokine genes, they observed over one dozen SNPs that are associated with the disease. Daly *et al*. (2001) showed that the region can be subdivided into 11 haplotype blocks, each with only two to four haplotypes describing more than 90% of haplotype diversity at this chromosomal region (Figure 1). Had the underlying block structure been previously known, they could have genotyped as few as 20 SNPs (termed *haplotype tag SNPs or htSNPs*) to test the entire 250 kb. One risk-associated haplotype was identified, but because of the existence of large haplotype blocks and the presence of considerable LD between them, the causative polymorphism (or even the causative gene) cannot be identified using genetic mapping alone. In order to identify which gene or which polymorphism is contributing to the disease, functional studies, such as expression analysis, targeted mutations in vitro, or knockout studies in mice are necessary. Using a different population might also be helpful. Oksenberg *et al*. (2004) used

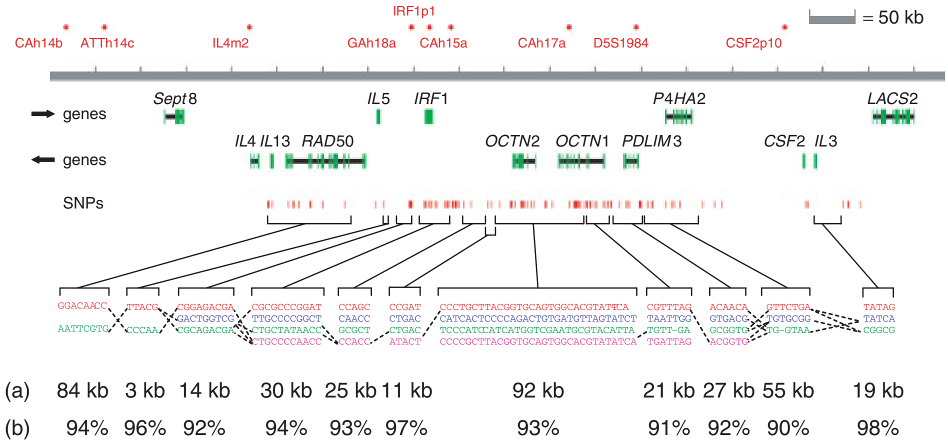


Figure 1 Haplotype structure at the IBD5 locus. (a) Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one. (b) Percentage of observed chromosomes that match one of the common patterns exactly. In addition, four markers fell between blocks, which suggests that the recombinational clustering may not take place at a specific base-pair position but rather in small regions. (Reproduced from Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232 by permission of Nature Publishing Group)

a cohort of African Americans to define a haplotype in the HLA region that is associated with multiple sclerosis. The African haplotypes in this region do not display the high degree of LD and haplotype structure observed in Caucasians. This enabled the research team to reduce the region to only one gene, *DRB1*, excluding *DQB1*. The fact that many disease-associated SNPs located in the vicinity of one or more genes can be present within a haplotype block, and that LD can be detected across blocks, clearly illustrates the importance of understanding the underlying haplotype structure of a region in the population studied. Using Alzheimer's disease as an example, several studies showed that the association between the ApoE gene and the disease could be identified by using haplotype information even though the true variant was not genotyped (Martin *et al*., 2000, Fallin *et al*., 2001).

Different analytical approaches can be used to construct a comprehensive LD/haplotype map of a selected region. Several methods have been described to identify haplotype blocks and will be explained in more details (*see* Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4). A color-coded pairwise D' map, constructed with programs such as GOLD (Abecasis and Cookson, 2000), is also very useful to visually examine regions of high and low LD (*see* Article 73, **Creating LD maps of the genome**, Volume 4). Finally, association can be measured using multilocus haplotypes instead of individual SNPs, resulting in an increase in power to detect indirect association and in a decrease in the number of tests performed, thus minimizing the correction applied to the p-value (Zhang *et al*., 2002a). Several studies have successfully mapped an associated gene or haplotype block to a common trait by using a dense set of SNPs to understand the haplotype structure. Some examples are listed in Table 1.

Although recombination hot spots have been observed directly in other species (yeast, *Drosophila*, mouse), it is labor-extensive in humans. Jeffreys *et al*. (2001) used sperm-typing to directly observe meiotic events and infer the existence of recombination hot spots in humans. They observed that 94% of the crossovers occurred at only six positions in the 300 kb HLA region on chromosome 6p. These hotspots for recombination are all between 1 and 2 kb in length. In yeast, hot spots tend to be associated with GC-rich chromosomal regions and form open domains in meiotic chromatin that allow access to the recombination machinery (reviewed in Petes, 2001). However, in humans, no similarity in the primary genomic sequence has yet to be predictive of the hot spots. Recombination hot spots have also been observed in the pseudoautosomal region of chromosome Y (Lien *et al*., 2000, May *et al*., 2002) and in the β-globin gene cluster (Schneider *et al*., 2002, Smith *et al*., 1998). It is still unknown whether these observations are representative of the recombination events in the remaining genome and for all populations. Kauppi *et al*. (2003) showed that although haplotype diversity was different in the HLA region between European and African populations, patterns of LD were similar. We note, however, that the average recombination rate is 1.65 times higher in females than in males, pointing to the fact that mechanisms involved are not homogeneous (Kong *et al*., 2002).

Gabriel *et al*. (2002) extensively genotyped 50 autosomal regions spanning 13 Mb of sequence and found that, on average, haplotype blocks extend for 22 kb in Asian and European populations and 11 kb in the Yoruba population from Nigeria. Using one marker every 7.8 kb, they observed that about 50% of the sequence was in blocks. Using denser sets of SNPs, the fraction of sequence included in blocks will increase but it will be accompanied by a drop in the average block size as additional blocks of smaller size are identified (Ke *et al*., 2004). Gabriel *et al*. (2002) defined blocks as regions in which 95% of the marker pairs show no strong evidence of historical recombination, the latter being based on confidence intervals between D' values. This accounts for some uncertainties, such as finite sample size and unphased diploid data. They observed that block boundaries between populations are extremely similar. This is in agreement with the recombination hot spot theory and it demonstrates that a blocklike model can be applied to the whole genome and to other populations as well. Other studies interpreted similar results differently, explaining that the majority of the blocklike patterns could be explained by population history (Patil *et al*., 2001, Phillips *et al*., 2003). However, the presence of the longest blocks had to be explained by some other mechanism, such as recombination hot spots. Both interpretations, population history and recombination hot spots, clearly play a role in shaping the genome but their relative contribution is still debated today.

Many different definitions of haplotype blocks have been published, and it is known that allelic frequency and SNP density will affect to some degree all block definitions (Zondervan and Cardon, 2004, Ke *et al*., 2004, Schulze *et al*., 2004). Denser genotyping of SNPs will reveal more blocks, shorter in length, while low allelic frequency SNPs will tend to break existing blocks into shorter ones. Also, since recombination does not occur all the time in recombination hot spots (i.e., LD can exist between adjacent blocks and recombination can occur inside blocks), block boundaries will vary between definitions. For these reasons, it is unclear yet

how to compare haplotype blocks between studies or whether the results reflect some underlying biological processes. For a review of the methods to identify and define haplotype blocks, *see* Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4. Because the Gabriel *et al*. (2002) approach is somewhat more easily understandable to biologists and because it does not need phased haplotype data, it has been more widely used up to now.

3. Building a haplotype map

Following these pilot studies, it became apparent that building a human haplotype map that would describe about 80-90% of the genome that is included in blocks and representing most of the human diversity would be extremely valuable. It was recognized that the existence of such a map would provide a valuable tool to investigators wanting to identify genes involved in common diseases. One would then have to select only a few SNPs in each block in a candidate region without losing much power in an association study (see Figure 2). It is estimated that this map would reduce by 10- to 30-fold the number of SNPs necessary to do any

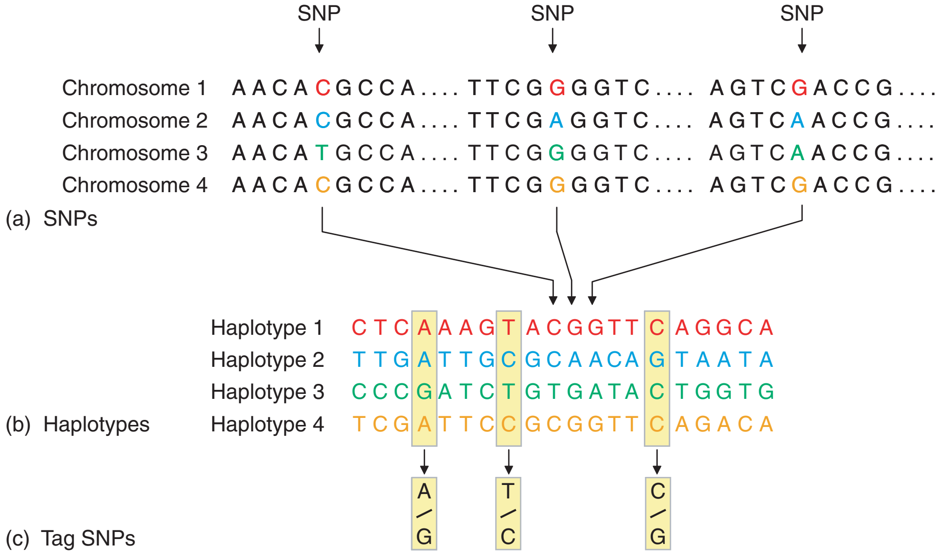


Figure 2 SNPs, haplotypes, and tag SNPs. (a) SNPs. Shown is a short stretch of DNA from four versions of the same chromosome region in different people. Most of the DNA sequence is identical in these chromosomes, but three bases are shown where variation occurs. Each SNP has two possible alleles; the first SNP in panel a has the alleles C and T. (b) Haplotypes. A haplotype is made up of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6000 bases of DNA. Only the variable bases are shown, including the three SNPs that are shown in panel a. For this region, most of the chromosomes in a population survey turn out to have haplotypes 1-4. (c) Tag SNPs. Genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely. For instance, if a particular chromosome has the pattern A-T-C at these three tag SNPs, this pattern matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population. (Reprinted with permission from Nature, The International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796. Copyright 2003 Macmillan Magazines Limited)

association study targeting common risk alleles. For whole-genome association studies, this may involve testing approximately 300 000 to 1 million SNPs, instead of the 10 to 15 million common SNPs in the human genome. The International HapMap Consortium, composed of teams from Canada, the United States, the United Kingdom, Nigeria, China, and Japan, was established in 2002 to build a map that will be useful for association mapping in any human population. Details of the approach were published in the first year of the project (International HapMap Consortium (2003)). As discussed above, although block boundaries are similar between populations of various ancestries, many differences, notably in block length and haplotype frequency, can be observed. Thus, four distinct populations (of European, Chinese, Japanese, and Yoruba ancestry) were initially chosen for their inclusion in the HapMap project. These samples are thought to include a significant amount of the overall human genetic variation and thus, the results should be applicable to most association studies. The human haplotype map should be completed by the end of 2005, but genotypes are made available to all researchers on the web as soon as they are produced. As of May 2004, more than 600 000 SNPs have been genotyped on HapMap samples.

Despite its initial successes, many questions remain unanswered. For example, will the HapMap results be applicable in a different population than those used to create the HapMap? It is known that African subpopulations can exhibit relatively greater differences between each other, as what is usually observed when comparing non-African populations. Likewise, founder populations of Caucasian origin may well present differences in haplotype structure (i.e., length of block) and diversity. It has been shown also that allelic frequency and SNP density can alter, sometimes dramatically, the observed block structure (Zondervan and Cardon, 2004, Ke *et al*., 2004, Schulze *et al*., 2004). The HapMap Consortium elected to use SNPs with a minor allelic frequency of at least 5% in each population used to build the map. This threshold is consistent with those used in the study of common variants causing common diseases (Lander, 1996, Lohmueller *et al*., 2003; *see also* Article 11, **Mapping complex disease phenotypes**, Volume 3). Low frequency variants could still, in principle, be identified using the map, although usually with much lower power to detect them (Zondervan and Cardon, 2004).

It is feasible to produce a haplotype map of a particular locus and/or in a different population. It is necessary to genotype a high-density set of SNPs on a subset of the samples and build a haplotype map that could be used to select tag SNPs to use in subsequent studies in larger cohorts from a population. The first step in building a haplotype map is to define the number of samples needed to build the map. The answer to this question depends on haplotype diversity, but in general, panels of 90 unrelated individuals or 30 trios should be sufficient to detect haplotypes with greater than 5% frequency. Using unrelated samples instead of families can maximize the number of independent chromosomes as many analysis programs can use unrelated samples to accurately determine haplotype phase, including PHASE (Stephens *et al*., 2001), EM method (Excoffier and Slatkin, 1995). However, the use of families can improve the phasing of haplotypes when LD is lower and also provide a measure of genotyping errors (using tests for Mendelian inheritance).

The second step is to select the required density of SNPs. Ke *et al*. (2004) observed that a density of atleast 1 SNP every 2 kb was needed to stabilize the

block boundaries. However, since LD is not homogeneous throughout the genome and because haplotype blocks vary greatly in length, a hierarchical approach was adopted by the International HapMap Consortium (2003). In Phase I of the HapMap project, 1 SNP every 5 kb was genotyped in every population sample to identify large blocks. This 5 kb scan is expected to generate a map with approximately 60% of the genome in blocks. Additional markers are added in regions where LD is too low (i.e., where it is not possible to accurately predict the genotype of a nearby SNP) to cover an additional 20-30% of the genome. This strategy should be valuable for any kind of haplotype map. As of March 2004, there were more than 7 million SNPs in dbSNP, with a large fraction discovered by several resequencing projects of the International HapMap Consortium (2003). More than one million SNPs will be validated by the HapMap project, providing a valuable resource to the human genetics community. It was also shown that using double-hit SNPs (i.e., SNPs independently observed two or more times in SNP discovery projects) improves the overall success rate and correlates with a higher average minor allelic frequency. The third step is to choose the appropriate genotyping platform. Many different platforms exist on the market with wide variation in throughput and price. Highly parallel processing, from a few SNPs to several thousands, has resulted in a reduction of the price per genotype. However, depending on the number of SNPs genotyped, some can be more advantageous than others (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4). The last step is to select a set of tag SNPs that will accurately describe each block in the selected samples and will be used subsequently to genotype the remaining samples (see below).

4. Selecting haplotype tag SNPs

One of the most important contributions of a haplotype map will be the description of a set of haplotype tag SNPs, specific for a population, that will improve significantly the efficiency of subsequent association studies. One straightforward method to pick htSNPs is presented in Figure 2. However, there can be as many htSNP sets as there are haplotype block definitions. Also, even though a blocklike description of the genome sounds appealing, there might be some problems if we rely on them to pick htSNPs: first, there will always be significant LD between adjacent blocks, so a strategy that would try to tag each block, especially very small ones, could be inefficient. Finally, it will be impossible to account for those regions that lie outside a block. Cardon and Abecasis (2003) showed that the best htSNP selection should not be dependent on the concept of blocks but on more general patterns of LD and haplotype diversity. This means that in a selected region, the chosen SNPs must capture the variation that exists at the unexamined sites by LD. Another important variable to consider is the balance between the reduction in the number of SNPs to genotype and the reduction in the power to detect an association.

Zhang *et al*. (2002b) noted that the power is reduced by about 4% when 25% of the total number of SNPs, selected by their method, is used, compared to a drop of 12% when a random SNP set of similar size is chosen instead. When 14% of the SNPs are used, the power drops by 9 and 21%, respectively. Also, power

loss was much greater using single-locus association tests than with a two-locus haplotype approach. Although other published methods seem to perform similarly, the different algorithms do not identify the same number of blocks or htSNPs. For example, Zhang *et al*. (2002a), using the same data as Patil *et al*. (2001), were able to reduce the number of haplotype blocks and htSNPs selected by more than 20%. Various other methods have recently been published that attempt to minimize the number of tag SNPs (Ke and Cardon, 2003, Meng *et al*., 2003, Carlson *et al*., 2004, Schulze *et al*., 2004, Sebastiani *et al*., 2003, Stram *et al*., 2003), pointing to the fact that the optimal way to mathematically define the problem has probably not yet been described or will vary on the basis of local LD in different regions of the genome. Since r^2, a measure of LD, is inversely proportional to the increase in the sample size required to detect an indirect association, the most useful approach, with the knowledge of the Haplotype Map, would be to pick the minimal set of htSNPs that describe all other markers using a r^2 threshold (ex:0.8). This allows the sample size needed to detect the association to be easily calculated (Carlson *et al*., 2003).

In the past decade, significant progress was made in defining the genetic basis of common diseases as well as in genotyping technologies and statistical methodologies. However, the mapping of common disease genes is still an arduous task today. The understanding of the haplotype structure of human chromosomes is crucial to improve the success of association studies. With results from the HapMap project available to all researchers, many studies that previously would have been done with a few nonvalidated SNPs can now be attempted with sufficient power to effectively screen hundreds of candidate genes and large chromosomal regions. It is important to realize that the haplotype mapping approaches are still in their infancy and need to be improved for the full potential of the Human Haplotype Map to be achieved: (1) genotyping costs need to decrease; (2) analysis methods must improve; and (3) pilot studies must evaluate the effect of using different populations or the impact of disease polymorphisms, with various penetrance and allelic frequency, on the success of association mapping of common traits. Despite the caveats, the Haplotype Map will be a powerful tool to systematically screen the genome and elucidate the genetic causes of common genetic diseases.

References

Abecasis GR and Cookson WO (2000) GOLD–graphical overview of linkage disequilibrium. *Bioinformatics*, **16**, 182–183.

Allen M, Heinzmann A, Noguchi E, Abecasis G, Broxholme J, Ponting CP, Bhattacharyya S, Tinsley J, Zhang Y, Holt R, *et al*. (2003) Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nature Genetics*, **35**, 258–263.

Cardon LR and Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, **19**, 135–140.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**, 518–521.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L and Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, **74**, 106–120.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Excoffier L and Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–927.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D and Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Research*, **11**, 143–151.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Giallourakis C, Stoll M, Miller K, Hampe J, Lander ES, Daly MJ, Schreiber S and Rioux JD (2003) IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *American Journal of Human Genetics*, **73**, 205–211.

Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, Gudmundsson G, Grant SF, Thorgeirsson G, *et al*. (2004) The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nature Genetics*, **36**, 233–239.

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, *et al*. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics*, **26**, 163–175.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**, 217–222.

Kauppi L, Sajantila A and Jeffreys AJ (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Human Molecular Genetics*, **12**, 33–40.

Ke X and Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**, 287–288.

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, *et al*. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, **13**, 577–588.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, **31**, 241–247.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139–144.

Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**, 234–236.

Lander ES (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.

Lander E and Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

Lien S, Szyda J, Schechinger B, Rappold G and Arnheim N (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *American Journal of Human Genetics*, **66**, 557–566.

Lohmueller KE, Pearce CL, Pike M, Lander ES and Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**, 177–182.

Lonjou C, Collins A and Morton NE (1999) Allelic association between marker loci. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 1621–1626.

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, *et al*. (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *American Journal of Human Genetics*, **67**, 383–394.

May CA, Shone AC, Kalaydjieva L, Sajantila A and Jeffreys AJ (2002) Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nature Genetics*, **31**, 272–275.

Meng Z, Zaykin DV, Xu CF, Wagner M and Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *American Journal of Human Genetics*, **73**, 115–130.

Mira MT, Alcais A, Nguyen VT, Moraes MO, Di Flumeri C, Vu HT, Mai CP, Nguyen TH, Nguyen NB, Pham XK, *et al*. (2004) Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature*, **427**, 636–640.

Mohlke KL, Lange EM, Valle TT, Ghosh S, Magnuson VL, Silander K, Watanabe RM, Chines PS, Bergman RN, Tuomilehto J, *et al*. (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Research*, **11**, 1221–1226.

Oksenberg JR, Barcellos LF, Cree BA, Baranzini SE, Bugawan TL, Khan O, Lincoln RR, Swerdlin A, Mignot E, Lin L, *et al*. (2004) Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. *American Journal of Human Genetics*, **74**, 160–167.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Peltonen L (2000) Positional cloning of disease genes: advantages of genetic isolates. *Human Heredity*, **50**, 66–75.

Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nature Reviews. Genetics*, **2**, 360–369.

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, *et al*. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics*, **33**, 382–387.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, *et al*. (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics*, **29**, 223–228.

Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Schneider JA, Peto TE, Boone RA, Boyce AJ and Clegg JB (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. *Human Molecular Genetics*, **11**, 207–215.

Schulze TG, Zhang K, Chen YS, Akula N, Sun F and McMahon FJ (2004) Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Human Molecular Genetics*, **13**, 335–342.

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS and Ramoni MF (2003) Minimal haplotype tagging. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9900–9905.

Service SK, Ophoff RA and Freimer NB (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate. *Human Molecular Genetics*, **10**, 545–551.

Smith RA, Ho PJ, Clegg JB, Kidd JR and Thein SL (1998) Recombination breakpoints in the human beta-globin gene cluster. *Blood*, **92**, 4415–4421.

Stefansson H, Sarginson J, Kong A, Yates P, Steinthorsdottir V, Gudfinnsson E, Gunnarsdottir S, Walker N, Petursson H, Crombie C, *et al*. (2003) Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. *American Journal of Human Genetics*, **72**, 83–87.

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, *et al*. (2002) Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics*, **71**, 877–892.

Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE and Thomas DC (2003) Modeling and E-M estimation of haplotype-specific

relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity*, **55**, 179–190.

The International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796

Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, *et al*. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature*, **418**, 426–430.

Weedon MN, Schwarz PE, Horikawa Y, Iwasaki N, Illig T, Holle R, Rathmann W, Selisko T, Schulze J, Owen KR, *et al*. (2003) Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *American Journal of Human Genetics*, **73**, 1208–1212.

Werner M, Herbon N, Gohlke H, Altmuller J, Knapp M, Heinrich J and Wjst M (2004) Asthma is associated with single-nucleotide polymorphisms in ADAM33. *Clinical and Experimental Allergy*, **34**, 26–31.

Zhang K, Calabrese P, Nordborg M and Sun F (2002a) Haplotype block structure and its applications to association studies: power and study designs. *American Journal of Human Genetics*, **71**, 1386–1394.

Zhang K, Deng M, Chen T, Waterman MS and Sun F (2002b) A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7335–7339.

Zhang Y, Leaves NI, Anderson GG, Ponting CP, Broxholme J, Holt R, Edser P, Bhattacharyya S, Dunham A, Adcock IM, *et al*. (2003) Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nature Genetics*, **34**, 181–186.

Zondervan, KT and Cardon, LR (2004) The complex interplay among factors that influence allelic association. *Nature Reviews. Genetics*, **5**, 89–100.

Specialist Review

YAC-STS content mapping

Claudia Gosele, Heike Zimdahl and Norbert Hubner
Max-Delbruck-Center for Molecular Medicine, Berlin, Germany

1. STS marker

Sequence tagged site (STS) markers represent short unique DNA sequences for which a specific PCR (polymerase chain reaction) assay can be designed, so that any DNA sample can be easily tested for the presence or absence of this specific DNA fragment. One of the hallmarks of an STS is that it maps to a single location in the genome. STSs were proposed as a "common language for physical mapping" (Olson *et al*., 1989), and have emerged as principal markers used in a variety of mammalian maps (*see* Article 9, **Genome mapping overview**, Volume 3). The principle of STS mapping is to create clone maps based on STS content, allowing the correlation between different maps.

A YAC (yeast artificial chromosome)-based STS content mapping approach was first used to construct a map of human chromosome 7 (Green *et al*., 1991) in 1991. The proposed strategy was based on the mapping of large segments of human DNA using YACs as the source of cloned DNA and STSs as landmarks to order these clones and anchor them with radiation hybrid maps. The principle disadvantage of any STS marker approach – compared to hybridization-based markers – lies in the cost of acquiring single-copy sequences and synthesizing oligonucleotides specific to each STS.

2. STS marker development

PCR-based STS content mapping is a highly robust and straightforward technique. It can be applied to all kinds of cloning systems and mapping panels. The scale of STS development can vary from creating a single STS corresponding to an interesting gene sequence for the purpose of locating the gene in the genome, to creating a large number of STSs from a particular chromosome or subchromosomal region as part of a physical mapping project.

Regardless of the origin of STSs or scale of a project, the major steps of STS development are:

1. acquisition of sequence
2. assessment of sequence by comparison of databases

3. selection of PCR primers
4. development of a PCR assay
5. assessment of the uniqueness of an STS by limited mapping
6. assessment of STS quality.

Many genomes contain abundant repetitive sequences, including interspersed repetitive repeats, satellite sequences, and gene families that should be avoided in the development of an STS. Sequences can be compared to databases of known repeat sequences, for example, by using the web-accessible program RepeatMasker (Smit *et al*., 1996–2004).

The primer design should be done on unique sequences. Several algorithms are available for selecting oligonucleotides for PCR (e.g., PRIMER, OLIGO). PCR products that are used to detect an STS are usually 100–1000 bp size, with an average size of approximately 200 bp. If multiplex PCR is used to score multiple STSs simultaneously, a range of product sizes should be selected to facilitate resolution of products from the reaction mixture.

When primers are selected from a cDNA sequence, it is important to select oligonucleotides that are likely to lie within a single exon, to ensure that genomic DNA can serve as a template for the amplification of the STS. Since PCR-based STS content mapping demands a primer design and an oligonucleotide for each marker studied, the costs of a project are increased. Furthermore, PCR has to be done on every clone of a library, followed by the detection of products, which results in a large number of reactions. For screening whole libraries, the collection of clones in pools can circumvent this problem and reduces drastically the number of PCR reactions to be done (see Figure 1).

The next step in STS generation includes the development of a robust PCR assay that can be carried out under standardized reaction conditions. A robust and standardized assay is especially important for projects that involve the generation of more than a small number of STSs because a large number of PCR assays will be carried out.

As an alternative to PCR, genomic libraries can be spotted on high-density gridded filters and several thousands of clones can be screened in a single hybridization step (see Section 4.2).

3. Large-insert clone libraries

With the availability of one or more closely linked DNA markers from a genomic region of interest, one can begin to develop a contig of overlapping clones that spans the region. A cloned contig not only provides information on physical distances but can also be used as the raw material from which positional cloning of a phenotypically defined locus can proceed. The generation of a contig is pursued most efficiently by screening a large-insert genomic library. Although a number of systems for generating large-insert libraries have been described, to date, bacterial artificial chromosomes (BAC) and YAC cloning systems have been used most widely.

The development of YAC cloning technology, first implemented by David Burke and Maynard Olson at Washington University in St. Louis (Burke *et al*., 1987), has

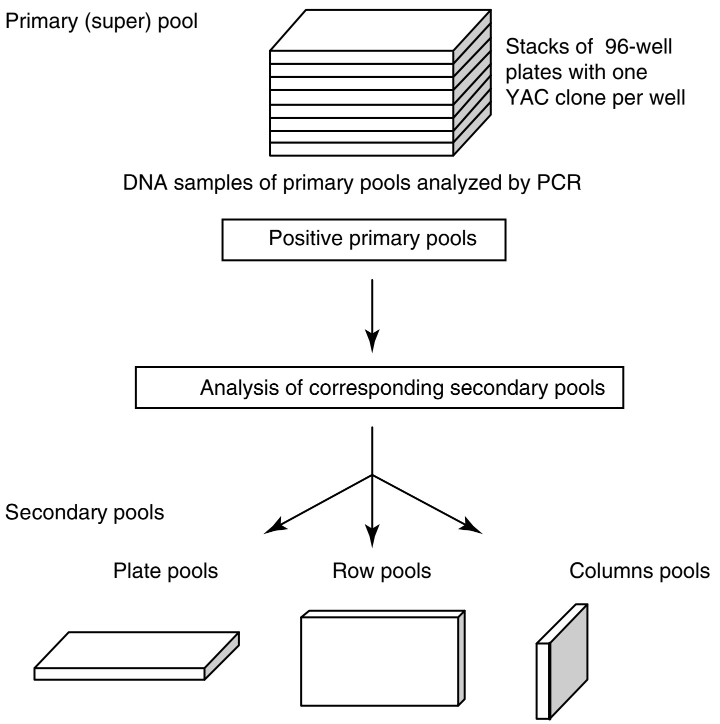


Figure 1 Three-dimensional pooling of clone libraries: A YAC library is usually stored in 96-well plates: one stack consists of 8 plate pools, 8 row pools, and 12 column pools, and PAC/BAC libraries are stored in 384-well plates: one stack consists of 8 plate pools, 16 row pools, and 24 column pools. By pooling clone libraries, the number of PCRs for the library screening is significantly reduced

enabled the cloning of very large fragments of exogenous DNA that range in size up to 2 Mb and thus directly enhanced the relationship between genetic, physical, and functional mapping of genomes. YAC cloning systems are based on yeast plasmids, containing DNA sequences that function as telomeres (TEL), as well as containing yeast origin of replication (ARS) and centromere segments (CEN). "Artificial" yeast chromosomes are formed by ligation of random, large fragments of genomic DNA between two arms that contain, in one case, a telomere and a centromere, and in the other case, a telomere alone, with selectable drug-resistance markers on both arms. These YAC constructs are transfected back into yeast where they will move alongside host chromosomes into both daughter cells at each mitotic division.

The construction of a YAC library proceeds in a manner that is very different from that of most other types of genomic libraries. Every clone in the library must be picked individually and placed into a separate compartment (e.g., of a microtiter dish). This process is extremely time consuming and labor intensive, but once a library has been formed with individual clones in individual wells, it is essentially immortal. For this reason and others, it makes good sense to screen established libraries for a gene of interest rather than to create a new library.

The first human YAC library to be described had a 2.2-fold genomic coverage and an average insert size of ∼265 kb, and was distributed freely to the entire scientific community (Burke *et al*., 1987).

Although YAC clones have facilitated the construction of long-range physical maps, it should be mentioned that the YAC cloning system is not perfect. A percentage of clones within a YAC library are chimeric; that is, their inserts are composed of two or more unrelated genomic fragments that have become ligated together as an artifact of the cloning process. The preidentification of chimeric clones is essential before one can begin to generate a physical map. The disadvantages (e.g., the chimerism and the instability) can limit the utility of YAC libraries and restrict their purposes.

Two other systems for cloning large genomic inserts have been described more recently, which offer high clonal stability, reduced cloning biases, and are easily purified for DNA sequencing: the PAC (bacteriophage P1-derived artificial chromosomes) (Ioannou *et al*., 1994) and BAC (bacterial artificial chromosomes) (Shizuya and Kouros-Mehr, 2001).

The PAC system is based on the use of the bacteriophage P1 as a cloning vector (Pierce *et al*., 1992; Pierce and Sternberg, 1992). This system has been used to obtain a mouse genomic library with average inserts in the range of 75–95 kb with a maximum cloning capacity of 100 kb. The P1 cloning system has two advantages over YACs: first, it has much more efficient cloning rates, and second, like other bacterial cloning systems, it allows the efficient purification of large amounts of clone DNA away from the rest of the bacterial genome. The utility of this cloning system in the analysis of genomic organization within the H2 region has been demonstrated (Gasser *et al*., 1994).

The BAC system (bacterial artificial chromosome) is derived from the well-studied *E. coli* F factor, which is essentially a naturally occurring single-copy plasmid (Shizuya *et al*., 1992). This plasmid has been converted into a vector that allows the cloning of inserts with more than 300 kb of DNA, and with a reported average size range of 200–300 kb. The BAC system has the same advantages as P1 and the added advantage of a larger potential insert size.

4. Physical mapping

Physical mapping is a means to map genes or DNA sequences on a chromosome without relying on meiotic segregation (genetic mapping). Physical maps provide molecular access to chromosomal regions of interest and therefore facilitate the positional cloning of genes. Ideally, the series of clones should be gap-free and form a contiguous clone array. For the analysis of a specific chromosomal region, like QTL regions, a high-resolution physical map is desirable. A series of identified clones are assembled to provide a full representation of the region of interest.

Common physical mapping strategies (*see* Article 9, **Genome mapping overview**, Volume 3) are based on the PCR screening of large-insert cloning libraries using appropriate markers. An extremely powerful and convenient tool for ordering sequences according to their chromosomal position is a radiation hybrid panel (*see* Article 14, **The construction and use of radiation hybrid maps in**

genomic research, Volume 3). The localization of any STS can be achieved by PCR amplification and scoring of a particular radiation hybrid panel. The obtained patterns are compared with patterns of previously mapped markers held on a central server.

4.1. PCR-based STS mapping

STS content mapping involves scoring a series of clones for the presence or absence of particular STSs. STS content mapping is often used as a tool to assemble clone contigs. Large-insert genomic libraries (e.g., YAC libraries) are usually screened by PCR to amplify specific sequences contained within the genomic clones. These STSs serve as unique identifiers to ascertain whether the DNA fragments are located within other genomic clones to establish an overlap between clones. The number of PCRs for library screening can be significantly reduced by pooling YAC clones (see Figure 1).

Chromosomal walking by clone–clone hybridization is not practically feasible with mammalian YACs: the large amount of repetitive DNA in the inserts means that blocking of the repetitive DNA signal during hybridization is technically difficult. Instead, techniques are used to recover short-end fragments from individual YACs, for example, by restriction enzyme digestion or PCR amplification. Usually, the YAC DNA is cleaved with a restriction enzyme that is known to cleave the YAC vector sequence. Among the cleavage products, there will be fragments containing both the unknown terminal sequence from the insert DNA and the adjacent known vector sequence. Such sequences can be amplified using various PCR-based methods by using a primer binding to the vector sequence to permit access to an adjacent uncharacterized sequence (e.g., Inverse-PCR, Vectorette PCR).

YAC insert end fragments can then be used as hybridization probes to screen colony filters from a YAC library (see Section 4.2 below), or are sequenced to design oligonucleotide primers to permit a PCR assay for this sequence. PCR can then be used again to screen pooled libraries for identifying YAC with overlapping sequences. All positive clones from a YAC, or other large-insert libraries can be sized by PFGE, and fragments at both ends of each insert can be isolated rapidly by several standard protocols (Riley *et al*., 1990; Cox *et al*., 1993). End fragments from each clone should be used as probes to perform an initial test of the possibility of chimerism. This can be accomplished by probing appropriate somatic cell hybrid lines to determine whether both ends map to the same chromosome as the original DNA marker used to isolate the clone; if appropriate somatic cell hybrid lines are not available, one can also test the segregation of the end fragments on a panel of 20 interspecific (or intersubspecific) backcross samples. If the two end fragments show complete concordance in transmission, this can be taken as strong evidence for nonchimerism; in contrast, two or more recombination events would be highly suggestive of a chimeric clone. Chimeric clones need not be discarded; it is just necessary to be aware of their nature in any interpretation of the data that they generate.

The process of deriving YAC clones from a library can be brought to a halt when the clones that have already been obtained include the locus being sought.

It is only possible to reach this conclusion when the derived contig extends over markers that map apart from the locus on both of its sides. In other words, the contig must extend across the two closest recombination break points that define the outer limits of localization. If cloning is begun with a very dense map of markers placed onto a high-resolution cross, this endpoint is likely to be reached more quickly. With real luck, it might even be reached with the first set of YACs obtained in the initial screening of the library.

The interrepeat or interspersed repetitive sequence (IRS)-PCR is another means of isolating sequences that are specific to a chromosomal region of interest, for example, using somatic cell hybrids. IRS-PCR is an exceptional type of PCR in which genomic sequences located between two highly repetitive SINE elements are amplified (Ledbetter *et al*., 1990) (see Figure 2). The mammalian genome contains a large amount of highly repeated DNA sequence families, which are largely transcriptionally inactive. A wide variety of different repeats are known and they are classified into two major types of organization: tandemly repeated and interspersed repeated sequences. Tandemly repeated noncoding DNA is mainly grouped into three subclasses depending on the average size: satellite, mini-, and microsatellite DNA. Interspersed repetitive noncoding DNA sequences are not clustered, but are dispersed and a number of different classes exist. Most of the DNA families contain some members that are capable of undergoing retrotransposition (Deininger, 1989; Daniels and Deininger, 1985). Two major classes have been discerned on the basis of repeat unit length: SINEs (short interspersed nuclear elements) with an average size of 0.1 to 0.3 kb and LINEs (long interspersed nuclear elements) with an average size between 0.3 and 8 kb. The Alu element, with more than one million copies, is the most abundant SINE in humans. Because of the high copy number, the Alu gene family comprises more than 10% of the human genome (Lander *et al*., 2001).

IRS-PCR strategies have been applied to various species using the SINE sequences in human (Alu repeat) (Nelson *et al*., 1991), mouse (B1 repeat) (Hunter *et al*., 1996; McCarthy *et al*., 1995), rat (ID repeat) (Gosele *et al*., 2000), and zebrafish (DANA/mermaid repeat) (Shimoda *et al*., 1996). Since these SINE elements are present at a very high frequency in the respective genome, two of such repeats will often be found in close proximity and sometimes in opposite orientations. A single primer corresponding to a sequence close to the end of the repeat consensus sequence can bind to each of two closely located, oppositely orientated repeat sequences (Nelson *et al*., 1989). IRS-PCR can be applied on any genomic DNA template. If the starting DNA is complex (e.g., genomic DNA, cell hybrid DNA), the IRS-PCR will generate a series of bands that cannot be

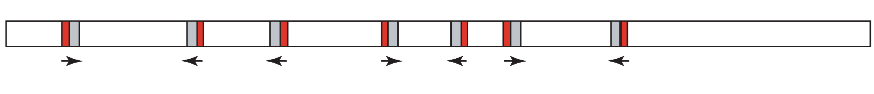


Figure 2 IRS-PCR permits amplification of DNA sequences located between two closely positioned but oppositely orientated repeat elements. A single primer corresponding to a sequence close to the end of the repeat consensus sequence can bind to each of two closely located, oppositely orientated repeat sequences. Amplification is carried out with a single primer complementary to the 5′-3′ repetitive element

resolved by conventional agarose gel electrophoresis. The size of amplified IRS-PCR products ranges from 200 to 2000 bp. If IRS-PCR is carried out on low-complexity templates (e.g., YAC, PAC or BAC clones), one or a few products will be generated, which can directly be exploited as markers (Schalkwyk *et al*., 2001). The majority of IRS probes consist of unique single-copy DNA sequences (except for incorporated primers). Therefore, IRS-PCR products from large-insert genomic clones can be compared with equivalent products from other library clones to check for overlapping clones based on IRS amplification products in common, and also used as hybridization probes for screening large-insert clone libraries (see Section 4.2). The generation of large numbers of IRS markers is not only rapid but also cost-efficient, because there is no requirement to sequence markers or to design locus-specific primers.

4.2. Hybridization-based STS mapping

Once YAC clones are obtained by PCR-based screening of YAC libraries, YAC end fragments can be isolated as described above. YAC insert end sequences can then be used as hybridization probes to screen colony filters from a YAC, PAC, or BAC library to identify adjacent clones.

An alternative strategy for integrated physical and genetic mapping is based on the interrepeat or interspersed repetitive sequence (IRS)-PCR system, which enables a high-throughput screening of large numbers of clones in a time and cost-efficient way by hybridization.

The IRS-PCR-based physical mapping strategy relies on the ability to detect clone overlaps by hybridization of an individual IRS-PCR product to IRS-PCR product pool filter. Therefore, coordinate pools that define a library are amplified with the repetitive element primer. These IRS-PCR pool products are spotted onto nylon membranes in ordered arrays. Individual IRS-PCR products of clones (previously mapped) are utilized as hybridization probes to identify overlapping clones. Overlapping clones are then amplified again via IRS-PCR and PCR products are used as hybridization probes (see Figure 3). Clone contigs are built by repeating these rounds of PCR and hybridization. Chromosome walking based on IRS-PCR and hybridization is bidirectional and, therefore, highly efficient.

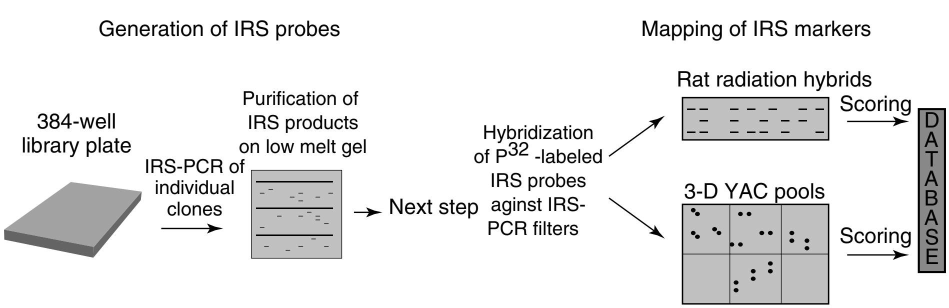


Figure 3 Schematic representation of the IRS-PCR-based physical mapping strategy for the construction of an integrated radiation hybrid and physical map

The main advantage is that this approach permits the simultaneous screening of multiple probes and/or libraries in one hybridization step, avoiding running of gels, sequencing, and primer design. This technology was utilized for the construction of clone contigs in mouse (Hunter *et al*., 1994), physical mapping of the mouse (Hunter *et al*., 1996; McCarthy *et al*., 1995; Schalkwyk *et al*., 2001) and the rat genome (Krzywinski *et al*., 2004). For the physical mapping of the rat genome, two mapping methods were combined to gain information about the proximity of marker loci: YAC libraries were screened by hybridization-based assays to identify clones containing a given locus. Nearby loci tend to be present in many of the same clones, allowing proximity to be inferred. Marker-content linkage can be detected over distances of about 800 kb, given the average insert size of the YAC library used (Figure 4).

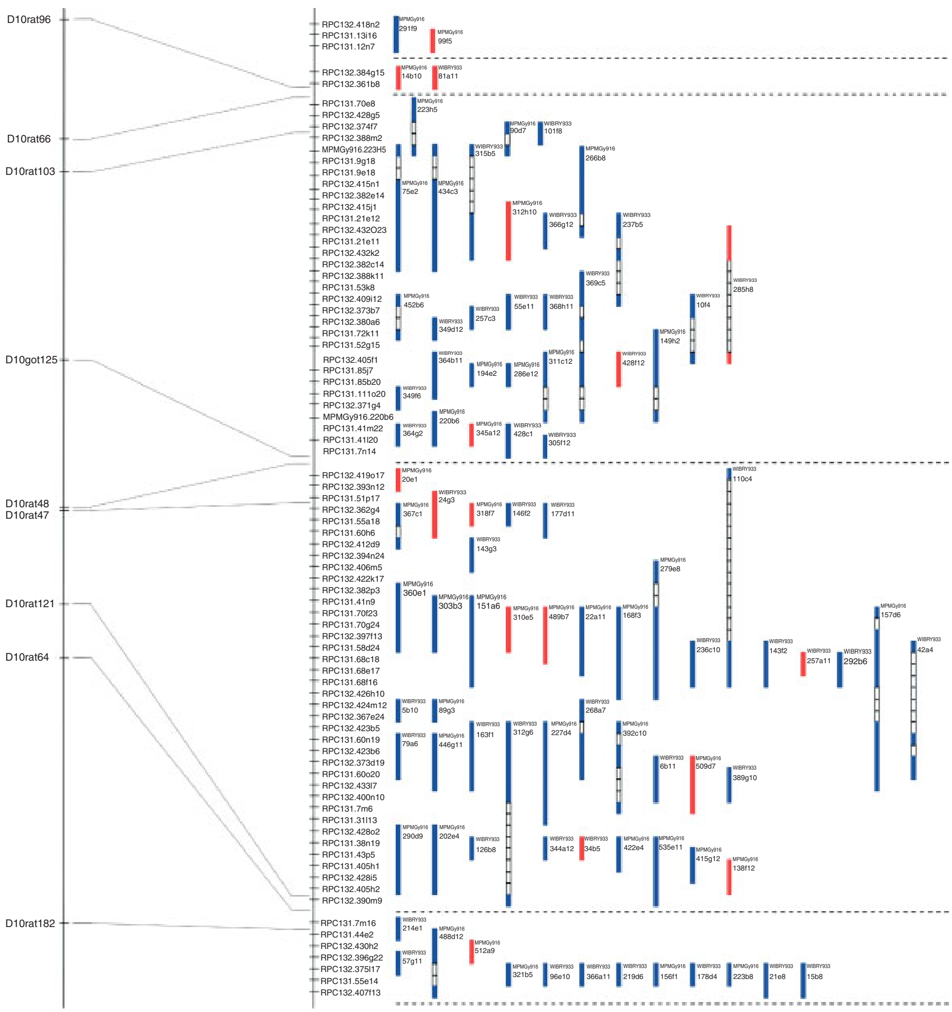


Figure 4 YAC contig of a region of the rat chromosome 10, constructed by IRS-PCR-based physical mapping

Hybrid cell lines, each containing many chromosomal fragments produced by radiation breakage, are screened to identify those hybrids that have retained a given locus. Nearby loci tend to show similar retention patterns, allowing proximity to be inferred. RH linkage can be detected for distances of about 2–3 Mb, given the average fragment size of the RH panel used. For the construction of a physical map and assembly of contigs, YAC clones with positive hybridization signals are considered. The number of YAC clones has to be pruned with considerable care toward chimeric clones, an inherent problem with any YAC library (Green *et al*., 1999). The final map of each chromosome can be constructed by integrating the YAC-linkage information with the known radiation hybrid map positions of the IRS markers using, for example, the co2 software package (Hudson *et al*., 1995). Doubly linked contigs are identified and then single linkage information is used to join doubly linked contigs known to lie nearby.

The primary goal of physical mapping is to assemble a comprehensive series of DNA clones with overlapping inserts. Clone-based laboratory methods maintain an important component to study large genomes through applications such as fluorescent in situ hybridization and comparative genomic hybridization. Physical maps provide an ordered, high-resolution, redundant clone set spanning the entire genome, which will be an important resource for the easy identification and access to clones spanning regions of interest in the relevant genome. The generation of in-depth physical maps will therefore continue to be a desired component for functional genomics.

Further reading

Chumakov I, Rigault P, Guillou S, Ougen P, Billaut A, Guasconi G, Gervy P, LeGall I, Soularue P, Grinas L, *et al*. (1992) Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature*, **359**, 380–387.

References

Burke DT, Carle GF and Olson MV (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, **236**, 806–812.

Cox RD, Whittington J, Shedlovsky A, Connelly CS, Dove WF, Goldsworthy M, Larin Z and Lehrach H (1993) Detailed physical and genetic mapping in the region of plasminogen, D17Rp17e, and quaking. *Mammalian Genome*, **4**, 687–694.

Daniels GR and Deininger PL (1985) Repeat sequence families derived from mammalian tRNA genes. *Nature*, **317**, 819–822.

Deininger PL (1989) SINEs: short interspersed repeated DNA elements in higher eukaryotes. In *Mobile DNA*, Berg DE and Howe MM (Eds.), American Society for Microbiology: Washington, pp. 619–636.

Gasser DL, Sternberg NL, Pierce JC, Goldner-Sauve A, Feng H, Haq AK, Spies T, Hunt C, Buetow KH and Chaplin DD (1994) P1 and cosmid clones define the organization of 280 kb of the mouse H-2 complex containing the Cps-1 and Hsp70 loci. *Immunogenetics*, **39**, 48–55.

Gosele C, Hong L, Kreitler T, Rossmann M, Hieke B, Gross U, Kramer M, Himmelbauer H, Bihoreau MT, Kwitek-Black AE, *et al*. (2000) High-throughput scanning of the rat genome using interspersed repetitive sequence-PCR markers. *Genomics*, **69**, 287–294.

Green ED, Hieter P and Spencer FA (1999) Yeast artificial chromosomes. In *Genome analysis-A laboratory manual*, Birren B, *et al*. (Ed.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, pp. 479–487.

Green ED, Mohr RM, Idol JR, Jones M, Buckingham JM, Deaven LL, Moyzis RK and Olson MV (1991) Systematic generation of sequence-tagged sites for physical mapping of human chromosomes: application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics*, **11**, 548–564.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al*. (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.

Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, Baptista R, Kruglyak L, Xu SH, *et al*. (1995) An STS-based map of the human genome. *Science*, **270**, 1945–1954.

Hunter KW, Ontiveros SD, Watson ML, Stanton VP Jr, Gutierrez P, Bhat D, Rochelle J, Graw S, Ton C, Schalling M, *et al*. (1994) Rapid and efficient construction of yeast artificial chromosome contigs in the mouse genome with interspersed repetitive sequence PCR (IRS-PCR): generation of a 5-cM, >5 megabase contig on mouse chromosome 1. *Mammalian Genome*, **5**, 597–607.

Hunter KW, Riba L, Schalkwyk L, Clark M, Resenchuk S, Beeghly A, Su J, Tinkov F, Lee P, Ramu E, *et al*. (1996) Toward the construction of integrated physical and genetic maps of the mouse genome using interspersed repetitive sequence PCR (IRS-PCR) genomics. *Genome Research*, **6**, 290–299.

Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA and de Jong PJ (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics*, **6**, 84–89.

Krzywinski M, Wallis J, Gosele C, Bosdet I, Chiu R, Graves T, Hummel O, Layman D, Mathewson C, Wye N, *et al*. (2004) Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. *Genome Research*, **14**, 766–779.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Ledbetter SA, Nelson DL, Warren ST and Ledbetter DH (1990) Rapid isolation of DNA probes within specific chromosome regions by interspersed repetitive sequence polymerase chain reaction. *Genomics*, **6**(3), 475–481.

McCarthy L, Hunter K, Schalkwyk L, Riba L, Anson S, Mott R, Newell W, Bruley C, Bar I, Ramu E, *et al*. (1995) Efficient high-resolution genetic mapping of mouse interspersed repetitive sequence PCR products, toward integrated genetic and physical mapping of the mouse genome. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5302–5306.

Nelson DL, Ballabio A, Victoria MF, Pieretti M, Bies RD, Gibbs RA, Maley JA, Chinault AC, Webster TD and Caskey CT (1991) Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: identification of clones associated with a disease locus. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6157–6161.

Nelson DL, Ledbetter SA, Corbo L, Victoria MF, Ramirez-Solis R, Webster TD, Ledbetter DH and Caskey CT (1989) Alu polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA sources. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 6686–6690.

Nusbaum C, Slonim DK, Harris KL, Birren BW, Steen R, Stein LD, Miller J, Dietrich WF, Nahf R, Wang V, *et al*. (1999) A YAC-based physical map of the mouse genome. *Nature Genetics*, **22**, 388–393.

Olson M, Hood L, Cantor C and Botstein D (1989) A common language for physical mapping of the human genome. *Science*, **245**, 1434–1435.

Pierce JC, Sauer B and Sternberg N (1992) A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: improved cloning efficacy. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 2056–2060.

Pierce JC and Sternberg NL (1992) Using bacteriophage P1 system to clone high molecular weight genomic DNA. *Methods in Enzymology*, **216**, 549–574.

Riley J, Butler R, Ogilvie D, Finniear R, Jenner D, Powell S, Anand R, Smith JC and Markham AF (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Research*, **18**, 2887–2890.

Schalkwyk LC, Cusack B, Dunkel I, Hopp M, Kramer M, Palczewski S, Piefke J, Scheel S, Weiher M, Wenske G, *et al*. (2001) Advanced integrated mouse YAC map including BAC framework. *Genome Research*, **11**, 2142–2150.

Shimoda N, Chevrette M, Ekker M, Kikuchi Y, Hotta Y and Okamoto H (1996) Mermaid, a family of short interspersed repetitive elements, is useful for zebrafish genome mapping. *Biochemical and Biophysical Research Communications*, **220**, 233–237.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8794–8797.

Shizuya H and Kouros-Mehr H (2001) The development and applications of the bacterial artificial chromosome cloning system. *The Keio Journal of Medicine*, **50**, 26–30.

Smit AFA, Hubley R and Green P (1996–2004) RepeatMasker at **http://www.repeatmasker.org**.

Specialist Review

The construction and use of radiation hybrid maps in genomic research

Mathieu Gautier and André Eggen
Laboratoire de Génétique Biochimique et de Cytogénétique, Jouy-en-Josas, France

1. Introduction

At the beginning of the sixties, Barski *et al*. (1960) reported the occurrence of spontaneous cellular fusion events. A few years later, new methodologies using Sendai virus inactivated by UV (Yerganian and Nell, 1966) or polyethylene glycol (PEG) treatments made it possible to induce cell fusion and to produce heterocaryotic somatic hybrid cells (Pontecorvo, 1975). In addition, hybrid cells, originally called recombinant cells, could be separated from the parental cells from which they originated by culturing the cells in selective media. In 1975, Goss and Harris proposed for the first time the use of these techniques for genetic mapping purposes: they had observed that lethally X-ray irradiated donor cells could fuse to a receptor cell giving rise to a viable hybrid cell line. This so-called radiation hybrid (RH) cell line possessed a heterocaryon with chromosomes corresponding to a mosaic of chromosome fragments from the donor and the receptor cells: the stronger the irradiation dose, the shorter the average expected size of the donor chromosome fragments was. The principle is thus very similar to the classical linkage mapping strategy since X-ray induced breakage points mimic recombination points produced during meiosis.

In this review, we will consider the main principles of the construction and characterization of RH panels, their advantages to other mapping tools for the development of high-resolution genetic and comparative maps, and their possible contributions in different genome mapping projects.

2. Construction of RH panels: principles

2.1. Production and selection of RH cells

As shown in Figure 1, nucleotide biosynthesis can be accomplished through two biological pathways, that is, the main and the salvage pathways. In selective HAT (Hypoxanthin Aminopterin Thymidine) medium, aminopterin blocks the main pathway while the two precursors (hypoxanthin and thymidine) are available for the

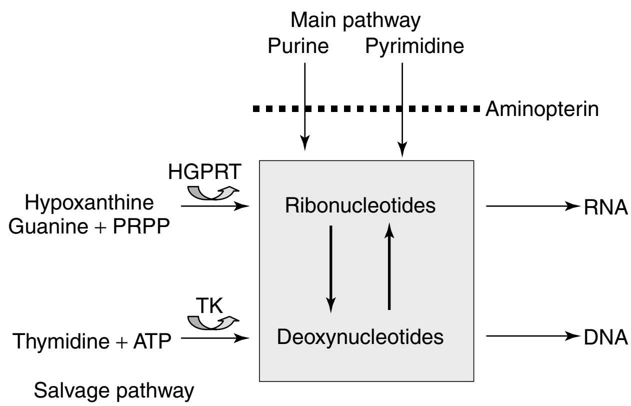


Figure 1 Nucleotide biosynthesis pathway. In the selective HAT medium, the main pathway is blocked by aminopterin (A of HAT), a structural analog of folic acid. Cells need the two precursors hypoxanthine (H of HAT) and thymidine (T of HAT) to produce respectively ribo- and deoxyribonucleotides through the salvage pathway. Mutant cells deficient, either for HGPRT (hypoxanthine-guanine phosphoribosyl transferase) or TK (thymidine kinase), cannot use the salvage pathway and thus will die

salvage pathway. To construct an RH cell line, the chosen receptor cell line (usually hamster or mouse) is deficient either for thymidine kinase (TK) or hypoxanthine-guanine phosphoribosyl transferase (HGPRT). After X-ray irradiation of the donor cells from the species of interest, usually with a dose between 3000 and 12 000 γ-ray Rad, a fusion with the receptor cells is induced. When grown in selective HAT medium, nonfused (tk^- or $hgprt^-$) receptor cells lacking one of the two key enzymes for the salvage pathway and lethally irradiated nonfused donor cells will be counterselected. Only RH cells in which the deficient *tk* or *hgprt* gene from the receptor cell has been complemented by its functional counterpart carried by one of the integrated chromosomal donor fragments will then survive (Figure 2).

During the culture of RH cell lines, chromosomal segments originating from the donor cells are randomly eliminated, while chromosomes from the receptor cell are conserved. To our knowledge, the mechanism behind this phenomenon has not yet been elucidated. Consequently, each independent RH cell line constituting an RH panel (usually composed of about one hundred lines) will contain a different set of chromosomal segments from the donor genome. Nevertheless, a bias will remain for the genomic region containing the marker gene of selection (e.g., in man, *tk* and *hgprt* are respectively located on HSA17 and HSAX chromosomes and, therefore, the corresponding regions will be preferably retained).

From a practical point of view, to allow for experimental replication and data comparisons using the same lines, it is important to extract sufficiently large quantities of DNA from each line since additional rounds of culture of the RH cell lines will lead to a set of donor chromosomal segments different from the original one. The main principles of the production and selection of RH cell lines are summarized in Figure 2.

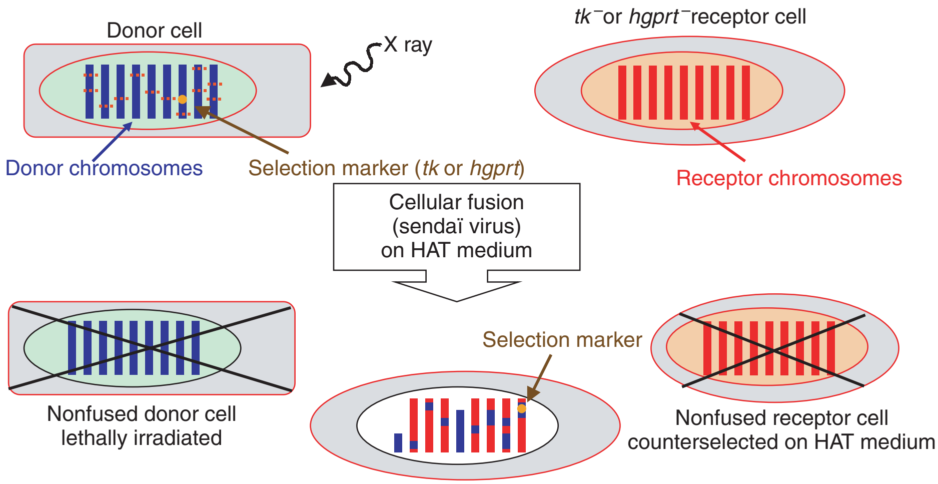


Figure 2 Schematic representation of the construction of RH cell lines

2.2. *"Haploid" and "diploid" RH panels*

Although the principle of producing radiation hybrids was known since the middle of the 1970s, the use of RH panels to construct maps remained limited for about 15 years owing to the paucity of available genes and markers and the fact that PCR was not yet a mature technology. In 1990, Cox and coworkers were the first to demonstrate the feasibility of producing an RH panel and its efficiency to construct an RH map of human chromosome 21 (HSA21). They also presented the first principles of the statistical tools necessary for mapping a marker. For this large-scale RH panel, the donor cells were obtained from a somatic hybrid cell line carrying a "haploid" copy of HSA21. After irradiation with a dose of 8000-Rad X ray and fusion with a rodent receptor cell line, the authors produced 103 haploid RH cell lines, which were found to randomly retain between 30 and 60% of human chromosome 21.

At that point, the generalization of this procedure to produce whole-genome RH panels (WGRHP) was not straightforward since monohybrid somatic cell lines were difficult or nearly impossible to produce for most species. Thus, too many RH lines would have to be produced for the panel to be efficient. A new strategy, consisting of using donor cells derived from diploid cell lines (human fibroblasts), was then developed resulting in "diploid" RH cell lines (Walter *et al*., 1994). It is interesting to note that this methodology is very similar to the initial proposition made by Goss and Harris (1975). This new strategy was then used to generate RH panels for many different species such as mouse (Schmitt *et al*., 1996; McCarthy *et al*., 1997), cattle (Womack *et al*., 1997; Rexroad *et al*., 2000; Williams *et al*., 2002), dog (Priat *et al*., 1998), pig (Yerle *et al*., 1998; Yerle *et al*., 2002), rat (Watanabe *et al*., 1999; McCarthy *et al*., 2000), cat (Murphy *et al*., 1999), horse

Table 1 Overview of some whole-genome radiation hybrid panel (WGRHP) available in different species

Species	Reference	Irradiation dose (Rad)	Number of lines	Mean estimated retention frequency	kb/cR[a]	Resolution kb[b]
Cat	Murphy *et al*. (1999) Murphy *et al*. (2001)	5000	93	0.39	195	538
Cattle	Womack *et al*. (1997) Band *et al*. (2000)	5000	101	0.22	330	1500
Cattle	Rexroad *et al*. (2000)	12 000	88	0.30	n.c.	n.c.
Cattle	Williams *et al*. (2002)	3000	94	0.23	75	347
Chicken	Morisson *et al*. (2002) Pitel *et al*. (2004)	6000	90	0.22	43.7–63	221–318
Dog	Priat *et al*. (1998)	5000	126	0.21	166	627
Horse	Kiguwa *et al*. (2000)	3000	94	0.28	n.c.	n.c.
Horse	Chowdhary *et al*. (2002) Chowdhary *et al*. (2003)	5000	92	0.44	200	494
Macaque	Murphy *et al*. (2001)	5000	93	0.33	330	1500
Man	Gyapay *et al*. (1996)	3000	93	0.32	208	699
Man	Stewart *et al*. (1997)	10 000	83	0.16	29	218
Mouse	Schmitt *et al*. (1996)	3000	164	0.18	3500	11 856
Mouse	McCarthy *et al*. (1997)	3000	94	0.28	98	372
Pig	Yerle *et al*. (1998)	7000	126	0.35	37	84
Pig	Yerle *et al*. (2002)	12 000	90	0.35	14	44
Rat	Watanabe *et al*. (1999)	3000	96	0.27	106	409
Zebra fish	Geisler *et al*. (1999)	3000	94	0.18	61	361

[a]In some cases, we assume 1 cM is equivalent to 1 Mb to estimate the ratio kb/cR.
[b]The resolution was estimated as the average size of retained fragments (=100 times the ratio kb/cR following the definition of the unity cRay) divided by the product of the number of lines and the average retention frequency.

(Kiguwa *et al*., 2000; Chowdhary *et al*., 2002), macaque (Murphy *et al*., 2001), zebra fish (Geisler *et al*., 1999), and chicken (Morisson *et al*., 2002). Table 1 presents an overview of the different WGRHP panels available for the species cited above.

3. RH mapping methodology

3.1. *Definition and control of the parameters of interest*

RH mapping methodology is mostly inspired by the linkage mapping methodologies leading to a similarity in terminology in both cases. The two important parameters to be considered in RH mapping are the probability of breakage between two markers and the retention probability of the donor chromosomal segment (Figure 3).

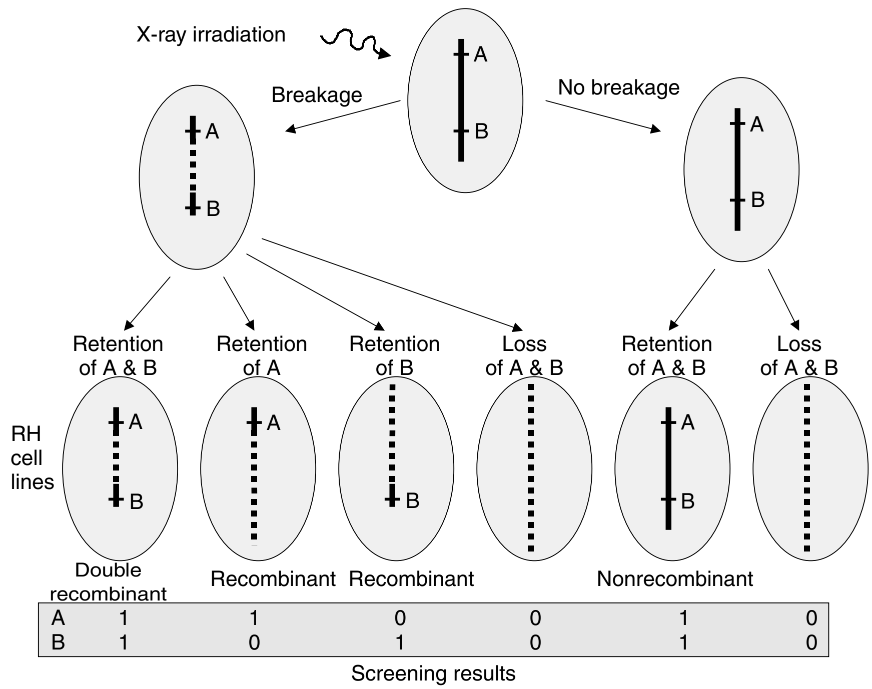


Figure 3 RH mapping principles. Breakage points generated by X-ray irradiation in the donor cells mimic meiotic recombinations, and it becomes possible for a given chromosomal segment (here between markers A and B) to sort out nonrecombinant "parental" haplotypes (no breakage between A and B and the resulting segment is retained or eliminated) and "recombinant" haplotypes (breakage between A and B and only one marker is retained). The analogy with linkage mapping can be further prolonged with the identification of "double recombinant" RH cell lines (breakage between A and B and they are retained or eliminated together). Observable results from RH panel screening for the presence/absence of markers are shown on the bottom of the figure

3.1.1. Breakage probability

The breakage probability is dependent both on the physical distance separating markers and the irradiation dose used to generate the panel: for a given X-ray dose, the closer the two markers are, the lower the probability of breakage between them and the larger the probability of coretention or coelimination. The breakage probability varies from 0 if the markers are at the same position to 1 if markers are segregating independently since their co-retention is only dependent on the retention probability.

Distances in RH maps are measured in $cRay_{xRad}$. 1 $cRay_{xRad}$ corresponds to a breakage probability of 1 % between two markers in a panel constructed with a γ-ray dose of x Rad. In theory, it is possible to control the resolution of the panel by controlling the irradiation dose; however, it should be noted that the reproducibility using the same irradiation dose among laboratories appears to be far from perfect and the DNA fragmentation due to the irradiation dose seems to depend on donor cells. Therefore, the resolution of a 3000-Rad panel may be the same as that of a

5000-Rad panel or even higher (see Table 1). Moreover, the expected resolution has to be adjusted according to the number of markers available or expected.

3.1.2. Retention probability

The different X-ray induced "recombinant" or "nonrecombinant" chromosomal segments from the donor cell can be detected only if they are retained in the RH cell lines considered (see Figure 3). Thus, the probability of retention introduces a second level of control to achieve a good resolution for the RH panel. This parameter can be compared to the number of individuals needed in a linkage mapping experiment to obtain a sufficient number of informative meiosis. Nevertheless, since the mechanisms of elimination of chromosomal segments during RH cell growth remain poorly understood, the retention probability in the RH panel can be controlled during its construction by operating a selection among the RH cell lines available. This should be done carefully to avoid introducing a bias in further analyses (for instance, by screening a small set of independent markers to estimate the overall retention probability of each RH cell line produced).

3.2. Parameters estimation

Statistical methodologies for RH mapping are identical to the ones used classically in linkage mapping and developed in the beginning of the 1990s (Boehnke *et al*., 1991; Cox *et al*., 1990; Lunetta and Boehnke, 1994; Lange *et al*., 1995). A description of these methodologies follows.

Screening the RH panel for a set of N markers permits the identification, for each pair of markers A_i and A_j (i and j varying from 1 to N), of four different types of RH cell lines (respectively $A_i^+A_j^+$, $A_i^+A_j^-$, $A_i^-A_j^+$, and $A_i^-A_j^-$ corresponding to the lines having retained respectively both A_i and A_j, A_i but not A_j, A_j but not A_i, and neither A_i nor A_j; Figure 3). However, it is not possible to estimate the parameters of interest directly from these four different populations of each of the four line types since it is not possible to directly distinguish between breakage and marker retention probabilities. Nevertheless, the expected number for each class can be calculated using as unknown parameters the breakage probability (θ_{ij}) between each couple of markers A_i and A_j, the retention probabilities r_i (respectively r_j) of marker A_i (respectively A_j) and the coretention probabilities r_{ij} ($i \neq j$) of A_i and A_j. These parameters are then estimated by maximizing the likelihood of the observations (see Cox *et al*., 1990 for detailed equations). Finally, assuming randomness of breakage events and no interference between them, breakage occurrence can be modeled as a Poisson process. The distance d_{ij} (in cRay_x, x being the irradiation dose in Rad) between markers A_i and A_j is thus $d_{ij} = -\log(1 - \theta_{ij})$. This function is analogous to the Haldane mapping function used in linkage mapping and computations of likelihood functions require the assumption of absence of multiple recombinations between markers that are responsible for nonadditiveness of two-point distances for physically distant markers. However, in

their pioneering experiment (14 markers covering 20 Mb on HSA21), Cox *et al*. (1990) showed that this effect appeared to be nonsignificant or negligible.

3.3. Linkage group construction (two-point analysis)

The first step in the construction of a map is to identify markers belonging to the same chromosome or to the same genomic region. Thus, two-point analyses consist of identifying, at a given threshold, a group of markers that are genetically linked. As in linkage mapping, for each of the $N(N-1)/2$ pairs of markers A_i and A_j, linkage is evaluated by calculating the Lod score (Lod_{ij}) corresponding to the log ratio of the likelihoods $L^1{}_{ij}$ of the data assuming linkage between A_i and A_j (alternative hypothesis H_1) and $L^0{}_{ij}$ of the data under the null hypothesis (H_0) of no corresponding linkage. Parameter values are those estimated as before for $H_{1,t}$ while a breakage probability of 1 is assumed to calculate H_0.

A linkage group at the S significance threshold will be defined by the L markers A_k (k varying from 1 to L with $L \leq N$) such that at least one marker A_l (l varying from 1 to L) gives $Lod_{kl} \geq S$. The number of linkage groups is thus increasing with S. It should be noted that for a given threshold (for instance $S = 3$), the linkage criterion is generally more stringent than in classical linkage mapping (Cox *et al*., 1990).

3.4. Determination of the order of markers inside a linkage group

As for linkage mapping, two families of methods are used to order markers inside linkage groups: nonparametric and parametric methods (Boehnke, 1992; Boehnke *et al*., 1991; Lange *et al*., 1995).

Nonparametric methods are based on an intuitive parsimony criterion consisting in finding the order among markers, which results in the minimum number of breakages in the RH cell lines of the panel. For instance, let us consider $N = 10$ markers screened on a cell line giving the result vector $\mathbf{h} = [1\ 1\ 1\ 9\ 0\ 9\ 0\ 0\ 1\ 1]$ (1 if the marker is present, 0 if the marker is absent, and 9 if the status of the marker is unknown). To explain **h**, at least two breakage points are necessary (one between marker 3 and 5 and one between 8 and 9, unknown breakpoints are ignored). The minimal number of breakages is called Obligate Chromosomal Break (OCB) and represents a minor of the actual number of breakages since double recombinants are ignored. To evaluate the OCB, one needs to count the number of times 0 (respectively 1) is followed by a 1 (respectively 0) in the result vector. The main advantage of this method is that the model considered is not restrictive (the only constraint is to ignore double recombinants), rather intuitive and not computationally intensive. However it provides no information about distances between markers.

Parametric approaches are generally based on the maximum likelihood principle. Thus, they define a stochastic model to sort the different possible orders according to the likelihood of the data. Several models have been proposed that differ by the number of parameters they impose for the estimation. In general, the hypothesis

aims at restraining the number of retention probabilities among markers. In the fuller model (Cox *et al*., 1990) named "general retention model", if N markers are considered, retention probabilities are estimated for all the $(N(N+1)/2)$ possible chromosomal segments (N possible segment with 1 marker; $N/2$ or $(N-1)/2$ if N is uneven containing $N-1$ markers; ... ; 1 possible segment containing all N markers). This model thus includes $(N^2+3N-2)/2$ parameters ($N-1$ breakage probabilities and $(N(N+1)/2)$ retention probabilities). When N gets bigger, it becomes quite computationally intensive, overparameterized and can only be applied in the case of "haploid" RH cell lines. Therefore, other simpler models have been proposed: the "equal retention probability model", the "centromeric or telomeric retention model", the "left-endpoint model", and the "selected locus model" (Bishop and Crockford, 1992; Chakravarti and Reefer, 1992; Lawrence and Morton, 1992; Boehnke *et al*., 1991; Boehnke, 1992; Lunetta *et al*., 1996). Some of these models are nested and thus can be compared relatively to each other.

Maximization of the likelihood for the different orders needs algorithmic computation. If comparison of the likelihood's among different possible orders permits to sort them, differences between the log-likelihoods of two consecutive orders may, however, be nonsignificant (for instance, at the significance threshold of 3). To circumvent this problem, a "framework map" can be constructed by selecting the K markers among the N ones such that the best order found has a log-likelihood superior to the one of the second order with a difference corresponding to the chosen threshold. Several software propose options to compute framework maps.

In the case of "diploid" hybrid cell lines, it is not possible (in general) to distinguish between lines having one copy of each marker and lines having two copies of the marker (each one from one of the two chromosome homologs of the donor cell). Thus, for parametric models, likelihood computation requires a hidden Markov chain algorithm (Lange *et al*., 1995). Nevertheless, in most cases, analysis of data from a "diploid" RH panel using haploid models does not seem to introduce differences in the best final order found (Ben-Dor *et al*., 2000) except a small underestimation of the distances between markers (about 5%).

Finally, when considering N markers, there are $N!/2$ possible orders to explore. Thus, whichever is the model chosen, evaluating all these orders to find the best one becomes quickly impossible. Some algorithms derived from combinatorial optimization under constraint (here either the number of OCB or the likelihood according to the model) were developed to decrease the number of orders to explore. A first class of methods based on the complete set of markers was proposed such as the "branch and bound", "stepwise locus ordering", or "simulated annealing" method (Nijenhuis and Wilf, 1978; Kirkpatrick *et al*., 1983; Barker *et al*., 1987). Recently, Ben-Dor *et al*. (2000) developed an algorithm based on the "Travelling Salesman Problem" (Garey and Johnson, 1979; Cormen, 1990). All these methods are heuristic and except for the ⟨ branch and bound ⟩ method, which is computationally intensive and practically impossible for a large set of markers ($N>10$), they do not guarantee that the best order will be found. Other heuristic (or metaheuristic) methods were proposed to try to improve a given order and to test if it is not suboptimal such as the "flip algorithm", "Tabu search", or "genetic algorithm" (Glover, 1986; Hansen, 1986; Holland, 1973; Barker *et al*., 1987).

In the end, once the best order is found, it is still possible to check the data by identifying unlikely recombinants, which can reveal a genotyping mistake. This type of a posteriori verification must however be undertaken carefully to avoid bias data.

3.5. Software

Several software packages are publicly available, and differ by the optimization algorithm or the options proposed. The most frequently used are:

- RHMAP (**http://www.spn.umich.edu/group/statgen/software**) contains three different programs, which can be freely downloaded: RH2PT (two-point analysis), RHMINBRK (order determination by a nonparametric approach), and RHMAXLIK (order determination by a parametric approach with almost all the models available).
- RHMAPPER (**http://www.genome.wi.mit.edu/ftp/pub/software/rhmapper**), which can also be freely downloaded. It uses a parametric strategy and a hidden Markov model to perform maximum likelihood calculations on multipoint maps (either on "haploid" or "diploid" panels). It is particularly suitable for large-scale mapping projects.
- RHO (**http://www.cs.technion.ac.il/Labs/cbl/research.html**), which should be used via a web interface. It uses the heuristic described in Ben-Dor *et al*. (2000), which can be applied on either parametric or nonparametric models.
- Carthagene (**www.inra.fr/bia/T/CarthaGene/**), which can be freely downloaded. This software is very user friendly and proposes many different construction procedures. It uses a parametric model with calculation times greatly decreased by an improvement of the EM algorithm (Schiex *et al*., 2002). However, like RHO it assumes that the panel is "haploid", but this does not seem to have a great influence on the reliability of the orders found (see above).

4. Advantages of RH mapping

RH mapping methodologies have met great success in many different species. This can be explained by several advantages as compared with classical linkage mapping strategies.

First, a panel consisting of 100 RH cell lines is in most cases sufficient to offer a good representation of the genome of interest, and its resolution (understood as the threshold at which two closely related markers will be distinguished) is higher than that of linkage mapping analysis. Indeed, in linkage mapping, the resolution is directly dependent on the number of informative meiosis in the pedigree analyzed (assuming an average correspondence of 1 cM and 1 Mb, to distinguish two markers, 1 Mb apart, it is necessary to have in theory 100 informative meiosis). If some markers are not informative in all the families of the pedigree analyzed, the number of individuals needed can therefore exceed greatly the number of informative meiosis wanted. Thus, the size of the experiment very often limits the resolution to the order of the centimorgans. In contrast and as explained above, a control of the irradiation dose and to a lower extent, the number of lines of the

RH panel make it possible to achieve a very fine resolution (up to less than 100 or 50 kb), the limit soon being represented by the number of markers available. Additionally, as far as we know, no hot or cold spot of X-induced breakages have been reported in genome-wide studies. As such, RH distances appear to be closely related with actual physical distances, and RH mapping is thus generally considered as a physical mapping method.

In practice, an RH panel is screened using several types of molecular methods: enzymatic expression analysis, probe hybridization, or more frequently PCR screening, which is easy and fast. One big advantage is that markers do not have to be polymorphic and hence all kinds of STSs (single sequence tags) and, in particular, coding sequences such as ESTs (expressed sequence tags) can be easily mapped on a broad scale. Additionally, detailed maps can be built for the non-pseudoautosomal part of sexual chromosomes such as the Y chromosome in mammals (Liu *et al*., 2002). These characteristics make it possible to use RH panel as an efficient and powerful tool to draw comparative maps through the use of comparative anchoring markers (O'Brien *et al*., 1993; Yang and Womack, 1998; Everts-van der Wind *et al*., 2004). The main principle is to choose markers in coding regions, permitting an easier identification of orthologies among genomes when the whole-genome sequence has not been sequenced. The only limiting factor of RH mapping is that if the species of the donor cells is closely related to the species of the receptor cells, the genome of the receptor cell may interact with the chosen probe. This can be avoided either by defining probes in the untranslated region of a gene, which has a lower level of sequence similarity (Wilcox *et al*., 1991) or in the case of PCR-based probes, by amplifying introns (using exonic probes) to generate an interspecific polymorphism (the intron length being less conserved). It is also possible to use more complex detection methods such as SSCP "single-strand conformation polymorphism", but then it starts to become more time consuming and labor intensive.

Finally, RH mapping constitutes an efficient tool to speed up positional cloning of genes affecting traits of interest, particularly if the sequence of the genome considered is not yet available or if only limited genome coverage is available (up to 2x coverage). RH mapping makes it possible to integrate both linkage and comparative maps thus to exploit efficiently both information sources. Indeed, the linkage mapping–related methodology permits the identification of genomic regions involved in the genetic determinism of the variation of a trait (QTL) using molecular markers and phenotypic information recorded on individuals of a given pedigree. An RH map including both markers used in linkage maps and comparative anchoring markers will permit anchorage of the identified genomic region on the genome of a different reference species and thus to benefit from the functional (identification of putative candidate genes) or, more generally, the mapping information available for this species.

5. Conclusion

Even if the completion of several whole-genome sequences appears to challenge the use of RH panels for positional cloning experiments, it should be noted that

they have played an active and important part in the history and success of these different genome projects, for example in man (Olivier *et al*., 2001) or recently, in rat (Kwitek *et al*., 2004). Indeed, with its high mapping resolution, contigs sequence assembly can be greatly speeded up or some inconsistencies resolved by screening on a given panel some relevant markers (from BAC end sequences or EST for instance).

Moreover, for most species for which no whole-genome sequence is available, the construction and use of an RH panel constitute a powerful tool for positional cloning strategies and more generally for making progress in genomic approaches. Notably, it produces a very fine resolution intermediate between that of linkage maps and BAC-based physical maps. The ease of screening and the fact that markers do not need to be polymorphic make it possible to develop fine whole-genome comparative maps. Thus, the species of interest can benefit from another better-characterized species to quickly build a physical map (Murphy *et al*., 2001). The only remaining limit is the number of available genetic markers, but the resolution of the panel can be adjusted by controlling the irradiation dose.

Further reading

Flaherty L and Herron B (1998) The new kid on the block--a whole genome mouse radiation hybrid panel. *Mammalian Genome*, **9**(6), 417–418.

Hawken RJ, Murtaugh J, Flickinger GH, Yerle M and Robic A (1999) A first-generation porcine whole-genome radiation hybrid map. *Mammalian Genome*, **10**(8), 824–830.

References

Band MR, Larson JH, Rebeiz M, Green CA, Heyen DW, Donovan J, Windish R, Steining C, Mahyuddin P and Womack JE (2000) An ordered comparative map of the cattle and human genomes. *Genome Research*, **10**(9), 1359–1368.

Barker D, Green P, Knowlton R, Schumm J, Lander E, Oliphant A, Willard H, Akots G, Brown V and Gravius T (1987) Genetic linkage map of human chromosome 7 with 63 DNA markers. *Proceedings of the National Academy of Sciences of the United States of America*, **84**(22), 8006–8010.

Barski G, Sorieul S and Cornefert F (1960) Production dans des cultures in vitro de deux souches cellulaires en association, de cellules de caractère ”hybride”. *Comptes Rendus de l'Académie des Sciences (Paris)*, **251**, 18.

Ben-Dor A, Chor B and Pelleg D (2000) RHO–radiation hybrid ordering. *Genome Research*, **10**(3), 365–378.

Bishop DT and Crockford GP (1992) Comparisons of radiation hybrid mapping and linkage mapping. *Cytogenetics and Cell Genetics*, **59**(2–3), 93–95.

Boehnke M (1992) Multipoint analysis for radiation hybrid mapping. *Annals of Medicine*, **24**(5), 383–386.

Boehnke M, Lange K and Cox DR (1991) Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics*, **49**(6), 1174–1188.

Chakravarti A and Reefer JE (1992) A theory for radiation hybrid (Goss-Harris) mapping: application to proximal 21q markers. *Cytogenetics and Cell Genetics*, **59**(2–3), 99–101.

Chowdhary BP, Raudsepp T, Honeycutt D, Owens EK, Piumi F, Guérin G, Matise TC, Kata SR, Womack JE and Skow LC (2002) Construction of a 5000(rad) whole-genome radiation

hybrid panel in the horse and generation of a comprehensive and comparative map for ECA11. *Mammalian Genome*, **13**(2), 89–94.

Chowdhary BP, Raudsepp T, Kata SR, Goh G, Millon LV, Allan V, Piumi F, Guerin G, Swinburne J, Binns M, *et al*. (2003) The first-generation whole-genome radiation hybrid map in the horse identifies conserved segments in human and mouse genomes. *Genome Research*, **13**(4), 742–751.

Cormen TH, Lieserson CE, Rivest RL and Stein C (1990) *Introduction to Algorithms*, MIT Press: Cambridge, pp. 1028.

Cox DR, Burmeister M, Price ER, Kim S and Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, **250**(4978), 245–250.

Everts-van der Wind A, Kata SR, Band MR, Rebeiz M, Larkin DM, Everts RE, Green CA, Liu L, Natarajan S, Goldhammer T, *et al*. (2004) A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research*, **14**(7), 1424–1437.

Garey MR and Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman: New York, pp. 338.

Geisler R, Rauch GJ, Baier H, van Bebber F, Bross L, Dekens MP, Finger K, Fricke C, Gates MA, Geiger H, *et al*. (1999) A radiation hybrid map of the zebrafish genome. *Nature Genetics*, **23**(1), 86–89.

Glover F (1986) Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, **13**, 533–549.

Goss SJ and Harris H (1975) New method for mapping genes in human chromosomes. *Nature*, **255**(5511), 680–684.

Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, Prud'homme JF, Dib C, Auffray C, *et al*. (1996) A radiation hybrid map of the human genome. *Human Molecular Genetics*, **5**(3), 339–346.

Hansen P (1986) The steepest ascent mildest heuristic for combinatorial programming. *Congress on Numerical Methods in Combinatorial Optimization Capri, Italy*.

Holland JH (1973) Genetic algorithms and the optimal allocation of trials. *SIAM Journal of Computing*, **2**(2), 88–105.

Kiguwa SL, Hextall P, Smith AL, Critcher R, Swinburne J, Millon L, Binns M, Goodfellow PN, McCarthy LC, Farr CJ, *et al*. (2000) A horse whole-genome-radiation hybrid panel: chromosome 1 and 10 preliminary maps. *Mammalian Genome*, **11**(9), 803–805.

Kirkpatrick S, Gelatt CD and Vecchi MP (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kwitek AE, Gullings-Handley J, Yu J, Carlos DC, Orlebeke K, Nie J, Eckert J, Lemke A, Andrae JW, Bromberg S, *et al*. (2004) High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence. *Genome Research*, **14**(4), 750–757.

Lange K, Boehnke M, Cox DR and Lunetta KL (1995) Statistical methods for polyploid radiation hybrid mapping. *Genome Research*, **5**(2), 136–150.

Lawrence S and Morton N (1992) Physical mapping by multiple pairwise analysis. *Cytogenetics and Cell Genetics*, **59**(2–3), 107–109.

Liu WS, Mariani P, Beattie CW, Alexander LJ and Ponce De Leon FA (2002) A radiation hybrid map for the bovine Y Chromosome. *Mammalian Genome*, **13**(6), 320–326.

Lunetta KL and Boehnke M (1994) Multipoint radiation hybrid mapping: comparison of methods, sample size requirements, and optimal study characteristics. *Genomics*, **21**(1), 92–103.

Lunetta KL, Boehnke M, Lange K and Cox DR (1996) Selected locus and multiple panel models for radiation hybrid mapping. *American Journal of Human Genetics*, **59**(3), 717–725.

McCarthy LC, Bihoreau MT, Kiguwa SL, Browne J, Watanabe TK, Hishigaki H, Tsuji A, Kiel S, Webber C, Davis ME, *et al*. (2000) A whole-genome radiation hybrid panel and framework map of the rat genome. *Mammalian Genome*, **11**(9), 791–795.

McCarthy LC, Terrett J, Davis ME, Knights CJ, Smith AL, Critcher R, Schmitt K, Hudson J, Spurr NK and Goodfellow PN (1997) A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Research*, **7**(12), 1153–1161.

Morisson M, Lemiere A, Bosc S, Galan M, Plisson-Petit F, Pinton P, Delcros C, Feve K, Pitel F, Fillon V, *et al*. (2002) ChickRH6: a chicken whole-genome radiation hybrid panel. *Genetics, Selection, Evolution*, **34**(4), 521–533.

Murphy WJ, Menotti-Raymond M, Lyons LA, Thompson MA and O'Brien SJ (1999) Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics*, **57**(1), 1–8.

Murphy WJ, Page JE, Smith C, Desrosiers RC and O'Brien SJ (2001) A radiation hybrid mapping panel for the rhesus macaque. *The Journal of Heredity*, **92**(6), 516–519.

Nijenhuis A and Wilf HS (1978) *Combinatorial Algorithms for Computers and Calculators*, Academic Press: New York, pp. 308.

O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA and Copeland NG (1993) Anchored reference loci for comparative genome mapping in mammals. *Nature Genetics*, **3**(2), 103–112.

Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, Brady SD, Bushard JM, Bustos VI, Chu A, *et al*. (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science*, **291**(5507), 1298–1302.

Pitel F, Abasht B, Morisson M, Crooijmans RP, Vignoles F, Leroux S, Feve K, Bardes S, Milan D, Lagarrigue S, *et al*. (2004) A high-resolution radiation hybrid map of chicken chromosome 5 and comparison with human chromosomes. *BMC Genomics*, **5**(1), 66.

Pontecorvo G (1975) Production of mammalian somatic cell hybrids by means of polyethylene glyco (PEG) treatment. *Somatic Cell Genetics*, **1**(4), 397–400.

Priat C, Hitte C, Vignaux F, Renier C, Jiang Z, Jouquand S, Cheron A, Andre C and Galibert F (1998) A whole-genome radiation hybrid map of the dog genome. *Genomics*, **54**(3), 361–378.

Rexroad CE, Owens EK, Johnson JS and Womack JE (2000) A 12,000 rad whole genome radiation hybrid panel for high resolution mapping in cattle: characterization of the centromeric end of chromosome 1. *Animal Genetics*, **31**(4), 262–265.

Schiex T, Chabrier P, Bouchez M and Milan D (2002) *Boosting EM for radiation hybrid and genetic mapping*. Lecture Notes in Computer Science, 2149.

Schmitt K, Foster JW, Feakes RW, Knights C, Davis ME, Spillett DJ and Goodfellow PN (1996) Construction of a mouse whole-genome radiation hybrid panel and application to MMU11. *Genomics*, **34**(2), 193–197.

Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, Hadley D, Harris M, Hussain S, *et al*. (1997) An STS-based radiation hybrid map of the human genome. *Genome Research*, **7**(5), 422–433.

Walter MA, Spillett DJ, Thomas P, Weissenbach J and Goodfellow PN (1994) A method for constructing radiation hybrid maps of whole genomes. *Nature Genetics*, **7**(1), 22–28.

Watanabe TK, Bihoreau MT, McCarthy LC, Kiguwa SL, Hishigaki H, Tsuji A, Browne J, Yamasaki Y, Mizoguchi-Miyakita A, Oga K, *et al*. (1999) A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genetics*, **22**(1), 27–36.

Wilcox AS, Khan AS, Hopkins JA and Sikela JM (1991) Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Research*, **19**(8), 1837–1843.

Williams JL, Eggen A, Ferretti L, Farr CJ, Gautier M, Amati G, Ball G, Caramori T, Critcher R, Costa S, *et al*. (2002) A bovine whole-genome radiation hybrid panel and outline map. *Mammalian Genome*, **13**(8), 469–474.

Womack JE, Johnson JS, Owens EK, Rexroad CE, Schlapfer J and Yang JP (1997) A whole-genome radiation hybrid panel for bovine gene mapping. *Mammalian Genome*, **8**(11), 854–856.

Yang YP and Womack JE (1998) Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. *Genome Research*, **8**(7), 731–736.

Yerganian G and Nell MB (1966) Hybridization of dwarf hamster cells by UV-inactivated Sendai virus. *Proceedings of the National Academy of Sciences of the United States of America*, **55**(5), 1066–1073.

Yerle M, Pinton P, Delcros C, Arnal N, Milan D and Robic A (2002) Generation and characterization of a 12,000-rad radiation hybrid panel for fine mapping in pig. *Cytogenetic and Genome Research*, **97**(3–4), 219–228.

Yerle M, Pinton P, Robic A, Alfonso A, Palvadeau Y, Delcros C, Hawken R, Alexander L, Beattie CW, Schook LB, *et al*. (1998) Construction of a whole-genome radiation hybrid panel for high-resolution gene mapping in pigs. *Cytogenetics and Cell Genetics*, **82**(3–4), 182–188.

Specialist Review

Linkage mapping

Mark E. Samuels
Dalhousie University, Halifax, NS, Canada

Marie-Pierre Dubé
Institut de Cardiologie de Montréal, Montreal, QC, Canada

1. Introduction and scope

The purpose of this chapter is to provide a practical guide to linkage mapping for the identification of genes predisposing to human disease (or other interesting phenotypes). The emphasis will be on technical issues and pedigree-based analysis. More theoretical concerns, particularly those relating to methods in statistical genetics, will be covered in depth elsewhere (*see* Article 48, **Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping**, Volume 1, Article 52, **Algorithmic improvements in gene mapping**, Volume 1, Article 58, **Concept of complex trait genetics**, Volume 2, and Article 11, **Mapping complex disease phenotypes**, Volume 3). Alternative approaches such as linkage disequilibrium (LD) and SNP-based association mapping are covered in other chapters (*see* Article 12, **Haplotype mapping**, Volume 3, Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3, Article 69, **Reliability and utility of single nucleotide polymorphisms for genetic association studies**, Volume 4, Article 73, **Creating LD maps of the genome**, Volume 4, and Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4).

2. General approaches to linkage mapping

Linkage as a formal term refers to the mapping of a predisposing polymorphism or mutation at a genetic locus through the analysis of chromosomal segments transmitted to individuals with some known degree of relationship (Ott, 1991; Terwilliger and Ott, 1994) (see Figure 1). The enabling principle for linkage mapping in humans is the use of anonymous polymorphic DNA variants, called *markers*, as tags for these chromosomal segments (Botstein *et al*., 1980). Using these tags, one can detect the correlated inheritance of a particular trait with that of closely linked marker loci.

The genetic mapping described here results preferentially from the analysis of pedigrees showing Mendelian segregation of the trait of interest (*see* Article 51,

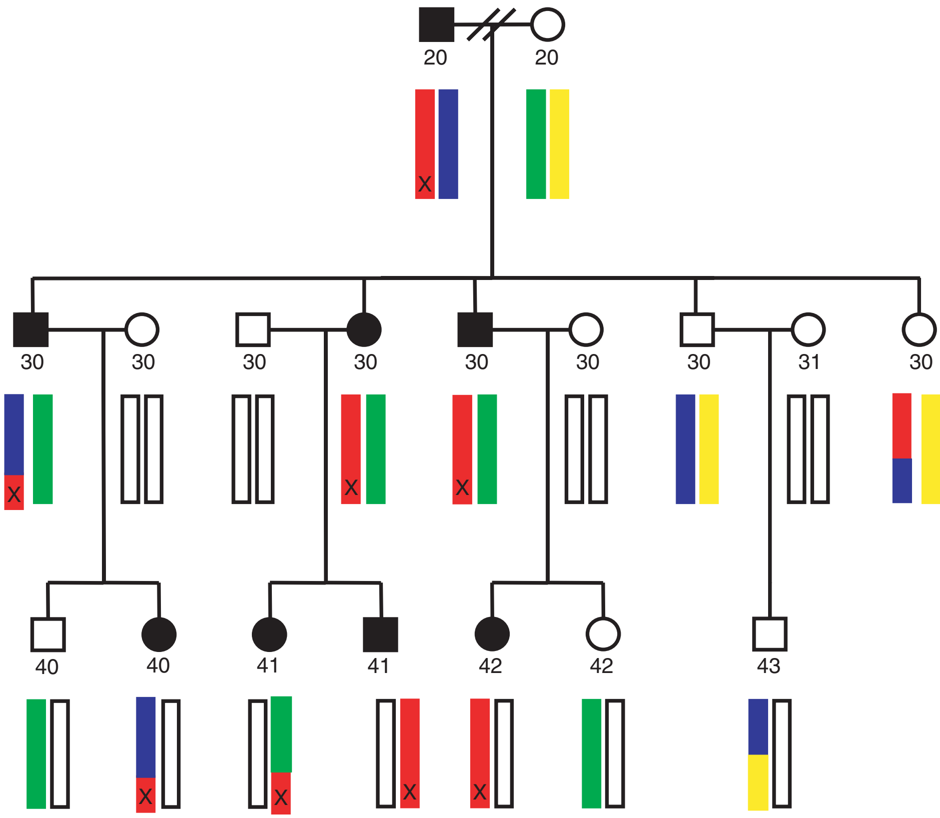


Figure 1 Schematic visualization of a pedigree segregating a biological phenotype of interest. According to convention, square symbols are male, circles are female. Affected individuals are shaded in black. The four hypothetical founding haplotypes for a given chromosome are indicated in red, blue, green, and yellow. Additional copies of this chromosome, introduced by spouses marrying into the pedigree, are displayed as unshaded bars. In this example, it is presumed that the affected founder is known for this pedigree (male in top generation). A causal mutation at a specific locus is presupposed by the X on the red haplotype. Recombination events reduce the extent of the red haplotype transmitted through the pedigree. In this idealized example, there is perfect cosegregation of the mutation (X) and a surrounding segment of red haplotype, in all affected individuals

Choices in gene mapping: populations and family structures, Volume 1 and Article 77, **Mechanisms of inheritance**, Volume 2). By this we mean: phenotypes are relatively straightforward to characterize; transmission in families is generally unilineal and unambiguous (although this can be confounded in some populations that exhibit high degrees of consanguinity); and underlying sequence variants usually confer obvious and severely deleterious effects on gene function (or in rarer cases obvious and severe gain of gene function). For the detection of linkage, large families have more statistical power than small ones, but these are not always available, especially for traits with low penetrance, delayed age of onset, or for traits of complex etiology (Haines and Pericak-Vance, 1998; *see also* Article 58, **Concept of complex trait genetics**, Volume 2 and Article 11, **Mapping complex disease phenotypes**, Volume 3).

3. Properties of genetic markers

Over the years, a variety of different genetic markers have been used for mapping purposes. For the past decade, the markers of choice for linkage have been microsatellite repeats, also known as VNTRs (variable number of tandem repeats) or STRs (short tandem repeats) (see Figure 2) (Litt and Luty, 1989; Taylor *et al*., 1989; Beckmann and Soller, 1990; Weber, 1990). They consist of stretches of repeating units such as CACACA or GATAGATAGATA, embedded within unique sequences in various chromosomal locations. For the most part, they lie outside the coding exons of genes, since varying repeat lengths other than triplets would otherwise lead to frameshift mutations. A particular repeat marker may be unambiguously detected using appropriately designed PCR primers in the surrounding unique sequence. These repeats frequently vary in length in different individuals, presumably because of occasional mistakes made by the replication machinery. Such events are relatively infrequent, so that these repeats are stable enough to be used in studies spanning multiple generations in a pedigree. They are not wholly stable however, and the identity of microsatellite alleles by state is usually not sufficient to infer identity by descent in individuals of distant

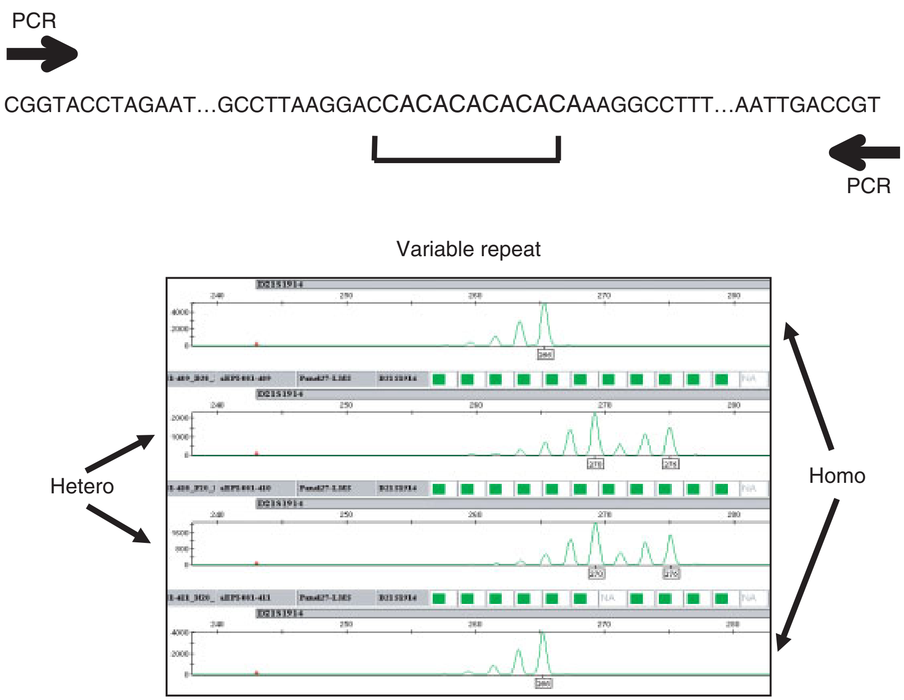


Figure 2 Typical dinucleotide microsatellite repeat marker. A (CA)n repeat is embedded within a unique sequence context, which provides for the development of marker-specific PCR amplification primers. Below is a chromatogram for a real dinucleotide repeat, D21 S1914, located on chromosome 21. Genotypes are shown for four individuals, first and fourth are homozygotes, second and third are heterozygotes. Relative allele sizes (i.e., names) are indicated in small black boxes below the highest molecular weight peak for each allele. (Courtesy of J. Thompson.)

or unknown genealogical relationship (with exceptions in founder populations). In order to be useful in linkage analysis, a marker must have multiple alleles present in a population. The more alleles the better; however, owing to the limitations of analytical approaches, it is usually best to use markers with no more than 12–15 alleles. The information content of markers is commonly measured using the heterozygosity value and the polymorphism information content (PIC) value. Dinucleotide microsatellites are typically more informative than tri- or tetranucleotide repeats, with heterozygosities as high as 0.7–0.8.

It is estimated that the human genome has 5000–10 000 such microsatellite repeats. For these markers to be used in a genetic mapping study, their relative order and distance along each chromosome must be known. Several genetic maps have been generated, through marker genotyping in large families with multiple meioses, and using recombination data to orient and locate markers with respect to each other (Weissenbach *et al*., 1992; Gyapay *et al*., 1994; Broman *et al*., 1998; Yu *et al*., 2001; DeWan *et al*., 2002; Kong *et al*., 2002). Now almost all markers can be placed unambiguously on the assembled human genome sequence.

The human nuclear genome consists of approximately 3600 centimorgans (cM) in genetic distance (averaged over both sexes) (Kong *et al*., 2002). Thus, to cover the genome at 10-cM resolution requires 360 genetic markers, assuming each is fully informative; 5-cM resolution requires 720 markers. Good microsatellites of high information content come close to these limits, hence genome-wide mapping panels have on the order of 400 (10 cM) to 800 (5 cM) markers respectively. Such sets are commercially available (Reed *et al*., 1994; Lindqvist *et al*., 1996).

The general approach to a linkage mapping experiment is to perform a whole-genome scan at approximately 5- or 10-cM density on a set of samples from one or more families transmitting the phenotype of interest. Following statistical analysis, potential regions of linkage are identified on various chromosomes. Each of these is subjected to genotyping of additional microsatellite markers at increased density, followed by reanalysis. Ideally, only a single region survives the second round of mapping. A third round of genotyping with an increased density of markers, potentially exhausting all microsatellite repeats in a region, may follow. Owing to the relatively high cost of complete genome scans, often only subsets of sampled individuals are genotyped in the initial phase, focusing on those carrying the most definitive phenotypic state.

In some situations, linkage analysis may begin with or be restricted to specific genes with a higher biological probability of involvement in the phenotype, often termed candidate genes. The general principles of mapping are the same, but practically this reduces the scope of genotyping to smaller sets of markers near these genes, with potentially significant cost savings. Often, this approach is used to exclude genes already known to mutate to the phenotype of interest, in newly ascertained families.

For fine-mapping of recombination events in specific meioses, commercially available genome scan panels have insufficient resolution. Therefore, laboratories involved in linkage mapping must develop additional custom markers. Public databases include many thousands of potentially available microsatellite markers that can be used for such fine-mapping. These are typically identified with a “D” number such as D1 S2134, indicating the chromosome such as chr1 plus the specific

marker number – however, some nonpolymorphic PCR amplification products or sequence tagged sites (STSs) were historically assigned D numbers; thus, such numbers are not automatically indicative of microsatellite repeat status. Moreover, some useful markers have never been assigned D numbers but retain only the numbers from the projects in which they were developed, such as Utah markers (UT numbers) (Utah Marker Development Group, 1995), Marshfield markers (Mfld numbers) (Broman *et al*., 1998; Weber, 1990; Weber and Broman, 2001; DeWan *et al*., 2002), and Genethon markers (AFM numbers) (Weissenbach *et al*., 1992; Gyapay *et al*., 1994; Reed *et al*., 1994). Current databases attempt to unify all known markers and aliases for each marker, but it is not necessarily true that two differently named markers are truly different.

Once all identified markers in a region have been exhausted, the genome sequence in chromosomal regions of interest can be directly examined for additional microsatellite motifs using standard bioinformatics tools. The total potential resolution of microsatellite markers is on the order of 0.2 to 1 cM, which is typically sufficient for positional cloning purposes. Recall that the smallest definable interval containing a putative causal variant is a function of recombination events that have occurred in meioses between the patients who have been sampled. The actual size of the recombinant interval is not dependent on the density of markers being used to analyze it. Only the resolution with which the exact site of recombination is mapped is affected by the density of markers employed.

Recently, commercial mapping panels of single-nucleotide polymorphisms (SNPs) have begun to come into use for linkage mapping (Tsai *et al*., 2003; Sellick *et al*., 2004). A single microsatellite marker is often considered to have equivalent information content to 3–4 SNPs. SNP technologies will not be reviewed here (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1), however, whole-genome mapping sets are now commercially marketed. These sets initially contained in the range of 4–10 000 SNPs, roughly equivalent to a 5-cM microsatellite genome scan, and are designed for family-based linkage analysis. More recently, SNP panels of 100 000 markers have been developed. As SNPs are generally biallelic, it is intrinsically more straightforward to generate allele calls, so manual review may be unnecessary at least for standardized marker sets. SNPs are believed to be stable over long periods of time. A given SNP is usually presumed to have arisen once only during evolution (although some nucleotide positions may turn out to be unstable and mutate repeatedly). Thus, identity of state for an SNP site in two individuals is considered to be indicative of identity by descent. For fine-mapping equivalent to microsatellites at about 1-cM resolution, hundreds of thousands to millions of potential SNPs are available in public databases, however, the informativeness of these must be evaluated for specific patient samples in a family-mapping study (Sachidanandam *et al*., 2001; Holden, 2002). Even if SNP sets become the standard tool for low-resolution genome scanning, microsatellites will probably continue to play a useful role in fine-mapping for this reason.

4. Microsatellite genotyping

Microsatellites are typically assayed following PCR. One of the PCR primers is usually tagged with a fluorescent dye, and the products of PCR are resolved

electrophoretically either on polyacrylamide gels or by capillary electrophoresis (Ziegle *et al*., 1992; Gelfi *et al*., 1994; Reed *et al*., 1994; Gyapay *et al*., 1996; Lindqvist *et al*., 1996; Mansfield *et al*., 1996; Ghosh *et al*., 1997; Mansfield *et al*., 1997; Vainer *et al*., 1997; Wenz *et al*., 1998; Delmotte *et al*., 2001; Wenz *et al*., 2001). Although the concept is straightforward, there are potential technical pitfalls.

PCR primers must amplify the marker in question with high specificity. Ideally, both primers should lie in unique sequence; however, in practice, this is sometimes difficult to achieve as microsatellites often lie in or near repetitive elements. Hence, PCR conditions may require optimization to generate sufficiently specific products. For genome scan mapping panels, standardized conditions have been developed and are available, although laboratories should be prepared to reoptimize if needed. In developing custom microsatellite markers for fine-mapping, laboratories must usually develop their own PCR conditions. One may move PCR primers in addition to altering reaction conditions, as long as the primers remain specific to the repeat unit under development. Indeed, the exact primer sequences of commercial marker kits may be proprietary and different from public database primers for those markers.

When DNA is extracted from patient blood samples, both maternal and paternal chromosomes are recovered, hence both alleles of a microsatellite marker are observed. On occasion, genotyping may employ DNA from single sperm cells, or from cell lines reduced to haploidy through cell fusion and chromosome loss. But for linkage analyses, blood samples are the usual source. Inactivation of X chromosomes in females presents no problem.

A typical dinucleotide chromatogram is shown in Figure 2, with examples of homozygous and heterozygous individuals. Although unique products have been amplified, note that there are multiple peaks even in the homozygotes. Extra so-called stutter peaks are observed, smaller than the full-length product, and presumed to result from enzymatic skipping during PCR. The spacing of stutter peaks is equivalent to the type of repeat unit. Stutter peaks do not generally impede genotyping. The size of a microsatellite allele is usually defined by the position of the largest molecular weight peak.

The enzymes typically used in PCR have a tendency to add additional, nontemplated nucleotides at the 3′ ends of products, to a variable extent. Thus, microsatellite chromatograms historically have suffered from the problem of "peak-splitting". This is separate from and in addition to the observation of stutter peaks. If this problem is severe enough, particular markers may be wholly useless. Specific added sequence elements can reduce the intrinsic variability of nontemplated addition. These sequence elements are added to the 5′ end of the nonlabeled PCR primer, so that variability in nontemplated addition is reduced at the 3′ end of the labeled strand, which is the strand visualized by the instrumentation (see Figure 2) (Brownstein *et al*., 1996; Magnuson *et al*., 1996).

Despite optimization, some markers do routinely give extra peaks, presumably because of additional priming sites in the genome. Such markers may still be useful if these peaks are sufficiently reproducible. However, automated genotyping programs may require additional training to deal with them. In some cases, extra peaks fall into the expected allele range for other markers multiplexed with the marker in question.

Microsatellite genotype calls, usually given in base pairs, are not exact but are relative to internal size standards, and as such are only indirect readouts of the actual number of repeats in a given allele in a given sample. These size standards may be purchased from commercial suppliers or synthesized in the laboratory (Brondani and Grattapaglia, 2001). Unfortunately, the interpolation of allele sizes is dependent on the specifics of the electrophoretic system used. Thus, genotypes are difficult to compare between different instrument platforms, and often between different laboratories' versions of the same marker. One solution to this is to normalize all allele calls for a given marker to a standard DNA sample, such as a CEPH control DNA. This technique allows data to be pooled across multiple platforms, although standardized calls may need to be created independently for each different instrument.

To increase efficiency, multiplexing is typically performed. Unfortunately, microsatellites have proved recalcitrant to pre-PCR multiplexing. Therefore, multiplexing of microsatellites is usually performed after PCR and prior to electrophoresis.

Since multiplexed markers are subjected to electrophoresis in the same lane or capillary, it is critical to associate specific chromatogram peaks with the correct marker. Allelic size ranges are determined for a specific PCR primer pair used to amplify a marker, either based on public information or else by test genotyping a set of random control DNAs. Thus, peaks for a specific marker have an expected size range where genotypes are called. However, novel alleles are often observed when large numbers of experimental samples are subsequently genotyped for that marker. Some of these alleles may fall outside the expected range for that marker. In this case, trained software must be updated to incorporate the new alleles, which can be problematic if there is overlap with other markers that were previously multiplexed in the same lane. This problem is often identified through the failure of a marker neighboring the actual marker with the allelic expansion. In the worst case, individual markers may have to be removed from a multiplexed panel and electrophoretically analyzed separately. To minimize potential for subsequent allelic overlap, when a new panel of multiplexed markers is developed, a gap should be provided between the known ranges of size-adjacent markers.

Multiplexing also relies on the availability of multiple fluorescent dye tags with different emission spectra. Markers with overlapping size range but different dye tags may thus be pooled. Commercial systems typically permit four different dyes to be multiplexed, one of which is used for the internal size standard. The various fluorescent dyes alter the mobility of DNA fragments, so that the apparent electrophoretic mobility of a given marker will change if a different dye is substituted. A similar and even more severe problem may arise if the dye tag or spacer structure is altered on the internal size standard, in which case ALL marker mobilities may have to be redefined.

Commercial mapping panels have been optimized for marker dye color and spacing so that as many as 15–20 different microsatellites may be assayed in the same lane, significantly reducing cost and enhancing throughput. With effort, custom panels can also be highly multiplexed, although this may not necessarily be cost-effective.

Following electrophoresis and data collection, actual genotype calls must be made. This can be performed either fully manually or semiautomatically. Commercially available software packages exist for automated genotype calling (Applied Biosystems GeneMapper, SoftGenetics GeneMarker), but while effective, these packages require caution in actual use. In practice, some amount of manual review is always necessary, particularly for markers with complex chromatograms or extra peaks. Nonetheless, current software genotyping programs can be very efficient in reducing the required amount of manual trace review for well-behaved markers or markers with which laboratories have extensive experience.

It is recommended that if primers are redesigned, version numbers be used explicitly. It can be highly confusing if multiple versions of a marker, with slightly different primer sequences and or dye types, have the same name in a laboratory system. Version numbers may need to be removed prior to statistical analysis, since public database and map names will now be inconsistent with the internal identifiers.

For the research laboratory not equipped for whole-genome microsatellite mapping, there are several outsourcing alternatives. However, fine-mapping of potential linkages with custom markers is almost always the next step. Outsourcing such custom genotyping is more problematic, and laboratories with serious interest in linkage mapping are encouraged to develop at least some capabilities for internal genotyping.

For microsatellite PCR, 5–20 ng of high-quality genomic DNA are required for each marker. Thus, whole-genome scans at 5-cM resolution with follow-up demand 5–20 μg of DNA per patient. These quantities can routinely be achieved using fresh blood samples in the tens of milliliters, or equivalent frozen white cells (buffy coats), or from cell culture of immortalized fibroblast lines. In cases in which a blood draw is not possible, buccal (cheek) swabs may sometimes be obtained, yielding sufficient DNA for small numbers of reactions only. Recently, several protocols have been developed for whole-genome amplification, particularly suited for whole-genome SNP analysis since so many more markers are required. The utility of these protocols for microsatellite genotyping is not fully validated.

High-volume genotype data must be appropriately archived and made available to statistical geneticists. Integrating clinical, pedigree, and genotype data sets can be surprisingly challenging. Moreover, statistical analysis programs generally require very specific formatting of data. Unfortunately, there are few appropriate commercial database prototypes serving the needs of human geneticists, although ProgenyLab is a relatively recent entry in this area. Laboratories expecting to perform large amounts of linkage mapping are highly encouraged to develop the integrated database systems.

5. Statistical genetic analysis of linkage

The essence of linkage analysis is to detect the cosegregation of a particular chromosomal segment (defined through marker genotyping) with the phenotypic state of interest, in a set of related patients such as a single family (Ott, 1991; Terwilliger and Ott, 1994). The question is whether any particular chromosome

segment in the genome cosegregates with the phenotype more frequently than one would expect by chance alone. To determine this probability usually requires elaborate mathematical analysis. The required statistical methodologies are discussed in detail elsewhere (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1, Article 52, **Algorithmic improvements in gene mapping**, Volume 1, and Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3). Here, we give only the briefest overview of statistical genetics, to place the discussion of genotyping into a procedural context.

Statistical genetic tests are traditionally broadly subdivided into two main categories, those in which explicit modeling assumptions are made concerning the behavior of a presumptive causal allele and those in which no such assumptions are made. These are termed parametric (or model-based) and nonparametric (or model-free) analysis respectively. The terms "model-based" and "model-free" are preferred, however, as most methods labeled nonparametric do nonetheless rely on some genetic assumptions. In model-based analysis, assumptions are made that the disease gene–population frequency and the penetrance of the disease alleles in homozygotes and heterozygotes can be accurately estimated. When the mode of action of the disease gene cannot be predicted with confidence, such as is the case for complex diseases, model-free analyses are typically used. Generally, these simply test for excess sharing or preferential transmission of particular marker alleles in family units.

The most commonly used statistic for model-based linkage analysis is the maximum likelihood ratio. This tests the hypothesis of disease and marker cosegregation versus the null hypothesis of random segregation. For historical reasons of convenience, the base 10 logarithm of the ratio of the likelihoods is used and referred to as the LOD score (log of odds) (Morton, 1955). The conventional significance threshold used in linkage analysis is $\text{LOD} \geq 3$ for Mendelian diseases. The genome-wide significance threshold is generally set slightly higher to $\text{LOD} = 3.3$ for complex trait analysis. It is also possible to determine the significance of a test applied to a particular data set empirically using computer simulations. To this end, replicates of the family collection are generated by computer, with random genotypes based on correct inheritance, allele frequencies, and marker recombination fractions. The linkage testing procedure is conducted in each simulated dataset and the maximum LOD score or p-value is noted. The genome-wide threshold of significance is taken as a score that is exceeded in fewer than 5% of replicates.

Statistical linkage analysis can be performed using either a single genetic marker at a time (two-point linkage, that is, disease locus and marker locus), or alternatively using multiple genetic markers simultaneously (multipoint linkage). The advantage of using multiple markers is that the phase of markers can be estimated with more precision, and this can add considerable power to the test. Multipoint linkage calculations, however, require significant computational resources when sufficiently large pedigrees or numbers of markers are analyzed. Exact solutions of multipoint linkage are incomputable for very large pedigrees or marker sets with the commonly employed tools such as LINKAGE, FASTLINK, ALLEGRO, GENEHUNTER, MERLIN, and VITESSE (Lathrop *et al*., 1984; Cottingham *et al*., 1993; Schaffer *et al*., 1994; O'Connell and Weeks, 1995; Kruglyak *et al*., 1996;

Gudbjartsson *et al*., 2000; Markianos *et al*., 2001; Sobel *et al*., 2001; Abecasis *et al*., 2002). Researchers usually accept the limitations on pedigree size or number of markers, which can be examined simultaneously. There are programs such as SIMWALK, LOKI, and MCSIM that estimate inheritance vectors using approximation methods (Weeks *et al*., 1995; Heath, 1997; Thomas *et al*., 2000). Such programs have been shown to give accurate LOD scores in the majority of cases, and provide a valid alternative when exact computations are impossible (*see* Article 52, **Algorithmic improvements in gene mapping**, Volume 1).

Linkage results may be presented numerically in tabular format, but for multipoint analysis, it is common to report results graphically, with scores for parametric or nonparametric linkage plotted as a function of position on a chromosome. In this way, recombination events that unlink chromosomal segments from the phenotype appear as drops in the linkage statistic (Figure 3).

Although direct visualization of allelic phases, haplotypes, and recombination events is not theoretically necessary for locus mapping, in practice, it is widely used for manual review of linkage analysis results (Figure 4). The definition of the haplotypes in a pedigree is achieved by the phasing of alleles at each genotyped marker. By phase, we mean the two alleles of each marker must be assigned as having been transmitted from the paternal or maternal parent. For fully informative markers the process is simple, but for real markers in incompletely sampled pedigrees, phasing of alleles requires mathematical techniques. As with multipoint LOD score calculation, phase determination for large pedigrees and marker sets is computationally restrictive. An added difficulty in visualizing multimarker haplotypes is incorporating them into pedigree drawings. Pedigree

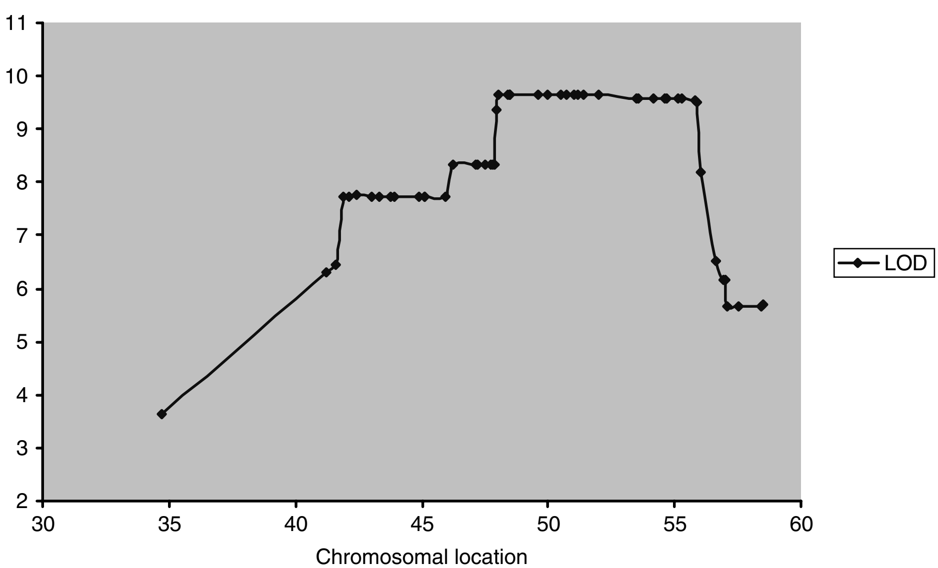


Figure 3 Multipoint linkage. mpLOD score as calculated by the MCSIM algorithm is plotted versus chromosomal location for a dense set of fine-mapping markers near the FH3 hypercholesterolemia locus (Reproduced from Timms *et al*. (2004) by permission of Springer-Verlag)

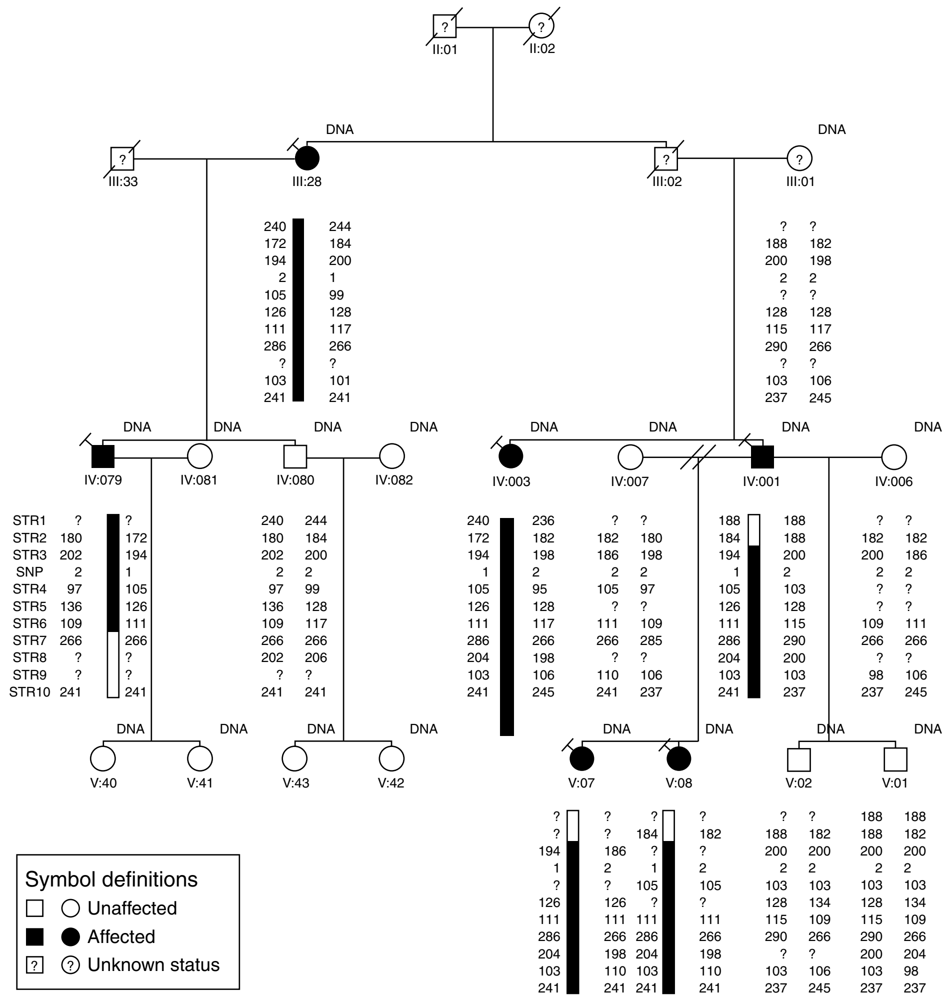


Figure 4 Haplotype visualization. A typical pedigree is shown with marker alleles phased using Genehunter. Each individual is uniquely identified by generation (in Roman numerals) and place (in Arabic numerals). Filled symbols designate affected individuals, open symbols are unaffected individuals, question marks inside symbols indicate individuals whose diagnosis is either unknown or ambiguous. Individuals with DNA samples collected are indicated. Genetic markers (anonymized) are listed in chromosomal order on the left. Beneath each genotyped individual, allele sizes are given for each marker; question marks here indicate a failure to call an allele for that marker/individual combination. Alleles have been phased so that chromosomal haplotypes may be visualized directly, although in this example only one haplotype is explicitly identified by a black bar. Recombination events may be observed in individuals IV:079 and IV:001

drawing packages such as Cyrillic, Progeny, or PedDraw are all usable, though all have limitations in dealing with large haplotypes or complex pedigrees. Haplotype reconstruction analyses can also be performed using approximation algorithms as implemented in SIMWALK or MCSIM (which must be interpreted cautiously), or else smaller pedigrees or subsections of pedigrees can be haplotyped exactly using GENEHUNTER or MERLIN and manually assembled.

As individual SNPs are insufficiently informative for most types of linkage analysis, replacement of microsatellites with SNP-based linkage mapping sets will demand multipoint linkage and haplotyping analyses. Current software packages such as Genehunter and Vitesse have not yet been extensively tested in such scenarios, however, it is anticipated that they will be applicable. New tools for SNP analysis include SNPLink, ALOHOMORA and HaploPainter (Ruschendorf and Nurnberg, 2005; Thiele and Nurnberg, 2005; Webb *et al*., 2005).

All linkage algorithms require correct Mendelian inheritance patterns of the individual marker alleles. This can be tested prior to linkage analysis using the PedCheck program, which verifies the structural integrity of the pedigrees and the Mendelian inheritance of alleles irrespective of phenotypic status (O'Connell and Weeks, 1998). When errors are detected, they may sometimes be corrected by review of the raw genotype data. However, some inheritance errors cannot be explained by any obvious technical mistakes. In such cases, it is advised to eliminate the allele calls of all pedigree members involved in the nuclear families generating the error, or in more severe cases eliminating a marker completely from analysis. One source of such inheritance errors is spontaneous mutation of a microsatellite allele to a different repeat length. Much theoretical attention has been given to the topic of unidentified genotype errors in data sets.

In monogenic as well as complex disorders, mutations in different genes can result in a similar phenotype, so that groups of families displaying a shared phenotype may not segregate a causal variant in the same gene (genetic or locus heterogeneity). Care must be taken therefore when pooling pedigrees for linkage analysis. It may be possible, in some instances, to subgroup pedigrees according to subtle phenotypic differences. Alternatively, robust statistical analyses that allow for locus heterogeneity in the calculation of heterogeneity LOD scores can be used. Those methods will improve the power of linkage detection. Alternatively, different families may segregate different mutations in the same gene for a given phenotype (allelic heterogeneity). In this case, different families will generate linkage to the same chromosomal interval although not sharing the same marker haplotype. In special cases, such as French Canada, Newfoundland, Finland, and so on, identical chromosomal segments or haplotypes can often be detected in different family units whose genealogical relationship may not be known (de la Chapelle and Wright, 1998; Laan and Paabo, 1998; Arcos-Burgos and Muenke, 2002). Such populations are frequently referred to as founder populations or population isolates.

One special subset of linkage analysis is homozygosity mapping. The general principles of recessive trait mapping are the same as for dominant traits. Accurate statistical analysis requires estimates of mutation and marker allele frequencies, which are not easily obtained in advance. However, in special cases, one can assume that affected individuals have received two copies of the same mutant allele. Such examples are common for specific populations with known high rates of consanguinity caused by either geographical or cultural factors, including founder populations (Lander and Botstein, 1987; Sheffield *et al*., 1998). Homozygous haplotype mapping can be applied by manual inspection if all affected patients from a population are homozygous for the same marker alleles for several successive markers in a genome scan. Failure to detect perfect shared marker homozygosity does not rule out the hypothesis, since recombination events may have unlinked

marker alleles from the disease allele in some affected individuals in a data set. Moreover, even in relatively isolated populations, it is possible for multiple mutations in a gene to be segregating, leading to haplotype heterogeneity and failure of homozygosity mapping. Nonetheless, this can be an extremely powerful approach.

6. Positional cloning

The ultimate purpose of linkage mapping is to define recombinant intervals sufficiently small to support direct molecular screening of DNA sequences for causal variants. It is beyond the scope of this chapter to discuss this process, positional cloning, in full detail. However, there is lively interest in significantly reducing the cost of DNA sequencing, and it is not impossible that something like a large scale gene-screening approach could become cost-effective within the next few decades, which in theory could obviate the need for mapping (*see* Article 7, **Single molecule array-based sequencing**, Volume 3).

One can approximate the LOD score equivalent to a given interval size with the formula: size in cM = 60/LOD, such that a LOD score of 3.0 is on average equivalent to an interval of 20 cM (Sham, 1998). Even in the case of a monogenic disorder, with a well-defined chromosomal locus and constrained recombinant boundaries, it has historically been a significant challenge to identify causal variants. With the advent of the human-genome project, in the best case laboratories can simply resequence annotated genes within an interval (*see* Article 23, **The technology tour de force of the Human Genome Project**, Volume 3 and Article 24, **The Human Genome Project**, Volume 3). More typically, time and resources must still be devoted to clarifying gene content. Novel genes and new exons of incompletely defined gene transcripts, still arise at a significant rate in positional cloning projects, although this should become less of an issue within the next several years.

There are numerous approaches to mutation detection. Ultimately, the gold standard remains direct DNA sequencing, but several other physicochemical techniques (denaturing HPLC, mismatch scanning, chip-based resequencing, etc.) have been developed as alternatives for first-pass analysis. This remains a fast-moving field with technical improvements ongoing. Even for direct sequencing, software tools for more highly automated detection of sequence variants are still in development (Polyphred, Sequencher, Staden Tracediff, Softgenetics MutationSurveyor, etc.).

For linkage-based positional cloning, rare variants, typically resulting in obvious changes in gene function (frameshifts, stop codons, missense changes in conserved or biochemically validated amino acid residues, changes in conserved splice junction elements, etc.), are generally well accepted as causal for rare phenotypes especially when they are absent or at vanishing small frequencies in the general population (usually such a mutation is defined as undetected in at least 100 random control individuals). The overall mutation rate in the general population is such that for severe loss-of-function mutations there are usually many different allelic mutations detected for the same phenotype. Once a positional cloning project

has led to provisional identification of a small number of mutations, follow-up validation of new mutations in additional patients often ensues rapidly. In the cases where phenotypes are proposed to arise from unusual types of mutations, such as specific gain-of-function changes, validation may be more challenging.

7. Future of linkage mapping

We believe that linkage mapping still plays an important role in disease gene identification. Of the proposed 25–30 000 human genes (Jaillon *et al*., 2004), fewer than 2000 have identified genetic variants associated with a clear phenotype in the OMIM database. For discovering the phenotypic effects of mutation in the remaining genes, especially for high penetrance mutations such as severe loss-of-function, traditional family-based linkage analysis is still the most efficient technology available. Homozygosity mapping of recessive phenotypes in appropriate populations, through a hybrid linkage/LD approach, can be even more efficient in the gene discovery process.

8. Electronic databases

These indicate some of the most commonly used sites for sources of genetic and genomic data and services.

Genome project data, maps, annotations:

NCBI **http://www.ncbi.nlm.nih.gov/**
UCSC **http://genome.ucsc.edu/**
Ensembl **http://www.ensembl.org/**

Genetic markers and maps:

Marshfield **http://research.marshfieldclinic.org/genetics/**
Genethon **http://www.cephb.fr/ceph-genethon-map.html**
GDB **http://www.gdb.org/**
Decode **http://www.decode.com/**
NCBI **http://www.ncbi.nlm.nih.gov/**
SNP consortium **http://snp.cshl.org/**
Hapmap **http://www.hapmap.org/**

Genetic linkage analysis software:

Rockefeller University **http://linkage.rockefeller.edu/** and **http://linkage.rockefeller.edu/soft/list.html**
MIT **http://www.broad.mit.edu/humgen/soft.html**
UCLA **http://www.biomath.medsch.ucla.edu/faculty/klange/software.html**
NCBI **http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html**
Univ of Pittsburgh **http://watson.hgen.pitt.edu/register/soft_doc.html**

Genetic diseases and mutations:

OMIM **http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM**
Human Genome Variation database **http://hgvbase.cgb.ki.se/**
Cardiff **http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html**
Weizmann **http://bioinfo.weizmann.ac.il/cards/index.html**
HUGO **http://www.gene.ucl.ac.uk/nomenclature/**
UK HGMP Resource Center **http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-mutation.html**

Commercial genotyping services and marker suppliers:

Decode **http://www.decode.com/**
Marshfield **http://research.marshfieldclinic.org/genetics/**
Australian Genome Research Facility **http://www.agrf.org.au/**
Center for Inherited Disease Research **http://www.cidr.jhmi.edu/**
Montreal **http://www.cgdn.generes.ca/eng/core/genotyping.html**
Illumina **http://www.illumina.com/**
Affymetrix **http://www.affymetrix.com/**
Applied Biosystems **http://www.appliedbiosystems.com/**

References

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Arcos-Burgos M and Muenke M (2002) Genetics of population isolates. *Clinical Genetics*, **61**, 233–247.

Beckmann JS and Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. *Biotechnology (N Y)*, **8**, 930–932.

Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

Broman KW, Murray JC, Sheffield VC, White RL and Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics*, **63**, 861–869.

Brondani RP and Grattapaglia D (2001) Cost-effective method to synthesize a fluorescent internal DNA standard for automated fragment sizing. *Biotechniques*, **31**, 793–795, 798, 800.

Brownstein MJ, Carpten JD and Smith JR (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques*, **20**, 1004–1006, 1008–1010.

Cottingham RW Jr., Idury RM and Schaffer AA (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics*, **53**, 252–263.

de la Chapelle A and Wright FA (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 12416–12423.

Delmotte F, Leterme N and Simon JC (2001) Microsatellite allele sizing: difference between automated capillary electrophoresis and manual technique. *Biotechniques*, **31**, 810, 814–816, 818.

DeWan AT, Parrado AR, Matise TC and Leal SM (2002) The map problem: a comparison of genetic and sequence-based physical maps. *American Journal of Human Genetics*, **70**, 101–107.

Gelfi C, Orsi A, Righetti PG, Brancolini V, Cremonesi L and Ferrari M (1994) Capillary zone electrophoresis of polymerase chain reaction-amplified DNA fragments in polymer networks: the case of GATT microsatellites in cystic fibrosis. *Electrophoresis*, **15**, 640–643.

Ghosh S, Karanjawala ZE, Hauser ER, Ally D, Knapp JI, Rayman JB, Musick A, Tannenbaum J, Te C, Shapiro S, *et al*. (1997) Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group. *Genome Research*, **7**, 165–178.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Gyapay G, Ginot F, Nguyen S, Vignal A and Weissenbach J (1996) Genotyping Procedures in Linkage Mapping. *Methods*, **9**, 91–97.

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M and Weissenbach J (1994) The 1993–94 Genethon human genetic linkage map. *Nature Genetics*, **7**, 246–339.

Haines JL and Pericak-Vance MA (Eds.) (1998) *Approaches to Gene Mapping in Complex Human Diseases*, Wiley-Liss: New York.

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

Holden AL (2002) The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*, **32,** S22–S26.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al*. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, **31**, 241–247.

Kruglyak L, Daly MJ, Reeve-Daly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Laan M and Paabo S (1998) Mapping genes by drift-generated linkage disequilibrium. *American Journal of Human Genetics*, **63**, 654–656.

Lander ES and Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Lathrop GM, Lalouel JM, Julier C and Ott J (1984) Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3443–3446.

Lindqvist AK, Magnusson PK, Balciuniene J, Wadelius C, Lindholm E, Alarcon-Riquelme ME and Gyllensten UB (1996) Chromosome-specific panels of tri- and tetranucleotide microsatellite markers for multiplex fluorescent detection and automated genotyping: evaluation of their utility in pathology and forensics. *Genome Research*, **6**, 1170–1176.

Litt M and Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, **44**, 397–401.

Magnuson VL, Ally DS, Nylund SJ, Karanjawala ZE, Rayman JB, Knapp JI, Lowe AL, Ghosh S and Collins FS (1996) Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotechniques*, **21**, 700–709.

Mansfield ES, Vainer M, Enad S, Barker DL, Harris D, Rappaport E and Fortina P (1996) Sensitivity, reproducibility, and accuracy in short tandem repeat genotyping using capillary array electrophoresis. *Genome Research*, **6**, 893–903.

Mansfield ES, Vainer M, Harris DW, Gasparini P, Estivill X, Surrey S and Fortina P (1997) Rapid sizing of polymorphic microsatellite markers by capillary array electrophoresis. *Journal of Chromatography A*, **781**, 295–305.

Markianos K, Daly MJ and Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, **68**, 963–977.

Morton NE (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics*, **7**, 277–318.

O'connell JR and Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

O'Connell JR and Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, **63**, 259–266.

Ott J (1991) *Analysis of Human Genetic Linkage*, Johns Hopkins University Press: Baltimore.

Reed PW, Davies JL, Copeman JB, Bennett ST, Palmer SM, Pritchard LE, Gough SC, Kawaguchi Y, Cordell HJ, Balfour KM, *et al*. (1994) Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genetics*, **7**, 390–395.

Ruschendorf F and Nurnberg P (2005) ALOHOMORA: a tool for linkage analysis using 10K SNP array data. *Bioinformatics*, **21**, 2123–2125.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Schaffer AA, Gupta SK, Shriram K and Cottingham RW Jr. (1994) Avoiding recomputation in linkage analysis. *Human Heredity*, **44**, 225–237.

Sellick GS, Longman C, Tolmie J, Newbury-Ecob R, Geenhalgh L, Hughes S, Whiteford M, Garrett C and Houlston RS (2004) Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. *Nucleic Acids Research*, **32**, e164.

Sham P (1998) *Statistics in Human Genetics*, John Wiley & Sons: New York.

Sheffield VC, Stone EM and Carmi R (1998) Use of isolated inbred human populations for identification of disease genes. *Trends in Genetics*, **14**, 391–396.

Sobel E, Sengul H and Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Human Heredity*, **52**, 121–131.

Taylor GR, Noble JS, Hall JL, Stewart AD and Mueller RF (1989) Hypervariable microsatellite for genetic diagnosis. *Lancet*, **2**, 454.

Terwilliger JD and Ott J (1994) *Handbook of Human Genetic Linkage*, Johns Hopkins University Press: Baltimore.

Thiele H and Nurnberg P (2005) HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, **21**, 1730–1732.

Thomas A, Gutin A, Abkevich V and Bansal A (2000) Multipoint linkage analysis by blocked Gibbs sampling. *Statistics in Computing*, **10**, 259–269.

Timms KM, Wagner S, Samuels ME, Forbey K, Goldfine H, Jammulapati S, Skolnick MH, Hopkins PN, Hunt SC and Shattuck DM (2004) A mutation in PCSK9 causing autosomal-dominant hypercholesterolemia in a Utah pedigree. *Human Genetics*, **114**, 349–353.

Tsai Y-Y, Pugh EW, Boyce P, Doheny KF, Fan Y-T, Scott AF, St. Hansen M, Oliphant A, Loi H, Mei R, *et al*. (2003) *American Society of Human Genetics*, Vol. 73s, American Journal of Human Genetics: Los Angeles.

Utah Marker Development Group (1995) A collection of ordered tetranucleotide-repeat markers from the human genome. *American Journal of Human Genetics*, **57**, 619–628.

Vainer M, Enad S, Dolnik V, Xu D, Bashkin J, Marsh M, Tu O, Harris DW, Barker DL and Mansfield ES (1997) Short tandem repeat typing by capillary array electrophoresis: comparison of sizing accuracy and precision using different buffer systems. *Genomics*, **41**, 1–9.

Webb EL, Sellick GS and Houlston RS (2005) SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. *Bioinformatics*, (Epub ahead of print).

Weber JL (1990) Human DNA polymorphisms and methods of analysis. *Current Opinion in Biotechnology*, **1**, 166–171.

Weber JL and Broman KW (2001) Genotyping for human whole-genome scans: past, present, and future. *Advances in Genetics*, **42**, 77–96.

Weeks DE, Sobel E, O'connell JR and Lange K (1995) Computer programs for multilocus haplotyping of general pedigrees. *American Journal of Human Genetics*, **56**, 1506–1507.

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G and Lathrop M (1992) A second-generation linkage map of the human genome. *Nature*, **359**, 794–801.

Wenz HM, Dailey D and Johnson MD (2001) Development of a high-throughput capillary electrophoresis protocol for DNA fragment analysis. *Methods in Molecular Biology*, **163**, 3–17.

Wenz H, Robertson JM, Menchen S, Oaks F, Demorest DM, Scheibler D, Rosenblum BB, Wike C, Gilbert DA and Efcavitch JW (1998) High-precision genotyping by denaturing capillary electrophoresis. *Genome Research*, **8**, 69–80.

Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, *et al*. (2001) Comparison of human genetic and sequence-based physical maps. *Nature*, **409**, 951–953.

Ziegle JS, Su Y, Corcoran KP, Nie L, Mayrand PE, Hoff LB, McBride LJ, Kronick MN and Diehl SR (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics*, **14**, 1026–1031.

Short Specialist Review

Microarray comparative genome hybridization

Robert A. Holt and Martin Krzywinski
Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada

1. Introduction and history

Comparative genome hybridization (CGH) is a method for genome-wide detection of chromosomal differences (*see* Article 11, **Human cytogenetics and human chromosome abnormalities**, Volume 1) between a sample and control that are due to DNA copy number changes. Briefly, total genomic DNA from a "test" and a "reference" individual are labeled with different fluorescent dyes and cohybridized to a representation of the genome in the presence of CoT-1 DNA (At a given temperature, the rate of DNA renaturation depends on concentration (Co) and time (t). CoT-1 DNA represents a rapidly reassociating and thus a highly repeat-enriched fraction of genomic DNA. It is typically derived by denaturing sheared gDNA at a concentration of 3 mM, reassociating for 5.5 min, and then isolating the reassociated double-stranded product.), which is used to block repetitive sequence. The ratio of signals emitted from different loci provides a map of variation in copy number in the genome of the "test" individual. Originally, metaphase chromosome spreads were used as the genome representation (Kallioniemi *et al*., 1992) for CGH, and in this format the technique has been widely used (Nacheva *et al*., 1998; Brown and Botstein, 1999; James, 1999; Weiss *et al*., 1999; Ness *et al*., 2002; Albertson and Pinkel, 2003) for the analysis of tumors (*see* Article 14, **Acquired chromosome abnormalities: the cytogenetics of cancer**, Volume 1) and developmental abnormalities such as mental retardation and congenital anomalies. A number of experimental and analytical modifications (*see* Article 58, **CGH data analysis**, Volume 7) have been proposed to increase the resolution, such as standard reference intervals (Kirchhoff *et al*., 1999), and precision, such as four-color CGH (Karhu *et al*., 1999). Microarray comparative genome hybridization (maCGH) represents an evolution of the classical method, whereby chromosome spreads are replaced by DNA fragments of known genomic location spotted on a microarray slide. There are several variations of the maCGH format, in which either BACs (Bacterial Artificial Chromosomes), cDNAs, or oligonucleotides are used as the DNA target. Regardless of format, maCGH offers distinct advantages over both classical CGH and other microscopic cytogenetic methods. The resolving power of maCGH is considerably greater than the maximum of approximately

5 Mb achievable by G-banding (*see* Article 12, **The visualization of chromosomes**, Volume 1) or 1–2 Mb (amplifications) and 10 Mb (deletions) by conventional CGH (Bentz *et al*., 1998; Kirchhoff *et al*., 1999). The only theoretical limits on resolution are the number, size, and sampling density of the targets on the array. Further, the method is more scalable than microscopic methods, allowing the parallel and quantitative (Moore *et al*., 1997; Kirchhoff *et al*., 1998; Quackenbush, 2002; Geller *et al*., 2003) evaluation of large numbers of samples, and does not require intact chromosomes for analysis. The most significant limitations of maCGH are (1) the genomic location of amplified DNA sequence is not known and (2) unless chromosomes are first separated by flow cytometry then labeled and hybridized individually (a process called array painting (Fiegler *et al*., 2003a)), the assay is blind to chromosomal aberrations that do not result in copy number changes, such as balanced translocations. Nonetheless, maCGH has proven its utility through the detection of DNA copy number changes in tumors (Albertson *et al*., 2000; Bruder *et al*., 2001; Struski *et al*., 2002; Nakao *et al*., 2004; Cai *et al*., 2001; Zhao *et al*., 2004), children with mental retardation and various dysmorphic syndromes (Shaw-Smith *et al*., 2004; Veltman *et al*., 2002; Xu and Chen, 2003; Yu *et al*., 1997), and molecular evolution (Locke *et al*., 2003).

2. BAC arrays

Presently, the construction of genomic microarrays is dominated by the use of BACs as the target for hybridization. Several of the BAC libraries that provided critical positional information for guiding sequencing and assembly of the human genome (Lander *et al*., 2001; Venter *et al*., 2001) are available for array construction. Initial use of these resources provided first-generation arrays with approximately 1-Mb resolution (Snijders *et al*., 2001; Fiegler *et al*., 2003b). Recently, an optimal tiling set of clones providing coverage for the entire genome has been selected (Krzywinski *et al*., 2004) and a high-resolution BAC microarray has been manufactured using these clones (Ishkanian *et al*., 2004). Theoretical resolution of this clone set is based on the degree of clone overlap, and is calculated to be 75 kb.

BACs are desirable as hybridization targets not only because their genomic positions are known accurately but also because their large insert size (approximately 150–200 kb on average) allows integration of hybridization signal over a comparatively large region and gives sufficient sensitivity to routinely detect single copy number changes starting with only a few hundred nanograms of labeled test DNA (Albertson and Pinkel, 2003). Preparation and spotting of BACs on to arrays is made difficult by the low yield of DNA from BAC cultures and the large molecular weight of the DNA. Both factors are a detriment to handling DNA at the high concentration necessary for achieving good signal-to-noise ratio in hybridizations. These problems have been overcome by preparing a representation of each BAC clone by ligation-mediated PCR (LMPCR), whereby clones are fragmented, oligonucleotide adapters are ligated to the ends of fragments, and the fragments are amplified by PCR using adapter-specific primers. In this manner, a large and renewable quantity of DNA suitable for array printing is generated. LMPCR was the first reported technique for the preparation of clones for maCGH

and ratio data obtained from arrays composed of LMPCR BAC representations have been shown to be essentially identical to ratios reported on intact DNA from the same BACs (Pinkel *et al*., 1998). Degenerate oligo primed PCR (DOP-PCR) (Fiegler *et al*., 2003b) and rolling circle amplification (RCA) (Smirnov *et al*., 2004; Buckley *et al*., 2002) have also been successfully used in the preparation of BAC DNA for spotting. The principal drawbacks of BAC arrays include the ultimate limit of resolution determined by their large insert size and the continued necessity for using large amounts of CoT-1 DNA to block highly repetitive sequences (although numerical methods exist to mitigate this effect (Kirchhoff *et al*., 1997)). Further complications from repeat elements arise in telomeric and pericentromeric regions. While these regions often contain loci of interest, they are highly repetitive and therefore masked by CoT-1 DNA. Care must also be taken to avoid being misled by low-copy-repeat elements that are not masked by CoT-1 DNA. It is estimated that 5% of the human genome is made up of interspersed duplications (*see* Article 26, **Segmental duplications and the human genome**, Volume 3) (Eichler, 2001; Bailey *et al*., 2002) that represent, for example, homology between gene families, and these naturally occurring duplications can confound analysis of BAC maCGH data.

3. cDNA arrays

The use of cDNA clones as the target for hybridization in maCGH (Pollack *et al*., 1999; Kargul *et al*., 2001; VanBuren *et al*., 2002; Yamamoto *et al*., 2002) has obvious advantages in terms of the number and variety of clone sets and prefabricated arrays available for human studies, but also for studies of other model organisms, pathogens, disease vectors, novel therapeutics, and organisms of industrial importance for which no genome sequence or validated large insert genomic clone set is yet available. While CGH using cDNA arrays is informative only for coding sequence, concentrating resolving power on this fraction of the genome can be considered an advantage, particularly when gDNA and RNA are available from the same individual, allowing cointerrogation of gene dosage and gene expression at precisely the same loci. Information on copy number changes in gene regulatory regions or other nontranscribed regions may be missed, but the same is true for some of the current generation BAC arrays that do not offer complete genome coverage. The principal drawback in using cDNA clones as hybridization targets is limited sensitivity. Relatively large amounts of labeled DNA (up to 10 μg) are required for each hybridization and the resulting signal must be averaged over a number of clones to define local copy number (Pollack *et al*., 1999). While large genomic amplifications are readily detectable, cDNA arrays are generally not considered to be the best tool for detection of single copy number differences.

4. Oligonucleotide arrays

Two recent developments in the application of oligonucleotide arrays (*see* Article 57, **Low-level analysis of oligonucleotide expression arrays**, Volume 7) to DNA copy

number analysis have shown promise: representational oligonucleotide microarray analysis (ROMA) (Lucito *et al*., 2003; Sebat *et al*., 2004) and the use of Affymetrix SNP chips. ROMA is an interesting approach enabled entirely by completion of the reference human genome sequence. In ROMA, a representation of the genome sequence is prepared by digesting gDNA with a restriction enzyme (BglII) and fragments are amplified using the same basic procedure as LMPCR, described above. ROMA arrays are spotted with oligonucleotides (70mers) that are designated to have near-homogeneous annealing characteristics, and match unique (nonrepetitive) sequence present within computationally defined BglII fragments. Thus, the target sequence on the array is repeat-free, obviating the need for CoT-1 DNA as a blocking agent. The reduced complexity of the target and probe fractions improves signal-to-noise performance and reduces the amount of sample required for hybridization. In principle, the resolution is very high, but in practice a finite number (approximately 120 000) of repeat-free BglII fragment 70mers in the human genome places an upper limit on resolution. Resolution could be increased further, in theory, by digesting with more than one restriction enzyme, and in time, different restriction enzymes and enzyme combinations will likely be found that give an optimal number and spacing of targets. An 85 000-element ROMA array has been characterized (Lucito *et al*., 2003) and has been shown to be capable of detecting both known and novel single and multiple copy deletions and amplifications, including several less than 100 kb in length. The practice of ROMA is restricted to organisms that have a quality whole-genome sequence available, and the array design steps are demanding, but this approach holds much promise.

Single nucleotide polymorphism (SNP) (*see* Article 71, **SNPs and human history**, Volume 4) have recently been developed by Affymetrix for array-based genotyping (Kennedy *et al*., 2003), and these arrays have also shown some promise as a platform for evaluation of DNA copy number (Zhao *et al*., 2004; Bignell *et al*., 2004). These arrays contain allele-specific 25mer oligonucleotide probes complementary to SNPs predicted to be in the fraction of the genome represented by the digestion fragments generated by the enzyme used in sample preparation (typically Xba1 or HindIII). Current arrays formats contain up to 100 000 SNPs and provide resolution as low as approximately 30 kb. Multiple different oligonucleotide probes cover each polymorphic site on both the sense and antisense strand. Like ROMA, preparation of sample DNA for hybridization relies on LMPCR for sample complexity reduction and amplification, the difference being that for the Affymetrix experiments, XbaI or HindIII, rather than BglII, is the enzyme used for digestion of sample gDNA. It is important to note that the use of SNP arrays for DNA copy number evaluation is fundamentally different from ROMA, BAC or cDNA maCGH in two regards. First, the SNP chip assay does not rely on comparative hybridization. Rather, each sample DNA is individually labeled and hybridized. Copy number differences are detectable only in comparison to reference DNA samples evaluated in separate experiments and stored in a database provided by Affymetrix. Second, because alleles are present as separate array elements, the SNP chip platform uniquely enables loss of heterozygosity events that are caused by hemizygous deletion to be distinguished from those that are caused by copy number neutral events, such as deletion followed by subsequent duplication of the remaining locus. Loss of heterozygosity (LOH) is common in cancer cells (Vogelstein and

Kinzler, 1998), where many tumor suppressor genes are inactivated by mutation in one allele and hemizygous deletion of the other wild-type allele. However, other LOH mechanisms such as mitotic recombination or gene conversion do not lead to copy number changes, and it is important to be able to distinguish between these mechanisms. Further, genetic deletion syndromes such as Angelman's syndrome have different outcomes depending on whether the deleted allele is maternally or paternally inherited. Thus, the ability to distinguish parent-of-origin effects has implications for genetic diagnosis and counseling.

The present Affymetrix SNP chips detect homozygous deletions, hemizygous deletions, and amplifications simultaneously with LOH detection (Zhao *et al*., 2004; Bignell *et al*., 2004). Direct comparison with BAC and cDNA array analysis (Zhao *et al*., 2004) showed that the three platforms gave generally comparable copy number results, although the noise of individual measurements was greater on the SNP chip platform, and analysis of raw data using a Hidden Markov Model (*see* Article 98, **Hidden Markov models and neural networks**, Volume 8) was necessary to obtain the best inference of copy number. As with ROMA, high target density is possible with this approach, and a 100 000-element XbaI-based SNP array is presently under construction, which will likely prove to be a very powerful and useful tool.

5. Experimental considerations

While the platforms described above (BAC CGH, cDNA CGH, ROMA, Affymetrix SNP chips) have all been shown to have utility in evaluation of DNA copy number, there are several additional issues to be considered when investigating DNA copy number aberrations. These include the amount, integrity and source of sample and reference DNA, the sensitivity of the assay, and the prevalence of copy number polymorphisms. Regarding input sample and reference DNA, BAC array CGH and the ROMA-based methods require several hundred nanograms of genomic DNA per hybridization, and cDNA arrays require considerably more. While several hundred micrograms of DNA seem a modest amount, even this quantity can be difficult to obtain from clinical samples, particularly from microdissected tissue, or from postmortem tissue where DNA may have degraded. Regarding reference DNA, it is desirable, if not essential, to use the same reference DNA within a series of hybridizations from the same study, or to compare across studies. Thus, there is a need for a large repository of reference DNA in laboratories performing this assay. Ideally, this would be constitutional DNA from a single donor with a defined karyotype, but because this is impractical, pooled DNA from multiple individuals of the same gender (available commercially form Clontech or Novagen) is often used as reference. With sufficient numbers of individuals represented in a DNA pool, any individual karyotypic anomalies become negligible. While in principle DNA from selected cell lines would offer a renewable source of reference DNA, the prevalence of karyotypic anomalies in immortalized cells may make results difficult to interpret, and caution is advised. Of note, recent success in amplifying sample and reference DNA using Phi29 polymerase or Bst polymerase suggests that this approach may provide a practical solution where input DNA is

limiting for CGH experiments. Initial results show limited representational bias and background amplification if experimental conditions are carefully controlled (Lage *et al*., 2003). Even with appropriate input sample and reference DNA, a common observation is that maCGH generally does not achieve theoretical values for copy number differences (Albertson and Pinkel, 2003). For example, female test versus male reference comparisons typically give less than the expected 3:2 ratio for X-chromosome probes and measurable signal for the presumably absent Y chromosome. The reasons for dynamic range suppression are poorly understood, but may relate to the presence of somatic mosaicism (*see* Article 18, **Mosaicism**, Volume 1) (as in the case of tumor samples contaminated with surrounding stromal cells of normal karyotype) or, for clone-based arrays, deletion or insertion events that span less than the length of the target cDNA or BAC. Incomplete suppression of repetitive sequence may also be implicated. It is for these reasons that independent verification of all putative copy number changes by a second method such as FISH (*see* Article 22, **FISH**, Volume 1) or quantitative real-time PCR remains essential.

Copy number polymorphisms (CNPs) have the potential to confound CGH analysis. While there are only a small number of well documented CNPs in the human population, such as the Rh locus (Wagner and Flegel, 2000), the CYP2D6 locus (Meyer and Zanger, 1997), and the green color pigment locus (Nathans *et al*., 1986), as maCGH becomes broadly applied it is becoming clear that CNPs are not uncommon, and represent an important source of genetic variation (Sebat *et al*., 2004; Iafrate *et al*., 2004). DNA copy number variation is clearly a hallmark of tumor cells (Vogelstein and Kinzler, 1998), and there is also evidence that substantial levels of chromosomal anueploidy may exist in neurons (Rehen *et al*., 2001). Because many or perhaps most CNPs may be benign, a survey of common polymorphisms in different ethnic populations would provide a valuable resource for interpreting disease-focused CGH studies. Presently, however, owing to the unknown scope of copy number polymorphism, it is important that studies investigating CNP disease associations include an appropriate group of matched control individuals.

References

Albertson DG and Pinkel D (2003) Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics*, **12** Spec No 2, R145–R152.

Albertson DG, Ylstra B, Segraves R, Collins C, Dairkee SH, Kowbel D, Kuo WL, Gray JW and Pinkel D (2000) Quantitative mapping of amplicon structure by array cgh identifies cyp24 as a candidate oncogene. *Nature Genetics*, **25**, 144–146.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW and Eichler EE (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Bentz M, Plesch A, Stilgenbauer S, Dohner H and Lichter P (1998) Minimal sizes of deletions detected by comparative genomic hybridization. *Genes, Chromosomes & Cancer*, **21**, 172–175.

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, **14**, 287–295.

Brown PO and Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.

Bruder CE, Hirvela C, Tapia-Paez I, Fransson I, Segraves R, Hamilton G, Zhang XX, Evans DG, Wallace AJ, Baser ME, *et al*. (2001) High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (nf2) patients using microarray-cgh. *Human Molecular Genetics*, **10**, 271–282.

Buckley PG, Mantripragada KK, Benetkiewicz M, Tapia-Paez I, Diaz De Stahl T, Rosenquist M, Ali H, Jarbo C, De Bustos C, Hirvela C, *et al*. (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Human Molecular Genetics*, **11**, 3221–3229.

Cai WW, Chen R, Gibbs RA and Bradley A (2001) A clone-array pooled shotgun strategy for sequencing large genomes. *Genome Research*, **11**, 1619–1623.

Eichler EE (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, **17**, 661–669.

Fiegler H, Gribble SM, Burford DC, Carr P, Prigmore E, Porter KM, Clegg S, Crolla JA, Dennis NR, Jacobs P, *et al*. (2003a) Array painting: A method for the rapid analysis of aberrant chromosomes using DNA microarrays. *Journal of Medical Genetics*, **40**, 664–670.

Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IP, *et al*. (2003b) DNA microarrays for comparative genomic hybridization based on dop-pcr amplification of bac and pac clones. *Genes, Chromosomes & Cancer*, **36**, 361–374.

Geller SC, Gregg JP, Hagerman P and Rocke DM (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**, 1817–1823.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **36**, 949–951.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**(3), 299–303.

James LA (1999) Comparative genomic hybridization as a tool in tumour cytogenetics. *The Journal of Pathology*, **187**, 385–395.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kargul GJ, Dudekula DB, Qian Y, Lim MK, Jaradat SA, Tanaka TS, Carter MG and Ko MS (2001) Verification and initial annotation of the NIA mouse 15K cDNA clone set. *Nature Genetics*, **28**, 17–18.

Karhu R, Rummukainen J, Lorch T and Isola J (1999) Four-color cgh: A new method for quality control of comparative genomic hybridization. *Genes, Chromosomes & Cancer*, **24**, 112–118.

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, *et al*. (2003) Large-scale genotyping of complex DNA. *Nature Biotechnology*, **21**, 1233–1237.

Kirchhoff M, Gerdes T, Maahr J, Rose H, Bentz M, Dohner H and Lundsteen C (1999) Deletions below 10 megabasepairs are detected in comparative genomic hybridization by standard reference intervals. *Genes, Chromosomes & Cancer*, **25**, 410–413.

Kirchhoff M, Gerdes T, Rose H, Maahr J, Ottesen AM and Lundsteen C (1998) Detection of chromosomal gains and losses in comparative genomic hybridization analysis based on standard reference intervals. *Cytometry* **31**(3), 163–173.

Kirchhoff M, Gerdes T, Maahr J, Rose H and Lundsteen C (1997) Automatic correction of the interfering effect of unsuppressed interspersed repetitive sequences in comparative genomic hybridization analysis. *Cytometry*, **28**, 130–134.

Krzywinski M, Bosdet I, Smailus D, Chiu R, Mathewson C, Wye N, Barber S, Brown-John M, Chan S, Chand S, *et al*. (2004) A set of BAC clones spanning the human genome. *Nucleic Acids Research*, **12**, 3651–3660.

Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, Dillon D, Segraves R, Vossbrinck B, Gonzalez A, Pinkel D, *et al*. (2003) Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-cgh. *Genome Research*, **13**, 294–307.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D and Eichler EE (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Research*, **13**, 347–357.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research*, **13**, 2291–2305.

Meyer UA and Zanger UM (1997) Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269–296.

Moore DH II, Pallavicini M, Cher ML and Gray JW (1997) A t-statistic for objective interpretation of comparative genomic hybridization (cgh) profiles. *Cytometry*, **28**, 183–190.

Nacheva EP, Grace CD, Bittner M, Ledbetter DH, Jenkins RB and Green AR (1998) Comparative genomic hybridization: A comparison with molecular and cytogenetic analysis. *Cancer Genetics and Cytogenetics*, **100**, 93–105.

Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, Lafuente A, Wiencke JW, Terdiman JP and Waldman FM (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**(8), 1345–1357.

Nathans J, Thomas D and Hogness DS (1986) Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. *Science*, **232**, 193–202.

Ness GO, Lybaek H and Houge G (2002) Usefulness of high-resolution comparative genomic hybridization (cgh) for detecting and characterizing constitutional chromosome abnormalities. *American Journal of Medical Genetics*, **113**, 125–136.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, **23**, 41–46.

Quackenbush J (2002) Microarray data normalization and transformation. *Nature Genetics*, **32**(Suppl), 496–501.

Rehen SK, McConnell MJ, Kaushal D, Kingsbury MA, Yang AH and Chun J (2001) Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13361–13366.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, *et al*. (2004) Microarray based comparative genomic hybridisation (array-cgh) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *Journal of Medical Genetics*, **41**, 241–248.

Smirnov DA, Burdick JT, Morley M and Cheung VG (2004) Method for manufacturing whole-genome microarrays by rolling circle amplification. *Genes, Chromosomes & Cancer*, **40**, 72–77.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263–264.

Struski S, Doco-Fenzy M and Cornillet-Lefebvre P (2002) Compilation of published comparative genomic hybridization studies. *Cancer Genetics and Cytogenetics*, **135**, 63–90.

VanBuren V, Piao Y, Dudekula DB, Qian Y, Carter MG, Martin PR, Stagg CA, Bassey UC, Aiba K, Hamatani T, *et al*. (2002) Assembly, verification, and initial annotation of the NIA mouse 7.4K cDNA clone set. *Genome Research*, **12**, 1999–2003.

Veltman JA, Schoenmakers EF, Eussen BH, Janssen I, Merkx G, van Cleef B, van Ravenswaaij CM, Brunner HG, Smeets D and van Kessel AG (2002) High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *American Journal of Human Genetics*, **70**, 1269–1276.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Vogelstein B and Kinzler KW (1998) *The Genetic Basis of Human Cancer*, McGraw-Hill: New York.

Wagner FF and Flegel WA (2000) Rhd gene deletion occurred in the rhesus box. *Blood*, **95**, 3662–3668.

Weiss MM, Hermsen MA, Meijer GA, van Grieken NC, Baak JP, Kuipers EJ and van Diest PJ (1999) Comparative genomic hybridisation. *Molecular Pathology*, **52**, 243–251.

Xu J and Chen Z (2003) Advances in molecular cytogenetics for the evaluation of mental retardation. *American Journal of Medical Genetics*, **117C**, 15–24.

Yamamoto H, Imsumran A, Fukushima H, Adachi Y, Min Y, Iku S, Horiuchi S, Yoshida M, Shimada K, Sasaki S, *et al*. (2002) Analysis of gene expression in human colorectal cancer tissues by cdna array. *Journal of Gastroenterology*, **37**(Suppl 14), 83–86.

Yu LC, Moore DH II, Magrane G, Cronin J, Pinkel D, Lebo RV and Gray JW (1997) Objective aneuploidy detection for fetal and neonatal screening using comparative genomic hybridization (cgh). *Cytometry*, **28**, 191–197.

Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*, **64**, 3060–3071.

Short Specialist Review

Linkage disequilibrium and whole-genome association studies

Karen L. Novik and Angela R. Brooks-Wilson
Genome Sciences Centre, Vancouver, BC, Canada

1. Introduction

Complex diseases are those that involve multiple genetic loci as well as environmental or lifestyle effects (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2). Such diseases often affect a substantial proportion of the population. Uncovering the genetic components of such diseases is a current challenge for human genetics.

2. Complex diseases and association studies

As an example, let us consider the situation for one complex disease, breast cancer. It has been estimated that less than 2% of all breast cancer cases are caused by rare inherited mutations in two genes called *BRCA1* and *BRCA2* and that these genes account for only 20% of the excess familial risk of the disease (Anglian Breast Cancer Study Group, 2000). Clearly, other breast cancer susceptibility genes remain to be identified, especially for the sporadic form of the disease that is rarely the result of mutations in the *BRCA* genes (Gayther *et al*., 1998). Segregation analysis suggests that there could be many different breast cancer loci, each contributing a small effect (Antoniou *et al*., 2001).

The association study is a recent method that can be used to identify complex disease genes (for review see Cardon and Bell, 2001). This method compares the frequency of genetic variants in unrelated cases (who have a given disease) and controls (who are free of disease) to identify variants or regions that are putatively involved in disease etiology. If such variants are *associated* with disease, further characterization is necessary to demonstrate a causal role in the disease process. Association studies can use a candidate gene approach to investigate polygenic disorders; the choice of candidates is based on previous biological and/or genetic insights into that disease. BRCA1 and BRCA2 for example, have roles in mammalian DNA double-strand break (DSB) repair, and this has provided a rationale for breast cancer association studies that have focused on numerous members of this molecular pathway. Variants and/or haplotypes of BRCA2,

XRCC2, XRCC3, and Ligase 4 have all been associated with modest risks of breast cancer and population-attributable risks (the proportion of a population's breast cancer that is due to a particular genetic variant) of up to 2% (Healey *et al*., 2000; Kushel *et al*., 2002). Although candidate gene approaches have successfully identified low-risk variants for breast cancer susceptibility, this approach is only just beginning to address the genetics of this complex disease.

3. Whole-genome association and linkage disequilibrium

Whole-genome association studies, in contrast to candidate gene–based studies, do not require existing knowledge of the relevance of specific genes, pathways, or biological hypotheses in order to identify the genetic determinants of disease. Whole-genome scans use genetic markers of at least moderate allele frequency distributed across the genome. There are currently 5 million validated SNPs distributed across the 3 billion base pairs of human sequence (single-nucleotide polymorphism database – dbSNP build 123 – **www.ncbi.nlm.nih.gov/SNP**), and directed resequencing efforts show that more SNPs exist than are currently in the public domain (Carlson *et al*., 2003). Is it necessary to include all of this genetic variation in a whole-genome scan for genetic association? Fortunately, the genetic phenomenon of linkage disequilibrium (LD) reduces the number of variants necessary for a whole-genome association study. LD, otherwise known as nonrandom association of alleles, can be used to correlate genetic variation with phenotypic traits (*see* Article 73, **Creating LD maps of the genome**, Volume 4). LD between alleles of physically linked markers is an indication of their recombination history in the population, and can be affected by numerous contributing factors such as recombination rate, mutation age, genetic drift, ethnic diversity and natural selection. LD can vary significantly within and between different populations, in particular, Europeans show greater LD than African populations (for review, see Ardlie *et al*., 2002). Furthermore, LD varies between and across whole chromosomes (Reich *et al*., 2001; Patil *et al*., 2001; Dawson *et al*., 2002). Many studies suggest that the human genome is organized into haplotype blocks that show high LD, interspersed with shorter regions of high recombination and consequently low LD (Gabriel *et al*., 2002; Ardlie *et al*., 2002 and references therein). Certainly, chromosomes 21 and 22 both show this blocklike LD structure (Patil *et al*., 2001; Dawson *et al*., 2002). Common haplotypes can represent most of the genetic variation across relatively large regions of the genome. These haplotypes (including the known and unknown variation) can be genotyped by using a small number of "haplotype tagging" SNPs (htSNPS) that suffice to specify all reasonably common haplotypes in the population of interest (*see* Article 12, **Haplotype mapping**, Volume 3). Thus, LD, in the form of haplotypes, can be used to reduce the number of SNPs needed to genotype a particular genomic region or the entire genome. An international collaborative effort, the HapMap project, is underway to determine the size and boundaries of the human haplotype blocks. This project is now midway through the typing of 600 000 SNPs (on average 1 every 5000 bp) in each of the three populations (International HapMap Consortium, 2003; Couzin, 2004). It has already become clear that this number of SNPs will be insufficient to produce a refined haplotype map representative of all populations (Gabriel *et al*., 2002).

The number of markers necessary to conduct a whole-genome scan for association will be a function of the average size of a haplotype block in the human genome and the number of markers necessary per block to specify all reasonably common haplotypes in populations of interest. Estimates of the number of markers required range from 100 000 to 1 million SNPs (Gabriel *et al*., 2002; International HapMap Consortium, 2003; Carlson *et al*., 2003). Until the completion of the HapMap, the best current prediction of the average size of an LD block for European populations is in the range of 10–30 kb; blocks in African populations are generally smaller (Gabriel *et al*., 2002; Ardlie *et al*., 2002 and references therein). The feasibility of whole-genome scanning for association will also depend upon the lower limit of odds ratio (OR) that is desirable to identify for a given disorder. Major genetic effects can be detected using smaller case/control groups; subtle effects such as a doubling of risk (OR = 2) require larger sample sizes. The sample size used for a study thus determines whether it will simply skim off the larger genetic effects, neglecting smaller ones, or whether it will be a more thorough assessment of the genome in terms of both major and subtle genetic risk factors. The optimal sample size required for a meaningful whole-genome scan is also impacted by statistical corrections required to adjust for multiple testing.

4. Gene-environment interactions

A more comprehensive understanding of the causes of complex diseases will also depend on studies that incorporate gene-environment interaction. Such studies require both accurate environmental and/or lifestyle data for the same group of individuals that are characterized genetically, thus necessitating even larger sample sizes than purely genetic studies. The sample sizes of association studies are limited by issues such as the cost of phenotypic characterization of cases and controls for a given disorder, which can vary greatly between diseases.

5. Technology and cost

For whole-genome scans for association, cost will be a key consideration. Let us assume that a comprehensive genome scan is likely to involve approximately 500 000 markers and that such an experiment will include at least 1000 samples. The number of genotypes required is hence on the order of 5×10^8 genotypes. If costs were only 1 cent per genotype (this has yet to be achieved routinely), the hypothetical genome scan above would cost $5 million to complete. This figure is unrealistic for all but the largest research groups and, therefore, the per-genotype costs would have to be reduced severalfold to fit into the budgets of most laboratories.

The high cost of genome scanning could be decreased by two means, (1) the use of DNA pools rather than individual samples and (2) the use of very high orders of multiplexing or parallel genotyping of SNPs in individual DNA samples (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4). DNA pooling involves the mixing of precisely equal quantities of individual DNAs to form, for example, “case” and “control” pools, followed by a genotyping procedure

that can determine the allele frequency of each pool at each SNP tested. For analysis of DNA pools, the genotyping procedure used must be quantitative and as sensitive as possible. For complex diseases, it is likely that large numbers of genetic factors, many with subtle effects, will combine to produce disease susceptibility. Genotyping methods that are not sufficiently quantitative or sensitive to detect small differences in allele frequencies between pools would likely be inadequate to dissect many of the genetic factors underlying common complex diseases. To date, few methodologies have been shown, in peer-reviewed publications, to be truly quantitative. Two such methods are the MassARRAY system (Sequenom, Inc.) and pyrosequencing (Biotage AB), which have been shown to quantitatively measure differences in allele frequencies below 2% (Bansal *et al*., 2002; Herbon *et al*., 2003; Gruber *et al*., 2002). While pooling offers a reduction in the number of DNAs to be genotyped, some information is also lost, as differences within a pool can no longer be analyzed. In particular, it is less powerful than individual genotyping when known risk factors (such as smoking, age, sex) are being considered for each sample (Carlson *et al*., 2004). One way to counter this loss is to sort the samples into subpools on the basis of their respective risk factors. This will, however, increase the number of assays per marker, which is contrary to the rationale for pooling in the first instance (Carlson *et al*., 2004). Working from our earlier assumption of 500 000 markers and 1000 samples, a genome scan involving two DNA pools, analyzed in triplicate, would need to have a pool-genotype cost of $0.33 or less to give a total cost of $1 million (corresponding to a fivefold overall reduction in cost compared to individual sample genotyping). In addition, such technologies would need to be accompanied by a ready set of assays corresponding to a suitably dense series of markers.

Other technology developers have focused on highly parallel or highly multiplexed genotyping of individual samples using techniques that need not be precisely quantitative, but are capable of reliably distinguishing heterozygotes. Such technologies include capillary electrophoresis-based methods such as the Applied Biosystem SNPlex™ system (48-plex), the Illumina BeadArray (1536-plex) and chip array-based methods such as those of ParAllele (10 000 nonsynonymous SNPs), Affymetrix (currently 100 000 SNPs per chip), or Perlegen (up to 1.5 million SNPs). Under our assumption of 500 000 SNPs and 1000 samples, the cost of genotyping individual samples using these methods would need to be approximately 0.2 cents per genotype to bring the cost of such a study to $1 million and be attractive to a wide variety of laboratories. The cost of several high-throughput methods is now on the order of cents per genotype (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4). Until this is reduced further, however, whole-genome scans for association will remain in the domain of an exclusive few laboratories or companies with the resources to cover the current costs.

6. What is the current status of human genome scans?

Various corporate groups refer to unpublished genome scans for association. Academic groups, in contrast, have published several papers reporting "genome scans"

for association using only a few thousand markers that are clearly inadequate in number to be considered a representative of the entire human genome. While both types of report represent some progress, it seems that genome scans for association at the present time remain unproven.

In the meantime, as the HapMap moves toward completion and commercial groups vie aggressively to produce faster and cheaper genotyping methods, academic researchers continue to carry out hypothesis-driven candidate gene studies; basing their intelligent guesses on the current understanding of human disease biology. In the future, comparing these results to those of whole-genome scans for association may tell us how much – or how little – we understand about our own genome.

References

Anglian Breast Cancer Study Group (2000) Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, **83**, 1301–1308.

Antoniou AC, Pharoah PDP, McMullan G, Day NE, Ponder BA and Easton D (2001) Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population based study. *Genetic Epidemiology*, **21**, 1–18.

Ardlie KG, Kruglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **3**, 299–309.

Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P and Braun A (2002) Association testing by DNA pooling: an effective initial screen. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16871–16874.

Cardon LR and Bell JI (2001) Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91–99.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**, 518–521.

Carlson CS, Eberle MA, Kruglyak L and Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.

Couzin J (2004) Consensus emerges on HapMap strategy. *Science*, **304**, 671–673.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, *et al*. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**, 544–548.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Gayther SA, Pharoah PDP and Ponder BAJ (1998) The genetics of inherited breast cancer. *Journal of Mammary Gland Biology and Neoplasia*, **3**, 365–376.

Gruber JD, Colligan PB and Wolford JK (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing. *Human Genetics*, **110**, 395–401.

Healey CS, Dunning AM, Teare MD, Chase D, Parker L, Burn J, Chang-Claude J, Mannermaa A, Kataja V, Huntsman DG, *et al*. (2000) A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. *Nature Genetics*, **26**, 362–364.

Herbon N, Werner M, Braig C, Gohlke H, Dutsch G, Illig T, Altmuller J, Hampe J, Lantermann A, Schreiber S, *et al*. (2003) High-resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex disease. *Genomics*, **81**, 510–518.

International HapMap Consortium (2003) The International HapMap project. *Nature*, **426**, 789–796.

Kushel B, Auranen A, McBride S, Novik KL, Antoniou A, Lipscombe JM, Day NE, Easton DF, Ponder BA, Pharoah PD, *et al*. (2002) Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Human Molecular Genetics*, **11**, 1399–1407.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Short Specialist Review

Fingerprint mapping

Jacqueline E. Schein and Martin I. Krzywinski
Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

1. Introduction

Physical maps constructed from fingerprinted clones have been widely used in genomic research, for both genome-wide and region-specific analyses. As with other clone-based physical map construction strategies, one starts with a library of randomly arrayed clones, each clone containing an unknown fragment of DNA derived from the genome of interest, and identifies experimentally clone relationships that describe their proximity to one another in the intact genome. On the basis of these established relationships, an ordered set of overlapping clones representing the underlying genome is generated. In fingerprint map construction, these clone relationships are determined by comparing the characteristic patterns of DNA fragments generated by restriction digests of the cloned DNA (the clone "fingerprint"). Any two clones sharing a large fraction of their DNA are expected to have very similar fingerprint patterns. Therefore, by comparing the similarity of fingerprint patterns of all clones within a set of fingerprinted clones, those with significant similarity can be inferred as being derived from DNA from overlapping segments of the genome. The number and pattern of shared restriction fragments allows the clones to be ordered with respect to each other, thereby reconstructing contiguous regions of the genome. Because they are clone based, these maps provide a sequence-ready resource for genome sequencing efforts (*see* Article 8, **Genome maps and their use in sequence assembly**, Volume 7) and an entry point for cloning and functional analysis of genes of interest. They represent and elucidate the underlying structure of the genome being studied and can be integrated with other genomic and genetic data, such as genetic markers and genomic or gene-based sequences, allowing correlation with whole-genome sequence assemblies as well as other types of genome maps, including cytogenetic maps, genetic linkage maps (*see* Article 15, **Linkage mapping**, Volume 3), radiation hybrid (RH) maps (*see* Article 14, **The construction and use of radiation hybrid maps in genomic research**, Volume 3), and sequence tag site (STS) maps (*see* Article 13, **YAC-STS content mapping**, Volume 3). For an overview of genome mapping, see Article 9, **Genome mapping overview**, Volume 3. The specific details of fingerprint map construction will be discussed here, beginning with a description of how this approach to physical mapping evolved.

2. Origins of fingerprint mapping

2.1. Molecular tools derived from basic research

Fingerprint mapping is the determination of the relative positions of restriction endonuclease sites along a DNA molecule. The concept of restriction mapping is therefore by definition contingent on the existence of restriction endonucleases. Thus, a critical step in the evolution of this mapping technique was the discovery, isolation and characterization of these enzymes. Evidence of the existence of restriction enzymes was first observed in the early 1950s through the phenomenon of host-controlled variation, in which the ability of bacterial viruses to reproduce in certain host strains was dependent upon the host in which they had previously reproduced. This mechanism of host specificity in *Escherichia coli* was found to involve both DNA modification and DNA restriction activities. In 1968, a restriction enzyme from *E. coli K*, active against λ DNA, was the first restriction endonuclease to be highly purified and characterized. Purification and characterization of additional restriction enzymes rapidly followed (for early reviews, see Meselson *et al*., 1972; Nathans and Smith, 1975).

The potential of using restriction endonuclease digestion to characterize genomic DNA was first demonstrated in the early 1970s on SV40 DNA and the replicative form of ϕX174. These studies showed that specific viral DNA cleavage products could be generated by endonuclease digestion, that these products could be separated by polyacrylamide gel electrophoresis and individually identified, and that the number and size of the fragments produced could be used to characterize the viral DNA. The rationale behind these DNA cleavage experiments was twofold; (1) specific fragmentation of viral DNA chromosomes could potentially be used to generate small, unique DNA fragments that would be amenable to sequencing and (2) if such specific fragments could be produced, then the potential existed to order them with respect to each other and therefore provide a framework (i.e., a physical map) on which to map the location of specific genetic functions in the viral DNA. Indeed, several known ϕX174 activities had been successfully mapped to specific restriction fragments identified in the initial cleavage study. The DNA fragment patterns derived from restriction endonuclease digestion and electorphoretic separation (i.e., the fingerprint) were additionally found to be sufficiently sensitive and reproducible that they could be used to distinguish between different strains of SV40 (Nathans and Danna, 1972), in what may possibly be the first comparative genome mapping experiment.

The first genome restriction map was generated for the SV40 genome (Danna *et al*., 1973), using partial DNA digestion with restriction endonuclease isolated from *Haemophilus influenzae* and subsequent complete digestion of the partial digest products with two additional restriction endonucleases. This resulted in a circular map composed of the relative positions of the cleavage sites within the DNA molecule. Using similar techniques, restriction maps for the genomes of a number of other small DNA viruses were also constructed, including those of the polyoma virus, λ, øX174, and adenovirus. A simple method for fragment separation on agarose gels and visualization using ethidium bromide was also developed during this time (Sharp *et al*., 1973). Restriction mapping of DNA molecules became

a standard method for the direct characterization of small DNA chromosomes. The fundamental reagents and techniques required for fingerprint mapping had thus been established, using in part molecular tools that had been developed as a result of unrelated research into the mechanisms underlying bacterial host–pathogen interactions.

2.2. From viruses to humans: fingerprinting large genomes

The large size of bacterial and eukaryotic chromosomes, and the number and size of restriction digest fragments generated from these larger DNA molecules, made direct application of the restriction mapping techniques developed for the smaller viral DNA genomes problematic. Two primary technological advances provided the means to fingerprint map these large DNA molecules; the development of pulsed-field gel electrophoresis (PFGE) (Schwartz and Cantor, 1984) for the separation of large DNA fragments, and the development of recombinant DNA technology (Jackson *et al*., 1972; Cohen *et al*., 1973) to reduce large segments of genomic DNA into a number of smaller, more easily manipulated cloned genomic fragments.

These technologies led to the development of two approaches to fingerprint map large regions of DNA. In one method, described as a "top-down" or landmark mapping approach, intact genomic DNA (i.e., a whole chromosome) was digested with enzymes that cut rarely in the genome, generating large DNA fragments that were then separated by size on agarose gels by PFGE. These fragments were typically mapped relative to each other by hybridization of DNA probes, such as gene-based sequences or probes specific for restriction fragment ends. Because these restriction endonuclease recognition sites occur infrequently within the DNA sequence, this fingerprinting method generates a long-range but low-resolution "macrorestriction" map of the genome. Restriction maps for the genomes of *E. coli* (Smith *et al*., 1987), *Saccharomyces cerevisiae* (Link and Olson, 1991), and *Schizosaccharomyces pombe* (Fan *et al*., 1989) were generated using this approach. However, since this method requires the isolation of intact chromosomal DNA and the separation and detection of all fragments generated from a restriction digest of this DNA, it was not particularly well suited to the mapping of larger eukaryotic genomes. Additionally, it did not provide reagents that could be readily applied to functional studies or to sequencing strategies.

The second method utilized a "bottom-up" approach, in which many copies of the genome were first fragmented into smaller pieces of DNA, cloned into a bacterial vector and propagated in a suitable bacterial host. These smaller DNA fragments were easily isolated with standard molecular procedures and were amenable to restriction fingerprinting using the same general techniques applied to the viral DNA genomes. Thus, the restriction fingerprinting task was transformed from application to an entire eukaryotic chromosome to that of a series of easily manipulated DNA fragments. This approach was therefore more suited in terms of high-throughput laboratory techniques than the top-down approach, with the additional benefit of providing higher resolution due to the increased density with which restriction sites could be sampled along the DNA. It does, however, represent a

more complex task in terms of assembling a global fingerprint map from the individually fingerprinted DNAs (*see* Article 19, **Restriction fragment fingerprinting software**, Volume 3 and Article 1, **Contig mapping and analysis**, Volume 7). A variety of different strategies employing this basic approach have been used to construct fingerprint maps for eukaryotic genomes. The application of this methodology to the construction of whole-genome fingerprint maps was pioneered in the model organisms *Caenorhabditis elegans* (Coulson *et al*., 1986) and *S. cerevisiae* (Olson *et al*., 1986). The approach was soon employed in the generation of maps for other model organisms, including those of *E. coli* (Kohara *et al*., 1987; Knott *et al*., 1989), *Arabidopsis thaliana* (Hauge *et al*., 1991; Marra *et al*., 1999), and *Drosophila melanogaster* (Siden-Kiamos *et al*., 1990; Hoskins *et al*., 2000). Large regions of human chromosomes were also mapped with a variation of this approach (Carrano *et al*., 1989; Marra *et al*., 1997). Ultimately, as the molecular and computational techniques employed in random-clone fingerprinting and map assembly matured, a clone-based fingerprint map for the entire human genome was achieved (McPherson *et al*., 2001). Fingerprinted clone maps have been constructed for a number of additional mammalian and plant species, including those for the laboratory mouse (Gregory *et al*., 2002), laboratory rat (Krzywinski *et al*., 2004), rice (Tao *et al*., 2001), and maize (Cone *et al*., 2002) genomes. These maps have played, and continue to play, important roles in genome sequencing efforts. For more information on the use of physical maps in genome projects, (*see* Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3, Article 24, **The Human Genome Project**, Volume 3, and Article 8, **Genome maps and their use in sequence assembly**, Volume 7). The remainder of this review will discuss more specifically the current strategies used for fingerprint map construction that have evolved from the pioneering work of the past 30 years.

3. Fundamentals of fingerprint map construction

3.1. *Overview of the fingerprinting process*

The bottom-up approach for constructing fingerprint maps, also referred to as a contig-building strategy, can be divided into a fingerprint data generation (wet-lab) component and a contig construction (computational) component. The process is outlined in Figure 1, and encompasses the following basic steps: (1) construction of a large-insert clone library representing many copies of the genome, (2) DNA purification and restriction endonuclease digestion of a number of clones that together represent redundant coverage of the genome, (3) size separation of the restriction fragments by electrophoresis, (4) restriction fragment detection and size determination, (5) comparison of restriction fragment patterns between all clones to determine similarity, (6) assembly of clones with highly similar restriction fragment patterns into groups of ordered, overlapping clones (referred to as "contigs"), and (7) comparison of fingerprint patterns between clones at contig edges to identify moderate but still significant similarities, indicating joins between individual contigs, and thereby constructing larger contiguous regions of the genome. The end

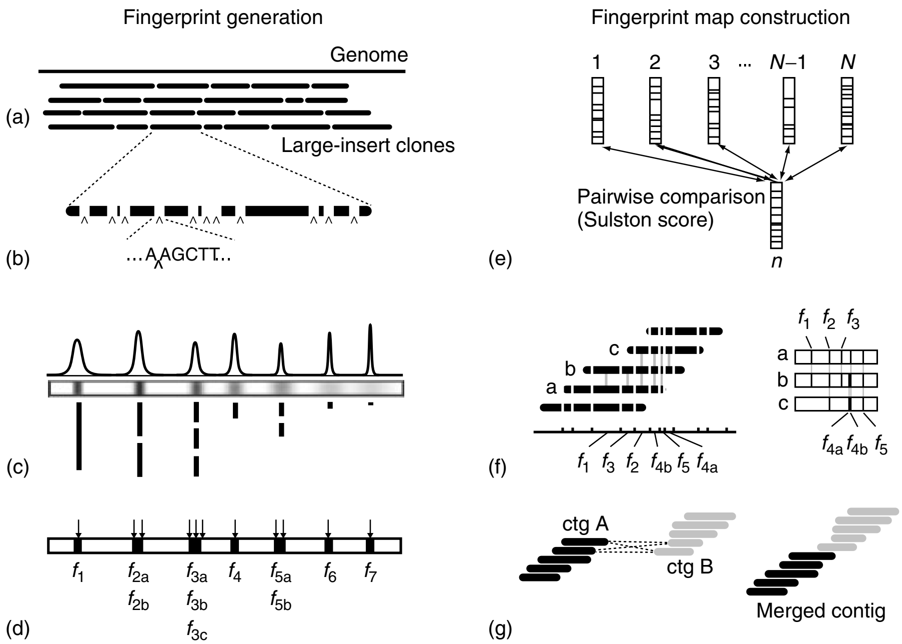


Figure 1 Overview of clone fingerprint data generation and map construction. The two components are shown, fingerprint generation on the left and map construction on the right. Fingerprint generation (a–d): (a) Generation of a large-insert clone library that represents the genome at a high level of redundancy. (b) Clones are sampled randomly from the library and digested with restriction endonuclease, here illustrated with the enzyme *Hin*dIII, with recognition sequence A|AGCTT. (c) Size separation of the restriction fragments by electrophoresis. Stylized data are depicted, with electrophoresis progressing from left to right. Top, chromatogram derived from fluorescently labeled fragments separated on automated sequencer; middle, fragments separated on an agarose gel and visualized with fluorescent DNA dye; bottom, actual restriction fragments. (d) Fragment detection and size determination. Each detected fragment is denoted with f_n where n indicates a particular fragment size. Note that multiple fragments of the same size can be detected on agarose gels. Size determination is made by comparison and interpolation to the fragment pattern of an analytical marker (not shown), composed of DNA fragments of known size. Map construction (e–g): (e) Fingerprint data are stored as an ordered list of fragment sizes and/or mobilities for each clone (depicted here as a size ordered set of fragments). Comparison of fingerprints between all clone pairs is first performed to determine the similarity of fragment patterns. (f) Clones with highly similar fingerprint patterns (depicted on the right) are grouped into ordered sets representing overlapping clones (depicted on the left). Order within the contig is deduced by the progression of ordered fragments across fragment patterns (right, bottom). (g) Map contiguity may be increased by subsequent comparison of fingerprint patterns between clones at contig edges (depicted on the left) to identify moderate but still significant similarities that can join contigs into a single structure (depicted on the right)

result of this process is a physical map represented by sets of ordered, overlapping clones. Depending on the fingerprinting technique used, the map may also reflect the underlying restriction fragment map of the genome. Assembly of the fingerprint data into contigs (steps 5–7, above) is performed with the assistance of a program called Fingerprint Contigs (FPC) (Soderlund *et al*., 1997; Soderlund *et al*., 2000). The details of the computational aspects involved in using clone fingerprint data to assemble contigs is described elsewhere (*see* Article 19, **Restriction fragment fingerprinting software**, Volume 3 and Article 1, **Contig mapping and analysis**, Volume 7) and will not be covered here.

One might expect at the end of this contig-building process that each chromosome will be represented by a single contig; however, in practice, this is not achieved owing to the effect of a number of technical factors that may each contribute to varying degrees, including reduced representation (or lack of representation) in the library of certain genomic regions, lack of genome coverage as a result of the random sampling approach, and unrecognized clone overlap. These are discussed in more detail below.

3.2. Fingerprinting methods

There are two basic clone fingerprinting techniques that have evolved from the early work in *C. elegans* and *S. cerevisiae*, differentiated primarily by the method in which restriction fragments are separated and detected. In one method, restriction fragments are resolved by size on agarose gels and detected by staining with a sensitive DNA dye. In the other method, fragment separation is achieved using polyacrylamide electrophoresis and fragments detected via either radioactive or fluorescent labels.

The agarose gel-based technique (Marra *et al*., 1997; Schein *et al*., 2004) was developed from the method used to construct the *S. cerevisiae* fingerprint map, and was the first method to be widely applied to genome physical map construction. In this method, clone DNA is digested to completion, typically using a single enzyme with a 6-bp recognition site, and the fragments separated by electrophoresis on agarose gels. Analytical marker standards with known fragment sizes are loaded in frequent intervals along the gel to provide a sizing standard. Restriction fragments between approximately 600 and 30 000 bp can be resolved and reliably detected (Fuhrmann *et al*., 2003). Essentially, all restriction fragments generated from each clone (typically on the order of 23 fragments per 100 kb for a single enzyme digest) are detected with this method, providing the potential of deriving an ordered restriction map of each contig. In practice, however, the restriction map is only partially ordered, consisting of a series of fragment "bins", each bin containing one or more fragments. The relative order of the bins is determined, but the order of multiple fragments within a bin is not. The detection of all fragments and their sizes has several advantages: insert sizes can be determined individually for each fingerprinted clone, which can be a useful constraint when including end sequences of the BACs into a genome sequence assembly or when assessing BAC end sequence alignments to a genome sequence assembly; the estimated size of the overlap between any two clones can be calculated directly by summing the size of shared fragments detected, which has practical application when selecting from a contig a tiling set of clones, for example, a minimal tiling set of clones for sequencing or for representation on a genome array (*see* Article 16, **Microarray comparative genome hybridization**, Volume 3); verification of sequence assembly accuracy can be performed by comparison between experimental fingerprint fragments and an electronic digest of the corresponding sequence, which can be particularly useful in detecting collapses in the assembly due to the presence of repetitive sequences.

The polyacrylamide-based fingerprinting techniques currently used were developed from the method used to construct the *C. elegans* fingerprint map. In this

method, fragments are separated by electrophoresis on automated sequencers, either slab-gel based or, more commonly now, capillary based. Only those fragments that fall within a size range of approximately 70–500 bp are detected, and multiplets are not reliably detected. In order to generate a sufficient number of fragments within this size range, the DNA is digested with two or more enzymes. One of the enzymes cuts frequently within the genome and leaves a blunt end. The other enzymes typically have 6-bp recognition sites and leave an overhang. The vast majority of the resulting fragments have one blunt end and one end with an overhang, and detectable fragments represent approximately 15% of the clone DNA. The fragments are labeled at the 6-cutter end with one or more fluorescently labeled dideoxy nucleotides. There are several variations of this approach. In one method, a single 6-cutter enzyme is used, either Type II (Gregory *et al*., 1997) or Type IIs (Ding *et al*., 2001) and a single labeled nucleotide is added. A number of fragments similar to that with the agarose method are detected. In an alternative for the latter approach, the overhang is fully sequenced (Ding *et al*., 2001), linking several bases of sequence information to each detected fragment. In a second method, four different 6-cutter enzymes are used, each labeled with a different fluorescent base (Luo *et al*., 2003), which adds restriction enzyme site information to the fragment size for each detected fragment. This method generates on the order of 78 fragments per 100 kb. One advantage of these methods over the agarose method is increased sizing accuracy, which is typically on the order of 1 bp. The increased number of fragments and added information content of two of these methods also provides the possibility of detecting smaller clone overlaps than with the agarose-based method, which may result in greater map contiguity.

3.3. Factors affecting genome representation in clone libraries

Genomic clone libraries are typically constructed from genomic DNA that has been fragmented by partial restriction endonuclease digestion. The distribution of restriction enzyme recognition sites within a particular genome is therefore an important consideration prior to selection of an enzyme for use in library construction. If there exist regions in a genome where the distance between neighboring recognition sites for a particular enzyme is greater than the maximum fragment size that can be cloned, then these regions will not be represented in a genomic library constructed using that enzyme. If a single restriction endonuclease suitable for partial digestion of the DNA cannot be identified, then construction of two or more libraries, each generated using a different restriction enzyme, can compensate to some extent if the distribution of restriction sites for each enzyme within the genome is complementary (e.g., enzymes with different G/C content in their recognition sequences). Analysis of the fragment size range generated by a complete digestion of the genomic DNA with a candidate enzyme can indicate whether there are regions of the genome that will not be cloned. The size limit of cloneable fragments is of course dependent on the vector selected for library construction. Bacterial artificial chromosome (BAC) vectors (Shizuya *et al*., 1992) are currently the vectors of choice for constructing large-insert genomic libraries for purposes of restriction fingerprint mapping. BAC vectors are capable of cloning segments of foreign DNA of up to 300 kb,

although insert sizes generally range from 100 to 200 kb. The cloned DNA is stably maintained, the rate of chimeric constructs is very low, and the clones are easily manipulated in the laboratory. However, there may be genomic sequences that are not readily cloned or easily propagated within bacterial hosts (e.g., heterochromatic DNA), and this can result in some bias in genome representation in a library.

3.4. Redundant genome sampling in a random-clone approach

In a random-clone fingerprinting strategy, clones from a genomic library are arrayed and sampled at random, with no a priori knowledge of where the clone inserts originated in the genome. Each successive clone that is sampled from a library may represent a completely unique region of the genome or it may overlap in whole or in part with one or more previously sampled clones. The first clones sampled from a library each has a high probability of representing a unique region of the genome, so the rate at which unrepresented regions of the genome is sampled with each additional clone is high. As the number of sampled clones increases, the probability decreases that each additional clone contains previously unsampled, unique DNA, and the rate at which unrepresented regions of the genome are sampled begins to decrease with each additional clone. In order to achieve complete, or nearly complete, representation of the genome in a random-clone approach, it is therefore necessary to sample many more clones (redundant sampling) than would be required to represent the genome if the clones were simply laid end to end.

The level of redundant sampling undertaken for a fingerprint mapping project is a function of the desired level of genome representation, the fraction of shared DNA between clones that is required to detect true clone overlaps (i.e., the sensitivity of overlap detection), and the relative number of contig gaps that is deemed acceptable. Given a truly nonbiased, randomly arrayed clone library, approximately fivefold genome redundancy (5X coverage) is necessary to provide substantially complete representation of a genome (Michiels *et al*., 1987). At fivefold redundancy, on average each nucleotide is represented in five different clones or, put another way, each clone overlaps on average with four other clones. This would roughly equate to 80% shared DNA between adjacent clones in the genome, a relatively substantial overlap. However, this is a calculated average, which means that half of the adjacent clone pairs will overlap by something less than 80%. Thus, for example, if 80% shared DNA is the minimum amount of overlap required to differentiate between true clone overlaps and false-positive overlaps during fingerprint contig assembly, half of the adjacent clone pairs in the genome will fail to satisfy this requirement. This will result in a large number of contig gaps in the assembly due to undetected clone overlaps. To minimize the number of contig gaps, the effective genome coverage in sampled clones must be increased to a depth that ensures that the majority of the genome is represented by adjacent clone pairs that overlap by the required amount.

3.5. Clone overlap detection and contig gaps

For any particular fingerprinting project, the level of redundant clone coverage required is dependent on both the size of the genome and the sensitivity of detection

of clone overlap, the latter of which is based on fingerprint similarity and is a function of clone size and fingerprinting technique. Clone overlap is essentially calculated as the relative proportion of common fragments shared between two clone fingerprints. Since larger genomes require more clones to represent them than do smaller genomes, the probability that there are two unrelated clones sharing by chance a certain number of fragments of the same size is also increased. Thus, as the size of the genome increases, the likelihood of detecting false-positive overlaps given a particular requirement for clone similarity also increases. The required amount of calculated overlap between two clones that is accepted as representing true overlap for purposes of contig construction must therefore be increased for large genomes relative to smaller genomes, and this will affect the level of redundant coverage selected. Mathematical descriptions and analyses of the various effects of these factors have been described (Lander and Waterman, 1988; Branscomb *et al*., 1990). For fingerprint maps of mammalian-sized genomes, a number of clones representing 10–15X genome coverage are typically fingerprinted.

References

Branscomb E, Slezak T, Pae R, Galas D, Carrano AV and Waterman M (1990) Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics*, **8**, 351–366.

Carrano AV, de Jong PJ, Branscomb E, Slezak T and Watkins BW (1989) Constructing chromosome- and region-specific cosmid maps of the human genome. *Genome*, **31**, 1059–1065.

Cohen SN, Chang AC, Boyer HW and Helling RB (1973) Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3240–3244.

Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG, *et al*. (2002) Genetic, physical, and informatics resources for maize. On the road to an integrated map. *Plant Physiology*, **130**, 1598–1605.

Coulson AR, Sulston J, Brenner S and Karn J (1986) Towards a physical map of the genome of the nematode Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7821–7825.

Danna KJ, Sack GH Jr and Nathans D (1973) Studies of simian virus 40 DNA. VII. A cleavage map of the SV40 genome. *Journal of Molecular Biology*, **78**, 363–376.

Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, Lam JY, Kim YM and Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics*, **74**, 142–154.

Fan JB, Chikashige Y, Smith CL, Niwa O, Yanagida M and Cantor CR (1989) Construction of a Not I restriction map of the fission yeast Schizosaccharomyces pombe genome. *Nucleic Acids Research*, **17**, 2801–2818.

Fuhrmann DR, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, Chinwalla A, Hillier LW, Waterston RH, McPherson JD, *et al*. (2003) Software for automated analysis of DNA fingerprinting gels. *Genome Research*, **13**, 940–953.

Gregory SG, Howell GR and Bentley DR (1997) Genome mapping by fluorescent fingerprinting. *Genome Research*, **7**, 1162–1168.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al*. (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.

Hauge BM, Hanley S, Giraudat J and Goodman HM (1991) Mapping the Arabidopsis genome. *Symposia of the Society for Experimental Biology*, **45**, 45–56.

Hoskins RA, Nelson CR, Berman BP, Laverty TR, George RA, Ciesiolka L, Naeemuddin M, Arenson AD, Durbin J, David RG, *et al*. (2000) A BAC-based physical map of the major autosomes of Drosophila melanogaster. *Science*, **287**, 2271–2274.

Jackson DA, Symons RH and Berg P (1972) Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 2904–2009.

Knott V, Blake DJ and Brownlee GG (1989) Completion of the detailed restriction map of the E. coli genome by the isolation of overlapping cosmid clones. *Nucleic Acids Research*, **17**, 5901–5912.

Kohara Y, Akiyama K and Isono K (1987) The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, **50**, 495–508.

Krzywinski M, Wallis J, Gosele C, Bosdet I, Chiu R, Graves T, Hummel O, Layman D, Mathewson C, Wye N, *et al*. (2004) Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. *Genome Research*, **14**, 766–779.

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Link AJ and Olson MV (1991) Physical map of the Saccharomyces cerevisiae genome at 110-kilobase resolution. *Genetics*, **127**, 681–698.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD and Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.

Marra M, Kucaba T, Sekhon M, Hillier L, Martienssen R, Chinwalla A, Crockett J, Fedele J, Grover H, Gund C, *et al*. (1999) A map for sequence analysis of the Arabidopsis thaliana genome. *Nature Genetics*, **22**, 265–270.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**, 1072–1084.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Meselson M, Yuan R and Heywood J (1972) Restriction and modification of DNA. *Annual Review of Biochemistry*, **41**, 447–466.

Michiels F, Craig AG, Zehetner G, Smith GP and Lehrach H (1987) Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries. *Computer Applications in the Biosciences*, **3**, 203–210.

Nathans D and Danna KJ (1972) Studies of SV40 DNA. 3. Differences in DNA from various strains of SV40. *Journal of Molecular Biology*, **64**, 515–518.

Nathans D and Smith HO (1975) Restriction endonucleases in the analysis and restructuring of dna molecules. *Annual Review of Biochemistry*, **44**, 273–293.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R and Frank T (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7826–7830.

Schein J, Kucaba T, Sekhon M, Smailus D, Waterston R and Marra M (2004) In *Methods in Molecular Biology*, Vol. 255, Zhao S and Stodolsky M (Eds.), Humana Press: Totowa, pp. 143–156.

Schwartz DC and Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, **37**, 67–75.

Sharp PA, Sugden B and Sambrook J (1973) Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose–ethidium bromide electrophoresis. *Biochemistry*, **12**, 3055–3063.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli

using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8794–8797.

Siden-Kiamos I, Saunders RD, Spanos L, Majerus T, Treanear J, Savakis C, Louis C, Glover DM, Ashburner M and Kafatos FC (1990) Towards a physical map of the Drosophila melanogaster genome: mapping of cosmid clones within defined genomic divisions. *Nucleic Acids Research*, **18**, 6261–6270.

Smith CL, Econome JG, Schutt A, Klco S and Cantor CR (1987) A physical map of the Escherichia coli K12 genome. *Science*, **236**, 1448–1453.

Soderlund C, Humphray S, Dunham A and French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Research*, **10**, 1772–1787.

Soderlund C, Longden I and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences*, **13**, 523–535.

Tao Q, Chang YL, Wang J, Chen H, Islam-Faridi MN, Scheuring C, Wang B, Stelly DM and Zhang HB (2001) Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. *Genetics*, **158**, 1711–1724.

Short Specialist Review

Restriction fragment fingerprinting software

Carol A. Soderlund
University of Arizona, Tucson, AZ, USA

1. Introduction

A physical map provides an ordering of clones, markers, or both. A physical map may be built using marker-clone associations (*see* Article 13, **YAC-STS content mapping**, Volume 3), where the markers are ordered such that they are contiguous for each clone (e.g., Alizadeh *et al*., 1995; Soderlund and Dunham, 1995). Alternatively, a physical map can be built using restriction fragment fingerprinting. In this case, a clone is digested with one or more restriction enzymes and the resulting fragments are measured. Two clones may overlap if they have a sufficient number of similar fragments. Overlapping clones are arranged into contigs to position the clones relative to each other. Whole-genome fingerprinting was first performed in the late 1980s (Coulson *et al*., 1986; Olson *et al*., 1986). Techniques for agarose-based fingerprinting have been greatly improved in order to reduce the amount of error (Marra *et al*., 1997). An alternative fingerprinting method called *HICF* (*High Information Content Fingerprinting*, Ding *et al*., 1999, 2001; Luo *et al*., 2003) has recently emerged. The most popular software for assembling fingerprinted clones into contigs is FPC (FingerPrinted Contigs, Soderlund *et al*., 1997), which works with either agarose or HICF. The FPC V7.2 software, executables, tutorial, and web-based tools are available from **http://www.genome.arizona.edu/software/fpc**.

2. The FPC software

FPC takes as input files of clones, where each clone is represented by a set of restriction fragments (often referred to as *bands*). It compares all pairs of clones, counts the number of shared bands, and computes the Sulston score (Sulston *et al*., 1988), which is the probability that the shared bands are just a coincidence. The user sets a cutoff and all clone pairs that have a Sulston score below the cutoff are considered overlapping. The assembly algorithm clusters clones such that each clone in a contig has a good overlap with at least one other clone in the contig. It then orders the clones by building a consensus band (CB) map, which is an approximation of the way the bands are ordered along the underlying genome. The clones are aligned to the CB map to give them an approximate position.

The measurement of the bands is not exact; so to compensate for this, the user supplies a tolerance to be used by FPC. If two bands are of the same value within plus/minus of the tolerance, they are considered to represent a fragment of the same size. False positive (F+) and false negative (F−) bands can cause F+ and F− clone overlaps; therefore, the accuracy of the band measurement is very important. It also affects the positioning of the clones; the more error in the data, the more imprecise the clone coordinates (Soderlund *et al*., 2000).

The user-supplied cutoff must be set to reduce F+ and F− overlaps. F+ overlaps result in chimeric contigs, which can generally be detected by an abundance of Q (Questionable) clones, where a Q clone is one in which the ordering routine cannot align 50% or more of the bands to the CB map. F− overlaps cause the clones to assemble into many contigs. For example, given a cutoff that results in 70% overlap between clones (which is typical for agarose-based fingerprints), a genome size of 2400 Mb, clones of size 150 000, and a 17x coverage, the clones will assemble into 1574 contigs if the clones are evenly distributed (Lander and Waterman, 1988). Since some regions are not cloneable and the coverage of clones is not evenly distributed, the number of contigs will be much greater than 1574.

The main FPC automatic functions are: (1) Build contigs, (2) IBC (Incremental Build Contigs) adds new clones to existing contigs and merges contigs, (3) DQer reassembles contigs with over a given number of Q clones using a more stringent cutoff, which reduces F+ overlaps, and (4) End Merger compares clones at the end of contigs using a less stringent cutoff and automatically joins contigs (V7.2 only), which reduces F− overlaps. As these functions do not fix all F+ and F− overlaps, FPC also contains many interactive queries and edit functions so that the user can manually fix the remaining problems (Engler and Soderlund, 2002).

2.1. *Using markers and anchors in FPC*

Fingerprints can be assembled to order the clones relative to each other, but do not order contigs or position them on the chromosome. Genetic markers or radiation hybrid markers (*see* Article 14, **The construction and use of radiation hybrid maps in genomic research**, Volume 3) have order and location on the chromosomes. If these markers have been hybridized to fingerprinted clones, the data can be entered into FPC and used to anchor contigs to chromosomes. Unanchored markers, such as many of the ESTs, are often hybridized against the clones. These marker-clone associations can be entered into FPC, which gives the markers an approximate ordering. The presence of markers in the FPC map is also important for verifying fingerprint data and can be used in conjunction with the fingerprints for assembly. The contig display (see Figure 1) provides a versatile way of viewing the clones, markers, and anchors.

When BESs (BAC end sequences) or sequenced clones (draft or finished) are associated with clones in the map, additional sequenced markers can be added electronically. This is done using the FPC function BSS (Blast Some Sequence), which takes a file of markers, compares them against the sequences associated with FPC clones using BLAST (Altschul *et al*., 1997), megaBLAST (Zhang *et al*., 2000), or BLAT (Kent, 2002). The hits can be added to the FPC map as electronic markers.

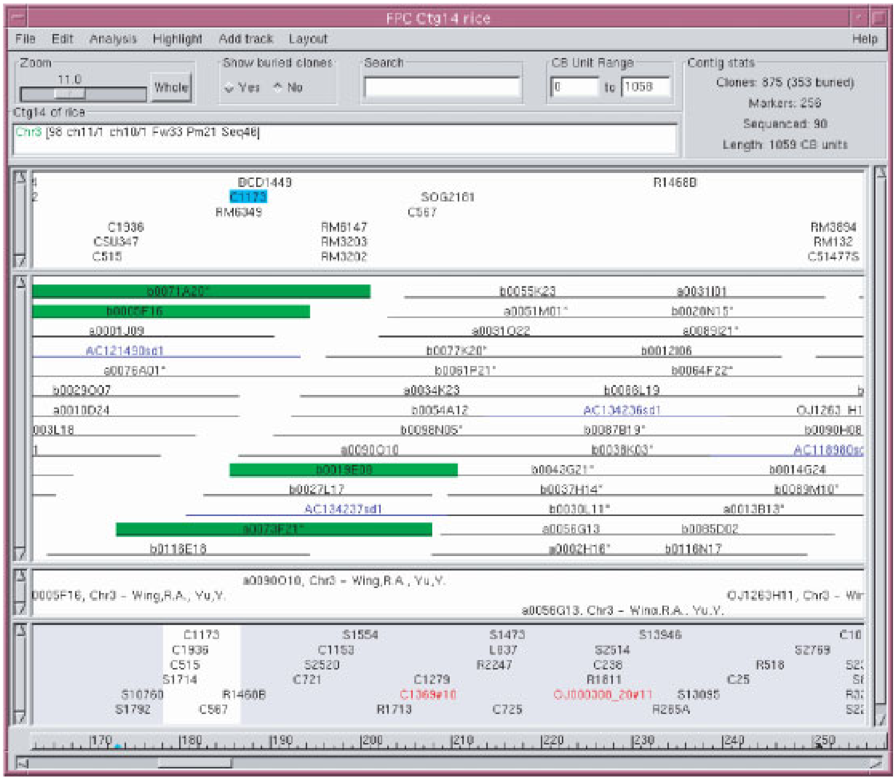


Figure 1 Each of the four regions with a scroll bar on the left is referred to as a *track*. The first track shows the markers. Selecting a marker highlights the clones that it is contained in, as illustrated by marker C1173. The second track shows the clones. The blue clones starting with “A” are sequenced clones from Genbank that have been digested in silico using FSD (FPC Simulated Digest, Engler *et al*., 2003). The third track shows remarks associated with clones or markers. The remarks shown here are attached to the simulated digest clones. The lowest track shows all anchors, which are markers that have a chromosome position. Anchors shown in red disagree with the majority of anchors as to the chromosome assignment. The chromosome assignment is shown above the first track and has been assigned by an FPC function based on majority rules

2.2. *Sequencing*

FPC is used for selecting clones for sequencing (e.g., The International Human Genome Mapping Consortium, 2001). Until recently, this has been performed interactively with FPC tools. A recent release provides a routine that automatically selects an MTP (Minimal Tiling Path, Engler *et al*., 2003), using sequence similarity or fingerprint overlap. When a draft sequence hits two BESs of clones that are near each other in FPC, this dual information provides a reliable overlap known in bases (e.g., Chen *et al*., 2004). For finding overlaps based on fingerprints, the algorithm looks for overlapping clones that are confirmed by two flanking clones and one spanner. An MTP is selected from the overlapping pairs using Dikstra’s shortest path algorithm (Dijkstra, 1959), giving precedence to sequence-based overlaps.

2.3. Agarose versus HICF

A commonly used implementation of agarose-based fingerprinting uses one 6-base enzyme and produces fragments with an average size of 4096 bases, which results in approximately 30–50 bands per clone. Typically, the program Image (Sulston *et al*., 1989) is used to determine the migration rate of the fragments and the corresponding sizes of each band. The accumulative size of fragments is used as the approximate size of the clone. A bottleneck with this method is the human time spent in interactively calling the bands in Image; this problem has recently been resolved with BandLeader (Fuhrmann *et al*., 2003).

HICF uses multiple enzymes and detects the terminal base of each fragment. Hence, two bands are considered the same if they have the same size and terminal base pair. The bands are run on a sequencing machine so that we have high-precision measurements of the bands. The bands sizes range from 50 to 500, and clones typically have over 100 bands; note that the bands only cover a subset of the clone, so they cannot be used to calculate the approximate size of the clone. Though FPC does not take base information as input, Ding *et al*. (1999) developed a simple scheme to encode the base in the fragment size.

Acknowledgments

This work was begun at the Sanger Centre, Hinxton, England. Continued work has been funded by grants USDA-IFAFS #11180 and NSF #0213764. Fred Engler, James Hatfield, William Nelson, Ian Longden, and Steven Ness have made significant contributions to the FPC software and documentation.

References

Alizadeh F, Karp RM, Weisser DK and Zweig G (1995) Physical mapping of chromosomes using unique probes. *Journal of Computational Biology*, **2**, 159–184.

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W and Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Chen R, Sodergren E, Weinstock G and Gibbs R (2004) Dynamic building of a BAC clone tiling path for the rat genome sequencing project. *Genome Research*, **14**, 679–684.

Coulson A, Sulston J, Brenner S and Karn J (1986) Towards a physical map of the genome of the nematode *C. elegans*. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7821–7825.

Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

Ding Y, Johnson M, Chen W, Wong D, Chen Y-J, Benson S, Lam J, Kim Y-M and Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics*, **74**, 142–154.

Ding Y, Johnson M, Colayco R, Chen Y, Melnyk J, Schmitt H and Shizuya H (1999) Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescent-labeled fingerprinting. *Genomics*, **56**, 237–246.

Engler F, Hatfield J, Nelson W and Soderlund C (2003) Locating sequence on FPC maps and selecting a minimal tiling path. *Genome Research*, **13**, 2152–2163.

Engler F and Soderlund C (2002) Software for physical maps. In *Genomic Mapping and Sequencing*, Dunham I (Ed.), Horizon Press, Genome Technology series: Norfolk, pp. 201–236.

Fuhrmann D, Krzywinski M, Chiu R, Saeedi P, Schein J, Bosdet I, Chinwalla A, Hillier L, Waterston R, McPherson J, *et al*. (2003) Software for automated analysis of DNA fingerprinting gels. *Genome Research*, **13**, 940–953.

Kent J (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Lander E and Waterman M (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Luo M-C, Thomas C, You F, Hsiao J, Shu O, Buell C, Malandro M, McGuire P, Anderson O and Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.

Marra M, Kucaba T, Dietrich N, Green E, Brownstein B, Wilson R, McDonald K, Hillier L, McPherson J and Waterston R (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**, 1072–1084.

Olson M, Dutchik J, Graham M, Brodeur G, Helms C, Frank M, MacCollin M, Scheinman R and Frank T (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7826–7830.

Soderlund C and Dunham I (1995) SAM: a system for iteratively building marker maps. *Computer Applications in the Biosciences*, **11**, 645–655.

Soderlund C, Humphrey S, Dunhum A and French L (2000) Contigs built with fingerprints, markers and FPC V4.7. *Genome Research*, **10**, 1772–1787.

Soderlund C, Longden I and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences*, **13**, 523–535.

Sulston J, Mallett F, Durbin R and Horsnell T (1989) Image analysis of restriction enzyme fingerprint autoradiograms. *Computer Applications in the Biosciences*, **5**, 101–132.

Sulston J, Mallet F, Staden R, Durbin R, Horsnell T and Coulson A (1988) Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biosciences*, **4**, 125–132.

The International Human Genome Mapping Consortium (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Zhang Z, Schwartz S, Wagner L and Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Synteny mapping

Simon G. Gregory
Duke University Medical Center, Durham, NC, USA

Comparisons between genomes reveal homologous sequences that reflect their common evolutionary origin and subsequent conservation. Segments of DNA that have function are more likely to retain their sequence than nonfunctional segments, as they are under the constraints of natural selection during evolution. Therefore, DNA segments that are conserved between species are more likely to encode similar function. Sequence comparisons between species provide information on gene structures and may reveal regulatory elements. Experience has shown that such comparisons benefit from the use of sequences from a variety of species representing a range of evolutionary divergence. Sequence conservation between species, within genic and nongenic regions, can be utilized for the construction of physical maps. These clone-based maps can underpin the generation of genome-wide sequence, provide regional coverage for directed sequencing efforts, or provide resources for genomic interrogation, for example, using fluorescence in situ hybridization (FISH), comparative genomic hybridization (CGH), or array CGH.

Similarity between genomes is evident at the level of long-range sequence organization where the order of multiple genes on a single chromosome is conserved, or where the chromosomal location of multiple genes, but not necessarily their precise order, is conserved (Nadeau and Taylor, 1984; DeBry and Seldin, 1996; Nadeau and Sankoff, 1998). In general, the degree of similarity at all levels is higher between species that are more closely related on an evolutionary scale, that is, diverged more recently from a common ancestor. Ultimately, comparison of the finished reference sequence of each organism is required to detect every conserved segment, and from this to deduce all the chromosome rearrangements (translocations, inversions, duplications, deletions, and gene conversion events) that have occurred between species. The ability to align the different genome maps over their entire length simultaneously defines the syntenic relationship between them at a new level of resolution and accelerates the process of sequence generation and other biological studies.

The recent revolution in large-scale genomic analysis has already yielded near-complete DNA sequences of a diverse range of organisms, including bacteria, yeast, worm, fly, dog, mouse, and man (Fleischmann *et al*., 1995; Churcher *et al*., 1997; The yeast genome directory, 1997; The C. elegans Sequencing Consortium, 1998; Adams *et al*., 2000; Lander *et al*., 2001; Venter *et al*., 2001; Waterston *et al*., 2002; Kirkness *et al*., 2003). Assembly of each large genome sequence to

date has been underpinned by production of a comprehensive map of overlapping large-insert bacterial clones (e.g., cosmids; Collins and Hohn, 1978) or bacterial artificial chromosome (BAC) clones (Shizuya *et al*., 1992; Coulson *et al*., 1986; Olson *et al*., 1986; Bentley *et al*., 2001; McPherson *et al*., 2001; Gregory *et al*., 2002) for sequencing, and in some cases also for integration with whole genome shotgun sequence data (Adams *et al*., 2000; Venter *et al*., 2001). Mapped clones provide invaluable information to identify and help eliminate incorrect assemblies between repetitive sequences, to provide substrates for targeted finishing (e.g., to >99.99% accuracy; Green, 2001; Dunham *et al*., 1999; Waterston and Sulston, 1998), and as a resource for experimental studies such as FISH (du Manoir *et al*., 1993) and metaphase and array-based CGH (Kallioniemi *et al*., 1992; Pinkel *et al*., 1998; Ishkanian *et al*., 2004). The study of other large genomes, particularly those with high levels of repetitive sequence (like that of the mouse), requires physical maps of a similar standard as a prerequisite for the production of finished sequence, either on a genome-wide scale or to provide access to any region of interest, which may be located in the map using landmarks such as known genes or genetic markers. Clones that are used for the assembly of these physical maps permit specific regions to be targeted for further investigation and, in particular, for the determination of the complete and accurate DNA sequence separately from other clones within the physical map. Because the source of the genomic sequence is generated clone by clone, problems encountered with sequence assemblies are similarly restricted to

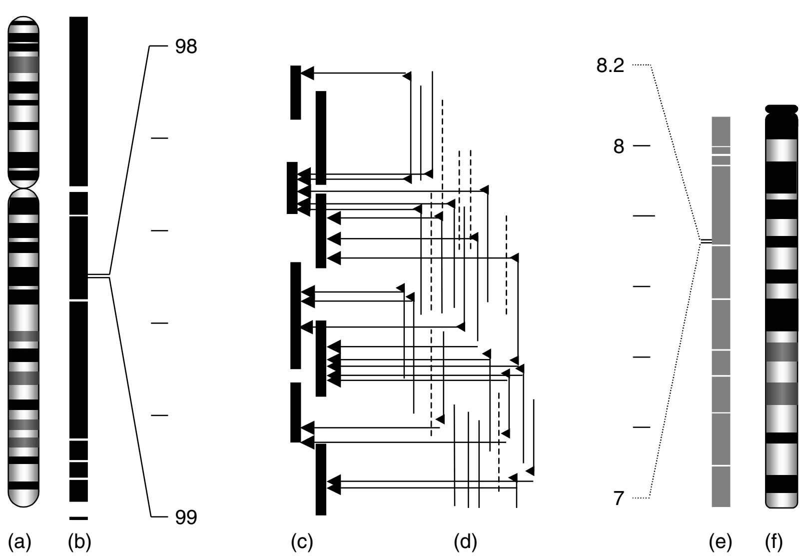


Figure 1 Construction of the physical map of the mouse genome using human genomic sequence as a reference. Finished human sequence from large-insert bacterial clones (c), originating from the physical map (b) of human chromosome 6 (a), provides the template for the alignment of mouse BAC end sequences (d) that had previously been assembled into fingerprint contigs. Contig assembly using the described strategy resulted in rapid assembly of sequence-ready contig coverage (e) of the mouse genome, including mouse chromosome 4 (f)

individual clones, greatly reducing the complexity of resolution of the problem compared to whole genome sequence assemblies.

The similarity in sequence organization between two genomes provides the opportunity for a reference genome, such as the finished sequence of the human genome, to be used as a framework to assemble the physical map of a second genome, such as the mouse (Gregory *et al*., 2002). The phasic construction of such a physical map of a second genome relies upon the existence of a highly redundant restriction digest database (>10-fold redundancy), the availability of BAC end sequences (BESs), and a genome-wide marker set. Initially, restriction fingerprints of the secondary organism are assembled within a database, such as Finger Printed Contigs (FPC) (Soderlund *et al*., 2000). BESs of the clones contained within these assembled contigs are then aligned to the reference genome, prior to inclusion of independently mapped genomic markers for correct positioning within the secondary organism (Figure 1). The juxtaposition of the clone contigs along the reference genome greatly accelerates the physical map construction process and develops a homology map between the two organisms. The proven success of assembling genome-wide physical maps, the cost of constructing a >10-fold genomic BAC library, and the ease with which genome-wide fingerprint databases can be assembled has led to the construction of several genomic fingerprint databases. While genome-wide fingerprint maps will facilitate the large-scale characterization of many varied species, the construction of small region specific sequence-ready maps will continue to be important for detailed interspecies sequence comparisons (Thomas *et al*., 2002).

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.

Bentley DR, Deloukas P, Dunham A, French L, Gregory SG, Humphray SJ, Mungall AJ, Ross MT, Carter NP, Dunham I, *et al*. (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature*, **409**, 942–943.

Churcher C, Bowman S, Badcock K, Bankier A, Brown D, Chillingworth T, Connor R, Devlin K, Gentles S, Hamlin N, *et al*. (1997) The nucleotide sequence of Saccharomyces cerevisiae chromosome IX. *Nature*, **387**, 84–87.

Collins J and Hohn B (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 4242–4246.

Coulson A, Sulston J, Brenner S and Karn J (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7821–7825.

DeBry RW and Seldin MF (1996) Human/mouse homology relationships. *Genomics*, **33**, 337–351.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al*. (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.

du Manoir S, Speicher MR, Joos S, Schrock E, Popp S, Dohner H, Kovacs G, Robert-Nicoud M, Lichter P and Cremer T (1993) Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Human Genetics*, **90**, 590–610.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

Green ED (2001) Strategies for the systematic sequencing of complex genomes. *Nature Reviews. Genetics*, **2**, 573–583.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al*. (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, *et al*. (2003) The dog genome: survey sequencing and comparative analysis. *Science*, **301**, 1898–1903.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Nadeau JH and Taylor BA (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 814–818.

Nadeau JH and Sankoff D (1998) Counting on comparative maps. *Trends in Genetics*, **14**, 495–501.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R and Frank T (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7826–7830.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Shizuya H, Birren B, Kim U, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8794–8797.

Soderlund C, Humphray S, Dunham A and French L (2000) Contigs Built with Fingerprints, Markers, and FPC V4.7. *Genome Research*, **10**, 1772–1787.

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012–2018.

The yeast genome directory (1997) *Nature*, **387**, 5.

Thomas JW, Prasad AB, Summers TJ, Lee-Lin SQ, Maduro VV, Idol JR, Ryan JF, Thomas PJ, McDowell JC and Green ED (2002) Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Research*, **12**, 1277–1285.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Waterston R and Sulston JE (1998) The Human Genome Project: reaching the finish line. *Science*, **282**, 53–54.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Short Specialist Review

Hitchhiking mapping

Christian Schlötterer
Veterinärmedizinische Universität Wien, Wien, Austria

1. The principle of hitchhiking mapping

Hitchhiking mapping is one approach toward the identification and characterization of genes with a beneficial effect in a given context (Schlötterer, 2002; Schlötterer, 2003). The underlying principle of hitchhiking mapping is that a beneficial mutation will either be lost or increased in frequency until it becomes fixed in the population. The spread of a beneficial mutation also affects neutral variation linked to the beneficial mutation ("hitchhiking"; Maynard Smith and Haigh, 1974)). As a consequence, the pattern of sequence variation in the affected genomic region differs from neutral expectations. Population genetics has provided a large repertoire of statistical tests for the identification of genomic regions deviating from neutral expectations (Kreitman, 2000; Otto, 2000; *see also* Article 7, **Genetic signatures of natural selection**, Volume 1).

One of the possible consequences of the spread of a beneficial mutation is a reduction in variability. Figure 1 depicts the reduction in variability around a selected site, obtained from an average over 100 independent computer simulations of a selection event at the same site. In this simulation, the target of selection was unambiguously identified as the genomic region with the most pronounced reduction in variability.

2. Different phases of a hitchhiking mapping study

Hitchhiking mapping studies are carried out on a genome-wide scale to identify those parts of the genome that carry a recent beneficial mutation. In the first phase, a large number of loosely linked markers are analyzed. On the basis of this primary screen, a number of loci are identified, which show the most extreme distortion in allele frequency spectrum. Given that a very large number of loci could be tested in such a primary screen, additional testing is required to distinguish false positives from genomic regions subjected to directional selection.

The second phase of hitchhiking mapping focuses on the genomic region flanking one of the candidate regions identified in the primary screen. As linked sites are more strongly correlated after a selective sweep than under a neutral evolution scenario, the pattern of variation at linked genomic regions could be used to verify genomic regions subjected to a recent selective sweep.

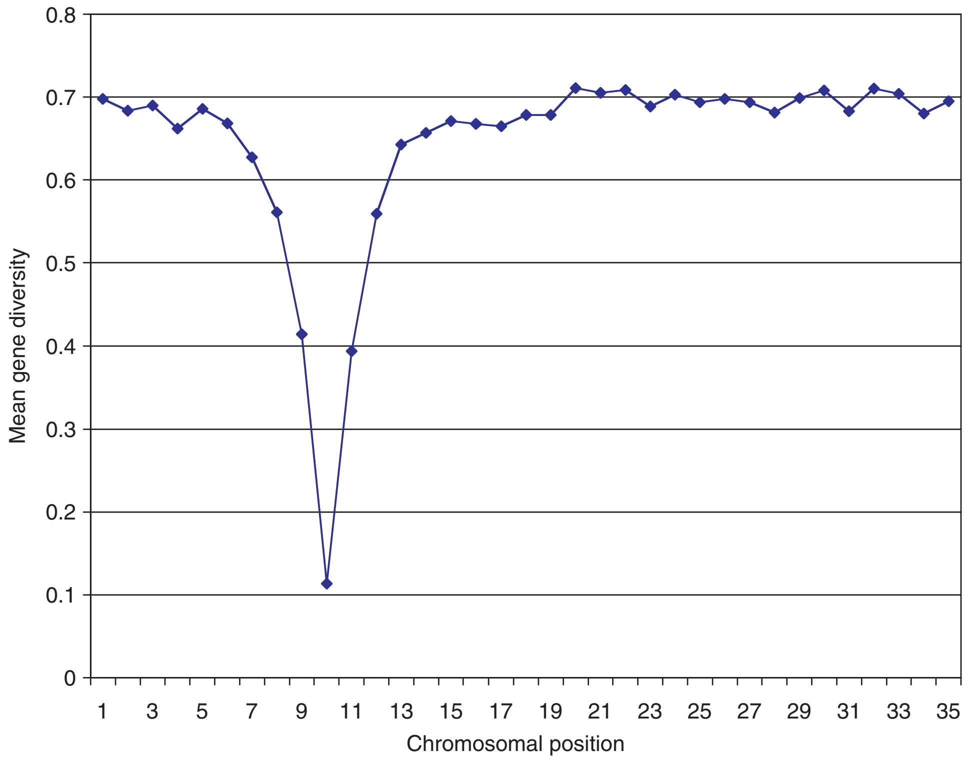


Figure 1 Mean gene diversity determined for 35 evenly spaced microsatellites over 100 simulation runs. For each of the simulations, a selective sweep was assumed to have occurred at the microsatellite No. 10, which shows the most pronounced reduction in variability. Computer simulations were performed with a computer program written by Y. Kim and modified for microsatellites by T. Wiehe. Simulation parameters were: microsatellite spacing = 12 kb, $\tau = 0.001$, $s = 0.001$, $\theta = 5$, $r = 5 \times 10^{-9}$

After the successful verification of a candidate region, the final step of a hitch-hiking mapping study involves a detailed analysis of the genomic region affected by the selective sweep. A comparison of multiple populations with and without a selective sweep could be highly informative for the identification of the molecular changes responsible for the selective sweep.

3. Which marker to use?

The primary screen of many, unlinked markers is greatly facilitated if a highly informative and cost-effective marker is used. Microsatellites are highly polymorphic markers present at a moderate density in most eukaryotic species, making them a good marker choice for first pass genome scans (Schlötterer, 2004). However, SNP (Akey *et al*., 2002) or DNA sequence analysis (Glinka *et al*., 2003) based genome scans have been performed. Microsatellites remain the best choice; the information content of single SNPs is lower than that for a microsatellite locus, and DNA sequencing is more expensive and complicated by the presence of indels.

The second phase requires polymorphism data for several linked genomic regions. Very often, microsatellites are not available at a high enough density. Therefore, DNA sequencing of short (400–800 bp) genomic regions is often the best strategy for the second hitchhiking mapping phase. High-density SNP analysis has also been shown to be informative (Sabeti *et al*., 2002).

The final phase of a hitchhiking mapping project requires a detailed analysis of the polymorphism in the candidate region, which is best achieved by DNA sequencing. Thus, different classes of markers and methods are preferable at the various stages of a hitchhiking mapping study.

4. Potential and limitations of hitchhiking mapping

Recent studies in yeast suggested that even the loss of gene function often does not result in a phenotype that is easily recognized under laboratory conditions (Winzeler *et al*., 1999). Thus, a large fraction of genes cannot be studied by classical genetic approaches. This applies, in particular, to ecologically relevant genes, which, by definition, are highly dependent on the ecological context in which an organism resides. Through the comparison of two groups of individuals adapted to different conditions (e.g., habitat, resistance against diseases, parasites, etc.), hitchhiking mapping provides the opportunity for the identification of genes that recently acquired a mutation, resulting in the phenotypic difference of interest. When the groups are unambiguously defined, hitchhiking mapping offers the advantage that no phenotype needs to be scored in the laboratory. Rather, natural selection has recognized the advantage of the beneficial mutation, which results in the typical molecular signature of a selective sweep. Therefore, hitchhiking mapping can identify even mutations with a subtle or environment-dependent phenotype.

One further advantage of hitchhiking mapping is that no experimental genetic crosses are required. Like linkage disequilibrium mapping, hitchhiking mapping builds upon meiotic recombination events that have occurred in natural populations. As a larger number of meiotic recombination events have occurred in natural populations, hitchhiking mapping could result in a higher mapping precision than quantitative trait locus (QTL) studies requiring experimental crosses.

The signature of a selective sweep is gradually lost as new mutations accumulate (Wiehe, 1998). Hence, hitchhiking mapping is limited to beneficial mutations that occurred in the recent past. Markers with a high mutation rate (such as microsatellites) are better suited for more recent selective sweeps than DNA sequence data. Nevertheless, in Drosophila, hitchhiking mapping was successfully applied to the detection of selective sweeps that occurred about 10 000 years (50 000–100 000 generations) ago. Both microsatellites and DNA sequence analysis detected the signature of the same selective sweep (Harr *et al*., 2002).

Probably the most challenging aspect of hitchhiking mapping is the functional verification of the identified alleles. As the phenotypic effects of these alleles are difficult to study, a comparison of putatively functionally diverged alleles is not straightforward. Nevertheless, at least for some of the identified genes, a sensitized background could be used to test the functional impact of naturally occurring alleles.

Acknowledgments

The laboratory of CS is supported through grants from the Fonds zur Förderung der wissenschaftlichen Forschung (FWF), European Union, and an EMBO young investigator award to CS.

References

Akey JM, Zhang G, Zhang K, Jin L and Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Glinka S, Ometto L, Mousset S, Stephan W and De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics*, **165**, 1269–1278.

Harr B, Kauer M and Schlötterer C (2002) Hitchhiking mapping - a population based fine mapping strategy for adaptive mutations in *D. melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12949–12954.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, **1**, 539–559.

Maynard Smith J and Haigh J (1974) The hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Otto SP (2000) Detecting the form of selection from DNA sequence data. *Trends in Genetics*, **16**, 526–529.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Schlötterer C (2002) Towards a molecular characterization of adaptation in local populations. *Current Opinions in Genetics and Development*, **12**, 683–687.

Schlötterer C (2003) Hitchhiking mapping - functional genomics from the population genetics perspective. *Trends in Genetics*, **19**, 32–38.

Schlötterer C (2004) The evolution of molecular markers-just a matter of fashion? *Nature Reviews. Genetics*, **5**, 63–69.

Wiehe T (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites. *Theoretical Population Biology*, **53**, 272–283.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al*. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Short Specialist Review

The Happy mapping approach

Francis Galibert
Génétique et Développement, Rennes, France

1. Introduction

In 1989, while a Ph.D. student at the University of Oxford, Paul Dear invented a new method for genome mapping, named "Happy mapping", which reflected the use of polymerase on minute amounts of DNA in the procedure. This method is basically an in vitro adaptation of the radiation hybrid (RH) method (Cox *et al*., 1990) in which random subsets of a genome of interest are integrated into the nucleus of a carrier cell to produce a panel of 80 or more independent hybrid cell lines.

This method offers several advantages over the RH method. First, it is possible to apply the Happy mapping method to any genome, including plant genomes, as no cell fusion is involved; second, a Happy panel can be produced in a few weeks as compared to several months for an RH panel; third, a Happy panel contains only the DNA of interest, which makes the analysis of markers easier as there is no possibility of interference with the genomic DNA of the carrier cell. Finally, the final computation of the vectors is simplified, and the resulting map is more robust as a result of the higher and more uniform retention value in each microtiter well of the Happy panel as opposed to the hybrid cell panel.

In the Happy method, minute amounts of genomic DNA, corresponding to less than a haploid equivalent of the genome of interest, are placed in the wells of a microtiter plate (Figure 1). Given the mass of a diploid mammalian genome (~5 pg per nucleus), this corresponds to an average of 2 pg of DNA per well. As a consequence of the limited amount of DNA in each well, only a subfraction of the whole genome is present and, as in the case of radiation hybrid cells, markers located close to each other on the genome tend to be found in the same wells of the microtiter plate. As for RH panels, the ability of a Happy panel to link two markers depends upon the size of the DNA fragments in the wells and, as a rule of thumb, the distance between two markers cannot exceed one-third to one-half of the mean size of the DNA fragments for them to be linked. Owing to the difficulty involved in manipulating very large DNA molecules, the Happy method tends to construct dense maps with at least one marker every megabase. To overcome this limitation, the construction of a Happy panel usually starts by incorporating entire cells or nuclei into agarose beads. The DNA is then gently extracted directly from these beads and subjected to pulsed-field gel electrophoresis. Following migration, bands

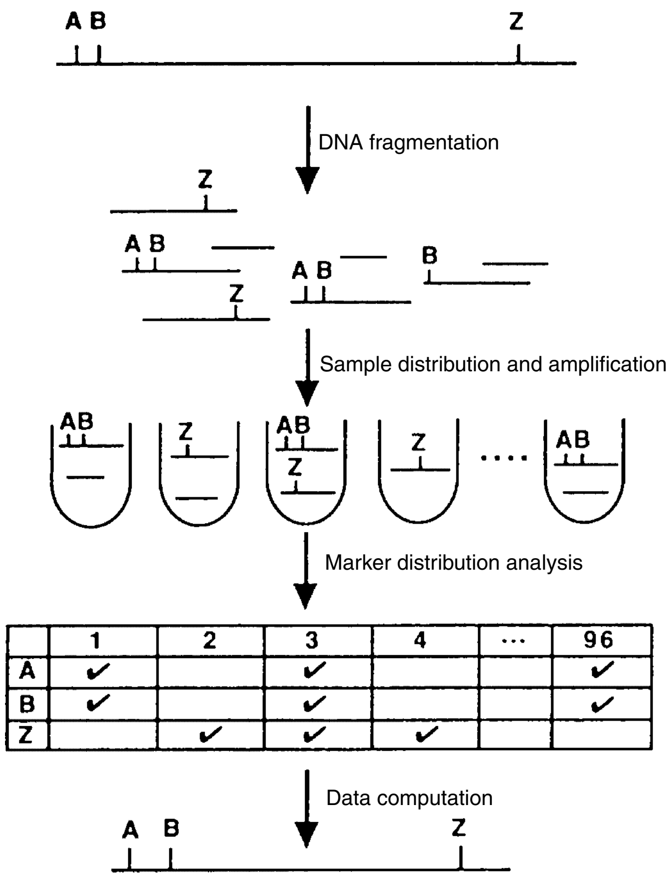


Figure 1 Schematic representation of the Happy mapping method

of the appropriate size – fragments bigger than a few Mb are difficult to obtain – are excised and placed in the wells of a microtiter plate. A series of preliminary experiments are usually necessary to adjust the quantity of DNA placed in each well (between 0.5 and 0.9 haploid genome equivalents). Up to this point, a panel can be readily constructed for any cell type and only requires a bit of practice.

The next stage in map construction is to analyze the distribution of the markers within the panel. This is generally done by PCR as for RH mapping. The immediate advantage of a "Happy panel" as compared to an RH panel is the absence of foreign DNA that could interfere with the DNA of interest during amplification. However, the limited amount of DNA raises a specific problem as the detection of one marker requires two PCR with two pairs of nested primers and a multiplex approach would allow the analysis of a small number of markers only. To overcome this limitation, a whole-genome PCR is required. This poses technical problems that have not yet been satisfactorily resolved. An efficient PCR should satisfy quantitative and qualitative goals. DNA must be amplified 10^7- to 10^8-fold to get enough material for the further mapping of 10^3 or more markers. Even more

importantly, the amplification has to be unbiased, that is, the composition of the DNA after amplification must correspond to its initial composition, such that all the markers present in a well before amplification are present afterward. During the last 10 years, several techniques using different cocktails of oligonucleotide primers and enzymes have been described, but none has met these requirements (Telenius *et al*., 1992; Zhang *et al*., 1992). In these studies, randomness of the product was obtained when a minimum of 30 to 50 DNA molecules was used, but not with just one molecule as in the case of a "Happy panel". Furthermore, the amount of DNA obtained in these studies was limited, prohibiting direct use of the DNA in Happy mapping. Through reamplification of the PCR products, sufficient material could be obtained, but as a sizable fraction (between 30 and 40% depending on the Happy panel) of the markers could not be mapped accurately, PCR-induced representational bias was suspected (De Ponbriand *et al*., 2002).

As a proof of concept of their method, Dear *et al*. (1998) mapped 1001 markers from human chromosome 14. To overcome the PCR bottleneck, they performed inter-Alu PCR with one degenerate primer specific to the repetitive Alu sequence. Although this provided excellent proof of principle, as the resulting map was subsequently shown to match the corresponding human sequence, this sort of map is of little value to gene hunters as these marker sequences are usually nonpolymorphic and not readily usable for synteny comparisons.

It is striking to note that apart from Paul Dear's group and ourselves, no one has ever published a map based on the "Happy technique" despite the potential advantages of this method. Owing to the absence of methods for obtaining sufficient quantities of unbiased amplified haploid DNA, the method has not been used to map any large genomes. Instead, Paul Dear and colleagues have produced genome maps of unicellular eukaryotes of less than 20 Mb (Abrahamsen *et al*., 2004) and of specific chromosomes such as Dictyostelium discoideum chromosome 6 (Konfortov *et al*., 2000). This involved a rather cumbersome two-step amplification approach. The first amplification with a limited yield was done according to Zhang *et al*. (1992) with a random 15-mer primer. In the second amplification step, aliquots of the first amplification product were subjected to multiplex PCR with between 20 and 200 pairs of primers corresponding to the PAC end sequences and additional markers (Piper *et al*., 1998). Finally, using an aliquot of this second amplification, marker-specific PCR was carried out to analyze the distribution of each marker within the Happy panel.

2. Future trends

Efforts to develop a PCR method that results in adequate yields and that can randomly amplify a minute amount of genomic DNA have been moderately successful.

However, the limited DNA yield obtained with the technique described by Zhang *et al*. (1992) has led Paul Dear and collaborators to develop a strategy on the basis of a two-step amplification combined with a multiplex PCR approach to map only a few hundred markers (Piper *et al*., 1998; Konfortov *et al*., 2000). Nevertheless, the potential of the approach and the robustness of the "Happy map" have been well established. Other recent developments, such as those based on DNA microarray

technology or the identification of new enzyme activities, should rekindle interest in the Happy method and lead to the proposal of novel applications.

The development of dedicated microarrays and the possibility of reducing the complexity of large genomes by only amplifying small restriction fragments (Kennedy *et al*., 2003) should lead to new opportunities. It may, for example, be possible to spot PCR fragments corresponding to markers of interest (i.e., corresponding to short restriction fragment produced by a six-nucleotide cutter enzyme) onto functionalized glass and to hybridize them to fluorescently labeled DNA extracted from the different samples of a Happy panel. This approach would detect markers present in each sample instead of asking in which samples each marker is present. If properly developed, it should be possible to use this strategy to derive dense maps of mammalian genomes for which we will shortly have access to sequence information generated by low-pass shotgun sequencing. Other amplification strategies will certainly be developed using the QBeta polymerase sold by several companies and which has been shown to support high yields of random amplification. This enzyme could be used either alone or in combination with the amplification method described by Zhang *et al*. (1992). The attractive possibility of dissecting identified DNA regions or chromosomal bands under a microscope (metaphase spreads) should make it possible to prepare localized markers. It should then be easy to map these limited number of markers on a Happy panel that could be constructed using the currently available method in a few weeks rather than in a few months for an RH panel. These are several possibilities that remain to be investigated now that it has been well established that the Happy mapping method can produce robust maps and could thus be a powerful alternative to the RH method.

Further reading

Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M, Vogel C, Teichmann SA, Ivens A and Dear PH (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. *Genome Research*, **13**, 1787–1799.

Dear PH and Cook PR (1989) Happy mapping: a proposal for linkage mapping the human genome. *Nucleic Acids Research*, **17**, 6795–6807.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng MQ, Liu C, Widmer G, Tzipori S, *et al*. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.

Cox DR, Burmeister M, Price ER, Kim S and Myers RM (1990) Radiation hybrid method: a somatic cell genetic method for constructing high-resolution maps of mammalian genomes. *Science*, **250**, 245–250.

De Ponbriand A, Wang XP, Cavaloc Y, Mattei MG and Galibert F (2002) Synteny comparison between apes and human using finemapping of the genome. *Genomics*, **80**, 395–401.

Dear PH, Bankier AT and Piper MB (1998) A high-resolution metric HAPPY map of human chromosome 14. *Genomics*, **48**, 232–241.

Kennedy GC, Matsuzaki H, Dong SL, Liu WM, Huang J, Liu GY, Xu X, Cao MQ, Chen WW, Zhang J, *et al*. (2003) Large-scale genotyping of complex DNA. *Nature Biotechnology*, **21**, 1233–1237.

Konfortov BA, Cohen HM, Bankier AT and Dear PH (2000) A high-resolution HAPPY map of Dictyostelium discoideum chromosome 6. *Genome Research*, **10**, 1737–1742.

Piper MB, Bankier AT and Dear PH (1998) Construction and characterization of a genomic PAC library of the intestinal parasite *Cryptosporidium parvum*. *Molecular and Biochemical Parasitology*, **95**, 147–151.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA and Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, **13**, 718–725.

Zhang L, Cui X, Schmitt K, Hubert R, Navidi W and Arnheim N (1992) Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 5847–5851.

Short Specialist Review

Digital karyotyping: a powerful tool for cancer gene discovery

Hai Yan and Darell Bigner
Duke University Medical Center, Durham, NC, US

Human beings are genetically diploid. Normally, there are 23 pairs of chromosomes in the nucleus of a somatic cell, and there are two copies of each gene at its specific genomic locus. However, chromosomal aneuploidy, and gene-specific amplification and deletion are commonly observed in human cancer cells. Although random chromosomal or subchromosomal changes can be attributed to genome instability in cancer cells (Lengauer *et al*., 1997), specific recurring gene copy number variations in cancer cells indicate oncogenes and tumor suppressor genes located within the gained and lost genomic regions, respectively. For example, amplification of Her2/neu in breast cancer (Slamon *et al*., 1989) and deletion of PTEN in glioblastoma (Steck *et al*., 1997; Li *et al*., 1997) have been shown to be the driving forces of tumorigenesis.

In the past decades, the quantitative detection of gene amplification and deletion in cancer genomes has been extensively applied to search for oncogenes and tumor suppressor genes. However, the rate of discovering new cancer genes has been unsatisfactory because of the lack of systematic methods that would enable high-resolution scans of the entire genome. The first accurate whole-genome cytogenetic analyses began in 1956 with a method to visualize and count human chromosomes (Tjio and Levan 1956) by histochemically staining the metaphase chromosomes to resolve 400 to 800 distinct chromosomal bands. Spectral karyotyping (Schrock *et al*., 1996) and multiplex-fluorescence in situ hybridization (Speicher *et al*., 1996) are the modern variations of classic karyotyping, which can reveal both numerical and structural aberrations but are limited in mapping resolution to >10 Mbps. Comparative genomic hybridization (CGH) (Kallioniemi *et al*., 1992) is one of the techniques that has been used successfully for measuring genetic dosage changes in the last decade. Limited by the power of resolution, these methods cannot detect genomic alterations at the single-gene level. Moreover, the large chromosomal regions identified by these methods usually contain a large number of genes. Isolating the gene that has a causal role in neoplasia from the many other genes located in the large altered genomic region constitutes a challenge to cancer researchers. Recently, the completion of the reference human genome sequence has made possible the development of several new techniques for measuring gene dosage at the level of single genes. These methods, including array CGH (Pinkel *et al*., 1998; Cai *et al*., 2002; Pollack *et al*., 1999; Solinas-Toldo, 1997) representational oligonucleotide microarray analysis (Lucito *et al*.,

2003; Sebat *et al*., 2004) single nucleotide polymorphism arrays (Lin *et al*., 2004) end-sequence profiling (ESP) (Volik *et al*. 2003) and digital karyotyping (Parrett and Yan, 2005; Wang *et al*., 2002; Shih Ie and Wang, 2005) as compared with conventional cytogenetics methods, provide an unprecedented mapping resolution that allows a precise localization of the amplified and deleted chromosomal regions.

Digital karyotyping accomplishes gene dosage screening of an entire genome by sequencing of representative, small DNA fragments, called tags, which are contained within the genome (Wang *et al*., 2002). These sequence tags (21 bp each), obtained from specific locations in the genome, contain sufficient information to match a tag sequence to its corresponding site in the genome. The tag density can be statistically analyzed to assess the relative genetic content of different loci. In practice, several detailed steps are needed to isolate the genomic tags. First, depending on the resolution desired, the genomic DNA from a tumor sample is cut by an endonuclease mapping enzyme, such as SacI, with a 6-bp recognition sequence, into representative fragments. The fragmented DNA molecules are then ligated to biotinylated linkers and are further digested with a second endonuclease fragmenting enzyme that recognizes even more frequent 4-bp sequences. DNA fragments containing biotinylated linkers are purified from the remaining fragments. New linkers containing a 6-bp site recognized by MmeI, a type IIS restriction endonuclease, are ligated to the fragmented DNA. The DNA fragments are then cleaved by MmeI, which releases 21-bp tags. Isolated tags are self-ligated to form ditags, PCR-amplified, concatenated, and cloned into bacteria. Every bacterial clone represents a homogeneous plasmid that contains a number of different tags. Practically, approximately 5000 to 7000 clones are sequenced from each tumor sample to establish a digital karyotyping library that collects a total of 160000 to 200000 tags. Tags are computationally extracted from sequence data and uniquely matched to a precise, assembled genomic sequence, allowing observed tags to be sequentially ordered along each chromosome. Tag densities are evaluated over moving windows to detect abnormalities in DNA sequence content in tumor samples. Theoretically, the number of experimentally derived genomic tags within any genomic region that contains the same number of virtual tags should be equal in a normal cell genome. In contrast, a cancer genome, with a significantly genomic copy number change in a given genomic region, then displays an abnormal number of observed tags in that genomic locus.

The major advantages of digital karyotyping over other gene dosage measurement methods are its higher resolution and its unbiased gene dosage readout. The number of virtual tags obtained using the mapping enzyme SacI and the fragmenting enzyme NlaIII is 842,202, based on the July 2003 Human Genome Sequence Assembly, giving a theoretical resolution of approximately 4 kb. Therefore, a resolution at single-gene level can be theoretically achieved if a sufficient number of experimental tags can be collected from a digital karyotyping library. However, the current resolution of digital karyotyping is limited by the number of tags that can be economically sequenced. Therefore, although digital karyotyping shows exquisite sensitivity and specificity for detecting gene dosage changes, practically, the analysis of 100000 filtered tags would be expected to reliably detect gene amplification of $\geq$100 kb, homozygous deletions of $\geq$600 kb, or a single gain or loss of regions

of $\geq$4 Mb in a diploid genome. Nevertheless, the resolution of digital karyotyping is the highest of current high-resolution, whole-genome screens and can provide a heretofore unavailable view of the DNA landscape of a cancer genome. Furthermore, digital karyotyping provides an unbiased gene dosage readout because digital karyotyping directly sequences and counts the tags, which are directly proportional to the amount of genetic material present.

The remarkable power of digital karyotyping for discovering cancer-related genes has been demonstrated by recent studies from several research groups. Wang *et al*. (2004) identified a 100 kb amplified region that contained the gene encoding thymidylate synthase (*TYMS*) in two of four metastatic colon cancers treated with 5-fluorouracil (5-FU). The high specificity of gene identification enabled them to perform further studies to show that the *TYMS* gene is amplified in approximately one-fourth of colon cancers treated with 5-FU, but was not amplified in tumors that had not been subjected to 5-FU therapy. These findings suggested that genetic amplification of *TYMS* is a mechanism of attaining 5-FU resistance. Shih Ie *et al*. (2005) identified a 1.8 Mb amplification region at 11q13.5 in three of seven ovarian carcinomas. Combined genetic and transcriptome analyses showed that *Rsf-1* was the only gene that demonstrated consistent overexpression in all of the tumors harboring the 11q13.5 amplification. Furthermore, these authors found that patients with *Rsf-1* amplification or overexpression had a significantly shorter overall survival than those without the amplification. Di *et al*. (2005), and Boon *et al*. (2005) applied digital karyotyping on medulloblastoma cell lines and discovered gene-specific amplifications containing the *OTX2* gene, a novel medulloblastoma oncogene. The characteristics of *OTX2* gene amplification and this gene's exclusive overexpression in medulloblastoma tumors make it a useful target for molecular-based therapy. Although only a small number of digital karyotyping libraries have been generated in these studies, each study has discovered altered genomic regions that are small enough to identify the critical genes immediately.

Digital karyotyping and other gene dosage screens are important for revealing specific types of genetic alterations in cancer genomes, but not all possible alternations can be detected by these screens. For example, epigenetic alterations such as DNA methylation and histone acetylation, in addition to subtle polymorphism expressional differences, all have the potential to influence tumor progression, yet would go undetected in gene dosage screens. To address these issues, novel genomic approaches, such as named methylation-specific digital karyotyping, have recently been developed (Hu *et al*., 2005). Moreover, genome rearrangements are important in cancer and other diseases, but cannot be detected by digital karyotyping, whereas ESP can complement these techniques by providing structural aberration maps.

Two important practical considerations for applying digital karyotyping technology are sample purity and cost. First, because tumor purity and homogeneity are especially critical for the detection of deletions, adequate pathologic review and sampling of specimens is important for selecting clinical samples. Techniques to obtain highly purified tumor material include affinity-purified cells, short-term tissue cultures, subcutaneous xenograft tumor cultures, and laser-capture microdissection. Second, a typical digital karyotyping library costs approximately

$10000 to $20000 because of large-scale tag sequencing. Therefore, currently, digital karyotyping is not suitable for large-scale, high-throughput screening. Nevertheless, although the technology is expensive, the high-resolution of data obtained from examining only a small collection of tumor cases may justify this expense.

It will now be possible, using various genomic approaches, to provide a bird's-eye view of the genomic landscape of cancer cells. Knowledge of the full spectrum of genetic alterations in the cancer genome at the single-gene level of resolution will enable a systems view of tumor biology and will lead to identification of novel prognostic markers and therapeutic targets.

References

Boon K, Eberhart CG and Riggins GJ (2005) Genomic amplification of orthodenticle homologue 2 in medulloblastomas. *Cancer Research*, **65**, 703–707.

Cai WW, Mao JH, Chow CW, Damani S, Balmain A and Bradley A (2002) Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nature Biotechnology*, **20**, 393–396.

Di C, Liao S, Adamson DC, Parrett TJ, Broderick DK, Shi Q, Lengauer C, Cummins JM, Velculescu VE, Fults DW, *et al*. (2005) Identification of OTX2 as a medulloblastoma oncogene whose product can be targeted by all-trans retinoic acid. *Cancer Research*, **65**, 919–924.

Hu M, Yao J, Cai L, Bachman KE, van den Brule F, Velculescu V and Polyak K (2005) Distinct epigenetic changes in the stromal cells of breast cancers. *Nature Genetics*, **37**, 899–905.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Lengauer C, Kinzler KW and Vogelstein B (1997) Genetic instability in colorectal cancers. *Nature*, **386**, 623–627.

Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R, *et al*. (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, **275**, 1943–1947.

Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH and Li C (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research*, **13**, 2291–2305.

Parrett TJ and Yan H (2005) Digital karyotyping technology: exploring the cancer genome. *Expert Review of Molecular Diagnostics*, **5**, 917–925.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Pollack JR, Perou CM Alizadeh AA Eisen MB Pergamenschikov A Williams CF Jeffrey SS Botstein D and Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, **23**, 41–46.

Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, *et al*. (1996) Multicolor spectral karyotyping of human chromosomes. *Science*, **273**, 494–497.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shih Ie M, Sheu JJ, Santillan A, Nakayama K, Yen MJ, Bristow RE, Vang R, Parmigiani G, Kurman RJ, Trope CG, *et al*. (2005) Amplification of a chromatin remodeling gene, Rsf-1/HBXAP, in ovarian carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14004–14009.

Shih Ie M and Wang TL (2005) Apply innovative technologies to explore cancer genome. *Current Opinion in Oncology*, **17**, 33–38.

Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, Keith DE, Levin WJ, Stuart SG, Udove J, Ullrich A, *et al*. (1989) Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707–712.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Speicher MR, Gwyn Ballard S and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, **12**, 368–375.

Steck PA, Pershouse MA, Jasser SA, Yung WK, Lin H, Ligon AH, Langford LA, Baumgard ML, Hattier T, Davis T, *et al*. (1997) Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nature Genetics*, **15**, 356–362.

Tjio J and Levan A (1956) The chromosome number of man. *Hereditas*, **42**, 1–6.

Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo WL, *et al*. (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7696–7701.

Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW and Velculescu VE (2002) Digital karyotyping. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16156–16161.

Wang TL, Diaz LA, Jr Romans K, Bardelli A, Saha S, Galizia G, Choti M, Donehower R, Parmigiani G, Shih Ie M, (2004) Digital karyotyping identifies thymidylate synthase amplification as a mechanism of resistance to 5-fluorouracil in metastatic colorectal cancer patients. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 3089–3094.

Introductory Review

The technology tour de force of the Human Genome Project

Elaine R. Mardis
Washington University School of Medicine, St. Louis, MO, USA

In the short span of a few years, genome sequencing centers around the world were required to undergo a transition that effectively doubled or tripled their normal yearly data production, in order to complete the Human Genome Project (HGP) well ahead of its anticipated completion date (Lander *et al*., 2001). Much of what enabled this increased scale of sequence production was a combination of technological discoveries, made in the preceding years, coupled with revolutionary instrumentation developments that occurred in a just-in-time fashion. What fueled our ability to efficiently sequence the genome, even at this hastened pace, was the creation of a high-resolution physical map (McPherson *et al*., 2001). Both the map and the sequence were necessary elements in generating a finished product of high fidelity and completeness, and both were enabled by a technology tour de force.

Early in the HGP, the importance of first generating a physical map of the human genome as an organizing framework for sequencing was recognized, and became the focus of activity for several groups (Olson *et al*., 1989). Overall, the process of physical map generation can be viewed as a stepwise process, whereby the genome is fragmented into relatively large pieces that can be carried by a host cell. These genome fragments are characterized with respect to sequence content (either restriction enzyme recognition sites or other unique sequences) and then fit back together, much like a puzzle, by virtue of shared sequences present in consistent order. As such, the physical map is a low-resolution construct of the genome, and early maps were typically generated for individual human chromosomes, where a single genome center was responsible for one or more chromosome maps. Many of the early chromosome maps were generated by characterizing chromosome-specific fragments in the yeast artificial chromosome (YAC) vector developed by Maynard Olson's group (Burke *et al*., 1987), and these YAC "clones" were characterized by a wide variety of sequence content approaches, as determined by the group generating the map. Just as many of these maps were approaching completion, however, the bacterial artificial chromosome (BAC) vector system (Shizuya *et al*., 1992) was developed to propagate large genomic fragments in a bacterial host, such as *Escherichia coli*. While BAC clones hold smaller pieces of foreign DNA than YACs (100 kb vs. 500 kb, on average, respectively), BACs are much less likely to delete or rearrange the inserted sequence and are more straightforward to harvest from their host cells (*E. coli*). Once scientists had placed fragments

representing a complete human genome into BAC clones (a "library"), the stage was set to revolutionize the physical map generation process by a clever combination of molecular biology methods and computer software. Mainly, this revolution took place by virtue of a scientific and logistic realization that the human genome needed to have a physical map that was generated with a consistent methodology, using a stable clone "currency" (BACs) that could, once characterized and localized, readily provide a genomic segment for DNA sequence determination. The approach, whole-genome physical mapping by restriction enzyme fingerprinting, was initially devised in our Center to provide a BAC-based physical map of the mustard weed genome (*Arabidopsis thaliana*) (Marra *et al*., 1999). The basic approach requires single restriction enzyme digestion of BAC clones whose total length is 10–15 times the genome size, separation of the resulting fragments on high-resolution agarose gels, staining and imaging the banding patterns, and entering the gel images into specialized software. This software then works to iteratively compare banding patterns for one BAC to all other BACs in the database, joining together those BACs that share a percentage of fragments above a preset threshold (Marra *et al*., 1997; Soderlund *et al*., 1997). The result of this process was the generation of "contigs", or collections of related BACs that recreated a specific region of the genome. From these contigs, BACs that represented the minimal amount of overlap between clones ("tile path") were selected for DNA sequencing. Manual review of data also enabled the joining of contigs to bridge gaps in the map. Although not highly automated, the brute force application of this approach (>300 000 BACs were fingerprinted for the human genome) resulted in a map that served as the main reference and coordination point for the sequencing of the genome (Lander *et al*., 2001).

And what of the DNA sequencing efforts required for the HGP's successful completion? How did we manage this massive increase in scale over a relatively short time period? The answer here lies in the cumulative efforts of many molecular biologists and engineers, both in academic and industrial settings, who provided the instrumentation, biochemistry, and technology to enable increased sequence production. Here, several key developments bear mentioning due to their significant impact. First is the development of cycle sequencing, an offshoot of polymerase chain reaction (PCR), which completely changed the face of DNA sequencing in terms of the efficacy with which reactions could be assembled and incubated. Cycle sequencing effectively is single primer PCR, where all components of the sequencing reaction are combined with the template DNA, and incubated with iterative temperature cycles that denature template molecules, anneal primers, and extend the annealed primers in succession, all without need for human interaction (McBride *et al*., 1989; Craxton, 1993). Prior to its introduction, large amounts of template DNA (~1 μg) were needed for each sequencing reaction, because only a single primer annealing and primer extension step could be done (due to the thermal lability of the enzyme). As such, the number of sequencing fragments was equal to the number of extended primers. Prior to loading, these reactions also needed to have the extended primer fragments melted away from the template strands by high-temperature incubation. Cycle sequencing decreased the input template amount by ~5-fold and eliminated several process steps. A second series of key developments centered around the fact that cycle sequencing reactions were

incubated in programmed thermal cycler instruments. These instruments initially accommodated either single or strip tubes, but ultimately (once developed), 96 well reaction plates with a uniform 8 × 12 tube configuration provided a convenient format. The development of 96 (and later 384) well plates with low thermal deformation, coupled with upstream steps that picked clones for sequencing into 96 tube format boxes (Panussis *et al*., 1996), and DNA preparation methods that could be accomplished in a 96 well format (Marziali *et al*., 1999; Mardis, 1994), paved the way to utilize liquid-handling robots that also were programmed to address 96 wells – these initially were "borrowed" clinical chemistry devices doing simple liquid-transfer steps, but ultimately this coupling of 96/384 well format sample containers and corresponding format liquid handling became integrated into more complicated robots that prepared and sequenced DNA in an assembly line fashion (Hawkins *et al*., 1997). A third significant contribution was provided by two related developments in the enzymology and fluorescent labeling of DNA sequence fragments. Sequencing enzymology was impacted by the creation of a mutant thermostable polymerase by Tabor and Richardson (1995), where a single amino acid change in the enzyme enhanced its binding affinity for modified nucleotides such as the dideoxynucleotide "terminators" used in sequencing. This dramatically reduced the time required for thermal cycling, the amount of expensive ddNTPs required per reaction, and improved the peak height disparities in sequence data. A second improvement, from the laboratory of Mathies (Ju *et al*., 1995a,b), utilized the technique of fluorescence resonance energy transfer (FRET) to label sequencing primers, which greatly enhanced the amount of light produced by each molecule and thus further reduced the amount of DNA needed for a sequencing reaction. Later, incorporation of the FRET-based labeling approach onto the ddNTPs (Rosenblum *et al*., 1997; Lee *et al*., 1997) enabled sequencing to transition from a four-tube reaction for each template (one primer and ddNTP mixture for each nucleotide) to a single-tube reaction where the identity of the fragment (A, C, G, or T) was solely coded by the ddNTP (with its corresponding fluorescent group) incorporated at its 3′-end.

Perhaps the most revolutionary progress in high-throughput sequencing during the HGP was that experienced by fluorescent DNA sequencing instrumentation. Initially introduced in mid-1980 (Smith *et al*., 1986a), these instruments used slab gels to separate the DNA sequencing fragments prior to detection and analysis of the resulting data (Smith *et al*., 1986b). Slab gels inherently limited the scale of operations obtainable for several reasons. First, the time, personnel, space, and logistics required to cast large numbers of polyacrylamide gels were limiting. Second, instrument setup with slab gels, including hand transfer of samples from microtiter plate wells into gel wells, was time consuming and error prone. Third, once the samples had run, the resulting gel image required the careful manual placement of tracking lines onto each lane/sample in order to properly extract the underlying sequence data. Again, this was a tedious, error prone and ultimately rate-limiting step, although software programs were developed to hasten the tracking line placement (Cooper *et al*., 1996). The introduction of glass capillary array–based DNA sequencing instruments largely addressed all of these limitations. Namely, the fragment separation matrix was injected into and expelled from the capillaries automatically, obviating the gel casting and associated steps; the samples were

automatically loaded on at the capillary ends by a process called *electrokinetic injection*, eliminating the sample transfer step; and the capillaries were fixed in space relative to the detector, and once localized, data collection occurred at a defined location for each capillary (Mardis, 1999). Steady improvements over time in terms of reducing the time required for separation and detection, enhancing the separation properties of the matrix that enabled longer read lengths, and increasing the stability of the capillaries, matrix, and buffer have enabled throughputs of up to 24 × 96 samples analyzed in a 24-h period, with almost completely unattended operation. This throughput potential stands in stark contrast to the maximum throughput of 3 × 96 samples per 24 h on the most advanced slab gel instrument available prior to the advent of capillary-based sequencers. As such, this final and often rate-limiting step in the process of sequence data acquisition has been successfully addressed, for now.

This brings us to the present day, anticipating what the future of genome sequencing holds and what challenges we will face. As such, the pursuit of sequencing technology is ongoing, which is as it should be. For example, now that we have completed the human sequence, the call for genome sequence from other organisms is steadily increasing, and we are applying the infrastructure, technology, and methods that were put in place for sequencing the human, toward these new pursuits (Stein *et al*., 2003; Waterston *et al*., 2002). Furthermore, now that the human genome sequence is available as a reference point, it makes sense to sequence additional human genomes as a means of coming to grips with the variation that is found when comparing two individuals, and with that found in a diseased state such as cancer (Ley *et al*., 2003).

References

Burke DT, Carle GF and Olson MV (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, **236**, 806–812.

Cooper ML, Maffitt DR, Parsons JD, Hillier L and States DJ (1996) Lane tracking software for four-color fluorescence-based electrophoretic gel images. *Genome Research*, **6**, 1110–1117.

Craxton M (1993) Cosmid sequencing. *Methods in Molecular Biology*, **23**, 149–167.

Hawkins TL, McKernan KJ, Jacotot LB, MacKenzie JB, Richardson PM and Lander ES (1997) A magnetic attraction to high-throughput genomics. *Science*, **276**, 1887–1889.

Ju J, Kheterpal I, Scherer JR, Ruan C, Fuller CW, Glazer AN and Mathies RA (1995a) Design and synthesis of fluorescence energy transfer dye-labeled primers and their application for DNA sequencing and analysis. *Analytical Biochemistry*, **231**, 131–140.

Ju J, Ruan C, Fuller CW, Glazer AN and Mathies RA (1995b) Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 4347–4351.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lee LG, Spurgeon SL, Heiner CR, Benson SC, Rosenblum BB, Menchen SM, Graham RJ, Constantinescu A, Upadhya KG and Cassel JM (1997) New energy transfer dyes for DNA sequencing. *Nucleic Acids Research*, **25**, 2816–2822.

Ley TJ, Minx PJ, Walter MJ, Ries RE, Sun H, McLellan M, DiPersio JF, Link DC, Tomasson MH, Graubert TA, *et al*. (2003) A pilot study of high-throughput, sequence-based mutational

profiling of primary human acute myeloid leukemia cell genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14275–14280.

Mardis ER (1994) High-throughput detergent extraction of M13 subclones for fluorescent DNA sequencing. *Nucleic Acids Research*, **22**, 2173–2175.

Mardis ER (1999) Capillary electrophoresis platforms for DNA sequence analysis. *Journal of Biomolecular Techniques*, **10**, 137–147.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**, 1072–1084.

Marra M, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterston RH (1999) zA map for sequence analysis of the Arabidopsis thaliana genome. *Nature Genetics*, **22**, 265–270.

Marziali A, Willis TD, Federspiel NA and Davis RW (1999) An automated sample preparation system for large-scale DNA sequencing. *Genome Research*, **9**, 457–462.

McBride LJ, Koepf SM, Gibbs RA, Salser W, Mayrand PE, Hunkapiller MW and Kronick MN (1989) Automated DNA sequencing methods involving polymerase chain reaction. *Clinical Chemistry*, **35**, 2196–2201.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Olson M, Hood L, Cantor C and Botstein D (1989) A common language for physical mapping of the human genome. *Science*, **245**, 1434–1435.

Panussis DA, Stuebe ET, Weinstock LA, Wilson RK and Mardis ER (1996) Automated plaque picking and arraying on a robotic system equipped with a CCD camera and a sampling device using intramedic tubing. *Laboratory Robotics and Automation*, **8**, 195–203.

Rosenblum BB, Lee LG, Spurgeon SL, Khan SH, Menchen SM, Heiner CR and Chen SM (1997) New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Research*, **25**, 4500–4504.

Shizuya H, Birren B, Kim U, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8794–8797.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH and Hood LE (1986a) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB and Hood LE (1986b) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.

Soderlund C, Longden I and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences*, **13**, 523–535.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al*. (2003) The genome sequence of caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biology*, **1**, E45.

Tabor S and Richardson CC (1995) A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6339–6343.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Introductory Review

The Human Genome Project

Tim Hubbard

Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

The Human Genome Project (HGP) has seen many important milestones in its history, but perhaps the most significant was in February 1996 when all the participants (Table 1) agreed that they would do their utmost to ensure that the genome sequence would be freely available to all. They agreed that sequence data would be released without restriction as swiftly as possible and that they would seek no patent protection. By any standards, it was a remarkable agreement. Potentially, large institutions could have patented genes as they went along. Funding agencies agreed with their research leaders that the best way to benefit humankind was to release the sequence swiftly and freely into the public domain. This approach has been vindicated by the massive growth in Internet access to the genome resources (see Section 5), which shows the value the worldwide community places on the sequence.

2. Starting and end points

Three separate proposals were made in the mid-1980s to sequence the human genome (Roberts, 2001). With an estimate of the genome size of 3 billion base pairs (bp) and a cost for sequencing each base of around US$ 10, many researchers fiercely opposed the concept, arguing that financial support would be diverted from other more immediate projects.

The Congress of the United States of America voted in 1988 to support the Human Genome Project, courageously accepting that even if the cost per base sequenced dropped 10-fold, the overall budget would be as much as US$ 3 billion simply for the sequence. The Project was formally announced in October 1990, funded by the US Department of Energy and the Institutes of Health. Around the world, interest in genome projects grew, resulting in the formation of the Human Genome Organization (HUGO), which acted in the early days to discuss priorities and procedures. As the HGP developed, participants joined from the United Kingdom, France, Germany, Japan, and China (Table 1).

From the outset, the HGP had as its goals not only to sequence the human genome but to develop new technologies, to set paradigms using model organisms, to develop mechanisms transfer genome knowledge to the research community, and to consider ethical, legal, and social issues. Most of the scientific goals of

Table 1 Participants in the Human Genome Project

The institutions that form the Human Genome Sequencing Consortium include:	
The Wellcome Trust Sanger Institute,	The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK
Whitehead Institute/MIT Center for Genome Research,	Cambridge, MA, US
Washington University School of Medicine Genome Sequencing Center,	St. Louis, MO, US
Joint Genome Institute, U.S. Department of Energy,	Walnut Creek, CA, US
Baylor College of Medicine Human Genome Sequencing Center,	Department of Molecular and Human Genetics, Houston, TX, US
RIKEN Genomic Sciences Center,	Yokohama-city, Japan
Genoscope and CNRS,	UMR-8030, Evry Cedex, France
Genome Therapeutics Corporation (GTC) Sequencing Center,	Genome Therapeutics Corporation, Waltham, MA, US
Department of Genome Analysis,	Institute of Molecular Biotechnology, Jena, Germany
Beijing Genomics Institute/Human Genome Center,	Institute of Genetics, Chinese Academy of Sciences, Beijing, China
Multimegabase Sequencing Center,	The Institute for Systems Biology, Seattle, WA, US
Stanford Genome Technology Center,	Stanford, CA, US
Stanford Human Genome Center and Department of Genetics,	Stanford University School of Medicine, Stanford, CA, US
University of Washington Genome Center,	Seattle, WA, US
Department of Molecular Biology,	Keio University School of Medicine, Tokyo, Japan
University of Texas Southwestern Medical Center at Dallas[a]	Dallas, TX, US[a]
University of Oklahoma's Advanced Center for Genome Technology,	Dept. of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, US
Max Planck Institute for Molecular Genetics,	Berlin, Germany
Cold Spring Harbor Laboratory,	Lita Annenberg Hazen Genome Center, Cold Spring Harbor, NY, US
Gesellschaft für Biotechnologische Forschung mbH (GBF),	German Research Centre for Biotechnology Braunschweig, Germany
In addition, three institutions played a key role in providing computational support and analysis for the Human Genome Project:	
The National Center for Biotechnology Information at NIH, US	
The European Bioinformatics Institute in Cambridge, UK	
University of California, Santa Cruz, US	

[a]Sequencing center is no longer in operation.
The assembly of the genome sequence across chromosomes was also assisted by scientists at Neomorphic, Inc.

the project were significantly exceeded in terms of accuracy or amount of data collected (Table 2).

3. From 230K to 60 000K

In 1990, the longest single DNA sequence was some 230 000 bp, the genome of the cytomegalovirus. Within 10 years, more than 90% of the human genome would be sequenced and the longest contiguous sequence would have grown more than 100-fold, for human chromosomes 21 and 22. Today, the longest contiguous sequence is in excess of 60 000 000 bp.

Table 2

(a) Goals of the Human Genome Project[a]

Mapping and sequencing the human genome
Mapping and sequencing the genomes of model organisms
Data collection and distribution
Ethical and legal considerations
Research training
Technology development
Technology transfer

(b) Scientific achievements of the Human Genome Project[b]

Area	Goal	Achieved	Date
Genetic map	2–5-cM resolution map (600–1500 markers)	1-cM resolution map (3000 markers)	September 1994
Physical map	30 000 STSs	52 000 STSs	October 1998
DNA sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and cost of finished sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1400 Mb/year at <$0.09 per finished base	November 2002
Human sequence variation	100 000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene identification	Full-length human cDNAs	15 000 full-length human cDNAs	March 2003
Model organisms	Complete genome sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*	Finished genome sequences of *E. coli*, *S. cerevisiae*, *C. elegans, D. melanogaster*, plus whole-genome drafts of several others, including *C. briggsae*, *D. pseudoobscura*, mouse, and rat	April 2003
Functional analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid system for protein–protein interaction	2002

[a] *Source:* NHGRI (1990) Understanding Our Genetic Inheritance. The United states Human genome Project. The First Five Years: Fiscal Years 1991–1995. **http://www.genome.gov/10001477**.
[b] *Source:* NHGRI Human Genome Project Goals (2003) **http://www.genome.gov/11006945**.

This production of high-quality contiguous sequence is built upon a phased mapping and sequencing program established by the HGP during its development. The initial aim of the HGP was to produce by 2005 at least 95% of the gene-containing regions of the human genome at an error rate less than 0.01%, producing long runs of DNA sequence (Table 2).

Why should long runs be important? At that time, most of the sequences generated by labs around the world were shorter and of lower quality. But the human genome sequence was unlikely to be reproduced in full in the foreseeable future

and an archival gold standard was considered essential. Much more important, as biologists got to grips with human genes and variation in human genomes, it became apparent that many genes extended over tens or hundreds of thousands of base pairs, making longer sequence runs much more valuable. Similarly, as it was found that variation in the genome occurs at only about one in 1000 bp, accuracy also became an important criterion.

As the HGP progressed, the demand for accurate sequence was balanced against the growing hunger of the research community for more sequence. In many projects, a lower level of sequence quality is sufficient to allow researchers to isolate and analyze the gene(s) they are interested in. To satisfy this hunger, in 1998, the HGP adopted a plan first mooted in 1995 (Waterston and Sulston, 1998) to produce a draft or sketch of the genome as an intermediate product that could be generated rapidly. While not the finished product, the draft would allow much research to be accelerated and would ensure that the majority of the genome sequence was deposited in the public domain swiftly. At the same time, the HGP remained committed to go on and bring the draft sequence up to gold standard quality. The decision to rapidly create a draft was also partly made out of concerns about commercial plans to create a proprietary draft of the genome and the possible undesirable consequences of restrictions on its use (Sulston and Ferry, 2002).

4. Sequencing the genome

Today's high-throughput DNA sequencing is still based on the biochemical reactions and high-resolution methods for DNA analysis developed by Fred Sanger in the late 1970s (Sanger *et al*., 1977). Sanger's lab also pioneered approaches to apply this method to genome-sized sequence fragments and in 1982 published the first whole-genome shotgun (WGS) (Sanger *et al*., 1982) – a method in which a genome sequence is assembled from overlaps between many short sequence "reads" produced by breaking a genome apart at random. However, for very large genomes, it was considered that the very large number of fragments that would be generated by WGS approach would lead to serious assembly problems (*see* Article 25, **Genome assembly**, Volume 3), as would segmental duplications (*see* Article 26, **Segmental duplications and the human genome**, Volume 3). The HGP adopted a more complex hierarchical sequencing approach for this reason and because it facilitated finishing to high accuracy.

In hierarchical sequencing projects, large clones of DNA (called bacterial artificial chromosomes, or BACs) are first produced by "shotgun" fragmentation of the whole genome, each ∼100–200 kb in size. In the early phases of the HGP, a set of landmarks had been established on the 23 pairs of human chromosomes, and many of the BACs could be mapped to the genome using these landmarks. Through the construction of fingerprint contigs of BACs (*see* Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3) 300 000 BACs could be positioned on the genome through the small number of mapped BACs (Figure 1). From these, a minimal set of BACs was chosen that would cover the majority of the human genome (some regions of all large genomes are refractory to conventional cloning methods and are not represented in libraries of BACs). Each

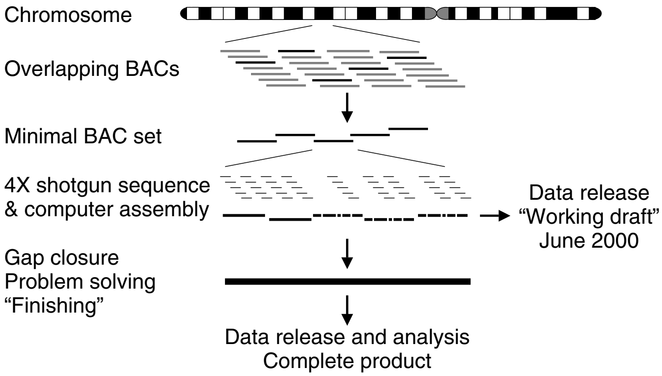


Figure 1 FPC (finger print contigs) are positioned and orientated on human chromosomes via Genethon markers and radiation hybrid (RH) Map. Each FPC is an assembly of BAC clone-based similarities between their restriction digest fingerprints

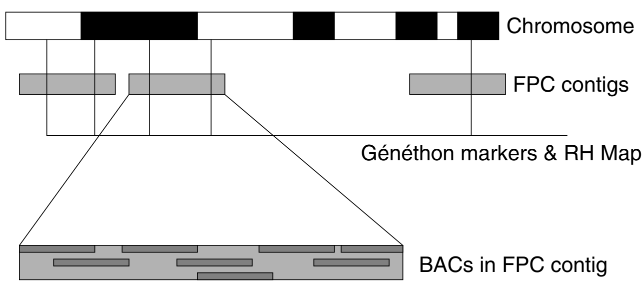


Figure 2 Strategy of HGP was based on selecting a minimal set of overlapping BAC clones to sequence, from libraries covering the genome many times. Each selected clone was then shotgun sequenced to a certain base coverage and then "finished" to close the remaining gaps and resolve problems in a semimanual fashion. A shotgun coverage of 4X means that on average each base in the clone will occur in four different reads

of the minimal set of clones was then subjected to shotgun sequencing (Figure 2). Each was fragmented in turn and subcloned into DNA plasmids that could be used for sequencing. The fragments were typically 2000 bp in length, and about 500 bases of sequence were "read" from each end. First-pass sequencing of 2000–4000 reads per BAC corresponds to reading each base about four times. Assembling these randomly obtained reads generates about 90–95% of the BAC sequence at an accuracy of about 99.9%. This is the draft – the sketch of our genome (International Human Genome Sequencing Consortium, 2001). The majority of the sequencing for the draft was achieved in a little over 12 months, representing an impressive distributed scale up and application of laboratory technology (*see* Article 23, **The technology tour de force of the Human Genome Project**, Volume 3).

Following the announcement of the draft sequence in June 2000, the HGP continued to improve the sequence quality. To fill gaps or to clear up ambiguities, new DNA clones were made using different methods, and sequence was generated

without using clones to fill gaps that appeared to be unclonable. Finishing a genome is a specialist skill and all the participants in the HGP devoted teams to these tasks.

By April 2003, the HGP had surpassed the original goals of the project. The sequence was of much higher quality than the standards set and the amount of the genome represented also exceeded the level of 95% of the gene-containing regions (International Human Genome Sequencing Consortium, 2004). Work continues today to improve the regions that are poorly represented. As the continuity and completeness of each chromosome has been finalized to what appears possible with current technology, each has been analyzed and published. To its credit, the HGP has never claimed to have finished the sequencing of the human genome.

While the human genome sequence has been finished, draft sequences of other vertebrate genomes have been generated using WGS techniques. Recently, the formally private assemblies from the commercial human sequencing project, which championed using WGS techniques for vertebrates (Venter *et al*., 2001), have been released. This has allowed the comparison of a WGS assembly with a clone assembly (International Human Genome Sequencing Consortium, 2004) for the same organism (She *et al*., 2004). The comparison shows that WGS techniques do cause large duplicated regions of genomes to be lost from the assembly with their associated genes (*see* Article 26, **Segmental duplications and the human genome**, Volume 3). It concluded that clone-based approaches are required to resolve such problems. Therefore, while WGS techniques have allowed rapid coverage of new genomes, users must be cautious about overinterpreting differences between genome sequences that may be assembly artifacts.

5. Finding value in our genome

The growth in sequence data in public databases has been exponential since the mid-1980s, requiring continuous development at the database centers to support it. Despite this, the human genome sequencing effort generated an unprecedented amount of sequence in a very short time. It has been a challenge to organize, analyze, and provide access to it, particularly in the face of a huge and growing demand (Figure 3). The HGP has driven the development of two major new public domain Web-based resources Ensembl (Hubbard *et al*., 2002) and UCSC (Kent *et al*., 2002) to organize and provide access to the human genome alongside the existing NCBI resources. The browsers have provided users with web access to an organized representation of the genome sequence, from the time when the genome consisted of around 100 000 fragments at the time of the draft, to the few hundred continuous blocks today. Each allows anyone to download the views, the underlying data, and even the software that makes them work. Each provides links to the experimental data that supports the genome structure and integrates information about sequence variation and disease association. As other vertebrate genomes have been sequenced, they have also provided integrated views of the similarities and differences between them through comparative analysis.

Beyond providing access to the genome sequence, the major focus has been locating the genes in the sequence and annotating their structure. The Ensembl project has been at the forefront, providing consistent gene sets by automatically

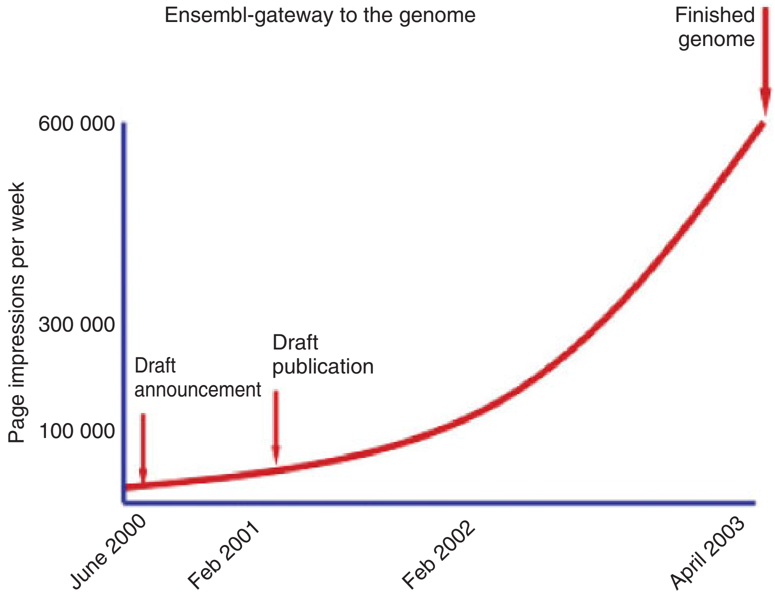


Figure 3 Growth in the use of Ensembl database. Ensembl (**http://www.ensembl.org**) is one of the three major "genome browsers" – Web-based resources that provide assembled information about genomes by creating dynamic web pages from huge supporting databases. The measures shown here are Page Impressions, which is the number of whole pages accessed (and is a more accurate and lower count than "Hits")

annotating the genome on the basis of the alignment of experimental evidence. As the genome has been finished and the collection of high-quality cDNA sequences has increased, a process of curating gene structures has begun to extend and refine this geneset, most notably by the Havana group as shown in their Vega database (Ashurst *et al*., 2005) and by the NCBI RefSeq annotators (Pruitt *et al*., 2003). Recently, Ensembl, NCBI, Havana, and UCSC have begun to collaborate to create the Consensus Coding Sequence (CCDS) resource, which identifies the subset of genes (14 795) for which the annotation of a CDS is agreed across all databases.

Annotation and analysis is very far from complete. Up to now, the focus has been the protein coding genes, and only now are efforts being made to identify and annotate noncoding RNA genes (*see* Article 27, **Noncoding RNAs in mammals**, Volume 3) and microRNAs (*see* Article 34, **Human microRNAs**, Volume 3). Separating functional genes from pseudogenes accurately remains problematic (*see* Article 29, **Pseudogenes**, Volume 3), as does identifying functional alternative transcripts (*see* Article 30, **Alternative splicing: conservation and function**, Volume 3). Ultimately, gene annotation must also extend to promoters (*see* Article 33, **Transcriptional promoters**, Volume 3).

The existence of the sequence and a core annotation has driven researchers worldwide to collect experimental data and perform computational analysis on a genomic scale. One of the most powerful aspects of the genome browsers has been the way that they encourage the integration of external data. Technologies such

as the distributed annotation system (DAS) (Dowell *et al*., 2001), which facilitate this, are based on the recognition that there is too much data that can be related to the genome to be stored in a single database. This virtual integration of annotation will drive progress in our understanding the genome by allowing researchers to easily share and compare different data.

6. Epilogue

The HGP established a clear set of goals (Table 2), within which targets were set and regularly updated in a series of five-year plans. Given the knowledge and capacity for sequencing and analysis at the outset of the Project, these were enormously challenging. Each was met and exceeded. With the goals of the HGP achieved, the centers involved in the project continue to improve the sequence of the human genome, to sequence genomes of other organisms and, above all, to grapple with the real quest of turning this information into biomedical benefit.

Beyond the genome itself, the boldness of the HGP project – to rapidly generate such a critical and huge dataset, through a coordinated public project with immediate public release – has not only transformed biology, but has had a major impact on what is viewed as possible through public research. The HGP has been a triumph for an open data access model of scientific research and has led to policy changes to support this better in future (Dennis, 2003; Editor, 2003). Coming at a time of the success of other open models in software development (open source) and in open access publishing, there are serious questions as to whether this can go as far as open source drug development (Economist, 2004).

Knowledge of our genome sequence means that biologists can do experiments that were not possible in 1999, and can study in detail diseases that had too complex a genetic basis before the catalog of genes was available. Only rarely do biological experiments transform our world. With its careful attention to quality, to free release, and to ethical issues, the HGP has produced a resource that can truly benefit all of humankind.

References

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, *et al*. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Research*, **33** Database Issue, D459–D465.

Dennis C (2003) Draft guidelines ease restrictions on use of genome sequence data. *Nature*, **421**, 877–878.

Dowell RD, Jokerst RM, Day A, Eddy SR and Stein L (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Economist 2004. An open-source shot in the arm? In *Economist*.

Editor (2003) Sacrifice for the greater good? *Nature*, **421**, 875.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Research*, **30**, 38–41.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D (2002) The human genome browser at UCSC. *Genome Research*, **12**, 996–1006.

Pruitt KD, Tatusova T and Maglott DR (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Research*, **31**, 34–37.

Roberts L (2001) The human genome. Controversial from the start. *Science*, **291**, 1182–1188.

Sanger F, Coulson AR, Hong GF, Hill DF and Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. *Journal of Molecular Biology*, **162**, 729–773.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL and Eichler EE (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.

Sulston J and Ferry G (2002) *The Common Thread*, Bantam Press: London.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Waterston R and Sulston JE (1998) The human genome project: reaching the finish line. *Science*, **282**, 53–54.

Specialist Review

Genome assembly

Michael C. Wendl, John W. Wallis, Shiaw-Pyng Yang, Asif T. Chinwalla and LaDeana W. Hillier
Washington University, St. Louis, MO, USA

1. Introduction

Historians will undoubtedly rank the Human Genome Project (HGP) among the great achievements of humanity. HGP (*see* Article 24, **The Human Genome Project**, Volume 3) represented an advancement of more than an order of magnitude in size over preliminary "model organism" projects for *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998) and *Drosophila melanogaster* (Adams *et al*., 2000), which themselves were breakthroughs. A notable aspect of HGP was the problem of assembling data to deduce the original sequence. In this article, we briefly outline the concept of assembly and the complications arising from various factors, especially sequence repeats. We then describe the two different approaches applied by investigators to generate and assemble human genome data. Finally, we outline the still-evolving assessment of these two projects and attempt to place them in a historic context.

2. The concepts of shotgun sequencing and assembly

In any project, the length of DNA that can be resolved by a single sequencing reaction is invariably a small fraction of the source DNA whose sequence is desired. At their most fundamental level, projects must therefore take the so-called shotgun approach (Sanger *et al*., 1980; Anderson, 1981; Deininger, 1983), whereby small random fragments are read and then reassembled to deduce the original sequence. The assembly process is enabled by oversampling the genome, so that many overlapping substrings of DNA sequence are available for reconstruction.

In a purely random sequence, the assembly problem would be relatively straightforward. The population of repeated sequence strings of a given size decreases in a power-law fashion. In other words, given a specific DNA fragment of length L, a genome would need to be about 4^L bases in size before we would expect to find another copy of this sequence. Disregarding sequencing errors for the moment, suppose we find two human reads that share a specific 30-bp sequence. In random sequence, such a segment would only be expected about every $4^{30} \approx 1.2 \times 10^{18}$ bases. Because the likelihood of finding two copies in the comparatively

paltry 3×10^9 human genome would then be exceedingly small, we would be quite confident in assembling our two reads into a sequence contig. Continuation of this process would result in a gradual conglomeration of reads into ever-larger contiguous regions. In the limit, the complete sequence would eventually be recovered.

Unfortunately, actual DNA contains significant biases, including repeat structures that occur in many largely identical copies. Many instances are much longer than the length of a single sequence read, for example, LINE elements and segmental duplications (Samonte and Eichler, 2002). Sequence data derived from such regions may not be able to be uniquely, and thus, correctly placed in contigs. Consequently, there will be many admissible assemblies, only one of which is correct (Myers, 1999). This phenomenon was especially problematic in the HGP because the human genome contains a large and diverse population of repeat structures.

There was considerable debate as to the most desirable technique for sequencing the human genome (Weber and Myers, 1997; Green, 1997), much of which related to the issue of assembly. The publicly funded project favored what has come to be called the *clone-by-clone approach* (Figure 1a). This method had been used successfully for the preliminary *C. elegans* project (C. elegans Sequencing Consortium, 1998). Here, a minimal set of large-insert clones, for example, Bacterial Artificial Chromosome (BAC) clones, are sequenced individually and assembled in conjunction with the aid of longer-range data, for example, fingerprint maps and clone end sequences. Investigators were optimistic because the clones and the scale of mapping data are typically larger than the scale of repeats. The latter are then essentially encapsulated, permitting a theoretically unambiguous assembly.

Conversely, a newly founded private venture called *Celera* favored a whole genome shotgun (WGS) methodology (Myers, 1999). Here, genomic DNA is

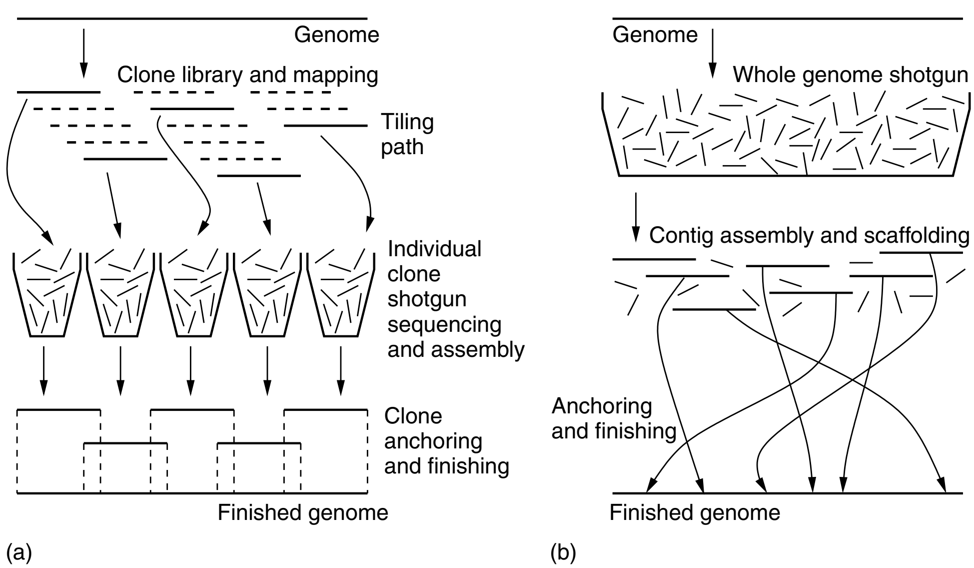


Figure 1 Diagrammatic representation of clone-by-clone sequencing (a) and whole genome shotgun sequencing (b)

sheared and sequenced directly (Figure 1b). Although the WGS technique had already been proven on a number of microbial genomes (e.g., Fleischmann *et al*., 1995), it was unclear whether success in assembling these low-repeat data would extend to the repeat-laden human genome. Proponents based their assessment of feasibility on a number of analyses. In particular, they predicted, based on the Lander and Waterman (1988) theory, that the number of gaps in the sequence would be small (Marshall and Pennisi, 1998). Simulations also suggested that the assembly problem would be tractable, despite the problem of repeats (Weber and Myers, 1997).

Ultimately, these two projects would proceed concurrently, but amid some controversy. WGS advocates maintained that the mapping phase of clone-by-clone sequencing was a slow, expensive, and unnecessary step. Conversely, sentiment in the public project was that a pure WGS assembly would not be well positioned for the downstream phase of completing the genome. In the following sections, we describe each of these two methods in greater detail. We will then draw a few comparisons between the results.

3. Whole genome shotgun approach

A purely WGS method does not utilize long-range mapping data or other nonsequence information to generate an assembly. From an algorithmic standpoint, the main design issue becomes correctly detecting and resolving repeated sequence in the native DNA. In their original proposal, Weber and Myers (1997) cited the so-called double-barrel shotgun approach as a mechanism of overcoming repeat difficulties. Here, inserts are size-selected such that their average length is greater than twice the average read length. These inserts can be sequenced from both ends, which yields not only coverage of the target, but also "mating" information for the pair of reads. That is, the separation distance between these reads is known, at least approximately, from the original insert size. As reads start coalescing into contiguous segments, this mate-pair information is useful in linking sequence contigs into larger "scaffolds". Moreover, it facilitates the encapsulation of repeat structures, as outlined above. This implies the necessity of obtaining mate-pair data over a variety of length scales.

One of the acknowledged liabilities of the WGS method is its susceptibility to errors that arise from corrupted mate-pairing information. Chimerism in clones is always a factor, although this can largely be eliminated by carefully designing and executing protocols. The larger contribution in early WGS projects came from sample tracking errors in the physical handling of DNA and lane tracking errors in analyzing gel images. Overall, such errors meant as many as 10% of the mate pairs were incorrect (Myers, 1999). By the time the pilot WGS project for Drosophila was under way, these mispairings could be reduced to substantially less than 1% (Myers *et al*., 2000). Consequently, mate-pair information could actively be utilized in the assembly process. On this basis, Myers *et al*. (2000) developed the assembly software, ultimately used for both the Drosophila and human genomes, which came to be called the *Celera assembler*.

3.1. Basic assembly methodology

The assembly process nominally begins with a computation that identifies all fragment overlaps. Input is in the form of a set of N fragments that have passed screens for vector and other contamination and that have been appropriately trimmed for base quality. Each new candidate fragment is compared to all existing fragments in the assembly, so that there are a total of $C_{N,2} = N\ (N - 1)/2$ comparisons. Matches of at least 40 bp and 94% similarity were viewed as significant. Such cases represent either a true fragment overlap or a repeat artifact (Figure 2). Misinterpreting the latter case as an overlap would effectively exclude an intervening segment of the original sequence.

In the next phase, sequence "unitigs" are constructed. Unitigs are defined by Myers *et al*. (2000) as local sequence assemblies whose overlaps are not disputed by any other data. However, this does not preclude the possibility of an erroneously assembled repeat region. Unitigs that exhibit uncharacteristically deep coverage are liable to be overcollapsed repeats. A probabilistic measure called the *A-statistic*, α, was devised to detect such cases (see Section 3.2). Large α values are associated with a small likelihood of a repeat-induced assembly error. Some of the more tractable repeat structures can be resolved at this point, using a dynamic programming algorithm to deduce repeat boundaries.

All unitigs determined to be free of repeat encumbrance are passed on to the "scaffolding" phase. Here, unitigs associated by mate-pair information are linked into scaffolds. In addition to yielding the order of unitigs, mate-pair data also allow them to be oriented relative to one another. A single mate pair is generally insufficient to declare a link, however, two or more imply a link to a statistically vanishing chance of error (Myers *et al*., 2000). Longer-range scaffolding is also performed on the basis of BAC end data. At the conclusion of this step, one largely has an ordered, scaffolded assembly of the nonrepetitive portion of the genome.

At this point, a series of increasingly aggressive steps of repeat resolution are undertaken. The first round attempts to place unitigs having relatively small α values and at least two links to other contigs. The next step concentrates on unitigs

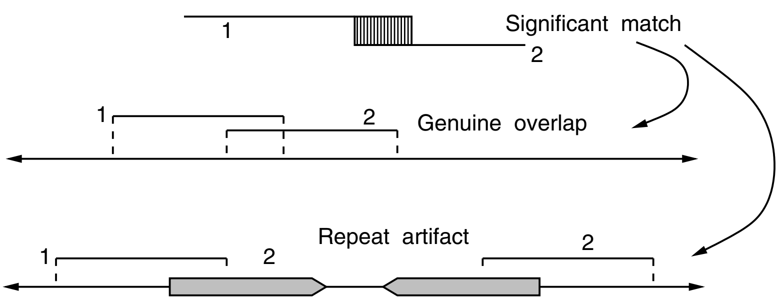


Figure 2 A significant match between two fragments represents a genuine overlap, or the collapse of a region due to a repeat structure. In the former case, fragments 1 and 2 actually derive from the same location in the source DNA molecule, while in the latter case they do not. If a repeat artifact were to be erroneously accepted as a valid overlap, the intervening region would be lost

with only a single link, except that it also requires the placement to be within the context of a tiling of other unitigs. In the original assembler used for the Drosophila project, gaps would be filled with the best overlap tiling of remaining unitigs; however, this step was omitted from the human assembler. At the conclusion of these steps, a consensus sequence is generated.

3.2. *Probabilistic repeat detection*

The primary metric used to detect collapsed repeat regions is the so-called A-statistic, denoted here by the symbol α. This statistic is derived according to a Poisson model for the number of reads one expects to find within a contig of length λ (Figure 3). Specifically, in a genome of estimated length G, the chance of a single read landing within the boundary of the given unitig is λ/G. This approximation neglects end effects, which is justified for large genomes. Presuming the N reads in the project to be independently and identically distributed (IID), one would then expect to find $\mu = N\lambda/G$ reads in the average unitig of length λ. In other words, in the absence of repeats, there would be an average of about μ reads from all contigs of length λ. Because N and λ/G are large and small, respectively, the population of reads essentially follows a Poisson process. That is, the probability of finding exactly k reads in a specific unitig of length λ is the Poisson probability

$$P(k, \mu) = \frac{\mu^k}{k!} \, e^{-\mu} \qquad (1)$$

where $e \approx 2.71828$ is Euler's number.

Now consider a unitig that is actually composed of two regions that have collapsed in the assembly because of a repeat structure (Figure 2). The average such contig would contain 2μ reads, so that the associated probability distribution is $P\ (k,\ 2\mu)$. The A-statistic is simply the base-10 log of the ratio of these two expressions.

$$\alpha = \log_{10}\left[\frac{P(k, \mu)}{P(k, 2\mu)}\right] = \cdots = \mu \log_{10}\ e - k \log_{10} 2 \qquad (2)$$

This expression relates the size of a contig to the number of reads it contains. Small values imply that a contig of given size contains more reads than what could reasonably be explained by random chance. One would then infer a collapsed

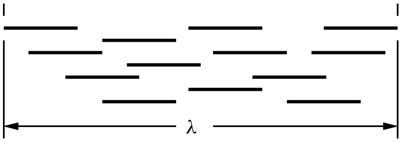


Figure 3 The A-statistic is based on the reads comprising a unitig of length λ bases

repeat. Extensive simulations indicated that $\alpha > 10$ almost always indicates the absence of repeat-related assembly error (Myers *et al*., 2000).

3.3. The human WGS assemblies

Venter *et al*. (2001) described two human shotgun assemblies: the whole genome assembly (WGA) and the compartmentalized shotgun assembly (CSA). Both of these assemblies combined native data generated by the WGS procedure, with additional data obtained from the public sequencing project (*see* Article 24, **The Human Genome Project**, Volume 3). WGS data comprised about 27 million reads derived from 2-, 10-, and 50-kb inserts. Variation in insert sizes provided linking information over several length scales. Public data were added in the form of 16 million synthesized reads, whose nature remains somewhat controversial (see Section 5). An additional 100 000 BAC end sequences were appropriated for long-range linking information.

The WGA assembly was performed essentially according to the algorithm described above. Parameters were selected such that error likelihood for each sequence overlap was about 10^{-17}. Unitigs were evaluated using equation 2, with the resulting set covering about 74% of the genome. After scaffolding and repeat resolution, BAC data were used to close additional gaps. The resulting assembly consisted of scaffolds spanning about 2.85 Gb with sequence coverage of about 2.59 Gb. There were about 11.3 million "chaff" reads that could not be incorporated into the assembly. Average scaffold and contig sizes were 1.5 Mb and 24 kb, respectively, and the total number of contigs was 221 000.

The CSA assembly is based on the premise of isolating smaller regions of the genome according to BAC contigs from the public project, creating localized assemblies using a combination of WGS and public data, and then tiling these assemblies into a larger result (Huson *et al*., 2001). A WGA assembly is then created from each resulting component. Here, about 6.2 million reads were ultimately "chaff". Average scaffold and contig sizes were 1.4 Mb and 23.2 kb, respectively, and the total number of contigs was 170 000. Anchoring of scaffolds to chromosomes was undertaken with the aid of physical mapping information from the public sequencing project (McPherson *et al*., 2001).

4. Clone-by-clone approach

The public sequencing effort explicitly set out with the goal of deciphering the genome in base-perfect form. That is, the error rate should be a maximum of one base in every 10 000. The approach would ideally be one of "divide and conquer" in two sequential steps. On a global level, the genome would be broken into a set of large-insert clones, predominantly BACs (Shizuya *et al*., 1992). At an average length exceeding 100 kb, they are larger than, and thus isolate most repeat structures. In the original plan, a physical map describing the order of these clones would then be generated via the high-throughput fingerprint technique (Marra *et al*., 1997). Here, clones are digested to create "fingerprints" of fragment lengths

(*see* Article 18, **Fingerprint mapping**, Volume 3). These lengths are inferred from band positions on electrophoretic gel images. A significant number of fragment length matches between two fingerprints indicate a likely overlap between their associated clones. The map would enable the selection of a minimally overlapping set of clones, each of which would then be locally shotgun sequenced and finished. With the given complement of information, generating the genome-wide consensus sequence would then essentially be a formality.

Investigators in the public effort deemed this approach more suitable than pure WGS on several grounds. First, WGS had not been proven on a repeat-rich genome and there was concern regarding the ability to use a draft assembly as a substrate for obtaining the base-perfect sequence. The outbred nature of the source DNA posed similar concerns. Moreover, previous projects had demonstrated that the mapping phase was a comparatively inexpensive component, usually less than 10% of total cost (Green, 1997). Improved mapping techniques had also been developed to increase throughput (Marra *et al*., 1997). The intermediate resolution afforded by BAC clones would also allow better targeting of underrepresented regions caused by inevitable cloning biases. Lastly, it would be more compatible with the distributed and international nature of the project.

A pilot phase proved the general feasibility of this approach by finishing a small portion of the genome (Lander *et al*., 2001). Around the same time, an alternative proposal was made to first generate a genome-wide draft sequence, while deferring the finishing phase to a later date. This was driven by two considerations. First, the research community was eager to obtain sequence data as quickly as possible. Second, the commercially funded Celera project sparked strong concerns about the privatization of the human genome and the rise of proprietary human databases (Pennisi, 1999). The public project thus proceeded along a modified version of the clone-by-clone approach as described below.

4.1. Clone selection

The modified project schedule actually necessitated constructing the physical map in parallel with the clone-sequencing phase rather than beforehand (McPherson *et al*., 2001). Consequently, clones could not be solely selected with guidance from a completed map to achieve a minimum tiling path. They were initially selected on the basis of having no overlap with already-sequenced clones or those in the sequencing pipeline. These "seed" clones formed a nonoverlapping set and thus functioned as nucleation sites for sequencing distinct regions of the genome. The fidelity of candidate seed clones was checked by comparing fingerprint bands to neighboring clones, to the clone size, and to its corresponding map fingerprint. Clones from contig ends were explicitly excluded because of their limited ability to confirm bands.

A clone registry was created for managing clones claimed by various participants for sequencing. Seed clones were then extended by picking mating clones having relatively high overlap scores (Sulston *et al*., 1988). Band confirmation and BAC end data were also used in this capacity. In cases where a seed clone sequence was available, overlaps could be further confirmed by comparison to the end sequence

of the candidate extension clone. The average overlap in the final clone set was found to be 47.5 kb (about 28%).

4.2. *Processing individual clones*

Individual large-insert clones were sequenced according to the standard shotgun methodology. Clones were typically sonicated and the resulting fragments were size-selected and subcloned into M13 or plasmid vectors. Details varied among the participants, for example, the proportions in which dye-primer and dye-terminator reactions were used and the type of sequencing platforms that were employed. Shotgun coverage was typically obtained to at least sixfold depth. Reads were assembled predominantly by the Phrap software package (Gordon *et al*., 1998), which uses an overlap-layout-consensus algorithm.

As with the WGS procedure, repeats tend to cause collapsed regions of misassembled reads. However, the localized nature of the assembly decreases the repeat problem substantially. Sequencing errors further complicate the process, which Phrap addresses through base-quality values assigned at the upstream base-calling step (Ewing *et al*., 1998). Phrap does not exploit mate-pair information in the way the Celera assembler does. Moreover, reads are predominantly derived from the same clone source, reducing problems associated with sequence polymorphism.

The first step of the Phrap assembly process determines which reads overlap by evaluating all pairwise subclone comparisons. Anomalies, such as unremoved vector and chimeric reads, are identified as well. Revised quality scores are then computed for each base. These reflect both the raw base-quality scores from the individual reads and any enhancements based on mutual overlap confirmation. The second step determines the optimal placement of overlapping reads to form a layout of the assembly. Here, sequence alignment scores are used in decreasing order to merge contigs. Some joints are rearranged at this stage as well. Finally, consensus sequence is generated for each base position from the columnized layout.

4.3. *Creating a draft assembly*

Assembly at the genome level involves combining the information from the clone overlap physical maps and validating those overlaps using the sequence. Given the draft nature of most of the clones and the fact that some were not yet represented in the map, this phase would not be trivial. After a uniform filtering procedure for contaminated sequences, the assembly approach proceeded in two steps: layout and merging.

The layout step sought to associate sequenced clones with clones on the physical map. Ideally, this would simply be a matter of mapping the names between the two clone sets. However, to minimize mistakes at this stage, two additional quality control procedures were instituted. Fingerprint accuracy was such that *in silico* digests could be used to confirm associations for sequence clones that had few enough contigs. BAC end sequences were also used in a similar capacity. A total of 25 403 clones were linked in this fashion. Fingerprint clone contigs were localized

to their respective chromosomal locations using fluorescence *in situ* hybridization (FISH) data (*see* Article 22, **FISH**, Volume 1) and a variety of radiation hybrid and genetic maps (Lander *et al*., 2001).

Software was then used in the merge phase to order and orient the entire collection of sequenced clones (Kent and Haussler, 2001). Sequence clones were initially aligned to those within the same fingerprint clone contig. Groups of overlapping clones were then built up into larger ordered "barges". Order and orientation were further pursued with the aid of additional information, for example, mRNAs, STSs, BAC end reads, and any other available linking information. A sequence path was then generated from the resulting structure.

The two-phase process used for the public sequencing effort resulted in the draft assembly described by Lander *et al*. (2001). They reported an overall N_{50} length for sequenced clone contigs of 826 kb. This is the length of the contig in which the average nucleotide resides. The N_{50} statistic varied by chromosome. In particular, it was a few orders of magnitude higher for chromosomes 21 and 22, which had already been completed (Dunham *et al*., 1999; Hattori *et al*., 2000). The overall assembly consisted of 4884 sequenced clone contigs and suggested a euchromatic genome size of 2.9 Gb.

5. Epilogue

The public and private genome assemblies were published simultaneously in February of 2001 (Lander *et al*., 2001; Venter *et al*., 2001). Celera's results quickly came under scrutiny based partially upon the amount of public sequencing data that had been appropriated, but more so on the manner in which it had been used. Waterston *et al*. (2002a) described the apparently nonrandom fashion in which synthetic "reads" were manufactured from the public genome assembly and tiled over Celera's own data. They further clarified the extent to which public marker and map data were used by Celera and the nature of the CSA assembly, which was the basis of all the biological analysis reported by Venter *et al*. (2001). Emphasizing the fact that Celera did not report an assembly comprised solely of their own shotgun data, Waterston *et al*. concluded that the published Celera assemblies did not represent a legitimate proof of concept of the WGS method applied to a complex genome. This view was echoed by other members of the scientific community (Green, 2002). Myers *et al*. (2002) rebutted this position, asserting that no linking information was retained in the synthetic reads and that the WGA and CSA assemblies were, in fact, pure WGS assemblies. Vigorous debate continues (Waterston *et al*., 2003; Adams *et al*., 2003; Istrail *et al*., 2004).

It will likely be some time yet before history conclusively places these efforts into a more objective relational context. Celera would certainly have averted any claims by simply eschewing public data. Their initial projections regarding the difficulty of such an assembly based on shotgun data alone, for example, number of gaps and amount of chaff, were clearly incorrect. This created an unexpected urgency for more data. While their use of public data was unquestionably legitimate, it has significantly blurred the comparison between these two fundamentally different approaches.

Regardless, the Celera assemblies might ultimately be relegated to scientific footnotes in the sense that they are unlikely to be extended into finished human sequences. The drive toward completing the genome is proceeding solely on the basis of the public data, and chromosomes are now being individually finished in fairly rapid succession (Deloukas *et al*., 2001; Heilig *et al*., 2003; Hillier *et al*., 2003; Mungall *et al*., 2003; Dunham *et al*., 2004; Grimwood *et al*., 2004). It is somewhat ironic that neither the WGS nor the clone-by-clone strategies appear likely to carry forward in pure form for future projects. Investigators currently favor hybrid approaches that combine aspects of both these techniques (Waterston *et al*., 2002b). The art of sequencing and assembling genomes is clearly an evolving one.

Related articles

Article 2, **Algorithmic challenges in mammalian whole-genome assembly**, Volume 7; Article 7, **Errors in sequence assembly and corrections**, Volume 7; Article 8, **Genome maps and their use in sequence assembly**, Volume 7

Acknowledgments

This work was supported by a grant from the National Human Genome Research Institute (HG002042).

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.

Adams MD, Sutton GG, Smith HO, Myers EW and Venter JC (2003) The independence of our genome assemblies. *Proceedings of the National Academy of Sciences*, **100**, 3025–3026.

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, **9**, 3015–3027.

C elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012–2018.

Deininger PL (1983) Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Analytical Biochemistry*, **129**, 216–223.

Deloukas P, Matthews LH, Ashurst J, Burton J, Gilbert JGR, Jones M, Stavrides G, Almeida JP, Babbage AK, Bagguley CL, *et al*. (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.

Dunham A, Matthews LH, Burton J, Ashurst JL, Howe KL, Ashcroft KJ, Beare DM, Burford DC, Hunt SE, Griffiths-Jones S, *et al*. (2004) The DNA sequence and analysis of human chromosome 13. *Nature*, **428**, 522–528.

Dunham I, Shimizu N, Roe BA, Chissoe S, Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, *et al*. (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.

Ewing B, Hillier L, Wendl MC and Green P (1998) Base-calling of automated sequencer traces using *Phred. I.* Accuracy assessment. *Genome Research*, **8**, 175–185.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of H. *influenzae* rd. *Science*, **269**, 496–512.

Gordon D, Abajian C and Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Green P (1997) Against a whole-genome shotgun. *Genome Research*, **7**, 410–417.

Green P (2002) Whole-genome disassembly. *Proceedings of the National Academy of Sciences*, **99**, 4143–4144.

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Gyamfi MT, *et al*. (2004) The DNA sequence and biology of human chromosome 19. *Nature*, **428**, 529–535.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al*. (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.

Heilig R, Eckenberg R, Petit J, Fonknechten N, Da Silva C, Cattolico L, Levy M, Barbe V, de Berardinis V, Ureta-Vidal A, *et al*. (2003) The DNA sequence and analysis of human chromosome 14. *Nature*, **421**, 601–607.

Hillier LW, Fulton RS, Fulton LA, Graves TA, Pepin KH, Wagner-McPherson C, Layman D, Maas J, Jaeger S, Walker R, *et al*. (2003) The DNA sequence of human chromosome 7. *Nature*, **424**, 157–164.

Huson DH, Reinert K, Kravitz SA, Remington KA, Delcher AL, Dew IM, Flanigan M, Halpern AL, Lai Z, Mobarry CM, *et al*. (2001) Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics*, **17**, S132–S139.

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR, *et al*. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proceedings of the National Academy of Sciences*, **101**, 1916–1921.

Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, **11**, 1541–1548.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD and Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research*, **7**, 1072–1084.

Marshall E and Pennisi E (1998) Hubris and the human genome. *Science*, **280**, 994–995.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, Jones MC, Horton R, Hunt SE, Scott CE, *et al*. (2003) The DNA sequence and analysis of human chromosome 6. *Nature*, **425**, 805–811.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KHJ, Remington KA, *et al*. (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

Myers EW, Sutton GG, Smith HO, Adams MD and Venter JC (2002) On the sequencing and assembly of the human genome. *Proceedings of the National Academy of Sciences*, **99**, 4145–4146.

Myers G (1999) Whole-genome DNA sequencing. *Computing in Science and Engineering*, **1**, 33–43.

Pennisi E (1999) Academic sequencers challenge Celera in a sprint to the finish. *Science*, **283**, 1822–1823.

Samonte RV and Eichler EE (2002) Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, **3**, 65–72.

Sanger F, Coulson AR, Barrell BG, Smith AJ and Roe BA (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology*, **143**, 161–178.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences*, **89**, 8794–8797.

Sulston J, Mallett F, Staden R, Durbin R, Horsnell T and Coulson A (1988) Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biosciences*, **4**, 125–132.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Waterston RH, Lander ES and Sulston JE (2002a) On the sequencing of the human genome. *Proceedings of the National Academy of Sciences*, **99**, 3712–3716.

Waterston RH, Lander ES and Sulston JE (2003) More on the sequencing of the human genome. *Proceedings of the National Academy of Sciences*, **100**, 3022–3024.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002b) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Weber JL and Myers EW (1997) Human whole-genome shotgun sequencing. *Genome Research*, **7**, 401–409.

Specialist Review

Segmental duplications and the human genome

Rhea U. Vallente
Washington State University, School of Molecular Biosciences, Pullman, WA, USA

Evan E. Eichler
University of Washington School of Medicine, Seattle, WA, USA

1. Introduction

The process of gene and genome duplication has played a major role in evolution and has sparked the interest of scientists for decades. Susumu Ohno's concept, through his highly influential book "*Evolution by gene duplication*" (Ohno, 1970), focused on the collective effect of duplications, mutations, and natural selection, wherein an increase in gene number followed by mutation may lead to organismal and functional complexity. In addition, genomes contain a large quantity of repetitive sequences, far in excess of that devoted to protein-coding genes (C-value paradox). At least 50–55% of the human genome is composed of repeat sequences, broadly composed of interspersed repeats, processed pseudogenes, simple sequence repeats, tandem repeats, and segmental duplications. Repeats have long been described as "junk", yet they actually play an important role in disease etiology, speciation, mutation and selection. These sequences may also shed light on chromosome structure and dynamics by reshaping the genome through rearrangements, creating entirely new genes and modulating the overall GC content.

Segmental duplications (also known as *low-copy repeat elements*) are a newly identified class of repetitive elements that consist of physically interspersed blocks of duplicated material in a genome. Duplicon is a specialized term applied to a segmental duplication whose extent of ancestral duplication is unambiguously delineated. In humans, these segments typically cluster in specific locations leading to structurally larger and more complex arrays that are later duplicated and rearranged into a mosaic patchwork architecture. Several of these clusters or nexi have been associated with the etiology of recurrent chromosomal rearrangements. The functional importance of these regions in genome evolution is only now beginning to become apparent.

2. Structural organization of segmental duplications in the human genome

Initial sequence analysis of the human genome has shown that ∼5–6% of the genome is composed of segmental duplications ranging in size from a few to

hundreds of kilobases, which are composed of a mosaic of duplicated segments originating from diverse areas of the genome (Cheung *et al*., 2001; Bailey *et al*., 2002a). The high degree of sequence identity (more than 60% of these duplications by mass show greater than 97% identity) suggest a recent origin. The first anecdotal reports on segmental duplications arose from routine physical mapping during the early phases of the Human Genome Project or as consequences of characterizing breakpoints associated with recurrent chromosomal structural rearrangements (Eichler, 1998; Inoue *et al*., 1999). On the basis of their distribution pattern, two types of segmental duplications may be distinguished. *Intrachromosomal duplications (chromosome-specific repeat regions (REPs) or low-copy repeat sequences (LCRs))* are genomic sequences interspersed along a single chromosome. On the other hand, segmental duplications located among nonhomologous chromosomes are termed *interchromosomal duplications* (*transchromosomal*). The amount of intrachromosomal and interchromosomal duplications varies dramatically among human chromosomes (Figure 1).

Several signature features distinguish intrachromosomal duplications, although finer details of the organization and distribution of subsequences in every chromosome vary. LCRs are often physically restricted to one chromosome and are situated in the proximal portions (within the first 10 Mb) of human euchromatic arms. Several chromosome-specific LCRs have been identified on a large number of chromosomes. Each block consists of smaller repeats (*duplicons*) that correspond to fragments of genes that originated from an ancestral locus. Most LCRs are relatively recent, as shown by their high range of sequence identity (97.5–99%), with variation reaching allelic levels or single-base changes. These interspersed intrachromosomal duplications predispose adjacent unique sequences to recurrent chromosomal structural rearrangements that are most often associated with diseases.

Interchromosomal duplications are physically biased to accumulate near heterochromatic regions (pericentromeric) and subtelomeric portions of chromosomes, possibly acting as a transition sequence between satellite repeat sequences and protein-coding sequences. Nearly half of all human chromosomes contain a 1–2-Mb zone of duplication extending from the satellite-repeat-containing centromere to the unique euchromatic regions. In the case of subtelomeric regions, blocks of interchromosomal duplications are usually situated within <100 kb of the telomeric T_2AG_3 tracts. A large number of partial- and whole-gene duplications has been characterized in detail. These segmental duplications share conserved exon-intron structure and have from 2 to 40 copies interspersed throughout the genome (Eichler, 2001). At least 30% of pericentromeric duplications originated from duplicatively transposed euchromatic regions (She *et al*., 2004a).

2.1. Pericentromeric duplications

Pericentromeric regions demarcate the transition between the heterochromatic satellite DNA at the centromere and the euchromatic gene-containing chromosome arm sequences. Such regions are sites of rapid evolutionary turnover, reduced gene expression, and suppressed genetic recombination. Directed mapping and sequencing of human pericentromeric regions reveal a preponderance of partial

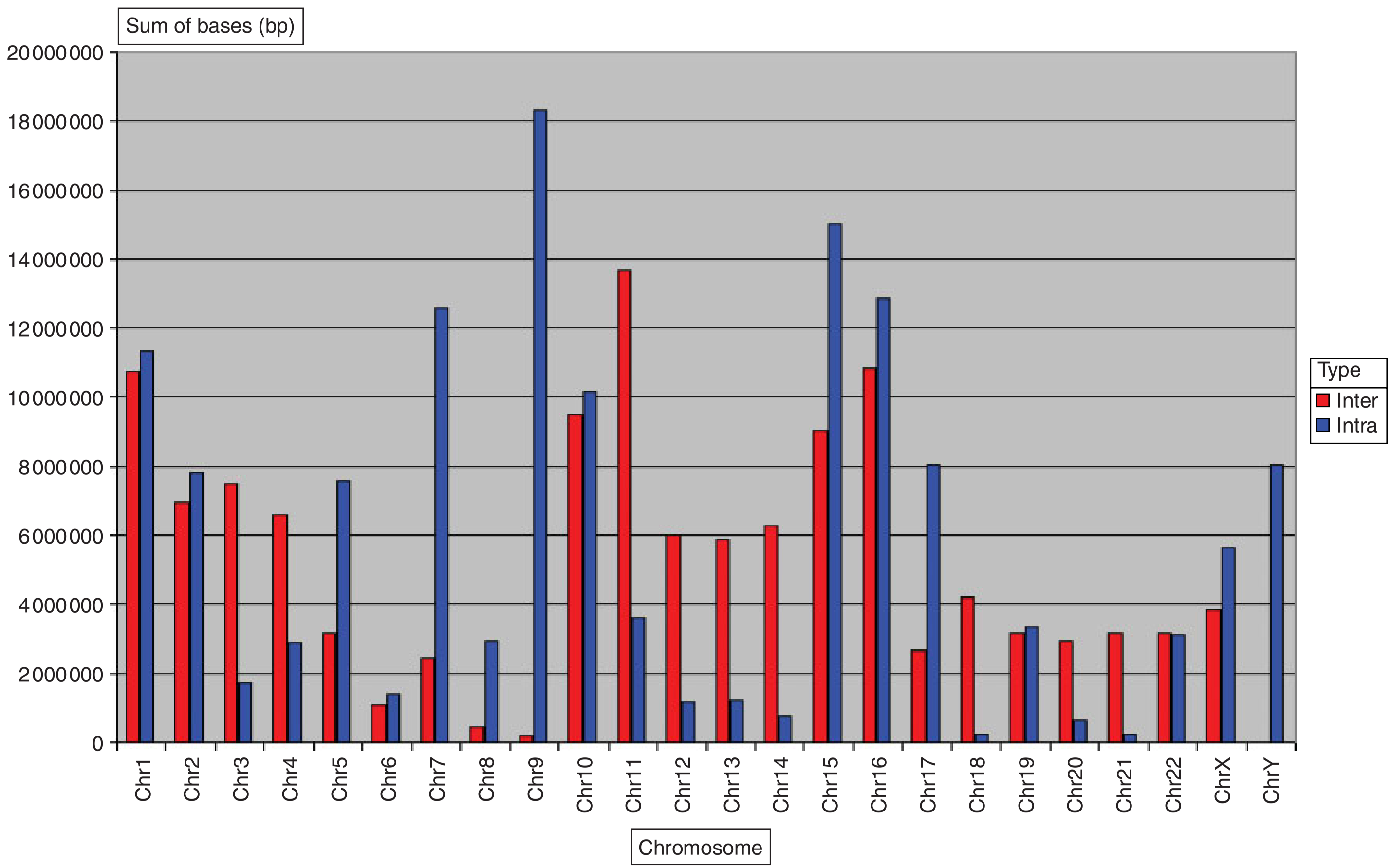


Figure 1 Segmental duplications in the human genome based on sequence identity and alignment length (>90%, >1 kb) of the finished human genome sequence (build35; May 2004). Inter- (red) and intra-chromosomal (blue) duplications vary per chromosome

gene duplications (Jackson *et al*., 1999; Loftus *et al*., 1999). Although the occurrence of mobile genetic elements and duplicated sequence within pericentromeric regions is a common property shared by distant species, the structure of the human genome appears to be unique in the proportion and extent of euchromatic blocks of duplications (Dunham *et al*., 1999; Hattori *et al*., 2000; Bailey *et al*., 2002b). Genome-wide assessment of segmental duplications revealed a sevenfold bias for duplications to map in these specific regions of the genome.

A proposed model of pericentromeric organization consists of two distinct domains (Guy *et al*., 2000): (1) a proximal domain that is satellite-rich, transcript-poor, and prone to interchromosomal duplication and (2) a distal domain that is satellite-poor and prone to intrachromosomal rearrangement. Similar tracts along the intrachromosomal duplications within the pericentromeric regions of chromosomes 21 and 22 have been reported (Orti *et al*., 1998; Edelmann *et al*., 1999; Ruault *et al*., 1999; Footz *et al*., 2001). An apparent exception to this two-domain model is that of KIAA0187, which had multiple rounds of pericentromeric duplications and dispersal through intrachromosomal rearrangements (Crosier *et al*., 2002). A fraction of these duplications show no evidence of an interstitial euchromatic origin and are now collectively termed pericentromeric interspersed repeats (PIRs), to distinguish them from duplicons of euchromatic origin (Horvath *et al*., 2003). An example is PIR4, a 49-kb element that localizes exclusively to more than half of all human pericentromeric regions. The frequent association of these elements with the boundaries of interspersed segmental duplications suggests that they may play a role in interchromosomal duplication and/or that they represent the original milieu wherein the first segmental duplications are duplicatively transposed.

2.2. *Subtelomeric duplications*

Subtelomeric regions immediately adjacent to terminal $(T_2AG_3)_n$ tracts are preferential sites (–three- to fourfold) for the accumulation of segmental duplications (Bailey *et al*., 2001). Genome-wide analysis of human subtelomeric regions (500 kb from the termini) found that ~10% of the region consisted of duplicated material located within the last 100 kb of the chromosomal arm (Riethman *et al*., 2004). The duplicated portions are composed of two distinct domains. The distal subdomain lies adjacent to the telomere and is composed of a mosaic of short segments of shared sequence homologies from many different chromosomes. These shared homologies are <2 kb in length and suggest the occurrence of frequent exchanges among all telomeres (Flint *et al*., 1997). The proximal subdomain is composed of much longer segments (10–40 kb) of higher sequence identity, which indicate a more recent origin. These recent duplications could be the result of unbalanced translocations occurring between telomeres during primate evolution.

3. Mechanisms of origin of segmental duplications

As opposed to whole-genome duplication, the mechanism(s) by which segmental duplications propagate within genomes are only beginning to be established.

Detailed characterization of segmental duplications have suggested that progenitor loci of these highly homologous segments most often occur outside the pericentromeric duplication zone (Regnier *et al*., 1997; Zimonjic *et al*., 1997; Horvath *et al*., 2001). Mechanisms involve duplicative transposition of apparently normal genomic DNA, often containing genes and smaller repetitive elements and range in length from <1 to 100 kb in length. Eichler and Horvath (Horvath *et al*., 2000; Eichler *et al*., 1997) proposed a two-step model for the dispersal of duplicated segments among multiple pericentromeric regions. The first step involves transposition seeding, wherein genomic segments from different chromosomal origins transposed to an ancestral pericentromeric region. An estimate for the time of occurrence for these events was ~5–15 million years ago, around the time of the divergence great-ape lineages. Multiple secondary events apparently duplicate larger blocks among pericentromeric regions of other nonhomologous chromosomes (Figure 2). These duplication events are more recent, and a few are restricted to the human lineage.

The mechanism for subtelomeric duplications likely differs from that observed in pericentromeric regions. Nonspecific, cryptic subtelomeric chromosomal rearrangements are known to occur frequently in humans, at a rate of 3–7.4% of patients with idiopathic mental retardation (Knight *et al*., 1997). During recent primate evolution, small unbalanced translocations may have occurred at a frequency that effectively contributed to the diversity of the subtelomeric regions through fixation. The sequencing analysis of the breakpoint sequences in ancestral subtelomeric regions supports this hypothesis (Martin *et al*., 2002), sharing strong yet discontinuous homology with ancestral copies at nonhomologous subtelomeric regions, as a result of sequence exchange among nonhomologous chromosomes during clustering of telomeres at the nuclear envelope (chromosomal bouquet formation). Segmental duplications at subtelomeric regions are thought to increase the gene copy number if functional genes are retained (Linardopoulou *et al*., 2001).

The mechanism by which intrachromosomal duplications have adopted an interspersed configuration within the human genome is unknown. Several studies however, point to Alu retroposons as playing a role in this process (Bailey *et al*., 2003; Babcock *et al*., 2003). Alu repeat sequences, particularly chimeric Alu repeats composed of two or more different subfamily sequence signatures, are enriched precisely at the boundaries of segmental duplications. This change supports an Alu–Alu recombination model for the origin of their interspersed configuration. Once interspersed segmental duplications have emerged, they are further propagated by nonallelic homologous recombination. A signature of such a recombination event is a change in copy number of sequence located between two segmental duplications. The rate of unequal crossing-over can rise dramatically in areas of tandemly repeated sequences. The activity of mobile elements and the nature of surrounding sequences influence the occurrence of duplications (Otto and Yong, 2002). Complex mechanisms have evolved to decrease the rate at which duplications establish within populations (Selker, 1999). Other general mechanisms that either silence duplicated sequences or degrade messenger RNA transcripts of duplicated genes are possibly present. These mechanisms have evolved as defense mechanisms to curb the proliferation and interference of transposable elements and dying duplication transcripts.

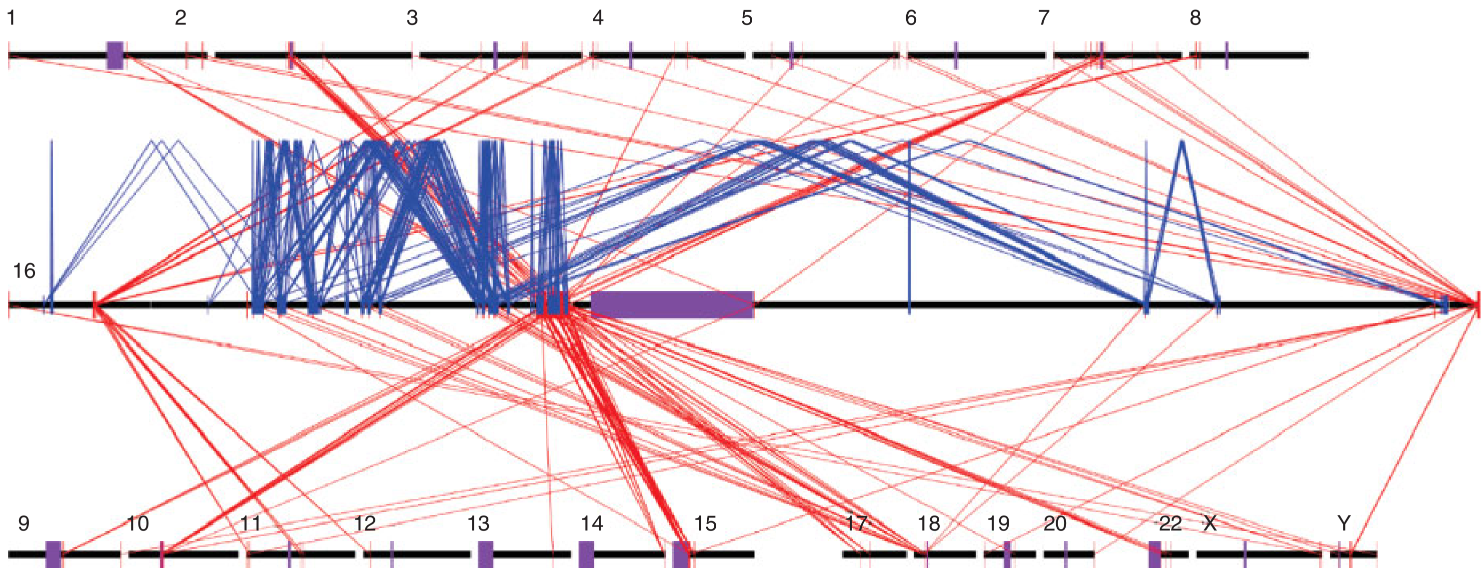


Figure 2 Recent segmental duplications on human chromosome 16. Large (>20 kb) highly similar (>90%) intrachromosomal (blue) and interchromosomal (red) segmental duplications are shown for chromosome 16. Chromosome 16 is magnified in scale relative to the other chromosomes. The intrachromosomal duplications shown may be sites of unequal homologous recombination resulting in large-scale rearrangements such as chromosomal region 16qh heteromorphisms. The centromere is colored purple. Cited from She *et al*. (2004b), **http://humanparalogy.gs.washington.edu/build35/chrom_cut.htm**

4. Screening tools for segmental duplications

Initially, identification and characterization of segmental duplications were based on anecdotal reports involving fluorescence in situ hybridization (FISH) probes hybridizing to unexpected multiple chromosomal sites or duplicated sequences existing at recurrent chromosomal breakpoints (Figure 3). While methods for screening mutations are well established, quantitative gene dosage measurements for larger (>100 bp) genomic regions may be underestimated (Armour *et al*., 2002). Computational, molecular, sequence, and cytogenetic methods, when used collectively, provide the most powerful assay to detect segmental duplications, with each method serving as a validation or complement to the other detection method. In certain laboratory settings, classical time-consuming methods such as Southern blot or other molecular tools screen for duplications and deletions. Recently, several new approaches have emerged that demonstrate high specificity and sensitivity in detecting recent duplications.

Computational methods provide a rapid means to detect duplicated DNA at a genome-wide level. In principle, it involves identification and extraction of high-copy repeats from the genomic sequence using RepeatMasker (Smit and Green, **http://repeatmasker.genome.washington.edu**), searching for similarities in the remaining unique sequences by BLAST (Altschul *et al*., 1997), reinserting repeats

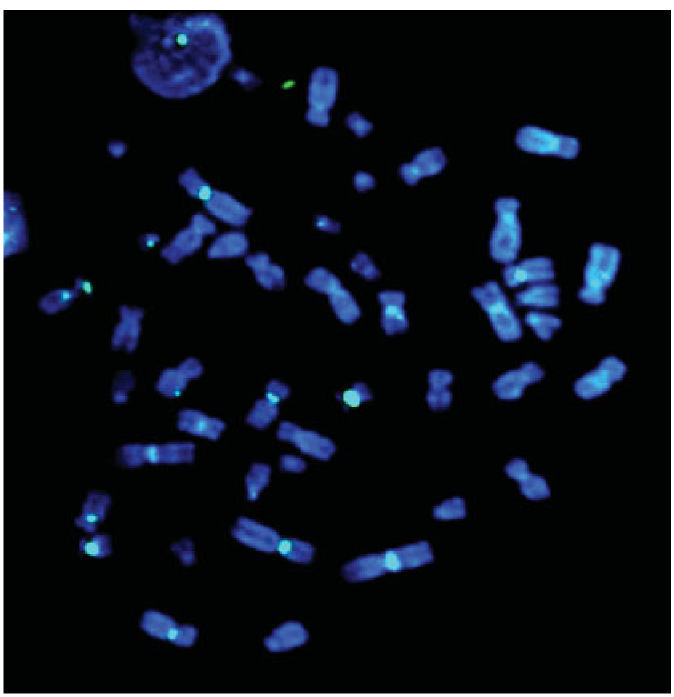


Figure 3 Localization of human chromosome 16p11-specific clone AC002037 in a metaphase chromosome spread shows multiple hybridization signals on chromosomal regions 2p11, 10p11, and 22q11

to generate pairwise alignments, heuristic trimming of the ends of the alignments, and generating global alignments with statistics (Bailey *et al*., 2001; Venter *et al*., 2001). Some selection criteria have been developed to minimize false-positives. Size ≥1 kb and sequence identity of ≥90% thresholds typically detect duplication events within the last 35 million years of primate evolution. Junction analyses also provide valuable hints on sequence exchange mechanisms between specific nonhomologous regions in the genome (Mefford *et al*., 2001; Mefford and Trask, 2002). More recently, assembly-independent approaches have identified potential duplications within genome sequence (Bailey *et al*., 2004a).

FISH has served as a straightforward experimental system for physically mapping DNA probes to metaphase chromosomes using antibody-based staining procedures. Probes generated from YACs, BACs, and PACs are either labeled directly using fluorochrome-conjugated nucleotides (e.g., fluorescein-dUTP; Cy5-dUTP) or indirectly using reporter molecules (e.g., biotin-dUTP; digoxigenin-dUTP) by nick-translation, random priming, or other molecular genetic techniques. FISH has provided a mechanism in determining the presence, number, and distinct location (in situ) of genetic material for characterization of intra- and interchromosomal rearrangements, regardless of their complexity. FISH fills the gap between molecular biology and chromosome banding analysis, resulting in a marked progress in cytogenetics research (for review, see (Trask, 2002)). Comparative mapping using region-specific probes has facilitated the identification of subchromosomal homologies, detailed reconstruction of genomic changes, and map evolutionary breakpoint regions in closely related species (Conte *et al*., 1998; Samonte *et al*., 1998). Advances in microscopy and signal detection hardware have allowed the low light level produced by FISH signals to be recorded and analyzed with increasing sensitivity. This 20-year-old technology has been most extensively used in the confirmation of segmental duplications (Rosenberg *et al*., 2001; Der-Sarkissian *et al*., 2002; Ravise *et al*., 2003).

A modified version of FISH that provides a much higher resolution is fibreFISH, which collectively refers to all hybridizations performed on released chromatin or DNA fiber, including chromatin FISH, halo FISH, visual mapping, direct visual hybridization (DIRVISH), and molecular combing. The degree of condensation is much less than metaphase chromosomes and interphase chromatin and is effective in the establishment of sequence order of BAC contigs, mapping ESTs, estimating and filling gaps in sequence-ready maps, and determining copy number of large-scale segments in the human genome (Ziolkowski *et al*., 2003; Pagel *et al*., 2004). FibreFISH is also used in medical genetics for the study of gene amplification, deletion and translocation, as well as the correlation of gene numbers with clinical phenotypes in diseases caused by additional copy numbers of genes. For segmental duplications, fibreFISH is most valuable in confirmation of closely spaced tandem duplicates and characterization of copy number variants in these regions.

Microarray-based comparative genomic hybridization (array-CGH or matrix-CGH) is a relatively new and revolutionary platform for high-resolution detection of DNA copy number aberrations. The improvement of this technique over a short time has provided many applications in research and high-throughput diagnostics. CGH was originally developed for genome-wide analysis of copy number imbalances

by cohybridization of differentially labeled whole genomic test and control DNA in a ratio of 1:1 on normal control metaphase spreads under conditions of in situ suppression hybridization (CISS). The resulting fluorescence intensity ratio is quantified by an imaging software that calculates a copy-number or molecular karyotype of the entire genome. However, the use of metaphase chromosomes provided low resolution (5–10 Mb for deletions and 2 Mb for duplications). The resolution of CGH improved by using arrayed subsets of genomic clones (BACs, PACs, and cosmids) and cDNA. Initial testing of the array-CGH focused on chromosome-wide screening of gene duplications and deletions and subsequently moved on to genome-wide scans and specific chromosomal regions. Resolution of CGH-arrays have improved over time, with an average resolution of 1.3 Mb (Fritz *et al*., 2002; Wilhelm *et al*., 2002). A recent genomic array format employing PCR repeat-free and nonredundant sequences operates with a resolution of ~23 kb (Mantripragada *et al*., 2003). Segmental duplications vary in size and degree of sequence identity, and hence the ability to detect gene dosage changes in these regions has been questioned. There is currently a limited yet promising number of investigations on these genomic regions (Locke *et al*., 2004; Shaw *et al*., 2004). Array-CGH has the power to discriminate between one and two DNA copies, as well as to detect homozygous deletions and unbalanced translocations. However, signal-to-noise ratio remains as a limiting factor in the reliability of microarray data. Although array-CGH holds considerable promise to revolutionize cytogenetics, the technology is not capable of detecting low-level mosaicism, balanced rearrangements or inversions, and for the time being, can only be seen as a supplementing technique for the examination of chromosome aberrations (Iafrate *et al*., 2004).

5. Role of duplications in human disease

Several studies have demonstrated that closely located segmental duplications are one of the factors predisposing to the occurrence of genomic disorders, mainly by acting as highly unstable hotspots for chromosomal rearrangements (Ji *et al*., 2000; Emanuel and Shaikh, 2001). Such recurrent chromosomal changes result in dosage imbalances caused by deletion or duplication of genes that lie within the rearranged segments (*segmental aneusomy*). Gene dosage imbalances further result in misalignment and nonallelic homologous recombination between segmental duplications. The term *genomic disorder* was coined to designate diseases presenting with gene dosage etiologies (Lupski *et al*., 1998). A subset of these disorders is presented in Table 1. The most common rearrangements in the human genome are the deletions associated with DiGeorge syndrome (DGS; OMIM 188400), velocardiofacial syndrome (VCFS; OMIM 192430), and conotruncal anomaly face syndrome (CAFS; OMIM 217095). The repeat units involved with deletions are usually small and retain the same orientation on the same chromosome, leading to slipped pairing and unequal crossing-over in meiosis. Tandem duplication of DNA can directly lead (through dosage effects) or indirectly (through subsequent microdeletion) to clinical phenotypes, including velocardiofacial syndrome (Edelmann *et al*., 1999) and Prader Willi/Angelman

Table 1 Examples of low-copy repeats (LCRs) associated with genomic disorders

Chromosomal localization	LCR name	Size of LCR	Associated genomic disorder	Size and type of rearrangement
7q11.23		320 kb	Williams–Beuren syndrome	1.6-Mb deletion
8p23	REPD	~1.3 Mb		~4.7-Mb inversion
	REPP	~0.4 Mb		
8q21	CYP11B1/2	10 kb	Hypertension?	Duplication
15q11-14	LCR15/HERC2	15 kb	Prader–Willi or Angelman's syndrome (PWS, AS)	4-Mb deletion
17p11.1	SMS-REP	200 kb	Smith–Magenis syndrome	3.7-Mb deletion
17p12	CMT1A-REP	24 kb	Charcot-Marie-Tooth disease type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsy (HNPP)	1.4-Mb duplication/ deletion
21q21.3 → qter		550 kb	Mild developmental delay and mild dysmorphic features	15.5-Mb deletion
22q11	LCR22	225–400 kb	Velocardiofacial syndrome/DiGeorge syndrome, der(22) syndrome, and cat-eye syndrome	3-Mb deletion
Xq28	RCP, GCP	39 kb	Color blindness	Deletion
Yq11.2	AZFa, AZFc	229 kb	Male infertility	800-kb deletion

syndromes (Christian *et al*., 1999). In addition, large tracts of sequence frequently transpose into pericentromeric locations and distribute between nonhomologous chromosomes in a centromere-specific manner (Eichler *et al*., 1999). Construction of a chromosomal duplication map of malformations (Brewer *et al*., 1999) showed that half of the bands associated with duplications or deletions are pericentromeric, representing triplo- and haplolethal regions of the human genome. In addition, hyperploidy is better tolerated than hypoploidy in the human genome, with phenotypes associated with deletions more severe than those associated with duplications. Statistical analyses used to create cytogenetic duplication and deletion maps have shown that breakpoints involved in duplications are not randomly distributed, but instead, cluster within certain chromosome regions. Such maps may be useful in obviating the need for whole-genome scans as a first approach to identify disease genes.

The presence of heterozygous inversions at regions flanked by the segmental duplications may also be predisposed to further rearrangement (Gimelli *et al*., 2003). These submicroscopic inversions would interfere with the normal homologous synapses and promote misalignment and abnormal recombination. Moreover, inversions are a common cryptic feature of unstable chromosomal regions in the general population, and can thus be considered genomic polymorphisms (Samonte *et al*., 1996; Osborne *et al*., 2001). Rearrangements involving duplicatively transposed genomic sequences have been termed segmental polymorphisms, structural variants, large copy-number variants (LCV) or copy-number polymorphisms (CNPs). The characterization of such variation at the sequence level will require unprecedented high-throughput screening at a genome-wide level not readily achieved by current technology.

6. Segmental duplications and primate genome evolution

Previous decades of comparative analysis of proteins and chromosomes in higher primates and humans have shown striking similarities. Cytogenetically, hominoid chromosomes only differ by some gross chromosomal rearrangements such as peri- and paracentric inversions, reciprocal translocations, band insertions, and chromosome fusions. Inversions were the most common rearrangements in the great ape and human karyotypes, while translocations, Robertsonian fusions, and fissions, appeared to be more common in lesser apes and lower primates. On the basis of the limits of resolution of chromosome banding and painting techniques, interchromosomal rearrangements occur minimally, while intrachromosomal rearrangements (inversions, transpositions) needed further delineation of chromosome subregions. Finer-detailed karyotypic differences were later described by ZOO-FISH mapping.

The identification of highly homologous segmental duplications in the human genome has triggered comparative screening of these large-scale segments in closely related species (Table 2). The role of segmental duplications as one of the major driving forces in primate genome evolution has recently been reviewed (Samonte and Eichler, 2002). Segmental duplications appear to be an ongoing process that has been active throughout recent primate evolution, with a growing list of primate evolutionary chromosomal rearrangements wherein segmental duplications have been involved or have been noted within the proximity of these events (Stankiewicz *et al*., 2001; Dennehey *et al*., 2004). Most of these analyses employed a combination of molecular biology, in silico, and cytogenetic techniques in characterizing paralogous copies in the specific regions of the genomes. Comparative primate FISH mapping validated genomic sequence movements associated with intrachromosomal rearrangements by determining the physical location and order of clones in chromosomal regions earlier grossly described by classical G-banding techniques. An example showing the progressive characterization of a chromosome rearrangement in the course of primate genome evolution is the telomeric fusion described by Yunis and Prakash (1982), involving integration of two ancestral chromosomes in a head-to-head fashion to form human chromosome 2. Subsequent inactivation of one of the two original chromosomes

Table 2 Human–great ape evolutionary chromosomal rearrangements associated with segmental duplications

Human chromosomal site involved	Estimated time of occurrence (million years ago, mya)	Size of duplicated region	Duplicons involved	Other duplications	References
2q13–2q14.1	<6 mya	>600 kb		9pter, 9p11.2, 9q13	Fan *et al*. 2002
15q11–q13	2–5 mya	~600 kb	LCR15, GLP, CHRNA7		Locke *et al*. 2003a
18q11	<6 mya	19 kb		18p11	Dennehey *et al*. 2004; Goidts *et al*. 2004
5q13.3 and 17p12	<8 mya		LCR17pA, LCR17pB		Stankiewicz *et al*. 2001

resulted in 46 chromosomes in the humans, whereas chimpanzee, gorilla, and orangutan have 48 chromosomes. The telomeric fusion was later confirmed by comparative FISH-mapping (Luke and Verma, 1992; Wienberg *et al*., 1994; Ijdo *et al*., 1991). It has recently been shown that >600 kb of sequence surrounding the fusion site consisted of highly homologous (96–99% sequence identity) sequence blocks that had been duplicated to several subtelomeric and pericentromeric regions of nonhomologous chromosomes prior to the fusion event (Fan *et al*., 2002). Segmental duplications may have possibly influenced the fusion of these two ancestral chromosomes producing the current architecture of human chromosome 2.

Large-scale differences among primate genomes have been grossly underestimated by karyotype analyses, and it has been suggested that copy number differences and qualitative differences in their distribution may be more significant for primate genome evolution than previously anticipated. In silico and comparative FISH mapping of a set of human chromosome 5–specific clones on several primate species showed an accumulation of segmental duplications in euchromatic regions and preferential duplicative transposition from the heterochromatic region of human chromosome 5 (Courseaux *et al*., 2003). The CMT1A-REP repeat, a pericentromeric segmental duplication responsible for a 1.5-Mb duplication and deletion associated with the Charcot-Marie-Tooth disease type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsies (HNPP), was present as a single copy in several primates screened by comparative FISH mapping, except for chimpanzees. This chimpanzee-specific duplication of an ancestral sequence occurred during primate evolution (Keller *et al*., 1999). A human-specific pericentric inversion in chromosome region 15q11-q13 characterized by comparative FISH mapping and sequence analyses showed involvement of at least two duplicons, CHRNA7 and GLP/LCR15 (Locke *et al*., 2003a).

As more low-copy repeat-induced chromosomal rearrangements are identified, the notion that genome evolution is influenced by both macro- as well as microrearrangements is strengthened. Comparison of human and great-ape genomes using array comparative genomic hybridization (array-CGH) identified 63 sites of putative copy-number variation in both euchromatic and heterochromatic regions (Locke *et al*., 2003b). Examination of only 12% of the genome showed large-scale differences of hundreds of kilobases, suggesting that microrearrangements are frequent between humans and great apes. Moreover, nearly 50% of validated copy-number variation mapped to regions of segmental duplications strongly suggest that a nonrandom distribution of such events. A recent genome-wide gene-based screen for gene duplication across five hominoid species has indicated that most of the copy number changes occurred in the past 15 million years, with the human species showing the most pronounced expansions (Fortna *et al*., 2004).

Paralogous sequences have consistently reshaped the primate genome, creating putative mosaic genes and causing rearrangements that have affected genomic stability. Regions that were involved in independent chromosome rearrangements during primate evolution were also prone to intragenomic duplication. One plausible explanation is that large-scale rearrangements and cytogenetically cryptic deletions and duplications result from the same genomic turnover processes.

7. Duplications in other species

The availability of a number of eukaryotic genome sequences makes it possible to examine patterns and types of gene duplication that have structured genomes resulting in species-specific genome architecture. As confirmed from comparative primate analyses, both micro- and macroarrangements have been the major force in primate genome evolution, with mammalian genomes described as mosaics composed of rearrangement-prone fragile regions and conserved solid regions (Pevzner and Tesler, 2003). Assessment of mouse and rat genome sequences for content, structure, and distribution of segmental duplications have provided the opportunity to compare the nature and pattern of these highly homologous sequences and to determine their role in shaping the mammalian genome. Such comparisons to orthologous segments in other genomes would also provide a means for establishing the relative order and timing of the duplications. Interestingly, both rodent genomes (mouse, 1.7–2%; rat, 2.92%) showed reduced amounts of segmental duplications than that of the human genome (5–6%) and instead, tandem or clustered duplications were more frequent. Such estimates are only preliminary, especially in light of the fact that these regions of the euchromatin are among the last to be adequately resolved (Eichler *et al*., 2004). Thus, the structure and precise composition of segmental duplications within the mouse and rat genomes remain to be determined.

Pericentromeric duplications are observed to be a common feature in the mouse and rat genomes (Tuzun, 2004; Bailey *et al*., 2004b; Thomas, 2003). In the mouse, this similar mosaic architecture of sequence modules generates chimeric transcripts. However, unlike the human genome, insertions and deletions within duplicated segments are more common in the mouse, while short repeated sequences flanking duplicons appear to be absent in the mouse genome. On the basis of the estimates of sequence divergence in the mouse genome, it appears that segmental duplications are relatively of recent origin (3–7 million years ago) and have occurred at several distinct periods. A model for the evolution of the pericentromeric regions of mouse chromosomes 5 and 6 have been proposed to involve fission of an ancestral chromosome, followed by the repair on each side of the break and creation of two new centromeres and chromosomes. Since all mammals as well as the chicken genome show proximity of segmental duplications to centromeres, it is likely that unusual aspects of centromere biology, including recombination differences and/or increased rates of interchromosomal gene conversion, play an important role in the biased distribution.

8. Summary

The human genome has been changing through a variety of mutational processes. Recent segmental duplication is one such process that appears to be relevant to our understanding of both human genome evolution and genetic disease. Assessment of other eukaryotic genomes will provide a broader view of the relationship of duplications to mechanisms of chromosome evolution and gene innovation. Understanding the origin and propagation of segmental duplications will provide fundamental insight into how genes and genomes change during the course of time.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Armour JA, Barton DE, Cockburn DJ and Taylor GR (2002) The detection of large deletions or duplications in genomic DNA. *Human Mutation*, **20**(5), 325–337.

Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, Jurka J and Morrow BE (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Research*, **13**(12), 2519–2532.

Bailey JA, Baertsch R, Kent WJ, Haussler D and Eichler EE (2004a) Hotspots of mammalian chromosomal evolution. *Genome Biology*, **5**, R23.

Bailey JA, Church DM, Ventura M, Rocchi M and Eichler EE (2004b) Analysis of segmental duplications and genome assembly in the mouse. *Genome Research*, **14**(5), 789–801.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW and Eichler EE (2002a) Recent segmental duplications in the human genome. *Science*, **297**(5583), 1003–1007.

Bailey JA, Liu G and Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics*, **73**(4), 823–834.

Bailey JA, Yavor AM, Massa HF, Trask BJ and Eichler EE (2001) Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research*, **11**(6), 1005–1017.

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M and Eichler EE (2002b) Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. *American Journal of Human Genetics*, **70**(1), 83–100.

Brewer C, Holloway S, Zawalnyski P, Schinzel A and FitzPatrick D (1999) A chromosomal duplication map of malformations: Regions of suspected haplo- and triplolethality – and tolerance of segmental aneuploidy – in humans. *American Journal of Human Genetics*, **64**(6), 1702–1708.

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, *et al*. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, **409**(6822), 953–958.

Christian SL, Fantes JA, Mewborn SK, Huang B and Ledbetter DH (1999) Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Human Molecular Genetics*, **8**(6), 1025–1037.

Conte RA, Samonte RV and Verma RS (1998) Evolutionary divergence of the oncogenes GLI, HST and INT2. *Heredity*, **81**(Pt 1), 10–13.

Courseaux A, Richard F, Grosgeorge J, Ortola C, Viale A, Turc-Carel C, Dutrillaux B, Gaudray P and Nahon JL (2003) Segmental duplications in euchromatic regions of human chromosome 5: A source of evolutionary instability and transcriptional innovation. *Genome Research*, **13**(3), 369–381.

Crosier M, Viggiano L, Guy J, Misceo D, Stones R, Wei W, Hearn T, Ventura M, Archidiacono N, Rocchi M, *et al*. (2002) Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. *Genome Research*, **12**(1), 67–80.

Dennehey BK, Gutches DG, McConkey EH and Krauter KS (2004) Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics*, **83**(3), 493–501.

Der-Sarkissian H, Vergnaud G, Borde YM, Thomas G and Londono-Vallejo JA (2002) Segmental polymorphisms in the proterminal regions of a subset of human chromosomes. *Genome Research*, **12**(11), 1673–1678.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al*. (1999) The DNA sequence of human chromosome 22. *Nature*, **402**(6761), 489–495.

Edelmann L, Spiteri E, McCain N, Goldberg R, Pandita RK, Duong S, Fox J, Blumenthal D, Lalani SR, Shaffer LG, *et al*. (1999) A common breakpoint on 11q23 in carriers of the constitutional t(11;22) translocation. *American Journal of Human Genetics*, **65**(6), 1608–1616.

Eichler EE (1998) Masquerading repeats: Paralogous pitfalls of the human genome. *Genome Research*, **8**(8), 758–762.

Eichler EE (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, **17**(11), 661–669.

Eichler EE, Archidiacono N and Rocchi M (1999) CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Research*, **9**(11), 1048–1058.

Eichler EE, Budarf ML, Rocchi M, Deaven LL, Doggett NA, Baldini A, Nelson DL and Mohrenweiser HW (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: A phenomenon of pericentromeric plasticity. *Human Molecular Genetics*, **6**(7), 991–1002.

Eichler EE, Clark RA and She X (2004) An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nature Reviews Genetics*, **5**(5), 345–354.

Emanuel BS and Shaikh TH (2001) Segmental duplications: An ‘expanding’ role in genomic instability and disease. *Nature Reviews Genetics*, **2**(10), 791–800.

Fan Y, Newman T, Linardopoulou E and Trask BJ (2002) Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13–2q14.1 and paralogous regions. *Genome Research*, **12**(11), 1663–1672.

Flint J, Bates GP, Clark K, Dorman A, Willingham D, Roe BA, Micklem G, Higgs DR and Louis EJ (1997) Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Human Molecular Genetics*, **6**(8), 1305–1313.

Footz TK, Brinkman-Mills P, Banting GS, Maier SA, Riazi MA, Bridgland L, Hu S, Birren B, Minoshima S, Shimizu N, *et al*. (2001) Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: A search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Research*, **11**(6), 1053–1070.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, *et al*. (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology*, **2**(7), E207.

Fritz B, Schubert F, Wrobel G, Schwaenen C, Wessendorf S, Nessling M, Korz C, Rieker RJ, Montgomery K, Kucherlapati R, *et al*. (2002) Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Research*, **62**(11), 2993–2998.

Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X, *et al*. (2003) Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Human Molecular Genetics*, **12**(8), 849–858.

Goidts V, Szamalek JM, Hameister H and Kehrer-Sawatzki H (2004) Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Human Genetics*, **115**, 116–122.

Guy J, Spalluto C, McMurray A, Hearn T, Crosier M, Viggiano L, Miolla V, Archidiacono N, Rocchi M, Scott C, *et al*. (2000) Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Human Molecular Genetics*, **9**(13), 2029–2042.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al*. (2000) The DNA sequence of human chromosome 21. *Nature*, **405**(6784), 311–319.

Horvath JE, Bailey JA, Locke DP and Eichler EE (2001) Lessons from the human genome: Transitions between euchromatin and heterochromatin. *Human Molecular Genetics*, **10**(20), 2215–2223.

Horvath JE, Gulden CL, Bailey JA, Yohn C, McPherson JD, Prescott A, Roe BA, de Jong PJ, Ventura M, Misceo D, *et al*. (2003) Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Molecular Biology and Evolution*, **20**(9), 1463–1479.

Horvath JE, Schwartz S and Eichler EE (2000) The mosaic structure of human pericentromeric DNA: A strategy for characterizing complex regions of the human genome. *Genome Research*, **10**(6), 839–852.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **36**(9), 949–951.

Ijdo JW, Baldini A, Ward DC, Reeders ST and Wells RA (1991) Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(20), 9051–9055.

Inoue K, Osaka H, Imaizumi K, Nezu A, Takanashi J, Arii J, Murayama K, Ono J, Kikawa Y, Mito T, *et al*. (1999) Proteolipid protein gene duplications causing Pelizaeus-Merzbacher disease: Molecular mechanism and phenotypic manifestations. *Annals of Neurology*, **45**(5), 624–632.

Jackson MS, Rocchi M, Thompson G, Hearn T, Crosier M, Guy J, Kirk D, Mulligan L, Ricco A, Piccininni S, *et al*. (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Human Molecular Genetics*, **8**(2), 205–215.

Ji Y, Eichler EE, Schwartz S and Nicholls RD (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Research*, **10**(5), 597–610.

Keller MP, Seifried BA and Chance PF (1999) Molecular evolution of the CMT1A-REP region: A human- and chimpanzee-specific repeat. *Molecular Biology and Evolution*, **16**(8), 1019–1026.

Knight SJ, Horsley SW, Regan R, Lawrie NM, Maher EJ, Cardy DL, Flint J and Kearney L (1997) Development and clinical application of an innovative fluorescence in situ hybridization technique which detects submicroscopic rearrangements involving telomeres. *European Journal of Human Genetics*, **5**(1), 1–8.

Linardopoulou E, Mefford HC, Nguyen O, Friedman C, van den Engh G, Farwell DG, Coltrera M and Trask BJ (2001) Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location. *Human Molecular Genetics*, **10**(21), 2373–2383.

Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, Roe B, Rocchi M and Eichler EE (2003a) Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biology*, **4**(8), R50.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D and Eichler EE (2003b) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Research*, **13**(3), 347–357.

Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG and Eichler EE (2004) BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *Journal of Medical Genetics*, **41**(3), 175–182.

Loftus BJ, Kim UJ, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, *et al*. (1999) Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics*, **60**(3), 295–308.

Luke S and Verma RS (1992) Origin of human chromosome 2. *Nature Genetics*, **2**(1), 11–12.

Lupski JR (1998) Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, **14**(10), 417–422.

Mantripragada KK, Buckley PG, Benetkiewicz M, De Bustos C, Hirvela C, Jarbo C, Bruder CE, Wensman H, Mathiesen T, Nyberg G, *et al*. (2003) High-resolution profiling of an 11 Mb segment of human chromosome 22 in sporadic schwannoma using array-CGH. *International Journal of Oncology*, **22**(3), 615–622.

Martin CL, Wong A, Gross A, Chung J, Fantes JA and Ledbetter DH (2002) The evolutionary origin of human subtelomeric homologies – or where the ends begin. *American Journal of Human Genetics*, **70**(4), 972–984.

Mefford HC and Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. *Nature Reviews. Genetics*, **3**(2), 91–102.

Mefford HC, Linardopoulou E, Coil D, van den Engh G and Trask BJ (2001) Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Human Molecular Genetics*, **10**(21), 2363–2372.

Ohno S (1970) *Evolution by gene duplication*. London: George Allen and Unwin, 160.

Ohno S, Wolf U and Atkin NB (1968) Evolution from fish to mammals by gene duplication. *Hereditas*, **59**, 169–187.

Orti R, Potier MC, Maunoury C, Prieur M, Creau N and Delabar JM (1998) Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. *Cytogenetics and Cell Genetics*, **83**(3–4), 262–265.

Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, *et al*. (2001) A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics*, **29**(3), 321–325.

Otto SP and Yong P (2002) The evolution of gene duplicates. *Advances in Genetics*, **46**, 451–483.

Pagel J, Walling JG, Young ND, Shoemaker RC and Jackson SA (2004) Segmental duplications within the glycine max genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes. *Genome*, **47**(4), 764–768.

Pevzner P and Tesler G (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 7672–7677.

Ravise N, Dubourg O, Tardieu S, Aurias F, Mercadiel M, Coullin P, Ruberg M, Catala M, Lesourd S, Brice A, *et al*. (2003) Rapid detection of 17p11.2 rearrangements by FISH without cell culture (direct FISH, DFISH): A prospective study of 130 patients with inherited peripheral neuropathies. *American Journal of Medical Genetics*, **118A**(1), 43–48.

Regnier V, Meddeb M, Lecointre G, Richard F, Duverger A, Nguyen VC, Dutrillaux B, Bernheim A and Danglot G (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Human Molecular Genetics*, **6**(1), 9–16.

Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, Mudunuri U, Paul S and Wei J (2004) Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Research*, **14**(1), 18–28.

Rosenberg MJ, Killoran C, Dziadzio L, Chang S, Stone DL, Meck J, Aughton D, Bird LM, Bodurtha J, Cassidy SB, *et al*. (2001) Scanning for telomeric deletions and duplications and uniparental disomy using genetic markers in 120 children with malformations. *Human Genetics*, **109**(3), 311–318.

Ruault M, Trichet V, Gimenez S, Boyle S, Gardiner K, Rolland M, Roizes G and De Sario A (1999) Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene*, **239**(1), 55–64.

Samonte RV and Eichler EE (2002) Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics*, **3**(1), 65–72.

Samonte RV, Conte RA, Ramesh KH and Verma RS (1996) Molecular cytogenetic characterization of breakpoints involving pericentric inversions of human chromosome 9. *Human Genetics*, **98**(5), 576–580.

Samonte RV, Conte RA and Verma RS (1998) Molecular phylogenetics of the hominoid Y chromosome. *Journal of Human Genetics*, **43**(3), 185–186.

Selker EU (1999) Gene silencing: Repeats that count. *Cell*, **97**(2), 157–160.

Shaw CJ, Shaw CA, Yu W, Stankiewicz P, White LD, Beaudet AL and Lupski JR (2004) Comparative genomic hybridisation using a proximal 17p BAC/PAC array detects rearrangements responsible for four genomic disorders. *Journal of Medical Genetics*, **41**(2), 113–119.

She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, *et al*. (2004a) The structure and evolution of centromeric transition regions within the human genome. *Nature*, **430**(7002), 857–864.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL and Eichler EE (2004b) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.

Stankiewicz P, Park SS, Inoue K and Lupski JR (2001) The evolutionary chromosome translocation 4;19 in Gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Research*, **11**(7), 1205–1210.

Thomas JW, Schueler MG, Summers TJ, Blakesley RW, McDowell JC, Thomas PJ, Idol JR, Maduro VV, Lee-Lin SQ, Touchman JW, *et al*. (2003) Pericentromeric duplications in the laboratory mouse. *Genome Research*, **13**(1), 55–63.

Trask BJ (2002) Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Reviews Genetics*, **3**(10), 769–778.

Tuzun E, Bailey JA and Eichler EE (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Research*, **14**(4), 493–506.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Wienberg J, Jauch A, Ludecke HJ, Senger G, Horsthemke B, Claussen U, Cremer T, Arnold N and Lengauer C (1994) The origin of human chromosome 2 analyzed by comparative chromosome mapping with a DNA microlibrary. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, **2**(5), 405–410.

Wilhelm M, Veltman JA, Olshen AB, Jain AN, Moore DH, Presti JC Jr., Kovacs G and Waldman FM (2002) Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Research*, **62**(4), 957–960.

Yunis JJ and Prakash O (1982) The origin of man: A chromosomal pictorial legacy. *Science*, **215**(4539), 1525–1530.

Zimonjic DB, Kelley MJ, Rubin JS, Aaronson SA and Popescu NC (1997) Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(21), 11461–11465.

Ziolkowski PA, Blanc G and Sadowski J (2003) Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome. *Nucleic Acids Research*, **31**(4), 1339–1350.

Noncoding RNAs in mammals

Timothy Ravasi and David A. Hume
University of Queensland, Queensland, Australia

1. Introduction

As the genome-sequencing revolution has progressed through more complex organisms, the first outcome of each new genome sequence has been an estimated gene count. What appears to be an obsession with enumeration has largely ignored the question of the definition of a gene. From a geneticist's viewpoint, a gene is a unit of function that is defined by a mutant phenotype. On that basis, a gene can be defined as the minimal segment of DNA that is able to complement a mutant phenotype. In prokaryotes, and simple eukaryotes, the boundaries of a gene, a segment of DNA that must include the control elements in order to provide complete complementation, are relatively easy to define. As we have climbed up the level of genomic complexity to plants and animals, this definition of a gene has become untenable. Saturation mutagenesis of a mammalian genome is clearly a long way off and were it possible, analysis of the precise nature of each lethal phenotype is a major project and nonlethal mutations with no apparent phenotype present even greater challenges. As a consequence, genome scientists have accepted the de facto definition that a gene is the segment of DNA that encodes a protein; and in doing so have accepted the view that the only functional output of the genome (since gene is a functional definition) is a protein. Indeed, a focused effort by the NIH to produce a mammalian gene collection is directed exclusively to the identification of cDNAs that contain a complete open-reading frame (Baross *et al*., 2004), and a recent large-scale expression dataset in mouse and human focuses solely on protein-coding transcripts (Su *et al*., 2004). Even this approach leaves aside the major impact of alternative splicing on the proteome, and begs the question of whether a distinct cDNA encoding a different form of a protein, controlled by a distinct promoter and expressed in a distinct tissue, constitutes a separate gene.

It was recognized at a very early stage in genome-sequencing projects that the sequence could not be adequately annotated without extensive analysis of the transcriptome, the transcriptional output of the genome. There have been several different approaches to the analysis of transcriptional output. High-throughput approaches such as expressed sequence tag (EST) sequencing, and SAGE, followed by computational assembly into contigs (e.g., in UniGene/TIGR Gene Index) have generated large estimates of the number of separate transcripts that can be

encoded by a mammalian genome (Liang *et al*., 2000; Quackenbush *et al*., 2000; Quackenbush *et al*., 2001; Fahrenkrug *et al*., 2002; Lee *et al*., 2002; Sonstegard *et al*., 2002). The outputs have been useful in identifying putative exons in genome sequence, but do not help in identifying genuine contiguous transcripts, since one can have no confidence that the segments brought together in silico actually coexist in a real transcript. In part to deal with this problem, and in part to generate full-length cDNA physical clones, a number of groups attacked the task of isolating and sequencing a complete collection of the cDNAs representing the mammalian transcriptome (Kawai *et al*., 2001; Okazaki *et al*., 2002; Dalla *et al*., 2003; Ota *et al*., 2004). This is a greater challenge than the genome, since each individual cDNA must be generated from a particular tissue source of mRNA. Even taking account of the uncertainties with annotating genuine coding regions within the sequenced cDNAs, and problems with cloning artifacts leading to cDNA truncation at the 5′ or 3′ end, it became evident from the analysis of the first large sets (Kawai *et al*., 2001) that even amongst polyadenylated "mRNAs" there is a very significant component of full-length transcripts that do not encode a protein. As the systematic cloning of full-length cDNAs has progressed, and focused on less-abundant transcripts, the size of the validated noncoding RNA component of the transcriptome has continued to grow (Okazaki *et al*., 2002; Shabalina and Spiridonov, 2004). The existence of noncoding RNAs has been further validated with the appearance of genome-tiling microarrays (Kampa *et al*., 2004; Bertone *et al*., 2004). The isolated, or predicted, full-length cDNAs have focused mainly on polyA-plus mRNA and the size of transcripts that have been isolated and fully sequenced has been constrained by the ability to generate full-length cDNAs, to clone them into sequencing vectors, and to sequence them. A recent study using high-density genome-tiling arrays, and separately analyzing nuclear and cytoplasmic RNA, and polyA-plus and polyA-minus RNA populations, has revealed that up to 14.6% of the genome is transcribed in a single human cell type (Gingeras T, personal communication).

What does it all mean? Is the "junk DNA" transcribed into "junk RNA", or do all these transcriptional outputs of the genome actually have a function? In this review, we will present an analysis of the possible classes of noncoding RNAs that have been identified, overview evidence that their expression is regulated, and discuss some of the possible functions. The review will focus especially on the data from the mouse, the mammalian species in which the transcriptome has been most extensively sampled.

2. How many noncoding RNAs are there?

For the purpose of this review, we will take as a given that a very large fraction of the animal genome, from drosophila to man, is actively transcribed, in many cases from both strands. (Eddy, 2001; Shabalina and Spiridonov, 2004; Kampa *et al*., 2004; Stolc *et al*., 2004; Bertone *et al*., 2004). Combined with full-length and EST sequencing efforts, the evidence indicates that 50% or more of the transcripts generated do not contain an open-reading frame that is likely to be translated into a protein. Whereas the size of the predicted protein-coding transcriptome is

unlikely to grow a great deal beyond the current estimate of 30 000 “genes” in mammals, since the size can be inferred with some confidence from comparisons of completed genomes, the size of the predicted ncRNA (noncoding RNA) component will probably continue to grow as the depth of coverage increases. The ncRNAs that have been identified are generally less abundant than protein-coding RNAs. In a recent global analysis in drosophila using genome-tiling arrays (Stolc *et al*., 2004), protein-coding transcripts and genes associated with a known mutant phenotype were more highly expressed and 76% were differentially expressed at specific developmental stages. Conversely, putative ncRNAs, whether derived from intergenic or intragenic regions, were generally expressed at lower levels and only 15% showed developmental regulation. This observation cannot imply that any putative ncRNA lacks a function. A single molecule of RNA could be all that is required to act in cis in control of transcription. It says simply that they tend to be less abundant, and more will be identified as one digs deeper into expressed sequences. However, it does suggest that with depth of coverage one can drift into the territory of what skeptics would call “transcriptional noise”, at the extreme where the RNA is actually present at less than copy per cell but is still detectable.

Although in situ transcription assays are possible, and might eventually be used to complement array-based approaches, the detection of the transcriptional output of the genome generally requires the synthesis of cDNA as an intermediate, either for generation of probes to use in microarrays, or for cloning and sequencing. In undertaking any such analysis, the outcome is dependent upon the method used to purify the RNA, and the approach to cDNA synthesis. If the intent is to identify mature, processed transcripts that find their way into the cytoplasm, as opposed to primary nuclear transcripts, they can be enriched by stringent separation of cytoplasmic from nuclear RNA, and by purification using oligo-dT affinity chromatography to capture polyadenylated mRNA. Separation of nucleus from cytoplasm cannot be 100% efficient, especially from solid tissues rather than cell lines, and one can have no confidence of excluding leakage of nuclear transcripts into the “cytoplasm” when cells are lysed. If there is any systematic attempt to normalize the mRNA pool by subtraction to reduce the incidence of abundant transcripts, the inevitable consequence is that nuclear “contamination” will be enriched in the process.

The issue of nuclear contamination raises the issue of the extent of representation of partially processed nuclear RNAs in the putative ncRNA pool, and in turn whether these should be excluded from the ncRNA definition. There is a tacit assumption in all analysis of the transcriptome that intronic sequences are removed from the primary transcript by splicing, rapidly decay, and are unlikely to make a significant contribution to detectable outputs of the transcriptome. By extension, spliced introns per se are presumed to have no function and are not considered in analyzing the scale of the transcriptome. However, this assumption has seldom been systematically tested. In a landmark study, Clement *et al*. (2001) detected all three spliced introns from the coding region of the Pem homeobox gene in total RNA, and found that the half-lives of these Pem introns ranged from 9 to 29 min, which is in the same range as short-lived mRNAs such as those encoding c-fos and c-myc. Even more challenging to the prevailing dogma, all three of the spliced Pem introns were present in the cytoplasmic fraction. Although introns are regarded generally

as junk, these findings provide some support to the view that subsets of intron-derived RNAs excised by splicing can accumulate to significant levels, and could form part of a transcriptional control network (Mattick, 2001). A detailed study of the imprinted GNAS imprinted locus in the mouse also confirms that a number of transcripts with retained introns accumulate, and their expression is selectively controlled from one or other selectively imprinted parental allele (Holmes *et al*., 2003). If subsets of spliced introns or partly processed (or partly degraded) primary transcripts do accumulate selectively in the nucleus or cytoplasm, they will have been sampled only to a limited extent in transcriptome analysis based upon oligo-dT purification or priming, since this will sample only those that fortuitously include an A-rich string. Furthermore, they will not contain a consensus polyadenylation sequence nor will they be flanked by a classical promoter. The set of putative ncRNAs in the RIKEN mouse transcriptome set certainly contains a substantial component of such transcripts, which are routinely annotated as "immature". But where they have been tested, they appear to be reproducibly represented in purified RNA pools, and are therefore unlikely to represent fortuitous cloning of a partially spliced primary RNA.

Leaving aside the contribution of intron-derived RNAs to the detected transcriptome, at least some of the ncRNAs represented in cDNA collections and ESTs probably arise from the incomplete capture of a genuine full-length cDNA that lacks the complete open-reading frame. Even with the technical advances in full-length cDNA synthesis and capture of the 5′ end, the generation of a full-length cDNA requires reverse transcriptase to initiate at the 3′ end, commonly from an oligo-dT primer, and to extend through secondary structures in the RNA to a genuine 5′ end. On the basis of the capture of full-length coding regions of known protein-coding transcripts, the best outcomes for capture of the 5′ end are in the range of 70–80% full length. That is, the 5′ end of 70–80% of cDNAs in the best case represents the genuine transcription start site. Subsets of putative ncRNAs, lacking a predicted CDS, are annotated as "partial" based upon sequence overlap with another longer cDNA, or with an EST assembly. In no case can this be considered definitive. A subset of these ncRNAs with apparently truncated 5′ ends arise from internal promoters, and we cannot infer that they have no function solely because they lack the ORF of the corresponding "full-length" transcript.

Recognition of the potential "artifacts" associated with cytoplasmic RNA purification and cDNA production leads annotators to try to subdivide the putative ncRNAs based upon other features. A sizable subset of the RIKEN full-length mouse transcripts annotated as noncoding are bounded by consensus polyadenylation sequences at the 3′ end, and by consensus promoter regions at the 5′ end (e.g., CpG islands). These ncRNA are considered as mRNA-like, and this set contains many of the known functional ncRNAs (Numata *et al*., 2003). Unfortunately, the absence of these features says nothing about whether a putative ncRNA is "real" or a cloning artifact. We know so little about the function and regulation of ncRNAs that we cannot exclude the possibility that they are not capped in a conventional manner (e.g., like rRNAs), or that they utilize alternative transcriptional termination mechanisms (Eddy, 2001). If the ncRNAs that lack these conventional mRNA-like features prove also to be genuine transcriptional outputs, then their representation

in current cDNA collections will again be very low, since the current strategies for cDNA cloning are designed to exclude them.

In overview, our current knowledge does not permit us to distinguish cloning artifact from possible ncRNA based upon sequence analysis. However, many of the putative ncRNAs are strongly supported by expression analysis. In a recent study (Ravasi *et al*., 2005), we showed using RT-PCR, Northern blots, and microarray analysis that the large majority of the ncRNAs in the RIKEN clone collection are indeed reproducibly expressed, and many show tissue-specific expression. A subset is acutely regulated in macrophages stimulated with lipolysaccharide (LPS) (Ravasi *et al*., 2005), and the promoters directing expression of these transcripts show features in common with the promoters of LPS-stimulated protein-coding genes (unpublished).

If some of the existing putative cloned noncoding cDNAs are artifacts, or the corresponding transcripts are expressed as such low levels as to be best classified as transcriptional noise, the functional ncRNA output of the mammalian genome may be lower than what the current figures suggest. The expression analysis suggests this is not the case, and the genome-tiling data supports the view, supported by all the arguments above, that the number of ncRNAs encoded reproducibly by a mammalian transcriptome is actually very much larger than the current estimates.

3. The functions of noncoding RNAs

3.1. Mutations

It is inherently impossible to prove the absence of function, but skeptics about the function of ncRNA point to the absence of phenotypes associated with mutations. The number of non-protein-coding RNAs that have been functionally validated by mutagenesis screens in any eukaryote is small. The specific examples associated with mammalian imprinting will be discussed below. There are two well-cited examples in humans of genetic diseases associated specifically with loss of a noncoding RNA but these are actually RNA components of functional complexes (Eddy, 2001; Shabalina and Spiridonov, 2004). Conversely, in the ROSA-26 locus in mice, there are two highly expressed ncRNA transcripts convergent with a protein-coding transcript. Homozygous mutations of the ROSA-26 selectively ablates both ncRNAs, and mutants are born at less than the expected Mendelian ration, but there is no other phenotype evident, not even altered expression of the convergent-coding transcript (Zambrowicz *et al*., 1997). The majority of natural or chemically induced mutagenesis in mammals causes point mutations or deletions. A single deletion or mutation can, of course, radically disrupt protein-coding capacity of the encoded transcript but might have much less effect on the function of a ncRNA (Eddy, 2001). If ncRNAs do have a function, additional evidence that it does not require stringent sequence conservation comes from the fact that the large majority of identified ncRNAs are weakly conserved between mammalian species (Numata *et al*., 2003). So, function is likely to be inferred only by complete ablation of the transcript. In organisms such as drosophila, which also

produce many ncRNAs, the major approach to mutagenesis is P-element transposon insertion, yet even in this organism few phenotypes have been associated with ablation of an ncRNA. Given the possibility of high-throughput RNAi screens in this species (Boutros *et al*., 2004), it will certainly be feasible in future to assess whether any putative ncRNAs have nonredundant functions. Redundancy is obviously a secondary question. Many protein-coding genes, especially those that form part of multigene families, are redundant inasmuch as mutation has no apparent phenotype. Without knowing their functions, we cannot eliminate the possibility that ncRNAs form families containing short functional motifs, or with interdependent regulation/compensation, so that no single ablation abolishes the function of the ncRNA family. Support for this view comes from a recent global clustering analysis of putative ncRNA molecules, evocatively entitled "into the heart of darkness" (Bejerano *et al*., 2004). This study of the human genome identified 12 000 nonsingleton clusters dense in significant sequence similarities.

3.2. The transcriptional noise view

A genome-tiling analysis of sites on human chromosomes bound by the general transcription factor Sp1 revealed numerous binding sites associated with the apparent initiation of noncoding RNAs (Kampa *et al*., 2004). Sp1 sites are associated with transcription initiation in TATA-less promoters, and a significant proportion of the ncRNAs in the mouse transcriptome originate from CpG islands (Numata *et al*., 2003). A skeptical view is that the CG-rich Sp1 recognition motif occurs frequently in genomic sequence in CpG islands and with some measure of randomness binds Sp1 and initiates the production of transcripts that represent background noise; in essence, just a reflection of the large size of the mammalian genome. An extension of this view is that many such sites occur within Alu repeats, and other ancestral retrotransposons that retain cryptic, or low level, intrinsic promoter activity.

True randomness is an unlikely scenario, because the majorities of Sp1 sites in genomic DNA, and of repeat elements, are assembled into inactive chromatin, and are methylated and/or inaccessible to transcription factor binding. Transcriptionally active regions of the genome are associated with acetylated histones, and are sensitive to digestion by DNase (DNase hypersensitive sites). A more plausible origin for many ncRNAs lies within specific enhancers that occur in 5′ or 3′ flanking regions or within the introns of most protein-coding "genes". These sequences are commonly conserved across mammalian species. There are numerous examples in the literature of enhancers acting as promoters, not surprisingly since they commonly bind the same sets of transcriptional activators as conventional promoters. For example, in the globin locus, a distal enhancer initiates transcription of a long transcript that extends through to its downstream promoter elements. Ling *et al*. (2004) showed recently that interruption of this transcript greatly reduces the activity of the linked downstream promoter. Cook (2003) put forward the interesting idea that transcription arising from enhancers acts to focus the enhancer and linked promoter regions into the active "transcription factories" within the nucleus, thereby facilitating the looping out of intervening sequences

and concentrating the transcription regulators around the transcription start site of the target promoter. Since enhancers commonly bind tissue-restricted transcription factors, the production of ncRNAs from such elements will display the same regulation as the transcripts that are controlled by the enhancer. In the global study of drosophila transcription, expression of putative ncRNAs correlates significantly with expression of the closest annotated exons (Stolc *et al*., 2004). In some cases, this may mean that the putative ncRNAs are simply unannotated exons of the adjacent gene. But equally, this pattern would occur if the enhancers that control expression of the adjacent genes also direct transcription of ncRNAs.

4. Possible functions of ncRNAs

Broadly speaking, ncRNAs might function as catalytic RNAs, through their ability to base-pair with DNA or RNA of related sequence (or through triplex formation), or by binding to proteins. Some of the proposed and validated functions are discussed below. We will not deal in detail here with microRNAs and RNA interference, which has been reviewed elsewhere (Eddy, 2001; Mallory and Vaucheret, 2004).

4.1. Imprinting, silencing, and activation: cis-acting effects of ncRNAs

The archetypal functional ncRNA in mammals is *xist*, which functions in cis to initiate global chromatin remodeling and silencing on the X-chromosome (Brown and Chow, 2003; Ganesan *et al*., 2004). Precisely how this function is mediated remains unclear. The cis-acting function of ncRNAs in gene silencing is also evident in a number of imprinted loci. The best-characterized example is the IGF2R receptor locus, in which a very large antisense transcript is implicated in allele-specific transcription control. Ablation of the promoter driving this transcript (AIR) was found to abolish imprinting and in a separate transgenic manipulation, the role of antisense per se was excluded by directing production of AIR from an alternative promoter that does not overlap the IGF2R promoter (Sleutels *et al*., 2002; Sleutels *et al*., 2003). A subtle variant of this theme is seen in the ube3A locus, in which there is brain-specific silencing of the paternal allele that appears to be dependent upon a brain-specific antisense ncRNA transcript (Landers *et al*., 2004). There are at least 70 known imprinted loci in the mouse genome, and probably many more (Nikaido *et al*., 2003), and most express well-validated, imprinted, noncoding RNA transcripts (Goldmit and Bergman, 2004; Peters and Beechey, 2004). Although the X-chromosome is the only chromosome showing global monoallelic inactivation, the same phenomenon occurs in many regions of autosomes. The best characterized examples include odorant and pheromone receptors, immunoglobulin and T-cell receptors and many other receptors on the surface of immunocompetent cells, and many inducible lymphokines and cytokines (Chess *et al*., 1994; Mostoslavsky *et al*., 1998; Singh *et al*., 2003a,b; Ensminger and Chess, 2004a,b; Gimelbrant *et al*., 2005). The phenomenon is not restricted to inducible genes, or sensors

in the nervous and immune systems. Widespread evidence of asymmetric DNA replication, a feature of the X-chromosome, on regions of the autosomes, suggests that random and/or regulated allelic inactivation is also widespread. Recently, Gimelbrant *et al*. (2005) used this criterion to identify many additional genes that show this feature, and highlighted the fact that the widely expressed p120 beta-catenin gene can be expressed monoallelically, or biallelically, in different cell types. One example of allelic inactivation and asymmetric replication that was studied in detail was the tlr4 locus, which encodes a molecule needed for the recognition of bacterial lipopolysaccharide by macrophages. In mice that carry a nonfunctional mutant allele caused by a point mutation of tlr4, 50% of macrophages express the mutant allele and, consequently, are unresponsive to LPS. However, in heterozygotes in which one tlr4 allele is completely ablated, all macrophages express the wild-type tlr4 (Pereira *et al*., 2003). This finding indicates that there is an allele-counting mechanism, as in X-chromosome inactivation, that ensures that only one allele is active. In all probability, this counting mechanism also involves ncRNA.

In view of these many examples of cis-acting effects of ncRNA, one cannot dismiss the transcripts derived from enhancers as noise. Such ncRNA might act to interrupt the chromosome-wide, or local, cis-acting effects of negative regulators. In the case of X-chromosome inactivation, the antisense gene *Tsix* determines X-chromosome choice and represses the noncoding silencer, *Xist.* In principle, Tsix action may involve RNA, the act of transcription, or local chromatin. Recent studies concluded that the processed antisense RNA does not act alone and that Tsix function specifically requires antiparallel transcription through *Xist* (Heard, 2004; Shibata and Lee, 2004). This example could provide a model for the actions of the many antisense transcripts that have been detected in large cDNA collections, and which may be derived from intronic or 3′ enhancer-like elements.

4.2. Trans-acting effects of ncRNA

With the rapid advances in RNAi technology, much of the focus on actions of ncRNA has been directed toward RNA interference. Many microRNAs are derived by processing of transcribed introns (Eddy, 2001), and there is an obvious possibility that many sense–antisense pairs identified in the mouse transcriptome (Kiyosawa *et al*., 2003) might function in a regulatory loop. A recent paper highlighted the identification of large numbers of intronic antisense transcripts in human cells, and correlated their levels with the state of differentiation of prostate cancer cells (Reis *et al*., 2004). The existence of large numbers of such transcripts has been confirmed in a global study of the human genome using tiling arrays (Bertone *et al*., 2004). It is not necessarily the case that such functional pairs will always show discordant regulation. In fact, in the expression analyses that we have performed in activated macrophages, sense and antisense pairs show no absolute trend toward concordant or discordant regulation, if anything more commonly showing concordance (unpublished). From a mechanistic viewpoint, antisense transcription could act to interfere with transcription initiation (directing transcriptional silencing), through transcription termination (it is not at all clear

what happens if transcription occurs on both strands simultaneously), RNA decay, or translational interference, but there is little direct evidence for any of these mechanisms in mammals as a major regulatory mechanism.

The other possible mechanism of action of ncRNA lies in their ability to bind to proteins. In the specific case of the small number of transcribed pseudogenes, they may act to regulate the translation or stability of the corresponding protein-coding transcript. At last one specific example has been documented (Hirotsune *et al*., 2003). The same mechanism might arise via the transcription of truncated noncoding transcripts. The noncoding imprinted H19 RNA associated with specific posttranscriptional stabilization of thioredoxin by an unknown mechanism (Lottin *et al*., 2002). RNA molecules can also bind in a sequence-specific manner to DNA-binding proteins. Members of the large C2H2 zinc finger transcription factor family, including the ubiquitous Sp1 transcription factor, bind with high affinity to DNA–RNA duplexes (Shi and Berg, 1995) and some transcription factors, such as Wt1, can also bind single-stranded RNA (Zhai *et al*., 2001). The capacity of RNA to bind to regulatory proteins including transcription factors in a sequence-specific manner provides many possible avenues for participation in regulatory networks.

Non-protein-coding RNAs clearly exist as reproducible transcriptional output of the genome. Views on the functions of ncRNAs range from the skeptical or dismissive, to the big picture perspective, which sees them as an essential component of a control network that is essential in the leap in complexity from prokaryotes to eukaryotes (Mattick, 2001). The latter view is theoretically attractive, but difficult to test experimentally. The addition of ncRNA to our view of the transcriptome provides a challenge to the perspective of the mammalian genome as being composed of genes and intergenic regions. To describe the set of transcripts encoded by the mouse genome, the FANTOM consortium adopted a definition of a transcriptional unit (TU) a cluster of transcripts with overlapping sequence bounded by the most extreme transcription start point and termination sequence. As more ncRNA are identified, these boundaries have extended, the number of TU has collapsed, and the number of TU with partial or complete antisense TU has increased. The boundaries of "genes" defy definition; we have found ourselves dealing with transcript forests separated by transcript deserts. A consequence of this pattern is that the numbers of ncRNAs that have actually already been knocked out in the mouse genome at least, and probably also in drosophila, is much greater than one might have recognized. Many knockout constructs or P-element insertions will have simultaneously ablated overlapping sense or antisense ncRNAs, which could contribute to a phenotype (or the lack of a phenotype) either by altering the function or expression of neighboring genes in cis, or other genes/alleles in trans. In the future, a detailed knowledge of the full transcriptional output of a locus will be an essential part of the interpretation of any functional genomic study in eukaryotes.

Acknowledgments

This work was funded by grants from the National Health and Medical Research Council of Australia (NHMRC) to D.A.H.

References

Baross A, Butterfield YS, Coughlin SM, Zeng T, Griffith M, Griffith OL, Petrescu AS, Smailus DE, Khattra J and McDonald HL (2004) Systematic recovery and analysis of full-ORF human cDNA clones. *Genome Research*, **14**(10B), 2083–2092.

Bejerano G, Haussler D and Blanchette M (2004) Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics*, **20**(Suppl 1), I40–I48.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Heidelberg Fly Array Consortium, Paro R and Perrimon N (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science*, **303**(5659), 832–835.

Brown CJ and Chow JC (2003) Beyond sense: The role of antisense RNA in controlling Xist expression. *Seminars in Cell and Developmental Biology*, **14**(6), 341–347.

Chess A, Simon I, Cedar H and Axel R (1994) Allelic inactivation regulates olfactory receptor gene expression. *Cell*, **78**(5), 823–834.

Clement JQ, Maiti S and Wilkinson MF (2001) Localization and stability of introns spliced from the Pem homeobox gene. *The Journal of Biological Chemistry*, **276**(20), 16919–16930.

Cook PR (2003) Nongenic transcription, gene regulation and action at a distance. *Journal of Cell Science*, **116**(Pt 22), 4483–4491.

Dalla E, Verardo R, Lazarevi D, Marchionni L, Reid JF, Bahar N, Klari E, Marcuzzi G, Marzio R, Belgrano A, *et al*. (2003) LNCIB human full-length cDNAs collection: Towards a better comprehension of the human transcriptome. *Comptes Rendus Biologies*, **326**(10–11), 967–970.

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nature Reviews. Genetics*, **2**(12), 919–929.

Ensminger AW and Chess A (2004a) Bidirectional promoters regulate the monoallelically expressed Ly49 NK receptors. *Immunity*, **21**(1), 2–3.

Ensminger AW and Chess A (2004b) Coordinated replication timing of monoallelically expressed genes along human autosomes. *Human Molecular Genetics*, **13**(6), 651–658.

Fahrenkrug SC, Smith TP, Smith BA, Freking BA, *et al*. (2002) Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mammalian Genome*, **13**(8), 475–478.

Ganesan S, Silver DP, Drapkin R, Greenberg R, Feunteun J and Livingston DM (2004) Association of BRCA1 with the inactive X chromosome and XIST RNA. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **359**(1441), 123–128.

Gimelbrant AA, Ensminger AW, Qi P, Zucker J and Chess A (2005) Monoallelic expression and asynchronous replication of p120 catenin in mouse and human cells. *The Journal of Biological Chemistry*, **280**, 1354–1359.

Goldmit M and Bergman Y (2004) Monoallelic gene expression: A repertoire of recurrent themes. *Immunological Reviews*, **200**, 197–214.

Heard E (2004) Recent advances in X-chromosome inactivation. *Current Opinion in Cell Biology*, **16**(3), 247–255.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K-i, Wynshaw-boris A and Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**(6935), 91–96.

Holmes R, Williamson C, Peters J, Denny P, R GER Group, GSL Members, and Wells C (2003) A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Research*, **13**(6B), 1410–1415.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, **14**(3), 331–342.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, *et al*. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**(6821), 685–690.

Kiyosawa H, Yamanaka I, Osato N, Kondo S and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*, **13**(6B), 1324–1334.

Landers M, Bancescu DL, Le Meur E, Rougeulle C, Glatt-Deeley H, Brannan C, Muscatelli F and Lalande M (2004) Regulation of the large (approximately 1000 kb) imprinted murine Ube3a antisense transcript by alternative exons upstream of Snurf/Snrpn. *Nucleic Acids Research*, **32**(11), 3480–3492.

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, *et al*. (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research*, **12**(3), 493–502.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL and Quackenbush J (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics*, **25**(2), 239–240.

Ling J, Ainol L, Zhang L, Yu X, Pi W and Tuan D (2004) HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *The Journal of Biological Chemistry*, **279**, 51704–51713.

Lottin S, Vercoutter-Edouart AS, Adriaenssens E, Czeszak X, Lemoine J, Roudbaraki M, Coll J, Hondermarck H, Dugimont T and Curgy JJ (2002) Thioredoxin post-transcriptional regulation by H19 provides a new function to mRNA-like non-coding RNA. *Oncogene*, **21**(10), 1625–1631.

Mallory AC and Vaucheret H (2004) MicroRNAs: Something important between the genes. *Current Opinion in Plant Biology*, **7**(2), 120–125.

Mattick JS (2001) Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Reports*, **2**(11), 986–991.

Mostoslavsky R, Singh N, Kirillov A, Pelanda R, Cedar H, Chess A and Bergman Y (1998) Kappa chain monoallelic demethylation and the establishment of allelic exclusion. *Genes and Development*, **12**(12), 1801–1811.

Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M, RIKEN GER Group, GSL Members, *et al*. (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Research*, **13**(6B), 1402–1409.

Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, RIKEN GER Group, GSL Members, Hayashizaki Y, and Tomita M (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Research*, **13**(6B), 1301–1306.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**(6915), 563–573.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics*, **36**(1), 40–45.

Pereira JP, Girard R, Chaby R, Cumano A and Vieira P (2003) Monoallelic expression of the murine gene encoding Toll-like receptor 4. *Nature Immunology*, **4**(5), 464–470.

Peters J and Beechey C (2004) Identification and characterisation of imprinted genes in the mouse. *Briefings in Functional Genomics and Proteomics*, **2**(4), 320–333.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**(1), 159–164.

Quackenbush J, Liang F, Holt I, Pertea G and Upton J (2000) The TIGR gene indices: Reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, **28**(1), 141–145.

Ravasi T, Suzuki H, *et al*. (2005) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Proceedings of the National Academy of Sciences of the United States of America*, in press.

Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, Egidio CM, Paquola AC, Machado AA, Festa F, *et al*. (2004) Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene*, **23**(39), 6684–6692.

Shabalina SA and Spiridonov NA (2004) The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, **5**(4), 105.

Shi Y and Berg JM (1995) Specific DNA-RNA hybrid binding by zinc finger proteins. *Science*, **268**(5208), 282–284.

Shibata S and Lee JT (2004) Tsix transcription- versus RNA-based mechanisms in Xist repression and epigenetic choice. *Current Biology*, **14**(19), 1747–1754.

Singh N, Bergman Y, Cedar H and Chess A (2003a) Biallelic germline transcription at the kappa immunoglobulin locus. *The Journal of Experimental Medicine*, **197**(6), 743–750.

Singh N, Ebrahimi FA, Gimelbrant AA, Ensminger AW, Tackett MR, Qi P, Gribnau J and Chess A (2003b) Coordination of the random asynchronous replication of autosomal loci. *Nature Genetics*, **33**(3), 339–341.

Sleutels F, Tjon G, Ludwig T and Barlow DP (2003) Imprinted silencing of Slc22a2 and Slc22a3 does not need transcriptional overlap between Igf2r and Air. *The EMBO Journal*, **22**(14), 3696–3704.

Sleutels F, Zwart R and Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**(6873), 810–813.

Sonstegard TS, Capuco AV, White J, Van Tassell CP, Connor EE, Cho J, Sultana R, Shade L, Wray JE, Wells KD, *et al*. (2002) Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. *Mammalian Genome*, **13**(7), 373–379.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Emilio Barbano P, *et al*. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, **306**(5696), 655–660.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(16), 6062–6067.

Zambrowicz BP, Imamoto A, Fiering S, Herzenberg LA, Kerr WG and Soriano P (1997) Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(8), 3789–3794.

Zhai G, Iskandar M, Barilla K and Romaniuk PJ (2001) Characterization of RNA aptamer binding by the Wilms' tumor suppressor protein WT1. *Biochemistry*, **40**(7), 2032–2040.

Short Specialist Review

The distribution of genes in human genome

Giorgio Bernardi
Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Naples, Italy

1. Genome compartmentalization

Thirty years ago, we discovered that the bovine genome (neglecting satellite DNAs, which consist of tandemly repeated short sequences) was made of DNA molecules (about 30 kb in size in our work) of fairly uniform GC levels (GC is the molar percentage of guanine+cytosine in DNA). These DNA molecules belonged to a small number of families that covered a wide GC range (Filipski *et al*., 1973). Indeed, each family was comparable in compositional heterogeneity to prokaryotic genomes, the least heterogeneous genomes of living organisms. This was the first evidence that a mammalian genome was a mosaic of DNA segments belonging to different compositional families.

These initial observations were quickly extended to other vertebrates. The genomes of warm-blooded vertebrates essentially showed the compositional features just described for the bovine genome (Thiery *et al*., 1976). For example, in the human genome, five families of DNA molecules (neglecting satellite and ribosomal DNA) were identified and physically separated from each other. They were called L1, L2, H1, H2, and H3, in order of increasing GC level; the first three families comprised about 88% of DNA, the last two only about 12% (see Figure 1).

Further, investigations showed that the DNA molecules derived from much larger, fairly homogeneous DNA regions (at least 300 kb in size; Macaya *et al*., 1976), which were called *isochores* (for compositionally equal regions). The existence and the features of isochores were confirmed and visualized 25 years later (Pavliček *et al*., 2002) using the draft sequence of the human genome as soon as it became available (Lander *et al*., 2001; Venter *et al*., 2001; see Figure 2).

2. Genome phenotypes

In contrast to warm-blooded vertebrates, cold-blooded vertebrates showed a less striking heterogeneity, because their GC-richest isochores were not as GC-rich as in warm-blooded vertebrates (Figure 3). Most interestingly, the compositional patterns

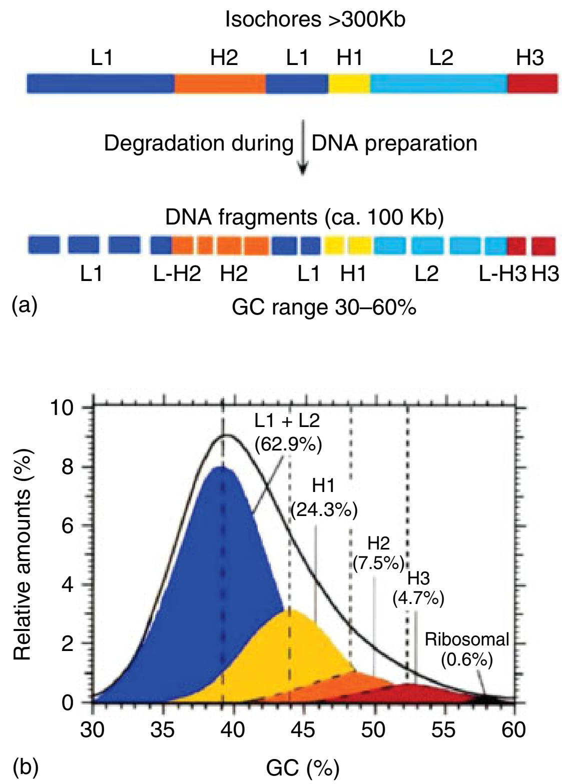


Figure 1 (a) Scheme of the isochore organization of the human genome. This genome, which is typical of the genome of most mammals, is a mosaic of large DNA segments, the isochores, which are compositionally fairly homogeneous and can be partitioned into a small number of families, light, or GC-poor (L1 and L2), and heavy, or GC-rich (H1, H2, and H3). Isochores are degraded during DNA preparation to fragments of 50–100 kb in size. The GC range of these DNA molecules from the human genome is extremely broad, 30 to 60%. (Reprinted with permission from the Annual Reviews of Genetics, Volume 29. Copyright 1995 by Annual Reviews **www.annualreviews.org**). (b) The CsCl profile of human DNA is resolved into its major DNA components, namely, DNA fragments derived from each one of the isochore families (L1, L2, H1, H2, and H3). Modal GC levels of isochore families are indicated on the abscissa (broken vertical lines). The relative amounts of major DNA components are indicated. Satellite DNAs (which form only a very few percent of the human genome) are not represented. (Reprinted from Zoubak *et al*., The gene distribution of the human genome, *Gene*, **174**, 95–102, Copyright 1996, with permission from Elsevier)

of the genomes of vertebrates are mimicked by the compositional patterns of coding sequences (that only represent 1–2% of the genomes in most vertebrates). Both compositional patterns amount to "genome phenotypes" (see Figure 3). This is a new concept compared to the "classical phenotype", which is represented by form and function, or, in molecular terms, by proteins and their expression.

It is important to note that an isochore organization of the genome is not limited to vertebrates, but is very widespread in eukaryotes, being also found in plants, insects, trypanosomes, and so on (see Bernardi, 2004).

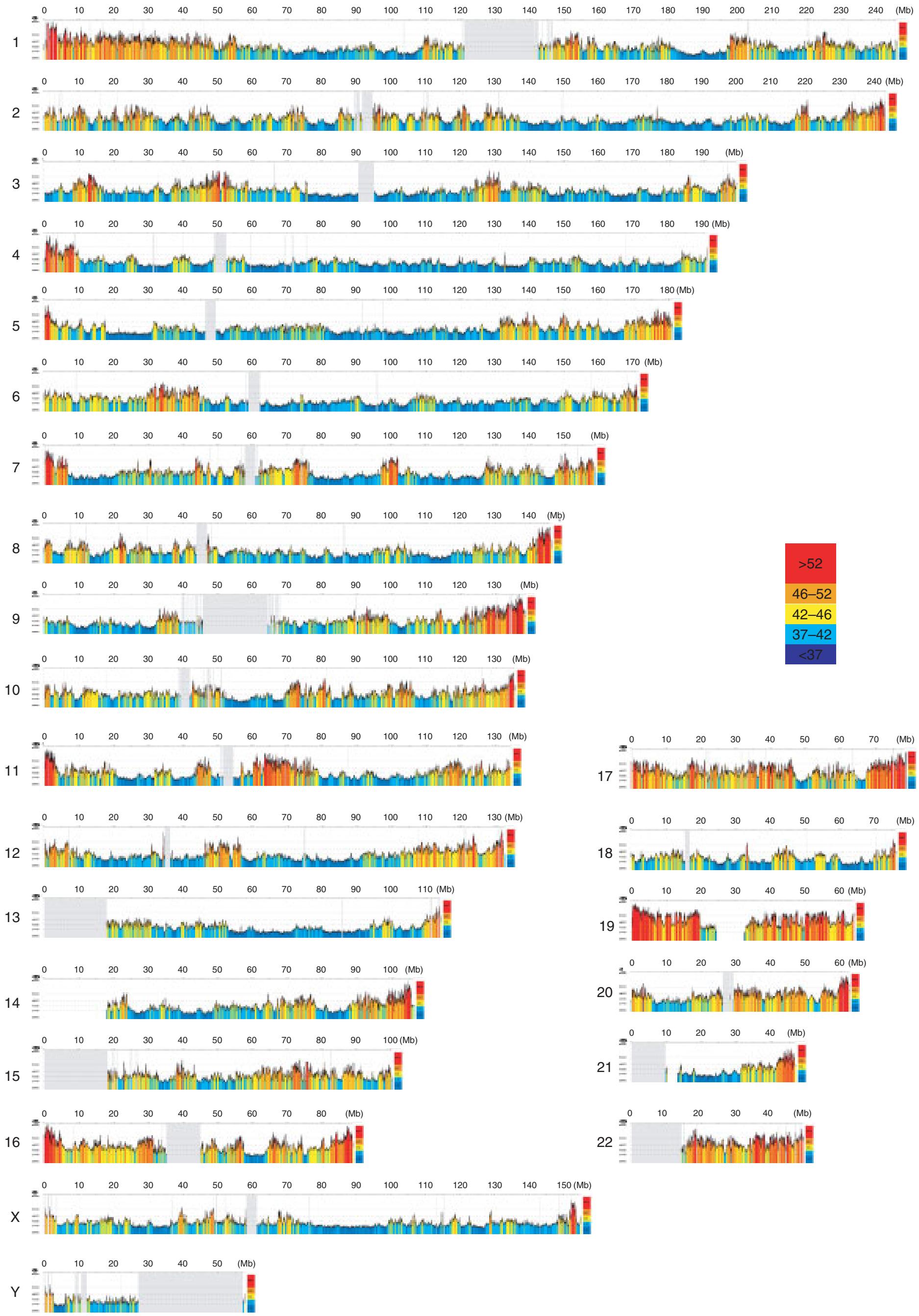


Figure 2 A color-coded compositional map of human chromosomes, representing 100 kb moving window plots that scan the human genome sequence. Color codes span the spectrum of GC levels in five steps, from ultramarine blue (GC-poorest isochores) to scarlet red (GC-richest isochores). Gray vertical lines correspond to the gaps present in the euchromatic part of the chromosomes. Gray bands to centromeres (Modified from Costantini *et al*., 2005)

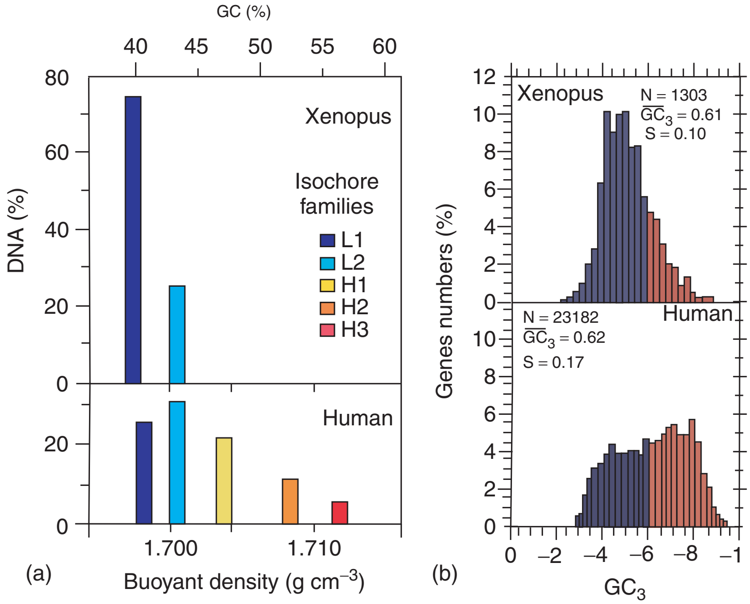


Figure 3 (a) Isochore families from *Xenopus* and human, as deduced from density gradient centrifugation. (b) Compositional patterns of coding sequences (represented by GC_3 values averaged per coding sequence; GC_3 is the GC level of third codon positions) for *Xenopus* and human. (Modified from Bernardi, 1995)

3. The genomic code

An obvious question is whether there is any correlation between the composition of coding and contiguous noncoding sequences. The answer is positive. Indeed, GC-rich coding sequences are flanked by GC-rich noncoding sequences, whereas GC-poor coding sequences are flanked by GC-poor noncoding sequences (Figure 4a). In fact, the *genomic code* (as these correlations were called) extends further, since compositional correlations also hold between exons and introns (Figure 4b) and among codon positions (see Figure 4c), the latter being valid from prokaryotes to mammals and being, therefore, called *universal correlations* (D'Onofrio and Bernardi, 1992). Finally, correlations also exist between the composition of coding sequences and the hydrophobicity and secondary structure (aperiodic, helix, strand) of the encoded proteins (Figure 5).

It should be stressed that the genomic code provided the first evidence for the eukaryotic genome being an *integrated ensemble*. One could also say that the genomic code has shown, to paraphrase Galileo, that the book of the genome is written in a mathematical language. Obviously, the genomic code cannot be reconciled with the idea of noncoding sequences being "junk DNA" (Ohno, 1972) and with the idea that "repeated sequences" like Alus and LINEs are "selfish DNAs" (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Indeed, if the base composition of coding sequences is under selection for thermal stability (see Bernardi,

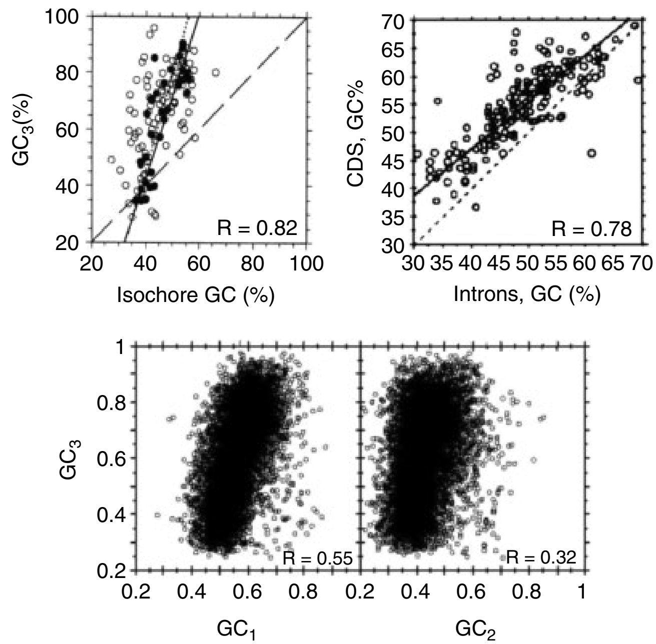


Figure 4 (a) Correlations between GC_3 of human genes and the GC level of DNA fractions or YACs (Yeast Artificial Chromosomes) in which the genes were localized (filled circles), or of 3′ flanking sequences (open circles; from Zoubak *et al*., 1996). (b) Correlations between GC levels of human coding sequences (CDS) and of the corresponding introns (Modified from Clay *et al*., 1996). (c) Correlations between GC_3 and GC_1 or GC_2 for 20 148 human genes (Modified from Jabbari *et al*., 2003)

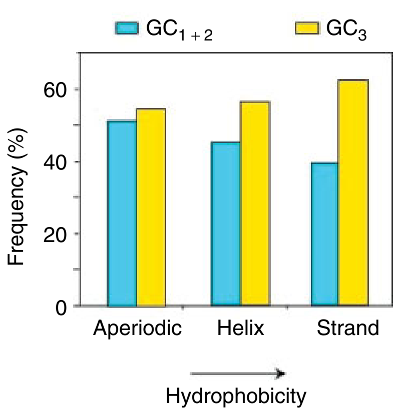


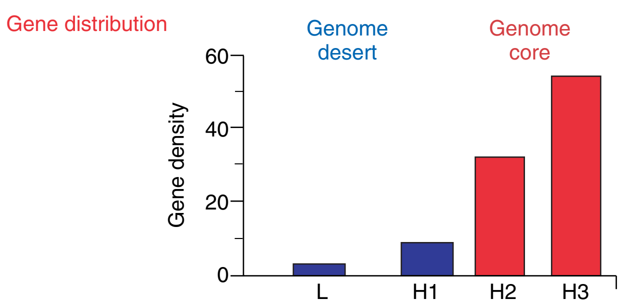
Figure 5 Histogram of the frequencies of GC_3 and GC_{1+2} in the three secondary structures of the proteins, which were ordered according to increasing hydrophobicity (Reprinted from D'Onofrio *et al*., 2002 The base composition of the genes is correlated with the secondary structures of the encoded proteins, *Gene,* **300**, 179–187, Copyright 2002, with permission from Elsevier)

2004), so must be noncoding sequences, since their composition is correlated with that of coding sequences.

4. Gene distribution in isochores and chromosomes

Two most important properties of vertebrates (as well as of other eukaryotes) concern the distribution of genes in the genome and in chromosomes and the correlations of such gene distributions with other structural and functional properties (Figure 6). The first indication that gene distribution was strikingly nonuniform (in contrast with the then current views) was published 20 years ago (Bernardi *et al*., 1985). Further work (Mouchiroud *et al*., 1991; Zoubak *et al*., 1996) identified two gene spaces, a *genome core* (the GC-rich isochore families H2 and H3) characterized by a high gene density and a *genome desert* (the GC-poor isochore families L1, L2, and H1), where gene density is low (see Figure 6). The genome desert and the genome core represent about 88% and 12%, respectively, of the genome, whereas the gene number is approximately the same in the two "gene spaces".

The two gene spaces are associated, as just mentioned, with different structural, functional, and evolutionary properties (see Figure 6). Among the former ones, the size of introns and untranslated regions (UTR) are large in the genome desert, but small in the genome core. The distribution of the DNA from the two gene spaces is quite different in metaphase chromosomes, the gene-dense regions of H2 and H3



Correlations with structure and function

Intro, UTR size	Large	Small
Chromatin structure	Closed	Open
GC Heterogeneity	Low	High
Gene experssion	Low	High
Replication timing	Late	Early
Recombination	Low	High

Figure 6 Gene distribution in the human genome. The major structural, functional, and evolutionary properties associated with each gene space are listed

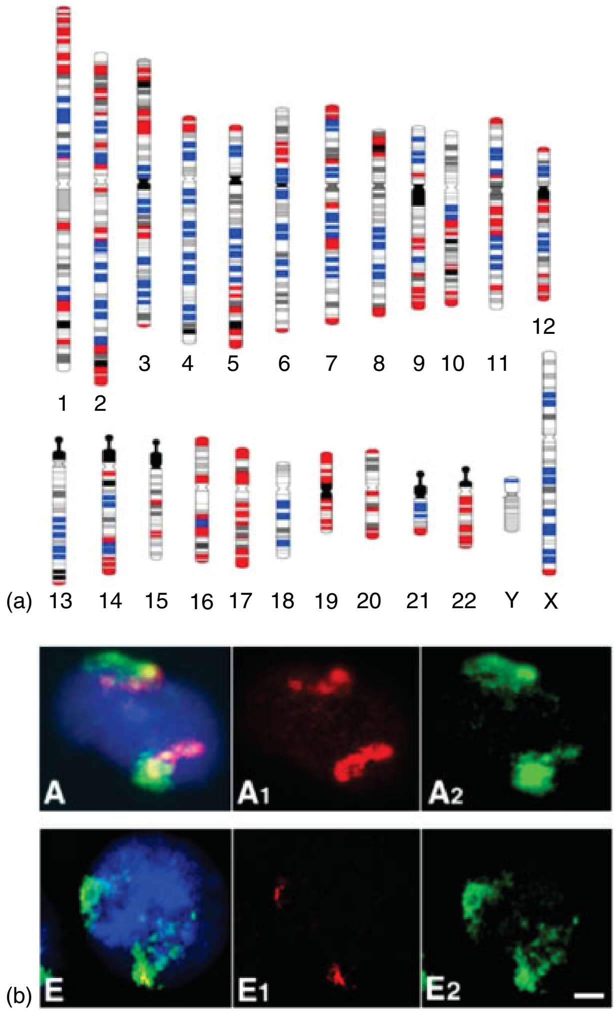


Figure 7 (a) Identification of the GC-poorest and the GC-richest chromosomal bands. Human karyotype at a resolution of 850 bands showing the chromosomal bands containing the GC-poorest (L1+, blue bars) and the GC-richest isochores (H3 +, red bars). The former correspond to G (iemsa) bands, the latter to R (everse) bands. The "intermediate" bands (the 50% of 80 bands that are neither GC- and gene-poorest nor GC- and gene-richest) are shown in white for the $H3^-$ R bands, in gray (according to Francke, 1994) for the $L1^-$ G bands (Modified from Federico *et al*., 2000). (b) Nuclear distribution of chromosomal regions characterized by different GC levels. 15–20 Mb chromosomal regions corresponding to GC-richest bands 6p21 (A) and 12q21 GC-poorest (E) were cohybridized with the corresponding chromosome probes. Band and chromosome DNAs were biotin- (red signals) and digoxigenin- (green signals) labeled, respectively. Nuclei were DAPI stained (blue). (Reprinted from Saccone *et al*., 2002 Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene*, **300**, 169–178, Copyright 2002, with permission from Elsevier)

isochores being predominant in or near telomeres, the gene-poor regions mainly near the centromeres (Figure 7a). Most interestingly, in the interphase nucleus, the gene-dense regions of chromosome are very "open" and centrally located, whereas the gene-poor regions are densely packed against the nuclear membrane (Figure 7b).

This latter finding not only accounts for the fact that gene-dense regions have a high transcription level compared to gene-poor regions but also for the need of the open chromatin of the former to be stabilized thermodynamically by a GC increase at the emergence of warm-blooded vertebrates. In contrast, the gene-poor regions are stabilized by their closed chromatin structure. Other distinctive functional properties concern replication timing (late and early in the genome desert and in the genome core, respectively) and recombination, which is frequent in the first case and rare in the second.

In summary, the bimodal gene distribution revealed by the compositional approach used in our work and confirmed by all subsequent sequencing investigations (*see* Article 28, **The distribution of genes in human genome**, Volume 3) is strongly correlated with essential structural, functional, and evolutionary properties of the genome.

References

Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Review of Genetics*, **29**, 445–476.

Bernardi G (2004) *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution*, Elsevier: Amsterdam.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M and Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–957.

Clay O, Cacciò S, Zoubak S, Mouchiroud D and Bernardi G (1996) Human coding and non-coding DNA: compositional correlations. *Molecular Phylogenetics and Evolution*, **5**, 2–12.

Costantini M, Saccone S, Federico C, Auletta F and Bernardi G (2005) Understanding chromosomal bands. (paper in preparation)

D'Onofrio G and Bernardi G (1992) A universal compositional correlation among codon positions. *Gene*, **110**, 81–88.

D'Onofrio G, Ghosh TC and Bernardi G (2002) The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene*, **300**, 179–187.

Doolittle WF and Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.

Federico C, Andreozzi L, Saccone S and Bernardi G (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Research*, **8**, 737–746.

Filipski J, Thiery JP and Bernardi G (1973) An analysis of the bovine genome by Cs2SO4-Ag+ density gradient centrifugation. *Journal of Molecular Biology*, **80**, 177–197.

Francke W (1994) Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenetics and Cell Genetics*, **6**, 206–219.

Jabbari K, Cruveiller S, Clay O and Bernardi G (2003) The correlation between GC_3 and hydropathy in human genes. *Gene*, **317**, 137–140.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Macaya G, Thiery JP and Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of Molecular Biology*, **108**, 237–254.

Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C and Bernardi G (1991) The distribution of genes in the human genome. *Gene*, **100**, 181–187.

Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven Symposia in Biology*, **23**, 366–370.

Orgel LE and Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.

Pavliček A, Paces J, Clay O and Bernardi G (2002) A compact view of isochores in the draft human genome sequence. *FEBS Letters*, **511**, 165–169.

Saccone S, Federico C and Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene*, **300**, 169–178.

Thiery JP, Macaya G and Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *Journal of Molecular Biology*, **108**, 219–235.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Zoubak S, Clay O and Bernardi G (1996) The gene distribution of the human genome. *Gene*, **174**, 95–102.

Short Specialist Review

Pseudogenes

David Torrents
European Molecular Biology Laboratory, Heidelberg, Germany

Pseudogenes are genomic regions that derive from genes but have no function (Balakirev and Ayala, 2003; Mighell *et al*., 2000; Vanin, 1985; Zhang and Gerstein, 2004). Because they have been traditionally assumed to present no biological relevance, the identification and classification of pseudogenes has attracted little attention and was initially restricted to evolutionary studies. Over the years, a significant number of pseudogenes have been accidentally uncovered during the analysis of gene-containing regions in different genomes, but their sequences and locations were rarely reported to public databases. It is in only recently that scientists became aware of the importance of identifying and annotating pseudogenes. Part of this growing attention derives from the need to provide basic annotation of recently sequenced genomes, which should include pseudogenes. But it is from the observation that a significant fraction of likely nonfunctional regions (pseudogenes) were being wrongly annotated as functional genes during preliminary analyses of completed large genomes (Gibbs *et al*., 2004; Lander *et al*., 2001; Waterston *et al*., 2002), that the identification and classification of pseudogenes has become extremely relevant. It is now required for any automatic approach for predicting genes to take into account the presence of pseudogenes and to include some rules to distinguish them. Finally, because of the increasing number of reports providing evidence that some regions previously classified as pseudogenes are actually functional, normally involved in the regulation of gene expression (Korneev *et al*., 1999; Healy *et al*., 1991; Hirotsune *et al*., 2003), collections of dead genes become an excellent source for the identification of additional examples of this type.

Although pseudogenes can exceptionally arise from the direct inactivation of functional genes, the vast majority of them are formed through gene duplication. In mammals, two major mechanisms of gene duplication coexist, with different impact on pseudogene formation and gene evolution: (1) The s*egmental duplication* of genes results in the formation of identical copies, mostly through unequal crossing-over during meiotic recombination (Alberts *et al*., 1994). This type of duplication might also involve the copy of the promoter and other regulatory regions along with the coding region. Therefore, favored by an initial period of functional redundancy and relaxed selective constraint, expressed duplicates might undergo sequence modifications that lead to the acquisition of new, or more specialized functions. Although the ultimate fate of most segmental duplicated genes

is unclear in terms of their preservation and the acquisition of new functions, it is commonly accepted that, one of the copies usually becomes a *nonprocessed pseudogene* through the accumulation of lethal mutations, whereas the other retains the original (or eventually enhanced) function (Force *et al*., 1999; Ohno, 1970; Lynch and Conery, 2000). Alternatively segmental duplications might only involve fragments of genes that will likely result in the formation of incomplete and hence nonfunctional gene copies. (2) A second mechanism of gene duplication corresponds to *retrotransposition*, which generates *processed* gene copies. This duplication mechanism results in the copy of mature RNAs, through their retrotranscription and integration back into the genome, using the machinery of retrotransposable elements (Esnault *et al*., 2000). In contrast to some segmental gene duplicates, processed mRNA copies carry no signals for transcription initiation. Because mRNA copies insert randomly in the genome, that is, likely far from active promoters, retrotransposed gene copies are rarely ever expressed and can therefore be considered "dead on arrival". However, there are isolated examples of functional genes that have arisen through the retrotransposition of mRNAs (Lander *et al*., 2001; Brosius, 1999; Burki and Kaessmann, 2004). The absence of introns and the presence of flanking repeats and polyA tail are clear characteristics that can be used (when detectable) to distinguish retrotransposed gene copies from segmental duplicated copies.

But how are pseudogenes recognized? Following the general definition, a given genomic region can be cataloged as pseudogenic when it shows homology to an active gene and has no function. Whereas homology is relatively easy to identify through the detection of significant sequence similarity, the absence of function remains impossible to prove, given all possible (new and subtle) types of biological roles in which a particular region can be involved. Nevertheless, a few characteristics that have been found in a number of gene duplicates have been accepted as strong indicators for the absence of functionality. These features have been used to report, not only single cases, but also large collections of pseudogenes from different organisms.

The first reports describing pseudogenes relied on the absence of transcription to propose the absence of functionality. In fact, the term *pseudogene* was first introduced in 1977 to describe a tandem copy of the *Xenopus laevis* 5 S RNA gene for which evidence of expression could not be initially found (Jacq *et al*., 1977). The same region was later shown to be actually expressed but with "inefficient termination" of transcription (Miller and Melton, 1981). Soon after, structurally identical copies of the rabbit and human globin genes were identified next to their functional paralogs and classified as pseudogenes, using again questionable arguments to propose nonfunctionality (Fritsch *et al*., 1980; Hardison *et al*., 1979; Lauer *et al*., 1980). Because the impossibility to detect expression of segmental duplicated genes can derive from methodological limitations, this argument is nowadays considered insufficient to suggest nonfunctionality. Even the classification of retrotransposed gene copies as nonfunctional or pseudogenic requires additional evidence, even though they are unlikely to ever be transcribed.

The detection of truncations, that is, in-frame stop codons or frameshifts, within duplicated protein coding regions is accepted as the most convincing evidence to disprove functionality, as their presence is generally not compatible with

the synthesis of complete and operative proteins. Through systematic searches of truncated and nearly complete retrotransposed gene copies, two independent approaches identified in human genome, 3500 (Ohshima *et al*., 2003) and 8000 (Zhang *et al*., 2003) processed pseudogenes. Similar procedures detected processed pseudogenes in other genomes, including puffer fish (Dasilva *et al*., 2002), fruit fly (Harrison *et al*., 2003), worm (Harrison *et al*., 2001), and yeast (Harrison *et al*., 2002). But these approaches, despite presenting a high specificity, overlooked an important fraction of pseudogenes, which include those with an apparently intact coding region, but with other type of lethal mutations (e.g., replacements of functionally essential amino acids, or disrupted or missing promoters), pseudogenes arising from segmental duplications, and all pseudogenes with incomplete coding regions.

At the same time, another study used a different principle to evaluate the presence or absence of functionality for detectable gene copies in human (Torrents *et al*., 2003), which did not depend either on the presence of truncations or on the mechanism of duplication, that is, it covered both retrotransposed and segmental duplicated pseudogenes. 20 000 complete and partial gene copies identified between predicted and known functional genes were evaluated as to functionality, only taking into account the ratio of silent (synonymous, K_S) to amino acid replacement (nonsynonymous, K_A) substitutions, which indicates the associated levels of selective constraint (Li *et al*., 1981). K_A/K_S ratios of pseudogenes and those of the vast majority of genes are generally different, as mutations in genes causing amino acid replacements with functional consequences are selected against, in contrast to mutations occurring in pseudogenes. This functionality test revealed that nearly all identified gene copies were neutrally evolving and therefore nonfunctional. The regions identified with this approach, despite constituting the largest set of nonfunctional gene copies identified so far, were estimated to correspond to a fraction of the complete population of human pseudogenes, which is likely to exceed the number of genes. Following the same detection and classification strategy, a similar number of pseudogenes were identified in mouse and rat (Gibbs *et al*., 2004). From the comparative analysis of human and mouse orthologous DNA blocks, the same study showed that the majority (>70%) of the pseudogenes are located far from their functional paralogs, which is consistent with a retrotranspositional origin. This distinction also revealed that although the number of both processed and nonprocessed pseudogenes correlates with the size of the chromosomes in human, their intrachromosomal distribution differs: processed pseudogenes are more abundant close to telomeres, nonprocessed pseudogenes are normally enriched in gene dense regions.

All mammalian genomes investigated so far appear to have a high and similar number of detectable pseudogenes (~20 000), suggesting that they share similar mechanisms (and rates) for the formation and death of this type of regions (Gibbs *et al*., 2004; Torrents *et al*., 2003). On the other hand, other vertebrates, such as chicken, appears to heave nearly an undetectable number of processed pseudogenes (ICGSC, 2004), which is likely due to the lack of interaction between the machinery of active retrotransposons with host mRNAs. Similarly, a number of searches within nonvertebrate genomes revealed in general a low number of both processed and nonprocessed pseudogenes (Harrison *et al*., 2003; Harrison *et al*., 2001; Harrison

et al., 2002; Zdobnov *et al*., 2002), which could be in agreement with the observed size constraints associated to their genomes (Petrov and Hartl, 2000).

Between the years 2001 and 2003, important progress has been achieved in the identification and classification of pseudogenes. Nevertheless, we expect that the sequencing of more genomes, and particularly the increasing availability of new experimental data revealing atypical forms of functionality, will provide, in a close future, additional criteria for the difficult task of distinguishing between functional and pseudogenic gene duplicates. This will then allow significant improvements to be made in the construction of pseudogene catalogs and to investigate their actual impact on gene and genome evolution.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson JD (1994) *Molecular Biology of the Cell*. Garland Publishing: New York.

Balakirev ES and Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? *Annual Review of Genetics*, **37**, 123–151.

Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**(1), 115–134.

Burki F and Kaessmann H (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genetics*, **36**(10), 1061–1063.

Dasilva C, Hadji H, Ozouf-Costaz C, Nicaud S, Jaillon O, Weissenbach J and Crollius HR (2002) Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(21), 13636–13641.

Esnault C, Maestre J and Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, **24**(4), 363–367.

Force A, Lynch M, Pickett FB, Amores A, Yan YL and Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**(4), 1531–1545.

Fritsch EF, Lawn RM and Maniatis T (1980) Molecular cloning and characterization of the human beta-like globin gene cluster. *Cell*, **19**(4), 959–972.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, *et al*. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**(6982), 493–521.

Hardison RC, Butler ET III, Lacy E, Maniatis T, Rosenthal N and Efstratiadis A (1979) The structure and transcription of four linked rabbit beta-like globin genes. *Cell*, **18**(4), 1285–1297.

Harrison P, Kumar A, Lan N, Echols N, Snyder M and Gerstein M (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *Journal of Molecular Biology*, **316**(3), 409–419.

Harrison PM, Milburn D, Zhang Z, Bertone P and Gerstein M (2003) Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Research*, **31**(3), 1033–1037.

Harrison PM, Echols N and Gerstein MB (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Research*, **29**(3), 818–830.

Healy MJ, Dumancic MM and Oakeshott JG (1991) Biochemical and physiological studies of soluble esterases from Drosophila melanogaster. *Biochemical Genetics*, **29**(7–8), 365–388.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A and Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**(6935), 91–96.

ICGSC (international-chicken-genome-sequencing-consortium) (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**(7018), 695–716.

Jacq C, Miller JR and Brownlee GG (1977) A pseudogene structure in 5 S DNA of Xenopus laevis. *Cell*, **12**(1), 109–120.

Korneev SA, Park JH and O'Shea M (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *The Journal of Neuroscience*, **19**(18), 7711–7720.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lauer J, Shen CK and Maniatis T (1980) The chromosomal arrangement of human alpha-like globin genes: sequence homology and alpha-globin gene deletions. *Cell*, **20**(1), 119–130.

Li WH, Gojobori T and Nei M (1981) Pseudogenes as a paradigm of neutral evolution. *Nature*, **292**(5820), 237–239.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–1155.

Mighell AJ, Smith NR, Robinson PA and Markham AF (2000) Vertebrate pseudogenes. *FEBS Letters*, **468**(2–3), 109–114.

Miller JR and Melton DA (1981) A transcriptionally active pseudogene in xenopus laevis oocyte 5 S DNA. *Cell*, **24**(3), 829–835.

Ohno S (1970) *Evolution by Gene Duplication*. George Allen and Unwin: London, p. 160.

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y and Okada N (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, **4**(11), R74.

Petrov DA and Hartl DL (2000) Pseudogene evolution and natural selection for a compact genome. *The Journal of Heredity*, **91**(3), 221–227.

Torrents D, Suyama M, Zdobnov E and Bork P (2003) A genome-wide survey of human pseudogenes. *Genome Research*, **13**(12), 2559–2567.

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics*, **19**, 253–272.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–562.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, *et al*. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science*, **298**(5591), 149–159.

Zhang Z, Harrison PM, Liu Y and Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Research*, **13**(12), 2541–2558.

Zhang Z and Gerstein M (2004) Large-scale analysis of pseudogenes in the human genome. *Current Opinion in Genetics & Development*, **14**(4), 328–335.

Short Specialist Review

Alternative splicing: conservation and function

Mikhail S. Gelfand
Institute for Information Transmission Problems, Moscow, Russia

Evgenia V. Kriventseva
BASF Plant Science GmbH, Ludwigshafen, Germany

At least half of human genes seem to be alternatively spliced (Lander *et al.*, 2001). This estimate is mainly based on the comparison of genomic DNA with EST (expressed sequence tag, *see* Article 78, **What is an EST?**, Volume 4) sequences (Mironov *et al.*, 1999; Brett *et al.*, 2000), and thus is subject to uncertainty stemming from the fact that the ESTs do not necessarily correspond to functional mRNA. Even if experimental artifacts such as underspliced transcripts could be eliminated, there remains a problem of errors by the splicing machinery itself, the so-called aberrant splicing (*see* Article 87, **Manufacturing EST libraries**, Volume 4).

Indeed, the normalization of mRNA concentrations during construction of clone libraries leads to the sequencing of ESTs arising from rare mRNA isoforms. Further, computational analysis has demonstrated the existence of numerous cancer-specific ESTs (more exactly, ESTs corresponding to cancer-specific alternatively spliced isoforms, *see* Article 82, **Using ORESTES ESTs to mine gene cancer expression data**, Volume 4) (Wang *et al.*, 2003; Sorek *et al.*, 2003; Xie *et al.*, 2002; Xu and Lee, 2003), the emergence of which could be due to the general disruption of control mechanisms in cancerous cell lines. Although one could claim that almost all human genes show some evidence of alternative splicing, when stricter criteria are considered (e.g. at least two ESTs supporting an alternative splicing event), the fraction of alternatively spliced genes decreases to 17–28% (Kan *et al.*, 2002).

A new twist to this discussion was added when several groups attempted to compare alternative splicing of human and mouse genes (Thanaraj *et al.*, 2003; Modrek and Lee, 2003; Modrek *et al.*, 2001; Nurtdinov *et al.*, 2003). Surprisingly, it turned out that a considerable fraction of human genes have alternatively spliced isoforms, which are not conserved in mouse.

Two different approaches have been applied to compare human and mouse alternative splicing. One of them was a direct comparison of human and mouse ESTs. This approach demonstrated that at least 15% of human splice junctions (introns) are conserved in mouse (Thanaraj *et al.*, 2003). A similar estimate was

made for different types of elementary alternatives considered separately, at that, exon skipping events were shown to be more conserved than alternative splicing sites (Sugnet *et al*., 2004). However, as the mouse EST data at least is far from saturation, this is clearly a lower bound on the fraction of conserved alternative splicing.

The other approach is based on aligning human protein isoforms to mouse genomic DNA using spliced alignment algorithms (Mironov *et al*., 2001; *see also* Article 15, **Spliced alignment**, Volume 7) or simply BLAST (Altschul *et al*., 1997; *see also* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7). At that, an isoform is assumed to be conserved if the alternative region aligns to the mouse genome without frameshifts and is bounded by the standard GT-AG dinucleotides. It is clear that this definition yields an upper estimate on the number of conserved isoforms, since these conditions are necessary but not sufficient: an isoform may be nonexistent due to changes in splicing site positions other than GT-AG, or to changes in regulatory sites such as splicing enhancers. Further, this definition does not take into account nonconserved exon skipping events.

This approach, applied in (Nurtdinov *et al*., 2003), demonstrated that at least half (55%) of 166 alternatively spliced human genes had isoforms not conserved in their mouse orthologs. This was due to about 25% of unconserved elementary alternatives. Notably, similar results were obtained for elementary alternatives confirmed by mRNAs (24% nonconserved) and by ESTs only (31%).

A much larger sample in a similar setting was analyzed in (Modrek and Lee, 2003), where only cassette exons were considered. All such exons were divided into exons included in the major isoforms (i.e., present in the majority of ESTs overlapping the relevant region) and the minor form exons. The former were found to be conserved in 98% of cases, whereas only about a quarter (27%) of the latter were conserved. Similar results were obtained in a smaller-scale human–rat comparison. The average conservation of both types of exons, 75%, is remarkably close to the degree of conservation of elementary alternatives reported in (Nurtdinov *et al*., 2003).

An important question is whether these nonconserved alternatives are real or arise from splicing errors, and so on. The number of documented functional nonconserved alternatively spliced isoforms is not large (Nurtdinov *et al*., 2003). In fact, it has been suggested that most nonconserved isoforms are not functional (Sorek *et al*., 2004). The fraction of nonconserved cassette (skipped) exons, identified by a combination of EST analysis and genomic comparisons, was similar to that of the two studies mentioned above (75%). However, it was demonstrated that most nonconserved exons (79%) either led to a frameshift (because their length did not contain an integer number of triplets) or contained an in-frame stop codon. By contrast, only 27% of conserved cassette exons interrupted the reading frame. The difference decreased when exons supported by multiple ESTs were considered (46% interrupting exons, among exons supported by at least five ESTs).

Does that mean that the majority of nonconserved isoforms are nonfunctional? Frame interruption per se does not make an isoform nonfunctional. Indeed, about 40% of both human (Modrek *et al*., 2001) and mouse (Zavolan *et al*., 2002) alternative isoforms identified from EST and full-length cDNA analysis have an

interrupted reading frame, and a slightly smaller estimate (22%) was obtained in the analysis of published experimental data (Thanaraj and Stamm, 2003). An intermediate number of alternative isoforms (35%) was reported in (Lewis *et al*., 2003); moreover, it was demonstrated that most such isoforms would be subject to nonsense-mediated mRNA decay, as the stop codon occurred more than fifty nucleotide upstream of the 5′-most exon–exon junction. As this trend persisted after the filtering of less-reliable isoforms, it is likely that the frame-interrupting isoforms are functional; one suggested possibility was that they are involved in the regulation of splicing, translation, and mRNA degradation.

Indeed, a different line of evidence for functionality of nonconserved isoforms was considered in (Modrek and Lee, 2003). In many cases, the minor form nonconserved exons not only were supported by multiple ESTs but also demonstrated evidence for tissue-specific expression, and constituted a majority in this tissue.

Thus, an open question seems to be not the reality of nonconserved isoforms but their functionality. A large-scale proteomic study will be necessary to determine whether these isoforms are translated and yielded protein products.

In any case, alternative splicing was demonstrated to have a major effect on the protein structure (Kriventseva *et al*., 2003). Indeed, when compared with a random model, alternative splicing was shown to prefer shuffling complete protein domains instead of disrupting domains or falling into interdomain regions and to target functional sites when it is occurring within a domain. Indeed, alternative splicing often involves domains implicated in protein–protein interactions (Resch *et al*., 2004). Further, it was shown that alternative splicing has a tendency to remove gene regions encoding signal peptides and single transmembrane segments, thus producing secreted, membrane-bound, and cytozolic isoforms of proteins (Xing *et al*., 2003; Cline *et al*., 2004).

Thus, alternative splicing is a major mechanism for generating protein diversity, both in extant organisms and in evolution. The latter explanation is supported by additional observations: evidence for positive selection based on analysis of synonymous and nonsynonymous nucleotide substitutions (Iida and Akashi, 2000) and the fact that all Alu-derived protein-coding regions of human genes are alternatively spliced (Sorek *et al*., 2004). Indeed, an elegant theory of Modrek and Lee (2003) states that alternative splicing provides an organism with a possibility to experiment with new protein functions while not disrupting the old protein. If a new variant proves to be beneficial, its fraction may increase due to subtle changes in regulatory sites.

However, this does not explain why generation of protein variability cannot be obtained by gene duplication. Another less-appreciated role of alternative splicing could be that of maintaining *protein identity*. Indeed, in many cases, a cell needs proteins that are different in some domains and exactly identical in others. The most obvious example of this is given by membrane, secreted, and intracellular isoforms of various receptors. The recognition or ligand-binding domain should be the same, whereas the membrane anchor or a signal peptide is encoded by alternative exons. It is clear that such an arrangement cannot be obtained by gene duplication, as this would require an expensive mechanism for maintaining the identity of those DNA fragments that should encode identical domains.

Overall, computational comparative analysis of alternative splicing is a hot and important topic. The next step probably would be in merging the diverse approaches, aimed at the description of all aspects of the alternative splicing phenomenon: evolution of the exon–intron structure and of sequence in alternatively spliced regions, regulation, consequences for the protein structure and function, and so on. And, it is clear that such analyses will not be restricted to the study of mammals (human–mouse–rat). Other appealing groups of already available genomes are the two nematodes (*Caenorhabditis elegans* and *Caenorhabditis briggsae*, *see* Article 44, **The *C. elegans* genome**, Volume 3) and also fruit flies (*Drosophila melanogaster*, *Drosophila pseudoobscura*, and others) with the malarial mosquito *Anopheles gambiae* serving as an outlier; to be complemented, as sequencing of eukaryotic genomes progresses, by chicken, fishes (*Takifugu rubrupes*, *Danio rerio*, *see* Article 46, **The Fugu and Zebrafish genomes**, Volume 3), honeybee, and plants.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J and Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**, 83–86.

Cline MS, Shigeta R, Wheeler RL, Siani-Rose MA, Kulp D and Loraine AE (2004) The effects of alternative splicing on transmembrane proteins in the mouse genome. *Pacific Symposium on Biocomputing* 17–28.

Iida K and Akashi H (2000) A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene*, **261**, 93–105.

Kan Z, States D and Gish W (2002) Selecting for functional alternative splices in ESTs. *Genome Research*, **12**, 1837–1845.

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS and Sunyaev S (2003) Increase of functional diversity by alternative splicing. *Trends in Genetics*, **19**, 124–128.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lewis BP, Green RE and Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 189–192.

Mironov AA, Fickett JW and Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Research*, **9**, 1288–1293.

Mironov AA, Novichkov PS and Gelfand MS (2001) Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors. *Bioinformatics*, **17**, 13–15.

Modrek B and Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, **34**, 177–180.

Modrek B, Resch A, Grasso C and Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, **29**, 2850–2859.

Nurtdinov RN, Artamonova II, Mironov AA and Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, **12**, 1313–1320.

Resch A, Xing Y, Modrek B, Gorlick M, Riley R and Lee C (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *Journal of Proteome Research*, **3**, 76–83.

Sorek R, Basechess O and Safer HM (2003) Expressed sequence tags: clean before using. Correspondence re: Z. Wang *et al*., computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research*, **63**, 655–657. *Cancer Research*, **63**, 6996; author reply 6996–6997.

Sorek R, Shamir R and Ast G (2004) How prevalent is functional alternative splicing in the human genome? *Trends in Genetics*, **20**, 68–71.

Sugnet CW, Kent WJ, Ares M, Jr. and Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing*, 66–77.

Thanaraj TA, Clark F and Muilu J (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, **31**, 2544–2552.

Thanaraj TA and Stamm S (2003) Prediction and statistical analysis of alternatively spliced exons. *Progress in Molecular and Subcellular Biology*, **31**, 1–31.

Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH and Lee MP (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research*, **63**, 655–657.

Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A and Mintz L (2002) Computational analysis of alternative splicing using EST tissue information. *Genomics*, **80**, 326–330.

Xing Y, Xu Q and Lee C (2003) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Letters*, **555**, 572–578.

Xu Q and Lee C (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, **31**, 5635–5643.

Zavolan M, van Nimwegen E and Gaasterland T (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Research*, **12**, 1377–1385.

Short Specialist Review

Overlapping genes in the human genome

Izabela Makalowska
Pennsylvania State University, University Park, PA, USA

Viruses have very compact genomes. Yet, the discovery of the overlapping genes in bacteriophage phiX174 in1976 (Barrell *et al*., 1976) came as a surprise. It took another decade before similar phenomena were noticed in higher eukaryotes. In 1998, in the same issue of Nature, Spencer *et al*. (Spencer *et al*., 1986) published two overlapping genes in Drosophila and Williams and Fried (1986) reported the same pattern in mouse. However, overlapping genes in mammalian genomes are unexpected phenomena. Why, one can ask, despite vast genome-less genomic area do some genes overlap in mammalian genome? Regardless of numerous reports about overlapping genes in human, until recently, overlapping genes were not considered to be important or a large-scale event in human and other vertebrates' genomes. The only exceptions were genes within genes, which relatively early were believed to be a common feature of nuclear genomes. Large-scale EST and genomic sequence studies led to the conclusion that overlapping genes, other than genes embedded in another gene intron, are commonly present in higher eukaryotes.

Overlapping genes may be divided into several types. The major division would depend on the genes' direction, and we can observe overlapping genes on the same strand and on opposite strands. The latter category, an antitranscript overlapping gene, is our main focus. Among these, we can observe genes that share the same locus but the overlaps are only between exonic region in one gene and intronic area in another and genes that share not only locus but also exonic sequences. Depending on which parts of two genes share the genomic region, we can also categorize these overlaps as head to head, when the overlap is in the 5′ region of both genes, or tail to tail when the overlap involves 3′ regions. Special types already mentioned are embedded genes, a case in which one gene lies completely in the area of another one. Sometimes we can observe that several genes are embedded in the genomic locus of a gene on the opposite strand. For example, intron 27 of human neurofibromatosis type 1 gene (NF1) contains three other genes: *OMG*, *EVI2B*, and *EVI2A*. Examples of different categories of overlapping genes are shown in Figure 1.

The total number of overlapping genes in human and other nuclear genomes is still unknown. So far, reported numbers vary from 774 pairs (Veeramachaneni *et al*., 2004) to well above 2000 (Yelin *et al*., 2003). The discrepancy is mostly

caused by different type of data used and both numbers may actually be correct. In the first studies, only protein coding cDNAs were analyzed and in the second one, all human ESTs including nonprotein coding genes were considered. From the studies of mouse overlapping genes (Kiyosawa *et al*., 2003), we know that about 75% of genes overlap involve at least one noncoding gene. However, we may expect the number of overlapping genes in human genome to be higher and a good fraction of them have not yet been discovered owing to incomplete sequence data.

The biologic functions of natural antisense transcripts, their involvement in physiological processes, and gene regulation in living organisms are barely known. There is speculation that they form a double-stranded RNA to downregulate the expression of sense RNA molecules. However, so far, large-scale analysis has not shown any significant correlation in terms of function, localization, or process between two members of sense–antisense pairs and the distribution of overlapping genes regarding these parameters was found to be not significantly different from the rest of human genes (Yelin *et al*., 2003). Forming a double-stranded RNA to downregulate the expression of an RNA molecule may be a good explanation for some fraction of overlapping genes, especially when noncoding RNA is involved. But in protein coding genes, it would make the overlap quite hazardous since it could lead to RNA degradation and in extreme cases formation of antiparallel heteroduplex RNA could completely block the expression of both genes. Escape from such evolutionary pressure could be in differential expression and there are multiple examples of overlapping, protein coding genes showing clearly different temporospatial expression. However, there are also instances of overlapping genes being expressed at the same time and place. Therefore, regulation of expression of overlapping genes either does not always involve a counterpart gene or there are different mechanisms for doing this, not just by forming double-stranded RNA.

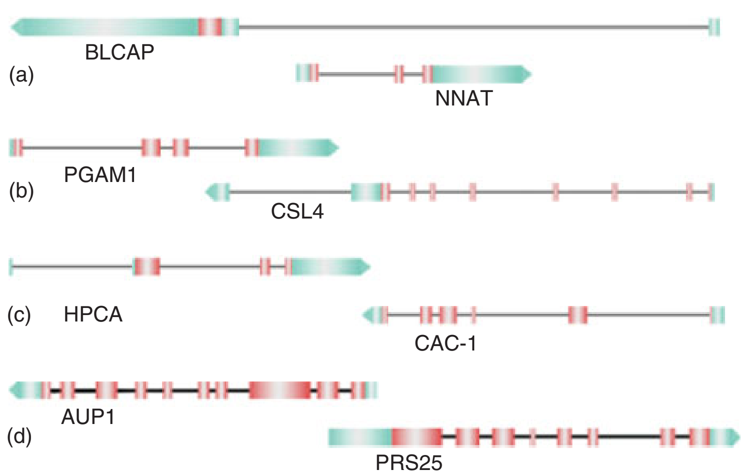


Figure 1 Examples of human overlapping genes. Red color indicates coding sequence and green color shows untranslated regions. (a) Imbedded gene, (b) genes sharing genomic regions with overlap between exon of one gene and intron of another, (c) tail-to-tail exons overlap, (d) head-to-head exons overlap. Red color denotes coding sequences and green color indicates untranslated regions

The evolution of overlapping genes is also unknown. We do not understand either the mechanism or the meaning of the origination of overlapping genes in higher eukaryotes. Keese and Gibbs (1992) suggested that overlapping genes arise as a result of overprinting – a process of generating new genes from preexisting nucleotide sequences. This process took place after the divergence of mammals from birds, and overlapping genes represent young, phylogenetically restricted genes encoding proteins with diverse functions, and are therefore specialized to the present lifestyle of the organism in which they are found.

Shintani *et al*. (1999) suggested that the overlap between the two genes studied by them, *ACAT2* and *TCP1*, arose during the transition from therapsid reptiles to mammals, and that the overlap could have happened in one of two ways. In one scenario, the rearrangement may have been accompanied by the loss of a part of the 3′-untranslated region, including the polyadenylation signal, from one gene. By chance, however, the 3′-UTR of the new neighbor on the opposite strand contained all the signals necessary for termination and transcription so that the translocated gene could continue to function. Alternatively, the two genes could have become neighbors through the rearrangement but did not overlap at first. Later, one of the genes lost its original polyadenylation signal, but was able to use a signal that happened to be present on the noncoding strand of the other gene. None of these hypotheses fully explains the nature of overlapping genes evolution. There is evidence that some human overlapping genes do not have othologs in other genomes (Makalowska *et al*., 2005), which support the hypothesis suggested by Keese and Geebs. On the other hand, there are many instances where genes overlap in one organism but are located next to each other, without any overlap, in another species.

Studies of overlapping *MINK* and *CHRNE* genes (Dan *et al*., 2002) provide some support for second hypothesis; however, it will only work well with genes overlapping at 3′ end. Results of this study show that mutations in the polyadenylation signal of the *CHRNE* gene resulted in the adoption of alternative signal conserved in the 3′-UTR of the *MINK* gene located downstream on the opposite strand. As an outcome, we observe an overlap between the last exons of these two genes. Analysis of genomes of different orders of placental mammals demonstrated that the CHRNA/MINK overlap occurred at least three times independently during the course of mammalian evolution, and all happened in distinct ways. One of these events most likely happened after the cercopithecoid/hominoid split. This means that many overlaps could be relatively young and the pattern of relation between genes does not have to be conserved among mammals. Confirmation of this may come from large-scale studies of human and mouse overlapping genes done by Veeramachaneni *et al*. (2004). Out of 255 human overlapping gene pairs, only in 95 cases were genes that were also overlapping in mouse and in 150 cases, genes were overlapping in human but not in mouse although both the genes were mapped next to each other.

It was expected that genes that overlap should be more conserved among species. Lipman (1997) explained the higher conservation of noncoding sequences of some genes exactly by the presence of antisense transcripts and therefore higher level of evolutionary pressure. However, genes that overlap in both human and mouse do not show statistically significant difference in the level of conservancy than

nonoverlapping genes (Veeramachaneni *et al*., 2004). The relation between overlap and level of conservation was also not observed by Dan *et al*. (2002). 3′-UTR of *MINK* gene was extremely conserved in all species, not in only those where the overlap occurred, and the 3′-UTR of *CHRNE* had overall faster evolutionary pace.

In conclusion, we can say that overlapping genes brought a new light and many questions into molecular evolution, genome structure, and gene functions. Despite many studies, the total number of such genes in higher eukaryotes, mechanism of their origination and functional significance are still awaiting full explanation.

References

Barrell BG, Air GM and Hutchison CA, III (1976) Overlapping genes in bacteriophage phiX174. *Nature*, **264**, 34–41.

Dan I, Watanabe NM, Kajikawa E, Ishida T, Pandey A and Kusumi A (2002) Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acids Research*, **30**, 2906–2910.

Keese PK and Gibbs A (1992) Origins of genes: "big bang" or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 9489–9493.

Kiyosawa H, Yamanaka I, Osato N, Kondo S and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*, **13**, 1324–1334.

Lipman DJ (1997) Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Research*, **25**, 3580–3583.

Makalowska I, Lin C and Makalowski W (2005) Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*, **29**(1), 1–12.

Shintani S, O'HUigin C, Toyosawa S, Michalova V and Klein J (1999) Origin of gene overlap: the case of TCP1 and ACAT2. *Genetics*, **152**, 743–754.

Spencer CA, Gietz RD and Hodgetts RB (1986) Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature*, **322**, 279–281.

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R and Makalowska I (2004) Mammalian overlapping genes: the comparative perspective. *Genome Research*, **14**, 280–286.

Williams T and Fried M (1986) A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. *Nature*, **322**, 275–279.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, *et al*. (2003) Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, **21**, 379–386.

Short Specialist Review

Comparisons with primate genomes

Mariano Rocchi and Nicoletta Archidiacono
University of Bari, Bari, Italy

The complete sequence of human, mouse, and rat genomes is now available, and sequence comparison has started to unveil the forces that shaped mammalian genomes. Whole-genome comparison among these three genomes is very interesting, but almost fruitless if used to delineate the recent evolutionary history of the human genome. Comparison of our genome with those of our closest relatives, the primates, is, with no doubt, much more rewarding in this respect. Unfortunately, only the chimpanzee genome, at draft level, is available.

Other approaches, however, have been exploited to unveil aspects of our recent evolution. Cytogenetics offers the opportunity of looking at genomes through variations of the pieces in which genomes are organized, the chromosomes. Cytogenetics came of age in the sixties when some technical achievements made it possible to culture peripheral human lymphocytes and prepare a large number of good-quality metaphases. Chromosomal count became an easy task and chromosomal syndromes as well as karyotype number of some primate species could be determined. It was noted, for instance, that the chimpanzee had 48 chromosomes, two more than man. The banding techniques era in seventies introduced a more powerful analytical tool in chromosome identification, and De Grouchy first suggested that the chromosomal number difference between humans and chimpanzee was due to the fact that two ancestral chromosomes, conserved in chimpanzee, fused to generate the human chromosome 2.

Dutrillaux did an extensive use of banding techniques in an attempt at a comprehensive view of karyotype evolution in primates, which are divided into prosimians, Platyrrhini (New World Monkeys, NWM), and Catarrhini (Old World Monkey and apes). Apes include lesser apes (gibbons) and great apes (orangutan, gorilla, chimpanzee). Yunis and Prakash (1982) reported a detailed analysis of chromosome similarities among great apes. This paper, which appeared in *Science*, defined all the changes detectable at high-resolution banding level. Figure 1 shows a comparative karyotype of human and great apes (chimpanzee, gorilla, and orangutan).

Heterochromatic blocks of rapidly evolving satellite DNA can differentiate even close species. For instance, heterochromatic DNA blocks are associated only with the centromeric domains of autosomes in humans, while in great apes, they are also telomerically and interstitially located (Figure 2). In addition, differences other

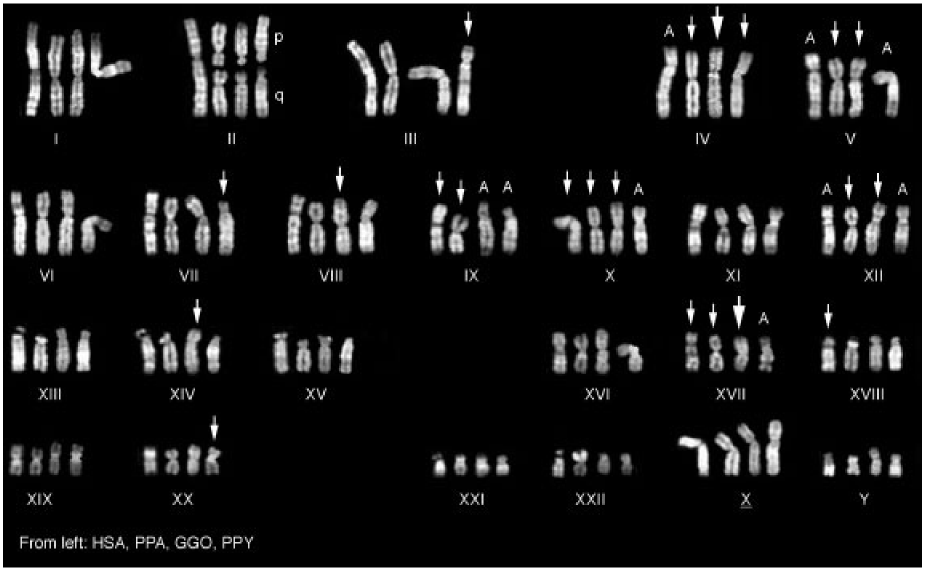


Figure 1 Comparative karyotype of human and great apes (pigmy chimpanzee, *Pan paniscus*, PPA; gorilla, *Gorilla gorilla*, GGO; Borneo orangutan, *Pongo pygmaeus pygmaeus*, PPY). Pigmy chimpanzee (bonobo) is one of the two chimpanzee species. The other one is the common chimpanzee (*Pan troglodytes*, PTR). They have an almost identical karyotype. The QM banded chromosomes are numbered according to the phylogenetic nomenclature (Roman numbers), which was introduced to better show the correspondence to human chromosomes, which is not evident using the specific chromosome number that is given according to the size of the chromosome. Note the chromosome 2 that resulted from the fusion of two ancestral acrocentric chromosomes. "A" stands for ancestral form; arrows point to derivative chromosomes, which usually differ from the ancestral form by an inversion. The ancestral form is not indicated when the chromosomes are identical. Minor changes were not considered. Gorilla shows the only translocation present in great apes (big arrow), involving chromosomes V and XVII. Some differences are only due to large heterochromatic blocks present in great apes (see Figure 2)

than rearrangements or heterochromatic blocks can strongly affect the structure of chromosomes. It has been recently shown, in this respect, that repeat-expansion in humans can account for up to 20% of DNA content difference between lemur and humans (Liu *et al*., 2003). Banding patterns analysis, therefore, could be misleading, especially when the species under study are not closely related.

The Fluorescence In Situ Hybridization (FISH) techniques introduced a completely new tool in cytogenetic investigations, and, being based on sequence homology, solved many of the problems posed by chromosomal similarities just based on visual inspection. These studies were pioneered by the J. Wienberg group and the T. Cremer (Jauch *et al*., 1992) groups in Germany. Figure 3 is an example of a cohybridization experiment using human painting probes on a gibbon metaphase. This single experiment discloses that gibbon chromosomes are very rearranged with respect to humans. The use of whole-chromosome paints and partial-chromosome paints, indeed, has the advantage of producing rapid results, but lacks resolution, and marker order remains frequently undetermined.

The human genome sequence generated by the public Consortium was achieved using a "hierarchical" approach. A minimal collection of overlapping clones,

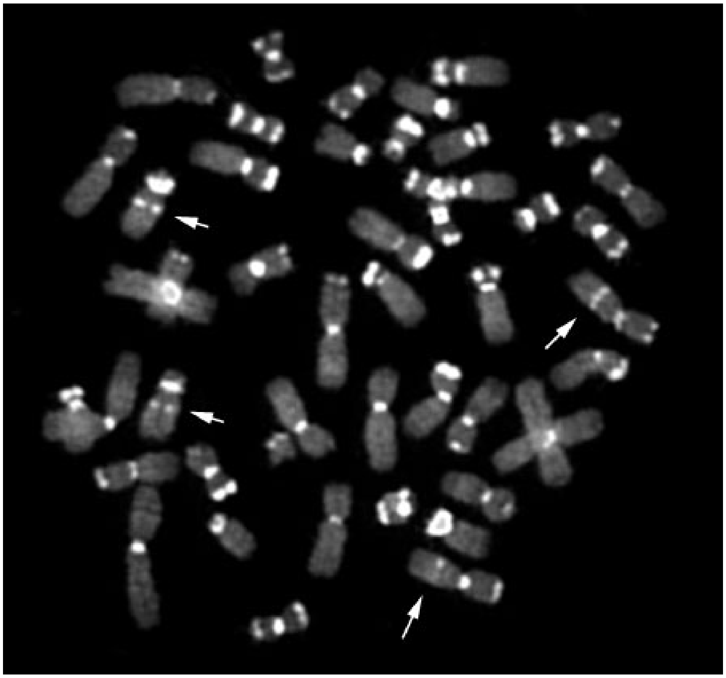


Figure 2 Metaphase form common chimpanzee (*Pan troglodytes*, PTR), showing heterochromatic blocks located on centromeres, telomeres, and in interstitial loci of chromosomes VII (arrow) and XIII (short arrow). The chromosomes are DAPI stained after the denaturation and hybridization procedure used for FISH (Fluorescence In Situ Hybridization)

covering the entire genome, was sequenced. As an intermediate step toward the definition of this "golden path", thousands of BAC/PAC clones were fingerprinted and end-sequenced. As a result, even if a specific clone was not chosen for complete sequencing, its position on the human sequence is defined at single base-pair resolution. In conclusion, for each region of the human genome, several overlapping probes are available and can be obtained from various sources, from the Pieter de Jong laboratory (Oackland) in particular, where most of the BAC/PAC probes were generated. The University California Santa Cruz browser (**http://genome.ucsc.edu**) specifically shows this large collection of end-sequenced clones. These resources are invaluable for molecular cytogenetics because they produce a clean and locus-specific signal in FISH experiments.

These resources have been used extensively to study marker order conservation in primates and led to the discovery of an unprecedented biological phenomenon: the evolutionary centromere repositioning. That is, the movement of the centromere along the chromosome not accompanied by any chromosomal rearrangement that would account for its movement. The first evolutionary centromere repositioning example was documented on chromosome 9 (Montefalcone *et al*., 1999). At present, several other examples have been reported (Ventura *et al*., 2001; Eder *et al*., 2003). The phenomenon appears to not be limited to primates (Band *et al*., 2000).

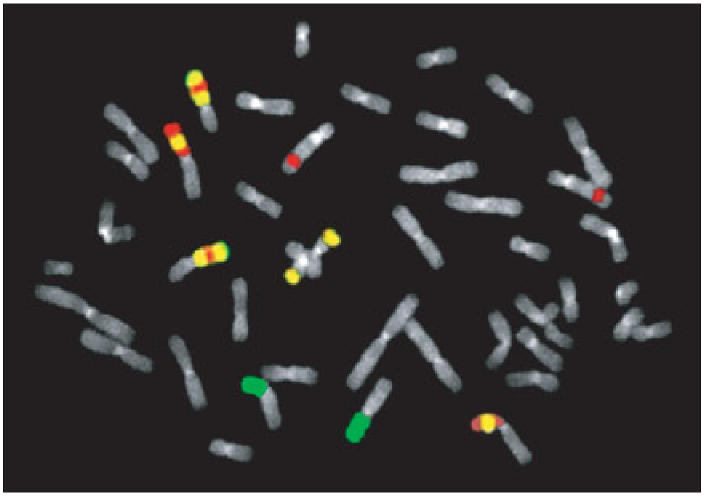


Figure 3 FISH experiment on a lar gibbon (*Hylobates lar,* HLA) metaphase using whole-chromosome paint specific for chromosome 3 (red) and a partial-chromosome paint specific for the short arm of this chromosome (3p) and for the short arm of chromosome X (green). 3p is part of the entire chromosome. 3p chromosomal segments of HLA are stained with red and green and, therefore, they appear yellow in the merged image. Portion of 3q are just red and Xp remain green. The experiment shows that HLA sequences corresponding to the human chromosome 3 are scattered in four different chromosomes but organized in at least eight distinct syntenic blocks

A centromere repositioning implies the inactivation of an ancestral centromere and the seeding of a neocentromere. The available examples of centromere inactivation suggest a common scenario accompanying their silencing. The strong constraint against recombination acting on normal centromeres progressively weakens following inactivation. Very likely, nonallelic homologous exchanges trigger a rapid elimination of satellite DNA, while pericentromeric duplications are dispersed over a longer range, up to 10 Mb in size. Similarly, evolutionary neocentromeres rapidly progress toward the “normal” complex organization typical of a mammalian centromere: they acquire a large block of centromeric heterochromatin and pericentromeric segmental duplications.

Neocentromeres are fully functioning centromeres that are formed ectopically, most frequently on acentric fragments generated as a result of cytogenetic rearrangements whose mitotic survival was rescued by neocentromere activation. The first well-documented neocentromere case was described in 1997, on chromosome 10 (du Sart *et al*., 1997). Since then, more than 50 neocentromeres have been described, many of which are clustered in clear hotspots, like the one at the region 15q24-26 (Amor and Choo, 2002). The evolutionary history of the ancestral 14/15 association disclosed an ancestral centromere that inactivated about 25 million years ago, after the great apes/Old World monkeys diverged. This inactivation has followed a noncentromeric chromosomal fission of an ancestral chromosome, which gave rise to phylogenetic chromosomes 14 and 15. Mapping of the ancient centromere and two neocentromeres in 15q24-26 established that the neocentromere domains map to duplicons, copies of which flank the centromere in Old World

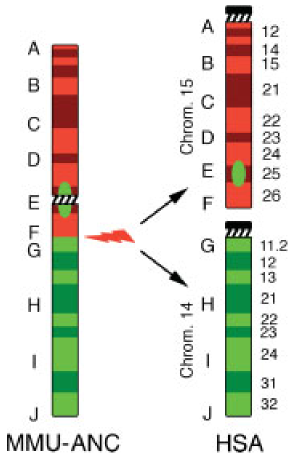


Figure 4 The diagram shows the evolutionary history of the ancestral 14/15 association. A fission event, which occurred about 25 million years ago, gave rise to chromosomes 14 and 15. The ancestral 14/15 association is conserved in macaque (*Macaca mulatta*, MMU). The event triggered the birth of two neocentromeres and the inactivation of the old centromere. Letters on the left are probes (BAC clones) used to define marker arrangement along the chromosomes

Monkey species that bear the ancestral 14/15 association (Figure 4) (Ventura *et al*., 2003). This suggests that the neocentromere at 15q24-26 may be due to the persistence of duplications accrued within the ancient pericentromere.

This is the first clear sample of an association between neocentromeres and ancestral centromeres. At present, we have a very rough picture of the evolutionary history of human chromosomes. Future studies are very promising in further clarifying this intriguing connection.

References

Amor DJ and Choo KH (2002) Neocentromeres: role in human disease, evolution, and centromere study. *American Journal of Human Genetics*, **71**, 695–714.

Band MR, Larson JH, Rebeiz M, Green CA, Heyen DW, Donovan J, Windish R, Steining C, Mahyuddin P, Womack JE, *et al*. (2000) An ordered comparative map of the cattle and human genomes. *Genome Research*, **10**, 1359–1368.

du Sart D, Cancilla MR, Earle E, Mao J, Saffery R, Tainton KM, Kalitsis P, Martin J, Barry AE and Choo KHA (1997) A functional neo centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. *Nature Genetics*, **16**, 144–153.

Eder V, Ventura M, Ianigro M, Teti M, Rocchi M and Archidiacono N (2003) Chromosome 6 phylogeny in primates and centromere repositioning. *Molecular Biology and Evolution*, **20**, 1506–1512.

Jauch A, Wienberg J, Stanyon R, Arnold N, Tofanelli S, Ishida T and Cremer T (1992) Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8611–8615.

Liu G, Program NC, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED and Eichler EE (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Research*, **13**, 358–368.

Montefalcone G, Tempesta S, Rocchi M and Archidiacono N (1999) Centromere Repositioning. *Genome Research*, **9**, 1184–1188.

Ventura M, Archidiacono N and Rocchi M (2001) Centromere emergence in evolution. *Genome Research*, **11**, 595–599.

Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N and Jackson MS (2003) Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Research*, **13**, 2059–2068.

Yunis JJ and Prakash O (1982) The origin of man: a chromosomal pictorial legacy. *Science*, **215**, 1525–1530.

Short Specialist Review

Transcriptional promoters

Wyeth W. Wasserman
University of British Columbia, Vancouver, BC, Canada

1. Introduction

Transcription, the first step in the flow of genetic information from DNA to RNA to protein, acts as the gatekeeper controlling the influence of genes upon the phenotype of cells. When the three-dimensional structure of chromatin and the presence of appropriate catalytic proteins are permissive, biochemical protein machinery is assembled within regions of genes termed promoters. While this summary is focused on human gene transcription, and more specifically to transcription mediated by RNA polymerase II, the properties of transcription in other systems will be briefly addressed.

2. Biochemistry of transcript initiation

The biochemical mechanisms of transcription of human protein-coding genes by RNA polymerase II are among the most closely studied of any cellular process. As such the process is richly described in dedicated textbooks (Latchman, 2003) and detailed review articles (Butler and Kadonaga, 2002).

Cis-regulatory elements in the promoter of a gene are bound by trans-acting proteins (Figure 1). These elements can include the broadly recognized TATA box sequence that frequently occurs approximately 30 bp before a site of transcript initiation (the Initiator site), as well as the common downstream promoter element. At these core elements, the basal machinery of transcription is formed, including such basal protein complexes as TFIIA, TFIIB, and TFIID. Studies have revealed diverse gene-to-gene variation in the specific characteristics of transcript initiation, so individual promoters may be composed of different combinations of the elements (and consequently more dependent upon specific subsets of trans-acting proteins). There are indications that some of the trans-acting proteins act in specific cellular contexts, in a manner similar to the "sigma" proteins that control transcription in bacteria (Hochheimer and Tjian, 2003). For clarity, there is no absolute requirement for either an upstream TATA element, the Initiator element, or for the downstream element.

On the basis of the characteristic elements, minimal promoters are often attributed to the region from −50 bp to +50 bp on either side of the transcription start site.

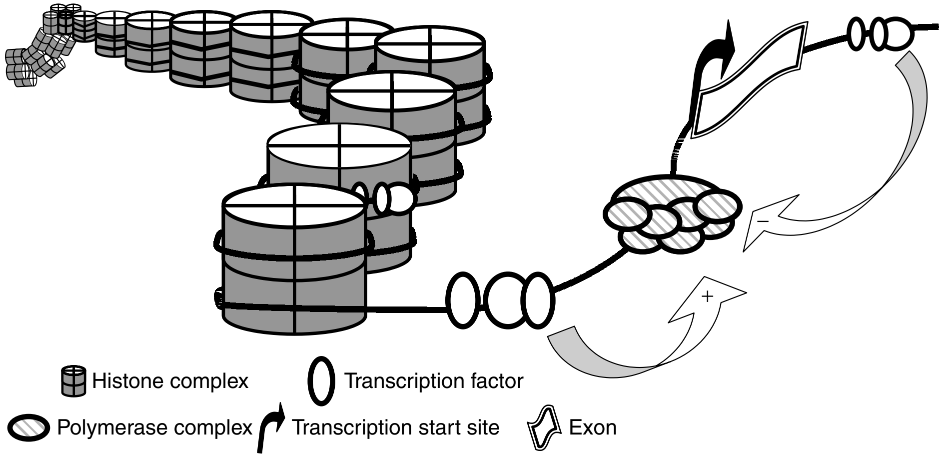


Figure 1 Transcriptional promoters: Transcription rates are defined by the activation and repression signals from regulatory modules composed of clusters of transcription factors bound to DNA. While illustrated as naked DNA, the true biochemical mechanisms occur in the context of higher orders of histone-DNA chromatin structure

However, in recent years, it has become apparent that many genes, particularly genes lacking a strong consensus TATA motif, have multiple transcription start sites within the same promoter, destroying the paradigm of the +1 position.

Of equal importance to local cis-regulatory sequences are epigenetic regulatory mechanisms governing the structure of chromatin. DNA in active regions of transcription can be subject to distinctive patterns of methylation (*see* Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1). Within intrapromoter regions, methylation is substantially reduced compared to the rest of the human genome. As methylated cytosines (on ring position five) 5′ to guanine nucleotides tend to undergo deamination to form a TG dinucleotide, there is a strong selection against CG dinucleotides throughout much of the genome. In many promoter regions, trans-acting proteins act to maintain an open chromatin conformation by altering methylation, which results in a higher retention of CG dinucleotides within these promoter regions. Not all promoters are linked to the retention of CG dinucleotides, and there are some indications that promoters with CG signatures tend to utilize a wide range of initiation sites (Butler and Kadonaga, 2002).

3. Genomics and promoter identification

The identification of promoter locations (*see* Article 22, **Eukaryotic regulatory sequences**, Volume 7) is critical to deciphering the regulatory programs of cells. Most sequence-specific transcription factors bind to regulatory elements proximal to promoters (Hannenhalli and Levy, 2001), so detailed analysis of regions flanking promoters (*see* Article 16, **Searching for genes and biologically related signals in DNA sequences**, Volume 7) can reveal important insights into the biochemical mechanisms governing expression. Furthermore, many genes contain alternative

promoters, which can contribute to the creation of different open reading frames directly or indirectly via promotion of specific alternative splicing events. Thus, the identification of functional promoters in the human genome remains an imperative.

Techniques for promoter identification have matured. Traditionally, molecular methods have relied on the identification of the longest transcript based on the creation of complementary DNA (cDNA) replicates of mRNA. Such methods have included run-on assays with oligo priming followed by extension by reverse transcriptase, and more recently by PCR with long cDNA collections in a process termed RACE. These methods are constrained by the capacity of the reverse transcriptase enzyme to successfully produce full-length DNA copies; however, the enzyme is easily blocked by RNA structure. Thus, most cDNAs do not extend to the original 5′ terminal position.

On the basis of the 5′ terminal cap structure of mRNA, a new technique using an antibody to recognize the cap has been developed to recover full-length transcripts (Okazaki *et al*., 2002). Originally applied to the generation of full-length cDNA copies, the method has recently been modified to allow for the recovery of 5′ terminal tags in a process analogous to serial analysis of gene expression (SAGE) (*see* Article 103, **SAGE**, Volume 4), a new method termed cap analysis of gene expression or CAGE (Shiraki *et al*., 2003). Early observations from these high-throughput studies indicate that many genes are transcribed from a pool of start sites spread over the promoter region.

The specific TSS is determined by the constraints upon the polymerase complex – TATA and DPE contribute to positioning – when they are well defined, a single TSS is primarily used, however, if they are not well defined, a stochastic positioning of the polymerase results in a wide range of TSS. In short, the PCR-based RACE techniques give misleading indications that the 5′ most position of a transcript is of the greatest importance, while such positions may be rarely used for initiation in the case of a stochastic process.

4. Bioinformatics methods

Bioinformatics-based approaches for promoter identification (*see* Article 19, **Promoter prediction**, Volume 7) loosely fall into three categories: ab initio prediction (*see* Article 66, ***Ab initio*** **structure prediction**, Volume 7) of specific transcription start sites; identification of regions in a genome likely to contain promoters; and mapping of known transcripts onto genome sequences.

On the basis of the properties of regulatory elements in minimal promoters, a diverse set of algorithms has been created for the ab initio prediction of initiation sites. A broad assessment (Fickett and Hatzigeorgiou, 1997) demonstrated that the performances of most algorithms were comparable to predictions of TATA elements generated with a position-specific scoring matrix (PSSM) (Bucher, 1990). The PSSM models the binding preferences for the TATA Binding Protein, generating a single quantitative score that is analogous to binding energy. The TATA profile renders a prediction on the order of one per 500 bp, providing little motivation to conduct laboratory experiments based solely on a prediction. The poor predictive performance of the transcription start site prediction methods reflects the diversity

of promoter structures and the above-mentioned observation that many promoters contain multiple initiation sites.

A new generation of bioinformatics algorithms address a slightly different goal, the identification of regions in a gene likely to contain promoters. These tools have proved highly successful in identifying the likely promoter regions in genes. While the implementation details vary widely, the algorithms generally identify CpG islands within a gene; those regions containing CG dinucleotides at the expected frequency based on the mononucleotide frequency of G and C. These methods perform poorly for the prediction of promoters that do not have associated CG retention.

Ultimately, the best methods for promoter identification are based on transcript data (Bajic *et al*., 2004). The compilation of CAGE data, EST sequences (*see* Article 78, **What is an EST?**, Volume 4), and full-length transcripts are rapidly building to a robust repository. The reference database dbTSS (Suzuki *et al*., 2004), which combines evidence from diverse transcripts, has emerged a primary source of human promoter data.

5. Considerations in other systems

Transcription by human polymerase II is merely one example of the range of transcription mechanisms used across organisms. In particular, operons are used extensively (*see* Article 20, **Operon finding in bacteria**, Volume 7). By producing multiple genes in a single polycistronic transcript, bacteria can minimize the need for regulatory control sequences while ensuring that genes that act together are produced concurrently.

In yeast, most regulatory sequences are located in close proximity to the transcribed gene and introns are not present; most promoters can be predicted simply by identifying open reading frames and selecting sequences adjacent to promoters.

6. Future

The activity of promoters is determined by a combination of DNA accessibility and the influence of trans-acting activator proteins. Our understanding of the structural constraints on promoters remains limited. Present research is suggesting that active genes may have regulated positions in the nucleus that influence activity. As well, there is growing evidence of a complex set of silencing mechanisms to maintain many promoters in an inactive state. The relationship between three-dimensional structure and promoter activity will be revealed in the coming years.

Finally, there is growing evidence of extensive variation between individuals in the expression of genes. Genetic sequence variation (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4) within promoters could have dramatic influence on the activity of promoters, and likely contributes to phenotypic diversity.

References

Bajic VB, Tan SL, Suzuki Y and Sugano S (2004) Promoter prediction analysis on the whole human genome. *Nature Biotechnology*, **22**(11), 1467–1473.

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, **212**(4), 563–578.

Butler JE and Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, **16**(20), 2583–2592.

Fickett JW and Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Research*, **7**(9), 861–878.

Hannenhalli S and Levy S (2001) Promoter prediction in the human genome. *Bioinformatics,* **17**(Suppl 1), S90–S96.

Hochheimer A and Tjian R (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes & Development*, **17**(11), 1309–1320.

Latchman D (2003) *Eukaryotic Transcription Factors*, Academic Press.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**(6915), 563–573.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26), 15776–15781.

Suzuki Y, Yamashita R, Sugano S and Nakai K (2004) DBTSS, DataBase of transcriptional start sites: progress report 2004. *Nucleic Acids Research*, **32**(Database issue) D78–D81.

Short Specialist Review

Human microRNAs

Sam Griffiths-Jones
The Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

In 2001, three groups published independent reports of the discovery of a large class of tiny noncoding RNAs (ncRNAs, *see* Article 27, **Noncoding RNAs in mammals**, Volume 3) in worm, fly, and human (Lagos-Quintana, 2001; Lee, 2001; Lau, 2001). The ncRNAs were named microRNAs (miRNAs) and the reports generated widespread interest. The founder member of the miRNA class had been identified some years earlier: the *lin-4* gene controls the timing of *Caenorhabditis elegans* larval development. Ambros and colleagues showed that *lin-4* codes for two small RNA products, one around 22 nucleotides (nt) in length, and the second around 60 nt (Lee, 1993). The latter was proposed to be a precursor for the shorter sequence. Further work showed that the mature lin-4 RNA was able to bind complementary regions of the lin-14 mRNA to repress its translation (Wightman, 1993; Lee, 1993). Ideas that lin-4 was an anomaly, specific to the worm, were dispelled by the identification of a second short RNA, let-7, found to regulate the transition from late-larval to adult in *C. elegans* (Reinhart, 2000). let-7 was found to be widely conserved in humans, flies, and many other bilateral animals (Pasquinelli, 2000). Subsequent discoveries have put the total number of known miRNAs in worms, flies, mammals, fish and plants above 1000, many of which have been shown to be expressed in tissue specific patterns (reviewed in He, 2004). miRNAs are implicated in gene regulatory roles in processes as important and diverse as development, fat metabolism, and oncogenesis (reviewed in Bartel, 2004; Ambros, 2004).

I briefly summarize the current understanding of miRNA biogenesis and message targeting – the reader is directed to many excellent review articles cited here for more in-depth coverage of these topics. This review then focuses on the human miRNA gene complement, with emphasis on conservation and gene organization.

2. Biogenesis and target regulation

Maturation from primary transcript to mature miRNA has been thoroughly characterized over the past two years (see Figure 1). The mature miRNA (often designated miR), around 21 nt in length, is excised from a larger precursor transcript (around

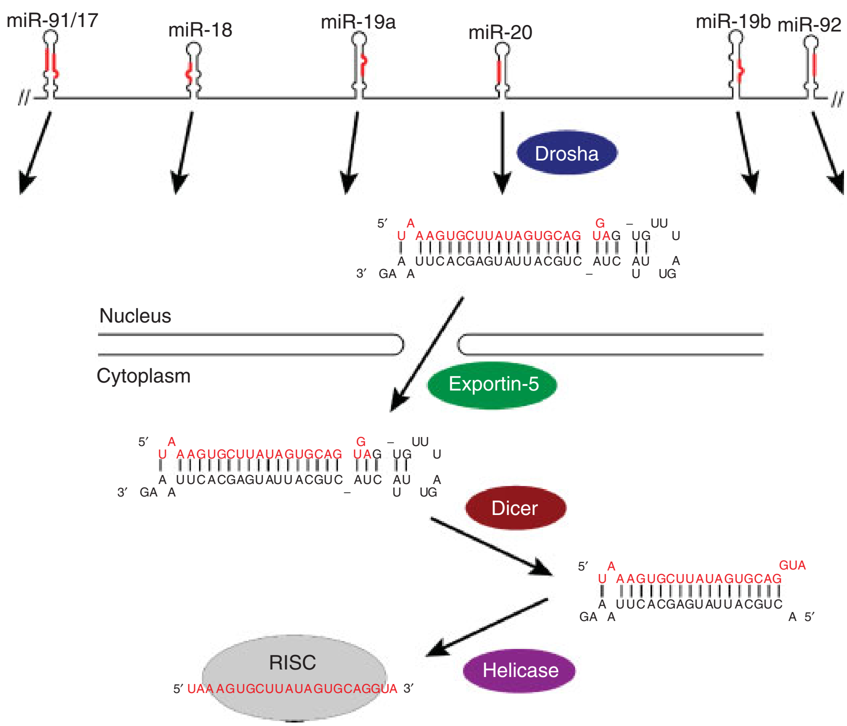


Figure 1 Summary of the biogenesis of a miRNA cluster (see text for description of processing steps)

70 nt in animals), which adopts a hairpin conformation by intramolecular base pairing. Usually, only one arm of the hairpin gives rise to a mature miRNA sequence. The opposite arm, designated miR* is sometimes observed at a lower frequency in cloning studies. The precursor hairpin (termed pre-miRNA) is itself processed from a larger primary transcript (termed pri-miRNA), which may be kilobases in length. The first processing step of pri-miRNA to pre-miRNA is catalyzed by an RNase III endonuclease, DROSHA, in the nucleus, and defines one end of the mature miR (Lee, 2003). The excised precursor is exported to the cytoplasm by exportin-5 (Kim, 2004) where the other end of the mature ~21 nt miRNA is cut by DICER, another RNase III endonuclease (Lee, 2002). The resulting miR/miR* duplex has a characteristic two nucleotide 3′ overhang at each end. The duplex is unwound by an unknown helicase before the mature miR is recruited into the RNA-induced silencing complex (RISC), which targets it to complementary sequences in untranslated regions (UTRs) of other mRNAs. The current hypothesis is that miRNAs may have two modes of action, depending in part on the degree of complementarity to the target regions in the message. Exact base pairing causes the messenger transcript to be degraded by RNAi, while more extensive mismatching causes translational repression (reviewed in Meister, 2004). Only a handful of mammalian miRNA targets have been experimentally confirmed (reviewed in Ambros, 2004; He, 2004).

3. Human miRNA genes

The miRNA Registry (Griffiths-Jones, 2004) provides a catalog of experimentally determined miRNA sequences and their predicted homologs in closely related organisms. At the time of writing, over 200 miRNA genes have been discovered in human, mouse, and rat, giving rise to around 190 unique mature miRNAs. Well over half of these sequences have experimental evidence of expression in human cell lines. Some miRNAs are expressed from several hairpin precursors, transcribed from different genomic loci.

By statistical analysis of the results from computational predictions and cloning studies, Lim *et al*. (2003) have estimated the total number of miRNA genes in the human genome to be not more than 255. Although this is perhaps fewer than expected, miRNAs still account for around 1% of the total human gene count, in similar proportions to some other large regulatory gene families, such as homeobox transcription factors and KRAB box zinc fingers (International Human Genome Sequencing Consortium, 2001). This also suggests that, in a very short space of time, the experimental biologists have identified well over three-quarters of the mammalian miRNA gene complement. However, if a subset of miRNA genes proves difficult to clone, perhaps due to low abundancy or specific spatial or temporal expression, and is missed by computational prediction methods, the total number of human miRNAs may eventually be shown to be larger.

The comprehensive nature of the known miRNA set has encouraged the first analyses of genomic contexts and cross-species conservation. The data show that a small number of miRNAs, including let-7, are well conserved from worms and flies through fish, birds, and mammals (Pasquinelli, 2000). Almost all human miRNAs are conserved in order and orientation in the mouse and rat genomes, and nearly 40% map to syntenic regions of the draft chicken genome (International Chicken Genome Sequencing Consortium, 2004). This seems to be in contrast with other classes of ncRNAs, such as the spliceosomal RNAs, whose genes have not been well conserved in syntenic regions of even very closely related genomes.

4. miRNA gene structure

Examining the genomic context of miRNA sequences shows a variety of gene structures and may suggest subtly different methods of biogenesis. The first miRNA reports demonstrated that some genes are arranged in clusters – in *C. elegans*, seven miRNA genes, numbered miR-35 through miR-41, cluster within 1 kb on chromosome II (Lau, 2001). It is suggested that such clustering may be indicative of polycistronic expression, with seven miRNAs processed from the same primary transcript. Indeed, the first primary transcript to be thoroughly characterized contains three mammalian miRNAs: miR-23a, miR-27a, and miR-24 (Lee, 2004). This work experimentally confirms previous predictions that miRNA genes are transcribed by RNA polymerase II (pol II), and that the primary transcripts are, like protein-coding messages, capped and polyadenylated. Moreover, the miR-23a∼27a∼24 primary transcript is long – almost 2 kb. Analysis of the human miRNA set shows that as many as 70 miRNA genes are located within 2 kb of another miRNA, and are

thus candidates for coexpression from the same transcript. The arrangement of the largest cluster of six miRNAs in human is shown in Figure 1. The evolution of this cluster has been the subject of detailed investigation: mammals appear to have three paralogous copies of the miR-17 cluster – on chromosomes 13, X, and 3 – while fish genomes contain four copies (Tanzer, 2004).

The second major class of miRNAs are processed from the introns of other transcripts, usually coding for proteins, but sometimes for noncoding RNAs. As many as 80 human miRNAs are located in introns of annotated protein-coding transcripts (Rodriguez, 2004), usually in the sense orientation, implying utilization of host gene transcription rather than dedicated promoter sequences. Embedding in a host transcript provides a convenient mechanism for coordinate expression of miRNA and protein. Many of the hosts fall into paralogous gene families (Weber, 2005). For example, the human genome contains four pantothenate kinase genes, numbered PANK1 to PANK4. miR-103 appears to be expressed from two predicted stem-loop precursors, located in intron 5 of both PANK2 (chr 20) and PANK3 (chr 5). Intron 5 of PANK1 (chr 10) contains the closely related miR-107. No experimentally confirmed miRNA overlaps the PANK4 transcript. Other miRNAs may be expressed from unrelated hosts, complicated further by different tissue expression profiles. For example, miR-7 has three predicted genes in the human genome. miR-7-3 is located in an intron of *pituitary gland specific factor 1*, a gene expressed exclusively in the pituitary gland (Rodriguez, 2004). An alternative precursor locus is located in an intron of a ubiquitously expressed gene, *heterogeneous nuclear ribonucleoprotein K* (*hnRPK*), in an miR/host gene relationship that is conserved in *Drosophila* (Bartel, 2004). miR-7-2 does not overlap any annotated transcripts and thus may be expressed from its own promoter under different regulatory conditions. The three *mir-7* gene contexts are shown schematically in Figure 2. Almost all identified human miRNA/host gene relationships are conserved in mouse and rat.

A number of miRNAs may also overlap with annotated mRNA-like transcripts that do not appear to code for a protein (Rodriguez, 2004). miR-15a and miR-16 are located in the same intron of the well-annotated DLEU2 noncoding transcript, in a region implicated in B cell chronic lymphocytic leukemias (CLL). The intron-encoded miRNAs have been shown to be deleted or downregulated in more than two-thirds of CLL cases (Calin, 2002). miR-206 appears to be transcribed as part of the longer 7H4 transcript, which is selectively found at the neuromuscular junction (Velleca, 1994). The data suggest that some miRNA host genes may be so-called inside-out genes, where the host transcript is expressed purely to yield the processed RNA from the introns, similar to some small nucleolar RNAs.

5. Conclusions

The past three years have seen staggering advances in our understanding of miRNAs, their genes and biogenesis. Estimations of gene counts suggest that experimental verification, sometimes effectively combined with computational prediction, have revealed the majority of miRNA genes in a number of model organisms, including in mammals, in an incredibly short space of time. Recent studies also shed light on the implications of miRNA discovery on the complexity of gene

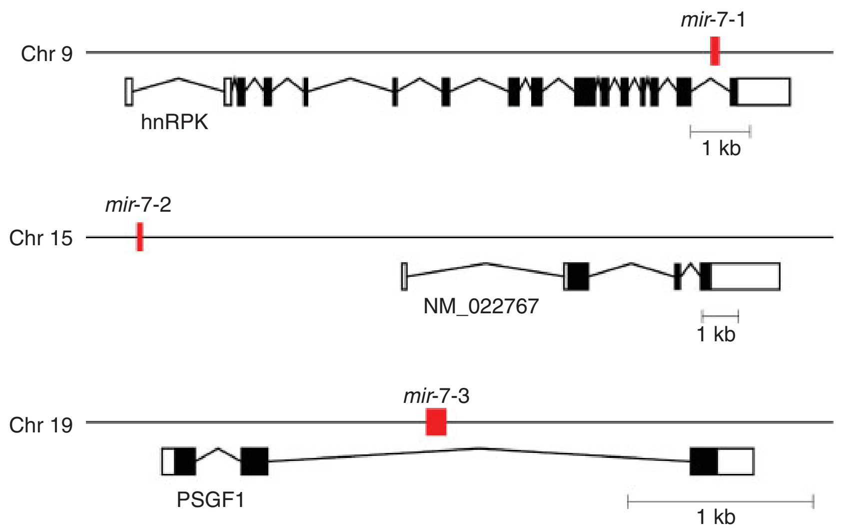


Figure 2 The three human *mir-7* genes on chr 9 (82.04 Mb, reverse strand), chr 15 (86.88 Mb), and chr 19 (4.72 Mb). All sequences are depicted from the 5′ to the 3′ end. The transcripts shown are *heterogeneous nuclear ribonucleoprotein K* (*hnRPK*), a putative nucleolar exonuclease (REFSEQ: NM_022767), and *pituitary specific growth factor 1* (*PSGF1*). Black boxes are exons, open boxes show untranslated regions. The miRNA precursor hairpin is shown in red

regulation, especially during development. A number of disease states, including some cancers, may yet involve miRNAs.

The diversity of miRNA gene structures provides some tantalizing examples of genes with distinct and parallel output, and hints at complex regulatory networks. A long primary transcript may express a single miRNA or multiple distinct miRNAs in a polycistronic cluster. The same transcript may also encode a protein, or a spliced mRNA-like noncoding RNA. miRNAs expressed from autonomous transcripts or coexpressed with a protein may then regulate the expression of a wide range of genes, possibly including other miRNA host genes; clustered miRNAs may coordinately regulate the same targets, or participate in different pathways. The same mature miRNA may originate from multiple genomic loci, under independent regulatory control, and it is also clear that any single targeted message may be regulated by more than one miRNA. Alternative splicing of host transcripts is likely to add a further layer of complexity to the story. These findings have fundamental implications for our understanding of what constitutes a gene, and how parallel output may be involved in vital regulatory mechanisms. We can expect the next three years to see equally significant advances in understanding of gene regulation by this extraordinary class of ncRNA, as their targets are identified and described.

6. Data sources

Discussion of miRNA gene counts and genomic arrangement is based on data from the miRNA Registry, release 5.0 (**http://www.sanger.ac.uk/Software/Rfam/**

mirna/). Genome assemblies and gene predictions relate to EnsEMBL human build 24.34e.1 and mouse build 24.33.1, available from (**http://www.ensembl.org/**).

Acknowledgments

Sam Griffiths-Jones is funded by the Wellcome Trust.

References

Ambros V (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, *et al*. (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15524–15529.

Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Research*, **32**, D109–D111.

Kim VN (2004) MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends in Cell Biology*, **14**, 156–159.

He L and Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews. Genetics*, **5**, 522–531.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

International Human Genome Sequencing Consortium (2001) Initial Sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lagos-Quintana M, Rauhut R, Lendeckel W and Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lau NC, Lim LP, Weinstein EG and Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, *et al*. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.

Lee RC and Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.

Lee RC, Feinbaum RL and Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lee Y, Jeon K, Lee JT, Kim S and Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, **21**, 4663–4670.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH and Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, **23**, 4051–4060.

Lim LP, Glasner ME, Yekta S, Burge CB and Bartel DP (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.

Meister G and Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P, *et al*. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR and Ruvkun G (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.

Rodriguez A, Griffiths-Jones S, Ashurst JL and Bradley A (2004) Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, **14**, 1902–1910.

Tanzer A and Stadler PF (2004) Molecular evolution of a microRNA cluster. *Journal of Molecular Biology*, **339**, 327–335.

Velleca MA, Wallace MC and Merlie JP (1994) A novel synapse-associated noncoding RNA. *Molecular and Cellular Biology*, **14**, 7095–7104.

Weber MJ (2005) New human and mouse microRNA genes found by homology search. *FEBS Journal*, **272**, 59–73.

Wightman B, Ha I and Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, **75**, 855–862.

Short Specialist Review

Endogenous retroviruses

Jens Mayer
University of Saarland, Homburg, Germany

Vertebrate genomes harbor considerable sequence portions of retroviral origin that are stably inherited genome components. Retroviruses typically infect somatic cells of a host organism. They reverse transcribe their RNA genome to double-stranded DNA, and integrate the DNA into the host cell genome, forming a so-called provirus, that serves to produce RNA transcripts and retroviral proteins for further infectious virus. In the evolutionary past, different retroviruses formed proviruses in the genome of germ-line cells. Since germ-cell genomes comprise the inherited genetic material and are precursors to somatic cells, incorporated proviruses could have been transmitted through generations and became part of somatic cells. The stably inherited proviruses are called *endogenous retroviruses* (ERVs). During evolution, ERVs were also transmitted to new species. In most cases, exogenous counterparts of ERVs are no longer present.

ERVs often resemble the former exogenous retrovirus proviral structure. Two long terminal repeats (LTRs) enclose *gag*, *protease*, *polymerase*, and *envelope* genes. The LTRs represent autonomous units for transcription regulation, initiation, and termination, while retroviral genes produce structural proteins and enzymatic activities required during retroviral replication. Owing to the long-time presence in the host genome, ERVs often accumulated nonsense mutations, rendering them coding-deficient and transcriptionally silent. Mutational events often deleted considerable portions of proviral genes. Several exceptions exist regarding transcriptional activity and coding-capacity, though (see below). Furthermore, proviruses were frequently reduced to so-called "solitary LTRs" by homologous recombination between proviral 5′ and 3′ LTRs. Solitary LTRs regularly outnumber the actual provirus count for a given ERV family by far (Gifford and Tristem, 2003).

Distinct ERV families can be identified in a genome. Members of a family are more similar in sequence among each other than to other ERV families. Thus, endogenization of diverse exogenous retroviruses occurred many times during evolution. Most ERV families subsequently increased proviral copy numbers in the host genome, that is, additional proviruses formed after initial provirus formation, and were likewise fixed in the population. It is thought that retroviral RNA was transcribed from a proviral element and was subsequently reverse transcribed and integrated into the same genome. Some ERV families expanded to several thousand proviral copies that way. Formation of new proviruses seems very low or does not occur in humans. Differences in proviral integration patterns can only be observed

on an evolutionary timescale. Various human proviruses are missing in, for instance, the chimpanzee genome, that evolutionarily separated about 5 million years ago (Buzdin *et al*., 2002). The mouse genome seems to hold more active ERV families, and some proviral loci in the mouse can be expressed as infectious virus. A recent analysis estimated approximately 10% of the mouse genome to be of retroviral origin (Mouse Genome Sequencing Consortium, 2002; *see also* Article 47, **The mouse genome sequence**, Volume 3).

The human genome harbors a variety of endogenous retroviruses, commonly referred to as HERVs (for human ERVs), that were studied to some extent. In fact, 8% of the human genome mass is estimated to be derived from retroviral sequences (International Human Genome Sequencing Consortium, 2001). This portion includes retroviral sequence families composed of more or less intact proviruses (HERVs in sensu strictu), and families with HERV sequence portions. The latter are exemplified by SINE-R (a short interspersed repeated sequence and a longer HERV portion) and SVA elements (a composite retrotransposon containing SINE-R portions, a variable number of tandem repeats, and an Alu element). Manifold HERV families were identified in the past by various means; cross-hybridization with exogenous retrovirus probes, isolation of HERV RNA transcripts, degenerated primers, proviral sequence characteristics, and so on. Although there is no standardized nomenclature, HERV families are often named according to the amino acid specificity of the tRNA that primed reverse transcription of retroviral RNA by binding to the primer binding site (PBS). The tRNA's corresponding amino acid single letter code is appended to HERV. Recent surveys cataloged between 30 and 50 HERV families. The widely employed reference sequence database for repetitive elements, Repbase, lists more than 200 different HERV and LTR families (Jurka, 2000; *see also* Article 9, **Repeatfinding**, Volume 7). It appears that many HERV sequences and families await more detailed genomic characterization. Various HERV families were found to have entered the genome before the separation of existing primate lineages and species; about 30–40 million years ago. Other HERV families appear much older, as homologous sequences are also found, for instance, in the mouse. Evolutionary ages of ERVs can be estimated from sequence divergence between a provirus' 5′ and 3′ LTRs. They were identical in sequence when the provirus formed and accumulated random sequence mutations over time (Dangel *et al*., 1995). Alternatively, specific probes may be employed in hybridization experiments to identify ERV family-positive evolutionary clades.

There is strong experimental evidence of ERVs having profoundly impacted host genomes during evolution as a result of proviral amplification. One aspect concerns the influence on cellular genes when proviral LTRs – autonomous transcriptional units – integrated in close proximity (*see* Article 33, **Transcriptional promoters**, Volume 3). Several instances were revealed in the human genome where HERV LTRs act as major or alternative promoters for cellular genes, provide poly-A or splice signals (*see* Article 30, **Alternative splicing: conservation and function**, Volume 3 and Article 23, **Alternative splicing in humans**, Volume 7). HERV sequence portions are then present in cellular mRNAs. As an example, an HERV LTR is the dominant promoter for human β1,3-galactosyltransferase gene expression in the colon (Dunn *et al*., 2003). When positioned in close proximity to gene

promoters, LTRs may also act as transcriptional enhancers or suppressors, owing to their inherent transcription factor binding sites. Very likely, similar studies in the mouse will reveal mouse genes affected by mouse ERVs.

ERV-encoded proteins can also provide important functions for host organisms. Two ERV-derived genes, *Fv-4* and *Fv1*, have been identified in the mouse to effectively block infection and provirus formation, respectively, of specific exogenous retroviruses (Best *et al*., 1997). At least one HERV protein may be involved in physiological processes in human. In the developing placenta, expression of the HERV-W Env protein strongly increases during fusion of trophoblasts, forming a syncytiotrophoblast layer. HERV-W Env also fuses cell membranes when expressed in cell culture, thus producing syncytia. HERV-W Env therefore may be crucially involved in human placenta development (Blond *et al*., 2000). Remarkably and hitherto unexplained, the so-called HERV-K(HML-2) family propagated intact retroviral genes and functional proteins among its proviruses for several million years (Mayer *et al*., 1999). Antibodies specific for HERV-K(HML-2) Gag and Env proteins are frequently present at higher levels in germ cell tumor patients, potentially serving as a molecular marker for that tumor type (Sauter *et al*., 1996). The HERV-K(HML-2) encoded Rec protein associates with the promyelocytic zinc finger protein and may be involved in germ cell tumorigenesis (Boese *et al*., 2000).

Taken together, ERVs are important genome components. They boarded the host genome at one time period in evolution and often considerably amplified in copy number within the genome. By doing so, they affected the genome in diverse aspects and thus contributed significantly to the evolution of host genomes.

Further reading

Coffin JM, Hughes SH and Varmus HE (1997) *Retroviruses*, Cold Spring Harbor Laboratory Press: Plainview.

Sverdlov ED (1998) Perpetually mobile footprints of ancient infections in human genome. *FEBS Letters*, **428**, 1–6.

Volff JN (Ed.) Single topic issue: Retrotransposable elements and genome evolution. *Cytogenetic and Genome Research*, in press).

References

Best S, Le Tissier PR and Stoye JP (1997) Endogenous retroviruses and the evolution of resistance to retroviral infection. *Trends in Microbiology*, **5**, 313–318.

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F and Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *Journal of Virology*, **74**, 3321–3329.

Boese A, Sauter M, Galli U, Best B, Herbst H, Mayer J, Kremmer E, Roemer K and Mueller-Lantzsch N (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. *Oncogene*, **19**, 4328–4336.

Buzdin A, Khodosevich K, Mamedov I, Vinogradova T, Lebedev Y, Hunsmann G and Sverdlov E (2002) A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs. *Genomics*, **79**, 413–422.

Dangel AW, Baker BJ, Mendoza AR and Yu CY (1995) Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics*, **42**, 41–52.

Dunn CA, Medstrand P and Mager DL (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12841–12846.

Gifford R and Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, **26**, 291–315.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics*, **16**, 418–420.

Mayer J, Sauter M, Racz A, Scherer D, Mueller-Lantzsch N and Meese E (1999) An almost-intact human endogenous retrovirus K on human chromosome 7. *Nature Genetics*, **21**, 257–258.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Sauter M, Roemer K, Best B, Afting M, Schommer S, Seitz G, Hartmann M and Mueller-Lantzsch N (1996) Specificity of antibodies directed against Env protein of human endogenous retroviruses in patients with germ cell tumors. *Cancer Research*, **56**, 4362–4365.

Genome archaeology

Samuel Aparicio
University of Cambridge, Cambridge, UK

1. Introduction

The availability of substantially sequenced animal genomes in the last 5 years has served as the framework not only for studies of individual genes, but has allowed much greater insight into the evolution of content and structure. Also at a very practical level, the availability of completed genomes has transformed the speed at which both evolutionary and functional studies can be undertaken – large amounts of time and effort that historically were spent on simply cloning a desired gene to study its sequence or mutate it, for example, are now entirely bypassed. Conceptually, the ability to compare long contiguous sequences directly (as opposed to inferring content from indirect studies of nucleic acids), and to know the entire contents of a given genome, has permitted more precise conclusions. The human (Lander *et al.*, 2001; Venter *et al.*, 2001), mouse (Waterston *et al.*, 2002; *see also* Article 47, **The mouse genome sequence**, Volume 3), pufferfish (Aparicio *et al.*, 2002; *see also* Article 46, **The Fugu and Zebrafish genomes**, Volume 3), and recently chicken genomes (Hillier *et al.*, 2004), have added considerable depth to the annotation of protein-coding gene loci and noncoding regulatory and structural elements. The basis for all such comparisons is that unconstrained sequences are quickly randomized by mutations, and thus the presence of conserved sequence elements over large distances strongly implies conserved function. The value of comparing multiple vertebrates has thus become realized. Outlined here are some of the key questions that have been or are being addressed at a macrogenomic scale from the genome-sequencing programs of metazoans.

2. Evolution of content

There are two major questions: how many gene loci are there in a given organism and what mechanisms have accounted for the numbers of gene loci. The question of the number of gene loci in a given organism is one of the oldest questions, and this has yielded to whole-genome sequencing. Before sequencing of complex animal genomes was initiated, estimates ranged widely, especially for humans, where the predictions ranged from 25 000 gene loci to over 140 000. Comparative genomics of partially sequenced genomes provided (Aparicio, 2000; Ewing and

Green, 2000; Roest Crollius *et al*., 2000) a good guide to this almost a year before the draft sequencing of humans announced the ballpark result of 30 000–35 000 loci – much fewer than what most people had predicted. Even this turned out to be an apparent overestimate – as the annotation of genomes has improved, the number of gene loci has drifted downward, currently to around 25 000 for humans, although recent methods (Bertone *et al*., 2004) of detecting rare transcripts suggest that the number may yet be revised upward again – but only by a small percent. The figure of approximately 25 000 should be compared with invertebrate metazoans, for example, flies and worms in which the gene content was found to be between 12 000 and 18 000 genes approximately (*see* Article 44, **The *C. elegans* genome**, Volume 3 and Article 45, **The *Drosophila* genome(s)**, Volume 3). Estimates for teleost fishes from completed genomes (Aparicio *et al*., 2002; Jaillon *et al*., 2004) at the base of the vertebrate tree are similar to those of humans. Clearly, the leap from invertebrates to vertebrates did not result in expansion of more than about one-third of the total gene complement. However, whole-genome sequencing has made it clear that substantial evolution has taken place within gene families and, crucially, to the means by which gene expression is controlled (see following text), suggesting that the apparent complexity of mammals has as much to do with these mechanisms as with increases in the total number of gene loci.

Related to the number of gene loci has been considerable insight into the processes by which the number of genes has increased through evolution. Several mechanisms are thought to have influenced this process, principally, tandem gene duplication, retrotransposition of genes, and whole-genome (or chromosomal) duplication. The process of whole-genome duplication is thought to occur via the existence of a chromosomally polyploid ancestor, which (by mechanisms presently not understood), fixes its chromosome complement in a diploid state, thus doubling the gene complement of a given lineage. While genome duplication is therefore the most far-reaching mechanism in sheer scale, it has also been the hardest to detect, because gene duplicates tend to be rapidly eliminated during evolution, unless fixation occurs through selective processes. Sequencing of yeasts and the worm provided early evidence that these organisms were degenerate polyploids (organisms that have experienced a round of genome duplication, followed by extensive loss of duplicate genes). The proposal by Ohno, that the content of vertebrate genomes had been shaped through successive rounds of genome duplication, was therefore a subject of some controversy until substantially completed fish and mammal genomes became available. The availability of these genomes has shown first that vertebrates certainly underwent one round of genome duplication, or possibly two. Second, it has also become clear that teleost fishes, a vertebrate subfamily, also underwent an additional round of genome duplication close to the time of their origin (Amores *et al*., 1998; Christoffels *et al*., 2004; Van de Peer, 2004). Duplication events were identified in both of these cases, by searching for linked groups of orthologous genes in which the linkages were shared between distant vertebrate species. Molecular phylogenetic dating methods applied to these clusters established the approximate timing of these events. Although attempts had been made at such analyses before the availability of sequenced genomes, it had proven difficult to establish a sufficient number of linkage groups with orthologous genes.

Evolution of protein-coding families has also become apparent as the result of genome-wide comparisons, for example, the selective expansion of olfactory receptor GPCRs in rodents, the expansion of immune receptor families in mammals, and the (unexplained) excess of kinase proteins in teleosts. Comparisons between vertebrate genomes have revealed how genomic structural elements such as repetitive elements have evolved. For example, when humans and fugu were first compared, it became clear that while humans have a much greater proportion of repetitive DNA sequences, these are mostly of a restricted class, retrotranscribed repeats capable of endoduplication and reinsertion around the genome. Fugu, however, have a much greater diversity of repetitive DNA sequences, with almost every described class of metazoan repeat sequence represented – but collectively, these sequences comprise less than 15% of the genome, as opposed to 40% in humans. The evolutionary and functional significance of this observation remains unclear.

3. Evolution of function

As implied above, evolved complexity has arisen not only from the number and types of gene loci encoded, but from increasingly complex regulation of gene expression. Comparisons of genomes show that this happens at multiple levels, from regulation of alternative splicing, through the mechanisms of transcriptional modification and posttranslational modification. Detecting regulatory signals in the genome is much harder, in part because no equivalent of linear code exists for these and partly because the signal elements are, for the most part, made of small degenerate sequences combined together. Comparisons of noncoding sequences were proposed many years ago as one means of detecting these elements, and such comparisons are now revealing the locations of regulatory sequences (*see* Article 48, **Comparative sequencing of vertebrate genomes**, Volume 3). Genome-wide comparisons have revealed both conserved and novel regulatory functions. A recent finding has been the presence of very highly conserved noncoding sequence regions between distantly related vertebrates (Bejerano *et al*., 2004)

For example, in the last 5 years, antisense RNA regulation has gained prominence in vertebrates, as a mechanism that may regulate gene expression. The encoding of antisense RNAs and micro-RNAs as potential regulatory features has recently been shown to be conserved amongst vertebrates (Lim *et al*., 2003). Although the exploration of the functions of micro-RNAs is still in its infancy, the fact that they are conserved as elements in genomes separated by 450 Myr of evolution strongly suggests that they are required.

In contrast, other mechanisms appear to be vertebrate specific – while measures of splicing complexity are still controversial, it is clear that vertebrates have expanded tissue-specific core spliceosome elements (hnRNP and SR proteins) in comparison with invertebrates. Promoter CpG island methylation is known to be a vertebrate feature (as distinct from methylation of viral DNA sequences that occur in all species), and the complexity of this regulation increases from basal vertebrates to mammals (epigenetic imprinting is thought to be specific to placental mammals, for example). Not surprisingly, the content of gene families effecting some of these functions has coevolved.

4. Conclusion

We are at the cusp of a new discipline delving into the science of genomes. We can expect the diversity of sequenced animal genomes to increase substantially over the coming years. The availability of additional sequenced genomes will allow us to further elaborate our understanding of both genome content and function, and provide enormous opportunities for elaborating our understanding of the evolution of the animal kingdom.

References

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, *et al*. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science*, **282**(5394), 1711–1714.

Aparicio SA (2000) How to count ... human genes. *Nature Genetics*, **25**(2), 129–130.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**(5585), 1301–1310.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS and Haussler D (2004) Ultraconserved elements in the human genome. *Science*, **304**(5675), 1321–1325.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samantha M, Weissman S, *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**(5705), 2242–2246.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S and Ventatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Molecular Biology and Evolution*, **21**(6), 1146–1151.

Ewing B and Green P (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, **25**(2), 232–234.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, *et al*. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**(7018), 695–716.

Jaillon O, Aury JM, Brunet F, Petit JL, Strange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**(7011), 946–957.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lim LP, Glasner ME, Yekta S, Burge CB and Bartel DP (2003) Vertebrate microRNA genes. *Science*, **299**(5612), 1540.

Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier, *et al*. (2000) Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics*, **25**(2), 235–238.

Van de Peer Y (2004) Tetraodon genome confirms Takifugu findings: most fish are ancient polyploids. *Genome Biology*, **5**(12), 250.

Venter JC, Adams MD, Myers EW, Li P, Mural RJ, Sutton CG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Waterston RH, Lindblad-Toh K, Birney E, Rodgers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–562.

Introductory Review

Functional analysis of genes

Rick Woychik and Carol Bult
The Jackson Laboratory, Bar Harbor, ME, USA

1. Introduction

The Human Genome Project started a revolution in biology. The initial effort to sequence the genomes of humans and many model organisms is now essentially complete (*see* Article 43, **Functional genomics in *Saccharomyces cerevisiae***, Volume 3, Article 44, **The *C. elegans* genome**, Volume 3, Article 45, **The *Drosophila* genome(s)**, Volume 3, Article 46, **The Fugu and Zebrafish genomes**, Volume 3, and Article 47, **The mouse genome sequence**, Volume 3). Researchers can now use numerous Internet resources and databases to access the annotated genomes of human, the laboratory mouse, yeast, Drosophila, and many other organisms. These electronic catalogs of genes with their associated nucleotide and protein sequences serve as a powerful launching off point for experimental and computational approaches to understanding the biological role and function of genes and gene products.

One of the most powerful ways of studying the function of a gene is to alter its expression within an organism. Indeed, analyzing changes in the expression of a gene to study its biological function in vivo has become a mainstay in biological research. If the expression of a gene is changed in a specific way (e.g., knocked out) and the organism develops a specific phenotype (e.g., a disease state), then a connection can be made between that gene and function in the organism.

There are many model organisms that can serve as platforms for determining gene function (*see* Article 38, **Mouse models**, Volume 3, Article 39, **The rat as a model physiological system**, Volume 3, Article 41, **Mouse mutagenesis and gene function**, Volume 3, and Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3), although only some of these are also useful for studying gene function with genetics approaches. Here we will cover a variety of model organisms including yeast (*Saccharomyces*), fly (*Drosophila*), worm (*Caenorhabditis*), zebra fish (*Danio*), mice (*Mus*), and rat (*Rattus*). That functional homologs for most human genes can be found within these species allows researchers to translate the understanding of gene function between organisms and ultimately to humans.

Experimental genetics is commonly used to study gene function because spontaneous and induced mutations often cause changes in the expression of a gene. Mutations in or near a gene can alter the level of transcription and/or translation,

and sequence variations within the coding sequence of a gene can change the nature of the protein product. In humans, this approach to understanding gene function is limited to new spontaneous mutations. In model organisms, however, the genome can be manipulated through experimental mutagenesis, and the differential expression of genes can be used to dissect the function of genes.

2. Experimental mutagenesis

Experimental genetic approaches involve the intentional generation of new germline mutations and then screening for the phenotypic effects of these mutations in the organism. This approach represents one of the biggest advantages of using model organisms over studying humans directly. Mutations are typically created through treatment with chemicals or radiation, or through insertional or targeted mutagenesis. It is also possible to generate alterations in the expression of genes through transgenesis and a new technology called RNAi. Here we will briefly summarize these approaches.

2.1. Radiation- and chemical-induced mutations

The ability to create new mutations with radiation and chemical agents has been one of the most powerful tools for studying gene expression in most model organisms. Radiation typically produces large chromosomal rearrangements such as translocations, inversions, and deletions that remove large numbers of genes. Chemical mutagens, on the other hand, are most often chosen for their ability to produce subtle mutations within genes. Various chemical agents are available for mutagenesis, and one that is becoming increasingly popular is ethylnitrosourea (ENU) (*see* Article 38, **Mouse models**, Volume 3 and Article 41, **Mouse mutagenesis and gene function**, Volume 3). ENU usually produces single base-pair substitutions within the DNA. The advantage of a point mutagen is that a variety of different types of mutations can be generated, from amorphic alleles that completely inactivate the gene, to hypomorphic alleles with decreased levels of expression of that gene, as well as other types of mutations. To understand the full phenotypic potential of mutations on gene function, researchers often try to generate a complete allelic series for a given gene. While most methods for inducing mutations involve treatment of the whole animal, chemical mutagenesis of embryonic stem (ES) cells in the laboratory mouse has also proven to be an effective means of producing mutant mice (*see* Article 41, **Mouse mutagenesis and gene function**, Volume 3).

2.2. Insertional mutagenesis

Most organisms have species-specific mobile genetic elements that are capable of "hopping" around the genome. Many mobile elements are retroviral-like elements with long terminal repeats (LTRs) at both ends that contain strong promoter and

regulatory elements. Integration of mobile elements is known to occur throughout the genome. When the integration reaction occurs within a gene, it can change the expression of the gene, most often inactivating or substantially downregulating the expression of the gene. This form of mutagenesis is called insertional mutagenesis. When the element inserts upstream of a gene, the promoters within the element can ectopically activate the expression of the downstream gene. This is referred to as promoter insertion mutagenesis.

Experimental protocols have been developed for insertional mutagenesis with mobile genetic elements in all model organisms. One of the most widely used approaches involves the use of P-elements in *Drosophila* in a process called hybrid dysgenesis. In this case, crossing two different lines of *Drosophila* (*see* Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3) mobilizes endogenous P-elements and efficiently creates new insertional mutations throughout the genome. Also, many spontaneous mutations in various organisms arise by insertional mutagenesis through the process of mobile genetic element hopping. The advantage of mutations created by insertional mutagenesis is that the mutant locus is "tagged" with the mobile genetic element. The fact that they are marked in this way makes it much easier to characterize the mutation at the molecular level.

2.3. Targeted mutations

Another form of experimental mutagenesis, called targeted mutagenesis, involves the creation of a new mutation within a specific gene. This approach can be used in yeast and mice (*see* Article 41, **Mouse mutagenesis and gene function**, Volume 3 and Article 43, **Functional genomics in *Saccharomyces cerevisiae***, Volume 3), but is not possible in most other organisms. With this form of mutagenesis, the germline can be changed in highly specific ways. It is possible to generate targeted mutations that range from single point mutations within a gene all the way to sizable deletions along the chromosome. It is also possible to "knock in" the coding regions of one gene downstream of regulatory elements of another gene. One of the newest advancements in the technology makes it possible to create "conditional" targeted mutations, where, for example, the gene is knocked out only in certain cell types or at certain times during development.

2.4. Transgenesis

One of the most powerful tools that has been developed in recent years involves the introduction of cloned fragments of DNA back into the genome of an organism. When the DNA fragment integrates randomly into the genome and becomes a stable heritable trait, the resulting organism is referred to as a "transgenic". This approach for studying the function of genes can be used in all the model organisms listed above, although the term "transgenic" is most often used in the context of studies with zebra fish, mice, and rats (*see* Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3). In mice and rats, transgenics are most often produced by microinjection of DNA directly into an early stage embryo,

although infection with specially designed viral vectors is also used. The transgenic technology is used primarily for the purpose of testing the dominant effects of introducing a gene or modified gene construct into an organism. It is also used for a variety of other applications, including the introduction of large fragments of cloned genomic DNA back into the germline of the organism.

2.5. RNA interference

In recent years, a new phenomenon has been discovered that can be used to alter gene expression in a targeted fashion. The method is called posttranscriptional gene silencing (PTGS). The active molecule in PTGS is a double-stranded RNA (dsRNA) with sequence homology to a specific target gene within the cell. The target gene is silenced by selective destruction of its mRNA. This process has come to be known as RNA interference, or RNAi. Experimental approaches for using RNAi to regulate the expression of genes have been developed for all model organisms, with the most remarkable being in *Caenorhabditis elegans*. In this organism, a genome-wide library of *Escherichia coli* clones have been developed, where each clone expresses a dsRNA for an individual gene in *C. elegans*. Feeding *E. coli* expressing dsRNA for a specific gene causes an inhibition of that gene in an otherwise wild-type organism.

Most often, the RNAi technology produces a "knockdown" of the expression of a given gene and does not completely inactivate expression. Therefore, it cannot be routinely used to study the null phenotype of a "knockout" of that gene. Nevertheless, the approach of using RNAi is proving to be tremendously useful as a tool for rapidly downregulating specific genes to study their function in various organisms. This has been particularly the case in *C. elegans*, where the knockdown of any gene can be quickly tested by feeding the appropriate clone of *E. coli* to the worms.

3. Naturally occurring sequence variations to study gene function

Unlike experimental mutagenesis that involves the generation of new mutations within genes, it is also possible to study the function of genes by exploiting the naturally occurring sequence variations that arise from spontaneous mutations in the germline of organisms (*see* Article 40, **Farm animals**, Volume 3). These spontaneous mutations create the genetic diversity that causes the different biological and disease-specific traits within individuals in a population.

The genes impacted by spontaneous mutations can be identified through a process called positional cloning. Positional cloning involves scanning the DNA of individuals in a test population with molecular markers that are capable of differentiating the parental alleles at hundreds of loci throughout the genome. Analysis of the data will reveal the region of the genome, called the critical region, that cosegregates with the mutant phenotype. Genes within the critical region are evaluated, and eventually the single gene that contains a sequence variant that

causes the mutant phenotype is identified. This process of positional cloning has become commonplace in the genetics community in recent years and was greatly facilitated by having access to the genome sequence and knowledge of the position of essentially all genes. It is the only way to conduct molecular genetics experiments in humans.

Compared to experiments in humans, however, studying spontaneous mutations is enhanced in most model organisms by the availability of genetically pure inbred lines. Inbred lines are those that are produced by sequential brother–sister mating such that all of the offspring within that line are essentially genetically identical. Examples of inbred lines are C57BL/6 J in mice, Sprague–Dawley in rats, and C32 and SJD in zebrafish. Other types of special genetic lines in several model organisms are based on shuffling genome segments from inbred lines. Recombinant inbred, or RI, lines are stable lines that are generated by crossing different inbred lines followed by brother–sister mating for many generations to stabilize the genome. Each RI line has a variable-sized chromosomal segment throughout the genome from each of the parental lines. Congenic lines are those that have a single chromosomal region of varying length from one inbred line, with the rest of the genome being derived from a second inbred line. Consomic lines are those that harbor one full chromosome from one inbred line with all of the other chromosomes being from a second inbred background.

Each inbred line has a different assortment of alleles that are fixed in the genome. Not every gene is different between each of these lines, but the various inbred strains each contain allelic differences within genes that often result in dramatically different phenotypes. Unlike simple Mendelian traits that are associated with a mutation in a single gene, most of the phenotypic differences between inbred lines arise from the complex interactions of allelic variations within multiple genes. These are referred to as complex traits. The individual loci that contribute to a complex trait are called complex trait loci, or QTLs (*see* Article 38, **Mouse models**, Volume 3). Examples of complex traits are blood pressure, atherosclerosis, osteoporosis, diabetes, and obesity.

There are considerable efforts under way to discover the allelic variants with the genes that underlie complex traits. These most often involve controlled breeding experiments between different inbred lines that exhibit extremes for a given phenotype, such as high and low blood pressure, although the direct analysis of RI, consomic, and congenic lines is also being used to study QTLs and complex traits. While it still remains a challenge to identify the allelic variants within specific genes that contribute to complex traits, the genome resources that are now available from the genome project are now making it possible to begin to characterize QTLs at the molecular level (*see* Article 38, **Mouse models**, Volume 3).

4. Genetics and model organisms

Each of the six model organisms described here has its own advantages and disadvantages. The mouse is one of the most versatile of model organisms for genetics experiments (*see* Article 38, **Mouse models**, Volume 3). In addition to the conservation of genome content and organization with that of the human genome, the mouse

has an anatomy and physiology that, in many cases, resembles that in humans. It has a rich history of genetic analysis, and is the focus of considerable current work in experimental genetics. There are many resources that maintain spontaneous, insertional, transgenic, chemical- and radiation-induced mutants, and there are inbred, congenic, consomic, and recombinant inbred lines available that take advantage of normal sequence variations in mice. Most notably, other than yeast, the laboratory mouse is the only organism where it is possible to efficiently alter gene expression by generating targeted mutations by homologous recombination.

There are some experiments that require a larger rodent than the mouse. In these cases, the rat is often used (*see* Article 39, **The rat as a model physiological system**, Volume 3). Like the mouse, there are spontaneous mutants, inbred lines, and congenic lines available, and it is possible to create transgenic lines of rats. There are protocols available for conducting mutagenesis experiments with chemicals such as ENU. However, the size of the rat resource does not match that of the mouse, and it is not possible to produce targeted mutations, by homologous recombination.

There are also certain types of experiments that are uniquely well suited for other model organisms. The relatively simple genome, rapid generation time, ability to generate and maintain large numbers of offspring, accessibility to the developing embryo, efficient strategies for chemical and mobile genetic element experimental mutagenesis, and extensive genetics resources make *Drosophila* the organism of choice for many experiments for studying gene function (*see* Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3). The fact that the early embryo is transparent makes *Danio* a perfect model for studying the earliest stages of embryonic development, and *Caenorhabditis* has numerous advantages for experiments in developmental biology because there are less than 1000 cells in the entire adult organism, and the fate of each cell has been mapped (*see* Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3). Finally, for other experiments that involve cellular processes, *Saccharomyces* may prove to be the organism of choice (*see* Article 43, **Functional genomics in *Saccharomyces cerevisiae***, Volume 3).

5. Databases as research tools to study gene function

The nature of biological research has changed with the genome project. New tools to collect large complex data sets are now available, and teamwork involving multiple groups of investigators at different institutions is becoming the norm. Databases now serve as centralized electronic repositories for data collected around the world. The components are now in place to support a new paradigm for conducting biological research that embraces the complexity associated with most biological systems. This new approach for biological research is referred to as "integrative biology", where data sets from multiple organisms are collected and mined computationally to reveal patterns in the data that would not have emerged if each data set was analyzed in isolation from the others. Ultimately, access to these new informatics tools should help the biological community understand how whole sets of genes function in complex biological systems.

One key component to integrative biology is to develop databases that serve as data integration hubs for each of the various model organisms. Most of the model organisms that are commonly used in biomedical research have a model organism-specific community database to support data integration and manual curation. The Mouse Genome Informatics (MGI; **http://www.informatics.jax.org**), for example, is a widely the recognized community database for the laboratory mouse. MGI integrates genetic and genomic data for the mouse in order to service its mission to facilitate the use of the mouse as a model system for understanding human biology and disease processes. The MGI database provides extensive information about mouse genes, sequences, alleles, mutant phenotypes, QTLs, strain polymorphisms, gene and protein function, developmental gene expression patterns, and curated mammalian homology (especially for rat and human) data. Data pertaining to all types of mouse lines described in this report are available to the research community (e.g., new ENU mutants, transgenic, targeted mutations, etc.) and are accessible from the MGI database. MGI is also a platform for computational assessment of integrated biological data with the goal of identifying candidate genes associated with complex phenotypes. Similar integrated databases for yeast (**http://www.yeastgenome.org/**), the fly (**http://flybase.bio.indiana.edu/**), the worm (**http://www.wormbase.org/**), zebra fish (**http://zfin.org**), and the rat (**http://rgd.mcw.edu/**) are also available. Multiple integrated electronic resources are available for accessing genetic and genomic data for humans (**http://www.ncbi.nlm.nih.gov/genome/guide/human/**; **http://www.ensembl.org**; **http://genome.ucsc.edu**).

In addition to building information systems that focus on the integration of diverse data for a single organism, it is also critical to the advancement of biomedical research to be able to compare functional genomic data across organisms. Developing standards that support this comparative genomics approach to functional genomics is currently a major focus of activities in the bioinformatics community and one such effort is being led by the Gene Ontology (GO) Consortium (**http://www.geneontology.org**). This GO Consortium is developing controlled vocabularies of terms related to the molecular function, biological process, and cellular location of gene products. The GO group both develops the terms and their definitions and then annotates genes and gene products using the terms. Consistent use of biological terms across different organisms means that researchers can retrieve data for a single organism or for multiple organisms accurately and consistently. Being able to compare data across species makes it possible to use and play to the strengths of all the different model organisms to study the function of genes.

6. Summary

Studying the function of genes is a multidimensional process that involves many organisms and a large variety of experimental approaches. No single organism is best suited for all experiments, and no single experimental strategy is appropriate for all genes. In the end, integrated electronic resources with data collected from all genes in yeast, flies, worms, mice, zebra fish, rats, and human will best serve the needs of the biomedical research community to study gene function.

Specialist Review

Mouse models

Michelle E. Goldsworthy and Roger D. Cox
Medical Research Council Harwell, Harwell, UK

1. Introduction

1.1. Why use the mouse as a model system?

The mouse has very well-defined genetics, with genetically homogenous inbred strains and a completed genome sequence. In addition, mice have a short breeding time that allows easy generation of sufficient numbers of individuals for an experiment, defined housing conditions including diet and health status, well-developed and still evolving phenotyping technologies, and a broad range of genetic manipulation tools. Figure 1 illustrates these advantages as compared to heterogeneous human populations.

1.2. What can the mouse really model?

We would like the mouse to model human disease but we obviously know that there are differences between these two species, which means that we must be realistic in our expectations. Among the many similarities are a similar sized genome and number of genes and broadly similar physiology. There are obvious differences in some aspects of basic biochemistry (mice have high HDL-cholesterol, for example), life span, and size. A mouse with mutations in the same genes that are mutated in human may not always fully recapitulate the human disease but it is important to remember that the differences may be informative and the similarities sufficient to allow investigation of some of the mechanisms involved that might otherwise be impossible. Thus, the mouse may serve as a model at different levels from fully representing a human disease and its genetics at one extreme to yielding insights into particular molecules or molecular interactions that form part of a disease process, or just as relevantly normal biology, conserved in humans at the other extreme. At this latter extreme, other lower organisms (for example, *Caenorhabditis elegans*) may also serve. Figure 1 illustrates some of these concepts. Through the mapping and cloning of genes underlying a phenotype in mouse, such as for example obesity, a whole new field may be born that gives fundamental insight into basic mechanisms, such as the control of food intake and metabolism. While these discoveries may not have identified the precise genetic factors contributing

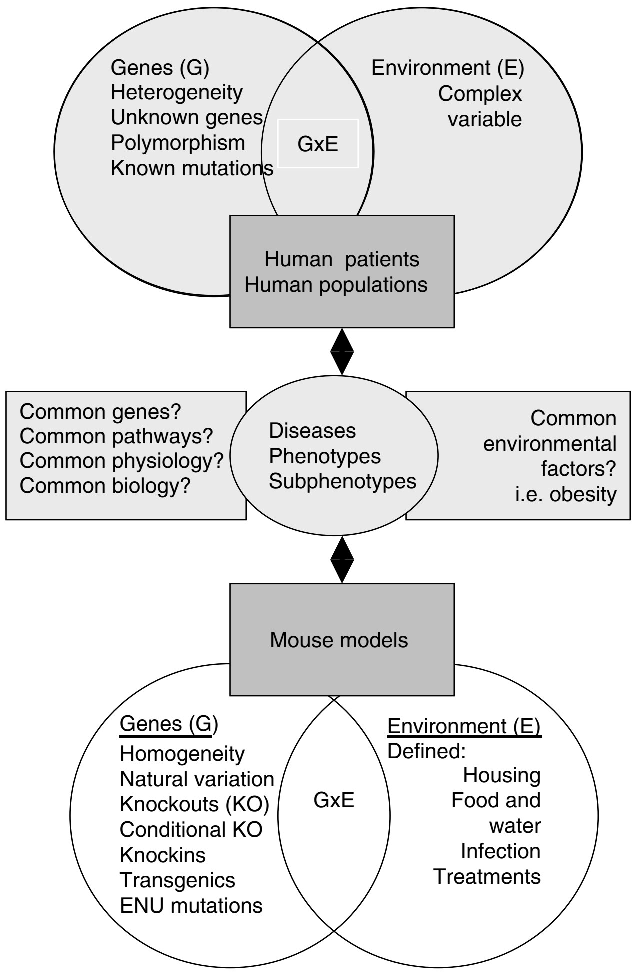


Figure 1 The expression of a phenotype or disease depends on the interaction of genes and environment. In the mouse, these factors can be well defined but their use as models for human disease analysis depends on there being at least some similarity in the genes, pathways, physiology, and biology of the process being studied – which is often true

to common human obesity disorders, they have at least described some of the fundamental pathways that are common to both species. Finally, in pathways that may be associated with human disease, where we know some of the genes that are involved in the mouse, it has been possible to investigate the basic processes that also operate in human but can only easily be modeled in another species. Using the example of obesity and type 2 diabetes, we will further elaborate on some of these points.

In type 2 diabetes, one finds both impairment of insulin secretion and insulin resistance (for a recent review, see Ashcroft and Rorsman, 2004). Insulin resistance is characterized by the impaired ability of insulin to inhibit hepatic glucose production and to stimulate glucose uptake by skeletal muscle. Insulin also fails to

suppress lipolysis in adipose tissue, which in turn increases circulating nonesterified fatty acid (NEFA) concentrations that stimulate gluconeogenesis, triglyceride synthesis, and glucose production in the liver. Insulin resistance generally increases in parallel with increasing body-fat mass with increased levels of obesity being largely responsible for the rise in prevalence of type 2 diabetes in the developed world. Although no mouse model of type 2 diabetes to date faithfully mimics all aspects of type 2 diabetes and obesity in humans, single gene spontaneous and induced mutations, transgenic mice and genetic differences between inbred mouse strains (QTLs) have all added to the understanding of the underlying metabolic pathways and disease processes some of which will be discussed in this review.

2. Spontaneous single-gene mutations

Obesity is a complex trait influenced by genetic factors in addition to age, gender, diet, and exercise. Body weight and appetite are mainly regulated by neuropeptides released from the hypothalamus in response to signals from fat, muscle, the liver, and gastrointestinal tract, and physiologically obesity can be the result of changes in either appetite, nutrient turnover, or thermogenesis. Disruption in energy balance homeostasis can also cause hyperleptinemia and hyperinsulinemia with obesity itself perhaps the greatest risk factor for the development of type 2 diabetes, coronary heart disease, and other metabolic disorders.

A number of spontaneous single-gene mutations result in obesity in the mouse and the mapping and cloning of these kick started this field, resulting in fundamental and novel discoveries. These include *obese* (ob) and *diabetic* (db) mouse models. Obesity in these models results from mutations in the hormone leptin and its receptor respectively. Leptin is a soluble hormone secreted from adipocytes that signal through the pro-opiomelanocortin (POMC) and neuropeptide Y (NPY) neurones in the hypothalamic arcuate nucleus (ARC), which normally controls feeding. POMC neurones also respond to a variety of other factors including dopamine, gonodal steroids, glucose, and insulin (Baskin *et al*., 1999), which would imply a broader role for these neurons in communicating peripheral information regarding energy homeostasis to centers controlling feeding and metabolism. Lack of biologically active leptin or leptin receptor effectively prevents normal signaling of energy stores by the adipose tissue to the hypothalamus, resulting in hyperphagia, decreased energy expenditure, and the accumulation of excess fat. The replacement of exogenous leptin in ob/ob and wild-type mice results in weight loss, decreased food intake, and increased physical activity, whereas as expected db/db mice are unaffected. Both ob/ob and db/db mouse strains exhibit in addition to obesity hyperinsulinemia, insulin resistance, and ultimately hyperglycemia becoming severely diabetic. Additionally, the interruption of leptin or α-melanocyte stimulating hormone (α-MSH) signaling in the hypothalamus is thought to be the primary cause of obesity in agouti (A^y) and Mc4r null mice.

The *tubby* stain of obese mice arose as a spontaneous mutation in the Jackson mouse colony (Coleman, 1990) and phenotypically is characterized by maturity-onset obesity accompanied by retinal and cochlear degeneration (Ohlemiller, 1997). The development of obesity in tubby mice is relatively mild and resembles weight

gain in humans more closely than that observed in *obese* (ob) and *diabetes* (db) mouse strains and resembles the phenotype observed in Ushers, Bardet-Biedl, and Alstrom syndromes in humans. Positional cloning of the *tub* gene identified it is a member of a novel gene family and appears to be specifically expressed within the nervous system, in the key hypothalamic nuclei that have been implicated in the central control of energy homeostasis (Kapeller, 1999). $Tub^{tmlRok-}$ knockout (KO) mice are phenotypically indistinguishable from tubby mice, indicating that the tubby phenotype is due to loss of function (Stubdal, 2000). *Tub* has been shown to be tyrosine phosphorylated in response to insulin and insulin growth factor-1. In vitro *tub* is phosphorylated by purified insulin receptor kinase, Abl, and JAK 2 but not by epidermal growth factor receptor and Src kinases, data that suggests that *tub* may function as an adaptor protein linking the insulin receptor and possibly other protein tyrosine kinases to SH2-containing proteins (Kapeller, 1999).

Agouti (A^y), a classic coat color mutant encoding the agouti signaling peptide is an antagonist of melanocortin 4-receptor (Mc4r) signaling; obesity in these mice is the result of aberrant antagonism of the hypothalamic Mc4r. The targeted disruption of other melanocortin receptors expressed in the CNS, for example $Mc4r^{tmlDhu}$ knockout mice or their peptides have confirmed this pathway as essential for the control of energy homeostasis (Huszar, 1997). $Mc4r^{tmlDhu}$ mice fed a moderately high fat diet were utilized in investigating the relative importance of Mc4r compared to leptin in control of food intake, food restriction and response to increased fat, energy expenditure, and activity (Butler, 2001). In order to investigate adaption to food restriction, $Mc4r^{tmlDhu-}$, $Mc3r^{tmlCone}$, $lep^{ob}/\ lep^{ob}$ and wild-type mice were fed a restricted diet for 96 h followed by reintroduction of ad libitum feeding. Absolute weight loss was similar across all groups and the kinetics of return to prediet weight was similar; this would imply therefore that neither Mc4r or Mc3r seem necessary for the sensing of nutritional defects. Feeding in response to diet was also studied, all four groups were swapped from a low to moderate fat diet, food intake and weight were increased in $Mc4r^{tmlDhu}$ mice but not in $Mc3r^{tmlCone}$ or in wild-type controls. No increase in weight or consumption was observed in $lep^{ob}/\ lep^{ob}$ mice, suggesting that Mc4r but not Mc3r might have a leptin-independent effect on food intake. Additionally, feed efficiency, that is weight gain per kilocalorie ingested was markedly increased in $Mc4r^{tmlDhu}$ mice but not in $lep^{ob}/\ lep^{ob}$ or in wild-type controls, which would suggest a specific interaction between Mc4r and fat content in the diet. Energy expenditure and increase in activity in response to moderate fat diet were also studied. Energy expenditure was measured using indirect calorimetry; mice were fed a low fat diet for 3 days followed by moderate fat diet for 3 days. Basal and total O_2 consumption during low fat chow consumption was low in ob/ob compared with wild-type and $Mc4r^{tmlDhu}$ mice. Following the switch to moderate fat chow, wild-type and ob/ob mice responded by increasing basal and total oxygen consumption with no response observed in $Mc4r^{tmlDhu}$ null mice. Activity measured by wheel running was also increased following moderate fat chow in wild type compared to $Mc4r^{tmlDhu}$ null mice. Therefore, Mc4r seems to be needed for regulation of activity-based energy expenditure in response to diet as well as being an important factor in the stimulation of diet induced thermogenesis. This study shows some of the advantages of using the mouse specifically to manipulate molecules of interest

and control environmental factors as well as being able to apply sophisticated phenotyping techniques to sufficient numbers of individuals to generate statistically valid results.

The obesity story illustrates the power of the mouse to yield insights into the fundamental mechanisms of control of energy homeostasis. The mouse as a model system is key to the understanding of this field. These various obesity mutants are not per se models for human obesity. For example, relatively few patients have mutations in leptin or its receptor, but they are models for investigating these highly conserved pathways – research that may lead ultimately to new therapeutic interventions (Flier, 2004).

3. Tissue-specific knockouts

In processes where genes are known to be involved, the mouse provides an extensive toolkit to investigate their action in more detail?. For example, the relative contribution of components of insulin signaling pathways, glucose uptake and insulin resistance have been studied via the generation of global or tissue-specific knockouts. A global homozygous knockout of the insulin receptor in mouse leads to fatal ketoacidosis in neonates; however, heterozygotes exhibit mild hyperglycemia with a corresponding hyperinsulinemia with some 10% of adult mice eventually developing diabetes dependent on background strain (Accili *et al*., 1996). A number of tissue-specific insulin receptor knockouts were generated using the Cre/loxP system and they have elucidated some of the complex interactions between various tissues during the development of impaired glucose tolerance and type 2 diabetes (summarized in Table 1). Surprisingly, muscle specific insulin receptor knockout (MIRKO) mice appeared to have normal glucose tolerance although insulin-stimulated muscle glucose uptake and glycogen synthesis were severely impaired accompanied by an increase in insulin-stimulated glucose transport in fat (Bruning, 1998; Kim *et al*., 2000). In addition, MIRKO mice have elevated fat deposits, serum triglycerides, and free fatty acid levels; thus, insulin resistance in muscle leads to dislipidemia and obesity but not to diabetes. Liver-specific insulin receptor knockout (LIRKO) in contrast exhibited severe insulin resistance, glucose intolerance and a failure of insulin to suppress hepatic glucose production (Michael *et al*., 2000). Mice additionally had marked hyperinsulinemia and an almost sixfold increase in β-cell mass. β-cell insulin receptor knockout βIRKO) mice lack glucose induced first phase insulin release and develop glucose intolerance with age (Kulkarni *et al*., 1999).

Global knockouts of signaling molecules downstream of the insulin receptor such as Irs1, Irs2, and Irs3 have also been studied. Irs1 is a principal substrate for insulin and insulin-like growth factor (IgF1) receptors. After tyrosine phosphorylation at several sites, Irs1 binds to and activates phosphatidylinositol-3-kinase (PI3K). Mice homozygous for Irs1 knockouts as expected showed no Irs1 phosphorylation or PI3K activity and a 50% reduction in intrauterine growth, impaired glucose tolerance, and a decrease in insulin-stimulated glucose uptake (Araki *et al*., 1994). Irs2 appears to act as an alternative substrate of the insulin receptor, and in Irs1 knockout mice, Irs2 substitutes for Irs1 leading to a marked increase in β-cell

Table 1 A summary of the many tissue and gene specific mouse knockout models of type 2 diabetes

Gene-/tissue-specific knockouts	Description of phenotype	References
$IR^{-/-}$	Diabetic ketoacidosis, early postnatal death	Joshi *et al*., (1996); Accili *et al*. (1996)
MIRKO	Elevated fat mass, serum triglycerides, and free fatty acids. Normal glucose, glucose and insulin tolerance	Bruning *et al*. (1998); Kim *et al*. (2000)
LIRKO	Severe glucose intolerance, failure to suppress glucose production, insulin resistance, and hyperinsulinemia	Michael *et al*. (2000)
βIRKO	Severe loss of insulin secretion in response to insulin, progressive impairment of glucose tolerance	Kulkarni *et al*. (1999)
NIRKO	Normal brain development and survival, diet sensitive obesity, and increased insulin and triglyceride levels	Bruning *et al*. (2000)
IRS-$1^{-/-}$	Growth retardation, mild insulin resistance, β-cell hyperplasia, hyperinsulinemia	Araki *et al*. (1994)
IRS-$2^{-/-}$	Severe insulin resistance, reduced β-cell mass, early diabetes	Withers *et al*. (1998); Kubota *et al*. (2000)
$IR^{+/-}$ IRS-$1^{+/-}$	Severe insulin resistance in muscle and liver, β-cell hyperplasia	Kido *et al*. (2000)
$IR^{+/-}$ IRS-$2^{+/-}$	Severe insulin resistance in liver but mild insulin resistance in muscle, moderate β-cell hyperplasia	Kido *et al*. (2000)
$IR^{+/-}$ IRS-$1^{+/-}$ IRS-$2^{+/-}$	Severe insulin resistance in liver and muscle, marked β-cell hyperplasia	Kido *et al*. (2000)
$GLUT4^{-/-}$	Moderate insulin resistance, glucose intolerance, and decreased adipocyte mass	Katz *et al*. (1995)
$GLUT4^{-/+}$	Insulin resistance, diabetes, hypertension	Stenbit *et al*. (1997)
MG4KO	Insulin resistance, glucose intolerance	Zisman *et al*. (2000)
FG4KO	Insulin resistance in muscle and liver, glucose intolerance, hyperinsulinemia	Abel *et al*. (2001)
$GLUT2^{-/-}$	Glucose intolerance, impaired insulin secretion, diabetes, and early death	Guillam *et al*. (1997)
$GK^{-/-}$ or β-cell $GK^{-/-}$	Diabetic ketoacidosis, early postnatal death	Grupe *et al*. (1995); Terauchi *et al*. (1995)
$GK^{-/+}$ or β-cell $GK^{+/-}$	Impaired insulin secretion, mild diabetes	Grupe *et al*. (1995); Terauchi *et al*. (1995)
Liver $GK^{-/-}$	Mild hyperglycemia	Postic *et al*. (1999)

mass resulting in increased insulin secretion. In contrast, knockout of Irs2 leads to hyperglycemia and insulin resistance; this is accompanied by a twofold reduction in β-cell mass with insulin secretion in response to insulin that decreased as hyperglycemia became more severe (Withers *et al*., 1998; Kubota *et al*., 2000) and the animals developed diabetes by 10 weeks of age. Differences in severity of disease in different knockout models have been attributed to modifiers in the

background strain (Terauchi *et al*., 2003). Irs3 (an isoform restricted to adipose tissue) null mice (Liu, 1999) in contrast showed normal body weight and normal blood glucose and insulin levels.

One physiological effect of insulin is stimulation of glucose uptake, with increased glucose transport in skeletal muscle and adipose tissue required for normalizing postprandial hyperglycemia. One result of insulin signaling leading to glucose uptake is translocation of GLUT-4 (the glucose transporter expressed in skeletal muscle, heart, and adipocytes) to the plasma membrane and uptake of circulating glucose (for a review of the insulin signaling pathway, see Bevan, 2001). GLUT-4 knockout mice, however, do not develop diabetes, exhibiting only moderate insulin resistance and impaired glucose tolerance (Katz *et al*., 1995; Stenbit *et al*., 1997). The loss of GLUT-4 however did result in growth retardation, reduced fat tissue, cardiac hypertrophy, and a shortened lifespan. Like the insulin receptor, tissue-specific knockouts of GLUT-4 in muscle and fat have been created. Muscle-specific knockouts (MG4KO) mice developed severe insulin resistance and glucose intolerance. Unexpectedly, fat-specific knockouts (FG4KO) showed, as well as impaired insulin-stimulated glucose uptake in adipocytes, insulin resistance in both muscle and liver (Abel *et al*., 2001). The mechanism by which this apparent fat cell defect alters insulin action in muscle and liver remains unclear but it is consistent with the view of the adipocyte as an endocrine cell (see Flier, 2004 for a recent review). Knockout studies indicate the glucose transporter 2 (GLUT2 Slc2a2) along with glucokinase is involved in the highly regulated process of insulin secretion by the β-cell. GLUT-2 (Slc2a2) deficient mice developed hyperglycemia and hypoinsulinemia, leading to severe diabetes and death within the first 3 weeks following birth (Guillam *et al*., 1997). $Slc2a2^{tmlThor}$ mice exhibited severely impaired glucose tolerance and isolated islet studies showed impaired glucose stimulated insulin secretion. $Slc2a2^{tmlThor}$ mice could be rescued by transgenic reconstitution of Slc2a2 expression in the β-cells alone (Thorens, 2000).

These examples illustrate the potential of the mouse to model specific deficiencies of known molecules, which when combined with tissue specific elimination, reveals some unexpected tissue interactions. The obvious limitation of null mutations is that they may mask some important interactions if the gene involved is essential for some earlier step such as embryogenesis.

4. Multigenic models

A panel of mutations affecting different aspects of a pathway or interacting system can relatively easily be combined by breeding in the mouse to generate a multigenic model with the purpose of investigating gene and pathway interactions and potentially generating more representative models of human disease. There are, for example, mutants that reduce HDL-cholesterol in the mouse that could be combined with other risk factor genes for cardiovascular disease.

In the diabetes field, knockout strains of mice have been combined in order to produce polygenic models of diabetes. Mutant mice heterozygous null for both the insulin receptor (Insr) and insulin receptor substrate 1 (Irs1) genes developed severe insulin resistance in muscle and liver with 40% of mice becoming overtly diabetic

by 6 months of age in contrast to each mutation alone (Bruning, 1997). In this model, insulin resistance was accompanied by compensatory β-cell hyperplasia and hyperinsulinemia in an effort to overcome insulin resistance. Doubly heterozygous Irs1 and Irs2 knockout mice developed insulin resistance in the liver with less pronounced β-cell hyperplasia and diabetes (Kido *et al*., 2000), while triple heterozygous mice for Insr, Irs1, and Irs2 developed severe early onset diabetes, characterized by insulin resistance in both muscle and liver accompanied by a compensatory increase in β-cell mass. Double knockouts of Irs1$^{-/-}$ and Irs3$^{-/-}$ unlike Irs1$^{-/-}$ alone were characterized by marked hyperglycemia (Laustsen, 2002) and thus would appear to compensate for each other's functions, no such compensation, however, was observed in Irs2$^{-/-}$ Irs$^{-/-}$ mice (Terauchi *et al*., 2003). In order to further characterize the interaction between Irs2 and β-cell function, Withers (1999) crossed insulin-like growth factor-1 receptor (Igf1r) and Irs2 knockout mice. Igf1r null mice die soon after birth, have elevated blood glucose levels and histological analysis showed failure of α- and β-cell of the pancreas to organize into mature islets, suggesting that like Irs2, Igf1r had a role in the development of normal function islets. Similar to Irs2, null animals Igf1r$^{+/-}$ Irs$^{-/-}$ animals showed glucose intolerance, polyuria, and weight loss. The reduction in β-cell area observed was more pronounced than that observed in Irs2 null animals alone suggesting that Igf1r/Irs2 signaling pathway is critical for β-cell proliferation and function. Tissue-specific knockouts have also been used to reconstitute a polygenic model of type 2 diabetes. Mice homozygous null for the IRS-1 gene but heterozygous for a β-cell-specific glucokinase knockout, which individually did not give rise to an overt diabetic phenotype, developed hyperglycemia, β-cell hyperplasia, and eventually diabetes with age (Terauchi *et al*., 1997).

5. Mutagenesis

Knockout models of type 2 diabetes and obesity, although providing information on disease processes, do not mimic human disease, which is multigenic in nature and likely arises from common sequence variants rather than from null mutations. Disease in humans is predominately caused by the additive effects of multiple genes and environmental influences. Random mutagenesis using the point mutation *N*-ethyl-*N*-nitrosourea (ENU) is well established in the mouse, and large-scale phenotype driven ENU screens are increasingly being used for the systematic analysis of gene function throughout the mouse genome. Point mutations induced by ENU mutagenesis can lead to both hypomorph, hypermorph, gain of function, and dominant negative mutants, which may be more relevant than null alleles in human disease. Mutagenesis screens to date have utilized free-fed blood biochemistry in biochemical screens to identify mutants with disruption in glucose homeostasis (Nolan, 2000; Hough, 2002). While free-fed blood biochemistry is a quick assay, it may bias the mutations identified to autosomal dominant models of MODY (Toye, 2004). Adopting additional phenotyping protocols such as an intraperitoneal glucose tolerance test (IPGTT) or insulin suppression tests will identify more subtle insulin-resistant and glucose-intolerant phenotypes. Further sensitizing the screen utilizing existing knockout models, for instance, those predisposed to insulin resistance

without developing diabetes, will identify interacting genes, which alone would not have been identified in a standard dominant or recessive screen. The knockout models can be chosen to identify interacting genes in specific aspects of diabetes, for instance, an insulin receptor knockout and the insulin signaling pathway. Such approaches may identify novel genes not considered to be diabetes candidate genes. Furthermore, environment conditions can also be manipulated such as feeding a high-fat diet in order to exacerbate disease or identify susceptibility loci.

6. Quantitative trait loci (QTL) studies

Numerous quantitative trait loci (QTL) studies have been carried out in mouse taking advantage of an array of differing inbred mouse strains for a wide variety of diabetes related phenotypes including, glucose tolerance, blood plasma glucose, and insulin concentrations, body weight, lipid concentrations, and hypertension. To date, some 75 QTLs for obesity and 85 QTLs for body weight have been identified in various mouse lines and map to every chromosome except Y (for recent reviews, see Brockmann and Bevova, 2002; Cox and Brown, 2003). Few genes to date have been identified; however, the availability of complete genome sequence of the mouse is likely to speed up the search for candidate genes in QTL loci. Advances in the QTL field are likely to center around the use of more sophisticated breeding schemes that take advantage of the wide and significant differences between multiple inbred strains such as the use of heterogeneous stocks (Hss) (Mott, 2000).

7. RNA interference (RNAi)

Although not yet feasible in mouse, large-scale RNAi screens have been successfully applied to fat regulation in simpler organisms such as *C. elegans*. Ashrafi *et al*. (2003) have screened the expression of some 16 757 worm genes and assayed for modulations in fat storage. Utilizing a double-stranded RNAi bacterial library that contains some 86% of the 19 000 *C. elegans* genes expressed in *Escherichia coli*, which was fed to the worms as food. Additionally, a dye, Nile red, was also fed to the worms, which enables the visualization of fat storage droplets in living worms. They identified 305 gene inactivations that caused reduced body fat and 112 genes that caused increased fat storage. Although a number of genes identified in the screen were already known to have a key role in mammalian fat or lipid metabolism, over 50% of the fat regulatory genes identified in the study have mammalian homologs that have not been previously implicated in the regulation of fat storage and may present interesting genes for further study in rodent models.

8. Concluding remarks

The mouse has a proven track record as a model organism that has made significant contribution to the understanding of disease and biology relevant to the human.

It is not the only model organism by any means but for the study of human disease, it is perhaps one of the most widely used. We have illustrated the multiple approaches that can be brought to bear to model human disease using as examples the study of type 2 diabetes and obesity in the mouse. Spontaneous mutations and knockouts have given insight into the relative contributions of differing genes in the regulation of glucose homeostasis, insulin resistance β-cell function, and adiposity. Tissue-specific knockouts have gone some way into dissecting the relative contributions of multiple tissues and organs to disease progression. Combinations of different mutants, which individually would not produce disease have resulted in polygenic models of type 2 diabetes. Sensitized ENU mutagenesis screens may yield additional novel diabetes genes. The mouse has only been one of the approaches available to us to investigate human disease, but together with studies in human populations, other model organisms and in vitro systems, it has opened unprecedented opportunities in this new century to really tackle the problem of common human disease.

Further reading

Aizawa S, Eto K, Kimura S, Nagai R, Tobe K, Lienhard GE and Kadowaki T (2003) Impact of genetic background and ablation of insulin receptor substrate (IRS)-3 on IRS-2 knock-out mice. *Journal of Biological Chemistry*, **278**(16), 14284–14290.

References

Abel ED, Peroni O, Kim JK, Kim YB, Boss O, Hadro E, Minnemann T, Shulman GI and Kahn BB (2001) Adipose-selective targeting of the GLUT4 gene impairs insulin action in muscle and liver. *Nature*, **409**(6821), 729–733.

Accili D, Drago J, Lee EJ, Johnson MD, Cool MH, Salvatore P, Asico LD, Jose PA, Taylor SI and Westphal H (1996) Early neonatal death in mice homozygous for a null allele of the insulin receptor gene. *Nature Genetics*, **12**(1), 106–109.

Araki E, Lipes MA, Patti ME, Bruning JC, Haag B, Johnson RS and Kahn CR III (1994) Alternative pathway of insulin signalling in mice with targeted disruption of the IRS-1 gene. *Nature*, **372**(6502), 186–190.

Ashcroft F and Rorsman P (2004) Type 2 Diabetes mellitus: not quite exciting enough? *Human Molecular Genetics*, **13**, R21–R31.

Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer J and Ruvkun G (2003) Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. *Nature*, **421**, 268–272.

Baskin DG, Figlewicz Lattemann D, Seeley RJ, Woods SC, Porte D Jr and Schwartz MW (1999) Insulin and leptin: Dual adiposity signals to the brain for the regulation of food intake and body weight. *Brain Research*, **848**(1–2), 114–123.

Bevan P (2001) Insulin signalling. *Journal of Cell Science*, **114**, 1429–1430.

Brockmann GA and Bevova MR (2002) Using mouse models to dissect the genetics of obesity. *Trends in Genetics*, **18**, 367–376.

Bruning JC, Gautam D, Burks DJ, Gillette J, Schubert M, Orban PC, Klein R, Krone W, Muller-Wieland D and Kahn CR (2000) Role of brain insulin receptor in control of body weight and reproduction. *Science*, **289**(5487), 2122–2125.

Bruning JC, Michael MD, Winnay JN, Hayashi T, Horsch D, Accili D, Goodyear LJ and Kahn CR (1998) A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM without altering glucose tolerance. *Molecular Cell*, **2**(5), 559–569.

Bruning JC, Winnay J, Bonner-Weir S, Taylor SI, Accili D and Kahn CR (1997) Development of a novel polygenic model of NIDDM in mice heterozygous for IR and IRS-1 null alleles. *Cell*, **88**(4), 561–572.

Butler AA, Marks DL, Fan W, Kuhn CM, Bartolome M and Cone RD (2001) Melanocortin-4 receptor is required for acute homeostatic responses to increased dietary fat. *Nature Neuroscience*, **4**(6), 605–611.

Coleman DL and Eicher EM (1990) Fat (fat) and tubby (tub): Two autosomal recessive mutations causing obesity syndromes in the mouse. *Journal of Heredity*, **81**(6), 424–427.

Cox RD and Brown SDM (2003) Rodent Models of genetic disease. *Current Opinions in Genetics and Development*, **13**, 278–283.

Flier JS (2004) Obesity wars: Molecular progress confronts an expanding epidemic. *Cell*, **116**, 337–350.

Grupe A, Hultgren B, Ryan A, Ma YH, Bauer M and Stewart TA (1995) Transgenic knockouts reveal a critical requirement for pancreatic beta cell glucokinase in maintaining glucose homeostasis. *Cell*, **83**(1), 69–78.

Guillam MT, Hummler E, Schaerer E, Yeh JI, Birnbaum MJ, Beermann F, Schmidt A, Deriaz N, Thorens B and Wu JY (1997) Early diabetes and abnormal postnatal pancreatic islet development in mice lacking Glut-2. *Nature Genetics*, **17**(3), 327–330.

Hough TA, Nolan PM, Tsipouri V, Toye AA, Gray IC, Goldsworthy M, Moir L, Cox RD, Clements S, Glenister PH, *et al*. (2002) Novel phenotypes identified by plasma biochemical screening in the mouse. *Mammalian Genome*, **13**(10), 595–602.

Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, Berkemeier LR, Gu W, Kesterson RA, Boston BA, Cone RD, *et al*. (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell*, **88**(1), 131–141.

Joshi RL, Lamothe B, Cordonnier N, Mesbah K, Monthioux E, Jami J and Bucchini D (1996) Targeted disruption of the insulin receptor gene in the mouse results in neonatal lethality. *The EMBO Journal*, **15**(7), 1542–1547.

Kapeller R, Moriarty A, Strauss A, Stubdal H, Theriault K, Siebert E, Chickering T, Morgenstern JP, Tartaglia LA and Lillie J (1999) Tyrosine phosphorylation of tub and its association with Src homology 2 domain-containing proteins implicate tub in intracellular signaling by insulin. *Journal of Biological Chemistry*, **274**(35), 24980–24986.

Katz EB, Stenbit AE, Hatton K, DePinho R and Charron MJ (1995) Cardiac and adipose tissue abnormalities but not diabetes in mice deficient in GLUT4. *Nature*, **377**(6545), 151–155.

Kido Y, Burks DJ, Withers D, Bruning JC, Kahn CR, White MF and Accili D (2000) Tissue-specific insulin resistance in mice with mutations in the insulin receptor, IRS-1, and IRS-2. *Journal of Clinical Investigation*, **105**(2), 199–205.

Kim JK, Michael MD, Previs SF, Peroni OD, Mauvais-Jarvis F, Neschen S, Kahn BB, Kahn CR and Shulman GI (2000) Redistribution of substrates to adipose tissue promotes obesity in mice with selective insulin resistance in muscle. *The Journal of Clinical Investigation*, **105**(12), 1791–1797.

Kubota N, Tobe K, Terauchi Y, Eto K, Yamauchi T, Suzuki R, Tsubamoto Y, Komeda K, Nakano R, Miki H, *et al*. (2000) Disruption of insulin receptor substrate 2 causes type 2 diabetes because of liver insulin resistance and lack of compensatory beta-cell hyperplasia. *Diabetes*, **49**(11), 1880–1889.

Kulkarni RN, Bruning JC, Winnay JN, Postic C, Magnuson MA and Kahn CR (1999) Tissue-specific knockout of the insulin receptor in pancreatic beta cells creates an insulin secretory defect similar to that in type 2 diabetes. *Cell*, **96**(3), 329–339.

Laustsen PG, Michael MD, Crute BE, Cohen SE, Ueki K, Kulkarni RN, Keller SR, Lienhard GE and Kahn CR (2002) Lipoatrophic diabetes in Irs1(-/-)/Irs3(-/-) double knockout mice. *Genes and Development*, **16**(24), 3213–3222.

Liu SC, Wang Q, Lienhard GE and Keller SR (1999) Insulin receptor substrate 3 is not essential for growth or glucose homeostasis. *Journal of Biological Chemistry*, **274**(25), 18093–18099.

Michael MD, Kulkarni RN, Postic C, Previs SF, Shulman GI, Magnuson MA and Kahn CR (2000) Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction. *Molecular Cell*, **6**(1), 87–97.

Mott R, Talbot C, Turri M, Collins A and Flint J (2000) A method of fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences*, **97**, 12649–12654.

Nolan PM, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, Gray IC, Vizor L, Brooker D, Whitehill E, *et al*. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genetics*, **25**(4), 440–443.

Ohlemiller KK, Hughes RM, Lett JM, Ogilvie JM, Speck JD, Wright JS and Faddis BT (1997) Progression of cochlear and retinal degeneration in the tubby (rd5) mouse. *Audiology and Neuro-otology*, **2**(4), 175–185.

Postic C, Shiota M, Niswender KD, Jetton TL, Chen Y, Moates JM, Shelton KD, Lindner J, Cherrington AD and Magnuson MA (1999) Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *Journal of Biological Chemistry*, **274**(1), 305–315.

Stenbit AE, Tsao TS, Li J, Burcelin R, Geenen DL, Factor SM, Houseknecht K, Katz EB and Charron MJ (1997) GLUT4 heterozygous knockout mice develop muscle insulin resistance and diabetes. *Nature Medicine*, **3**(10), 1096–1101.

Stubdal H, Lynch CA, Moriarty A, Fang Q, Chickering T, Deeds JD, Fairchild-Huntress V, Charlat O, Dunmore JH, Kleyn P, *et al*. (2000) Targeted deletion of the tub mouse obesity gene reveals that tubby is a loss-of-function mutation. *Molecular Cell Biology*, **20**(3), 878–882.

Terauchi Y, Iwamoto K, Tamemoto H, Komeda K, Ishii C, Kanazawa Y, Asanuma N, Aizawa T, Akanuma Y, Yasuda K, *et al*. (1997) Development of non-insulin-dependent diabetes mellitus in the double knockout mice with disruption of insulin receptor substrate-1 and beta cell glucokinase genes. Genetic reconstitution of diabetes as a polygenic disease. *The Journal of Clinical Investigation*, **99**(5), 861–866.

Terauchi Y, Matsui J, Suzuki R, Kubota N, Komeda K, Aizawa S, Eto K, Kimura S, Nagai R, Tobe K, *et al*. (2003) Impact of genetic background and ablation of insulin receptor substrate (IRS)-3 on IRS-2 knock-out mice. *Journal Biological Chemistry*, **278**(16), 14284–14290.

Terauchi Y, Sakura H, Yasuda K, Iwamoto K, Takahashi N, Ito K, Kasai H, Suzuki H, Ueda O, Kamada N, *et al*. (1995) Pancreatic beta-cell-specific targeted disruption of glucokinase gene. Diabetes mellitus due to defective insulin secretion to glucose. *Journal of Biological Chemistry*, **270**(51), 30253–30256.

Thorens B, Guillam MT, Beermann F, Burcelin R and Jaquet M (2000) Transgenic reexpression of GLUT1 or GLUT2 in pancreatic beta cells rescues GLUT2-null mice from early death and restores normal glucose-stimulated insulin secretion. *Journal Biological Chemistry*, **275**(31), 23751–23758.

Toye AA, Moir L, Hugill H, Bentley L, Quaterman J, Mijat V, Hough T, Goldsworthy M, Haynes A, Hunter AJ, *et al*. (2004) A new mouse model of type 2 diabetes, produced by N-Ethyl-Nitrosourea mutagenesis, is the result of a missense mutation in the glucokinase gene. *Diabetes*, **53**(6), 1577–1583.

Withers DJ, Gutierrez JS, Towery H, Burks DJ, Ren JM, Previs S, Zhang Y, Bernal D, Pons S, Shulman GI, *et al*. (1998) Disruption of IRS-2 causes type 2 diabetes in mice. *Nature*, **391**(6670), 900–904.

Withers DJ, Burks DJ, Towery HH, Altamuro SL, Flint CL and White MF (1999) Irs-2 coordinates Igf-1 receptor-mediated beta-cell development and peripheral insulin signalling. *Nature Genetics*, **23**(1), 32–40.

Zisman A, Peroni OD, Abel ED, Michael MD, Mauvais-Jarvis F, Lowell BB, Wojtaszewski JF, Hirshman MF, Virkamaki A, Goodyear LJ, *et al*. (2000) Targeted disruption of the glucose transporter 4 selectively in muscle causes insulin resistance and glucose intolerance. *Nature Medicine*, **6**(8), 924–928.

Specialist Review

The rat as a model physiological system

Dominique Gauguier
University of Oxford, Oxford, UK

1. Introduction

Much of our current understanding of integrative physiology is based on studies in the laboratory rat, and the body of physiological and pathophysiological data, as well as toxicological and pharmacological data, that are available for the rat is unparalleled in other species (Jacob and Kwitek, 2001). The rat has been the favored model for physiologists in various fields of biomedical research because of its large size, which facilitates invasive and repeated experimental interventions that remain technically difficult or impossible to apply in mice. It provides a broad range of unique and thoroughly studied models of common and prevalent human complex disorders, including type 1 and type 2 diabetes mellitus, essential hypertension, stroke, obesity and renal, neurological and autoimmune disorders. This chapter focuses on the description of rat models and tools used in the genetic study of quantitative traits (*see* Article 58, **Concept of complex trait genetics**, Volume 2), which is a particularly active field of research in the rat.

2. Rat models commonly used in genetic studies

Over 230 different inbred rat strains have been derived since the inbreeding of the first rat strain in 1909 – the same year that inbreeding of the first mouse strain began. Although genetic research has focused on well-characterized rat models of Mendelian and polygenic diseases (Table 1), an important proportion of inbred rat strains, showing potentially considerable genetic diversity, are still underutilized in physiological and genetics studies.

In contrast to mouse models (*see* Article 38, **Mouse models**, Volume 3), there are few rat models of disease traits spontaneously occurring as the consequence of a single gene mutation (see Section 7.1). In the absence of models spontaneously mirroring the etiology and pathogenesis of the most frequent and prevalent human diseases (e.g., diabetes mellitus, hypertension), a process of repeated breeding of increasingly affected animals isolated from an outbred stock has been used to produce rat strains exhibiting specific disease phenotypes (Table 1). This procedure

Table 1 Pathophysiological characteristics of major inbred rat strains used in QTL mapping experiments

Strain	Description	Original stock	Selection criterion[a] or main disease features[b]
ACI	August x Copenhagen Irish	–	Resistance to arthritis[b]
AS	Albino Surgery	Wistar	High blood pressure[a]
BBDP	Biobreeding diabetes prone	Wistar	Lymphopenia[b], type 1 diabetes mellitus[b]
BN	Brown Norway	–	High IgE[b], aortic IEL lesions[b], spike-wave discharges[b], resistance to mammary and liver cancers[b], resistance to arthritis[b]
CDS	Cohen diabetic sensitive	Unknown	Glucose intolerance on high-sucrose low-copper diet[a]
COP	Copenhagen	–	Resistance to mammary and liver cancers[b]
DA	Dark Agouti	–	Experimentally induced arthritis and autoimmune encephalomyelitis[b]
DRH	Donryu-RH	Donryu	Resistance to liver cancer[a]
E3	–	–	Resistance to pristane-induced arthritis[b]
FHH	Fawn Hooded	German-brown x White Lashley	High blood pressure[a]
F344	Fisher	–	Cancer susceptibility[b], spike-wave discharges[b], resistance to arthritis[b]
GAERS	Genetic absence epilepsy	Wistar	Absence seizures and spike-wave discharges[a] Rat from Strasbourg
GH	Genetically hypertensive	Wistar	High blood pressure[a]
GK	Goto-Kakizaki	Wistar	Glucose intolerance[a]
HTG	Hereditary hypertriglyceridemic	Wistar	Elevated plasma triglycerides[a]
KDP	Komeda diabetes prone	LETL	Type 1 diabetes mellitus[b]
LEW	Lewis	–	Experimental autoimmune encephalomyelitis[b]
LH	Lyon hypertensive	Sprague-Dawley	High blood pressure[a]
LN	Lyon normotensive	Sprague-Dawley	Normal blood pressure[a]
MHS	Milan hypertensive	Wistar	High blood pressure[a]
MNS	Milan normotensive	Wistar	Normal blood pressure[a]
MWF	Munich Wistar Fromter	Wistar	High number of superficial glomeruli[a], proteinuria[b], glomerulosclerosis[b], hypertensionb
OLETF	Otsuka Long Evans Fatty	Long Evans	Glucose intolerance[a] and obesity[b]
PKD(*Cy*)	–	Sprague-Dawley	Polycystic kidney disease[b]
RHA/Verh	Roman high avoidance	Wistar	Conditioned-avoidance response[a]
RLA/Verh	Roman low avoidance	Wistar	Conditioned-avoidance response[a]
SBH	Sabra hypertensive	Unknown	High blood pressure in response to unilateral nephrectomy[a]
SHR	Spontaneously hypertensive	Wistar	High blood pressure[a]
SHRSP	SHR Stroke prone	Wistar (SHR)	Stroke susceptibility[a]
SS	Salt sensitive	Sprague-Dawley	Salt induced high blood pressure[a]
SR	Salt resistant	Sprague-Dawley	Low blood pressure on high salt diet[a]
SDT	Spontaneously Diabetic Torii	Sprague-Dawley	Hyperglycemia and glucosuria[a]
WAG/Rij	Wistar Albino Glaxo	Wistar	Absence seizures and spike-wave discharges[a]
WBN/Kob	Wistar Bonn/Kobori	Wistar	Chronic pancreatitis and diabetes mellitus[b]
WF	Wistar Furth	Wistar	Susceptibility to DMBA-induced mammary cancer[b]
WKHA	Wistar Hyperactive	Wistar (SHR/WKY)	Normal blood pressure and hyperactivity[a]
WKY	Wistar Kyoto	Wistar	Normal blood pressure[a], resistance to mammary cancers[b]

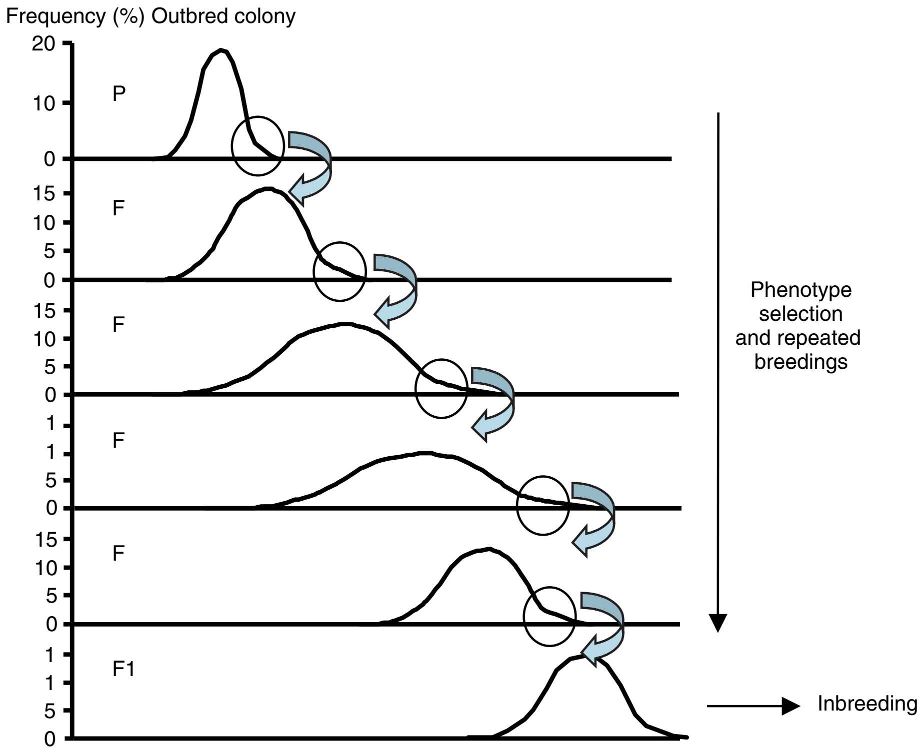


Figure 1 Selection of naturally occurring disease gene variants in inbred rats by phenotype screening and repeated breeding of outbred rats

(illustrated in Figure 1) is based on the assumption that naturally occurring alleles altering biological processes exist in outbred rats and can be concentrated in an inbred strain. A single pathophysiological criterion was applied for isolating animals carrying disease susceptibility or resistance alleles. Commonly used strains derive from either Wistar or Sprague-Dawley outbred stocks, and genetic diversity is therefore relatively limited.

Other rat models including chromosome substitution (consomic) strains, which are proposed as novel tools for assigning quantitative traits to an entire chromosome (Cowley *et al*., 2004), and congenic strains (described in Section 6.1) are being developed. Progress is also being made in the application of ENU mutagenesis and gene disruption technologies to the rat (Zan *et al*., 2003), which will undoubtedly have an important impact in physiological and functional genomics studies in the rat.

3. Rat genetic mapping panels

Most of the rat genetic studies are based on genetic and phenotypic analyses in classical segregating populations (F2 or first backcross cohorts). Crosses have

been generated using inbred strains either isolated from the same outbred stock (e.g., SHRxWKY, LHxLN, WKYxWKHA) or genetically unrelated, allowing increased polymorphism rate and the potential of contrasting alleles, and ultimately facilitating the detection and fine mapping of quantitative trait loci (QTLs) (*see* Article 11, **Mapping complex disease phenotypes**, Volume 3).

The panel of recombinant inbred (RI) strains derived from BN and SHR represents a permanent resource for QTL mapping (Pravenec *et al*., 1995). Although these strains primarily represent a mosaic of blood pressure (BP) contrasting alleles, other physiological characteristics in the founder strains (Table 1) provide opportunities for mapping multiple phenotypes and genotype-environmental interactions.

The rat genetically heterogeneous stock (HS), founded from BN/SsN, MR/N, BUF/N, M520/N, WN/N, ACI/N, WKY/N, and F344/N strains (Hansen and Spuhler, 1984), is in many respects similar to its mouse counterpart, and could be used to map the genetic basis of complex traits to subcentimorgan resolution (Mott and Flint, 2002). Well-established disease phenotypes in BN and F344 strains (Table 1) should make this panel amenable to QTL mapping.

4. Rat genetic and genomic resources

With the growing interest in disease gene mapping in the rat and comparative genomics, the production of rat genetic and genomic resources has closely followed those generated for the mouse. International efforts have progressively generated a large collection of rat microsatellite markers characterized for allele variations in strains commonly used in genetic studies (**http://www.well.ox.ac.uk/rat_mapping_resources**; **http://rgd.mcw.edu**). Polymorphism rates between strains vary from 20% for strains derived from the same outbred stock (e.g., SHRSP vs. WKY) to over 70% for genetically distant strains (e.g., SHR vs. BN or GK vs. BN). Results from polymorphism assays also showed the high genetic variability between colonies of supposedly identical strains (WKY, SHR, and SHRSP) (Bihoreau *et al*., 1997a). Three crosses (FHHxACI, SHRSPxBN, BNxGK) (Steen *et al*., 1999; Wilder *et al*., 2004) and a radiation hybrid (RH) panel (Steen *et al*., 1999; Watanabe *et al*., 1999) were used to localize the vast majority of markers in the rat genome.

The most significant recent developments in rat genomics are the completion of the rat genome sequence (Rat Genome Sequencing Project Consortium, 2004) and the emergence of rat single nucleotide polymorphism (SNP) markers (Zimdahl *et al*., 2004), which open important perspectives for refining genealogies of inbred rat strains and identifying disease genes.

5. Genetics of complex phenotypes in the rat

The multiple pathophysiological components participating in the onset and progression of complex diseases, their polygenic control, the involvement of environmental factors and gene–gene and gene–environment interactions underline the requirement of extensive biological screens for accurately assessing disease phenotypes (*see* Article 58, **Concept of complex trait genetics**, Volume 2). The possibility to

collect relatively large samples of biofluids and organs is a key element in rat quantitative genetic studies. To date, over 700 QTLs reflecting the effect of natural gene variants have been reported in the rat (**http://www.ensembl.org/Rattus_norvegicus**; **http://rgd.mcw.edu**). This section focuses on the most significant features of rat QTL studies illustrated in Figure 2, and their contribution to our understanding of the etiology of multifactorial diseases.

5.1. Susceptibility to experimentally induced disorders

Knowledge of strain-specific susceptibility and resistance to experimentally induced disorders (Table 1) has been particularly useful to investigate the genetics of common autoimmune and inflammatory processes and cancers in rat crosses. Immunization of LEW and DA rats with spinal cord tissue or myelin oligodendrocyte glycoprotein results in experimental autoimmune encephalomyelitis (EAE), a model of multiple sclerosis. Rheumatoid arthritis (RA) is induced in the DA rat by injection of cartilage antigens (collagen-induced arthritis – CIA) or nonimmunogenic substances with strong adjuvant effects (pristane-induced arthritis – PIA or oil-induced arthritis – OIA). Over 30 QTLs regulating RA susceptibility, chronicity, and severity have been mapped in crosses derived from the DA rat and either CIA-resistant (F344, BN, ACI, E3), PIA-resistant (E3, F344, LEW.1AV1), or OIA-resistant (LEW.1AV1) strains (Griffiths and Remmers, 2001; Olofsson *et al*., 2003a). The overlap between RA and EAE QTLs in the DA rat suggests the influence of common genes (Dahlman *et al*., 1998; Bergsteinsdottir *et al*., 2000). BN-specific enhanced IgE responsiveness to injections of aurothiopropanol sulfonate is under polygenic control (Mas *et al*., 2000). Susceptibility and resistance to dimethylbenzanthracene-induced mammary carcinogenesis, hepatocellular carcinoma, and estrogen-induced pituitary tumors have been mapped in the rat genome (Wendell *et al*., 2000; Lan *et al*., 2001; De Miglio *et al*., 2004).

In the field of type 2 diabetes mellitus (T2DM) and obesity, neonatal injection of streptozotocin (STZ), an alkylnitrosourea that specifically destroys pancreatic insulin-producing β-cells, dietary changes (e.g., high fat and cafeteria diets) and lesions of ventromedial hypothalamus nuclei have been used, primarily in rats, to induce permanent diabetes, insulin resistance, and obesity. Strain-specific resistance/susceptibility to diabetes/obesity induced by STZ or dietary changes, which has been applied to QTL mapping in mice, remains to be tested in the rat.

5.2. Spontaneously occurring disease phenotypes

Although neurobehavioral phenotypes are defined by procedures primarily developed in rats, only two studies have successfully explored the genetics of anxiety and activity in the rat (Moisan *et al*., 1996; Fernandez-Teruel *et al*., 2002). Invasive methods routinely used in rats to acquire repeated and prolonged cortical electroencephalogram recording provide interesting perspectives in neurogenetics as qualitative and quantitative information on spike-wave discharges (SWD) can be determined. Recent studies have demonstrated that SWD phenotypes in GAERS

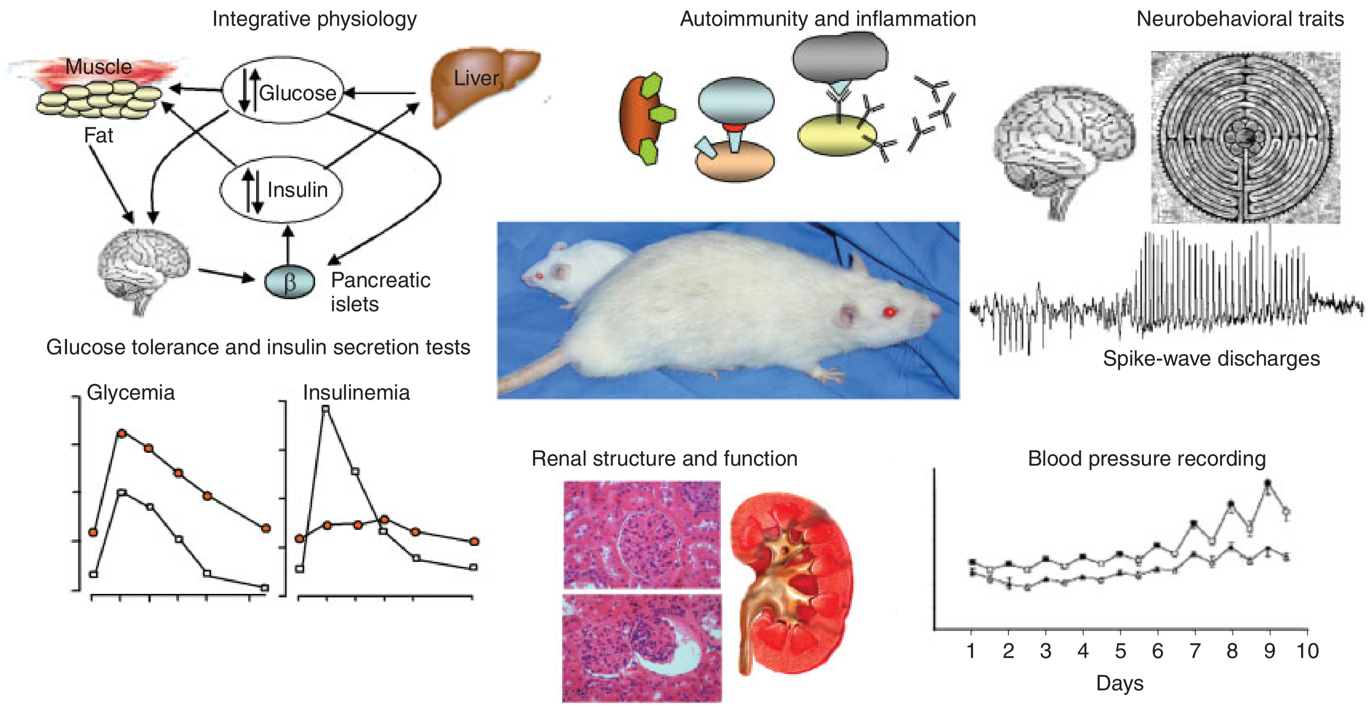


Figure 2 Examples of phenotypes mapped to the rat genome by QTL analysis. The size of the rat allowed the quantification of subphenotypes using procedures that are often technically difficult to perform in the mouse (in the background)

and WAG/Rij rat models of spontaneous absence seizures are under polygenic control and that different loci regulate SWD subtypes in the WAG/Rij rat (Gauguier *et al*., 2004; Rudolf *et al*., 2004).

Studies in the Bio-Breeding Diabetes Prone (BBDP) and Komeda Diabetes Prone (KDP) strains have generated a vast amount of information on the genetic and immunological basis of type 1 diabetes mellitus (T1DM) (autoimmune destruction of pancreatic beta-cells resulting in insulinopenia and severe hyperglycemia) (Jacob *et al*., 1992; Ramanathan and Poussier, 2001 Yokoi *et al*., 2002; Ramanathan *et al*., 2002).

Three rat strains (*Cy*, *Pck*, and *Wpk*) develop renal pathologies closely resembling human polycystic kidney diseases (PKD). Interestingly, modifier genes controlling disease variable severity, which is a phenomenon described in human PKD, have been mapped in the *Cy* rat (Bihoreau *et al*., 1997b; Bihoreau *et al*., 2002).

Essential hypertension (*see* Article 63, **Hypertension genetics: under pressure**, Volume 2) and T2DM (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2) are among the best examples of inherited quantitative phenotypes that have been extensively studied in the rat. The following sections emphasize the importance of the laboratory rat for acquiring intermediate phenotypes relevant to these disorders, which can be used in QTL detection experiments.

5.3. Genetic determinants of hypertension in the rat

Genetic studies of BP control have pioneered QTL mapping in the rat. The first successful localization of a genetic locus controlling BP quantitative variations was obtained with an RFLP marker of the renin gene in an SSxSR cross (Rapp *et al*., 1989). This line of research gained momentum with the increasing number of rat genetic markers, which allowed genome-wide scans of loci controlling cardiovascular traits in all spontaneously hypertensive rat strains (Rapp, 2000). The first evidence of the polygenic control of BP variables was obtained in a SHRSPxWKY cross (Hilbert *et al*., 1991; Jacob *et al*., 1991).

Our current knowledge of BP genetics in rats is based on data from over 25 different crosses, which allowed the mapping of over 190 QTLs in the rat genome (**http://rgd.mcw.edu**). The strongest evidence of replicated linkage to BP was obtained in chromosomes 1, 2, 10, and 13 (Rapp, 2000). Direct comparisons between results from these studies are problematic because different experimental conditions (e.g., age at phenotyping, methods for BP recording) and strain combinations were used. In this respect, extensive genetic studies in hybrids derived from SS rats bred to different strains, which were maintained in identical conditions and characterized for BP using identical procedures, strongly suggest that the genetic background of normotensive strains in experimental crosses influences QTL replication (Rapp, 2000).

A strategy of BP phenotype dissection initiated by Dubay *et al*. (1993) in hybrid rats of a segregating population was later extended to multiple BP-related biological parameters (Stoll *et al*., 2001) and phenotypes relevant to end-organ complications.

Although rat BP QTLs are often associated with effects on heart weight and left ventricular mass (Rapp, 2000), cardiac hypertrophy is a common but not inevitable complication of hypertension. Data from crosses between normotensive strains provide confirmatory evidence for the independent genetic control of cardiac mass and BP (Sebkhi *et al*., 1999; Gauguier *et al*., 2005). The existence of a causal link between hypertension and stroke in the SHRSP strain is also debated. The relationship between susceptibility to stroke and hypertension was addressed in two genetic studies, which gave fundamentally different results (Rubattu *et al*., 1996; Jeffs *et al*., 1997). Renal damage is present in almost all hypertensive rat strains (Rapp, 2000). Results from genetic studies in hypertensive rats strongly suggest that genetic factors independent of BP influence the progression and possibly the onset of renal disease (Brown *et al*., 1996; Gigante *et al*., 2003; Garrett *et al*., 2003). Of note, MNS rats develop glomerulosclerosis and proteinuria in the absence of hypertension.

Among abnormal vascular phenotypes observed in inbred rats, ruptures of the aortic internal elastic lamina (AIEL) spontaneously occur in BN rats, whereas other normotensive strains are devoid of these lesions (Behmoaras *et al*., 2005). Genetic studies in the BN rat allowed the mapping of independent QTLs for AIEL lesions (Harris *et al*., 2001) and variations of aortic elastin, collagen, and cell protein contents (Gauguier *et al*., 2005).

5.4. Type 2 diabetes mellitus, insulin resistance, and obesity

The laboratory rat is particularly appropriate for investigating the multiple metabolic and hormonal defects occurring in different tissues (pancreas, liver, fat, muscles) that contribute to insulin secretion deficiency and insulin resistance and ultimately hyperglycemia. Polygenic inheritance of diabetes-related traits has been demonstrated in rat models of spontaneous T2DM (GK, OLETF, SDT) (Galli *et al*., 1996; Gauguier *et al*., 1996; Kanemoto *et al*., 1998; Masuyama *et al*., 2003). The dissection of diabetes in subphenotypes in a GKxBN cross allowed the detection of independent QTLs for fasting glycemia and insulinemia, glucose tolerance, insulin secretion in response to glucose or arginine in vivo, adiposity and body weight (Gauguier *et al*., 1996). The genetics of other important pathophysiological components of diabetes in the GK strain (impaired beta cell, reduced β-cell mass function, dyslipidemia) is not explained by these results. A remarkable feature in the genetics of diabetes in these strains is the consistent mapping of diabetes QTLs to the same regions of chromosomes 1 and 2, despite differences in the procedures applied to assess glucose homeostasis and in the control strain (BN or F344) used to generate the experimental cross (Galli *et al*., 1996; Gauguier *et al*., 1996; Masuyama *et al*., 2003).

Among the common vascular complications of T2DM (neuropathy, retinopathy, and nephropathy), renal structural damage and albuminuria have been described in all rat models of spontaneous or experimentally induced diabetes and/or obesity, suggesting a permissive role of hyperglycemia, obesity, or hyperlipidemia on susceptibility to renal disease. The relative contribution of specific genetic factors and hyperglycemia on renal damage has not been addressed in diabetic models.

6. Strategies and tools for post QTL mapping studies

QTL detection in experimental crosses or consomic strains is only a preliminary stage in the genetic analysis of a quantitative phenotype. Further investigations are required to confirm the existence of a QTL, refine its chromosomal localization, and most importantly test the pathophysiological consequences of the underlying gene variant(s).

Congenic strains carrying segments of chromosomes harboring a QTL introgressed onto a permissive background (usually the reciprocal strain in the experimental cross) currently provides the most reliable way of progressing from mapping of a QTL to identification of the underlying gene (Rogner and Avner, 2003). Congenics designed using various strain combinations are now widely used for dissecting rat QTLs (Table 2).

From a fundamental perspective, the phenotypic impact of QTLs can be "weighed" by comparing quantitative traits in congenic strains and in segregating

Table 2 Rat congenic panels designed to dissect the pathophysiology of complex phenotypes and fine map QTL. Standard nomenclature (Recipient.Donor) indicates the origin of the donor alleles (QTL target) introgressed onto the genetic background of the recipient strain. A series of congenic strains were generally derived for different regions covering the QTL

Primary phenotype	Congenic series	Chromosomes
Hypertension	SHR.BN	1,5,8,13,19
	BN.SHR	2
	SHR.WKY, WKY.SHR	1,2,Y
	SHRSP.WKY	1,2
	WKY.SHRSP	2,10
	SS/jr.LEW	1,5,8,10,17
	SS/jr.MNS	2,10
	SS/jr.WKY	2
	SS/jr.SR/jr	3,7,9,13
	MNS.MHS, MHS.MNS	1 (*Add3*),4 (*Add2*),14 (*Add1*)
	SBN.SBH, SBH.SBN	1
Type 1 diabetes	BBDR.BBDP	4 (*lyp*)
	BB/OK.SHR	1,6,18,X
	F344.BBDP	4 (*lyp*)
Type 2 diabetes	BN.GK	2,8
	F344.GK	1
	F344.OLETF	5,7,8,9,11,12,14,16
	OLETF.F344	1
Cancer	WF.COP	2
	BN.F344, WF.WKY	5
Atopy	LEW.BN, BN.LEW	9,10
Nephropathy	LOU.BN	10
Arthritis	DA.F344, LEW1AV1.DA	10
	DA.E3	4,12
	DA.LEW1AV1, LEW1AV1.DA	4
	DA.PVG1AV1, PVG1AV1.DA	4
EAE	ACI.DA	10
Renal failure	ACI.FHH	1

populations (Garrett *et al*., 1998). Furthermore, studies in double congenics, which contain genomic regions covering two different QTLs, can help evaluate functional relationships (epistasis) between QTLs (Rapp *et al*., 1998). Congenics also provide opportunities to test the phenotypic impact of disease variants at a QTL when introgressed onto different genetic backgrounds.

From a biological angle, studies in congenics can test gender effects and gene-environment interactions, and characterize disease onset and progression using different or more refined phenotype procedures than those used in the original cross (Wallis *et al*., 2004). For example, a fundamental role of gene(s) at a RA QTL on arthritis susceptibility and progression was deduced from the reduced arthritogenic responses to collagen, pristane, squalene, and adjuvant in DA.PVG congenics (Backdahl *et al*., 2003). The most important outcome of studies in congenics lies in the possibility to translate statistical estimates supporting QTL localization to a genomic interval, thus defining anchor points in the rat genome sequence for candidate gene identification (see Section 7.2). However, the existence of variants in multiple genes contributing to a QTL effect is a hallmark of almost all examples of QTL dissection in congenics, making the selection of candidate genes problematic. Several genes independently contribute to rat QTLs for tumor resistance (Haag *et al*., 2003), hypertension (Garrett *et al*., 2001; Saad *et al*., 2001; Frantz *et al*., 2001), T2DM (Wallace *et al*., 2004), and atopy (Mas *et al*., 2004).

Among other strategies allowing fine QTL mapping, advanced intercross lines have been used for separating closely linked EAE QTLs (Jagodic *et al*., 2004). Genetic studies in the rat HS can also speed up disease gene identification (see Section 3), providing that the founder strains exhibit marked differences in phenotypes of interest. However, these strategies give limited information on the phenotypic effects of the causative gene variants.

Profiling gene expression is a powerful strategy for assisting the selection of candidate genes in a congenic interval. Commercial and custom arrays of oligonucleotides are now available for interrogating the level of tens of thousands of known and predicted rat transcripts. A major problem with the application of this technology to complex traits such as T2DM is knowing which tissue (pancreatic islets, liver, muscles, fat, nervous tissues) should be examined for differential gene expression. It can nevertheless be resolved in the rat by physiological studies that will indicate specific target tissue(s) and experimental conditions. The large amount of RNA that can be obtained from rat tissues and structurally or functionally different regions of an organ is also an advantage. Proteomics, which still requires relatively large amounts of biological material, is also an attractive technology for selecting candidate genes in rat congenic intervals.

7. Disease gene identification in rat models

A growing number of genes responsible for monogenic and polygenic traits have been identified in the rat (Table 3).

Table 3 Molecular basis of Mendelian disease and complex trait genes in rat models

Gene	Description	Disease strain	Mutation or sequence variation	Phenotype	References
Atp7b	Copper transporting ATPase	LEC	Partial deletion	Autosomal recessive hepatocellular necrosis	Wu *et al*. (1994)
Cblb	Casitas B-lineage lymphoma B	KDP	Nonsense mutation (Arg455X)	Type 1 diabetes	Yokoi *et al*. (2002)
Csf1	Macrophage colony stimulating factor	LEW (*tl*)	10-bp repeat insertion	Autosomal recessive osteopetrosis	Dobbins *et al*. (2002)
Cd36	Fatty acid translocase	SHR	Chromosomal deletion	Insulin resistance	Aitman *et al*. (1999)
Cyp11b1	11 beta-hydroxylase	SS/jr	Multiple amino acid changes	High blood pressure	Cicila *et al*. (1993)
Ide	Insulin degrading enzyme	GK	Missense variants (H18R, A890V)	Insulin resistance, glucose intolerance	Fakhrai-Rad *et al*. (2000)
Ian4	Immune-associated nucleotide 4	BBDP	Frameshift mutation in exon 3	Autosomal recessive lymphopenia	MacMurray *et al*. (2002), Hornum *et al*. (2002)
Lepr	Leptin receptor	Zucker (*fa*)	Missense mutation (Gln129Pro)	Autosomal recessive obesity	Phillips *et al*. (1996)
Ncf1	Neutrophil cytosolic factor 1	DA	Coding variants (M106V, M153T)	Arthritis severity	Olofsson *et al*. (2003b)
Mertk	Receptor tyrosine kinase	RCS	Frameshift mutation	Autosomal recessive retinal dystrophy	D'Cruz *et al*. (2000)
Pax-6	Homeobox paired box-6	rSey	Single base (G) exonic insertion	Autosomal recessive impaired brain development	Matsuo *et al*. (1993)
Pkhd1	Fibrocystin/polyductin	SD (*Pck*)	Splicing change (IVS35-2A-T)	Autosomal recessive PKD	Ward *et al*. (2002)
Ship2	Type-II SH2-domain-containing inositol 5-phosphatase	GK, SHR	Missense mutation (R1142C)	Insulin resistance	Marion *et al*. (2002)

7.1. Monogenic models

Recessive obesity in the Zucker *fatty* rat develops as a consequence of an amino acid substitution in the leptin receptor gene, which is also mutated in the *db/db* mouse (Phillips *et al*., 1996). Retinal degeneration in the Royal College of Surgeons (RCS) rat is caused by a mutation in a gene encoding a receptor tyrosine kinase, which leads to a loss of photoreceptors (D'Cruz *et al*., 2000). The *tl* (toothless) osteopetrotic rat carries a 10-bp insertion within the coding sequence of the macrophage colony stimulating factor gene (Dobbins *et al*., 2002). Autosomal recessive PKD in the PCK rat is caused by a splicing change in the fibrocystin/polyductin protein (Ward *et al*., 2002). Lymphopenia in the BBDP rat is caused by a frameshift mutation in the gene *Ian4* (also called *Ian5*) involved in immune mechanisms and the regulation of apoptosis (MacMurray *et al*., 2002; Hornum *et al*., 2002).

7.2. Polygenic models

Two T1DM susceptibility genes have been identified by positional cloning in BB and KDP rats. *Iddm2* maps to the $RT1^u$ haplotype of the class II MHC locus in the BB strain, but the precise mechanisms through which it predisposes to T1DM remains unknown (Ellerman and Like, 2000). A nonsense mutation in an ubiquitin-protein ligase involved in tyrosine kinase signaling pathways accounts for the T1DM locus *Kdp1* in the KDP rat (Yokoi *et al*., 2002).

Fine QTL mapping in congenics has proved an important strategy for detecting functional gene variants explaining, at least in part, central features of hypertension, T2DM, RA, and insulin resistance QTLs. In the SS strain, the causative role of amino acid substitutions in the 11 beta-hydroxylase (*Cyp11b1*) protein (Cicila *et al*., 1993) was eventually validated in a SS.SR congenic strain (Garrett and Rapp, 2003). Combining gene transcription profiling and phenotype investigations in SHR/*NCrj* congenic strains, a deletion in a fatty acid translocase gene (*Cd36*) was identified, which is probably relevant to insulin resistance in isolated adipocytes (Aitman *et al*., 1999). The absence of the mutation in the SHRSP/*Izm* colony believed to be the founder strain of SHR and SHRSP strains (Gotoda *et al*., 1999) suggests the occurrence of a de novo mutation in the SHR/*NCrj* strain. Functional variants have been recently described in the GK rat in genes localized in diabetes QTLs in rat chromosome 1 (Galli *et al*., 1996; Gauguier *et al*., 1996). Variants in the insulin degrading enzyme account for a diabetic phenotype in a F344.GK congenic strain (Fakhrai-Rad *et al*., 2000). A functional mutation specific to the GK and SHR strains was identified in the gene *Ship2*, which is involved in insulin stimulated glucose transport and lipid synthesis (Marion *et al*., 2002). In DA rats, a polymorphism in the neutrophil cytosolic factor gene (*Ncf1*) causes a decrease in oxygen burst, resulting in severe arthritis (Olofsson *et al*., 2003b).

8. Translating rat disease genes to human genetics

Rat genomic sequence data (Rat Genome Sequencing Project Consortium, 2004) have dramatically enriched the resources required for comparative genome

analyses between rat, mouse, and human, which were established previously by chromosomal mapping of rat gene and EST sequences (Watanabe *et al*., 1999; Wilder *et al*., 2004). Given the high degree of conservation of synteny and gene order in the three species, human chromosomal regions homologous to rat QTLs and congenic intervals can be easily identified.

A striking concordance has often been reported between the localization of rat QTLs and susceptibility loci for human essential hypertension (Stoll *et al*., 2000), epilepsy (Pinto *et al*., 2005), T2DM (Wallace *et al*., 2004), atopy (Mas *et al*., 2000), renal disorders (Bihoreau *et al*., 2002), and autoimmune/inflammatory diseases (Griffiths and Remmers, 2001). These findings emphasize the importance of investigations in the rat for prioritizing genetic studies in human to specific chromosomal regions.

Ultimately, genetic and functional genomic studies in rats can generate gene targets to be tested in human genetics. Mutations in the human orthologs of rat *Mertk* and *Pck* (PKHD1) have been found in patients with retinitis pigmentosa (Gal *et al*., 2000) and PKD (Ward *et al*., 2002), respectively. In the field of multifactorial disorders, the strongly significant association between an haplotype in SHIP2 (see Section 7.2) and components of the insulin resistance syndrome (T2DM, obesity, and hypertension) is the only example of successful translation of results from rat quantitative genetics to human complex diseases (Kaisaki *et al*., 2004).

9. Conclusions

Building on the wealth of physiological information that can be obtained in the rat, genetic loci cosegregating with phenotypes underlying key pathological features of human multifactorial disorders have been mapped in the rat genome. Ongoing studies in congenic strains have validated most rat QTLs, owing to their relatively modest phenotypic effects in experimental crosses, and reports identifying rat disease gene variants are now emerging in the literature. Parallel cross-disciplinary phenotyping in existing inbred rat strains and emerging congenic and consomic panels should further enhance the input of rat models in quantitative genetics research. Rat genomic sequence data and progress in comparative genomics will enhance the power of studies in the rat and lead, synergistically with studies in mice (*see* Article 38, **Mouse models**, Volume 3), to a better understanding of the role of natural variants in mechanisms involved in the etiology of human complex diseases.

Acknowledgments

Dominique Gauguier is Reader in Mammalian Genetics at the University of Oxford and holds a Wellcome Senior Fellowship in Basic Biomedical Science (057733). The author acknowledges support from the Wellcome Trust Functional Genomics Initiative CFG (Cardiovascular Functional Genomics) (066780) and BAIR (Biological Atlas of Insulin Resistance) (066786).

References

Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, Al-Majali KM, Trembling PM, Mann CJ, Shoulders CC, *et al*. (1999) Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genetics*, **21**, 76–83.

Backdahl L, Ribbhammar U and Lorentzen JC (2003) Mapping and functional characterization of rat chromosome 4 regions that regulate arthritis models and phenotypes in congenic strains. *Arthritis and Rheumatism*, **48**, 551–559.

Behmoaras J, Osborne–Pellegrin M, Gauguier D and Jacob MP (2005) Characteristics of the aortic elastic network and related phenotypes in seven inbred rat strains. *American Journal of Physiology Heart and Circulatory Physiology*, **288**, H769–H777.

Bergsteinsdottir K, Yang HT, Pettersson U and Holmdahl R (2000) Evidence for common autoimmune disease genes controlling onset, severity, and chronicity based on experimental models for multiple sclerosis and rheumatoid arthritis. *Journal of Immunology*, **164**, 1564–1568.

Bihoreau MT, Gauguier D, Kato N, Hyne G, Lindpaintner K, Rapp JP and Lathrop GM (1997a) A linkage map of the rat genome derived from three F2 crosses. *Genome Research*, **7**, 434–440.

Bihoreau MT, Ceccherini I, Browne J, Kränzlin B, Romeo G, Lathrop GM, James MR and Gretz N (1997b) Location of the first genetic locus, PKDr1, polycystic kidney disease in Han:SPRD cy/+ rat. *Human Molecular Genetics*, **6**, 609–613.

Bihoreau MT, Megel N, Brown JH, Kränzlin B, Crombez L, Tychinskaya Y, Broxholme J, Kratz S, Bergmann V, Hoffman S, *et al*. (2002) Characterisation of a major modifier locus for polycystic kidney disease (*Modpkdr1*) in the Han:SPRD(*cy*/+) rat in a region conserved with a mouse modifier locus for Alport syndrome. *Human Molecular Genetics*, **11**, 2165–2173.

Brown DM, Provoost AP, Daly MJ, Lander ES and Jacob HJ (1996) Renal disease susceptibility and hypertension are under independent genetic control in the fawn-hooded rat. *Nature Genetics*, **12**, 44–51.

Cicila GT, Rapp JP, Wang JM, St Lezin E, Ng SC and Kurtz TW (1993) Linkage of 11 beta-hydroxylase mutations with altered steroid biosynthesis and blood pressure in the Dahl rat. *Nature Genetics*, **3**, 346–353.

Cowley AW Jr, Roman RJ and Jacob HJ (2004) Application of chromosomal substitution techniques in gene-function discovery. *The Journal of Physiology*, **554**, 46–55.

Dahlman I, Lorentzen JC, de Graaf KL, Stefferl A, Linington C, Luthman H and Olsson T (1998) Quantitative trait loci disposing for both experimental arthritis and encephalomyelitis in the DA rat; impact on severity of myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis and antibody isotype pattern. *European Journal of Immunology*, **28**, 2188–2196.

D'Cruz PM, Yasumura D, Weir J, Matthes MT, Abderrahim H, LaVail MM and Vollrath D (2000) Mutation of the receptor tyrosine kinase gene *Mertk* in the retinal dystrophic RCS rat. *Human Molecular Genetics*, **9**, 645–651.

De Miglio MR, Pascale RM, Simile MM, Muroni MR, Virdis P, Kwong KM, Wong LK, Bosinco GM, Pulina FR, Calvisi DF, *et al*. (2004) Polygenic control of hepatocarcinogenesis in Copenhagen x F344 rats. *International Journal of Cancer*, **111**, 9–16.

Dobbins DE, Sood R, Hashiramoto A, Hansen CT, Widler RL and Remmers EF (2002) Mutation of macrophage stimulating factor (*Csf1*) causes osteopetrosis in the *tl* rat. *Biochemical and Biophysical Research Communications*, **294**, 1114–1120.

Dubay C, Vincent M, Samani NJ, Hilbert P, Kaiser MA, Beressi JP, Kotelevtsev Y, Beckmann JS, Soubrier F, Sassard J, *et al*. (1993) Genetic determinants of diastolic and pulse pressure map to different loci in Lyon hypertensive rats. *Nature Genetics*, **3**, 354–357.

Ellerman KE and Like AA (2000) Susceptibility to diabetes is widely distributed in normal class IIu haplotype rats. *Diabetologia*, **43**, 890–898.

Fakhrai-Rad H, Nikoshkov A, Kamel A, Fernstrom M, Zierath JR, Norgren S, Luthman H and Galli J (2000) Insulin-degrading enzyme identified as a candidate diabetes susceptibility gene in GK rats. *Human Molecular Genetics*, **9**, 2149–2158.

Fernandez-Teruel A, Escorihuela RM, Gray JA, Aguilar R, Gil L, Gimenez-Llort L, Tobena A, Bhomra A, Nicod A, Mott R, *et al*. (2002) A quantitative trait locus influencing anxiety in the laboratory rat. *Genome Research*, **12**, 618–626.

Frantz S, Clemitson J, Bihoreau MT, Gauguier D and Samani NJ (2001) Genetic dissection of region around the Sa gene on rat chromosome 1: evidence for multiple loci affecting blood pressure. *Hypertension*, **38**, 216–221.

Gal A, Li Y, Thompson DA, Weir J, Orth U, Jacobson SG, Apfelstedt-Sylla E and Vollrath D (2000) Mutations in MERTK, the human orthologue of the RCS rat retinal dystrophy gene, cause retinitis pigmentosa. *Nature Genetics*, **26**, 270–271.

Galli J, Li LS, Glaser A, Ostensson CG, Jiao H, Fakhrai-Rad H, Jacob HJ, Lander ES and Luthman H (1996) Genetic analysis of non insulin dependent diabetes mellitus in the GK rat. *Nature Genetics*, **12**, 31–37.

Garrett MR, Dene H and Rapp JP (2003) Time-course genetic analysis of albuminuria in Dahl salt-sensitive rats on low-salt diet. *Journal of the American Society of Nephrology*, **14**, 1175–1187.

Garrett MR, Dene H, Walder R, Zhang QY, Cicila GT, Assadnia S, Deng AY and Rapp JP (1998) Genome scan and congenic strains for blood pressure QTL using Dahl salt-sensitive rats. *Genome Research*, **8**, 711–723.

Garrett MR and Rapp JP (2003) Defining the blood pressure QTL on chromosome 7 in Dahl rats by a 177-kb congenic segment containing Cyp11b1. *Mammalian Genome*, **14**, 268–273.

Garrett MR, Zhang X, Dukhanina OI, Deng AY and Rapp JP (2001) Two linked blood pressure quantitative trait loci on chromosome 10 defined by Dahl rat congenic strains. *Hypertension*, **38**, 779–785.

Gauguier D, Behmoaras J, Argoud K, Wilder SP, Pradines C, Bihoreau MT, Osborne-Pellegrin M and Jacob MP (2005) Chromosomal mapping of quantitative trait loci controlling aortic elastin content in a cross derived from Brown Norway and LOU rats. *Hypertension*, **45**, 460–466.

Gauguier D, Froguel P, Parent V, Bernard C, Bihoreau MT, Portha B, Pénicaud L, Lathrop M and Ktorza A (1996) Chromosomal mapping of genetic loci associated with non insulin dependent diabetes in the GK rat. *Nature Genetics*, **12**, 38–43.

Gauguier D, van Luijtelaar G, Bihoreau MT, Wilder SP, Godfrey R, Vossen J, Coenen A and Cox RD (2004) Chromosomal mapping of genetic loci controlling absence epilepsy phenotypes in the WAG/Rij Rat. *Epilepsia*, **45**, 908–915.

Gigante B, Rubattu S, Stanzione R, Lombardi A, Baldi A, Baldi F and Volpe M (2003) Contribution of genetic factors to renal lesions in the stroke-prone spontaneously hypertensive rat. *Hypertension*, **42**, 702–706.

Gotoda T, Iizuka Y, Kato N, Osuga J, Bihoreau MT, Murakami T, Yamori Y, Shimano H, Ishibashi S and Yamada N (1999) Absence of Cd36 mutation in the original spontaneously hypertensive rats with insulin resistance. *Nature Genetics*, **22**, 226–228.

Griffiths MM and Remmers EF (2001) Genetic analysis of collagen-induced arthritis in rats: a polygenic model for rheumatoid arthritis predicts a common framework of cross-species inflammatory/autoimmune disease loci. *Immunological Reviews*, **184**, 172–183.

Haag JD, Shepel LA, Kolman BD, Monson DM, Benton ME, Watts KT, Waller JL, Lopez-Guajardo CC, Samuelson DJ and Gould MN (2003) Congenic rats reveal three independent Copenhagen alleles within the Mcs1 quantitative trait locus that confer resistance to mammary cancer. *Cancer Research*, **63**, 5808–5812.

Hansen C and Spuhler K (1984) Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcoholism, Clinical and Experimental Research*, **8**, 477–479.

Harris EL, Stoll M, Jones GT, Granados MA, Porteous WK, Van Rij AM and Jacob HJ (2001) Identification of two susceptibility loci for vascular fragility in the Brown Norway rat. *Physiological Genomics*, **28**, 183–189.

Hilbert P, Lindpaintner K, Beckmann JS, Serikawa T, Soubrier F, Dubay C, Cartwright P, De Gouyon B, Julier C, Takahasi S, *et al*. (1991) Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. *Nature*, **353**, 521–529.

Hornum L, Romer J and Markholst H (2002) The diabetes-prone BB rat carries a frameshift mutation in Ian4, a positional candidate of Iddm1. *Diabetes*, **51**, 1972–1979.

Jacob HJ and Kwitek AE (2001) Rat genetics: attaching physiology and pharmacology to the genome. *Nature Reviews Genetics*, **3**, 33–42.

Jacob HJ, Lindpaintner K, Lincoln SE, Kusumi K, Bunker RK, Mao YP, Ganten D, Dzau VJ and Lander ES (1991) Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. *Cell*, **67**, 213–224.

Jacob HJ, Pettersson A, Wilson D, Mao Y, Lernmark A and Lander ES (1992) Genetic dissection of autoimmune type I diabetes in the BB rat. *Nature Genetics*, **2**, 56–60.

Jagodic M, Becanovic K, Sheng JR, Wu X, Backdahl L, Lorentzen JC, Wallstrom E and Olsson T (2004) An advanced intercross line resolves Eae18 into two narrow quantitative trait loci syntenic to multiple sclerosis candidate loci. *Journal of Immunology*, **173**, 1366–1373.

Jeffs B, Clark JS, Anderson NH, Gratton J, Brosnan MJ, Gauguier D, Reid JL, Macrae IM and Dominiczak AF (1997) Sensitivity to cerebral ischaemic insult in a rat model of stroke is determined by a single genetic locus. *Nature Genetics*, **16**, 364–367.

Kaisaki PJ, Delepine M, Woon PY, Sebag-Montefiore L, Wilder S, Menzel S, Vionnet N, Marion E, Riveline JP, Charpentier X, *et al*. (2004) Polymorphisms in type-II SH2-domain-containing inositol 5-phosphatase (*INPPL1*, SHIP2) are associated with physiological abnormalities of the metabolic syndrome. *Diabetes*, **53**, 1900–1904.

Kanemoto N, Hishigaki H, Miyakiya A, Oga K, Okuno S, Tsuji A, Tagaki T, Takahashi E, Nakamura Y and Watanabe TK (1998) Genetic dissection of "OLETF", a rat model for non-insulin-dependent diabetes mellitus. *Mammalian Genome*, **9**, 419–425.

Lan H, Kendziorski CM, Haag JD, Shepel LA, Newton MA and Gould MN (2001) Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. *Genetics*, **157**, 331–339.

MacMurray AJ, Moralejo DH, Kwitek AE, Rutledge EA, Van Yserloo B, Gohlke P, Speros SJ, Snyder B, Schaefer J, Bieg S, *et al*. (2002) Lymphopenia in the BB rat model of type 1 diabetes is due to a mutation in a novel immune-associated nucleotide (Ian)-related gene. *Genome Research*, **12**, 1029–1039.

Marion E, Kaisaki PJ, Pouillon V, Gueydan C, Levy J, Bodson A, Krzentowski G, Daubresse JC, Mockel J, Behrends J, *et al*. (2002) The gene INPPL1, encoding the lipid phosphatase SHIP2, is a candidate for type 2 diabetes in rat and man. *Diabetes*, **51**, 2012–2017.

Mas M, Cavaillès P, Colacios C, Subra JF, Lagrange D, Calise M, Christen MO, Druet P, Pelletier L, Gauguier D, *et al*. (2004) Genetic control by two loci on chromosomes 9 and 10, of the Th2-immunopathological disorders triggered by aurothiopropanol sulfonate in the BN rat. *Journal of Immunology*, **172**, 6354–6361.

Mas M, Subra JF, Lagrange D, Pilipenko-Appolinaire S, Gauguier D, Druet P and Fournié GJ (2000) Rat chromosome 9 bears a major susceptibility locus for IgE response. *European Journal of Immunology*, **30**, 1698–1705.

Masuyama T, Fuse M, Yokoi N, Shinohara M, Tsujii H, Kanazawa M, Kanazawa Y, Komeda K and Taniguchi K (2003) Genetic analysis for diabetes in a new rat model of nonobese type 2 diabetes, Spontaneously Diabetic Torii rat. *Biochemical and Biophysical Research Communications*, **304**, 196–206.

Matsuo T, Osumi-Yamashita N, Noji S, Ohuchi H, Koyama E, Myokai F, Matsuo N, Taniguchi S, Doi H, Iseki S, *et al*. (1993) A mutation in the Pax-6 gene in rat small eye is associated with impaired migration of midbrain crest cells. *Nature Genetics*, **3**, 299–304.

Moisan MP, Courvoisier H, Bihoreau MT, Gauguier D, Hendley ED, Lathrop GM, James MR and Mormede P (1996) A major quantitative trait locus influences hyperactivity in the rat. *Nature Genetics*, **14**, 471–473.

Mott R and Flint J (2002) Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. *Genetics*, **160**, 1609–1618.

Olofsson P, Lu S, Holmberg J, Song T, Wernhoff P, Pettersson U and Holmdahl R (2003a) A comparative genetic analysis between collagen-induced arthritis and pristane-induced arthritis. *Arthritis and Rheumatism*, **48**, 2332–2342.

Olofsson P, Holmberg J, Tordsson J, Lu S, Akerstrom B and Holmdahl R (2003b) Positional identification of Ncf1 as a gene that regulates arthritis severity in rats. *Nature Genetics*, **33**, 25–32.

Phillips MS, Liu Q, Hammond HA, Dugan V, Hey PJ, Caskey CJ and Hess JF (1996) Leptin receptor missense mutation in the fatty Zucker rat. *Nature Genetics*, **13**, 18–19.

Pinto D, Westland B, de Haan GJ, Rudolf G, Martins da Silva B, Hirsch E, Lindhout D, Kasteleijn-Nolst Trenite DGA and Koeleman BPC (2005) Genome-wide linkage scan of epilepsy-related photoparoxysmal electroencephalographic response: evidence for linkage on chromosomes 7q32 and 16p13. *Human Molecular Genetics*, **14**, 171–178.

Pravenec M, Gauguier D, Schott JJ, Buard J, Kren V, Bila V, Szpirer C, Szpirer J, Wang JM, Huang H, *et al*. (1995) Mapping of quantitative trait loci for blood pressure and cardiac mass in the rat by genome scanning of recombinant inbred strains. *The Journal of Clinical Investigation*, **96**, 1973–1978.

Ramanathan S, Bihoreau MT, Patterson A, Marandi L, Gauguier D and Poussier P (2002) Thymectomy and radiation induced type 1 diabetes in non-lymphopenic BB rats. *Diabetes*, **51**, 2975–2981.

Ramanathan S and Poussier P (2001) BB rat lyp mutation and Type 1 diabetes. *Immunological Reviews*, **184**, 161–171.

Rapp JP (2000) Genetic analysis of inherited hypertension in the rat. *Physiological Reviews*, **80**, 135–172.

Rapp JP, Garrett MR and Deng AY (1998) Construction of a double congenic strain to prove an epistatic interaction on blood pressure between rat chromosomes 2 and 10. *The Journal of Clinical Investigation*, **101**, 1591–1595.

Rapp JP, Wang SM and Dene H (1989) A genetic polymorphism in the renin gene of Dahl rats cosegregates with blood pressure. *Science*, **243**, 542–544.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Rogner UC and Avner P (2003) Congenic mice: cutting tools for complex immune disorders. *Nature Reviews Immunology*, **3**, 243–252.

Rubattu S, Volpe M, Kreutz R, Ganten U, Ganten D and Lindpaintner K (1996) Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. *Nature Genetics*, **13**, 429–434.

Rudolf G, Bihoreau MT, Godfrey R, Wilder SP, Cox RD, Lathrop M, Marescaux C and Gauguier D (2004) Polygenic control of idiopathic generalized epilepsy phenotypes in the genetic absences rats from Strasbourg (GAERS). *Epilepsia*, **45**, 301–308.

Saad Y, Garrett MR and Rapp J (2001) Multiple blood pressure QTL on rat chromosome 1 defined by Dahl rat congenic strains. *Physiological Genomics*, **4**, 201–214.

Sebkhi A, Zhao L, Lu L, Haley CS, Nunez DJ and Wilkins MR (1999) Genetic determination of cardiac mass in normotensive rats: results from an F344xWKY cross. *Hypertension*, **33**, 949–953.

Steen RG, Kwitek-Black AE, Glenn C, Gullings-Handley J, Van Etten W, Atkinson OS, Appel D, Twigger S, Muir M, Mull T, *et al*. (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Research*, **9**, 1–8.

Stoll M, Cowley AW Jr, Tonellato PJ, Greene AS, Kaldunski ML, Roman RJ, Dumas P, Schork NJ, Wang Z and Jacob HJ (2001) A genomic-systems biology map for cardiovascular function. *Science*, **294**, 1723–1726.

Stoll M, Kwitek-Black AE, Cowley AW Jr, Harris EL, Harrap SB, Krieger JE, Printz MP, Provoost AP, Sassard J and Jacob HJ (2000) New target regions for human hypertension via comparative genomics. *Genome Research*, **10**, 473–482.

Wallace KJ, Wallis RH, Collins SC, Argoud K, Kaisaki PJ, Ktorza A, Bihoreau MT and Gauguier D (2004) Genetic dissection of a diabetes QTL in congenic lines of the Goto Kakizaki rat identifies a chromosomal region conserved with diabetes susceptibility loci in human 1q. *Physiological Genomics*, **19**, 1–10.

Wallis RH, Wallace KJ, Collins SC, McAteer M, Argoud K, Bihoreau MT, Kaisaki PJ and Gauguier D (2004) Enhanced insulin secretion and cholesterol metabolism in congenic strains of the spontaneously diabetic (type 2) Goto Kakizaki rat are controlled by independent genetic loci in rat chromosome 8. *Diabetologia*, **47**, 1096–1106.

Ward CJ, Hogan MC, Rossetti S, Walker D, Sneddon T, Wang X, Kubly V, Cunningham JM, Bacallao R, Ishibashi M, *et al*. (2002) The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nature Genetics*, **30**, 259–269.

Watanabe TK, Bihoreau MT, McCarthy LC, Kiguwa SL, Hishigaki H, Tsuji A, Browne J, Yamasaki Y, Mizoguchi-Miyakita A, Oga K, Hudson MR Jr, *et al*. (1999) A map of the rat genome containing 5,203 markers: 4,700 microsatellites and 605 genes in a rat, mouse and human comparative map. *Nature Genetics*, **2**, 27–36.

Wendell DL, Daun SB, Stratton MB and Gorski J (2000) Different functions of QTL for estrogen-dependent tumor growth of the rat pituitary. *Mammalian Genome*, **11**, 855–861.

Wilder SP, Bihoreau MT, Argoud K, Watanabe T, Lathrop M and Gauguier D (2004) Integration of the rat recombination and EST maps in the rat genomic sequence and comparative mapping analysis with the mouse genome. *Genome Research*, **14**, 758–765.

Wu J, Forbes JR, Chen HS and Cox DW (1994) The LEC rat has a deletion in the copper transporting ATPase gene homologous to the Wilson disease gene. *Nature Genetics*, **7**, 541–545.

Yokoi N, Komeda K, Wang HY, Yano H, Kitada K, Saitoh Y, Seino Y, Yasuda K, Serikawa T and Seino S (2002) Cblb is a major susceptibility gene for rat type 1 diabetes mellitus. *Nature Genetics*, **31**, 391–394.

Zan Y, Haag JD, Chen KS, Shepel LA, Wigington D, Wang YR, Hu R, Lopez-Guajardo CC, Brose HL, Porter KI, *et al*. (2003) Production of knockout rats using ENU mutagenesis and a yeast-based screening assay. *Nature Biotechnology*, **21**, 645–651.

Zimdahl H, Nyakatura G, Brandt P, Schulz H, Hummel O, Fartmann B, Brett D, Droege M, Monti J, Lee YA, *et al*. (2004) A SNP map of the rat genome generated from cDNA sequences. *Science*, **303**, 807.

Website references

http://www.ensembl.org/Rattus_norvegicus, (2005) *Rat Genome Annotation*, European Bioinformatics Institute: Hinxton.

http://www.hgsc.bcm.tmc.edu/projects/rat, (2005) *Rat Genome Project*, Baylor College of Medicine.

http://www.ratmap.gen.gu.se, (2005) *Rat Genome Database RatMap*, Göteborg University: Sweden.

http://www.well.ox.ac.uk/rat_mapping_resources, (2005) *Rat Genetic Data Repository, Wellcome Trust Centre for Human Genetics*, University of Oxford.

http://rgd.mcw.edu/, (2005) *Rat Genome Database*, Medical College of Wisconsin.

http://ratest.uiowa.edu/; (2005) *Rat EST project*; University of Iowa.

http://genome.ucsc.edu/, (2005) *UCSC Genome Browser; Rat Genome bioinformatics*, UC Santa Cruz.

Specialist Review

Farm animals

Leif Andersson
Uppsala University, Uppsala, Sweden

1. Introduction

Genome research in farm animals is justified because it may lead to important practical applications in the animal industry. A number of diagnostic DNA tests that are widely used have already been developed. Most of these concern monogenic traits or disorders but there are also several examples where a Quantitative Trait Loci (QTL) affecting a multifactorial trait has been exploited in the animal industry. Most of the QTL applications have been based on the principle of Marker Assisted Selection (MAS) where a marker bracket is used to alter the frequency of QTL alleles for which the molecular nature are unknown. There are some cases like *DGAT* controlling milk yield in cattle (Grisart *et al*., 2002; Grisart *et al*., 2004; Winter *et al*., 2002) and *IGF2* affecting muscle mass in pigs (Braunschweig *et al*., 2004; Van Laere *et al*., 2003) where the underlying mutation for a QTL has been revealed and a direct diagnostic test can be applied. However, practical animal breeding still relies primarily on traditional phenotypic selection despite the significant progress in animal genomics but this may change in the near future due to drastically reduced costs for high-throughput analysis of genetic markers.

Genome research in farm animals will contribute to our understanding of gene and genome evolution and will provide new basic knowledge, in particular, concerning the molecular basis for phenotypic variation. Historically, research on domestic animals has contributed considerably to basic biology. Charles Darwin used domestic animals as a proof-of-principle for his theory on phenotypic evolution caused by natural selection (Darwin, 1859). Furthermore, the inheritance of coat color in domestic animals was among the first traits to be used for genetic studies soon after Mendel's laws of heredity were rediscovered in the beginning of the twentieth century (Bateson, 1902; Spillman, 1906). Farm animals are not particularly well suited for studying simple monogenic disorders. Such deleterious mutations are rare in domestic animals since they tend to be eliminated from the breeding population. The best organisms for studying the phenotypic consequences of deleterious mutations with large phenotypic effects are humans, since such mutations often lead to clinical disorders, or experimental animals, since systematic screenings for induced deleterious mutations can be performed. However, for those mutations where one needs to compare hundreds of individuals with different genotypes to reveal a phenotypic effect, neither human materials nor

mutation screenings in model organisms are ideal. Most phenotypic variations in all species are due to mutations with mild phenotypic effects. Here farm animals provide unique opportunities because these species have been genetically modified by selective breeding, a process that has involved millions of individuals for thousands of years. In fact, no experimental organism has been genetically modified to the same extent as farm animals. Thus, farm animal populations harbor rich collections of nonmorbid mutations affecting phenotypic characters. Domestication can be viewed as an adaptation to a new environment. We now have the tools to unravel the molecular basis for this process. This research can shed light on many of the unanswered questions concerning the basis for phenotypic evolution. For instance, to which extent does loss-of-function mutations contribute to phenotypic evolution (Olson, 1999)? How important are regulatory mutations (King and Wilson, 1975) and epistatic interactions (Carlborg and Haley, 2004)? Does epigenetic inheritance play a significant role (Bjornsson *et al*., 2004)?

Genome research in farm animals can also make important contributions to human medicine, in particular, as regards metabolism, immune response, susceptibility to infectious disease, and reproductive traits. Metabolic traits are often altered in domestic animals since a common breeding goal is to promote the allocation of energy and nutrients into animal products such as meat, milk, and eggs. A good example is fat deposition in pigs. Before 1900, there was a strong preference for fat pigs because a high energy content in animal products was desired. However, during the last 50 years, there has been a stronger and stronger consumer demand for pig meat with low fat content, and the fat deposition has successively been reduced. In a cross between the wild boar and domestic pigs, it became apparent that the wild boar carried QTL alleles favoring fat deposition (Andersson *et al*., 1994). A high fat deposition is adaptive in the wild boar since it must use stored fat to survive periods of starvation. This situation resembles very much the "thrifty gene" hypothesis, which implies that a major reason for the problem with obesity and metabolic disorders in humans is that alleles favoring fat deposition had a selective advantage during periods in the past with inappropriate food supplies (Neel, 1962). Thus, the genes that once were associated with a selective advantage are today causing metabolic disorders. Genetic analysis of wild boar/domestic pig intercrosses can shed light on the genetic regulation of fat deposition and may reveal novel strategies for treatment of obesity and metabolic disorders in humans. Similarly, the intramuscular fat content has often been altered in meat-producing animals and fat content in skeletal muscle is of considerable interest in relation to the development of insulin resistance and Type II diabetes (Friedman, 2004). Disease resistance and immune response are other traits that have been altered during domestication as part of the adaptation to the farm environment with its higher density of animals and higher exposure to pathogens. There are many local breeds, in particular in tropical countries that have evolved resistance or tolerance to certain pathogens. Such breeds provide a largely unexploited resource for studies of the genetic basis for host–pathogen interactions. Finally, reproductive traits, such as age of sexual maturation, litter size, and seasonality, have been extensively modified in farm animals and genetic studies may shed light on the genetic regulation of reproduction. However, both disease resistance and reproductive traits are often difficult to study genetically due to a strong environmental influence on these traits.

2. Evolutionary history

All farm animals have, from an evolutionary point of view, a young history as domestications occurred within the last ∼10 000 years. Most of the neutral sequence polymorphisms we observe in these species have an origin well before domestication as can be shown by estimating the frequency of nucleotide substitutions that has accumulated since domestication. The neutral substitution rates in human and mouse have been estimated at 2.2×10^{-9} and 4.5×10^{-9} per site per year respectively (Mouse Genome Sequencing Consortium, 2002). Using the mean of these two figures, we can estimate that the average frequency of neutral nucleotide substitutions that has occurred between two haplotypes that diverged 10 000 years ago should be $\sim 6.5 \times 10^{-5}$ per site, that is, 6–7 single nucleotide polymorphisms (SNPs) per 100 kb. The average pairwise SNP rate within and between breeds of domestic chicken has recently been estimated at 5–6 SNPs/kb (International Chicken Polymorphism Map Consortium, 2004), and we expect that the corresponding figure for other domestic animals will be in the range 1–5 SNPs/kb. Thus, only a few percent of the neutral nucleotide substitutions we observe between two random chromosomes are expected to have occurred subsequent to domestication. Consequently, if one makes sequence comparisons between a wild ancestor and the corresponding domestic animals, for those species where the wild ancestor has not been extinct (e.g., red jungle fowl and wild boar), one finds that few sequence variants are unique for domestic populations. The exceptions are of course those mutations that have been under strong positive selection, or neutral sites in the near vicinity of selected sites, where a mutation may be common in the domestic animal but rare or completely absent in the wild ancestor. It is still an open question to which extent the response to phenotypic selection in domestic animals is due to mutations that were present in the wild ancestor before domestication or mutations that have occurred subsequent to domestication. The latter explanation is probably important, at least for mutations with major phenotypic effects. Those would often be associated with a selective disadvantage in the wild population.

Accumulating evidence shows that several farm animals originate from distinct subpopulations of the wild ancestor that diverged from each other long before domestication (Bruford *et al*., 2003). This is the case for cattle, which originates from two subspecies, denoted *Bos taurus* and *Bos indicus*, corresponding to cattle with or without hump and with a European and an Asian origin, respectively (Bruford *et al*., 2003). Similarly, pig domestication involved wild boar populations from both Europe and Asia (Giuffra *et al*., 2000), the clear separation of the European and Asian ancestor is evident from an average sequence divergence of 1.2% for the entire mitochondrial DNA (Kijas and Andersson, 2001). Therefore, modern breeds of farm animals may have a hybrid origin. This is well documented for African cattle (Freeman *et al*., 2004) and for several European breeds of domestic pigs (Giuffra *et al*., 2000). As an example, the characterization of the *IGF2* locus in pigs revealed that European domestic pigs carried *IGF2* haplotypes originating from both a European and an Asian wild ancestor and the two haplotype forms showed a sequence divergence of about 1% (1 SNP/111 bp; Van Laere *et al*., 2003).

3. Genomic resources

Basic resources for genome research such as large sets of genetic markers (microsatellites and SNPs), medium dense linkage maps, radiation hybrid panels, and BAC libraries have been established for all farm animals. A compilation of databases and websites providing information of these resources is given in Table 1. It should be noted that goat, sheep, and cattle are all ruminants and fairly closely related. Goat and sheep diverged about 5 million years before present and their common ancestor diverged from the cattle lineage about 20 million years before present. This means that genome projects in sheep and goat can take advantage of genome resources and information developed for cattle, the most well studied of these three species.

The ultimate genetic map is of course the complete genome sequence of the target species. A high-quality draft genome sequence at 6.6X coverage has recently been released for chicken as the first domestic animal and the first bird to have its genome sequenced (International Chicken Genome Sequencing Consortium, 2004). The data were generated by sequencing a red jungle fowl female from a partially inbred line. A comprehensive search for sequence polymorphism was carried out in parallel by shot-gun sequencing of three domestic birds, a layer, a broiler,

Table 1 Databases and resources for farm animal genomics

Database/Resource	Internet address	Comment
AvianNET	**http://www.chicken-genome.org/**	Chicken genome project
ARKdb	**http://www.thearkdb.org/**	Genome mapping data, all species
Bovine genome sequencing	**http://www.hgsc.bcm.tmc.edu/projects/bovine/**	
Cattle genome database, Australia	**http://www.cgd.csiro.au/**	
Chicken genome sequencing	**http://genome.wustl.edu/projects/chicken/**	
Dairy cattle QTL database	**http://www.vetsci.usyd.edu.au/reprogen/QTL_Map/**	Compilation of published QTL studies
Horse genome project	**http://www.uky.edu/Ag/Horsemap/**	
Goat database	**http://locus.jouy.inra.fr/cgi-bin/lgbc/mapping/common/intro2.pl?BASE=goat**	
Online Mendelian Inheritance in animals	**http://www.angis.org.au/Databases/BIRX/omia/**	The animal version of OMIM
TIGR gene indices	**http://www.tigr.org/tdb/tgi/**	EST data from several farm animals
U.S. Livestock Genome Mapping Projects	**http://www.genome.iastate.edu/**	Cattle, chicken, pig, horse, sheep
Wageningen University	**http://www.zod.wau.nl/vf/**	Chicken and pig genomics

Table 2 Animal resources for genome research in farm animals

Type of population	Traits[a]	Cost	Population size
Commercial	Standard	+	+++
Commercial	Special	++ − +++	++
Experimental herds	Special	++	+[b]
Experimental crosses	Special	+++	+[b]

[a]Standard traits refer to those traits registered for breeding purposes at no extra cost for the research project.
[b]Large experimental populations (thousands of individuals) may be used for chicken and, in some cases, also for pigs.

and a Silkie, each to 0.25X coverage (International Chicken Polymorphism Map Consortium, 2004). This revealed as many as 2.8 million SNPs for the chicken. A preliminary draft assembly at 3X coverage for cattle was released during 2004 and a high-quality draft genome sequence will be available by 2005 (**http://www.hgsc.bcm.tmc.edu/projects/bovine/**). A genome sequence with ∼1X coverage has been generated for the pig but the data are not yet publicly available (M. Fredholm, personal communication). No initiatives have yet been taken to sequence the genomes of goat, sheep, or horse.

4. Strategies for mapping trait loci

There are two routes for mapping trait loci in farm animals, namely, by using existing pedigrees from commercial populations or by generating experimental pedigrees (Table 2). Genome research in cattle and horse are almost exclusively based on existing pedigrees due to the high cost for making experimental pedigrees that are sufficiently large for gene mapping experiments. The chicken is the other extreme where almost all mapping efforts are carried out with experimental populations. The use of commercial populations can be very cost-effective since it is possible to collect large families with available phenotypic data at no other cost than those associated with sample collection. The phenotypic data are then limited to the traits recorded for breeding purposes such as milk quantity and milk composition in cattle or growth and meat content in pigs. More detailed phenotypic characterization can to some extent be added for research purposes. The collection of very large multigeneration families facilitates powerful statistical analysis. It is also possible to utilize the breeding values calculated on the basis of progeny testing that provide very accurate measures of the genotype of sires based on phenotypic data from a large number of progeny. This has been widely used for QTL mapping in dairy cattle based on the granddaughter design where the segregation of breeding values from grandsires to sons is evaluated (Georges *et al*., 1995; Weller *et al*., 1990).

The use of experimental populations has the advantage that environmental variation is better controlled and much more detailed phenotypic recordings can be made, but the maintenance of experimental populations are costly. Cross-breeding experiments are even more costly, but very powerful. Linkage analysis is then

facilitated by the high heterozygosity in the F_1 generation and the fact that the linkage phase between alleles at marker and trait loci is often the same in all F_1 animals. However, this is not always true since the founder populations are usually outbred, which means that some segregation at QTLs may occur within lines. Cross-breeding experiments are the only way to map trait loci that have been under very strong selection and therefore are fixed within lines.

An important difference between human genetics and animal genetics is that there is less genetic heterogeneity at trait loci in farm animals. The paradigm in human genetics is that there are many different mutations causing the same inherited disorder in different families. The reason for this is of course that there are many ways to knock out gene function and with the huge current effective population size in humans, many new deleterious mutations are generated each generation. In farm animals, it is much more common that one finds a single or a few causative mutations at trait loci. This is particularly true for those mutations that have been under strong phenotypic selection since they are spread rapidly across the population. The presence of a widespread causative mutation tracing back to a common ancestor facilitates the use of identical-by-descent (IBD) mapping to identify a common shared haplotype harboring the causative mutation (Andersson and Georges, 2004). The identification of a Quantitative Trait Nucleotide (QTN) at the *IGF2* locus underlying a major QTL in pigs is an excellent example where an IBD approach was used to map the causative mutation to a ∼15-kb interval (Van Laere *et al*., 2003).

5. Monogenic trait loci

The identification of genes underlying monogenic trait loci in farm animals is today often straightforward if a sufficient pedigree and/or population material is available for linkage or association analysis, respectively. In fact, the optimal design for mapping trait loci is to combine linkage and IBD mapping to achieve a high-resolution localization. A list of interesting trait loci for which the underlying causative mutation has been identified is given in Table 3.

The identification of causative genes and mutations are done by positional candidate cloning or by classical positional cloning. Positional candidate cloning is becoming more and more powerful with the continuous improvement of the functional annotation of vertebrate genes. This strategy implies that the trait locus is first mapped to a specific chromosomal region, then gene annotation data across species are extracted, a screen for mutations is done for the best candidate genes, and finally the candidate mutations are evaluated by genetic and functional analysis. For all species, except chicken, this approach must still involve a comparative mapping effort in order to take advantage of the near-complete genome sequences available for other species. This is usually straightforward since there is a high degree of conserved synteny and conserved gene order among closely related species like two mammalian species. However, it is not uncommon that linkage mapping assigns a locus to a region that harbors no obvious candidate gene because the function of the causative gene is still poorly understood. If so, a high-resolution mapping is required to reduce the critical

Table 3 Some particularly interesting trait loci in farm animals for which the causative mutation has been identified

Species	Trait	Gene	References
Cattle	Muscle hypertrophy	*MST*	Grobet *et al*. (1998)
	Fish odor in milk	*FMO3*	Lundén *et al*. (2003)
	Milk yield and fat content (QTL)	*DGAT*	Grisart *et al*. (2002), Winter *et al*. (2002)
	Milk yield and composition (QTL)	*GHR*	Blott *et al*. (2002)
Chicken	Plumage color and QTL for behavior	*PMEL17*	Keeling *et al*. (2004), Kerje *et al*. (2004)
Goat	Lack of horns, intersexuality	Noncoding region[a]	Pailhoux *et al*. (2001)
Horse	White color, megacolon	*EDNRB*	Metallinos *et al*. (1998), Santschi *et al*. (1998), Yang *et al*. (1998)
Pig	Malignant hyperthermia	*RYR1*	Fujii *et al*. (1991)
	Dominant white color, haematopoiesis	*KIT*	Giuffra *et al*. (2002), Marklund *et al*. (1998)
	Hypercholesterolaemia	*LDLR*	Hasler-Rapacz *et al*. (1998)
	Intestinal *E. coli* adherence	*FUT1*	Meijerink *et al*. (2000)
	Muscle glycogen content (QTL)	*PRKAG3*	Ciobanu *et al*. (2001), Milan *et al*. (2000)
	Muscle growth, size of heart, fat deposition (QTL)	*IGF2*	Van Laere *et al*. (2003)
Sheep	Fertility, ovulation rate	*BMP15*	Galloway *et al*. (2000)
		BMPR1B	Mulsant *et al*. (2001)
	Muscle hypertrophy	Noncoding region[a]	Charlier *et al*. (2002), Freking *et al*. (2002)

[a]These are apparently mutations in cis-regulatory element that influence the expression of one or more genes.

interval as much as possible and an IBD approach may be required to reduce the interval to a region significantly smaller than 1 Mb. This requires access to both sufficient animal material and high-density marker maps. The current efforts to generate complete genome sequences and large collections of polymorphisms, as recently accomplished for the chicken (International Chicken Genome Sequencing Consortium, 2004; International Chicken Polymorphism Map Consortium, 2004), will greatly facilitate the molecular characterization of trait loci in farm animals.

6. Quantitative trait loci (QTL)

A QTL is defined as a chromosomal region harboring one or several mutations affecting a multifactorial trait (Lynch and Walsh, 1998). The first genome scans for QTL detection in farm animals were carried out 10 years ago and involved growth and fatness traits in a wild boar/domestic pig intercross (Andersson *et al*., 1994) and milk production traits in dairy cattle (Georges *et al*., 1995). Since then, numerous QTL studies in farm animals have been reported. The statistical methodology for QTL analysis using different experimental designs is well established (Hoeschele, 2003; Jansen, 2003; Lynch and Walsh, 1998). User-friendly, Web-based software for QTL analysis is also available (Seaton *et al*., 2002). Perhaps the most important

component in the design of a QTL study is the size of the experiment since a material with too few progeny will lead to a low statistical power. The consequence of an underdimensioned experiment is that no QTLs are detected or that the estimated effects of the detected QTLs are inflated (Goring *et al*., 2001; Mackinnon and Georges, 1992); see Lynch and Walsh (1998) for recommendations on the required sample size for QTL detection. A general recommendation is that the larger sample size, the better, as a large sample size will allow a more sophisticated and powerful statistical analysis including the detection of genetic heterogeneity and epistatic interaction.

However, the major challenge is not to find QTLs but to unravel the underlying gene(s) and mutation(s) (Andersson and Georges, 2004). Gene identification is difficult for several reasons. First, the precision in QTL mapping is poor because the phenotype is determined by multiple QTLs in combination with environmental factors. Thus, there is no simple one-to-one relationship between a QTL and a phenotype. Second, a QTL with large effects may turn out to be caused by several linked QTLs each with a small effect. Third, most QTLs have mild phenotypic effects, which make it more difficult to spot a candidate mutation compared with mutations underlying monogenic disorders that often disrupt protein function or drastically reduce gene expression. Furthermore, QTLs may often be due to regulatory mutations that are much more difficult to find and functionally characterize compared with mutations in coding sequence.

The poor resolution in QTL mapping can be overcome by progeny testing and marker segregation analysis, which makes it possible to deduce the QTL status of the parental haplotypes with great confidence. This leads to a collection of haplotypes with known QTL alleles that in turn can be used for high-resolution mapping and sequencing. This may eventually lead to the causative mutation, at least in those favorable situations when the QTL is caused by a single mutation. This can be achieved by QTL analysis of extended pedigrees in commercial populations (Blott *et al*., 2002; Grisart *et al*., 2002; Winter *et al*., 2002). Available data also suggest that the same favorable QTL allele may be found in different breeds since there is often some gene flow between breeds (Van Laere *et al*., 2003), which should facilitate high-resolution QTL mapping. There is, therefore, a considerable interest to develop statistical methodology that can combine linkage and linkage disequilibrium mapping (Farnir *et al*., 2002; Meuwissen and Goddard, 2000; Meuwissen and Goddard, 2001; Perez-Enciso, 2003).

QTL mapping in F_2 intercrosses also suffer from a poor resolution in QTL assignments. This can be remediated by selective back-crossing or by the generation of advanced intercross lines (AIL) (Darvasi, 1998; Darvasi and Soller, 1995). Selective back-crossing first involves the identification of putative breeding sires that carry informative recombinant haplotypes. Second, progeny testings are done to determine the QTL status for the recombinant haplotypes (Marklund *et al*., 1999). The maintenance of an AIL allows the accumulation of recombinants that break up the strong linkage disequilibrium generated by cross-breeding. An AIL has the merit compared with selective back-crossing that it generates a material suitable for high-resolution QTL mapping for all the QTLs detected in the intercross.

Like in all other species, there are few QTLs in farm animals for which the underlying causative mutation have been identified. There are a couple of

examples where a mutation determining a monogenic trait also acts as a QTL for a multifactorial trait (Table 3). A missense mutation in the gene for the ryanodine receptor (*RYR1*) predisposes to malignant hyperthermia but also affects lean meat content in the pig (Fujii *et al*., 1991). Similarly, a missense mutation in the gene for the muscle-specific isoform of the regulatory $\gamma 3$ chain of AMP-activated protein kinase (*PRKAG3*) causes excess glycogen content in skeletal muscle and has an effect on lean meat content in the pig (Milan *et al*., 2000). Both these mutations have increased in frequency as a consequence of the strong selection for lean pigs, but diagnostic DNA tests have now been used to reduce their frequency in order to avoid the negative pleiotropic effects on other traits. A gene duplication and a splice mutation in the porcine *KIT* gene causes dominant white color and have pleiotropic effects on hematopoiesis (Marklund *et al*., 1998). More recently, it has been shown that the *PMEL17* locus in chicken both determines dominant white color and influences the risk of becoming a victim for feather pecking in the F_2 generation of a red jungle fowl/White Leghorn intercross (Keeling *et al*., 2004; Kerje *et al*., 2004).

Three mutations underlying QTLs in cattle and pig have been identified by a positional candidate cloning approach; in all three cases, an IBD approach facilitated the identification of the causative mutation. Missense mutations were detected in the genes for diacylglycerol transferase (*DGAT*) and the growth hormone receptor (*GHR*) as causing two major milk-production QTLs on chromosome 14 and 20, respectively (Blott *et al*., 2002; Grisart *et al*., 2002; Winter *et al*., 2002). Subsequent functional studies provided strong support for that the K232A mutation in *DGAT* in fact is causing the QTL effect (Grisart *et al*., 2004). A paternally expressed QTL with major effects on muscle mass, size of the heart, and back-fat thickness is located at the distal end of pig chromosome 2p (Jeon *et al*., 1999; Nezer *et al*., 1999). The QTL has been found in several intercrosses, including a wild boar/domestic pig intercross in which the QTL allele inherited from the domestic pig was associated with higher muscle mass and reduced back-fat thickness. Sequence analysis of ~28 kb of genomic DNA around the *IGF2* locus using 15 chromosomes with known QTL status led to the identification of the causative mutation in the form of a single point mutation in the middle of intron 3 in *IGF2* (Van Laere *et al*., 2003). The mutation occurs in an evolutionary conserved CpG island and in a 16-bp fragment that is completely conserved among eight mammalian species. Functional analysis revealed that the mutation disrupts the interaction with a nuclear factor, most likely a repressor, and leads to an upregulation of *IGF2* expression in postnatal skeletal and cardiac muscle but not in fetal muscle or in liver. The mutation has gone through a selective sweep in many commercial populations selected for lean meat content (Van Laere *et al*., 2003).

As high-density marker maps are developed and the cost for SNP typing is going down, it will be possible to carry out genome-wide association analysis. The number of markers required will vary from species to species and from population to population depending on the degree of linkage disequilibrium in the target population, but the number will be on the order of 10 000 loci or more. Furthermore, as the cost for DNA sequencing is going down, we will not only generate complete genome sequences for the farm animals but we will also be able to resequence individuals representing different breeds. Such comparative

sequence analysis should reveal footprints of the selection that has taken place during domestication and selective breeding (Andersson and Georges, 2004). The expected pattern is that a haplotype carrying one or more favorable mutations will be fixed or close to fixation in certain selected lines. Such data combined with genetic data from linkage or association analysis will be a gold mine for studying genotype–phenotype relationships.

References

Andersson L and Georges M (2004) Domestic animal genomics: deciphering the genetics of complex traits. *Nature Reviews. Genetics*, **5**, 202–212.

Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I, *et al*. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*, **263**, 1771–1774.

Bateson W (1902) Experiments with poultry. *Report to the Evolution Committee of the Royal Society. London*, **1**, 87–124.

Bjornsson HT, Fallin MD and Feinberg AP (2004) An integrated epigenetic and genetic approach to common human disease. *Trends in Genetics*, **20**, 350–358.

Blott S, Kim J-J, Moisio S, Schmidt-Küntzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, *et al*. (2002) Molecular dissection of a QTL: a phenylalanine to tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics*, **163**, 253–266.

Braunschweig MH, Van Laere A-S, Buys N, Andersson L and Andersson G (2004) IGF2 antisense transcript expression in porcine postnatal muscle is affected by a quantitative trait nucleotide in intron 3. *Genomics*, **84**, 1021–1029.

Bruford MW, Bradley DG and Luikart G (2003) DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics*, **4**, 900–910.

Carlborg O and Haley C (2004) Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, **5**, 618–625.

Charlier C, Segers K, Karim L, Shay T, Gyapay G, Cockett N and Georges M (2002) The callipyge mutation enhances the expression of coregulated imprinted genes in cis without affecting their imprinting status. *Nature Genetics*, **27**, 367–369.

Ciobanu D, Bastiaansen J, Malek M, Helm J, Woollard J, Plastow G and Rothschild M (2001) Evidence for new alleles in the protein kinase adenosine monophosphate-activated γ3-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. *Genetics*, **159**, 1151–1162.

Darvasi A (1998) Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genetics*, **18**, 19–24.

Darvasi A and Soller M (1995) Advanced intercross lines, an experimental population for fine genetic-mapping. *Genetics*, **141**, 1199–1207.

Darwin C (1859) *On the Origins of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*, John Murray: London.

Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L, Mni M, Moisio S, Simon P, *et al*. (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics*, **161**, 275–287.

Freeman A, Meghen C, Machugh D, Loftus R, Achukwi M, Bado A, Sauveroche B and Bradley D (2004) Admixture and diversity in West African cattle populations. *Molecular Ecology*, **13**, 3477–3487.

Freking BA, Murphy SK, Wylie AA, Rhodes SJ, Keele JW, Leymaster KA, Jirtle RL and Smith TP (2002) Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Research*, **12**, 1496–1506.

Friedman J (2004) Modern science versus the stigma of obesity. *Nature Medicine*, **10**, 563–569.

Fujii J, Otsu K, Zorzato F, de Leon S, Khanna VK, Weiler JE, O'Brien PJ and MacLennan DH (1991) Identification of a mutation in the porcine ryanodine receptor that is associated with malignant hyperthermia. *Science*, **253**, 448–451.

Galloway SM, McNatty KP, Cambridge LM, Laitinen MP, Juengel JL, Jokiranta TS, McLaren RJ, Luiro K, Dodds KG, Montgomery GW, *et al*. (2000) Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nature Genetics*, **25**, 279–283.

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, *et al*. (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*, **139**, 907–920.

Giuffra E, Kijas JMH, Amarger V, Carlborg Ö, Jeon J-T and Andersson L (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics*, **154**, 1785–1791.

Giuffra E, Törnsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL and Andersson L (2002) A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking *KIT*. *Mammalian Genome*, **13**, 569–577.

Goring HH, Terwilliger JD and Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, **69**, 1357–1369.

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, *et al*. (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Research*, **12**, 222–231.

Grisart B, Farnir F, Karim L, Cambisano N, Kim J-J, Kvasz A, Mni M, Simon P, Frere J-M, Coppieters W, *et al*. (2004) Genetic and functional demonstration of the causality of the DGAT1 K232A mutation in the determinism of the BTA14 QTL affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2398–2403.

Grobet L, Poncelet D, Royo LJ, Brouwers B, Pirottin D, Michaux C, Menissier F, Zanotti M, Dunner S and Georges M (1998) Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. *Mammalian Genome*, **9**, 210–213.

Hasler-Rapacz J, Ellegren H, Fridolfsson AK, Kirkpatrick B, Kirk S, Andersson L and Rapacz J (1998) Identification of a mutation in the low density lipoprotein receptor gene associated with recessive familial hypercholesterolemia in swine. *American Journal of Medical Genetics*, **76**, 379–386.

Hoeschele I (2003) Mapping quantitative trait loci in outbred pedigrees. In *Handbook of Statistical Genetics*, *Second Edition*, Balding DJ, Bishop M and Cannings C (Eds.), Wiley: England, pp. 477–525.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, **432**, 717–722.

Jansen RC (2003) Quantitative trait loci in inbred lines. In *Handbook of Statistical Genetics*, *Second Edition*, Balding DJ, Bishop M and Cannings C (Eds.), Wiley: England, pp. 445–476.

Jeon J-T, Carlborg Ö, Törnsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundström K, *et al*. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. *Nature Genetics*, **21**, 157–158.

Keeling L, Andersson L, Schütz KE, Kerje S, Fredriksson R, Carlborg Ö, Cornwallis CK, Pizzari T and Jensen P (2004) Feather-pecking and victim pigmentation. *Nature*, **431**, 645–646.

Kerje S, Sharma P, Gunnarsson U, Kim H, Bagchi S, Fredriksson R, Schütz K, Jensen P, von Heijne G, Okimoto R, *et al*. (2004) The *Dominant white*, *Dun* and *Smoky* color variants in

chicken are associated with insertion/deletion polymorphisms in the *PMEL17* gene. *Genetics*, **168**, 1507–1518.

Kijas JMH and Andersson L (2001) A phylogenetic study of the origin of the domestic pig estimated from the near complete mtDNA genome. *Journal of Molecular Evolution*, **52**, 302–308.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Lundén A, Marklund S, Gustafsson V and Andersson L (2003) A nonsense mutation in the FMO3 gene underlies fishy off-flavor in cow's milk. *Genome Research*, **12**, 1885–1888.

Lynch M and Walsh B (1998) *Genetics and Analysis of Quantitative Traits*, Sinauer associates: Sunderland.

Mackinnon MJ and Georges M (1992) The effects of selection on linkage analysis for quantitative traits. *Genetics*, **132**, 1177–1185.

Marklund S, Kijas J, Rodriguez-Martinez H, Ronnstrand L, Funa K, Moller M, Lange D, Edfors-Lilja I and Andersson L (1998) Molecular basis for the dominant white phenotype in the domestic pig. *Genome Research*, **8**, 826–833.

Marklund L, Nyström PE, Stern S, Anderssson-Eklund L and Andersson L (1999) Quantitative trait loci for fatness and growth on pig chromosome 4. *Heredity*, **82**, 134–141.

Meijerink E, Neuenschwander S, Fries R, Dinter A, Bertschinger HU, Stranzinger G and Vogeli P (2000) A DNA polymorphism influencing alpha(1,2)fucosyltransferase activity of the pig FUT1 enzyme determines susceptibility of small intestinal epithelium to *Escherichia coli* F18 adhesion. *Immunogenetics*, **52**, 129–136.

Metallinos DL, Bowling AT and Rine J (1998) A missense mutation in the endothelin-B receptor gene is associated with lethal white foal syndrome: an equine version of Hirschsprung disease. *Mammalian Genome*, **9**, 426–431.

Meuwissen TH and Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, **155**, 421–430.

Meuwissen TH and Goddard ME (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution*, **33**, 605–634.

Milan D, Jeon JT, Looft C, Amarger V, Thelander M, Robic A, Rogel-Gaillard C, Paul S, Iannuccelli N, Rask L, *et al*. (2000) A mutation in *PRKAG3* associated with excess glycogen content in pig skeletal muscle. *Science*, **288**, 1248–1251.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Mulsant P, Lecerf F, Fabre S, Schibler L, Monget P, Lanneluc I, Pisselet C, Riquet J, Monniaux D, Callebaut I, *et al*. (2001) Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5104–5109.

Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics*, **14**, 353–362.

Nezer C, Moreau L, Brouwers B, Coppieters W, Detilleux J, Hanset R, Karim L, Kvasz A, Leroy P and Georges M (1999) An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics*, **21**, 155–156.

Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, **64**, 18–23.

Pailhoux E, Vigier B, Chaffaux S, Servel N, Taourit S, Furet JP, Fellous M, Grosclaude F, Cribiu EP, Cotinot C, *et al*. (2001) A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nature Genetics*, **29**, 453–458.

Perez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics*, **163**, 1497–1510.

Santschi EM, Purdy AK, Valberg SJ, Vrotsos PD, Kaese H and Mickelson JR (1998) Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses. *Mammalian Genome*, **9**, 306–309.

Seaton G, Haley CS, Knott SA, Kearsey M and Visscher PM (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics*, **18**, 339–340.

Spillman WJ (1906) Inheritance of coat colour in swine. *Science*, **24**, 441–443.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Andersson G, *et al*. (2003) Positional identification of a regulatory mutation in *IGF2* causing a major QTL effect on muscle growth in the pig. *Nature*, **425**, 832–836.

Weller J, Kashi Y and Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science*, **73**, 2525–2537.

Winter A, Kramer W, Werner FA, Kollers S, Kata S, Durstewitz G, Buitkamp J, Womack JE, Thaller G and Fries R (2002) Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9300–9305.

Yang GC, Croaker D, Zhang AL, Manglick P, Cartmill T and Cass D (1998) A dinucleotide mutation in the endothelin-B receptor gene is associated with lethal white foal syndrome (LWFS); a horse variant of Hirschsprung disease. *Human Molecular Genetics*, **7**, 1047–1052.

Specialist Review

Mouse mutagenesis and gene function

Ralf Kühn
Institute for Developmental Genetics, Neuherberg, Germany

Wolfgang Wurst
Institute for Developmental Genetics, Neuherberg, Germany
Max-Plank-Institute for Psychiatry, Munich, Germany

1. Introduction

With the completion of the mouse and human genome sequences, a major challenge is the functional characterization of every gene within the mammalian genome and the identification of gene products and their molecular interaction network. The mouse offers many advantages for the use of genetics to the study of human biology and disease. Its development, body plan, physiology, behavior, and diseases (*see* Article 12, **Haplotype mapping**, Volume 3) have much in common, which is based on the fact that 99% of the mouse genes have a human ortholog. The investigation of gene function using mouse models is built on many years of technology development. A variety of mouse mutagenesis technologies, either gene- or phenotype-driven, are used as systematic approaches. The availability of the mouse genome sequences (*see* Article 47, **The mouse genome sequence**, Volume 3) supports gene-driven approaches such as gene-trap and targeted mutagenesis in embryonic stem (ES) cells, allowing an efficiency and precision of gene disruption unmatched among other mammals. Furthermore, chemical and transposon mutagenesis of the mouse genome allows to perform phenotype-driven screens for the unbiased identification of phenotype–genotype correlations involved in models of human disease. Taken together, the application of these approaches have already resulted in a worldwide collection of several thousand mouse mutants and will form the basis to generate mutations in every gene and to decipher their physiological function. In the following sections, we present a comprehensive review of gene- and phenotype-driven mutagenesis strategies applied to the mouse genome. Besides a summary of the basic principles of each approach, we emphasize on the latest and future developments of mouse mutagenesis.

2. Gene-trap mutagenesis

Gene-trap mutagenesis is based on the random integration of a gene-trap vector across the genome of ES cells and the disruption of coding sequences of genes through vector-specific elements. Gene-trap vectors simultaneously mutate a gene at the site of insertion, provide a sequence tag for the rapid identification of the disrupted gene, and mimic the expression of the tagged gene by a reporter gene. Since a single DNA or retroviral vector can be used to hit a large number of genes, gene trapping is a high-throughput insertional mutagenesis approach that enables to establish libraries of mutant ES cell clones rapidly and at low costs (Gossler *et al*., 1989; Friedrich *et al*., 1991; Skarnes *et al*., 1992; Wurst *et al*., 1995; Zambrowicz *et al*., 1998; Wiles *et al*., 2000; Hansen *et al*., 2003). The resulting databases of mutant genes provide the basis for the establishment of mutant mouse strains through germ-line chimeras raised from selected ES cell clones.

Typical gene-trap vectors are promoterless and contain a reporter–selector cassette that functions by generating a fusion transcript with the endogenous gene (Figure 1a). The most widely used βgeo cassette contains an ATG-less hybrid-coding region for the β-galactosidase reporter and the neomycin phosphotransferase selection marker (Friedrich *et al*., 1991). The inclusion of a splice acceptor element (SA) upstream of βgeo leads to the generation of a fusion transcript upon vector integration into an intron of a transcribed gene in ES cells. The fusion transcript will ideally be prematurely terminated at a polyadenylation signal sequence (polyA) placed downstream of the βgeo element. In case that the translational reading frames of the trapped transcript and the βgeo cassette are in line, a fusion protein is produced that confers resistance of the ES cell clone to the neomycin analog G418. To identify genes independent of reading frames, an internal ribosomal entry site (IRES) can be placed between SA and a βgeo version that includes an ATG start codon.

Upon introduction of a gene-trap vector into ES cells by electroporation or retroviral infection, the population is selected for G418 resistance such that only ES colonies harboring a productive vector integration into an active gene are able to survive. This stringent selection scheme is the basis for the high efficiency of gene-trapping technology since each resistant ES cell clone represents an independent integration event into a unique gene (Stanford *et al*., 2001; Floss and Wurst, 2002). However, gene-trap vectors that rely on the use of a splice acceptor element to express a resistance marker can identify all genes sufficiently transcribed in ES cells and that contain introns but not those genes that are not expressed in ES cells. To further extend the application of gene trapping to all genes, independent of their expression status in ES cells, poly(A)-trap vectors have been developed. These vectors contain a promoter-driven resistance gene followed by a splice donor element (SD) without a poly(A) sequence (Niwa *et al*., 1993; Zambrowicz *et al*., 1998). Thus, drug resistance is only obtained after successful vector integration into an endogenous gene and the capture of its poly(A) signal by the vector-derived transcript.

ES cells are commonly used as the substrate for insertional mutagenesis through gene-trap vectors. Recently, an alternative strategy has been developed that relies on the *Sleeping Beauty* transposable element as germ-line insertional mutagen (Izsvak

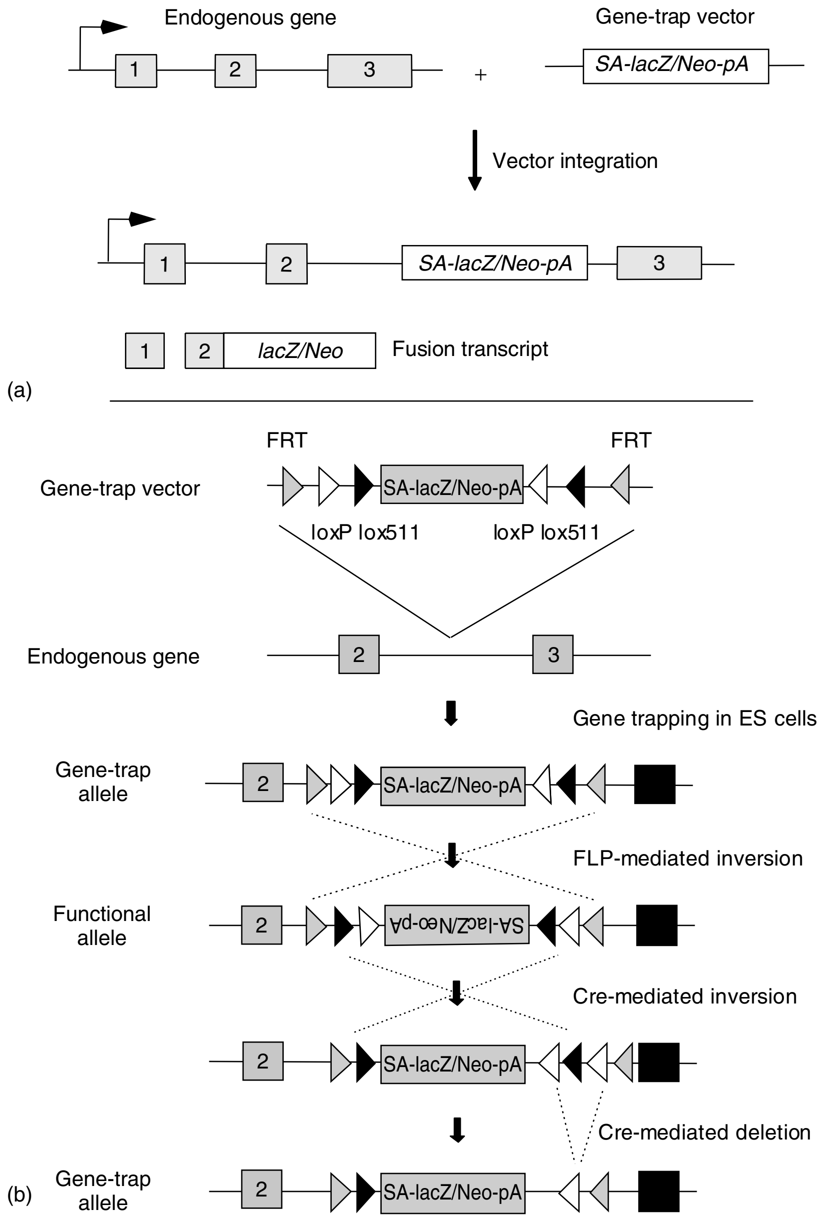


Figure 1 Gene-trap mutagenesis. (a) integration of a standard gene-trap vector into an endogenous gene. SA – splice acceptor sequence, pA – poly(A) sequence. (b) Integration of a conditional gene-trap vector into an endogenous gene. The initial gene-trap allele can be converted into a functional allele by FLP-mediated inversion of the mutagenic gene-trap cassette in ES cells or FLP deleter mice. This cassette can be irreversibly reinverted through Cre-mediated recombination in ES cells or Cre transgenic mice

et al., 2000; Horie *et al*., 2003; Carlson *et al*., 2003). In this in vivo approach, doubly transgenic mice harbor a transposase transgene and a transposon poly(A) gene-trap vector that can be mobilized by transposition. For this purpose, the mutagenic cassette of a poly(A) gene-trap vector is flanked by a pair of transposon terminal inverted repeats that contain transposase binding sites. This element is excised in the male germline of doubly transgenic mice and reintegrates into new genomic locations. Upon the outcross of transgenic males to wild-type females, the offspring contains new transposon insertions that are found at a frequency of ~2 per male gamete. The analysis of several hundred integration sites revealed that about 30% of transposition events occur locally clustered within 3 Mb near the donor site, while the other events are widely distributed across various chromosomes (Horie *et al*., 2003; Carlson *et al*., 2003). Thus, the in vivo transposition of gene-trap vectors can be applied to region-specific and genome-wide mutagenesis.

In addition to broadly acting gene and poly(A) traps, specific classes of proteins can be trapped with modified vectors. A secretory trap vector, designed to capture genes for secreted or membrane proteins expressed in ES cells, contains a membrane spanning domain fused to the amino terminus of the βgeo reporter–selector cassette (Skarnes *et al*., 1995; Mitchell *et al*., 2001). Thereby, the activation of βgeo depends on the production of fusion proteins that incorporate an N-terminal signal sequence or a transmembrane domain from the gene at the insertion site. To further increase the versatility of gene-trap mutagenesis, vectors were developed that include recognition sequences for the site-specific DNA recombinase Cre (Araki *et al*., 1999; Hardouin and Nagy, 2000). These vectors enable recombinase-assisted postinsertional modifications at the gene-trap locus in ES cells to drive the expression of any foreign cDNA from the specific promoter of the trapped gene.

The currently employed gene-trap vectors irreversibly modify the endogenous target genes, comparable to germ-line null mutations created by gene targeting in ES cells (Figure 1a). To avoid the potential embryonic lethality of targeted germ-line mutants, a scheme for conditional gene targeting has been introduced that enables to restrict gene inactivation to specific cell types by the use of the Cre/loxP recombination system (see Section (4)). The upcoming generation of gene-trap vectors is designed to combine the advantages of gene-trap and conditional mutagenesis by the development of conditional gene-trap vectors (Melchner and Stewart, 2004) (Figure 1b). These vectors rely on a reporter–selector cassette like βgeo that can be independently inverted by the site-specific recombinases Cre or FLP. For this purpose, the cassette can be flanked by a pair of FRT sites that allows to invert the cassette by FLP-mediated recombination upon the usual establishment of a gene-trap integration into an expressed gene. In its original sense, orientation of the cassette disrupts the expression of the trapped gene, whereas the inverted cassette should be nonmutagenic and enable gene expression at wild-type levels. In addition, the cassette must be flanked by a pair of wild-type and mutant loxP sites in a specific order and orientation such that an irreversible inversion is mediated by Cre recombinase (FLEx strategy; Schnütgen *et al*., 2003). When mice that harbor such a silent gene-trap cassette are crossed to a transgenic strain with tissue-specific expression of Cre, the gene trap should be reactivated in vivo and lead to the conditional disruption of the target gene.

The analysis of more than 5000 gene-trap events by Hansen *et al*. (2003) revealed that gene-trap insertions are dispersed throughout the genome and occur more frequently in chromosomes with a high density of genes. All functional classes of mammalian genes are amenable to gene trapping. However, the integration of gene-trap vectors is not entirely random and several preferred integration sites (hot spots), some of which occur many times, have been observed. About half of these hot spots are associated with the use of specific gene-trap vectors, while the other half occurs independent of the specific vector design. With the increasing size of a gene-trap library, the rate of trapping new genes is not linear but declines since multiple integrations accumulate and the pool of new trappable genes decreases. About 700 genes were trapped within a pool of the first 1600 integrations of a gene-trap vector (Hansen *et al*., 2003), while only 1 new gene was added every 35 tags in a library comprising over 100 000 insertions of a poly(A)-trap vector (Skarnes *et al*., 2004). Considering that over half of the multiple integrations are vector specific, the most effective way to saturate the genome with gene-trap insertions is the use of a variety of gene-trap vectors.

Ideally, the integration of a gene-trap vector leads to the functional knockout of the trapped endogenous gene. The mutagenicity of a given vector in respect of interrupting the endogenous transcript depends on the relative strength of the splice acceptor and poly(A) sites flanking the reporter–selector element. In addition, the mutagenicity of gene-trap vectors is likely also determined by their integration position within a trapped gene. The majority of retroviral gene-trap insertions occurs in the 5′-half of genes, whereas insertions of plasmid vectors are distributed more evenly over the entire gene-coding regions (Hansen *et al*., 2003). In some cases, the combination of a specific insertion site and gene-trap vector is not entirely effective such that the vector sequence can be excised from the endogenous transcript as part of the intron into which the integration has occurred. In this case, wild-type gene product is produced to some extent and may lead to a hypomorphic mutation. In a side-by-side comparison of 11 gene-trap mutants with targeted mutants affecting the same genes, all but one phenocopied the targeted mutants, while one strain showed a partial loss-of-function (Mitchell *et al*., 2001). The overall analysis of homozygous mouse mutants derived from gene-trap ES cell clones revealed that 60% of the strains exhibit obvious phenotypes and 30% of these phenotypes lead to embryonic lethality (Stanford *et al*., 2001; Hansen *et al*., 2003). This frequency is comparable with mutants generated by gene targeting, indicating that most gene-trap insertions result in null alleles.

Gene-trap mutagenesis in ES cells is a large-scale effort that cannot be completed by a single group but requires cooperation within the research community. Six gene-trap screens run by academic research centers across Europe and North America are combined in the International Gene Trap Consortium (IGTC; Table 1). The IGTC collection of gene-trap ES cell clones includes 27 000 independent insertions, which represent a 32% coverage of all mouse genes (Skarnes *et al*., 2004). The IGTC gene-trap ES cell clones are freely available to academic scientists, and the sequence tags of all insertions are mapped on the Ensembl mouse genome server (Table 1), providing a direct link to any gene of interest. A parallel effort was initiated by the biotechnology company Lexicon Genetics that developed a large library of gene-trap clones (OmniBank; Zambrowicz *et al*., 1998). This collection

Table 1 Web-based resources related to mouse mutagenesis and gene function

Gene trap	
International Gene Trap Consortium	**www.igtc.ca**
German Gene Trap Consortium	**www.genetrap.de**
Sanger Institute Gene Trap Resource	**www.sanger.ac.uk/PostGenomics/genetrap**
BayGenomics	**www.baygenomics.ucsf.edu**
Ensembl Mouse Genome Server	**www.ensembl.org/Mus_musculus**
Targeted mutants	
The Jackson Laboratory – Induced Mutant Resource	**www.jax.org/imr/notes.html**
Mouse Knockout and Mutation Database	**http://research.bmn.com/mkmd**
The Jackson Laboratory – Transgenic/Targeted Mutation Database	**http://tbase.jax.org**
Cre Mouse Database	**www.mshri.on.ca/nagy/cre.htm**
ENU mutagenesis	
ENU-Mouse Mutagenesis Screen Project	**www.gsf.de/ieg/groups/enu-mouse.html**
Harwell Mutagenesis Programme	**www.mut.har.mrc.ac.uk/**
EUMORPHIA	**www.eumorphia.org**
Mouse Clinical Institute	**www-mci.u-strasbg.fr**
German Mouse Clinic	**www.gsf.de/ieg/gmc**

of 200 000 sequence tags deposited in GenBank achieves close to 60% coverage of the mouse genome (Skarnes *et al*., 2004). The two efforts together presently cover nearly two-thirds of the mouse genome such that gene trapping has proven to be the most effective strategy to mutate a substantial fraction of all mouse genes. Therefore, gene trapping is the first choice of the European- and US-based initiatives that plan the complete mutagenesis of all mouse genes (Auwerx *et al*., 2004; Austin *et al*., 2004).

3. Gene-targeting mutagenesis

Gene targeting allows the introduction of predesigned, site-specific modifications into the mouse genome (Capecchi, 1989). It has been extensively used in the past decade for the preplanned disruption of genes in the murine germline, resulting in mutant strains referred to as *knockout mice*. Gene inactivation is achieved through the insertion of a selectable marker into an exon of the target gene or the replacement of one or more exons.

The mutant allele is initially assembled in a specifically designed gene-targeting vector such that the selectable marker is flanked at both sides with genomic segments of the target gene that serve as homology regions to initiate homologous recombination. The frequency of homologous recombination increases with the length of these homology arms. Usually, arms with a combined length of 10–15 kb are cloned into standard, high-copy plasmid vectors that accommodate up to 20 kb of foreign DNA. To select against random vector integrations, a negative selectable marker, such as the *Herpes simplex* thymidine kinase or diphteria toxin gene, can be included at one end of the targeting vector. Upon electroporation of such a vector into ES cells and the selection of stable integrants, clones that

underwent a homologous recombination event can be identified through the analysis of genomic DNA using a PCR or Southern blot strategy. Using such standard gene-targeting vectors, the frequency of homologous recombination falls into the range of 0.1–10% of stable transfected ES cell clones. This rate depends on the length of the vector homology region, the degree of sequence identity of this region with the genomic DNA of the ES cell line, and likely on the differential accessibility of individual genomic loci to homologous recombination. Optimal rates are achieved with longer homology regions and by the use of genomic fragments that exhibit sequence identity to the genome of the ES cell line, that is, both should be isogenic and derived from the same inbred mouse strain. Upon the isolation of recombinant ES cell clones, modified ES cells are injected into blastocysts to transmit the mutant allele through the germline of chimeras and to establish a mutant strain. Through interbreeding of heterozygous mutants, homozygotes are obtained that can be used for phenotype analysis. Most knockout strains have been generated one by one through the standard scheme described above; an approach that typically requires 1–2 years of hands-on work for vector construction, ES cell culture, and mouse breeding. Working protocols and technical details of the gene-targeting technology have been compiled in a recent manual (Nagy *et al*., 2003).

Since the first demonstration of homologous recombination in ES cells by Thomas and Capecchi in 1987, gene targeting has been successfully used to generate more than 2500 knockout mouse strains. Thus, almost 10% of the mouse genome is presently covered by targeted mutations. While all published mutants and their phenotypes are recorded in databases (Table 1), the availability and distribution of these strains becomes increasingly difficult. The Jackson Laboratories, as the largest distribution center, presently holds about 350 targeted mutants (Table 1).

Using the "classical" gene-targeting approach described above, germ-line mutants are obtained that harbor the knockout mutation in all cells throughout development. This strategy identifies the first essential function of a gene during ontogeny. If the gene product fulfills an important role in development, its inactivation can lead to embryonic lethality, precluding further analysis in adult mice. In the mean, about 30% of all knockout mouse strains exhibit an embryonic lethal phenotype; for specific classes of genes, for example, those regulating angiogenesis, this rate can reach 100%. To avoid embryonic lethality and to study gene function only in specific cell types, Gu *et al*. (1994) introduced a modified, conditional gene-targeting scheme that allows to restrict gene inactivation to specific cell types or developmental stages (Figure 2). In a conditional mutant gene, inactivation is achieved by the insertion of two 34-bp recognition (loxP) sites of the site-specific DNA recombinase Cre into introns of the target gene such that recombination results in the deletion of loxP-flanked exons. Conditional mutants initially require the generation of two mouse strains: one strain harboring a loxP-flanked gene segment obtained by gene targeting in ES cells (Figure 2a) and a second, transgenic strain expressing Cre recombinase in one or several cell types. The conditional mutant is generated by crossing these two strains such that target gene inactivation occurs in a spatial and temporal restricted manner, according to the pattern of recombinase expression in the Cre transgenic strain (Torres and Kühn, 1997; Nagy *et al*., 2003; Figure 2b). In addition, the loxP-modified strain can be converted into a conventional germ-line mutant through the cross to a Cre

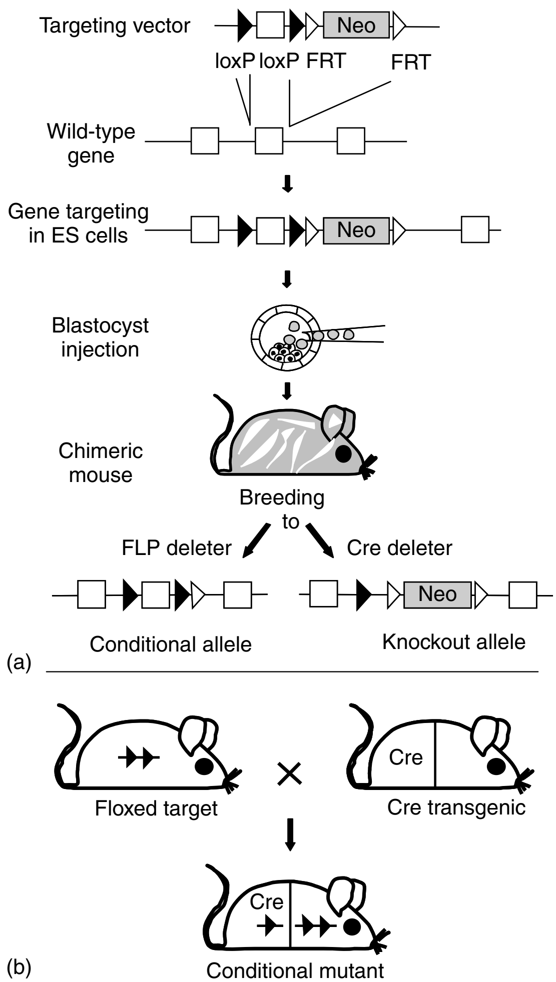


Figure 2 Conditional gene targeting. (a) A target gene exon is flanked by loxP sites, the selection marker (neo) is flanked by FRT sites. Upon gene targeting in ES cells and germ-line transmission, the conditional allele is generated upon deletion of the selection marker by a cross to FLP deleter mice. A cross to Cre deleter mice generates a germ-line knockout allele. (b) A mouse carrying a loxP-flanked (floxed) target gene is crossed to a transgenic mouse expressing Cre recombinase in specific cell types (Cre is restricted to the left-hand portion). In the resultant double transgenic mouse (bottom), recombination of the floxed target is restricted to cells expressing Cre. Filled triangles represent the loxP sites

deleter strain that expresses recombinase in germ cells (Figure 2a). Both types of mutants are often generated side by side from the same ES cell clone to investigate gene function during embryonic development and in the adult animal.

Conditional mutants have been used to address various biological questions that could not be resolved with germ-line mutants, often because a null allele results in an embryonic or neonatal lethal phenotype. For this purpose, more than 100

Cre transgenic strains with tissue-specific recombinase expression have been published that cover many cell types for which a specific promoter region is available (Table 1; Nagy *et al*., 2001); about 30 of these strains are available from the Jackson Laboratory. The characteristics of a given line can be identified by crossing to a Cre reporter strain that activates a reporter gene upon Cre-mediated deletion of a loxP-flanked transcriptional stop cassette (Soriano, 1999). Most Cre transgenic strains express recombinase from a constitutively active promoter, starting with the activation of the promoter region during development.

A smaller number of Cre mice allows to induce Cre activity in one or more cell types upon the administration of a small molecule inducer (Lewandoski, 2001). Transcriptional control of Cre has been achieved by the tetracycline-inducible gene expression system in transgenic mice using doxycycline as inducer (Utomo *et al*., 1999). This system requires the independent introduction of two genes coding for a doxycycline-regulated transactivator protein and for Cre recombinase, and thus numerous transgenic lines must be crossed and tested to identify the strains in which both genes are optimally regulated. Posttranslational control of Cre activity has been achieved by the expression of fusion proteins of Cre with mutant ligand-binding domains of the estrogen or progesterone receptor (Feil *et al*., 1997; Kellendonk *et al*., 1999; Branda and Dymecki, 2004). These ligand-binding domains are unresponsive to natural steroids but can be activated in mice by the administration of synthetic steroid antagonists.

Conditional alleles have been generated for more than 100 genes that lead to embryonic lethality in case of a germ-line knockout (Kwan, 2002). The generation of conditional alleles involves the same technology as the production of germ-line knockouts but the construction of gene-targeting vectors and mouse breedings require more time and efforts. For the construction of a conditional gene-targeting vector, a selection marker and a loxP site are inserted into one intron of the target gene while a second loxP sequence is placed into another intron. Upon homologous recombination, the selection marker gene is usually removed from the targeted allele to avoid its potential interference with the expression of the loxP-modified gene. For this purpose, the selection marker can be flanked with FLP recombinase recognition (FRT) sites and deleted from the genome by a cross of mice harboring the targeted allele to a transgenic strain that expresses FLP in germ cells (Rodriguez *et al*., 2000; Figure 2a). Upon removal of the selection marker, the loxP-flanked allele can be bred to homozygosity together with the required Cre transgene to obtain conditional mutants for phenotype analysis (Figure 2b).

Owing to the higher efforts required for the generation of conditional mutants, this technology has been mostly applied to genes that possess important functions in the adult but exhibit an embryonic lethal knockout phenotype. Besides the avoidance of embryonic lethality, a conditional mutant can reveal information about the function of a widely expressed gene in different tissues by combination with various Cre lines.

In addition to the use of Cre/loxP for gene inactivation, site-specific recombination has been applied to achieve other types of genome manipulation in ES cells or mice. These include the generation of large chromosomal deletions or inversions, of chromosomal translocations, gene replacement, recombinase-mediated cassette exchange, and the inversion of gene segments (Branda and Dymecki, 2004).

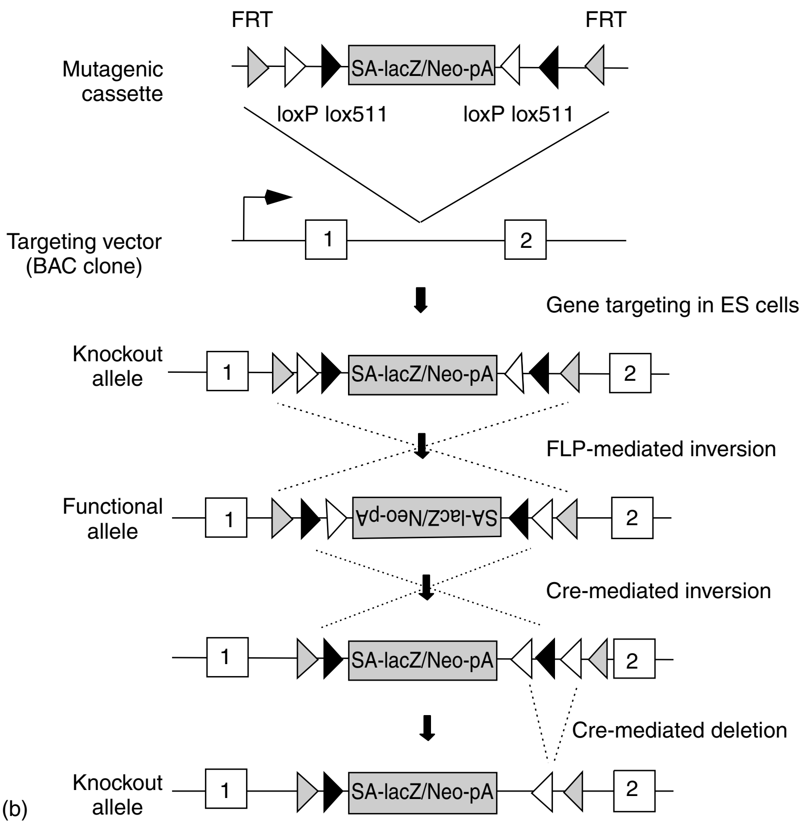
Gene targeting, in its first decade, has largely progressed in a one-by-one manner by the contributions of a large number of laboratories; for each mutant, 2–3 years of work are required for vector construction, ES cell culture, mouse breeding, and analysis. Thus, gene targeting is presently a low-throughput technology, in contrast to gene trapping that generates insertional mutations in ES cells at much larger numbers and with less effort because a single, generic vector can be used to mutagenize any gene. However, with the recent advance in technology, as described below, it is possible to produce targeted mutants faster and at larger scale.

The first advance is the development of novel DNA engineering strategies that rely on homologous recombination in bacteria, utilizing the phage-derived recombination protein pairs RecE/RecT (ET cloning; Muyrers *et al*., 2001) or Redα/Redβ (Recombineering; Copeland *et al*., 2001). These recombination functions are either carried on plasmids or have been inserted into the bacterial genome. For gene-targeting purposes, this technology enables to manipulate large genomic sequences cloned into BAC vectors without the use of restriction enzymes or ligation reactions (Angrand *et al*., 1999). Since the whole mouse genome is available in the form of sequenced BAC clones, all genes are readily accessible to these methods.

First, it is possible to subclone genomic fragments from BAC vectors into standard cloning plasmids as a basis of gene-targeting vector construction. In a second ET/recombineering step, a selection cassette that functions in bacteria as well as ES cells can be inserted at a preselected site to produce a vector for standard knockout alleles. This can be combined with a third step to introduce a loxP sequence at a distant site for the construction of conditional gene-targeting vectors (Muyrers *et al*., 2001; Liu *et al*., 2003). Thereby, starting from BAC clones and oligonucleotides, ET cloning/recombineering allows to construct gene-targeting vectors within a few weeks.

The second advance builds on these BAC manipulation methods and further simplifies their handling such that complete, modified BAC clones are directly used as gene-targeting vectors. Owing to the size of the vector homology arms of 100–200 kb, it is not practicable to identify recombinant ES cell clones by standard Southern blotting such that two alternative strategies were developed. The first, "Velocigene" procedure (Valenzuela *et al*., 2003), takes advantage of the fact that only one copy of the wild-type target gene remains in homologous recombined ES cells, while random BAC vector integrants retain both wild-type copies. To test whether ES cell clones harbor a random or targeted vector integration, their genomic DNA is assayed by quantitative PCR to determine the copy number of the wild-type allele in comparison to an internal standard. In the second approach (Yang and Seed, 2003), all ES cell clones are first screened for the presence of a random integration using a vector-specific PCR. The remaining clones are assayed by fluorescent in situ hybridization (FISH) for the presence of only two hybridization signals indicating homologous recombinants. The use of BAC clones considerably simplifies the construction of gene-targeting vectors. Furthermore, BAC-targeting vectors are reported to result in high frequencies of homologous recombination in ES cells (Valenzuela *et al*., 2003), even if the vector arms and the ES cell line are nonisogenic, that is, derived from different inbred strains.

It can be anticipated that the next technical step toward the streamlined production of targeted mutants will follow shortly, with the development of generic



(b)

Figure 3 Generic conditional gene targeting. A conditional mutagenic cassette, inserted by bacterial homologous recombination into a generic site (intron 1) of genomic BAC clones, serves as a gene-targeting vector in ES cells. The initial knockout allele can be converted into a functional allele by FLP-mediated inversion of the mutagenic cassette in ES cells or FLP deleter mice. This cassette can be irreversibly reinverted through Cre-mediated inversion and deletion in ES cells or Cre transgenic mice

conditional gene-targeting vectors. These vectors will combine BAC clones as backbone of targeting vectors with an invertable gene disruption cassette, as used for conditional gene-trap mutagenesis (Figure 3). The irreversible inversion of a mutagenic cassette through Cre recombinase can be achieved by the FLEx approach (Schnütgen *et al*., 2003) that combines wild-type and mutant lox sites with a Cre-mediated inversion and deletion step. In addition, the cassette can be (reversibly) inverted by FLP/FRT-mediated recombination such that the initial knockout configuration is converted into a conditional allele in ES cells or FLP deleter mice (Figure 3). Such a cassette could be inserted in a generic manner into the first intron of any target gene and would allow to produce conditional targeting vectors by a single ET cloning step. This technology may contribute to cover the whole mouse genome by a combination of gene-trap and targeted mutagenesis, as planned by the European mouse mutagenesis program (Auwerx *et al*., 2004).

4. ENU mutagenesis

ENU (ethylnitrosourea) mutagenesis is a phenotype-driven approach, whereby large numbers of mutations are induced at random and new mutants are identified through specific phenotype screens (Justice, 2000; Brown and Balling, 2001; Brown and Hardisty, 2003). Since no prior assumption is made about the underlying genes, ENU mutagenesis represents an unbiased way for the identification of genes and genetic pathways involved in biological processes.

At the beginning of a mouse ENU mutagenesis screen, high doses of ENU are repeatedly administrated to a group of males. Upon a transient phase of infertility, the testes of ENU-treated animals are repopulated by gametes derived from mutagenized stem cells. The chemical mutagen ENU generates point mutations by the transfer of its ethyl group on oxygen or nitrogen residues of DNA, resulting in mispairing and base-pair substitution upon DNA replication. The highest ENU-induced mutation rates occur in premeiotic spermatogonial stem cells. Given a specific locus mutation rate of around 10^{-3}, the likelihood curves indicate that one has to screen around 2000–3000 gametes to have a 90% or higher chance of observing one mutation. The mutagenized males are bred to wild-type females and progeny of the resulting G1 generation can be directly screened for the presence of novel phenotypes caused by dominant mutations (Soewarto *et al*., 2003). To identify recessive mutations, G1 males and females are crossed in defined breeding pairs to produce a second, G2 generation. G2 females are then backcrossed to their father to obtain a third (G3) generation in which recessive mutations, inherited by the father, can become homozygous (Figure 4). G3 progeny (around 20–30 in total) from several litters derived from each G1 male are screened for the biological parameters of interest to score for individuals that exhibit a mutant phenotype. Upon the identification of variant individuals, breeding allows to establish mutant strains that are maintained as breeding colonies for further analysis. Confirmed mutant strains that exhibit similar phenotypes may be based on independent mutations in the same gene; alternatively, different genes with similar functions could be affected. The number of genes affected in a group of mutants can be estimated by complementation analysis through the cross of different mutants.

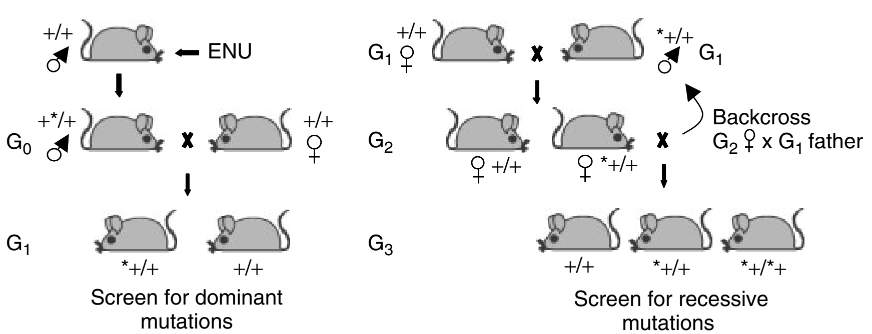


Figure 4 ENU mutagenesis. Breeding scheme to identify dominant and recessive mutants induced by ENU mutagenesis

To identify the mutant gene and the underlying mutation at the molecular level, a generic low-resolution mapping approach enables the rapid assignment of a mutation to a particular chromosomal segment (*see* Article 9, **Genome mapping overview**, Volume 3 and Article 15, **Linkage mapping**, Volume 3). The initial step of genetic mapping requires that mutants on the specific inbred background be outcrossed with a second mapping strain, resulting in F1 hybrid progeny that acquires one set of chromosomes from each of the parental strains. The mapping strain should exhibit numerous polymorphic markers interspersed throughout the genome. Meiotic recombination in F1 mice shuffles chromosomal segments and allows to identify in the next generation markers that cosegregate with the mutant phenotype and are closely linked to the mutant locus (linkage analysis). For the mapping of dominant mutations, F1 mice are further backcrossed to the mapping strain; to map recessive mutations, F1 mice are intercrossed in brother–sister matings, yielding progeny of which 25% are homozygous for the mutation. As an alternative to natural matings, large numbers of offspring can be produced from a single male through in vitro fertilization (IVF) of oocytes using fresh or frozen sperm (Thornton *et al*., 1999). The offspring of these matings are then analyzed for the linkage of the mutant phenotype with specific genetic markers. For this purpose, polymorphic genetic markers such as simple sequence length polymorphisms (SSLPs; microsatellites; Witmer *et al*., 2003) and single nucleotide polymorphisms (SNPs; Lindblad-Toh *et al*., 2000) are used that are anchored on the genome and can be analyzed by PCR amplification from genomic DNA. Provided that a sufficient number of mice is available, this mapping approach often allows to assign mutations to a region of less than 1 Mb of the genome, corresponding to 5–10 genes that can be considered as candidates for the mutant gene. In consideration of the results of biological analyses obtained from the mutant and computational prediction of gene functions, a likely candidate gene can be often defined. Finally, this candidate gene must be analyzed at the molecular level by sequencing of cDNA or PCR-amplified exon regions from mutant and control mice.

The analysis of 62 germ-line mutations derived from 24 genes revealed that ENU preferentially modifies A/T base pairs and mainly leads to T/A transversions and G/C transitions (Justice *et al*., 1999). Translated into protein products, 64% of these changes result in missense mutations, 10% lead to a premature stop codon, and 26% affect mRNA splicing. In most cases, the affected proteins exhibit a partial or complete loss-of-function but gain-of-function mutations, for example, by the loss of an inhibitory domain of a tumor suppressor gene (Moser *et al*., 1995), have been also described. An advantage of ENU mutagenesis is that it can evoke a range of mutations within a single gene that may affect protein function in different ways. Such allelic series can provide a fine structure dissection of protein function.

Since ENU mutagenesis is a phenotype-driven approach, mutant isolation relies on the spectrum and quality of available phenotype assays. The screen of a large number of animals further requires that first-line detection assays are broad, simple, and inexpensive. Visible phenotypes that affect the eye, coat, or size are most simple to detect. Standard test for reflexes, sight or hearing loss, balance, and coordination can be used to screen for motor and sensory organ phenotypes. X-ray analysis enables to examine skeletal and soft tissue development. Clinical tests performed on mouse blood can yield phenotypes relevant to hematological and immunological

disease, while clinical chemistry on serum components can diagnose multiple organ system anomalies. The development of new approaches in phenotyping and the standardization of primary and secondary phenotyping protocols for all body systems in the mouse is the subject of a dedicated research programme (EUMORPHIA; Table 1). Some of the participating phenotyping centers, like the Mouse Clinical Institute and the German Mouse Clinic (Table 1), also provide service units for the characterization of targeted and gene-trap mutants.

The results of two genome-wide screens for dominant mutations of the United Kingdom and German ENU mutagenesis programmes have been recently reported (Nolan *et al*., 2000; Hrabe de Angelis *et al*., 2000; Justice, 2000). Both groups applied a range of phenotype screens to identify mutants relevant to human disease by assessment of visible phenotypes, the detection of sensory, neuromuscular, and neurological defects by a stepwise assessment of many parameters (SHIRPA protocol) or hematological and clinical chemistry assays. From screening 40 000 mice together, these groups have isolated around 1000 new mutations. Mutants were found for any phenotypic area of interest, and the overall rate of recovery of dominant mutations was in the range of 2%. About 10 large-scale ENU mutagenesis programmes are maintained worldwide, each with a specific biological focus on neurological, behavioral, morphological, developmental, or immunological phenotypes (Brown and Balling, 2001). These ongoing efforts, mostly genome-wide screens for recessive mutations, have already yielded important models of human disease (Vreugde *et al*., 2002; Toye *et al*., 2004) and will further significantly contribute to draw a functional map of the genome.

Besides the mutagenesis of the wild-type genome, ENU can be further employed to screen for modifier genes (Nadeau, 2001) on the genetic background of an established mutant or transgenic strain that exhibits a specific phenotype. Such modifier screens are used to search for additional members of a genetic pathway that influence a given phenotype in positive or negative manner, resulting in the isolation of suppressor or enhancer mutants. This and other sophisticated screening procedures are routinely used in fruit flies (St Johnston, 2002; *see also* Article 42, **Systematic mutagenesis of nonmammalian model species**, Volume 3) to identify modifiers of preexisting genetic defects and may provide a paradigm for the future development of ENU mutagenesis screens in the mouse. The results of the first suppressor screen on the background of a targeted mouse mutant have been recently reported (Carpinelli *et al*., 2004).

A new opportunity to capitalize on chemical mutagenesis in the mouse is the development of gene-driven ENU screens. This approach is based on parallel archives of genomic DNA and frozen sperm from G1 mutant males derived from ENU mutagenesis programmes. The DNA samples are screened for mutations in the gene of interest by PCR amplification of exon regions and the identification of base-pair substitutions by denaturing high-performance liquid chromatography (DHPLC). To detect a single event in a given gene at the chance of 90% requires the screen of approximately 2500 samples (Coghill *et al*., 2002). Upon identification of a DNA sample carrying a mutation, the corresponding mouse mutant can be recovered from the frozen sperm of the affected male by IVF. The proof of principle for this approach has recently been achieved by screening for mutations in the connexin 26 gene (Coghill *et al*., 2002).

In a related approach, ES cells are treated in vitro with ENU or other chemical mutagens (Chen *et al*., 2000; Munroe *et al*., 2000). ENU is able to induce loss-of-function mutations in the *Hprt* locus at a frequency of 10^{-3}, without affecting the germ-line competence of treated ES cells. For phenotype-driven screens, chimeric mice can be produced from bulk cultures of mutagenized ES cells followed by the backcrossing or intercrossing of ES cell–derived offspring. Alternatively, the gene-driven approach enables to isolate mutations in nonselectable genes of interest at the cellular level by screening DNA or cDNA samples derived from libraries of mutagenized ES cell clones. Using a library of 2060 mutagenized ES cell clones, Vivian *et al*. (2002) could detect 29 allelic mutations of the *Smad2* and *Smad4* genes by DHPLC-based heteroduplex analysis of RT-PCR products covering the entire coding regions of both genes.

5. Future developments

Over the last decade, a rich variety of mouse mutagenesis technologies has emerged. Appreciation of the power of mouse genetics to inform about human physiology and disease, and the ease of producing mutant alleles, has led to large-scale gene-trap and ENU mutagenesis programs. These efforts, together with the individual production of targeted mutants, have started to build a functional map of the mammalian genome. Despite these efforts, the absolute number of mouse mutants to date remains only a 10% fraction of the ~30 000 genes in the mouse genome. To improve this situation, several initiatives toward a genome-wide project for the systematic mutational and functional analysis of all mouse genes have recently emerged. The European conditional mouse mutagenesis programme (EUCOMM; Auwerx *et al*., 2004) puts priority on the development of conditional gene trap and generic conditional gene-targeting strategies. The US-based Knockout Mouse Project (KOMP; Austin *et al*., 2004) also proposes to saturate the genome by a combination of gene trapping and gene targeting, with less emphasis on conditional mutagenesis. It can be expected that these initiatives will lay the ground to build a more complete functional map of the mammalian genome over the next decade and promise a bright future for mouse genetics.

References

Angrand PO, Daigle N, van der Hoeven F, Scholer HR and Stewart AF (1999) Simplified generation of targeting constructs using ET recombination. *Nucleic Acids Research*, **27**, e16.

Araki K, Imaizumi T, Sekimoto T, Yoshinobu K, Yoshimuta J, Akizuki M, Miura K, Araki M and Yamamura K (1999) Exchangeable gene trap using the Cre/mutated lox system. *Cellular and Molecular Biology (Noisy-le-Grand)*, **45**, 737–750.

Austin CP, Battey JF, Bradley A, Bucan M, Capecchi MR, Collins FS, Cook WC, Dove WF, Duyk GM, Dymecki S, *et al*. (2004) The knockout mouse project – a comprehensive plan for placing knockouts of all mouse genes into the public domain. *Nature Genetics*, **36**, 921–924.

Auwerx J, Avner P, Baldock R, Ballabio A, Balling R, Barbacid M, Berns A, Bradley A, Brown S, Carmeliet P, *et al*. (2004) EUCOMM – the European dimension for the mouse genome mutagenesis programme. *Nature Genetics*, **36**, 925–927.

Branda CS and Dymecki SM (2004) Talking about a revolution: the impact of site-specific recombinases on genetic analyses in mice. *Developmental Cell*, **6**, 7–28.

Brown SD and Balling R (2001) Systematic approaches to mouse mutagenesis. *Current Opinion in Genetics & Development*, **11**, 268–273.

Brown SD and Hardisty RE (2003) Mutagenesis strategies for identifying novel loci associated with disease phenotypes. *Seminars in Cell & Developmental Biology*, **14**, 19–24.

Capecchi MR (1989) The new mouse genetics: altering the genome by gene targeting. *Trends in Genetics*, **5**, 70–76.

Carlson CM, Dupuy AJ, Fritz S, Roberg-Perez KJ, Fletcher CF and Largaespada DA (2003) Transposon mutagenesis of the mouse germline. *Genetics*, **165**, 243–256.

Carpinelli MR, Hilton DJ, Metcalf D, Antonchuk JL, Hyland CD, Mifsud SL, Di Rago L, Hilton AA, Willson TA, Roberts AW, *et al*. (2004) Suppressor screen in Mpl-/- mice: c-Myb mutation causes supraphysiological production of platelets in the absence of thrombopoietin signaling. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 6553–6558.

Chen Y, Yee D, Dains K, Chatterjee A, Cavalcoli J, Schneider E, Om J, Woychik RP and Magnuson T (2000) Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. *Nature Genetics*, **24**, 314–317.

Coghill EL, Hugill A, Parkinson N, Davison C, Glenister P, Clements S, Hunter J, Cox RD and Brown SD (2002) A gene-driven approach to the identification of ENU mutants in the mouse. *Nature Genetics*, **30**, 255–256.

Copeland NG, Jenkins NA and Court DL (2001) Recombineering: a powerful new tool for mouse functional genomics. *Nature Reviews. Genetics*, **2**, 769–779.

Feil R, Wagner J, Metzger D and Chambon P (1997) Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochemical and Biophysical Research Communications*, **237**, 752–757.

Floss T and Wurst W (2002) Functional genomics by gene-trapping in embryonic stem cells. *Methods in Molecular Biology*, **185**, 347–379.

Friedrich G and Soriano P (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes & Development*, **5**, 1513–1523.

Gossler A, Joyner AL, Rossant J and Skarnes WC (1989) Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science*, **244**, 463–465.

Gu H, Marth JD, Orban PC, Mossmann H and Rajewsky K (1994) Deletion of a DNA polymerase beta gene segment in T cells using cell type-specific gene targeting. *Science*, **265**, 103–106.

Hansen J, Floss T, Van Sloun P, Fuchtbauer EM, Vauti F, Arnold HH, Schnutgen F, Wurst W, von Melchner H and Ruiz P (2003) A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9918–9922.

Hardouin N and Nagy A (2000) Gene-trap-based target site for cre-mediated transgenic insertion. *Genesis*, **26**, 245–252.

Horie K, Yusa K, Yae K, Odajima J, Fischer SE, Keng VW, Hayakawa T, Mizuno S, Kondoh G, Ijiri T, *et al*. (2003) Characterization of sleeping beauty transposition and its application to genetic screening in mice. *Molecular and Cellular Biology*, **23**, 9189–9207.

Hrabe de Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto D, Marschall S, Heffner S, Pargent W, Wuensch K, Jung M, *et al*. (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genetics*, **25**, 444–447.

Izsvak Z, Ivics Z and Plasterk RH (2000) Sleeping beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of Molecular Biology*, **302**, 93–102.

Justice MJ (2000) Capitalizing on large-scale mouse mutagenesis screens. *Nature Reviews. Genetics*, **1**, 109–115.

Justice MJ, Noveroske JK, Weber JS, Zheng B and Bradley A (1999) Mouse ENU mutagenesis. *Human Molecular Genetics*, **8**, 1955–1963.

Kellendonk C, Tronche F, Casanova E, Anlag K, Opherk C and Schutz G (1999) Inducible site-specific recombination in the brain. *Journal of Molecular Biology*, **285**, 175–182.

Kwan KM (2002) Conditional alleles in mice: Practical considerations for tissue-specific knockouts. *Genesis*, **32**, 49–62.

Lewandoski M (2001) Conditional control of gene expression in the mouse. *Nature Reviews. Genetics*, **2**, 743–755.

Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, *et al*. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics*, **24**, 381–386.

Liu P, Jenkins NA and Copeland NG (2003) A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Research*, **13**, 476–484.

Melchner H and Stewart AF (2004) Engineering of ES cell genomes with recombinase systems. In *Handbook of Stem Cells*, Robert L (Ed.), Elsevier.

Mitchell KJ, Pinson KI, Kelly OG, Brennan J, Zupicich J, Scherz P, Leighton PA, Goodrich LV, Lu X, Avery BJ, *et al*. (2001) Functional analysis of secreted and transmembrane proteins critical to mouse development. *Nature Genetics*, **28**, 241–249.

Moser AR, Luongo C, Gould KA, McNeley MK, Shoemaker AR and Dove WF (1995) ApcMin: a mouse model for intestinal and mammary tumorigenesis. *European Journal of Cancer*, **31A**, 1061–1064.

Munroe RJ, Bergstrom RA, Zheng QY, Libby B, Smith R, John SW, Schimenti KJ, Browning VL and Schimenti JC (2000) Mouse mutants from chemically mutagenized embryonic stem cells. *Nature Genetics*, **24**, 318–321.

Muyrers JP, Zhang Y and Stewart AF (2001) Techniques: Recombinogenic engineering-new options for cloning and manipulating DNA. *Trends in Biochemical Sciences*, **26**, 325–331.

Nadeau JH (2001) Modifier genes in mice and humans. *Nature Reviews. Genetics*, **2**, 165–174.

Nagy A, Gertsenstein M, Vintersten K and Behringer R (2003) *Manipulating the Mouse Embryo*, Third Edition, Cold Spring Harbour Laboratory Press: Cold Spring Harbour, New York.

Nagy A and Mar L (2001) Creation and use of a Cre recombinase transgenic database. *Methods in Molecular Biology*, **158**, 95–106.

Niwa H, Araki K, Kimura S, Taniguchi S, Wakasugi S and Yamamura K (1993) An efficient gene-trap method using poly a trap vectors and characterization of gene-trap events. *Journal of Biochemistry (Tokyo)*, **113**, 343–349.

Nolan PM, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, Gray IC, Vizor L, Brooker D, Whitehill E, *et al*. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genetics*, **25**, 440–443.

Rodriguez CI, Buchholz F, Galloway J, Sequerra R, Kasper J, Ayala R, Stewart AF and Dymecki SM (2000) High-efficiency deleter mice show that FLPe is an alternative to Cre-loxP. *Nature Genetics*, **25**, 139–140.

Schnütgen F, Doerflinger N, Calleja C, Wendling O, Chambon P and Ghyselinck NB (2003) A directional strategy for monitoring Cre-mediated recombination at the cellular level in the mouse. *Nature Biotechnology*, **21**, 562–565.

Skarnes WC, Auerbach BA and Joyner AL (1992) A gene trap approach in mouse embryonic stem cells: the lacZ reported is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. *Genes & Development*, **6**, 903–918.

Skarnes WC, Moss JE, Hurtley SM and Beddington RS (1995) Capturing genes encoding membrane and secreted proteins important for mouse development. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6592–6596.

Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, Young SG, Ruiz P, Soriano P, Tessier-Lavigne M, *et al*. (2004) A public gene trap resource for mouse functional genomics. *Nature Genetics*, **36**, 543–544.

Soewarto D, Blanquet V and Hrabe de Angelis M (2003) Random ENU mutagenesis. *Methods in Molecular Biology*, **209**, 249–266.

Soriano P (1999) Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nature Genetics*, **21**, 70–71.

Stanford WL, Cohn JB and Cordes SP (2001) Gene-trap mutagenesis: past, present and beyond. *Nature Reviews. Genetics*, **2**, 756–768.

St Johnston D (2002) The art and design of genetic screens: drosophila melanogaster. *Nature Reviews. Genetics*, **3**, 176–188.

Thornton CE, Brown SD and Glenister PH (1999) Large numbers of mice established by in vitro fertilization with cryopreserved spermatozoa: implications and applications for genetic resource banks, mutagenesis screens, and mouse backcrosses. *Mammalian Genome*, **10**, 987–992.

Torres RM and Kühn R (1997) *Laboratory Protocols for Conditional Gene Targeting*, Oxford University Press: Oxford.

Toye AA, Moir L, Hugill A, Bentley L, Quarterman J, Mijat V, Hough T, Goldsworthy M, Haynes A, Hunter AJ, *et al*. (2004) A new mouse model of type 2 diabetes, produced by N-ethyl-nitrosourea mutagenesis, is the result of a missense mutation in the glucokinase gene. *Diabetes*, **53**, 1577–1583.

Utomo AR, Nikitin AY and Lee WH (1999) Temporal, spatial, and cell type-specific control of Cre-mediated DNA recombination in transgenic mice. *Nature Biotechnology*, **17**, 1091–1096.

Valenzuela DM, Murphy AJ, Frendewey D, Gale NW, Economides AN, Auerbach W, Poueymirou WT, Adams NC, Rojas J, Yasenchak J, *et al*. (2003) High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nature Biotechnology*, **21**, 652–659.

Vivian JL, Chen Y, Yee D, Schneider E and Magnuson T (2002) An allelic series of mutations in Smad2 and Smad4 identified in a genotype-based screen of N-ethyl-N-nitrosourea-mutagenized mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15542–15547.

Vreugde S, Erven A, Kros CJ, Marcotti W, Fuchs H, Kurima K, Wilcox ER, Friedman TB, Griffith AJ, Balling R, *et al*. (2002) Beethoven, a mouse model for dominant, progressive hearing loss DFNA36. *Nature Genetics*, **30**, 257–258.

Wiles MV, Vauti F, Otte J, Fuchtbauer EM, Ruiz P, Fuchtbauer A, Arnold HH, Lehrach H, Metz T, von Melchner H, *et al*. (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nature Genetics*, **24**, 13–14.

Witmer PD, Doheny KF, Adams MK, Boehm CD, Dizon JS, Goldstein JL, Templeton TM, Wheaton AM, Dong PN, Pugh EW, *et al*. (2003) The development of a highly informative mouse Simple Sequence Length Polymorphism (SSLP) marker set and construction of a mouse family tree using parsimony analysis. *Genome Research*, **13**, 485–491.

Wurst W, Rossant J, Prideaux V, Kownacka M, Joyner A, Hill DP, Guillemot F, Gasca S, Cado D, Auerbach A, *et al*. (1995) A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics*, **139**, 889–899.

Yang Y and Seed B (2003) Site-specific gene targeting in mouse embryonic stem cells with intact bacterial artificial chromosomes. *Nature Biotechnology*, **21**, 447–451.

Zambrowicz BP, Friedrich GA, Buxton EC, Lilleberg SL, Person C and Sands AT (1998) Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature*, **392**, 608–611.

Specialist Review

Systematic mutagenesis of nonmammalian model species

Marcel van den Heuvel and David Sattelle
University of Oxford, Oxford, UK

1. Genes, mutants, and large-scale screens

Genetics (*see* Article 37, **Functional analysis of genes**, Volume 3 and Article 41, **Mouse mutagenesis and gene function**, Volume 3) used to be driven by an urge to understand the structure of the genome. How parts of the genome, and hence genetic traits, were linked on a chromosome provided important clues, resulting in an improved description of the morphology of the genome. However, following the discovery of the building blocks of DNA and the inevitable urge to fully decode the "book of life", genetics shifted its emphasis toward addressing the roles of individual genes. Genes had been shown to encode single proteins, each of which conferred a quite distinct function (e.g., an enzyme catalyzing a particular step in a metabolic pathway). Drawing on these basic findings, a new approach was developed to search for genes, which are required for a particular biological process. This necessitates a new concept of associated genes, based not on their location, but on their function instead. Hence, if we now wish to find all the genes involved in a particular process, our search for sets of functionally linked genes should cover the entire genome. To saturate the whole genome with mutants first requires mutagenizing, and then screening large numbers of animals. The extensive use of this approach has brought to center stage in biology two invertebrate animal species, the fruit fly, *Drosophila melanogaster* and the nematode worm, *Caenorhabditis elegans*. Their widespread use has less to do with their animal specifics and more with the ease and history of their use in the laboratory. The most important factor has been the ability to raise, study, and genetically characterize very large numbers of animals, each unique in its genetic composition. Using these animal "models", functional genetics in metazoan animals was initiated, leading to many large-scale functional mutagenesis screens over the last 30 years.

2. The fly and the worm: genetic model organisms used widely in mutagenesis studies

Each animal model has its genetic advantages as well as its limitations. For instance, Drosophila genetic screens for mutations affecting particular processes

are helped by the existence of so-called balancer chromosomes. These are the results of extensive genetic-building work to create functional chromosomes, which are however “completely” scrambled. Such chromosomes are useful to suppress recombination between homologous chromosomes and are therefore used in flies to stabilize chromosomes. In practice, this means that it is possible to keep a recessive mutation present in a genetic background (a specific chromosome) in a heterozygous, balanced state (the balancers also usually contain a lethal recessive mutation) over many (indefinite) generations. In addition, most of the balancer chromosomes also contain a dominant and several recessive markers, allowing the researcher to easily follow segregation of chromosomes, thereby removing the need for extensive testing to verify genetic background in a particular experiment. Drosophila, however, also suffers from some disadvantages; it has been difficult to generate a gene targeting approach (Rong *et al*., 2002). If one finds and wants to knock out a particular gene, the approach, therefore, still usually follows a random screen and genetic mapping. In addition, Drosophila represents an evolutionary highly developed insect, and thus might not be a good overall model for animal biology (Tautz, 2004). Animals that have been placed at evolutionary branchpoints may be more helpful in this aspect.

C. elegans, like Drosophila, has its limitations; it is an effective genome model for other rhabditid nematodes but sometimes comparisons with its more distant cousins among the nematodes can present problems, making further distant comparisons sometimes impossible. For example, there are a number of nematode genes that are absent from *C. elegans* and it appears that gene loss has played a key role in *C. elegans* genome evolution (Parkinson *et al*., 2004). Nevertheless, this 1-mm long, “simple” free-living nematode worm has an extremely rapid generation time (3 days at room temperature). The short life cycle, together with its hermaphroditic lifestyle, facilitate the maintenance and study of genetic strains. The consistency of development, the transparency of the worm, together with the relatively small number of cells, have made the first complete description of the cell lineage for an entire organism possible (Sulston *et al*., 1983; Sulston and Horvitz, 1977). The hermaphrodite has 558 cells at the first larval stage, rising to 959 in the adult. The adult nervous system has 302 neurons, and the “‘wiring diagram” of its 5000 synapses has been determined in exquisite detail (White *et al*., 1986). In *C. elegans*, the genetic map is anchored to (1) a physical map of the six chromosomes based on a combination of cosmid and yeast artificial chromosome (YAC) clones and (2) the entire genome sequence. Ready access to this outstanding resource has facilitated the analysis of gene function.

3. History

Gregor Mendel by publishing his “Versuche über Pflanzen-Hybriden” in 1865 provided an experimental demonstration that traits can be followed through subsequent generations, introducing, via plant breeding, concepts fundamental to what we now know as genetics. Thomas Hunt-Morgan and his followers took up this concept and for the first time used Drosophila to understand animal genetics. He can be regarded as the father of fly genetics. Sadly, one of the few remaining

descendants of his school of research, Ed Lewis, has recently died. He was awarded the Nobel Prize in 1995 for his work on the genetics of the *bithorax* complex of Drosophila genes and shared it with two other fly scientists, Christiane Nüsslein-Volhard and Eric Wieschaus. The mutants uncovered in the *bithorax* region were also perhaps the most extravagant example of a gene (or gene complex) leading to a clear trait: turning the fly's balance organs into wings (Lewis, 1978). The results had the superficial appearance of transporting Drosophila back in time, "creating" an evolutionarily ancient four-winged insect. This finding exemplifies the link between genes and evolution. Lewis experimented extensively with mutagenesis and showed that X-ray radiation cause chromosome deletions in Drosophila at any level. This work alerted the US government that radiation has no lower limit of effect (Scott and Lawrence, 2004). The two scientists with whom Ed Lewis shared his price set out to do an experiment that proved to be the jewel in the crown for Drosophila – to find all the genes involved in making a fly.

By the early 1960s, the central dogma of molecular biology (DNA makes RNA makes protein) was established and Sydney Brenner, who contributed so much to that core of knowledge, decided that a new experimental organism was required for a genetic approach to the organizing principles of the nervous system and embryonic development. He chose the nematode *C. elegans.* Among his criteria for selecting the worm were that it should be the simplest organism with traits of interest. Another was ease of manipulation. The first worm mutagenesis experiment was carried out in 1967, and in 1974 the results of the first characterization of ~100 genes were published (Brenner, 1974). This landmark publication was the basis for many subsequent mutagenesis experiments and screens. John Sulston and Bob Horvitz went on to determine the entire cell lineage of the worm (Sulston *et al*., 1983; Sulston and Horvitz, 1977) and in 1998 John Sulston, Alan Coulson, and colleagues at the Wellcome Institute Sanger Centre, UK, together with Bob Waterston and colleagues in the United States published the first draft sequence of the entire *C. elegans* genome, the first genome of a multicellular organism to be sequenced. Brenner, Sulston, and Horvitz shared the 2002 Nobel Prize in Physiology or Medicine for their discoveries.

4. Mutants: how to make a fly

Christiane Nüsslein-Volhardt and Eric Wieschaus took the idea of functional groupings of genes to a higher level by assuming that a set number of genes would be required to guide and steer the formation and development of the fly embryo. Perhaps stimulated by the findings of Ed Lewis on the *bithorax* complex, they saturated the whole fly genome with mutants, chromosome by chromosome, extracting and stabilizing any mutant that led to a recessive embryonic lethal endpoint. How this can be done is shown in Figure 1(a). They screened through the 1st (X), 2nd, 3rd, and 4th chromosomes in this way, screening in total probably around 40 000 lines, and isolated approximately 150 genetic complementation groups, each identifying several alleles in at least one gene involved in some aspect of embryonic development (Jürgens *et al*., 1984; Nüsslein-Volhard *et al*., 1984; Wieschaus *et al*., 1984). During the course of this work, they published

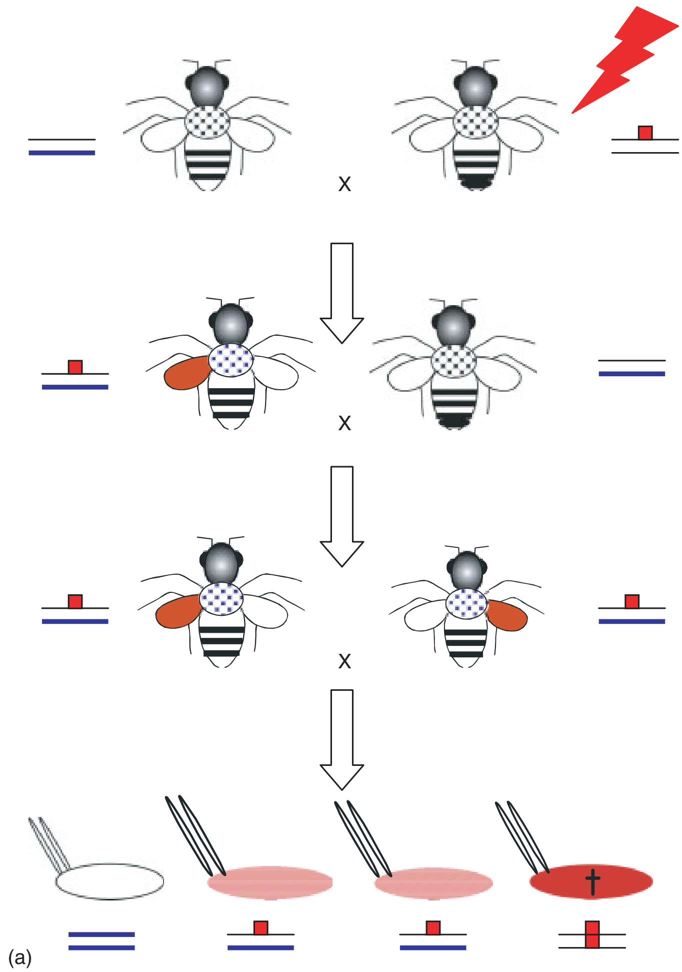


Figure 1 (a) Schematic outline of a simple F2 mutagenesis screen in Drosophila used to search for embryonic lethal recessive mutations. The male fly on the right is mutagenized (in red). The amount of mutagen is adjusted such that on average one hit per genome is reached. This is indicated as a single red box on one of this animal's chromosomes (both copies of a single chromosome are drawn). The male is crossed to a female with a balancer chromosome in her genetic make up, en masse (indicated as a blue line, with a "normal" chromosome as pair, black line). The mutagenized male chromosomes are thus balanced in the next generation (shown as a pair of a red box chromosome and a blue chromosome). The siblings for each of putative mutant chromosomes are intermated to create a stock (heterozygous for the original genetic defect), as well as generating offspring that can be studied. This is done initially as single pair crosses to stabilize the mutant background; the number of such (successful) crosses represents the final number of chromosomes screened. One-quarter of the offspring will be homozygous mutant (in red, shown as a pair of red boxed chromosomes) and should show the recessive phenotype (and if the phenotype is recessive lethal, die). (b) Schematic outline of a simple F2 screen in the nematode worm *Caenorhabditis elegans*. Wild-type worms are treated with the mutagen ethyl methanesulfonate (EMS). One-quarter of the F2 progeny will be homozygous for the mutated gene. Candidate mutant animals are isolated and cloned. The short life cycle of the worm, means that within very few days its F3 progeny can be examined to determine whether the candidate mutant breeds true

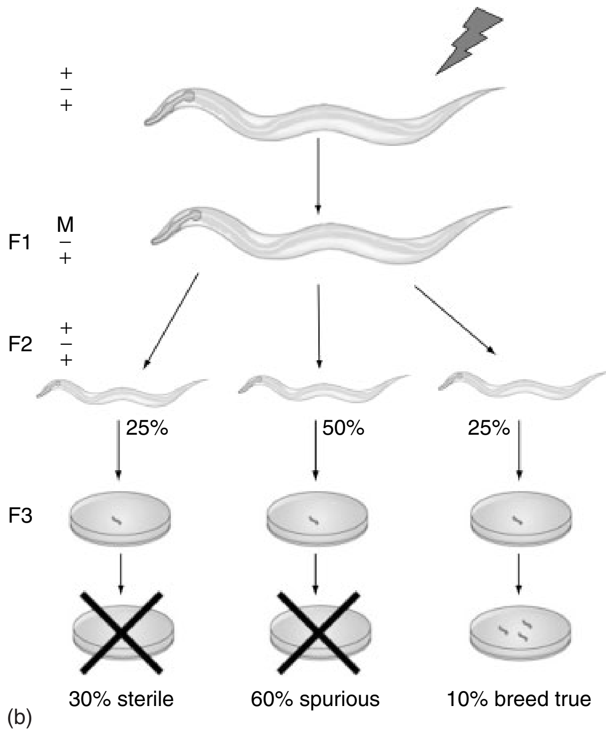


Figure 1 (*continued*)

a paper describing some remarkable findings (Nüsslein-Volhard and Wieschaus, 1980): segmentation in the fly appeared to be dependent on a simple building block pattern of gene activity. A first set of genes is required to assemble in large blocks, consecutive segments, covering the whole of the embryo. These large blocks in the thorax and abdomen of the future larvae are then again subdivided by another large group of genes that determine every other segment. Finally, there is a set of genes that seems to be required for each segment, mutants of which show segmentally repeated defects (Figure 2). Now, we may be used to the idea of the simplicity with which development takes its course, but it was an exciting and very surprising finding at that time.

However, this was not all that made these mutagenesis screens so important for the rest of the genetics and development community. With the advent of molecular biology, Drosophila played a defining role in the cloning of genes from its genome. In fact, possibly the first example of the route now almost standard of isolating a mutant, mapping the genome location and finding and cloning the gene, was not one of the mutants isolated in the large screens initiated by Jürgens, Wieschaus, and Nüsslein-Volhard but instead was a mutant isolated by Ed Lewis, *antennapedia* (*ant*) (Scott *et al.*, 1983). The cloning of the *ant* gene was also a primary example of clear evolutionary homology, a small protein domain in the encoded Antennapedia protein appeared to be homologous to a known bacterial DNA binding domain, now known as the homeo domain (Laughon and Scott,

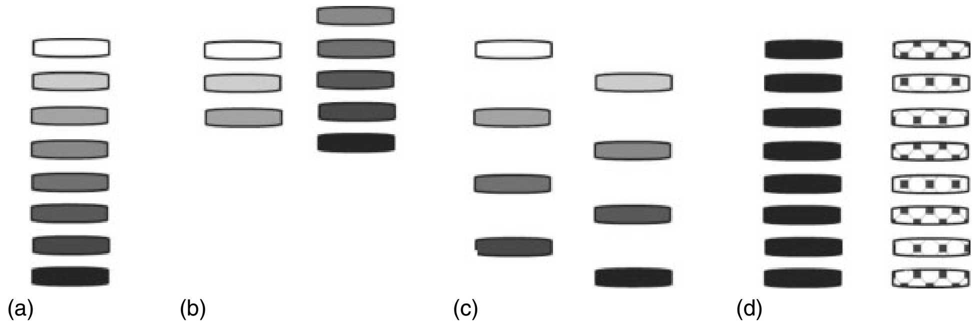


Figure 2 Schematic representation of mutant phenotypes of late embryos (cuticles) as described by Nüsslein-Volhard and Wieschaus (1980). (a) Normal cuticle pattern of abdominal segments, showing belts of denticles as boxes. Each belt in anterior to posterior direction has a unique character. (b) Phenotypes as observed in two hypothetical gap mutants, showing deletions of consecutive segments. (c) Phenotypes as observed in hypothetical pair rule mutants, showing deletions of every other segment. (d) Phenotypes as observed in hypothetical segment polarity mutants, each segment is altered in a similar manner (although the denticle belts still show differences in anterior to posterior direction across the animal)

1984). This was one of the first examples where a gene shown to function as a direct driver of a developmental process was evolutionarily conserved. It was certainly not the last.

Many of the genes isolated as mutants in the mutagenesis screens performed by Nüsslein-Volhard and Wieschaus have over the last decades proven to represent a set of evolutionarily ancient genes that are required to set up embryonic development in all animal phyla (and sometimes even in other phyla too). The fly thus has proven to be a very rich source of developmental genes, and almost all have been found through mutagenesis screens as described above. However, this is not the only contribution the fly has made to understanding developmental genes. Once a gene has been isolated and a mutant generated, this opens the door to functional analysis. Normally, such a mutant will have been isolated as part of a group of mutants leading to similar phenotypes (see Figure 2), and thus the gene can often be placed within a functional system. Further epistasis analysis, generating genetic combinations, then places the gene within the genetic (and often clear molecular) hierarchy in each group. The efforts to understand the genetic hierarchy leading to a patterned fly embryo, based on mutant screens has led to the isolation and functional characterization of large sets of genes and their encoded proteins, often enabling clear predictions regarding biochemical interactions. The processes driven by such gene hierarchies might not always be directly comparable from Drosophila to higher animals, such as human, and indeed evolutionarily they should not be but much can be learned from the analysis in flies (see (Tautz, 2004). Mutagenesis screens as described here had no prior genetic set up except to create a defined “normal” background amenable to mutagenesis and screening. They have been very useful in defining many processes and usually also resulted in a range of alleles for each gene. Refinement of the genetic background can however further assist in the recovery of more specific mutants and genes.

5. Mutants: what do you see?

The animal eye lends itself well to mutagenesis screens since, in most cases, visual abilities are not required in a laboratory environment. However comparing mutants and genes isolated from screens for mutants specifically affecting the fly eye, with genes now known to be required for eye development, has shown clearly that a large set of genes is required for eye development, and many of these are not represented by mutants with eye defects. Some of these genes are thought to be required at earlier steps in embryonic development (Cagan and Ready, 1989) and mutations would lead to severe developmental defects, that is, no fly, no eye. Thus, although the eye seems a perfect organ to screen for genes that could have a potential role in, for instance, human eye anomalies, it has not been easy to find these. However, an elegant deployment of the late differentiating Drosophila eye has provided a very useful model.

The tissue from which the eye of the fly develops remains a naïve field of cells until late in the larval stage when a front of differentiation sweeps through the eye field to generate the stereotype insect compound eye pattern (Tomlinson, 1985). It is thought therefore that the generation of the different cells within each unique entity of the compound eye is dependent on local cell interactions (Ready *et al*., 1976). Through traditional behavioral screens, a mutant in one of the neuronal, light-sensitive cells (in this case UV sensitive) had been isolated, called *sevenless* (*sev*) (the 7th omatidial cell, out of a total of eight cells, disappears, hence the name) (Harris *et al*., 1976). The gene was cloned, sequenced, and its expression localized (Hafen *et al*., 1987; Tomlinson *et al*., 1987). It appeared that the *sev* gene encoded a tyrosine kinase receptor. Such proteins were known as important in oncogenesis in humans and had been studied in cell culture. The appearance of an onco-protein in a developmental decisions process was seen as a prime example of a signaling pathway determining a cell's fate. Key questions remaining were: how does it function, what is its ligand and how does it signal to induce the R7 fate? Extensive screens for other genes leading to mutant phenotypes only affecting the R7 cell did not result in new mutants. To try and find these, a screen was designed using what we now know as a sensitized genetic background. Instead of looking for mutants in as clean and normal a genetic background as possible, the principle of these screens is to create a background already on the brink of failure within a very specific process. Subsequently, introducing mutations into such a background, for example by only reducing the activity of a gene by half (heterozygous), could tip the balance over and lead to a phenotype. Such screens are therefore based on dominant suppression or enhancement of a subtle phenotype. The mutagenesis screen as set up and performed by Simon *et al*. (1991) introduced a very useful concept into screening for specific biological processes in the fly and one that has been repeated many times (Haines and van den Heuvel, 2000). The screen in the fly eye using *sev* as well as modified allelic varieties of *sev*, resulted in the isolation of mutants encoding downstream components (Simon *et al*., 1991). Some of these genes had already been isolated as putative oncogenes in higher animals, such as Ras (for review see Varmus, 1984), again indicating that flies possess mechanisms for driving differentiation and proliferation, very similar to those of higher animals. In addition, further mutagenesis work led to the addition of new functional assays

designed to understand Ras function, through for instance the isolation of novel mutants in this signalling pathway and new functional interactions characterized (Gaul *et al*., 1992; Hariharan *et al*., 1991; Simon *et al*., 1991; Wolff and Ready, 1991), as well as leading to the isolation of downstream effectors (Brunner *et al*., 1994a; Brunner *et al*., 1994b; Chang *et al*., 1995; Lai and Rubin, 1992). The isolation of these mutants and characterization of the homozygous phenotype did confirm the prediction: most are homozygous embryonic lethal.

6. Mutants: discovering where drugs act

Most mutants of *C. elegans* have been generated using the chemical mutagen, EMS (ethyl methane sulphonate) to generate mutations in sperm and oocytes. Many of the mutations generated in Brenner's original screen were visible recessive mutations, identified in a simple F2 screen (Brenner, 1974) (Figure 1b). Worms with mutant phenotypes are then transferred to new plates to test whether the phenotype is transmissible to the next generation. In a typical screen 12 000 copies of any particular gene are assayed. The frequency of recovery of mutations is about one in 2000 copies of the gene. Using this approach, Brenner identified 619 mutants with visible phenotypes. An excellent account of *C. elegans* genetic screens is given by (Jorgensen and Mango, 2002).

An example of the utility of a chemistry-to-gene screen is that initiated by Brenner using levamisole, an antiparasitic drug (Brenner, 1974). Levamisole is a broad-spectrum antiparasitic drug widely used to eradicate roundworm infestations in livestock and humans (World Health Organization **http://www.who.int/topics/en/**). *C. elegans* resistant to 100-μM levamisole were characterized by their ability to migrate across an agar plate containing the drug, more quickly than sensitive wild type worms and it was noted that some had characteristic phenotypes such as uncoordinated movement or twitching. Brenner surmised that levamisole acts as an acetylcholine (ACh) agonist since its effect resembles the paralysis produced by the acetylcholinesterase (AChE) inhibitor, lannate. Later pharmacological studies on cut worms by Lewis *et al*. (1980) showed that levamisole-resistant mutants lack functional muscle nicotinic acetylcholine receptors (nAChRs) and the loss of high-affinity levamisole binding in two mutants (*unc-50* and *unc-74*) led to the notion that the affected genes encode for proteins involved in levamisole receptor processing and assembly (Lewis *et al*., 1987).

Now, nearly all of the 12 genes that mediate levamisole resistance have been identified. Five encode the nAChR subunits, UNC-29, LEV-1, UNC-38, UNC-63, and LEV-8 (Fleming, 1997; Culetto *et al*., 2004; Towers *et al*., 2005). *C. elegans* possesses one of the most extensive and diverse nAChR gene families known, consisting of at least 27 subunits that have been divided into subgroups based on sequence homology (Jones and Sattelle, 2004). As expected from the hypercontraction induced by levamisole, all the five nAChR subunits identified from this chemistry-to-gene screen are expressed in body wall muscle. Recent functional studies show that these subunits are important components of levamisole-sensitive nAChRs (Richmond and Jorgensen, 1999; Culetto *et al*., 2004; Towers *et al*., 2005).

The other levamisole-resistance genes appear to encode molecular components associated with nAChR assembly and function. These include UNC-50, a novel transmembrane protein, the mammalian homolog of which, UNCL, is an inner nuclear membrane RNA binding protein that increases cell-surface expression of vertebrate nAChRs ($\alpha 4/\beta 2$) when coexpressed with them in *Xenopus laevis* oocytes and COS cells (Fitzgerald *et al*., 2000). The *lev-10* gene has shown to encode a transmembrane protein required for postsynaptic aggregation of nAChRs at the neuromuscular junction (Gally *et al*., 2004). LEV-10 interacts extracellularly with either nAChR subunits, or nAChR-associated proteins. Other levamisole resistance genes regulate signal transduction downstream of nAChR activation. UNC-68 is the ryanodine receptor, which is expressed in body-wall muscles and is necessary for normal locomotion. It plays a role in intracellular Ca^{2+} regulation and is also a validated pesticide target site (Maryon *et al*., 1996). The *unc-22* and *lev-11* genes, respectively encode twitchin, which contains fibronectin typeIII, immunoglobulin, and protein kinase domains (Benian *et al*., 1993) and tropomyosin, which regulates muscle contraction (Kagawa *et al*., 1997).

Thus, a chemistry-to-gene screen using levamisole not only identifies a subset of nAChRs from a very large family that are targeted by an antiparasitic drug, but also provides insights into cellular pathways linked with nAChR function. Among the gene products acting upstream and downstream of the drug target is another validated pesticide target. Thus, a chemical based screen thereby demonstrating that this approach not only uncovers the target of a known chemical but can also be used to identify new drug targets. Furthermore, as demonstrated by studies on UNCL, such screens have the potential to shed light on the function of human genes.

Other screens in use in *C. elegans* include modifier (enhancer suppressor screens) screens that search for second site mutations that either exacerbate or ameliorate the phenotype of the first mutation (such as described above for Drosophila). Multigenerational, microscopy-based and laser ablation screens have been pursued successfully in the worm as well as screens for lethal mutants (see Jorgensen and Mango, 2002).

7. Genetics in forward and reverse gears

The worm has about 20 000 genes fewer than 10% of which have been identified by mutation. Of these, 575 have been cloned and characterized at the molecular level. About 70% have a mammalian counterpart and about 50% of human disease genes have a *C. elegans* equivalent (Culetto and Sattelle, 2000), hence the interest in further understanding the remaining genes. Genes that are specific to nematodes are also of interest both from an evolutionary perspective and from a practical viewpoint in designing new, safer antiparasitic drugs (2.9 billion people have parasitic nematode infections, there are substantial losses in livestock and crop damage worldwide is estimated at $80 billion). Genetics can move forwards through the mapping of a novel phenotype to discover the gene(s) involved and then addressing function. A genome-wide mutagenesis program is under way in *C. elegans* with the aim of generating a deletion mutant for every gene.

Now that we have the complete genome sequence for fly (Myers *et al*., 2000), worm and other organisms too, we can also deploy to powerful effect a genetic reverse gear (*see* Article 44, **The *C. elegans* genome**, Volume 3 and Article 45, **The *Drosophila* genome(s)**, Volume 3). An example of the new reverse genetics can be illustrated in the genomewide RNA interference (RNAi) screens carried out by, for example, the Ahringer (Kamath *et al*., 2003) and Plasterk (Simmer *et al*., 2003) laboratories. This work was based on the original breakthrough by Fire *et al*. (1998) who showed that introducing double-stranded RNA (dsRNA) for a particular gene into worms resulted in the knockdown of the function of that particular gene. Such high-throughput screens in *C. elegans* are facilitated by the finding that the dsRNA can be delivered to worms via the noninfective *E. coli* strain of bacteria on which they feed. The results from the large-scale screens covering 86% of the genome showed that 1722 (10.3%) showed phenotypes. Rerunning the screens with particular mutants hypersensitive to RNAi is now providing access to new genes (Simmer *et al*., 2003).

8. Genomes and mutagenesis

The completion of the Drosophila genome sequence has led to some new as well some modified mutagenesis protocols and of course has sped up the cloning of genes underlying already mapped mutants. The increasing evidence that many genes are required throughout the development and life of an animal at various places and at different times indicated that often gene knock out would not lead to a full spectrum of this gene's function. To bypass the early functions of genes, experimental strategies described above can be used but another way of bypassing earlier gene requirements is to make cells homozygous mutant for such a gene only at certain times and/or in selected places. Somatic homozygous clones can be induced by recombination for instance as a results of chromosome insult. However to regulate this, increase efficiency and localization, yeast recombination sites (and the required enzymatic activity) were introduced at fixed places through out the fly genome, the so-called FLP-FRT system (Chou, 1992; Ryder *et al*., 2004; Xu and Rubin, 1993) (Figure 3a). The end result of these experiments is a pool of heterozygous cells within which homozygous mutant (and homozygous marker expressing cells) clones are found. This technique is very well placed for analyzing the effects of signalling pathways in fields of cells.

However, recombination as induced by this system can also be used to drive the generation of chromosomal deletion stocks. By generating large numbers of individual insertions of the recombination enzyme (FLP) recognition site (FRT) across the genome, intrachomosomal recombination can be induced which leads to the loss of the chromosome area in between two FRT sites present in cis on the chromosome. Random FRT sites have now been mapped to their chromosomal location by sequence analysis and mapping onto the genome sequence. It is possible therefore to create a custom-built deletion across a defined chromosome area (see Parks *et al*., 2004 and **http://www.drosdel.org.uk/**) using this system.

The FLP-FRT lines are all based on the use of the P-element system to generate fly transgenics (Spradling and Rubin, 1982). Insertional mutagenesis has been used

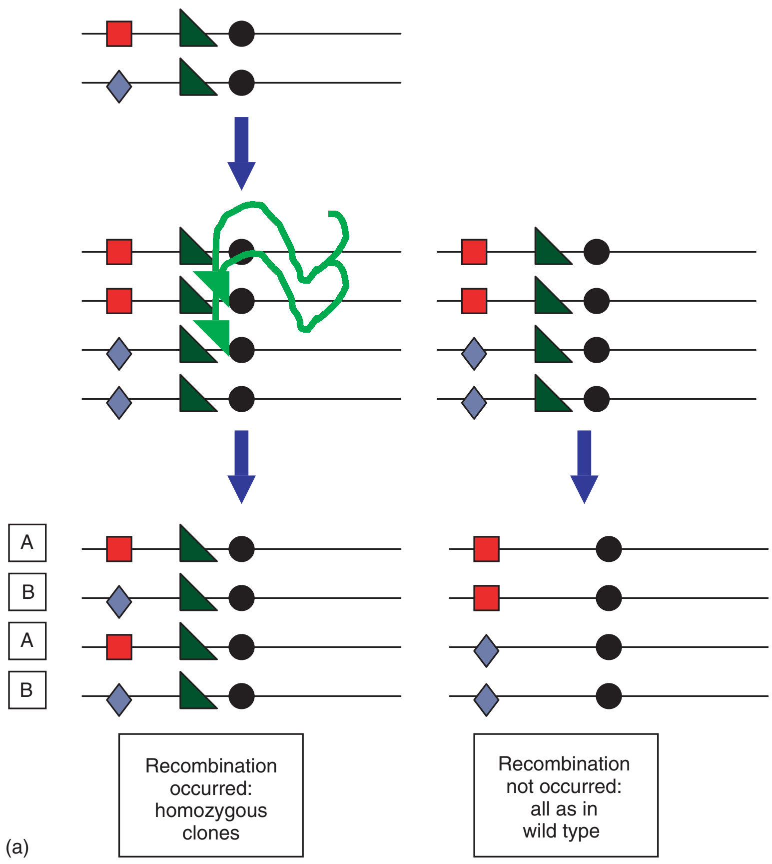


Figure 3 (a) Schematic outlining use of FRT-FLP system to generate clones. In a heterozygous mutant background (as represented by a pair of chromosomes, one carrying a mutation: red box, the other expressing a blue marker gene), where each of the arms on which your mutant allele is present, carry, close to the centromere (black circle), a recognition site for a yeast derived recombinase enzyme (Flippase, FLP); these sites are represented by a green triangle and also known as FRT sites. When the DNA is duplicated (cell undergoing division), a pulse of Flippase activity would lead to recognition of the FRT sites by Flippase (light green triangle with line) and induce recombination between the two chromosomes (shown on left). If the cells undergo division but no Flippase activity is present, only random recombination would take place (very rare). In either case, the chromosomes will be separated according to the normal pattern, where A and A are joined (and B and B). In the case where no recombination has taken place, this leads to two daughter cells with the same genetic build-up as the parent cell (shown on right). If recombination has taken place, the two daughter cells are different, one being homozygous for your mutation, the other homozygous for the marker. (b) Schematic outlining use of UAS-GAL4 system in flies. The Gal transcription factor protein recognizes very specifically, a binding site known as UAS (Upstream Activating Sequence). When this site is cloned in front of Your Favorite Gene (YFG), expression of GAL4 protein in a cell will turn on transcription of YFG and lead to production of product. To use this system efficiently, large numbers of GAL4 driver stocks have been generated, each expressing GAL4 in different places and at different times during the fly life cycle. These can be crossed to your animal transgenic for the UAS-YFG construct. In this way, general lethality by expression of YFG can be overcome since only progeny will be affected and the cross can be set up endlessly. In addition, highly specific cells can be targeted by using GAL4 lines expressed in such cells only. Further sophistication to the system has been added by using a GAL4 inhibitor, GAL80

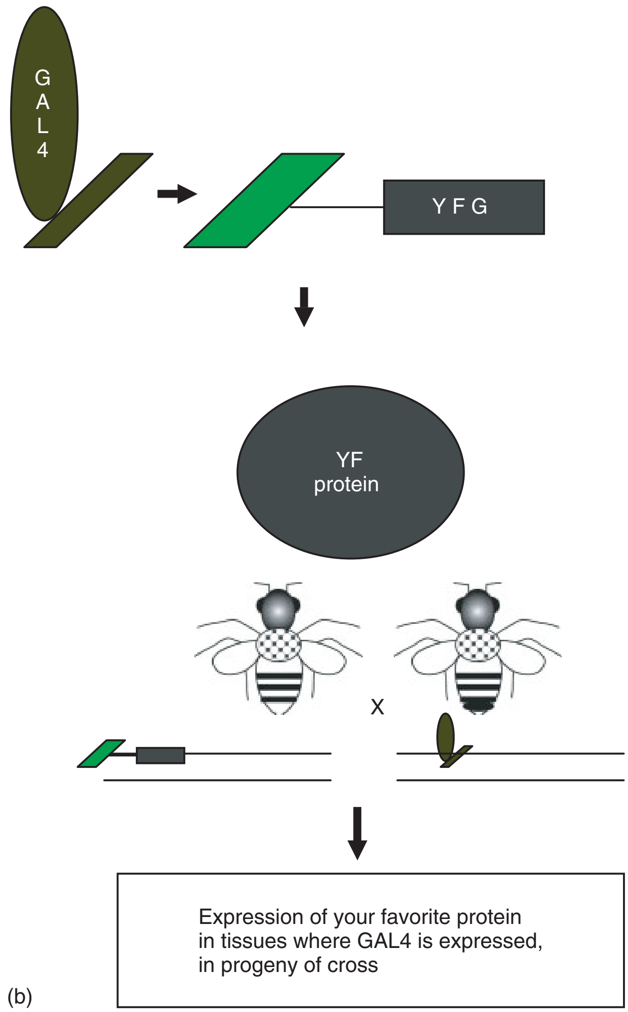


Figure 3 (*continued*)

extensively in Drosophila to target genes, generate gene expression tags as well as dominant enhancer traps (e.g., Kania *et al*., 1995; Kelso *et al*., 2004; Rio and Rubin, 1985; Rørth *et al*., 1998). Often such insertional elements are not fully random as to choice of insertion site, and new elements with either more random insertion preferences or just with altered choices have been developed (Bonin and Mann, 2004). All of such methods are however not specific to flies, although they are highly developed and often intricately linked to specific scientific questions. General methods of screening and analyzing gene function such as RNAi have also been applied in flies, mostly in cells (Kiger *et al*., 2003). However, new approaches by generating inducible constructs (UAS-GAL4 system, Brand and Perrimon, 1993) (Figure 3b) encoding hairpin loop mRNA stretches for a particular gene have been tested.

9. General conclusion and prospects

As the few examples given here illustrate, mutagenesis coupled with carefully designed screens of various kinds has proved extremely productive in ascertaining gene functions and in discovering the components of pathways involving the products of many genes. We have chosen examples of how such approaches have contributed to developmental biology and neurobiology but many other examples could have been selected. In the future, forward genetic screens will continue to be important especially when combined with expression profiling via DNA microarrays, the expanded use of gene reporters, RNAi, and the increasing use of interactome maps summarizing protein–protein interactions. The use of sensitized screens in both model organisms seems set to increase. A major challenge for the future will be addressing the functions of those genes (estimated as approximately two-thirds of the *C. elegans* genes) for which it is estimated that no visible, lethal, or sterile phenotype can be generated. As outlined in this chapter, the ability to look randomly for genetic aberrations that influence a biological process, without any preconceptions, is important and has led to a wealth of new discoveries.

Further reading

Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, *et al*. (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature*, **408**, 331–336.

References

Benian GM, L'Hernault SW and Morris ME (1993) Additional sequence complexity in the muscle gene, *unc-22*, and its encoded protein, twitchin, of *Caenorhabditis elegans*. *Genetics*, **134**, 1097–1104.

Bonin CP and Mann RS (2004) A piggyBac transposon gene trap for the analysis of gene expression and function in Drosophila. *Genetics*, **167**, 1801–1811.

Brand AH and Perrimon N (1993) Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, **118**, 401–415.

Brenner S (1974) The genetics of *Caenorhabditis elegans*. *Genetics*, **77**, 71–94.

Brunner D, Ducker K, Oellers N, Hafen E, Scholz H and Klambt C (1994a) The ETS domain protein pointed-P2 is a target of MAP kinase in the sevenless signal transduction pathway. *Nature*, **370**, 386–389.

Brunner D, Oellers N, Szabad J, Biggs WH III, Zipursky SL and Hafen E (1994b) A gain-of-function mutation in Drosophila MAP kinase activates multiple receptor tyrosine kinase signaling pathways. *Cell*, **76**, 875–888.

Cagan RL and Ready DF (1989) Notch is required for successive cell decisions in the developing Drosophila retina. *Genes and Development*, **3**, 1099–1112.

Chang HC, Solomon NM, Wassarman DA, Karim FD, Therrien M, Rubin GM and Wolff T (1995) phyllopod functions in the fate determination of a subset of photoreceptors in Drosophila. *Cell*, **80**, 463–472.

Chou TB and Perrimon N (1992) Use of a yeast site-specific recombinase to produce female germline chimeras in Drosophila. *Genetics*, **131**, 643–653.

Culetto E, Baylis HA, Richmond JE, Jones AK, Fleming JT, Squire MD, Lewis JA and Sattelle DB (2004) The *Caenorhabditis elegans* unc-63 Gene encodes a levamisole-sensitive nicotinic acetylcholine receptor alpha subunit. *Journal of Biological Chemistry*, **279**, 42476–42483.

Culetto E and Sattelle DB (2000) A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes. *Human Molecular Genetics*, **9**, 869–877.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Fitzgerald J, Kennedy D, Viseshakul N, Cohen BN, Mattick J, Bateman JF and Forsayeth JR (2000) UNCL, the mammalian homologue of UNC-50, is an inner nuclear membrane RNA-binding protein. *Brain Research*, **877**, 110–123.

Fleming JT (1997) *Caenorhabditis elegans* levamisole resistance genes *lev-1*, *unc-29*, and *unc-38* encode functional acetylcholine receptor subunits. *Journal of Neuroscience*, **15**, 5843–5857.

Gally C, Eimer S, Richmond JE and Bessereau JL (2004) A transmembrane protein required for acetylcholine receptor clustering in *Caenorhabditis elegans*. *Nature*, **431**, 578–582.

Gaul U, Mardon G and Rubin GM (1992) A putative Ras GTPase activating protein acts as a negative regulator of signaling by the Sevenless receptor tyrosine kinase. *Cell*, **68**, 1007–1019.

Hafen E, Basler K, Edstroem JE and Rubin GM (1987) Sevenless, a cell-specific homeotic gene of Drosophila, encodes a putative transmembrane receptor with a tyrosine kinase domain. *Science*, **236**, 55–63.

Haines N and van den Heuvel M (2000) A directed mutagenesis screen in Drosophila melanogaster reveals new mutants that influence hedgehog signaling. *Genetics*, **156**, 1777–1785.

Hariharan IK, Carthew RW and Rubin GM (1991) The Drosophila roughened mutation: activation of a rap homolog disrupts eye development and interferes with cell determination. *Cell*, **67**, 717–722.

Harris WA, Stark WS and Walker JA (1976) Genetic dissection of the photoreceptor system in the compound eye of *Drosophila melanogaster*. *Journal of Physiology*, **256**, 415–439.

Jones AK and Sattelle DB (2004) Functional genomics of the nicotinic acetylcholine receptor gene family of the nematode, *Caenorhabditis elegans*. *Bioessays*, **26**, 39–49.

Jorgensen EM and Mango SE (2002) The art and design of genetic screens: *Caenorhabditis elegans*. *Nature Reviews. Genetics*, **3**, 356–369.

Jürgens G, Wieschaus E, Nüsslein-Volhard C and Kluding H (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. II. Zygotic loci on the third chromosome. *Wilhelm Roux's Archives of Development Biology*, **193**, 283–295.

Kagawa H, Takuwa K and Sakube Y (1997) Mutations and expressions of the tropomyosin gene and the troponin C gene of *Caenorhabditis elegans*. *Cell Structure and Function*, **22**, 213–218.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, *et al*. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.

Kania A, Salzberg A, Bhat M, D'Evelyn D, He Y, Kiss I and Bellen HJ (1995) P-element mutations affecting embryonic peripheral nervous system development in Drosophila melanogaster. *Genetics*, **139**, 1663–1678.

Kelso RJ, Buszczak M, Quinones AT, Castiblanco C, Mazzalupo S and Cooley L (2004) Flytrap, a database documenting a GFP protein-trap insertion screen in Drosophila melanogaster. *Nucleic Acids Research*, **32**, D418–D420.

Kiger AA, Baum B, Jones S, Jones MR, Coulson A, Echeverri C and Perrimon N (2003) A functional genomic analysis of cell morphology using RNA interference. *Journal of Biology (Online)*, **2**, 27.

Lai ZC and Rubin GM (1992) Negative control of photoreceptor development in Drosophila by the product of the yan gene, an ETS domain protein. *Cell*, **70**, 609–620.

Laughon A and Scott MP (1984) Sequence of a *Drosophila* segmentation gene: protein structure homology with DNA binding proteins. *Nature*, **310**, 25–31.

Lewis EB (1978) A gene complex controlling segmentation in *Drosophila*. *Nature*, **276**, 565–570.

Lewis JA, Elmer JS, Skimming J, McLafferty S, Fleming J and McGee T (1987) Cholinergic receptor mutants of the nematode *Caenorhabditis elegans*. *Journal of Neuroscience*, **7**, 3059–3071.

Lewis JA, Wu CH, Levine JH and Berg H (1980) Levamisole-resistant mutants of the nematode Caenorhabditis elegans appear to lack pharmacological acetylcholine receptors. *Neuroscience*, **5**, 967–989.

Maryon EB, Coronado R and Anderson P (1996) unc-68 encodes a ryanodine receptor involved in regulating *C. elegans* body-wall muscle contraction. *The Journal of Cell Biology*, **134**, 885–893.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al*. (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

Nüsslein-Volhard C and Wieschaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature*, **287**, 795–801.

Nüsslein-Volhard C, Wieschaus E and Kluding H (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*, I. Zygotic loci on the second chromosome. *Wilhelm Roux's Archives of Development Biology*, **193**, 267–282.

Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, *et al*. (2004) A transcriptome analysis of the phylum Nematoda. *Nature Genetics*, **36**, 1259–1267.

Parks AL, Cook KR, Belvin M, Dompe NA, Fawcett R, Huppert K, Tan LR, Winter CG, Bogart KP, Deal JE, *et al*. (2004) Systematic generation of high-resolution deletion coverage of the Drosophila melanogaster genome. *Nature Genetics*, **36**, 288–292.

Ready DF, Hanson TE and Benzer S (1976) Development of the Drosophila retina, a neurocrystalline lattice. *Developmental Biology*, **53**, 217–240.

Richmond JE and Jorgensen EM (1999) One GABA and two acetylcholine receptors function at the *C. elegans* neuromuscular junction. *Nature Neuroscience*, **2**, 791–797.

Rio DC and Rubin GM (1985) Transformation of cultured Drosophila melanogaster cells with a dominant selectable marker. *Molecular and Cellular Biology*, **5**, 1833–1838.

Rong YS, Titen SW, Xie HB, Golic MM, Bastiani M, Bandyopadhyay P, Olivera BM, Brodsky M, Rubin GM and Golic KG (2002) Targeted mutagenesis by homologous recombination in D. melanogaster. *Genes and Development*, **16**, 1568–1581.

Rørth P, Szabo K, Bailey A, Laverty T, Rehm J, Rubin GM, Weigmann K, Milán M, Benes V, Ansorge W, *et al*. (1998) Systematic gain-of-function genetics in Drosophila. *Development*, **125**, 1049–1057.

Ryder E, Blows F, Ashburner M, Bautista Llacer R, Coulson D, Drummond J, Webster J, Gubb D, Gunton N, Johnson G, *et al*. (2004) The DrosDel collection: a set of P-element insertions for generating custom chromosomal aberrations in Drosophila melanogaster. *Genetics*, **167**, 797–813.

Scott MP and Lawrence PA (2004) Obituary: Edward B. Lewis (1918-2004). *Nature*, **431**, 143.

Scott MP, Weiner AJ, Hazelrigg TI, Polisky BA, Pirrotta V, Scalenghe F and Kaufman TC (1983) The molecular organization of the Antennapedia locus of Drosophila. *Cell*, **35**, 763–776.

Simmer F, Moorman C, van der Linden AM, Kuijk E, van den Berghe PV, Kamath RS, Fraser AG, Ahringer J and Plasterk RH (2003) Genome-wide RNAi of *C. elegans* using the hypersensitive rrf-3 strain reveals novel gene functions. *PLoS Biology*, **1**, E12.

Simon MA, Bowtell DD, Dodson GS, Laverty TR and Rubin GM (1991) Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the sevenless protein tyrosine kinase. *Cell*, **67**, 701–716.

Spradling AC and Rubin GM (1982) Transposition of cloned P elements into Drosophila germ line chromosomes. *Science*, **218**, 341–347.

Sulston JE and Horvitz HR (1977) Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology*, **56**, 110–156.

Sulston JE, Schierenberg E, White JG and Thomson JN (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, **100**, 64–119.

Tautz D (2004) Segmentation. *Developmental Cell*, **7**, 301–312.

Tomlinson A (1985) The cellular dynamics of pattern formation in the eye of Drosophila. *Journal of Embryology and Experimental Morphology*, **89**, 313–331.

Tomlinson A, Bowtell DD, Hafen E and Rubin GM (1987) Localization of the sevenless protein, a putative receptor for positional information, in the eye imaginal disc of Drosophila. *Cell*, **51**, 143–150.

Towers PR and Sattelle DB (2005) The *C. elegans lev-8* gene encodes a nicotinic acetylcholine receptor subunit (ACR-13) with roles in egg laying and pharyngeal pumping. *Journal of Neurochemistry*, **93**, 1–9.

Varmus HE (1984) The molecular genetics of cellular oncogenes. *Annual Review of Genetics*, **18**, 553–612.

White JG, Southgate E, Thomson JN and Brenner S (1986) The structure of the neuronal system of the nematode Caenorhabditis. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences*, **314**, 1–340.

Wieschaus E, Nüsslein-Volhard C and Jürgens G (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. III. Zygotic loci on the X-chromosome and fourth chromosome. *Wilhelm Roux's Archives of Development Biology*, **193**, 296–307.

Wolff T and Ready DF (1991) The beginning of pattern formation in the Drosophila compound eye: the morphogenetic furrow and the second mitotic wave. *Development (Cambridge, England)*, **113**, 841–850.

Xu T and Rubin GM (1993) Analysis of genetic mosaics in developing and adult *Drosophila* tissues. *Development*, **117**, 1223–1237.

Short Specialist Review

Functional genomics in *Saccharomyces cerevisiae*

Kara Dolinski and Olga Troyanskaya
Princeton University, Princeton, NJ, USA

1. Introduction

The availability of the complete *Saccharomyces cerevisiae* genome sequence, published in 1996 (Goffeau *et al*., 1996), stimulated much innovation in genome-scale experimental approaches, including the development and implementation of the many uses of DNA microarrays, the systematic deletion (Winzeler *et al*., 1999; Deutschbauer *et al*., 2002; Giaever *et al*., 2002; Steinmetz *et al*., 2002) and genetic footprinting studies (Smith *et al*., 1995; Smith *et al*., 1996), several genome-wide collections of gene and promoter fusions (e.g., Ross-Macdonald *et al*., 1997), as well as a number of genome-wide protein- and gene-interaction studies. The goals of these very different approaches are nonetheless the same: to determine the function of all the genes in the yeast genome. A number of bioinformatics approaches are beginning to be applied to predict gene function on the basis of these large-scale, heterogeneous data sets. Because genome-wide experimental methods often sacrifice specificity for scale, an integrated analysis of multiple types of experimental data is necessary to make accurate predictions of gene function on such a large scale. After discussing the functional genomics data currently available for *S. cerevisiae*, we describe how a probabilistic integration method can use these data to predict gene function.

2. Available functional genomics data in *S. cerevisiae*

The avalanche of functional genomics data began with the development of microarray technology in the mid-1990s (Schena *et al*., 1995; Shalon *et al*., 1996; Brown and Botstein, 1999). Gene expression microarrays generate mRNA expression profiles of genes on the scale of the entire genome. More recently, microarrays have now been applied to map protein–DNA interactions (e.g., Ren *et al*., 2000; Iyer *et al*., 2001; Lieb *et al*., 2001; Kurdistani *et al*., 2002; Kurdistani and Grunstein, 2003; Ng *et al*., 2003; Harbison *et al*., 2004) and to characterize protein interactions (Zhu *et al*., 2001; *see also* Article 97, **Seven years of yeast microarray analysis**, Volume 4). In addition to these various microarray experiments, high-throughput interaction studies have also been carried out, including large-scale

Table 1 Some sources of functional genomics data collections for *S. cerevisiae*

Database	Data type	References	URL
GRID	Genetic/physical interactions	Breitkreutz *et al*. (2003)	**http://biodata.mshri.on.ca/yeast_grid/servlet/SearchPage**
BIND	Genetic/physical interactions, Pathways	Bader *et al*. (2003)	**http://www.blueprint.org/bind/bind.php**
DIP	Physical interactions	Xenarios *et al*. (2002)	**http://dip.doe-mbi.ucla.edu/dip/Main.cgi**
MINT	Physical interactions	Zanzoni *et al*. (2002)	**http://160.80.34.4/mint/**
IntAct	Physical interactions	Hermjakob *et al*. (2004b)	**http://www.ebi.ac.uk/intact/index.html**
Deletion Consortium	Large-scale phenotype analysis	Winzeler *et al*. (1999); Giaever *et al*. (2002)	**http://www-sequence.stanford.edu/group/yeast_deletion_project/data_sets.html**
GEO	MicroArray	Edgar *et al*. (2002)	**http://www.ncbi.nlm.nih.gov/geo/**
ArrayExpress	MicroArray	Brazma *et al*. (2003)	**http://www.ebi.ac.uk/arrayexpress/**
YMGV	MicroArray	Marc *et al*. (2001)	**http://www.transcriptome.ens.fr/ymgv/**
SMD	MicroArray	Gollub *et al*. (2003)	**http://smd.stanford.edu/**
OPD	Mass Spec/Proteomics	Prince *et al*. (2004)	**http://bioinformatics.icmb.utexas.edu/OPD/**

genetic interaction screens (Tong *et al*., 2001; Ooi *et al*., 2003; Tong *et al*., 2004), two hybrid analyses (Uetz *et al*., 2000; Ito *et al*., 2001; Hazbun *et al*., 2003), mass spectrometry (Gavin *et al*., 2002; Ho *et al*., 2002), and combined mass spectrometry/chromatography (aka MudPIT) (Washburn *et al*., 2001; Washburn *et al*., 2002). A wealth of large-scale phenotype data are available in *S. cerevisiae*, including results from the Saccharomyces Genome Deletion Project (Winzeler *et al*., 1999; Deutschbauer *et al*., 2002; Giaever *et al*., 2002; Steinmetz *et al*., 2002) and from large-scale insertional mutagenesis (Smith *et al*., 1995; Smith *et al*., 1996; Kumar *et al*., 2002). In addition, there are databases available that contain quantitative data from analysis of morphological mutants (SCMD, see Saito *et al*., 2004) and growth aberrations (PROPHECY, see Warringer *et al*., 2003).

Table 1 lists some of the major collections of yeast functional genomics data. All these sites allow public data download in addition to a web interface, facilitating bioinformatics analysis of the data. The yeast databases MIPS and SGD also distribute some of these large-scale data sets. MIPS (**ftp://ftpmips.gsf.de/yeast/**) and SGD (**ftp://ftp.yeastgenome.org/pub/yeast/**) both provide data for bulk download via ftp. In addition, SGD has developed a lightweight version of SGD called SGD Lite (**http://sgdlite.princeton.edu/**), which has some of the yeast data in a PostgreSQL database meant for simple, local installation.

3. Standards in the functional genomics community

Adherence to community standards for data format, annotation, and distribution is essential to enable full utilization of data generated by large-scale experiments through comparison and integration of multiple data sets. The Open Biological Ontologies (OBO) site is a web page that provides links to various ontologies and standards projects, including the Microarray Gene Expression Data (MGED)

Society (see **http://www.mged.org/**, Spellman *et al*., 2002; Causton and Game, 2003) and the Gene Ontology project (GO, **http://www.geneontology.org**, Ashburner *et al*., 2000; *see also* Article 82, **The Gene Ontology project**, Volume 8). Also available through the OBO site are links to the Proteomics Standards Initiative, which has created standards for protein–protein interactions as well as mass spectrometry data (Hermjakob *et al*., 2004a), and BioPAX, which has been developing a common exchange format for pathways data (see **http://www.biopax.org/**).

4. General probabilistic integration of functional genomics data

To address the need for a generalizable method for comprehensive data integration, an approach should combine heterogeneous data types of various levels of accuracy in an algorithmic fashion and should also easily adapt to new data sources. Recently, several such approaches have been introduced, including methods that focus on modeling one or several specific data types (Friedman *et al*., 2000; Pavlidis and Noble, 2001; Ihmels *et al*., 2002; Imoto *et al*., 2002; Segal *et al*., 2003) and more general integration approaches (Jansen *et al*., 2003; Troyanskaya *et al*., 2003; Lanckriet *et al*., 2004).

The advantage of general approaches to data integration is their adaptability. For example, MAGIC (Multisource Association of Genes by Integration of Clusters), a general probabilistic approach to data integration, uses a Bayesian network architecture that can easily incorporate new data sources, datasets, and analysis methods (Troyanskaya *et al*., 2003). It incorporates expert knowledge in the prior probability parameters in the Bayesian framework or learns from available data, thus formally integrating relative accuracies of different experimental and computational techniques in the analysis and minimizing potential bias toward well-studied areas in its reasoning. In addition, Bayesian networks are generally robust to noise in prior probabilities and in training data. These characteristics of Bayesian networks yield high accuracy of gene function predictions (the prototype MAGICsystem can achieve accuracy of up to 83% for the high-stringency predictions), and the probabilistic nature of the system provides confidence levels for each output.

The MAGIC integration system takes as input groupings (or clusters) of genes based on each experimental data set (e.g., shared transcription factor binding sites, protein–protein interaction, or coexpression data). The system represents all input groupings as gene i–gene j pairs with corresponding scores s_{ij}. The score s_{ij} corresponds to the strength of each method's belief in the existence of a relationship between gene i and gene j. The score $s_{ij} > 0$ if gene i and gene j appear in the same cluster or grouping or if they interact on the basis of an experimental method, and it can be binary (e.g., results of coimmunoprecipitation experiments), continuous or discrete (e.g., $-1 \leq s \leq 1$ for Pearson correlation). MAGIC's Bayesian network combines evidence from input groupings and generates a posterior belief for whether each gene i–gene j pair has a functional relationship. For each pair of

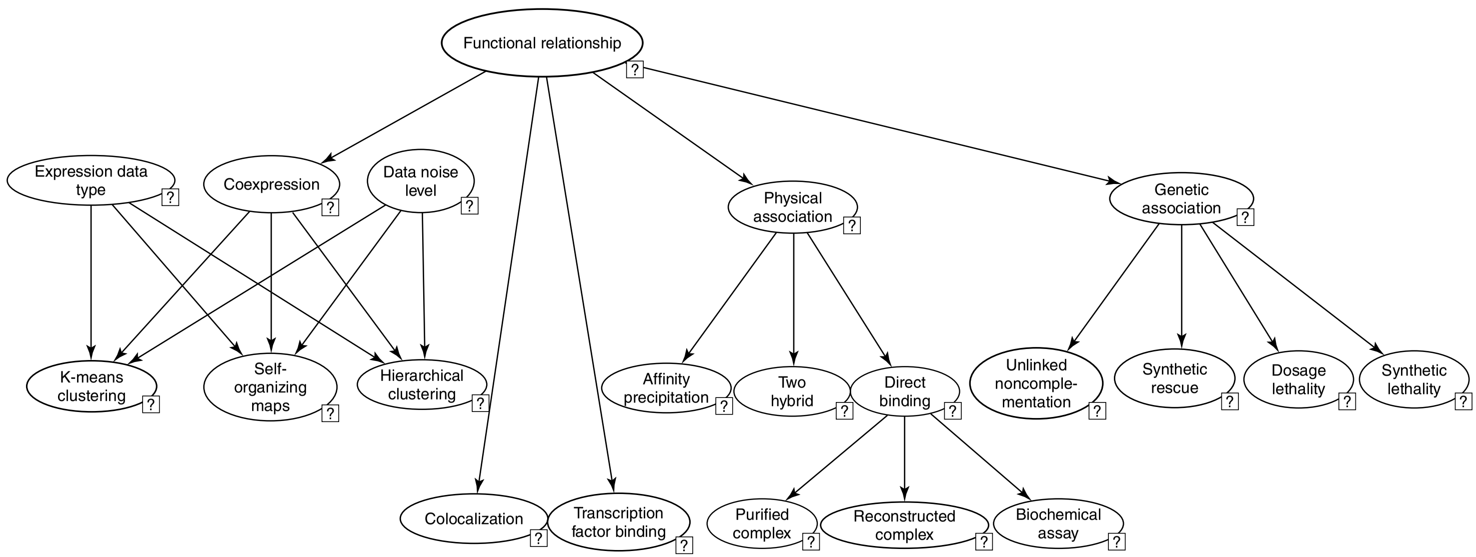


Figure 1 A structure for the Bayesian network in MAGIC. The network is instantiated with evidence (at the bottom nodes) for each pair of genes in the yeast genome, and the final confidence level is produced on the basis of the evidence for biological relationship available for each pair of genes and on the prior probabilities encoded in the network conditional probability tables. This figure is adapted from Troyanskaya *et al*. (2003)

genes, the network essentially asks the following question: "What is the probability, based on the evidence presented, that products of gene *i* and gene *j* interact or are involved in the same biological process?" MAGIC's Bayesian network structure (Figure 1) was determined through consultation with experts in yeast genomics and microarray analysis, and the prior probabilities can be either based on expert consultation or learnt from example data (such as functional annotations of genes with the Gene Ontology biological process annotations).

This success of function prediction based on heterogeneous data demonstrates the potential of sophisticated data integration algorithms. Further development of such computational methods combined with increased availability of large-scale functional genomics data in downloadable standardized formats should enable accurate prediction of function for most unknown yeast proteins. These predictions, followed by targeted laboratory experiments, may enable fast and relatively low cost annotation of the entire yeast genome.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Bader GD, Betel D and Hogue CW (2003) BIND: The biomolecular interaction network database. *Nucleic Acids Research*, **31**, 248–250.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, *et al*. (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**, 68–71.

Breitkreutz BJ, Stark C and Tyers M (2003) The GRID: the general repository for interaction datasets. *Genome Biology*, **4**, R23.

Brown PO and Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.

Causton HC and Game L (2003) MGED comes of age. *Genome Biology*, **4**, 351.

Deutschbauer AM, Williams RM, Chu AM and Davis RW (2002) Parallel phenotypic analysis of sporulation and postgermination growth in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15530–15535.

Edgar R, Domrachev M and Lash AE (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

Friedman N, Linial M, Nachman I and Pe'er D (2000) Using bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546–563.

Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, *et al*. (2003) The stanford microarray database: data access and quality assessment tools. *Nucleic Acids Research*, **31**, 94–96.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hazbun TR, Malmstrom L, Anderson S, Graczyk BJ, Fox B, Riffle M, Sundin BA, Aranda JD, McDonald WH, Chiu CH, *et al*. (2003) Assigning function to yeast proteins by integration of technologies. *Molecular Cell*, **12**, 1353–1365.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al*. (2004a) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**, 177–183.

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, *et al*. (2004b) IntAct: an open source molecular interaction database. *Nucleic Acids Research*, **32**, D452–D455.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y and Barkai N (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**, 370–377.

Imoto S, Goto T and Miyano S (2002) Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. *Pacific Symposium on Biocomputing*, **7**, 175–186.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M and Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF and Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Kumar A, Cheung KH, Tosches N, Masiar P, Liu Y, Miller P and Snyder M (2002) The TRIPLES database: a community resource for yeast molecular biology. *Nucleic Acids Research*, **30**, 73–75.

Kurdistani SK and Grunstein M (2003) Histone acetylation and deacetylation in yeast. *Nature Reviews. Molecular Cell Biology*, **4**, 276–284.

Kurdistani SK, Robyr D, Tavazoie S and Grunstein M (2002) Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nature Genetics*, **31**, 248–254.

Lanckriet GR, Deng M, Cristianini N, Jordan MI and Noble WS (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pacific Symposium on Biocomputing*, pp. 300–311.

Lieb JD, Liu X, Botstein D and Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics*, **28**, 327–334.

Marc P, Devaux F and Jacq C (2001) yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Research*, **29**, E63.

Ng HH, Robert F, Young RA and Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular Cell*, **11**, 709–719.

Ooi SL, Shoemaker DD and Boeke JD (2003) DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nature Genetics*, **35**, 277–286.

Pavlidis P and Noble WS (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, **2**, RESEARCH0042.

Prince JT, Carlson MW, Wang R, Lu P and Marcotte EM (2004) The need for a public proteomics repository. *Nature Biotechnology*, **22**, 471–472.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Ross-Macdonald P, Sheehan A, Roeder GS and Snyder M (1997) A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces*

cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 190–195.

Saito TL, Ohtani M, Sawai H, Sano F, Saka A, Watanabe D, Yukawa M, Ohya Y and Morishita S (2004) SCMD: Saccharomyces cerevisiae Morphological Database. *Nucleic Acids Research*, **32**, Database issue, D319–D322.

Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.

Shalon D, Smith SJ and Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, **6**, 639–645.

Smith V, Botstein D and Brown PO (1995) Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6479–6483.

Smith V, Chou KN, Lashkari D, Botstein D and Brown PO (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science*, **274**, 2069–2074.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, **3**, RESEARCH0046.

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, *et al*. (2002) Systematic screen for human disease genes in yeast. *Nature Genetics*, **31**, 400–404.

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Troyanskaya OG, Dolinski K, Owen AB, Altman RB and Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8348–8353.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Warringer J, Ericson E, Fernandez L, Nerman O and Blomberg A (2003) High-resolution yeast phenomics resolves different physiological features in the saline response. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15724–15729.

Washburn MP, Ulaszek R, Deciu C, Schieltz DM and Yates JR III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Analytical Chemistry*, **74**, 1650–1657.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al*. (1999) Functional characterization of the S cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, **30**, 303–305.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G (2002) MINT: a Molecular INTeraction database. *FEBS Letters*, **513**, 135–140.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Short Specialist Review

The *C. elegans* genome

Jonathan Hodgkin
University of Oxford, Oxford, UK

1. The *C. elegans* genome

The genome of the small nematode worm *Caenorhabditis elegans* contains just over 100 million base pairs. It is one of the smallest of all known animal genomes, and partly for this reason, it was the first fully sequenced genome for any multicellular organism, being essentially completed in 1998 (*C. elegans* Sequencing Consortium, 1998).

The nuclear genome of normal diploid *C. elegans* hermaphrodites is organized into five pairs of autosomes (size range 13–21 Mb) and one pair of X chromosomes (17 Mb). The alternative sexual form, the male, has only one X chromosome and there is no Y chromosome. The small mitochondrial genome (13.8 kb) has also been completely sequenced. Nematode chromosomes are unusual in that they are holocentric, with multiple kinetochores, and lack extended regions of centric heterochromatin. *Caenorhabditis elegans* also has relatively little repetitive sequence. As a result, it has been possible to achieve full sequence coverage, telomere to telomere, for all six chromosomes, resulting in a precise figure of 100 277 975 nucleotides for the entire nuclear genome.

Extensive annotation of the genome has revealed approximately 19 900 predicted protein-coding genes, as well as about 600 tRNA genes, 200 tRNA pseudogenes, 55 copies of the large ribosomal RNA genes, 110 copies of the 5S rRNA gene, and the usual eukaryotic sets of snRNAs, scRNAs, snoRNAs, and other functional RNA genes. Micro-RNA genes, which were first discovered in *C. elegans*, number at least 100. Repeat sequences account for about 6% of the genome; some of these form tandem arrays that may serve as kinetochores. Some of the other repeat sequences belong to known transposon families, most of which are currently inactive although mobilization can be achieved. Telomeres are conventional, with TTAGGC tandem repeats, but there are no specialized subtelomeric regions.

The total percentage of protein-coding DNA is high, for a multicellular organism (27%), and there appear to be relatively few pseudogenes and almost no processed pseudogenes. About one-quarter of all genes are organized into operons, in which between two and eight genes are cotranscribed from a common promoter as a polycistronic transcript (Blumenthal *et al*., 2002). This primary transcript is broken up into separate monocistronic mRNA molecules by trans-splicing: a small leader RNA, termed SL2, transcribed from separate loci, is spliced onto the 5′ end of

each cistronic sequence, in a process related to conventional cis-splicing. About 70% of all genes also undergo general trans-splicing to acquire SL1, a small leader RNA similar to SL2, at the 5′ end of the mRNA. Organization into operons seems to be a peculiarity of the *C. elegans* genome, rarely found in other invertebrate genomes, and possibly related to the compactness of the genome. Some operons contain genes of related function, but many do not.

Gene number is surprisingly high when compared to an estimate of about 14 000 genes for *Drosophila*, which is a substantially more complex animal in terms of anatomy, development, and behavior. Various factors may contribute to this large number of genes, such as a high level of duplication of genes and genomic regions, as compared to other eukaryotic genomes. Another contributory factor may be a relatively low level of alternative splicing – currently, only about 11% of *C. elegans* genes are known to generate more than one mRNA isoform, which is lower than the estimates for other animal species. Protein diversity may therefore be generated more by expanded gene families than by alternative splicing.

Some gene families appear to have undergone considerable expansion during the evolution of *C. elegans*. Notably, there are over 200 genes encoding DNA-binding proteins of the nuclear hormone receptor class, and about 1000 genes encoding seven-pass transmembrane proteins, which are probably G-protein coupled receptors (GPCRs). The large number of these predicted GPCRs may be related to the sophisticated chemosensory repertoire of *C. elegans*. Detection of odorants provides the worm with most of its sensory information about the environment.

Multiple postgenomic approaches are being used to verify and investigate all the genes predicted to exist in the *C. elegans* genome. Many of the predicted genes have no obvious homologs in other organisms, and their function is usually unknown. Large sets of ESTs (expressed sequence tags) have been generated, along with SAGE (serial analysis of gene expression) analyses, in order to define the transcriptome. Transcripts for predicted genes have been systematically searched for, by Reverse Transcription Polymerase Chain Reaction (RT PCR). Systematic expression studies have been undertaken by in situ hybridization, by microarray analysis, and by the generation of transgenic lines carrying reporter genes fused to Green Fluorescent Protein (GFP) or lacZ markers. Protein interaction maps have been constructed using high-throughput yeast 2-hybrid screens (Li *et al*., 2004).

In terms of function, large numbers of genes have been investigated by conventional forward mutagenesis, using chemical and transposon mutagenesis. Reverse genetic programs for systematically generating deletions in targeted genes have been established. Large-scale functional studies have been carried out by means of RNAi (RNA interference), which provides an effective means of suppressing expression of most, though not all, *C. elegans* genes. RNAi can be elicited most easily by feeding worms on bacteria expressing double-stranded RNA for a targeted gene, and “feeding libraries” have been generated that allow RNAi to be applied to most of the worm’s genes (Kamath *et al*., 2002). The resulting data allow preliminary assignment of function to about 23% of genes. RNAi tests on the remaining 77% reveal no obvious function, however, for a variety of possible reasons, such as subtle or redundant activities, or incomplete knockdown by RNAi.

The RNAi data, together with other information, provide evidence for long-range order in the *C. elegans* genome, with some clustering of genes affecting related

biological processes. There are also conspicuous differences between the autosomes and the X chromosome, with a reduced frequency of essential genes being found in the X chromosome. The five autosomes all show a similar arrangement of a distinct central region, or cluster, flanked by arms of roughly equal size. The cluster regions are somewhat more gene-dense and show reduced recombination levels; the genes in the clusters show higher levels of conservation and are more frequently essential, as compared to the genomic average. In contrast, the arm regions have higher levels of recombination, contain more gene families and more genes that lack obvious homologs in other organism. The origin and significance of these large-scale features are unknown.

Investigation of the *C. elegans* genome has been greatly aided by extensive sequencing of closely related nematode species. In particular, the genome of *Caenorhabditis briggsae* has been shotgun sequenced at 10x coverage, allowing assembly of most of its genome (Stein *et al*., 2003). The two species diverged possibly 100 million years ago, and the two genomes have experienced multiple rearrangements, mostly intrachromosomal. Comparison of the *C. elegans* and *C. briggsae* genomes has permitted confirmation of many predicted genes and identification of conserved noncoding regions such as regulatory sites. The model organism database WormBase (**http://www.wormbase.org/**) provides an integrated and continually updated overview of the genomes of *C. elegans* and related species.

References

Blumenthal T, *et al*. (2002) A global analysis *of Caenorhabditis elegans* operons. *Nature*, **417**, 851–854.

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Kamath RS, *et al*. (2002) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.

Li S, *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.

Stein LD, *et al*. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *Public Library of Science Biology*, **1**, 166–192.

The *Drosophila* genome(s)

Steven Russell and Casey M. Bergman
University of Cambridge, Cambridge, UK

The first release of the 118-Mb *Drosophila melanogaster* euchromatic genome sequence, generated by a combination of map-based and whole-genome shotgun strategies (Adams *et al*., 2000), has been updated several times, improving sequence and annotation quality (Celniker and Rubin, 2003). With the essential use of EST and cDNA sequences, release 4.0 predicts 13 472 genes producing 18 746 protein-coding transcripts. Less-stringent annotations suggest additional genes, at least some of which are supported by microarray and RNAi studies (Boutros *et al*., 2004). Along with the protein-coding complement of the genome, there have been improvements in annotating transposable element sequences (Kaminker *et al*., 2002), an effort to assemble the 60 Mb of heterochromatin (Hoskins *et al*., 2002) and computational attempts to identify microRNAs (Brennecke and Cohen, 2003). This ongoing work aims at generating a high-quality contiguous genome sequence and functional annotation encompassing the entire euchromatic genome with as much of the heterochromatin as possible. All sequence gaps in the euchromatin are scheduled to be closed by the beginning of 2005; currently, there are 23 gaps in release 4. Annotation is, of course, an ongoing effort. Physical gaps and large tandem repeat regions are also being completed as a part of finishing the heterochromatic sequence. As it stands, in terms of scaffold integrity and annotation accuracy, the *D. melanogaster* euchromatin represents a gold standard for genome sequences.

The success of the whole-genome shotgun strategy in *D. melanogaster* suggested that other *Drosophila* species could be rapidly sequenced as resources for comparative genome analysis. The first, *D. pseudoobscura*, was chosen on the basis of estimates that unconstrained DNA sequence divergence should be "saturated" between *D. pseudoobscura* and *D. melanogaster* (Bergman *et al*., 2002); thus, sequences conservation between these species should imply functional constraint. The first draft of the *D. pseudoobscura* sequence and whole genome comparisons with *D. melanogaster* are now available (**http://pipeline.lbl.gov/cgi-bin/gateway2?bg=dm1**). Two further efforts to sequence an additional 10 *Drosophila* species are underway; the first, focusing on developing resources for population genetic and evolutionary analysis of two species closely related to *D. melanogaster* – *D. simulans* and *D. yakuba*; the second, to sequence a panel of eight species spanning a range of divergence distances in the genus. As of August 2004, over two million sequencing reads and preliminary WGS assemblies exist

for five of these species (**http://rana.lbl.gov/drosophila/multipleflies.html**). The depth and richness of genome sequences now available in *Drosophila* will provide invaluable resources for the comparative analyses of genes and cis-regulatory sequences (Bergman *et al*., 2002; Grad *et al*., 2004).

Complementing the sequence(s), several genomics resources are available; these include 2-hybrid libraries (Giot *et al*., 2003), RNAi collections (Boutros *et al*., 2004), and oligonucleotide and amplicon microarrays (Johnston *et al*., 2004). Such resources increase the utility of the fly as a system for integrative biology, for exploring conserved aspects of development, and as a model for human diseases. The strength of *Drosophila* has always been its sophisticated array of genetic tools and the genome sequence has strengthened this. Transposon insertions are now easily mapped with base-pair precision to the genome sequence, and the Gene Disruption Project, which aims at generating insertions in every gene, has currently tagged over 40% of *Drosophila* genes (Bellen *et al*., 2004). In addition, collections of transposons carrying sites for the yeast site-specific FLP recombinase are being used to make precisely defined chromosomal aberrations such as second-generation deficiency kits (Ryder *et al*., 2004; Parks *et al*., 2004). Complementing classical approaches, forward genetic strategies utilizing transposons for GAL4-inducible gene expression and protein trapping are identifying genes involved in a variety of processes. Finally, the ongoing development of high-density SNP maps permits rapid association of mutant phenotypes with individual genes. Thus, the genome sequence has accelerated the large-scale genetic analysis of the fly by facilitating phenotype-to-genotype association.

As a model for human disease, the fly genome is excellent for uncovering components of conserved molecular pathways (Tickoo and Russell, 2002). One of the first discoveries from the genome sequence was the fly homolog of the p53 tumor suppressor. Unlike mammals, flies have a single p53 gene, greatly facilitating a functional analysis since problems of redundancy are obviated (Sutcliffe and Brehm, 2004). Over two-thirds of the genes implicated in human cancers have fly counterparts, and considerable progress is being made in understanding the biology and interactions of these genes in the fly. The Homophila database, linking entries in the Online Mendelian Inheritance in Man (OMIM) database with homologous sequences in the *Drosophila* genome currently contains over 1600 associations (Reiter *et al*., 2001). Complex processes, such as insulin signaling, the control of growth, and the regulation of longevity, are all beginning to yield to genome scale analysis in the fly. Similarly, with a sophisticated nervous system, *Drosophila* is a firmly established tool for studying conserved aspects of neurobiology. Using carefully controlled behavioral paradigms, genome-wide screens for genes involved in processes as diverse as alcoholism, drug addiction, and sleep are underway. In addition, there is considerable interest in using the fly as a model for neurodegenerative disease since the identification of homologs of genes implicated in human neuropathologies. Many groups are using genetic and microarray screens to identify gene function in neurodegenerative processes with the hope of uncovering targets for drug intervention in humans. Taken together, the *Drosophila* genome sequence has opened up a range of disease states and physiological processes to a systems level analysis in the fly and promises much in terms of increasing our understanding of human biology.

Finally, among model organisms, *Drosophila* presents a unique opportunity to link information encoded in the genome sequence to chromatin structure through the banding patterns of the polytene chromosomes. Since their discovery, many hypotheses, such as "one band-one gene", have been postulated to explain the banding pattern, although no general relationship has yet been discovered between the different chromatin states of the bands and interbands. This classical problem can now be addressed by linking the genome sequence to cytological maps of polytene chromosomes and testing the association of sequence features with banding patterns. Genomics approaches in *Drosophila* also allow the analysis of chromosome structure and organization as it relates to chromatin regulation and gene expression. For example, combining expression data with the genome sequence demonstrates the existence of gene expression "neighborhoods", linked clusters of neighboring genes with similar expression profiles (Spellman and Rubin, 2002). In addition, chromatin-immunoprecipitation (ChIP) experiments with genome tiling path microarrays allow genome-wide mapping of in vivo DNA-protein interactions (Sun *et al*., 2003). The continuing application of such approaches may present the key to unlocking the elusive relationship between chromatin structure and function and the underlying genetic or genomic organization of the chromosomes.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, Tsang G, Evans-Holm M, Hiesinger PR, Schulze KL, Rubin GM, *et al*. (2004) The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics*, **167**, 761–781.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al*. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biology*, **3**, RESEARCH0086.

Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Consortium HF, Paro R and Perrimon N (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science*, **303**, 832–835.

Brennecke J and Cohen SM (2003) Towards a complete description of the microRNA complement of animal genomes. *Genome Biology*, **4**, 228.

Bridges CB (1916) Non-disjunction as proof of the chromosomal theory of heredity. *Genetics*, **1**, 1–52.

Bridges CB (1935) Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *The Journal of Heredity*, **26**, 60–64.

Celniker SE and Rubin GM (2003) The *Drosophila melanogaster* genome. *Annual Review of Genomics and Human Genetics*, **4**, 89–117.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Grad YH, Roth FP, Halfon MS and Church GM (2004) Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*. *Bioinformatics*, **20**, 2738–2750.

Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, *et al*. (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biology*, **3**, RESEARCH0085. **http://www.dhgp.org/**

Johnston R, Wang B, Nuttall R, Doctolero M, Edwards P, Lu J, Vainer M, Yue H, Wang X, Minor J, *et al*. (2004) FlyGEM, a full transcriptome array platform for the *Drosophila* community. *Genome Biology*, **5**, R19.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, *et al*. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology*, **3**, RESEARCH0084.

Parks AL, Cook KP, Belvin M, Dompe NA, Fawcett R, Huppert K, Tan LR, Winter CG, Bogart KP, Deal JE, *et al*. (2004) Systematic generation of high-resolution deletion coverage of the *Drosophila melanogaster* genome. *Nature Genetics*, **36**, 288–297.

Reiter LT, Potocki L, Chien S, Gribskov M and Bier E (2001) A systematic Analysis of human disease-associated gene sequences in *Drosophila melanogaster*. *Genome Research*, **11**, 1114–1125.

Ryder E, Blows F, Ashburner M, Bautista-Llacer R, Coulson D, Drummond J, Webster J, Gubb D, Gunton N, Johnson G, *et al*. (2004) The DrosDel collection: a set of *P*-element insertions for generating custom chromosomal aberrations in *Drosophila melanogaster*. *Genetics*, **167**, 797–813.

Spellman PT and Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *Journal of Biology*, **1**, 5.

Sun LV, Chen L, Greil F, Negre N, Li TR, Cavalli G, Zhao H, Van Steensel B and White KP (2003) Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9428–9433.

Sutcliffe JE and Brehm A (2004) Of flies and men; p53, a tumor suppressor. *FEBS Letters*, **567**, 86–91.

Tickoo S and Russell S (2002) *Drosophila melanogaster* as a model system for drug discovery and pathway screening. *Current Opinion in Pharmacology*, **2**, 555–560.

Short Specialist Review

The Fugu and Zebrafish genomes

Greg Elgar
MRC Rosalind Franklin Centre for Genomic Research, Cambridge, UK

The Ray-finned fish, comprised primarily of teleosts, represent over half of the world's extant vertebrates and first arose about 450 million years ago (Figure 1). In 1968, Ralph Hinegardner assayed the cellular DNA content of over 200 teleost fishes (Hinegardner, 1968). Although he documented a wide range of genome sizes, the smallest belonged to the Tetraodontoid fish, or Pufferfish. His haploid measurements of 0.40 pg of DNA per cell equate to a genome size of less than 400 Mb. However, it was not until 1990 that these findings were applied to genomics, when Sydney Brenner initiated a program to characterize the Fugu genome at the sequence level (Brenner *et al*., 1993). Nine years later, *Fugu rubripes* became the second vertebrate, after human, to have its genome sequenced to draft status (Aparicio *et al*., 2002). This was a milestone in comparative genomics, because it hailed the advent of a number of whole-genome comparisons between the human genome and other vertebrates in order to identify similarities and differences in sequence that might be linked to function (*see* Article 48, **Comparative sequencing of vertebrate genomes**, Volume 3).

Besides containing a similar set of genes, Fugu genes tend to have the same intron/exon structure as their human counterparts. There are few exceptions to this, but paradoxically, where there are differences, the tiny Fugu genome contains additional introns. For example, the human *PKD1* gene contains 46 exons, whereas the Fugu gene possesses 54 (Sandford *et al*., 1997). There are also additional introns found in the Fugu dystrophin gene (Pozzoli *et al*., 2003).

Early mapping and sequencing studies sought to establish how much conservation of gene order, or synteny, could be found between Fugu and human genomes (*see* Article 20, **Synteny mapping**, Volume 3). At a time when there were very few detailed physical maps of the human genome, sequencing the more compact Fugu genome to look for candidate genes that could be mapped back to the equivalent human region was seen as a useful approach to the identification of candidate disease genes. Some early studies showed great promise (Trower *et al*., 1996), while others failed to find expected syntenic groups (Gilley *et al*., 1997), and it soon emerged that there was a high degree of regional variation.

Nevertheless, the advantages of sequencing the Fugu genome to identify genes in a region were obvious, and this was further accelerated through the use of sequence scanning approaches. This allowed very rapid analysis of the gene content of a significant portion of the Fugu genome (Elgar *et al*., 1999), providing a much more

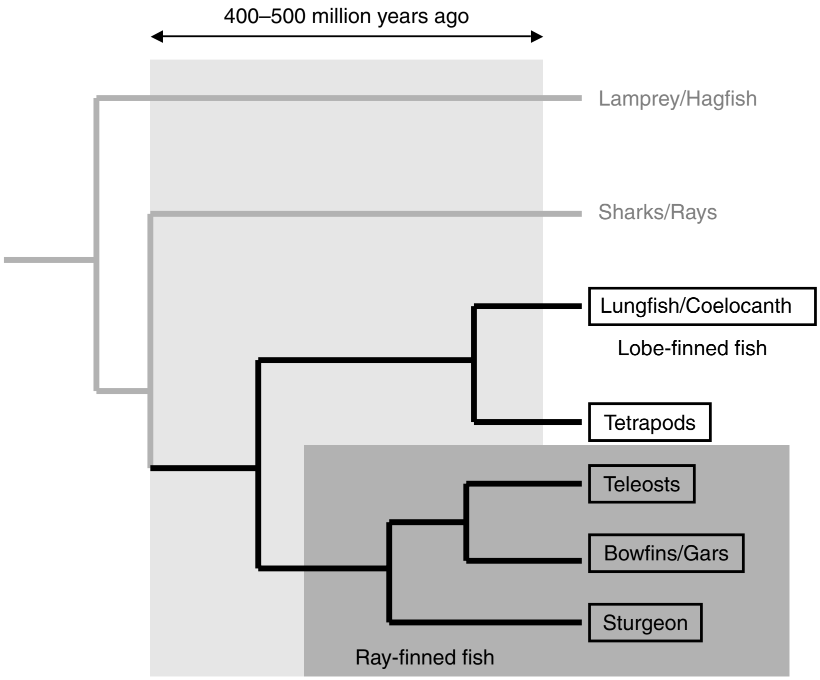


Figure 1 Schematic representation of early vertebrate evolution, indicating the relationship between tetrapods and teleost fish relative to other key orders of the fishes

detailed analysis of the structure of the Fugu genome and adding fuel to the argument for sequencing it in its entirety. In October 2000, a consortium of labs from the United States, United Kingdom, and Singapore was established to sequence the genome and 12 months later the first draft was released. Analysis of the whole-genome sequence, largely through comparison with the human genome, the only other vertebrate sequence available at the time, identified and discussed the relevance of similarities and differences between these two highly divergent genomes (Aparicio *et al*., 2002).

In support of earlier studies on smaller data sets, the Fugu genome was found to be rich in coding sequence, with small introns and short intergenic distances. Whereas the Fugu genome contains only about 500 introns greater than 10 kb in length, there are over 12 000 that exceed this size in the human genome. An exhaustive analysis of the repetitive portion of the Fugu genome found that, interestingly, although less than 15% of the genome is repetitive, a large number of different classes of repeats are identifiable. For example, there are at least 40 different families of transposable elements in the Fugu genome that have substitution rates low enough to suggest they are still active, compared to six such families in the human genome.

It is well established that the human genome has an isochore structure of low and high G+C content, which is generally reflected by regions of low and high gene density, and there is some evidence that the Fugu genome is also heterogeneous,

although not to the extent of mammals. Generally, Fugu DNA is more G+C rich than human DNA with an average across the genome of 45% compared with 41% for human. This is not simply the result of the increased density of coding sequence in the Fugu genome (which is generally much higher in G+C content) as there is little difference in average G+C between regions of very low gene density and regions of very high gene density.

A comparison of the Fugu and human proteomes provided insights into the evolution of different classes of proteins in vertebrates. While many proteins are very similar between the two genomes, about 25% of the proteome of each species is either unique or has evolved to such an extent that similarity is no longer easy to recognize using whole-genome approaches. Many of these proteins are immune related, such as cytokines, which would be expected to evolve rapidly, and more in-depth analyses have succeeded in identifying some of these, including CD4, interferons, and interleukins.

The analysis of the coding portion of the Fugu genome was made all the more difficult owing to the complete absence of cDNA data at the time. This has been redressed to an extent by another International collaboration, this time between Cambridge, Japan, and the Unites States, to sequence 24 000 ESTs, representing tags for about 10 000 Fugu genes (Clark *et al*., 2003).

With the availability of large contiguous regions, a more global assessment of synteny with mammalian genomes could be made. In agreement with many earlier reports on specific regions, the Fugu genome retains large blocks of conserved synteny with the human genome, but in general, these regions are highly scrambled due presumably to local rearrangements. Consequently, it is unusual to find more than two or three genes in the same order and orientation in both the Fugu and human genomes.

In 2004, just over two years after the analysis of the Fugu genome, the genome of a second pufferfish, *Tetraodon nigroviridis*, was released. The two genomes, unsurprisingly given their close evolutionary relationship, are remarkably similar, but importantly, as well as confirming the Fugu analysis, the Tetraodon genome data was compared with other genomes in an attempt to reconstruct the vertebrate proto-karyotype (Jaillon *et al*., 2004).

Perhaps the most exciting application of the Fugu genome in comparative genomics is in the identification, through conservation, of putative noncoding functional sequences. Early studies on the *Hox* genes demonstrated that this approach was valuable (Marshall *et al*., 1994), but once again there was some debate as to whether the Fugu genome was simply too evolutionarily distant to be of real use in this area. Nevertheless, a number of regions have been successfully identified and characterized using this approach (reviewed in Elgar, 2004).

All these regions had one thing in common; they were associated with genes involved in developmental regulation. Capitalizing on this, Woolfe *et al*. (2005) carried out a genome-wide survey of all highly conserved noncoding sequences in Fugu, examined their association with genes involved in development, and critically developed an assay that allowed these sequences to be functionally annotated. With a hint of irony, the organism that they used to functionally assay these highly conserved sequences was the zebrafish.

Whereas the motivation behind sequencing the Fugu genome derived solely from the need for a model organism for comparative genomic analyses, the need to sequence the zebrafish genome was driven by researchers already using the zebrafish as an experimental model in its own right. The Fugu genome may be small but as an experimental system it is far from ideal. These pufferfish are large marine teleosts that take two to three years to reach sexual maturity and only produce eggs and sperm seasonally. They are not suited to tank breeding and are expensive to maintain.

The zebrafish on the other hand, emerged primarily because of its advantages as an experimental system. It is a common, freshwater aquarium fish that is extremely easy to keep and maintain at relatively high densities. They are small (adults are about 3 cm in length) and reach sexual maturity at about 12 weeks. Moreover, and perhaps most importantly, they readily generate large numbers of fertilized eggs. Collection of newly fertilized eggs can be coordinated ready for injection/manipulation using a fixed light cycle. While being oviparous has clear advantages for observation of the developing embryo, the zebrafish has the added benefit that the embryo is virtually transparent. In fact, providing it is not permitted to dry out, an embryo will develop normally while being observed under a microscope.

As a result, the zebrafish genome was prioritized by the genomics community, and more importantly, by the funding agencies, as one of the key genomes to be sequenced after the human genome. Unfortunately, the zebrafish genome is rather large for a teleost fish, with an estimated size of between 1600 and 1700 Mb, making it four to five times larger than the Fugu genome. An additional difficulty associated with the whole genome shotgun assembly is that the DNA was derived from a thousand embryos, resulting in a very high level of polymorphism within the source DNA. Robust draft assemblies, tied to fingerprint maps, started to appear in 2003 with an estimated completion date of 2005. While its genome is not yet as complete as that of Fugu, there are both genetic linkage maps (Knapik *et al*., 1998; Hukriede *et al*., 1999) and radiation hybrid panels (Gates *et al*., 1999; Kelly *et al*., 2000) available for zebrafish, and in addition there is a large community of experimental biologists working on its biology. A large number of ESTs have also been sequenced, and as a result, the development of comprehensive zebrafish microarrays is well under way (Lo *et al*., 2003). The zebrafish is from the Ostariophysi, whereas the pufferfish are from the Acanthopterygii. While lack of fossil records makes it difficult to date this divergence, these two Superorders diverged very early in the teleost radiation, at least 100–150 and possibly over 200 million years ago. From an evolutionary point of view, the relative positions of these two fish are extremely fortuitous, as they are divergent enough to make genomic comparison informative, while capturing a significant proportion of teleosts between them. Consequently, if gene order across a region is conserved between Fugu and zebrafish, it is likely that it will also be syntenic in other teleost fish. The combination of genomics and experimental biology using these two fish provides some exciting opportunities and has already been used in order to dissect the regulatory region of the *SCL* gene (Barton *et al*., 2001).

With the sequence data generated from both the Fugu and zebrafish genomes, a pattern of genome duplication within the teleost lineage is emerging. This duplication must have been early to have occurred in both genomes and it is

now thought that the duplication might have been formative in the evolution of the teleost lineage itself (Taylor *et al*., 2003). This results in both zebrafish and Fugu genomes having a number of additional copies of genes, although it is not clear how many. A best guess estimate for the Fugu genome would be that it has retained about 10% of its duplicates, whereas this figure might be closer to 15% for the zebrafish. Further comparative analyses with the completed zebrafish genome should provide a more accurate estimate.

The Fugu and zebrafish genome sequences follow closely on the heels of the human genome and have provided templates for a variety of comparative analyses encompassing both coding and noncoding DNA. The completion of the zebrafish genome will not only improve the power of this form of analysis, but will also provide a tremendously valuable resource for the zebrafish experimental community.

References

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.

Barton LM, Gottgens B, Gering M, Gilbert JG, Grafham D, Rogers J, Bentley D, Patient R and Green AR (2001) Regulation of the stem cell leukemia (SCL) gene: A tale of two fishes. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6747–6752.

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B and Aparicio S (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature*, **366**, 265–268.

Clark MS, Edwards YJ, Peterson D, Clifton SW, Thompson AJ, Sasaki M, Suzuki Y, Kikuchi K, Watabe S, Kawakami K, *et al*. (2003) Fugu ESTs: New resources for transcription analysis and genome annotation. *Genome Research*, **13**, 2747–2753.

Elgar G, Clark MS, Meek S, Smith S, Warner S, Edwards YJ, Bouchireb N, Cottage A, Yeo GS, Umrania Y, *et al*. (1999) Generation and analysis of 25 Mb of genomic data from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Research*, **9**, 960–971.

Elgar G (2004) Identification and analysis of cis-regulatory elements in development using comparative genomics with the pufferfish, *Fugu rubripes*. *Seminars in Cell and Developmental Biology*, **15**, 715–719.

Gates MA, Kim L, Egan ES, Cardozo T, Sirotkin HI, Dougan ST, Lashkari D, Abagyan R, Schier AF and Talbot WS (1999) A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome Research*, **9**, 334–347.

Gilley J, Armes N and Fried M (1997) *Fugu* genome is not a good mammalian model. *Nature*, **385**, 305–306.

Hinegardner R (1968) Evolution of cellular DNA content in Teleost fishes. *The American Naturalist*, **102**, 517–523.

Hukriede NA, Joly L, Tsang M, Miles J, Tellis P, Epstein JA, Barbazuk WB, Li FN, Paw B, Postlethwait JH, *et al*. (1999) Radiation hybrid mapping of the zebrafish genome. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9745–9750.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al*. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

Kelly PD, Chu F, Woods IG, Ngo-Hazelett P, Cardozo T, Huang H, Kimm F, Liao L, Yan YL, Zhou Y, *et al*. (2000) Genetic linkage mapping of zebrafish genes and ESTs. *Genome Research*, **10**, 558–567.

Knapik EW, Goodman A, Ekker M, Chevrette M, Delgado J, Neuhauss S, Shimoda N, Driever W, Fishman MC and Jacob HJ (1998) A microsatellite genetic linkage map for zebrafish (*Danio rerio*). *Nature Genetics*, **18**, 338–343.

Lo J, Lee S, Xu M, Liu F, Ruan H, Eun A, He Y, Ma W, Wang W, Wen Z, *et al*. (2003) 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Research*, **13**, 455–466.

Marshall H, Studer M, Popperl H, Aparicio S, Kuroiwa A, Brenner S and Krumlauf R (1994) A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature*, **370**, 567–571.

Pozzoli U, Elgar G, Cagliani R, Riva L, Comi GP, Bresolin N, Bardoni A and Sironi M (2003) Comparative analysis of vertebrate dystrophin loci indicate intron gigantism as a common feature. *Genome Research*, **13**, 764–772.

Sandford R, Sgotto B, Aparicio S, Brenner S, Vaudin M, Wilson RK, Chissoe S, Pepin K, Bateman A, Chothia C, *et al*. (1997) Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Human Molecular Genetics*, **6**, 1483–1489.

Taylor JS, Braasch I, Frickey T, Meyer A and Van de Peer Y (2003) Genome duplication, a trait shared by ∼22,000 species of ray-finned fish. *Genome Research*, **13**, 382–390.

Trower MK, Orton SM, Purvis IJ, Sanseau P, Riley J, Christodoulou C, Burt D, See CG, Elgar G, Sherrington R, *et al*. (1996) Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 1366–1369.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, *et al*. (2005) Highly conserved non-coding sequences are associated with development. *PLoS Biology*, **3**, e7.

Short Specialist Review

The mouse genome sequence

Ian J. Jackson
Western General Hospital, Edinburgh, UK

The mouse, having long been established as the leading mammalian genetic model system (*see* Article 38, **Mouse models**, Volume 3), was highlighted early in the Human Genome Program as a priority candidate for genome sequencing. In 2002, a publicly funded consortium published a high-quality draft sequence from the mouse strain C57BL/6J (Mouse Genome Sequencing Consortium, 2002), which has become the reference sequence. The sequence was available via the Internet and utilized by many well before its publication. Publication of this sequencing effort was preceded by (and may well have been accelerated by) a commercial pay-for-use sequence that was a combination of a number of different strains (some data discussed in Mural *et al*., 2002).

The public effort used a whole genome shotgun (WGS) methodology, albeit integrated with a deep Bacterial Artificial Chromosome (BAC) contig (Gregory *et al*., 2002) by inclusion of the ends of these BACs in the shotgun assembly. The sequence could also be linked to high-resolution genetic maps through the BAC contig and through the many sequenced molecular markers used in the mapping. In recognition of the importance of the mouse genome, efforts continued in order to generate finished sequence from individual mouse BAC clones. The use of a WGS methodology led to the omission of some segmental duplications from the draft sequence, which have "collapsed" to single copy, or have been omitted from the assembly (Cheung *et al*., 2003; Bailey *et al*., 2004). Nevertheless, it appears that the mouse genome has less segmental duplication than the human sequence; 1–2% compared to 5–6% in humans.

The mouse genome sequence has considerable value in three areas:

- to facilitate the identification of the molecular basis of variation and mutation, principally through the identification and localization of genes in the sequence;
- as a means of discovering gene regulation and other control elements in the genome, by comparison with other vertebrate sequences;
- as a model for understanding the evolution of genomes.

Gene predictions using the draft sequence produce an estimate of mouse gene number of around 30 000, very similar to the human gene number (Mouse

Genome Sequencing Consortium, 2002; International Human Genome Sequencing Consortium, 2001). Not surprisingly, the vast majority (99%) of predicted mouse genes have homologs in the human genome. Perhaps more surprising is that only around 80% of mouse genes have clear 1:1 orthologs in the human sequence; that is, where a mouse gene has only a single human ortholog and vice versa. The absence of 1:1 orthology for so many genes is largely due to differential expansion of gene families in different species. Comparison with the rat genome sequence shows that substantial differential expansion has occurred even within the rodent lineage (Rat Genome Sequencing Project Consortium, 2004). Analysis of the function of gene families that have expanded in the mouse relative to humans suggests that the expansions have occurred through selection imposed by mouse- or rodent-specific lifestyles or behaviors. Thus, the mouse repertoire of olfactory receptors is greatly expanded relative to humans, although the receptor structures seem to cover the same range. The many more mouse receptors perhaps indicate that mice can better discriminate between closely related scents (Zhang and Firestein, 2002). Mice appear also to have expanded gene families, relative to humans, that encode certain reproductive functions, and others that are involved in innate immunity. The large cytochrome P450 gene family also shows numerous differences between mouse, humans, and rats, most likely because the toxin-metabolizing enzymes encoded by this family have been differentially selected in different species.

A high-quality genome sequence has greatly accelerated the identification of genes that are responsible for mouse mutant phenotypes (*see* Article 38, **Mouse models**, Volume 3). Before the sequence was available, the so-called positional cloning required that the candidate interval containing the mutation be reduced as much as reasonably possible by genetic mapping crosses. Mutation identification, now that the sequence is on hand, can begin with a lower resolution genetic cross from which all candidate genes in the interval can be identified and sequenced in the mutant strains. The much lower cost, and higher speed, of sequencing relative to animal husbandry has led to mutant gene identification using many fewer animals, which is an important animal welfare spin-off from the genome sequence.

A comparison of sequence from multiple inbred mouse strains shows that sequence differences between strains are not uniformly distributed across the genome (Wade *et al*., 2002). Instead, any pair of strains has large regions that have nucleotide differences (mainly single base differences) at an average rate of 1 per 250 bp, while other regions differ at only one base in 20 kb. Furthermore, comparison between any two strains will show a different genomic distribution of high and low variation than another pair. The laboratory mouse is a mixture of two subspecies of *Mus musculus*, *M. m. domesticus* and *M. m. musculus*. The distribution of variation probably reflects the origin of each genomic region within the strain; high variation being intersubspecific and low variation, intrasubspecific. One difference in 20 kb is a rather low level of intrasubspecies variation, when compared to the variation seen in other populations, such as humans, and probably reflects the small founder population of laboratory mice. This simplistic model leads to the hypothesis that much of the complex genetic diversity of phenotype between strains is due to their mosaic of subspecies origins. It may be possible to narrow the genomic interval containing particular Quantitative Traits Loci (QTLs) by comparing the pattern of sequence variation between multiple strains. A model

experiment that compared albino with pigmented strains has demonstrated that this approach can home in on within a few hundred kilobases of the tyrosinase gene, the gene that is known to cause the albino phenotype.

About 1.5% of the genome is coding sequence of genes and another 1% represents untranslated sequences of mRNA. When the mouse genome sequence is compared to the human sequence, about 5% aligns better than would be expected for a sequence that has randomly mutated over evolutionary time, and hence appears to be under selection. The excess of selected sequence over gene sequences probably consists of transcriptional and other regulatory elements (Dermitzakis *et al*., 2002). Comparison of the mouse and human with additional genomes gives more specificity to the comparison and is able to select potential regulatory regions. As an example of the utility of sequence comparisons, the mouse, human, and Fugu sequence around the *Pax6* locus has been compared and potential long-range regulatory elements identified. By generating transgenic mice in which these elements have been linked to reporter genes, Kleinjan *et al*. (2001) have demonstrated that they do indeed have regulatory potential.

As species have diverged during evolution, their respective genomes have undergone large-scale alterations, so that their gene content and gene order may stay roughly the same at a megabase scale, but at a larger, chromosomal scale there has been rearrangements of genomic segments. A comparison between the mouse and human genomes shows that there have been about 300 chromosomal rearrangements during the 75 million years of evolution that separated these species (Mouse Genome Sequencing Consortium, 2002). When the rat genome is analyzed in addition, it can be seen that the rate of rearrangement in the rodent lineage overall is much greater than in the evolutionary path to humans (Rat Genome Sequencing Project Consortium, 2004). When the rate of neutral sequence base changes is examined in these same lineages, it too is about twice as fast in the path to rodents than that to humans from their last common ancestor. Possibly, a more rapid generation time in rodents and their ancestors has resulted in more sequence changes when measured against time.

Like other sequenced mammalian genomes, a substantial fraction of the 2500 Mb of the mouse genome is made up of repetitive DNA. Most repetitive DNA is derived from transposable elements that have spread throughout the genome at various points in evolution. These transposons spread by multiplication of individual elements, which means that they can be categorized on the basis of their sequences into families. Comparison with the rat and human genomes allows these families to be classified as mouse specific, rodent-lineage specific, and ancestral; the last group being sequences that can be recognized as shared between all three mammals and must have been present in their last common ancestor (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004). About one-third of the mouse genome is made up of repetitive DNA that arose since humans and rodents diverged, of which about 350 Mb, or 14% of the total, is specific to the mouse lineage. Only about 5% of the genome is recognizable as ancestral repetitive DNA, apparently a much smaller fraction than the 22% seen in the human genome. However, the more rapid rate of sequence changes in the rodent lineage has probably caused the drift of ancestral sequences so that they are no longer recognizable.

In summary, the mouse genome is about 14% smaller than that of humans, although it has about the same gene content. Almost all of the 30 000 genes have orthologs in humans. The smaller DNA content of the mouse genome is not due to less-active recent transposition of repetitive sequences; in fact, there has been more activity in the mouse. Rather, there has been more deletion of nonfunctional DNA in the rodent lineage, which has removed ancestral as well as recent repetitive DNA.

The utility of the genome sequence has already been demonstrated many times by geneticists who have used it in the identification of genes underlying mutant or variant phenotypes. As more mammalian genomes are sequenced, the power of comparative genomics becomes more apparent, and a high-quality finished sequence should further promote the mouse as the principal mammalian model organism.

Further reading

Bradley A (2002) Mining the mouse genome. *Nature*, **420**, 512–514.

Boguski MS (2002) The mouse that roared. *Nature*, **420**, 515–516.

Nadeau JH (2002) Tackling complexity. *Nature*, **420**, 517–518.

References

Bailey JA, Church DM, Ventura M, Rocchi M and Eichler EE (2004) Analysis of segmental duplications and genome assembly in the mouse. *Genome Research*, **14**, 789–801.

Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, Heng HH, Koop BF and Scherer SW (2003) Recent segmental and gene duplications in the mouse genome. *Genome Biology*, **4**, R47.

Dermitzakis ET Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV and Antonarakis SE (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al*. (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kleinjan DA, Seawright A, Schedl A, Quinlan RA, Danes S and van Heyningen V (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Human Molecular Genetics*, **10**, 2049–2059.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, *et al*. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, **296**, 1661–1671.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Wade CM, Kulbokas EJ 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K and Daly MJ (2002) The mosaic structure of variation in the laboratory mouse genome. *Nature*, **420**, 574–578.

Zhang X and Firestein S (2002) The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience*, **5**, 124–133.

Short Specialist Review

Comparative sequencing of vertebrate genomes

Matthew E. Portnoy and Eric D. Green
National Human Genome Research Institute, Bethesda, MD, USA

1. Introduction

The past decade has brought astonishing growth in the generation of sequence data from eukaryotic genomes. This has largely been catalyzed by major technological and strategic advances in large-scale DNA sequencing (Green, 2001a), coupled with the intense effort to complete the Human Genome Project (*see* Article 24, **The Human Genome Project**, Volume 3) and, in particular, to finish the sequence of the first vertebrate genome – that of *Homo sapiens* (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001).

With a complete human genome sequence now available, attention has rapidly turned to understanding the functional information it encodes. Significant advances have been made in identifying the protein-coding portion of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001); however, this portion only reflects an estimated 1–2% of the ∼2.9-Gb human genome sequence. Importantly, an additional 3–4% of the human genome appears to be functional, but does not code for protein (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004); these sequences include elements that provide temporal and spatial control of gene expression (Wasserman and Sandelin, 2004) as well as those involved in chromosome dynamics. It is now apparent that a comprehensive cataloging of all functional elements in the human genome, especially those that do not directly code for protein, will require a multifaceted approach, involving the generation of additional laboratory- and computational-based data and the development of new paradigms for assimilating and analyzing the resulting complex data sets.

One of the most powerful approaches for identifying functional genomic elements involves the comparison of genome sequences from species at distinct evolutionary positions (Miller *et al*., 2004; Nobrega and Pennacchio, 2004; Boffelli *et al*., 2004; Pennacchio, 2003; Hardison, 2003). The resulting information provides a working knowledge of the precise sequence-level similarities and differences among genomes, which in turn can be used to gain insight about genome function. For example, sequences found to be common (or conserved) among species separated by large evolutionary distances (e.g., >50–100 million years) can be

considered candidates for serving a functional role; the process of identifying such conserved sequences has been termed *phylogenetic footprinting* (Duret and Bucher, 1997; Weitzman, 2003). In contrast, sequences found to be different among closely related species (e.g., primates) can be considered less likely to be functional; the process of "eliminating" such sequences from consideration (thereby leaving the remaining sequences as candidates for serving a functional role) has been termed *phylogenetic shadowing* (Boffelli *et al*., 2003; Boffelli *et al*., 2004). In short, strategies have emerged that involve the use of genome sequences from both closely and distantly related species to extract functional information by comparative sequence analysis.

Two complementary approaches have been used in recent years to generate vertebrate genome sequences (Green, 2001a) en route to comparative analyses. In whole-genome sequencing projects, data are generated across an entire species' genome (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001; Mouse Genome Sequencing Consortium, 2002; Aparicio *et al*., 2002; Rat Genome Sequencing Project Consortium, 2004); while such efforts are comprehensive with respect to the individual genome, they are limited in terms of the total number of different genomes that can be compared (because of the costs associated with sequencing an entire vertebrate genome). A subset of the whole-genome sequencing projects performed to date has involved species that are used as experimental models (the so-called reference genomes, such as mouse, rat, and zebrafish; *see* Article 39, **The rat as a model physiological system**, Volume 3, Article 46, **The Fugu and Zebrafish genomes**, Volume 3, and Article 47, **The mouse genome sequence**, Volume 3, respectively). In targeted sequencing projects, data are generated for discrete genomic regions, typically from multiple species (Thomas *et al*., 2003); while such efforts examine only a limited portion of the genome, they result in sequence comparisons that involve large collections of evolutionarily diverse species. The major vertebrate whole-genome and targeted sequencing efforts are listed at **www.intlgenome.org**.

2. Vertebrate whole-genome sequences

The central goals of the Human Genome Project included the generation of foundational information about the human genome and that of a handful of carefully selected other species, in particular, commonly used experimental models (Green, 2001b; *see also* Article 24, **The Human Genome Project**, Volume 3). Initially, only human and mouse were included among the vertebrates, but more recently that list has grown substantially (Table 1).

A finished human genome sequence was generated using a clone-based shotgun-sequencing strategy (International Human Genome Sequencing Consortium, 2001; *see also* Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3), whereby minimal overlapping sets of large-insert clones (mostly bacterial-artificial chromosome (BAC) clones) were individually subjected to random shotgun sequencing, followed by directed finishing (Green, 2001a). The prior construction of a BAC-based physical map of the human genome (McPherson

Table 1 Vertebrate whole-genome sequences

Common name	Species	Sequencing strategy[a]	Approximate coverage[b]	URL
Human	*Homo sapiens*	HSS	99%	**www.ncbi.nlm.gov/genome/guide/human** **genome.ucsc.edu** **www.ensembl.org/Homo_sapiens**
Chimpanzee	*Pan troglodytes*	WGS	4X	**genome.wustl.edu** **www.broad.mit.edu**
Rhesus macaque	*Macaca mulatta*	WGS/HSS	In progress	**www.hgsc.bcm.tmc.edu**
Cow	*Bos taurus*	WGS/HSS	In progress	**www.hgsc.bcm.tmc.edu**
Dog	*Canis familiaris*	WGS	In progress	**www.broad.mit.edu**
Rat	*Rattus norvegicus*	WGS/HSS	>90%	**www.hgsc.bcm.tmc.edu**
Mouse	*Mus musculus*	WGS/HSS	90–96%	**www.ncbi.nlm.gov/genome/guide/mouse** **genome.ucsc.edu** **www.ensembl.org/Mus_musculus**
Laboratory opossum	*Monodelphis domestica*	WGS	In progress	**www.broad.mit.edu**
Chicken	*Gallus gallus*	WGS	6.6X	**genome.wustl.edu**
Frog	*Xenopus tropicalis*	WGS	In progress	**www.jgi.doe.gov**
Zebrafish	*Danio rerio*	WGS/HSS	5.7X	**www.sanger.ac.uk**
Fugu	*Takifugu rubripes*	WGS	5.7X	**www.jgi.doe.gov**
Pufferfish	*Tetraodon nigroviridis*	WGS	6X	**www.genoscope.cns.fr**

[a]HSS = clone-based hierarchical shotgun sequencing (*see* Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3); WGS = whole-genome shotgun sequencing.
[b]Sequence coverage indicated as redundant coverage (e.g., 4X) or percent of total covered by assembled sequence (e.g., 99%).

et al., 2001) provided a key organizing framework for the long-range assembly of the whole-genome sequence.

Sequencing the genome of the most widely used experimental mammal – the laboratory mouse – began prior to completion of the Human Genome Project (Table 1; *see also* Article 47, **The mouse genome sequence**, Volume 3). The initial phase of this effort involved whole-genome shotgun sequencing and initial assembly (Mouse Genome Sequencing Consortium, 2002), with BAC-based finishing now ongoing. The mouse's critical role in biomedical research has prompted the finishing of its genome sequence to roughly the same quality as the completed human genome sequence. Interestingly, initial comparisons reveal that roughly 40% of the mouse genome sequence aligns with the corresponding (or orthologous) regions of the human genome sequence (Mouse Genome Sequencing Consortium, 2002), but only a small minority of that aligning sequence (totaling roughly 5% of the mouse or human genome) is actively conserved and presumed to be functionally important.

The third mammalian whole-genome sequence to be generated was that of the rat (Rat Genome Sequencing Project Consortium, 2004) (Table 1). This project

utilized an integrated whole-genome and BAC-based shotgun-sequencing strategy, which yielded a very high-quality draft sequence. However, at present, there are no plans to finish the rat genome sequence to the same quality standards as were used for the human or mouse genome sequences. Indeed, until the costs of sequence finishing decrease substantially, it is unclear which other vertebrate genomes (if any) will be sequenced as accurately and completely as the human genome.

Whole-genome draft sequences have also been generated for a trio of fish species: zebrafish (*Danio rerio*) and two types of pufferfish, *Fugu rubripes* (Aparicio *et al*., 2002) and *Tetraodon nigroviridis* (Table 1) (*see* Article 46, **The Fugu and Zebrafish genomes**, Volume 3). The zebrafish genome sequence will be critical for the rapidly growing research community using these fish as experimental models, whereas the two pufferfish species were selected for their notably compact genomes (among the smallest of all vertebrates). Among the primates, a whole-genome draft sequence has been generated for the chimpanzee, our closest relative in evolution, while that of the rhesus macaque is actively being produced. The current list of available or planned vertebrate whole-genome sequences includes that of two additional nonprimate eutherian mammals (dog and cow), a marsupial (the laboratory opossum *Monodelphis domestica*), a bird (the chicken), and an amphibian (the frog *Xenopus tropicalis*) (Table 1). As the costs of large-scale DNA sequencing decline, the list of generated vertebrate whole-genome sequences (such as in Table 1) will inevitably grow, with the strategies for generating and analyzing that sequence likely changing as well (see below).

The generation of an ever-growing set of vertebrate genome sequences, coupled with initial efforts to compare them in a rigorous fashion, has yielded spectacular amounts of data. To ensure that this information is readily accessible and comprehensible, especially to the general biomedical research community, several groups have developed "genome browser" systems. These consist of convenient navigational tools for accessing and utilizing the genomic sequence of different vertebrates as well as organized frameworks for assimilating relevant annotations, including those emanating from comparative analyses. The three most widely used genome browsers are the UCSC Genome Browser (**genome.ucsc.edu**), Ensembl (**www.ensembl.org**), and NCBI Map Viewer (**www.ncbi.nlm.nih.gov/mapview**). As a representative example, Figure 1 depicts a ∼150-kb segment of the human genome, as displayed by the UCSC Genome Browser. Note the ability to observe simultaneously the structure of known genes in the region (*RET* and *GALNACT-2*) as well as various other types of information (e.g., promoters, repetitive elements, and the results of comparative analyses with a handful of other species). Indeed, this figure illustrates how genomes will likely be viewed in the coming years, as increasingly detailed information about genes, other functional elements, regions of sequence conserved among various species, and other genome features of interest are assimilated layer by layer.

3. Multivertebrate sequences of targeted genomic regions

While the above whole-genome sequencing efforts are providing valuable data for comparative analyses, they are limited in the total number of vertebrates being

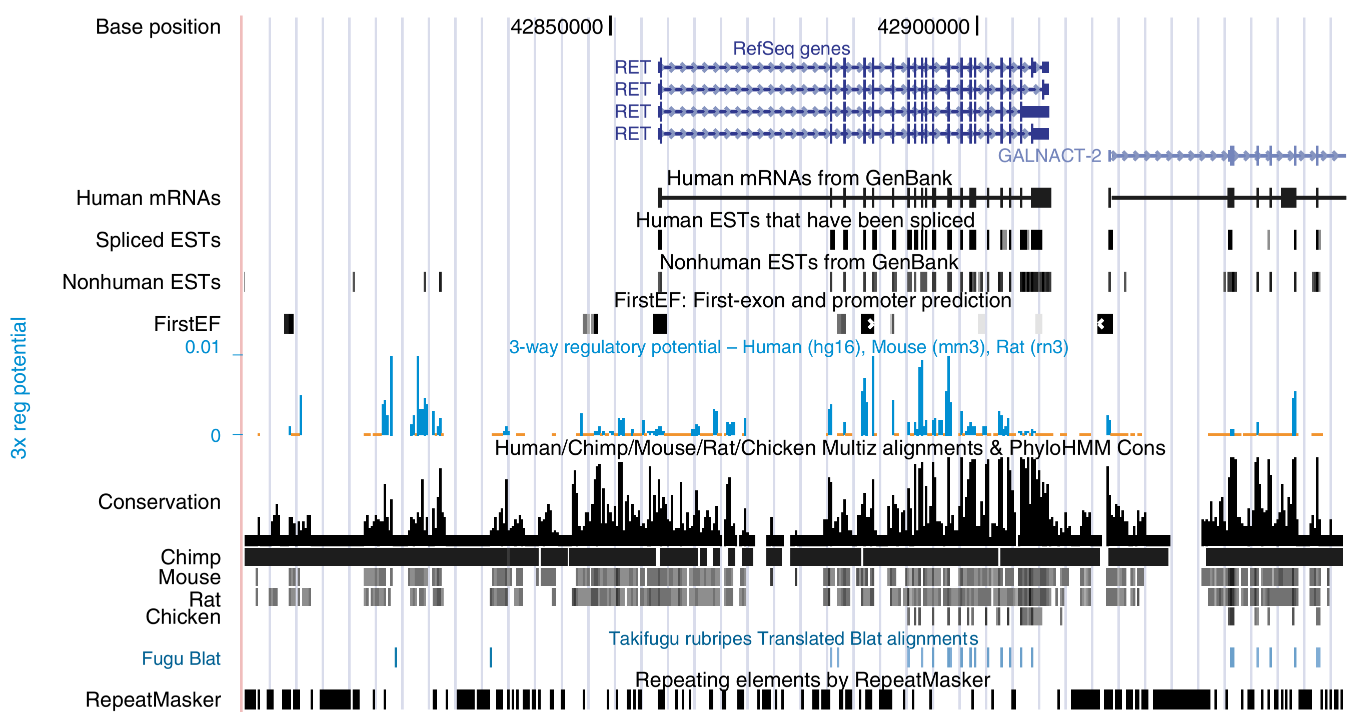


Figure 1 View of a roughly 150-kb segment of the human genome, as displayed on the UCSC Genome Browser. A segment of human chromosome 10 encompassing the *RET* gene is shown (July 2003 build of the human genome sequence, UCSC version hg16/NCBI build 34; coordinates chr10:42,800,001-42,950,000). Various annotations are displayed; in particular, note the tracks depicting the results of comparative sequence analyses (see **genome.ucsc.edu** for details)

studied (see Table 1). To complement these projects, the sequence of smaller, targeted genomic regions can be generated from a greater number of species, resulting in comparative sequence analyses with larger, more evolutionarily diverse collections of vertebrates (Thomas and Touchman, 2002). For example, the NISC Comparative Sequencing Program (see **www.nisc.nih.gov**) is currently sequencing more than 150 targeted regions of the human genome in multiple vertebrates, in some cases generating orthologous sequence data from over 30 species (Thomas *et al*., 2003).

These studies have already yielded some interesting findings. First, the resulting data have reflected the first-available genomic sequence for a number of vertebrates, providing new insights about the genetic blueprints of these species. These have included information about gene density, the relative degree of genome compression/expansion, the amounts and types of repetitive sequences, the extent and types of mutational events that have uniquely sculpted each genome, and the general patterns of conservation seen upon comparison with other species' sequences (Thomas *et al*., 2003). Second, the generation of orthologous sequences from large sets of vertebrates has catalyzed the development of computational methods for multispecies comparative sequence analyses. For example, new approaches have been developed for identifying sequences that are highly conserved across multiple species (called *Multispecies Conserved Sequences* or *MCSs*) (Margulies *et al*., 2003; Margulies *et al*., 2004). Interestingly, vertebrates differ with respect to the effectiveness of their sequence for detecting MCSs in the human genome (Margulies *et al*., 2003; Margulies *et al*., 2004). Finally, targeted sequencing projects are particularly well suited for genome-evolution studies because they can readily yield sequence data from carefully selected species of interest (Thomas *et al*., 2003). Such studies can include in-depth surveying of multiple species at a particular phylogenetic node (e.g., primates), which, at least for the foreseeable future, is not possible with whole-genome sequence data sets.

Multivertebrate sequencing of targeted genomic regions is playing an important role in the recently launched ENCODE (Encyclopedia of DNA Elements) project, which aims to identify all of the functional elements in the human genome (**genome.gov/ENCODE**). ENCODE's initial goal is to catalog comprehensively the functional elements in a selected 1% (~30 Mb) of the human genome, using a diverse set of experimental and computational approaches. These targeted ~30 Mb, which are distributed across 44 different genomic regions, are being sequenced in multiple vertebrates. The resulting sequences will be subjected to myriad comparative analyses, with the results in turn compared to various other types of data (computational and experimental) generated for the same ~30 Mb. Eventually, this process should provide important insights into the utility of multispecies sequence comparisons for unraveling the complexities of genome function.

4. Deducing genome function through sequence comparisons

Comparative analyses of whole-genome and targeted-genome sequences from various vertebrates have been shown to be valuable for the study of genome function. Simple alignments of orthologous sequences from two or more genomes

can be used to identify the presence and structure of genes (Batzoglou *et al*., 2000; Miller *et al*., 2004). More refined approaches have been developed for gene prediction that involve sequence comparisons; for example, using the human and mouse genome sequences, TWINSCAN (Korf *et al*., 2001) and SLAM (Alexandersson *et al*., 2003) have been used to produce a conservative estimate of 25,622 genes in the human genome (Flicek *et al*., 2003) and to detect roughly 80% of the predicted human exons in the NCBI RefSeq gene collection (see **www.ncbi.nlm.nih.gov/RefSeq**) (Alexandersson *et al*., 2003). In a more targeted fashion, human–mouse sequence comparisons directly led to the discovery of the apolipoprotein A5 gene (*APOA5*) (Pennacchio *et al*., 2001); subsequent functional studies showed the importance of *APOA5* in regulating triglyceride levels (Pennacchio, 2003).

Comparative sequence analyses are also proving critical for detecting conserved sequences outside of coding regions, which are candidates for functional noncoding elements (e.g., such as those regulating gene expression). For example, phylogenetic footprinting of sequences from a set of diverse mammals identified several regulatory elements upstream of the ε-globin gene (Gumucio *et al*., 1993). Similarly, human–mouse sequence comparisons of the interleukin gene cluster identified several conserved noncoding sequences, the longest of which was shown to be a cis coactivator of several nearby interleukin genes (Hardison, 2000; Loots *et al*., 2000). More global methods have now been developed for identifying sequences that are most highly conserved across multiple species (Margulies *et al*., 2003; Dermitzakis *et al*., 2003), many of which are likely to be functionally important. Interestingly, in the human genome, there are more such highly conserved sequences within noncoding regions compared to coding regions (Margulies *et al*., 2003; Margulies *et al*., 2004). By a different strategy (specifically, phylogenetic-shadowing methods), sequence comparisons involving closely related species have been used to identify potential regulatory elements (Boffelli *et al*., 2003; Boffelli *et al*., 2004).

5. Future prospects

The landscape of comparative vertebrate sequencing is changing rapidly. The major commitments to date for whole-genome sequencing mostly involve vertebrates associated with large research communities that will directly exploit the resulting sequence data. As such, the genome sequences of species such as human (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001), mouse (Mouse Genome Sequencing Consortium, 2002), rat (Rat Genome Sequencing Project Consortium, 2004), chicken, zebrafish, *Xenopus*, dog, and cow provide reference information of great value. In addition, each whole-genome sequence provides secondary value by contributing to an ever-expanding repertoire of comparative sequence analyses, which more broadly advance our knowledge of complex genomes. However, only a few remaining vertebrates can be regarded as true reference species, and thus a primary rationale for most future genome-sequencing projects will be the acquisition of data for comparative studies.

The current plans for vertebrate genome sequencing largely reflect these changing priorities. Ongoing or soon-to-be-initiated sequencing projects include a wider

sampling of vertebrates across the phylogenetic tree and exploration of a larger set of primates. Significantly, recent findings indicate that the identification of highly conserved sequences from evolutionarily diverse species can be accomplished with lower-quality draft sequence. Specifically, comparative analyses using low-redundancy genomic sequences (e.g., providing one- to twofold coverage) from a larger number of species appear to be more effective in identifying the most highly conserved genomic elements than those using high-redundancy sequences (e.g., providing two- to fourfold coverage) from a smaller number of species (unpublished data). These findings are prompting efforts to acquire low-redundancy genomic sequence from a large, diverse group of vertebrates. Such an endeavor would not be for the purpose of generating an assembled sequence of each genome but rather to amass a large data set that can be used predominantly for comparative analyses. This approach reflects a strategic shift from the sequencing projects performed under the auspices of the Human Genome Project, but one that resonates with the high-priority efforts to interpret the human genome sequence in a comprehensive fashion.

More futuristic views of comparative vertebrate sequencing largely depend on the knowledge of the relative costs of large-scale DNA sequencing. Should those costs continue to decline substantially, then the genomes of much larger collections of vertebrates (and indeed invertebrates as well) would inevitably be sequenced, with the resulting massive data sets greatly empowering comparative studies. Regardless, the lessons learned to date clearly indicate that the sequence of each species' genome contains a treasure trove of information about evolutionary history and that comparisons of those histories are critical for understanding the complexities of genome structure and function.

Acknowledgments

We thank Elliott Margulies, Bob Blakesley, Nancy Hansen, and Monica Janossy for the critical reading of this chapter.

References

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**, 1301–1310.

Alexandersson M, Cawley S and Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, **13**, 496–502.

Batzoglou S, Pachter L, Mesirov JP, Berger B and Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, **10**, 950–958.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L and Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Boffelli D, Nobrega MA and Rubin EM (2004) Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics*, **5**, 456–465.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C and Antonarakis SE (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, **302**, 1033–1035.

Duret L and Bucher P (1997) Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, **7**, 399–406.

Flicek P, Keibler E, Hu P, Korf I and Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Research*, **13**, 46–54.

Green ED (2001a) Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics*, **2**, 573–583.

Green ED (2001b) The human genome project and its impact on the study of human disease. In *The Metabolic and Molecular Bases of Inherited Disease*, Eighth Edition, Scriver CR, Beaudet AL, Sly WS, Valle D, Childs B, Kinzler KW and Vogelstein B (Eds.), McGraw-Hill: New York, NY, pp. 259–298.

Gumucio DL, Shelton DA, Bailey WJ, Slightom JL and Goodman M (1993) Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 6018–6022.

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics*, **16**, 369–372.

Hardison RC (2003) Comparative Genomics. *Public Library of Science Biology*, **1**, 156–160.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM and Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Margulies EH, Blanchette M, Haussler D and Green ED (2003) Identification and characterization of multi-species conserved sequences. *Genome Research*, **13**, 2507–2518.

Margulies EH, NISC Comparative Sequencing Program and Green ED (2004) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harbor Symposia on Quantitative Biology*, *Vol 68: The Genome of Homo Sapiens*, CSHL Press: Woodbury, NY, pp. 255–263.

Miller W, Makova KD, Nekrutenko A and Hardison RC (2004) Comparative genomics. *Annual Review of Genomics*, **5**, 15–56.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Nobrega MA and Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. *Journal of Physiology*, **554**, 31–39.

Pennacchio LA (2003) Insights from human/mouse genome comparisons. *Mammalian Genome*, **14**, 429–436.

Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM and Rubin EM (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*, **294**, 169–173.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Thomas JW and Touchman JW (2002) Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends in Genetics*, **18**, 104–108.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, **5**, 276–287.

Weitzman JB (2003) Tracking evolution's footprints in the genome. *Journal of Biology*, **2**, 9.

Short Specialist Review

The chimpanzee genome

Tarjei S. Mikkelsen
Broad Institute of MIT and Harvard, Cambridge, MA, US

1. Introduction

As our closest, extant evolutionary relative, the common chimpanzee (*Pan troglodytes*) offers a unique perspective on the human species and its history. All heritable, biological traits unique to our species, such as distinct anatomy, cognitive capacities, and some disease susceptibilities are ultimately caused by one or more discrete differences between the human and chimpanzee genomes. Comparative analysis can help reveal these differences, as well as the mutational processes and selective pressures that have generated them.

An initial draft sequence of the chimpanzee genome, based on 4x whole-genome shotgun (WGS) coverage (*see* Article 4, **Sequencing templates – shotgun clone isolation versus amplification approaches**, Volume 3) of a single donor, has been made publicly available by a US-based consortium (Chimpanzee Sequencing and Analysis Consortium, 2005). Additional WGS sequencing, as well as efforts to construct a BAC-based physical map, are underway. The draft WGS sequence assembly is supplied with nucleotide quality scores indicating the presence of potential sequencing errors (*see* Article 11, **Algorithms for sequence errors**, Volume 7), which are particularly important to take into account when comparing closely related species.

A growing number of additional genomic resources are also available, including a BAC-based assembly of chromosome 21 (formerly chromosome 22; McConkey, 2004) from a different individual (Watanabe, 2004); two chromosome Y sequences (Hughes *et al*., 2005; Kuroki, 2006); PCR amplified exons from over 13,000 known genes (Nielsen, 2005); cDNA sequences (Hellmann, 2003); and light WGS coverage of additional West African and Central African chimpanzees (Chimpanzee Sequencing and Analysis Consortium, 2005).

2. Magnitude and patterns of sequence divergence

Because of the relatively short time since the divergence of humans and chimpanzees, most nucleotides in our genomes are identical by descent, and an observed difference nearly always represents a single mutation. Most of the differences reflect random genetic drift, and thus they hold extensive information

about the mutational processes that have shaped our genomes in recent evolutionary history.

Single nucleotide substitutions are the most abundant type of differences. Overall, 1.23% of orthologous nucleotides differ, of which 85% or less represent fixed interspecies divergence, and the rest are due to intraspecies polymorphism. Substitutions are not uniformly distributed throughout the sequences, largely reflecting context-dependent variation in mutation rates. For example, although CpG dinucleotides constitute only 2% of the genomes, they account for a quarter of all nucleotide substitutions. On a larger scale, orthologous sequences situated within 10 Mb of a telomere have, on average, accumulated 15% more substitutions than the rest of the genome.

Nucleotide insertions and deletions (indels) are less abundant than substitutions, but affect significantly more sequence overall. In total, 5 to 6 million indels have resulted in the human and chimpanzee genomes each containing 40 to 45 Mb of euchromatic sequence not present in the other. The vast majority of indels are small (98.6% are shorter than 80 bp), but the largest few contain most of the affected sequence (approximately 70000 indels longer than 80 bp constitute 75% of the linage-specific sequences).

Transposable elements are the source of distinct indels in both the human and chimpanzee genomes. The major difference is the emergence of large, subterminal caps of satellite repeats on chimpanzee chromosomes (Yunis and Prakash, 1982). The euchromatic sequences also show evidence of linage-specific insertions of all major classes of transposable elements. The primate-specific Alu element has been three-fold more active in the human genome, whereas LINE-1 elements have been inserted at similar rates in both genomes. Transposable element insertions may have affected expression or splicing patterns of nearby genes (Britten, 1997).

Large-scale chromosomal rearrangements make up the least abundant, but most dramatic, differences between the two genomes. Early cytogenetic characterization revealed 9 pericentric inversions between human and chimpanzee chromosomes and a fusion of two ancestral chromosomes in the human lineage (Yunis and Prakash, 1982). Surveys of structural variation aided by the WGS assembly have refined the localization of these rearrangements, and revealed several hundred additional chromosomal inversions, segmental duplications and deletions (Newman *et al*., 2005). A significant fraction overlap known genes, and have consequently contributed to differences in expression levels of these genes between the two species.

3. Signatures of natural selection

Mutations provide the substrate upon which natural selection acts to mould the evolution of a species. The impact of natural selection can be inferred from patterns of sequence divergence that deviate from those expected under neutral drift (*see* Article 9, **Modeling protein evolution**, Volume 1).

Signatures of negative selection, or removal of deleterious alleles from a population, are easily recognizable by comparing orthologous genes in humans and chimpanzees. The average protein coding gene has accumulated only a single

amino acid substitution in each of the human and chimpanzee lineages since our divergence, and the mean ratio of nonsynonymous to synonymous substitutions across all orthologs is 0.23, indicating that at least 77% of amino-acid changing mutations are sufficiently deleterious to be removed by natural selection. This ratio is approximately 35% higher than observed between mouse and rat orthologs, indicating a general relaxation of evolutionary constraints in the primate lineages relative to the murid lineages, likely due to smaller effective population sizes, and consequently a greater impact of genetic drift, in humans and chimpanzees.

Signatures of positive selection, or rapid fixation of advantageous alleles, are more challenging to detect, but of even greater interest. Although effectively neutral alleles may well have phenotypic effects, it is generally thought that signatures of positive selection can help pinpoint the genetic changes most critical to our evolutionary history. The low divergence rates between human and chimpanzee orthologs greatly limit the power of statistical test for positive selection in any single gene. Grouping genes into relevant categories, such as cellular function or pathway membership, can increase statistical power at the cost of lower resolution. Orthologs involved in the immune- and reproductive systems dominate the categories showing an excess of nonsynonymous over synonymous substitutions, the most stringent test for adaptive selection. This is a common observation in higher organisms studied so far, reflecting sustained selective pressures on these systems throughout evolution.

Lineage-specific acceleration of rates of evolution can reveal more subtle signatures of positive selection. Accounting for the overall relaxation of constraints in primates, the rates of evolution of primate and murid orthologs are highly correlated, but there is detectable acceleration in some functional groups. Genes involved in spermatogenesis and the male reproductive system show the most significant primate-specific acceleration, potentially reflecting a particularly strong influence of sexual selection on primate evolution. There is significantly less evidence of lineage-specific acceleration between human and chimpanzee orthologs when the murid genomes are used as outgroups, reflecting more similar selection pressures within the primate lineages. But, notably, the functional category that shows the strongest acceleration in the human, compared to the chimpanzee, is the transcription factor, supporting the hypothesis that changes in gene regulation may be a key factor underlying rapid anatomical evolution in the human lineage (King and Wilson, 1975).

4. Applications and prospects

There is much to be learned from comparison of the human and chimpanzee genome sequences beyond what has been gleaned from initial surveys.

The chimpanzee genome sequence has a special role in informing studies of human population genetics (*see* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1). In particular, it was used to validate novel single nucleotide polymorphisms (SNPs) for a human haplotype map (International HapMap Consortium, 2005). The sequence can also be used to estimate regional mutation rates, and as an effective outgroup to classify segregating alleles in the human population as ancestral or derived. The initial WGS assembly

facilitated assignment of ancestral states to over 80% of all publicly available SNPs mapped to the human genome with 98% accuracy. High resolution maps of local mutation rates and ancestral allele assignments can be used to directly inform inferences about genetic drift and natural selection in recent human history (Article 7, **Genetic signatures of natural selection**, Volume 1).

In addition to informing scans for recent selection in anatomically modern human populations, comparative analysis of primate genomes provides a unique opportunity for elucidating the evolutionary history of the human and Africa great ape lineages close to the times of divergence. The chimpanzee genome likely holds important clues to the extent of gene flow and the mechanisms that generated reproductive barriers between our progenitor populations, as well as the pace and extent of adaptive evolution following speciation.

Most initial analyses of evolution and selection in the human and chimpanzee lineages have focused on protein coding sequences, and largely ignored other functional elements in the genomes. Progress in comparative genomics (*see* Article 48, **Comparative sequencing of vertebrate genomes**, Volume 3) and growing appreciation of the abundance and importance of *cis-regulatory* elements and noncoding RNA in mammalian genomes (*see* Article 27, **Noncoding RNAs in mammals**, Volume 3) is rapidly removing this bias. The chimpanzee genome sequence allows systematic analysis of recent evolution in noncoding sequences and their effects on gene expression and development. For example, the sequence can be used to design gene expression assays suitable for cross-species comparisons with hybridization probes specific to sequences that do not differ with human, or alternatively, to mask unsuitable probes on existing human microarrays.

The ultimate goal of the chimpanzee genome project is to identify the specific genetic alterations that underlie each phenotypic difference between humans and other primates. This is made particularly challenging by the practical and ethical limitations on any experimental investigation, but careful correlation of genetic differences with phenotypic and clinical data is contributing to a growing list of genotype-phenotype relationships (Varki and Altheide, 2005).

References

Britten RJ (1997) Mobile elements inserted in the distant past have taken on important functions. *Gene*, **205**, 177–182.

Chimpanzee Sequencing and Analysis Consortium (2005) The initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.

Hellmann I, *et al*. (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Research*, **13**, 831–837.

Hughes JF, Skaletsky H, Pyntikova T, Mix PJ, Graves T, Rozen S, Wilson RK and Page DC (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*, **437**, 100–103.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Kuroki Y, *et al*. (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathways. *Nature Genetics*, **32**, 158–167.

McConkey EH (2004) Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenetic and Genome Research*, **105**, 157–158.

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M and Eichler EE (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Research*, **15**, 1344–1356.

Nielsen R, *et al*. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, **3**, e170.

Varki A and Altheide TK (2005) Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res*, **15**, 1746–1758.

Watanabe H, *et al*. (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, **429**, 382–388.

Yunis JJ and Prakash O (1982) The origin of man: a chromosomal pictorial legacy. *Science*, **215**, 1525–1530.

Short Specialist Review

Functional annotation of the mouse genome: the challenge of phenotyping

Steve D. M. Brown
MRC Mammalian Genetics Unit, Harwell, UK

1. Functional annotation of the mammalian genome

With the completion of the human genome sequence, attention has turned to the functional annotation of the genes and other sequence elements that are encoded within the DNA. We remain largely ignorant of the detailed role of many genes in normal physiology, biochemistry and development and the genetic pathways involved. Moreover, there is a pressing need to elaborate the relationship between genes and disease susceptibility if the progress made in the human genome project is to be translated into a better understanding of pathophysiological mechanisms and a concomitant improvement in therapeutic approaches and health care. Deciphering the relationship between gene and phenotype in a mammalian organism represents one of the biggest challenges for genetics and biology in the twenty-first century. While some progress can be made through human genetic studies, studies of model organisms will be key to elaborating gene–phenotype space. In particular, the mouse will play a pivotal role as we embark upon a systematic and comprehensive functional annotation of the mammalian genome.

An extensive genetic toolbox has been developed for modifying the mouse genome, and we are now able to introduce mutations into coding and other sequences more or less at will. As a consequence, efforts are now underway to mutate every gene in the mouse genome (Austin *et al*., 2004; Auwerx *et al*., 2004), focusing mainly on a combination of two mutagenesis approaches – gene trapping (Stanford *et al*., 2001) and gene targeting (Glaser *et al*., 2005). The expectation is that over the next 5 years we will have access to comprehensive libraries of mouse mutants for every gene. Attention will then turn to mutating noncoding sequences, including transcribed noncoding sequences and putative regulatory elements. In order to interpret the relationship between gene (or DNA element) and phenotype, it is necessary, however, to determine the effects of each mutation on the various developmental and physiological pathways. It is also clear that any profound understanding of gene function will require that we undertake a comprehensive analysis of mouse phenotype that encompasses all adult body systems, as well as effects on developmental processes. Phenotyping even a minimal set of around

25 000 mutations, representing a single mutant allele at each gene in the mouse genome, will therefore be a phenomenal undertaking. While such a programme when complete would in itself represent a very significant milestone, it is only a useful beginning. It is also recognized that it will be important to generate and phenotype a number of mutant alleles at each gene locus with a range of potential effects if we are to develop a deep and systematic knowledge of gene function. Equally, the phenotypic effects of a mutant may vary according to genetic background, and there are persuasive arguments for determining the phenotype of each mutant on a variety of genetic backgrounds.

Given the enormous challenges involved in generating systematic datasets of phenotypes of mouse mutants, there has been much recent attention on the development of phenotyping approaches with a particular focus on the following:

- standardizing phenotyping approaches;
- developing high-throughput comprehensive phenotype screens;
- developing novel phenotyping platforms and new technological approaches to phenotyping;
- developing new standards for phenotype data representation.

Progress in each of these areas will bring us closer to our goal of providing a comprehensive phenotype database for gene function, which will be an intrinsic component of developing a systems biology of the mouse.

2. Standardization of phenotyping approaches

The phenotype assay is crucial to determining the measured output, and there is considerable evidence that the standard operating procedure (SOP) employed can have a marked impact upon the results of a particular test. The implication is that we need to standardize our approaches to phenotyping to ensure comparability of datasets across time and place. In addition, it is clear that environmental conditions, including cage environment (Tucci *et al*., 2006) and diet, may have a bearing upon the outcome of phenotype tests and need to be cataloged when acquiring phenotype data. Crabbe *et al*. (1999) found considerable variation between laboratories in behavioral test outcomes despite efforts to standardize procedures. Though the reasons remain unclear, it is possible that unrecognized factors in either test or environmental conditions contributed to the unwanted variation. Overall, it is clear that we need to study and document further the variables both in test and environment that contribute to variation in test outcome, and to further standardize the procedures for phenotyping platforms. Indeed, the Eumorphia project (European Union Mouse Research for Public Health and Industrial Applications, **http://www.eumorphia.org**, see below) has made a major effort to standardize and validate procedures for a variety of mouse phenotyping platforms both within and between laboratories (Brown, 2005). While many tests were validated, some SOPs demonstrated considerable variation in test output

between laboratories and thus require further examination and elimination of test variables.

3. Developing high-throughput comprehensive phenotype screens

One prevailing dogma in mouse phenotyping is to employ a hierarchical approach whereby initially rapid, comprehensive batteries of relatively unsophisticated tests are applied – the so-called primary screen. Subsequently, more sophisticated, but inevitably more time-consuming, secondary or tertiary screens may be carried out on specific animals depending upon the phenotypes revealed in the primary screen. The SHIRPA screen is an example of a test battery that employs this hierarchical approach (Rogers *et al*., 1997). However, there are no hard and fast boundaries between primary, secondary, and tertiary tests. Traditionally, there has been an inverse relationship between throughput and sophistication. However, one of the aims in mouse phenotyping is to encompass as many phenotype tests as possible within the envelope of the primary screen.

4. Developing novel phenotyping platforms and new technological approaches to phenotyping – the birth of the mouse clinic

There has been much progress in developing novel phenotyping platforms whereby we bring new technologies, along with advances in equipment and test design, to bear to enhance both the access and the throughput of even the most sophisticated tests. For example, even quite complex behavioral assays such as circadian rhythm screens have benefited enormously from automation and data capture and can be effectively utilized as primary screens. Microtechnologies and remote telemetric monitoring will aid a wide variety of phenotype procedures, while improvements in the application of a whole spectrum of imaging platforms (including MRI, SPECT, ultrasound, micro-CT) to the mouse are set to provide a new wealth of phenotype data (Brown *et al*., 2006). Individual systems have been the focus of innovative approaches, for example, the development of the optokinetic test in mice for the measurement of visual acuity (Thaung *et al*., 2002). New technologies such as Luminex will transform our ability to analyze a wide variety of blood proteins (de Jager *et al*., 2003). Moreover, the utilization of mice that carry reporter molecules will enhance cell lineage analysis and the monitoring of tissue structure.

The rapid expansion in the variety and complexity of phenotyping platforms poses a problem for the mouse genetics community. Even with the availability of well-documented and standardized procedures, it is unreasonable to expect every laboratory to have the expertise or the equipment to apply even a fraction of phenotyping platforms. The concept of the *mouse clinic* has emerged – institutions that possess a broad range of phenotyping expertise and offer these

as services to external users. Nevertheless, given the scale of the enterprise required to complete the functional annotation of the mouse genome, considerable investment in additional infrastructure will be required to meet future phenotyping demand.

5. New standards for phenotype data representation

Developing appropriate standards for representation of phenotype data is as critical as developing the standards in phenotype testing if we are to be able to apply the necessary computational tools to the datasets that will underpin and inform any systems analysis. However, the development of appropriate structures to represent phenotype data is only in the initial stage. One route that is being explored is to use ontological structures of the kind that have been employed for the Gene Ontology (Ashburner *et al*., 2000). Key to these developments in phenotype ontologies is the recognition that the phenotype assay is central to the description of phenotype – change the assay, even slightly, and the measured outcome may be different. For this reason, one approach that is being explored is to utilize a compound description of phenotype constructed from component ontologies that allows features such as the SOP, environmental conditions, genetic background, and so on, to be represented (Gkoutos *et al*., 2005).

Access to raw phenotype data is equally important and will allow us to mine important relationships in the context of genetic differences. To date, sizeable sets of raw phenotype data are available for the Mouse Phenome project ((Bogue and Grubb, 2004); **http://www.jax.org/phenome**) and at EuroPhenome (**http://www.europhenome.org**), where in both cases baseline data for a variety of inbred strains is available. Access to raw data coupled with standards on data representation and exchange will be critical if we are to harness the developments in phenotyping platforms and phenotype data acquisition to progress in dissecting gene–phenotype relationships and developing a systems biology of the mouse.

6. Conclusions

The challenges facing mouse genetics community as it embarks on the functional annotation of the mammalian genome are very considerable. First, there is the urgent need for further development of phenotyping platforms that will lead to improvements in the speed, cost, and sophistication of phenotyping tests. At the same time, it will be vital to ensure the standardization of phenotyping protocols, which will allow the generation of datasets from disparate research centers that can be shared and compared. Finally, we need to develop new informatics standards for data acquisition and representation that importantly include the phenotype test as a key variable. All of these challenges are being addressed as efforts get underway to undertake comprehensive phenotype analyses of mouse mutants from the worldwide mutagenesis programs.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppit JT, *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, Collins FS, Dove WF, Duyk G, Dymecki S, Eppig JT, *et al* (2004) The Knockout Mouse Project. *Nature Genetics*, **36**, 921–924.

Auwerx J, Avner P, Baldock R, Ballabio A, Balling R, Barbacid M, Berns A, Bradley A, Brown S, Carmeliet P, *et al* (2004) The European dimension for the mouse genome mutagenesis programme. *Nature Genetics*, **36**, 925–927.

Bogue MA and Grubb SC (2004) The mouse phenome project. *Genetica*, **122**, 71–74.

Brown SDM, The Eumorphia Consortium (2005) EMPRESS: standardised phenotype screens for functional annotation of the mouse genome. *Nature Genetics*, **37**, 1155.

Brown SDM, Hancock JM and Gates H (2006) Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. *PLoS Genetics*, **2**, e149.

Crabbe JC, Wahlsten D and Dudek BC (1999) Genetics of mouse behaviour: interactions with laboratory environment. *Science*, **284**, 1670–1672.

Gkoutos GV, Green ECJ, Mallon AM, Hancock JM and Davidson D (2005) Using ontologies to describe mouse phenotypes. *Genome Biology*, **6**, R8.

Glaser S, Anastassiadis K and Stewart AF (2005) Current issues in mouse genome engineering. *Nature Genetics*, **37**, 1187–1193.

de Jager W, te Velthuis H, Prakken BJ, Kuis W and Rijkers GT (2003) Simultaneous detection of 15 human cytokines in a single sample of stimulated peripheral blood mononuclear cells. *Clinical and Diagnostic Laboratory Immunology*, **10**, 133–139.

Rogers DC, Fisher EMC, Brown SDM, Peters J, Hunter AJ and Martin JE (1997) SHIRPA – A proposed protocol for the comprehensive behavioural and functional analysis of mouse phenotype. *Mammalian Genome*, **8**, 711–713.

Stanford WL, Cohn JB and Cordes SP (2001) Gene-trap mutagenesis: past, present and beyond. *Nature Reviews. Genetics*, **2**, 756–768.

Thaung C, Arnold K, Jackson IJ and Coffey PJ (2002) Presence of visual head tracking differentiates normal sighted from retinal degenerate mice. *Neuroscience Letters*, **325**, 21–24.

Tucci V, Lad H, Parker A, Polley S, Brown SDM, *et al* (2006) Gene/environment interactions differentially affect mouse strain behavioural parameters. *Mammalian Genome*, **11**, 1113–1120.

Introductory Review

Bacterial pathogens of man

Julian Parkhill
The Wellcome Trust Sanger Institute, Cambridge, UK

Bacterial pathogens of man can be found in a number of different phylogenetic groups of bacteria, although their distribution amongst all the known genera of prokaryotes is somewhat patchy, and there is still a great deal of debate as to whether there are any pathogenic archaea. Genomic study of these pathogens has begun to identify some common themes amongst these organisms, but none that are truly universal, and has also served to underline the diversity of mechanisms utilized for virulence and host interaction. Comparative genomic analysis has also begun to indicate the different mechanisms by which these organisms may have evolved their specific interactions with the human host. Bacterial pathogenicity is not a discrete state, and these interactions can range from commensalism, where the disease outcome is accidental, through occasional, opportunistic pathogenicity in a generalist organism, to specialist pathogens that are dependent on the host. However, it is becoming increasingly clear that organisms can move between these categories over the course of evolution.

Some families of bacteria contain large numbers of human and animal pathogens, with the whole group seemingly specialized for interactions with eukaryotic hosts, and with many members capable of interaction with multiple hosts. One such family is the Enterobacteria (*see* Article 51, **Genomics of enterobacteriaceae**, Volume 4), which is named after its usual niche in the guts of mammals. Within this group are commensals, such as the nonpathogenic *Escherichia coli* K12 (Blattner *et al*., 1997), broad host-range pathogens, such as *Salmonella enterica* serovar Typhimurium (*S. typhimurium*) (McClelland *et al*., 2001), and symbionts, such as *Buchnera* and *Wigglesworthia* (Akman *et al*., 2002; Shigenobu *et al*., 2000). This group also includes arguably the most virulent human bacterial pathogen, that which causes plague, *Yersina pestis* (Parkhill *et al*., 2001a), and its less virulent relatives (*see* Article 58, ***Yersinia***, Volume 4). Within this group, genomic comparisons have been particularly fruitful, allowing comparison of pathogens and nonpathogens (e.g., *E. coli* O157:H7 (Perna *et al*., 2001) and *E. coli* K12 (*see* Article 51, **Genomics of enterobacteriaceae**, Volume 4)), broad host-range pathogens and host-restricted pathogens (e.g., *S. typhimurium* and *S. typhi* (Parkhill *et al*., 2001b)), and low-virulence organisms against highly virulent pathogens (e.g., *Yersinia pseudotuberculosis* (Chain *et al*., 2004) and *Y. pestis*). The genomic sampling of this family is already fairly broad, and is likely to get much deeper, with projects under way that are set to sample certain subgroups many times

over (see GOLD; **http://www.genomesonline.org**). Many useful and interesting insights have already been gained from the comparative genomic study of these organisms, not least the concepts of core gene sets and pathogenicity islands. Most of the organisms share a common set of core genes, often organized in the same order and orientation around the genome. These core genes are responsible for housekeeping functions, such as transcription, translation, central metabolism, and so on, as well as some functions that may be important for survival within the intestinal niche, such as motility and chemotaxis. Interspersed with these conserved regions are blocks of genes that confer specialist functions, such as interactions with a specific host, or particular virulence phenotypes. These accessory genes can be present in small groups of one, two, or more genes (sometimes called islets), or they can be clustered in larger islands containing tens of genes. These large islands (called pathogenicity or genomic islands) often carry specific mechanisms allowing insertion and excision from the chromosome, and sometimes transfer between bacterial cells, and are therefore self-mobile. In addition to these accessory genes on the chromosome, the enterics often carry plasmids encoding antibiotic resistance and virulence genes, and these can also be self-mobile. Comparative genomic analysis of enteric bacteria has also revealed that some host adaptations are relatively recent in evolutionary terms, and that despite the widespread acquisition of accessory genes, host restriction and adaptation can also be due to the loss of genes through mutation (Parkhill *et al*., 2001a), a process first described in *Rickettsia*.

Another group of organisms with a large range of different hosts and diseases and in which horizontal exchange of accessory DNA is widespread is the Streptococci (*see* Article 57, **Genome-wide analysis of group A *Streptococcus***, Volume 4). Amongst these organisms, the most deeply sampled species to date is *Streptococcus pyogenes* (Banks *et al*., 2004). *Streptococcus pyogenes* can cause a remarkable range of diseases, ranging from scarlet fever through toxic shock syndrome and impetigo to acute rheumatic fever. Many of the virulence factors of these organisms are carried on chromosomally integrated bacteriophage (prophage), which are clearly capable of horizontal spread, and indeed many of the significant differences between strains of the species are in the number and type of the prophage carried. A second medically important Streptococcal species is *S. pneumoniae*, which is one of the causative agents of bacterial meningitis, but usually lives as a commensal in the throat. Several *S. pneumoniae* genomes have been, or are being, sequenced (Tettelin *et al*., 2001), including that of the strain in which DNA was first demonstrated to be the genetic material (Avery *et al*., 1944; Hoskins *et al*., 2001), and they indicate a far greater role for exchange of, and variability in, chromosomal genes in driving diversity in this species. Much of this is likely to be due to the fact that *S. pneumoniae* is a naturally competent organism, and is capable of taking up DNA from the environment and integrating it into its chromosome directly. Despite these, and other, studies, the link between genome content, virulence, and carriage in this species is far from clear, and several more genomic sequences are planned in order to attempt to shed more light on this. The genus *Streptococcus* also includes several important animal pathogens, and genomic projects for many of these are ongoing.

Like *S. pneumoniae*, several organisms often considered to be out-and-out pathogens are really commensals, and only cause disease accidentally, or when in a nonnatural host. A good example of this type of organism is *Neisseria meningitidis* (*see* Article 62, **The neisserial genomes: what they reveal about the diversity and behavior of these species**, Volume 4). While it is capable of causing two of the most fulminant and feared diseases, meningitis and septicaemia, its real niche is as a commensal of the human throat, a site to which it is extremely well adapted. Surprisingly, causing invasive disease is an evolutionary dead end for the specific organisms that escape into the blood stream – they cannot subsequently transmit to another host. For this reason, genomic investigation is aimed at those strains that are more likely to cause invasive disease, but the analysis must take into account the fact that adaptations identified in the genome are selected for commensal growth and transmission, not virulence.

A second type of accidental pathogen is that which is adapted to commensal existence in one host, but causes disease in another. A specific example of this is *Campylobacter jejuni*, which is a commensal of birds (its growth optimum is 42°C, the internal temperature of the chicken gut). When ingested by humans (in undercooked food), it causes gastroenteritis. *Campylobacter jejuni* is from the epsilon group of the proteobacteria, and its genome (Parkhill *et al*., 2000) revealed few of the well understood pathogenicity determinants that had mainly been discovered in the gamma proteobacteria (which includes the well-studied enterics), such as type-III and IV secretion systems, pili, and so on, and there was no evidence for classical pathogenicity islands (*see* Article 61, **Comparative genomics of the ε-proteobacterial human pathogens *Helicobacter pylori* and *Campylobacter jejuni***, Volume 4). Much of the basis of *Campylobacter* virulence is still unknown, although it is known that the flagellar system (which has been shown to secrete proteins in lieu on a type-III system (Konkel *et al*., 2004)) and surface polysaccharides are important. Perhaps surprisingly, one of *Campylobacter*'s closest sequenced relatives, *Helicobacter*, is an obligate resident of human beings (*see* Article 61, **Comparative genomics of the ε-proteobacterial human pathogens *Helicobacter pylori* and *Campylobacter jejuni***, Volume 4). *Helicobacter pylori* colonizes the stomach, and can cause ulcers and gastric cancer. The *H. pylori* genome is highly recombinogenic (Suerbaum *et al*., 1998), and encodes a well-defined toxin-encoding pathogenicity island. Curiously, *H. pylori* transmission seems to be primarily vertical, to the extent that population analysis of *H. pylori* can be used to define the structure and movements of human populations over millennia (Falush *et al*., 2003).

Another group of pathogens is the generalists, which are capable of survival and growth in many different environmental niches, including humans, other animals, and plants. These bacteria tend to have large and often complex genomes, and include species from the Burkholderiaceae and Pseudomonads. *Burkholderia pseudomallei*, for example, is generally a soil-dwelling saprophyte, but can infect humans (causing the disease meliodosis). The genome of *B. pseudomallei* is large (>7.2 Mb in two chromosomes), and contains evidence for large numbers of mobile islands, as seen in the Enterobacteriaceae, many of which are involved in metabolic diversity and environmental survival rather than simply pathogenicity (Holden *et al*., 2004). In this sense, *B. pseudomallei* may view the eukaryotic cell as

just another environment that can be exploited. Another opportunistic pathogen with similar abilities to survive in a wider environment and infect humans is *Pseudomonas aeruginosa* (Stover *et al*., 2000), which can cause lung infections in Cystic Fibrosis patients, and can also infect serious skin burns.

At the other end of the scale are pathogens that have specialized to the extent that they can no longer survive outside their chosen host. Classical examples of these are the Spirochetes (*see* Article 60, **Spirochete genomes**, Volume 4), which include the agents of syphilis (*Treponema pallidum*) and Lyme disease (*Borrelia burgdorferi*), and the Chlamydiales (*see* Article 59, **Chlamydiae**, Volume 4), which are obligate intracellular pathogens, and include the cause of trachoma (*Chlamydia trachomatis*). These organisms are so specialized that they are often very difficult to grow in the laboratory, and genomics is usually the most effective, and sometimes the only, way of getting genetic information about these organisms. These host-obligate organisms tend to have compact genomes, with a large degree of gene loss and evidence of metabolic streamlining. This process has been taken almost to completion in the Mycoplasmas (*see* Article 53, **The Mycoplasmas – a congruent path toward minimal life functions**, Volume 4). These organisms rely on their host for many nutrients and metabolites, and have reduced their metabolism to the extent that they can no longer make a cell wall. This is evident in their extremely small genomes (<600 kb in some cases), and it is this that has made them attractive tools for attempts to identify the smallest genome capable of supporting life (Hutchison *et al*., 1999).

Some bacterial genera contain organisms exhibiting many of these different characteristics, indicating the relative ease with which bacteria can move between categories during evolution. A good example of this is the Mycobacteria (*see* Article 52, **Genomics of the *Mycobacterium tuberculosis* complex and *Mycobacterium leprae***, Volume 4), which contains generalists with large genomes (e.g., *M. smegmatis*), specialist pathogens with smaller genomes and evidence for host-specialization with small genetic changes (e.g., *M. tuberculosis* and *M. bovis* (Cole *et al*., 1998; Garnier *et al*., 2003)), and host-obligate pathogens with highly reduced genomes (e.g., *M. leprae* (Cole *et al*., 2001)). These organisms are also extremely successful, causing some of the most common and feared infections in human history, including tuberculosis and leprosy.

In summary, it can be seen that bacterial pathogens are not a coherent group of organisms; they come from a wide variety of backgrounds and use a large diversity of strategies for host interaction. All of this makes the study of bacterial pathogen genomes an exciting and interesting pursuit, with the real potential for contributing to the control and elimination of human diseases (*see* Article 55, **Reverse vaccinology: a critical analysis**, Volume 4).

References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M and Aksoy S (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genetics*, **32**(3), 402–407.

Avery OT, MacLeod C and McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of the pneumococcal types. *The Journal of Experimental Medicine*, **79**, 137–158.

Banks DJ, Porcella SF, Barbian KD, Beres SB, Philips LE, Voyich JM, DeLeo FR, Martin JM, Somerville GA and Musser JM (2004) Progress toward characterization of the group A *Streptococcus* metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. *The Journal of Infectious Diseases*, **190**(4), 727–738.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**(5331), 1453–1474.

Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, *et al*. (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(38), 13826–13831.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**(6685), 537–544.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**(6823), 1007–1011.

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, *et al*. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**(5612), 1582–1585.

Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, *et al*. (2003) The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 7877–7882.

Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, *et al*. (2004) Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(39), 14240–14245.

Hoskins J, Alborn WE Jr, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, *et al*. (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. *Journal of Bacteriology*, **183**(19), 5709–5717.

Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO and Venter JC (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**(5447), 2165–2169.

Konkel ME, Klena JD, Rivera-Amill V, Monteville MR, Biswas D, Raphael B and Mickelson J (2004) Secretion of virulence proteins from *Campylobacter jejuni* is dependent on a functional flagellar export apparatus. *Journal of Bacteriology*, **186**(11), 3296–3303.

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, *et al*. (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, **413**(6858), 852–856.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, *et al*. (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**(6858), 848–852.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, *et al*. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**(6770), 665–668.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, *et al*. (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**(6855), 523–527.

Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, *et al*. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**(6819), 529–533.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y and Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**(6800), 81–86.

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, *et al*. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**(6799), 959–964.

Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E, Dyrek I and Achtman M (1998) Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(21), 12619–12624.

Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, *et al*. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**(5529), 498–506.

Introductory Review

Eukaryotic parasite genome projects

Neil Hall

The Institute for Genomic Research, Rockville, MD, USA

1. Introduction

Parasites cover a diverse range of organisms from bacteria such as *Mycobacterium leparae* (*see* Article 52, **Genomics of the *Mycobacterium tuberculosis* complex and *Mycobacterium leprae***, Volume 4) to unicellular protists such as the malaria parasite *Plasmodium falciparum* and multicellular eukaryotes such as nematode worms or arthropods; in this chapter, I will be concentrating primarily on work carried out on protist and nematode genomes. At the time of writing, there have been only two complete eukaryotic parasite genomes published (Gardner *et al*., 2002; Katinka *et al*., 2001). However, there are over 30 ongoing genome projects that are releasing data freely onto the Internet, and are already influencing how researchers are designing experiments and are contributing to the development of new treatments, diagnostic tests, and surveillance mechanisms. Parasite genomes can be large and complex and, unlike in the case of bacteria, few parasite genomes have yet been sequenced to completion, and researchers are often reliant on using partial genome data or Expressed Sequence Tag (EST) data (*see* Article 78, **What is an EST?**, Volume 4). In this chapter, I will attempt to describe how genome data has provided useful insights into parasite biology, evolution, and treatment, in each case highlighting specific examples.

2. Biological insights

In most eukaryotic organisms, each gene is associated with its own promoter region, which regulates the transcript level of the gene so that the protein encoded by it is only produced when required. When chromosome I of the *Leishmania major* genome was sequenced, it revealed one of the most unusual aspects of kinetoplastid biology; all of the genes are arranged in linear blocks, which are now believed to be transcribed as single polycictronic transcription units radiating from a single "strand switch region" (Myler *et al*., 1999). Hence, many genes appeared to share a single promoter. This model has many similarities to bacterial transcription in which genes are organized into functional operons that form a single transcript, although in *L.*

major the genes within the transcription units are not functionally clustered, so need not necessarily be coregulated. Now that the full genome of *L. major* has been sequenced, along with the genomes of the related parasites of *Trypanosome brucei* and *Trypanosome cruzi*, the causative agents of African sleeping sickness and Chagas' disease respectively (Degrave *et al*., 2001), researchers have observed that the same mechanism of transcription is present in all of these species and they are turning their attention to how the genes are regulated in systems that appear to lack conventional promoters. Because the genome data has been made freely available, scientists are able to use techniques that would otherwise be impossible, such as microarrays that allow them to measure expression levels of all of the genes in a genome (*see* Article 90, **Microarrays: an overview**, Volume 4 Section 8), and proteomics (see Volume 3) which measures the expression of all the proteins encoded by the genome. It now appears that the genes are regulated by post transcriptional mechanisms, whereby RNA transcript processing controls how much messenger RNA is translated into protein (reviewed in Horn, 2001).

3. Evolution

Many parasites occupy interesting positions in the phylogentic tree of life. Some protist parasites are believed by some to be examples of early branching eukaryotes, such as the kinetoplastids and the amitochondriates (*Entamoeba*, *Giardia*, *and Trichomonas*), whereas the nematodes are a highly diverse group of organisms that provide useful comparisons to the model free-living nematode *C. elegans* (*see* Article 44, **The *C. elegans* genome**, Volume 3). The intracellular parasite *Encephalitozoon cuniculii* is a microsporidial fungus, which has reduced its genome size as it has evolved as an obligate parasite, and numerous cellular functions have been taken on by the host. The genome is tiny for an eukaryote (2.9 Mb) encoding only 2000 genes (Katinka *et al*., 2001). It was believed that microsporidia did not have mitochondria and that they could represent a lineage of eukaryotes that branched before the acquisition of mitochondria. However, the genome project found many genes that appeared to have mitochondrial functions, and it is now thought that a cryptic mitochondrion, that has lost its genome and many of its functions, is present.

Plasmodium falciparum contains an organelle (intracellular structure) called the apicoplast that is found in almost all apicomplexan parasites. Sequencing of the apicoplast genome revealed that it is orthologous to the plant chloroplast (although clearly no longer photosynthetic), and it is believed to have evolved via a process of secondary endocytosis, whereby an algal cell became enslaved by another eukaryotic cell (McFadden, 2000). When the first data became available from the *P. falciparum* nuclear genome project, it was noticed that *Plasmodium* was littered with genes that appeared to be plant-like in origin. These genes were most likely transferred to the nuclear genome from the ancient algal cell. Many of these genes produce protein products that are directed back to the apicoplast. It is now estimated that nearly 10% of the genes in the *Plasmodium* genome encode products that end up in the apicoplast, demonstrating how important the organelle is for the parasite (Foth *et al*., 2003; Gardner *et al*., 2002). This finding has made the apicoplast an

attractive drug target as it carries out many metabolic functions that are not found in the vertebrate host.

4. Drug targets

A major objective of scientific research into parasitic diseases is to identify new drug targets or to understand better the mode of action of existing drugs. Up until recently, much of this study has relied on traditional biochemical methodology; however, since the advent of genomics, biochemists have been able to direct research based on a better understanding of the organism's metabolic potential.

Once a genome has been sequenced and annotated, all of the metabolic enzymes encoded within it can be identified. These genes are often well conserved and therefore simple to find and can be pieced together into complete metabolic pathways. After all of the pathways have been identified, one can focus on research of those enzymes that are potentially good targets for inhibition either because there are existing drugs that act on them or because they are absent from the host organism and hence drugs could better target the parasite. When the genome of the malaria parasite *P. falciparum* was sequenced, researchers identified genes encoding enzymes for the type II fatty acid biosynthesis pathway. This pathway in bacteria was the target for a drug called Triclosan, which has since been tested against *Plasmodium* and proven to be effective (Surolia and Surolia, 2001). Since then, whole genome analysis has highlighted seven new pathways that were, in some way, unique to the parasite compared to the human host, and hence may be good drug targets (Gardner *et al*., 2002).

5. Forward genetics

For parasites that go though a sexual cycle such as *T. brucei* and *P. falciparum*, genetic mapping can be used to pinpoint DNA sequences that contribute to heritable phenotypes: This is known as *positional cloning* or *forward genetics*. This technique becomes a powerful research tool once a genome sequence can be overlaid onto a genetic map as this positions genes relative to the map. In the case of *T. brucei*, a genetic map of the entire genome has been generated and is being used to map drug resistance and virulence traits (Tait *et al*., 2002). The genetic map of *P. falciparum* has been successfully employed to localize the gene conferring resistance to chloroquine (*see* Article 9, **Genome mapping overview**, Volume 3) (Su *et al*., 1999; Su and Wellems, 1996).

6. Antigen discovery

Early data from the *P. falciparum* genome project revealed that many of the major antigens were located at the telomeres of the chromosomes, most notably, the *VAR* genes that encode Pfemp1 proteins that are expressed on the surface of infected

red blood cells (Bowman *et al*., 1999; Gardner *et al*., 1998). Because of this, researchers concentrated efforts on studying the other telomeric genes, some of which have since been identified as being involved in an antigenic variation or host interaction such as the *RIFIN* and *STEVOR* genes. The complete genome sequence has allowed us to identify all of the genes in these families, which is an important step in identifying useful domains for vaccine design and understanding the complex processes involved in the host parasite interaction. Other *Plasmodium* species have been partially sequenced (*Plasmodium vivax*, *Plasmodium berghei*, *Plasmodium chabaudi*, and *Plasmodium yoelii*), and this has shown that these parasites have different telomere structures and different antigens encoded in them, known as *VIR* genes (del Portillo *et al*., 2001; Fischer *et al*., 2003; Janssen *et al*., 2001; Janssen *et al*., 2002) (also known as *BIR*, *CIR*, or *YIR* genes, depending on which species they are found in). These *VIR* genes are also thought to be involved in antigenic variation and expressed on the red blood cell surface (del Portillo *et al*., 2001).

It may be some time before genome data will be converted into marketable treatments, but sequencing projects have clearly heralded a new era in parasite research and are driving development into new drugs and vaccines, as well as contributing massively to our understanding of parasite molecular biology. In the next few years, it is likely that more than 30 parasite genomes will be sequenced so the full impact of this tidal wave of data is yet to be felt.

References

Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, *et al*. (1999) The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. *Nature*, **400**, 532–538.

Degrave WM, Melville S, Ivens A and Aslett M (2001) Parasite genome initiatives. *International Journal for Parasitology*, **31**, 532–536.

del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez CP, Schneider NK, Villalobos JM, Rajandream MA, Harris D, *et al*. (2001) A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. *Nature*, **410**, 839–842.

Fischer K, Chavchich M, Huestis R, Wilson DW, Kemp DJ and Saul A (2003) Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of Plasmodium chabaudi, a malaria species that undergoes antigenic variation in the laboratory mouse. *Molecular Microbiology*, **48**, 1209–1223.

Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF and McFadden GI (2003) Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science*, **299**, 705–708.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.

Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, *et al*. (1998) Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. *Science*, **282**, 1126–1132.

Horn D (2001) Nuclear gene transcription and chromatin in Trypanosoma brucei. *International Journal for Parasitology*, **31**, 1157–1165.

Janssen CS, Barrett MP, Lawson D, Quail MA, Harris D, Bowman S, Phillips RS and Turner CM (2001) Gene discovery in Plasmodium chabaudi by genome survey sequencing. *Molecular and Biochemical Parasitology*, **113**, 251–260.

Janssen CS, Barrett MP, Turner CM and Phillips RS (2002) A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **269**, 431–436.

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, *et al*. (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature*, **414**, 450–453.

McFadden GI (2000) Mergers and acquisitions: malaria and the great chloroplast heist. *Genome Biology*, **1**, REVIEWS1026.

Myler PJ, Audleman L, de Vos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, *et al*. (1999) Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2902–2906.

Su X, Ferdig MT, Huang Y, Huynh CQ, Liu A, You J, Wootton JC and Wellems TE (1999) A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. *Science*, **286**, 1351–1353.

Su X and Wellems TE (1996) Toward a high-resolution Plasmodium falciparum linkage map: polymorphic markers from hundreds of simple sequence repeats. *Genomics*, **33**, 430–444.

Surolia N and Surolia A (2001) Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum. *Nature Medicine*, **7**, 167–173.

Tait A, Masiga D, Ouma J, MacLeod A, Sasse J, Melville S, Lindegard G, McIntosh A and Turner M (2002) Genetic analysis of phenotype in Trypanosoma brucei: a classical approach to potentially complex traits. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **357**, 89–99.

Specialist Review

Genomics of enterobacteriaceae

Jeremy D. Glasner and Nicole T. Perna
University of Wisconsin, Madison, WI, USA

1. Introduction

Enterobacteria were among the earliest targets for genome sequencing and are still the most densely sampled clade of bacteria in the genomics arena. Twenty complete genome sequences are available for members of the family as of September 2004, and the NCBI list of genomes in progress includes another 20 enterobacteria. This is undoubtedly an underestimate of the sequences that will be available for enterobacteria in the coming two years. Among the complete genomes are representatives of nine genera, including *Escherichia*, *Shigella*, *Salmonella*, *Yersinia*, *Erwinia*, *Buchnera*, *Photorhabdus*, *Wigglesworthia,* and *Candidatus Blochmannia*. Genomes in progress will add additional strains and species of these, as well as additional genera such as *Klebsiella*, *Proteus*, *Citrobacter*, *Enterobacter*, *Dickeya*, and *Pantoea*. Among these are standard laboratory research strains, human pathogens, livestock pathogens, plant pathogens, and insect endosymbionts. Each genome sequence aids in understanding the biology of the individual organism, but some of the greatest insights come from comparisons between the genomes. For many of these genera, sequences are available for multiple species or strains, providing unique perspectives on genome-wide polymorphism. Enterobacteria are also very experimentally tractable, and we are beginning to see a boom in downstream research making use of the sequences.

2. A brief history

The first enterobacterium chosen for complete genome sequencing was *Escherichia coli* K-12 strain MG1655 (Blattner *et al*., 1997; Perna *et al*., 2002), motivated by the widespread use of this model prokaryote for molecular biological research and industrial applications. To this day, *E. coli* K-12 remains one of the best-studied organisms on earth, with direct experimental evidence for the functions of a large number of genes. Data from high-throughput postgenomic experimentation with other strains or species may quickly eclipse the volume of data currently available for *E. coli* K-12. However, data from proteomic and functional genomic techniques tend to generate hypotheses that must be tested by more detailed experimental characterization. For this reason, *E. coli* K-12 with its long history of experimental

investigation continues to play a special role in genomics of all enterobacteria, and perhaps all prokaryotes, as the ultimate source of most experimental evidence used to attribute gene functions in genome annotation.

Following *E. coli* K-12, two groups of organisms were next in line for genome sequencing and remain active areas of inquiry. These are pathogens of humans and other animals (reviewed in Whittam and Bumbaugh, 2002), which were chosen for their biomedical relevance and distinctive disease potentials. Continued sequencing efforts in this area are justified by the astounding diversity seen in initial comparisons of these genomes and the number of different species that colonize the guts and are often associated with disease of humans and livestock. The other group is insect symbionts that were chosen as a model for studying microbe–insect interactions and evolution by genome reduction (Tamas *et al*., 2002; Shigenobu *et al*., 2000; Moran, 2003). Genome-sequencing efforts for plant-pathogenic enterobacteria lagged behind, but this deficit is being addressed with one recently published (Bell *et al*., 2004) and at least three ongoing projects.

3. Evolution of genome content and size in enterobacteria

Sequenced genomes of enterobacteria vary over an order of magnitude in length (0.62–5.59 million base pairs (Mbp)) and correspondingly, in number of predicted genes (504–5476), as shown in Table 1. Typical genome architecture is a single bidirectionally replicating circular chromosome and one or more circular plasmids. One striking observation is the extreme difference in genome sizes observed between free-living organisms and the insect endosymbionts, *Buchnera*, *Wigglesworthia*, and *Blochmannia*. The apparent simplicity of the idea of genome reduction as a consequence of an obligate relationship with a host belies the complexity of the evolutionary dynamics behind even these streamlined genomes. There are several insightful reviews on the subject (Moran, 2003; Klasson and Andersson, 2004; Wernegreen, 2002). Comparative analysis of the five complete genomes from insect endosymbiotic bacteria revealed only 277 protein-coding genes in common (Gil *et al*., 2003).

All genes in the symbionts are also found in the free-living enterobacteria. These clearly represent a subset of the genes present in the most recent common ancestor (MRCA) of all enterobacteria since many genes not found in the symbionts have orthologs in more distantly related bacteria. Moran *et al*. estimate that there were 2425 genes in the MRCA genome (Moran and Mira, 2001). Robust phylogenetic evidence supports the idea of a common lineage leading to all the sequenced insect endosymbiotic enterobacteria (Lerat *et al*., 2003). While deletions of the ancestral chromosome clearly shaped this lineage, other enterobacteria show extensive evidence of both loss and gain of genes. Acquisition of clusters of genes from other strains and species by horizontal transfer appears to be relatively common in free-living enterobacteria. Not so long ago, this observation disrupted a paradigm of clonal evolution for bacteria in general, based on considerable work with enterobacteria in particular (Smith *et al*., 1993). The significant role recombination plays in enterobacterial genome evolution is now quite clear.

Table 1 Published complete genome sequences from enterobacteria

General information	Species and strain	Genome length (bp)	Number of plasmids	Number of genes[a]	References
Common K-12 laboratory strain	*Escherichia coli* strain MG1655	4 639 675	0	4269	Blattner *et al*. (1997)
Enterohemorrhagic pathogen associated with outbreak in Japan	*Escherichia coli* serotype O157:H7 strain Sakai	5 498 450	2	5470	Hayashi *et al*. (2001)
Example of virulent serogroup 2a associated with dysentery	*Shigella flexneri* serotype 2a strain 301	4 607 203	1	4186	Jin *et al*. (2002)
Example of virulent serogroup 2a associated with dysentery	*Shigella flexneri* serotype 2a strain 2457 T	4 599 354	0	4073	Wei *et al*. (2003)
Enterohemorrhagic pathogen from meat linked with outbreak	*Escherichia coli* serotype O157:H7 strain EDL933	5 528 445	2	5346	Perna *et al*. (2001)
Urinary tract pathogen associated with pyelonephritis	*Escherichia coli* strain CFT073	5 231 428	0	5379	Welch *et al*. (2002)
Standard laboratory model system for Salmonella research	*Salmonella enterica* serovar Typhimurium strain LT2	4 857 432	1	4451	McClelland *et al*. (2001)
Multiple drug resistant pathogen associated with typhoid fever	*Salmonella enterica* serovar Typhi strain CT18	4 809 037	2	4600	Parkhill *et al*. (2001a)
Model strain for work with typhoid fever associated Salmonellae	*Salmonella enterica* Serovar Typhi strain Ty2	4 791 961	0	4323	Deng *et al*. (2003)
Human pathogen associated with bubonic and pneumonic plague	*Yersinia pestis* strain KIM	4 600 755	2	4280	Deng *et al*. (2002)
Recently isolated human pathogen associated with plague	*Yersinia pestis* strain CO92	4 653 728	3	4008	Parkhill *et al*. (2001b)
Human avirulent strain isolated from a rodent	*Yersinia pestis* strain 91001	4 595 065	4	3895	Song *et al*. (2004)
Pathogen associated with chronic but mild intestinal disease	*Yersinia pseudotuberculosis* strain IP32953	4 744 671	2	3974	Chain *et al*. (2004)

(*continued overleaf*)

Table 1 (*continued*)

General information	Species and strain	Genome length (bp)	Number of plasmids	Number of genes[a]	References
Plant pathogen associated with soft-rot and blackleg of potatoes	*Erwinia carotovora atroseptica* strain SCRI1043	5 064 019	0	4491	Bell *et al*. (2004)
Symbiont of nematodes and insect pathogen	*Photorhabdus luminescens* strain TT01	5 688 987	0	4905	Duchaud *et al*. (2003)
Endocellular symbiont of the aphid species *Baizongia pistacea*	*Buchnera aphidicola* (*Baizongia pistacea*)	615 980	1	504	Tamas *et al*. (2002)
Endocellular symbiont of the pea aphid species *Acyrthosiphon pisum*	*Buchnera* sp. APS	640 681	1	564	Shigenobu *et al*. (2000)
Endocellular symbiont of the aphid species *Schizaphis graminum*	*Buchnera aphidicola* (*Schizaphis graminum*)	641 454	1	545	Tamas *et al*. (2002)
Endocellular symbiont of carpenter ants	*Blochmannia floridanus*	705 557	0	589	Gil *et al*. (2003)
Endocellular symbiont of tsetse flies	*Wigglesworthia brevipalpis*	697 724	0	611	Akman *et al*. (2002)

[a]Number of protein-coding genes on the main chromosome reported in the ASAP database as of September 15, 2004 (Glasner *et al*., 2003).

Multilocus enzyme electrophoresis profiles from numerous comparisons of enterobacterial strains and species showed strong linkage disequilibrium, supporting a common evolutionary history for genes distributed throughout the chromosome, hence, clonal evolution. At odds with this model were analyses of early segments of the *E. coli* K-12 genome sequence, which revealed a significant number of genes with atypical codon usage and other nucleotide biases suggesting origins from other bacteria with fundamentally distinct compositional signatures (Medigue *et al*., 1991). Also, before completion of the first genome, researchers coined the term "pathogenicity island" to describe clusters of genes associated with virulence found specifically in pathogenic strains (Morschhauser *et al*., 1994). These islands also showed unusual nucleotide composition, and some could be mobilized and transferred among strains or species. Comparisons of complete genome sequences provided a unified framework for all these observations and expanded the concept of islands.

Alignment of even relatively closely related pairs of genomes, like two strains of *E. coli,* reveals hundreds of lineage-specific islands interspersed among conserved segments of the chromosome. More than a quarter of an *E. coli* genome can lie within these islands (Perna *et al*., 2001; Hayashi *et al*., 2001; Welch *et al*., 2002). They encode a wide variety of activities associated with diverse processes including but not limited to pathogenesis, and some interesting examples of islands in particular genomes are highlighted below.

Despite the apparently high levels of horizontal transfer introducing new DNA into the chromosomes of enterobacteria, replacement of existing alleles with divergent copies from other lineages is uncommon. Genome-scale phylogenetic analyses of the genes from the core chromosome repeatedly support the same branching order, or topology, for subsets of enterobacteria, consistent with the multienzyme electrophoresis data (Daubin *et al*., 2003; Lerat *et al*., 2003). The "Core Genome Hypothesis" has been described as an application of the biological species concept to bacteria (Wertz *et al*., 2003).

4. *E. coli* and *Shigella*

E. coli and *Shigella* will be treated as a group in light of overwhelming phylogenetic and genomic support for the idea that they are the same species (reviewed in Lan and Reeves, 2002). Sequences are complete and published for laboratory *E. coli* K-12 strain MG1655, two enterohemorrhagic *E. coli* O157:H7 strains (EDL933 and O157 Sakai), a urinary tract-associated *E. coli* strain (CFT073), and two *S. flexneri* strains (301 and 2457 T).

O157:H7 genome sequencing and comparison to K-12 was instrumental in revealing the extent of horizontal gene transfer in this species. Genomes of O157:H7 are also of considerable interest as important human pathogens, and many genes potentially relevant to disease were discovered or better characterized by genome sequencing. O157:H7 strains cause a bloody-diarrhea (hemorrhagic colitis), and occasionally infections lead to fatal hemolytic uremic syndrome (HUS). Two independent O157:H7 genomes have been sequenced and show very few differences in gene content and chromosome organization. Examination of the O157:H7 genomes revealed about 1.5 Mb of sequences that are not conserved with *E. coli* K-12 clustered into over 150 islands. Genes encoded within these islands include a well-known type III secretion system (LEE), as well as a second, previously uncharacterized cryptic type III secretion system now known to be widely distributed, albeit with substantial divergence, among *E. coli* strains (Ren *et al*., 2004). The islands also included a collection of fimbrial and nonfimbrial adherence factors, putative toxins, and iron utilization systems, many of which have subsequently been characterized with clear roles in pathogenesis. The sequence of the 92-Kb virulence plasmid revealed a large putative toxin, with a newly demonstrated role in adherence to epithelial cells and type III secretion (Stevens *et al*., 2004). The O157:H7 genomes are also rich in prophages, with at least 18 regions from each genome showing similarity to known phages. These include the functional prophages (933 W in EDL933 and VT2-Sakai from O157Sakai) that encode the characteristic shiga-like toxins. Interestingly, many strains, including the

two genomes sequenced, include a second set of shiga-like toxin genes encoded in the analogous position of a cryptic prophage, and the phage regions flanking the toxins are conserved with many other lambda-family phages in these and other genomes (Ohnishi *et al*., 2001). The position of these genes, in a late-phage transcript, provides insight into the regulation and dissemination of the toxin genes in the *E. coli* population. This may prove important in understanding the emergence of other distinct lineages of shiga-like toxin producing *E. coli* (STEC). Genome sequences for other STEC strains could reveal other genes associated with convergent evolution of hemorrhagic phenotypes emerging from distinct lineages of *E. coli* that arose by independent horizontal transfers of similar islands. No sequencing projects for other STEC are currently underway, although at least one representative STEC genome may be included in a newly funded effort.

Shigella are also important human pathogens associated with dysentery, particularly in developing countries. The two sequenced *S. flexneri* genomes are relatively small (4.59 and 4.60 Mb) and show more examples of chromosome rearrangement than other sequenced *E. coli* genomes (Jin *et al*., 2002; Wei *et al*., 2003). The gene content of strains 301 and 2457 T are very similar, but several major rearrangements are detectable (Wei *et al*., 2003; Darling *et al*., 2004). At least 15 genome rearrangements differentiate strain 2457 T from *E. coli* K-12, compared to the single inversion seen between O157:H7 EDL933 and K-12. Alignment of *Shigella* and K-12 genomes delineates 37 islands larger than 1 kb in length. These islands along with the *Shigella*-associated plasmids encode many of the known and putative virulence factors. These include a type III secretion system associated with invasion, iron uptake systems, and adherence factors. This list sounds qualitatively similar to the genes that distinguished O157:H7 strains from K-12. However, comparisons of genes from the *Shigella* islands with homologs in O157:H7 reveal differences in chromosomal locations of the genes and high levels of sequence divergence. This suggests that these islands likely arose as independent horizontal transfer events.

The *S. flexneri* genomes have a large number of IS elements relative to other *E. coli* strains, and many of the genome rearrangements are bounded by copies of the same element. The IS abundance is also obvious on the virulence plasmids. The plasticity observed in *Shigella* genome sequence comparisons illuminates a possible underlying basis for high levels of PFGE polymorphism among epidemiologically relevant *Shigella* isolates. A large number of pseudogenes (372) relative to other *E. coli* are also related to IS activity and, interestingly, account for many of phenotypic characters used as *Shigella* diagnostics (Wei *et al*., 2003). Although serotype 2a is particularly virulent, these two genomes barely begin to represent the diverse *Shigella* relevant to human health. Sequencing of genomes from additional *Shigella* species (*S. dysenteriae* and *S. boydii*) is underway.

Extraintestinal *E. coli*, like pyelonephritis-associated strain CFT073, are also of interest. *E. coli* are responsible for 70–90% of urinary tract infections. Strain CFT073 is also a member of one of the earliest diverging branches of *E. coli,* and this phylogenetic position makes it particularly useful for interpreting the evolutionary history of the species. Co-occurrence of islands in the CFT073 genome at the same location as distinct islands in O157:H7 and/or K-12 suggests reuse of chromosomal sites for successive horizontal transfers. This could be due to

elevated insertion at these sites, prohibitive natural selection elsewhere, permissive selection at these sites, neutrality of replacement of previously acquired elements, or some combination of these factors. The genome data can be used to both form and address a series of testable hypotheses about the evolutionary dynamics of horizontal transfer. In terms of content, the CFT073 genome is notable for the absence of a type III secretion system, now a veritable diagnostic marker for gram-negative pathogens. It does, however, encode fimbrial adherence factors, including the well-characterized *pap* antigens, iron utilization systems, and a total of 1.2 Mb not seen in other sequenced *E. coli* genomes. Observed insertions, deletions, and rearrangements relative to islands characterized from other UTI associated *E. coli* may reflect the age of this lineage. Even horizontal transfers since the common ancestor of all *E. coli* could be 10 million years old (Reid *et al*., 2000). Other genome projects are underway for extraintestinal *E. coli* from this deep-branching lineage, including strain RS218, a K1-antigen *E. coli* associated with neonatal sepsis, and meningitis. Although this lineage of *E. coli* also includes diarrheal pathogens, no sequencing projects are yet underway for these strains.

5. Salmonella

Three complete *Salmonella enterica* genome sequences are available, one common laboratory strain of *S. enterica* Typhimurium (LT2) and two strains of *S. enterica* Typhi (CT18 and Ty2). *Salmonellae* are a common food-borne threat to public health and the livestock industry, and Typhoid fever caused by Typhi afflicts 16 million people a year. The *Salmonella* genomes tell a similar story to those of *E. coli*, with extensive differential horizontal transfer driving divergence between the Typhi and Typhimurium chromosomes. Among the more significant observations arising from the comparison of these genomes is the large number of pseudogenes in Typhi, distinguished by a stop codon that disrupts a continuous reading frame in Typhimurium and *E. coli* strains. Every sequenced genome of an enterobacterium contains some pseudogenes, but the elevated rate of pseudogene formation in Typhi has been suggested as a reflection of the reduction to a human-specific pathogen from an ancestor with a broader host range. There are 3000 or so orthologs shared between any pair of *Salmonella* and *E. coli* genomes. The genomes are largely colinear, except for a series of inversions about the origin and terminus of replication. Apart from those shared with *E. coli*, Typhi and Typhimurium have another 500 genes in common including the type III secretion systems, a number of effector proteins, and metal-ion transporters. Among the genes present in Typhi but not Typhimurium are chaperone–usher fimbrial systems, a hemolysin, homologs of the *Campylobacter* toxin *cdtB* and the *Bordetella pertussis* toxin genes *ptxA* and *ptxB*, a putative polysaccharide acetyltransferase, and many phage genes. Conversely, in Typhimurium but not Typhi are the pSLT plasmid genes and other fimbrial genes. Partial sequences are available for a number of other *Salmonella*, like *S.e*. Dublin, *S.e*. Enteritidis, and *S.e*. Paratyphi. Nonredundant microarrays representing the total set of genes identified in *Salmonella* genomes are providing a powerful tool for genotyping the many and diverse *Salmonella* serovars (Boyd *et al*., 2003).

6. Yersinia

Analyses of four complete genome sequences are available for representatives of the Genus *Yersinia*. Three are *Y. pestis*, the causative agent of bubonic and pneumonic plague, with selected strains from two distinct plague biovars: Medievalis (strains YPKIM) and Orientalis (CO92) (Parkhill *et al*., 2001b; Deng *et al*., 2002) and a recently complete sequence for strain 9001, an isolate that is virulent in mice but not in humans. *Y. pestis* is a relatively young species, having diverged from *Y. pseudotuberculosis* within the last 20 000 years (recently reviewed in Brubaker, 2004). Analyses of the complete sequence of *Y. pseudotuberculosis* strain IP32953 and PCR-based, microarray-based, and sequence-based comparisons with *Y. pestis* genomes have just been published, revealing that very few events were required to evolve plague, a highly virulent pathogen, from a less virulent ancestor (Chain *et al*., 2004). In total, only 32 genes, clustered in 6 islands, are found in all *Y. pestis* chromosomes surveyed, but not in *Y. pseudotuberculosis*. There are also two *Y. pestis*-specific plasmids (a third, pCD1 is shared by both species). *Y. pestis* genomes also tend to show a relatively large number of pseudogenes relative to *Y. pseudotuberculosis*. The *Y. pestis* 91001 genome is quite different from the other two *Y. pestis* and includes an additional plasmid containing genes for a type IV secretion system as well as many chromosomal changes, with 141 likely pseudogenes.

Given the close relationship of all *Y. pestis*, it is not surprising that the CO92 and KIM genomes are alignable across 95% of their length (Deng *et al*., 2002). Even with this limited variability among the complete *Y. pestis* genomes, the sequences are proving useful. Microarray-based genotyping has been used to characterize live plague vaccine strains and natural isolates (Hinchliffe *et al*., 2003; Zhou *et al*., 2004a; Zhou *et al*., 2004b). Together, these comparisons reveal roughly 25 islands polymorphic in the *Y. pestis* population. Although *Y. pestis* show a much lower degree of gene acquisition by horizontal transfer than *E. coli* strains, *Y. pestis* genomes show much higher levels of rearrangement. Alignment of the CO92 and KIM genomes divides the chromosomes into at least 27 blocks of sequences that are colinear. Most disruptions in colinearity result from a series of inversions around the origin and terminus of replication, typically bounded by repetitive elements (Deng *et al*., 2002). A large number of rearrangements also differentiate the 91001 genome from the other two pestis chromosomes. This propensity for rearrangement is not shared with *Y. pseudotuberculosis* (Chain *et al*., 2004).

Just over half of the genes in each *Y. pestis* genome have putative orthologs in *E. coli*, *Salmonella*, or *Erwinia*. These, almost certainly, are components of the core ancestral chromosome, including most basic metabolic genes, global regulators, and many transport proteins. Although colinearity with these more distantly related taxa appears to be limited to segments of one or a few operons, dotplots of these putative orthologs show a characteristic “X” shape (Eisen *et al*., 2000), reflecting a series of inversions about either the origin or terminus of replication relative to the ancestral state. Putative orthologs observed off the main diagonals of the plot may reflect small translocations, or differential loss or gain of paralogous sequences that are currently mistaken as orthologous. Resolution will require additional phylogenetic analysis, and, possibly, additional genome-scale data.

Genome sequencing is complete for *Y. enterocolitica,* a close relative of *Y. pestis* that causes gastroenteritis. Complete annotation and publication of the sequence is anticipated later this year.

7. Plant-pathogenic enterobacteria

Recently, the publication of a genome for the pytopathogen *Erwinia carotovora* ssp. atroseptica allowed for the first comparison of plant and animal pathogenic enterobacteria. The reported number of genes shared between *E. c. atroseptica* and any *E. coli*, *Yersinia,* and *Salmonella* genome is about 2300. Among those genes not found in the animal-associated organisms are a battery of plant cell wall–degrading enzymes and an impressive array of secretion systems: type II, type III, type IV, TAT, and fimbrial. The contents of the likely horizontally acquired segments of these genomes show similarities with the other enterobacteria. They include secretion systems and numerous iron-uptake systems. However, unlike the animal-associated enterobacteria, there are not large numbers of fimbrial operons, suggesting that these extracellular structures may be less important for the plant-pathogenic lifestyle. A genome project is underway for another soft-rot associated plant pathogen, *Erwinia chrysanthemi* strain 3937. Although the common usage of the genus *Erwinia* for these two organisms suggests a close relationship, a series of recent taxonomic revisions suggest that they are quite different, and the genomes are expected to show substantial divergence. An entirely separate lineage of plant pathogens, many also known as *Erwinia*, are associated with diseases quite distinct from soft-rot. Genome projects for two of these are underway, *Erwinia amylovora*, the causative agent of “fire blight” of apples and pears, and *Pantoea stewartii*, a pathogen of sweet corn and maize. The *P. stewartii* genome includes 10–13 stable plasmids totaling a quarter of the genome.

8. Insect-associated enterobacteria

Photorhabdus luminescens is an enterobacterium found as a symbiont of soil entomopathogenic nematodes and a pathogen of diverse insects. The *P. luminescens* life cycle includes a symbiotic stage in the nematode gut and a virulent stage in the insect larvae that is killed through toxemia and septicemia. The *P. luminescens* genome (Duchaud *et al*., 2003) encodes a large number of adhesins, hemolysins, lipases, proteases, and antibiotic biosynthesis genes. These proteins are likely to play a role in the elimination of competitors, colonization of the host, and invasion and utilization of the insect cadaver. The genome encodes more predicted toxin-encoding genes than any other complete genome. The diversity of toxins may be related to pathogenicity on a wide variety of insect species. A number of secreted proteins were identified that likely play a role in bioconversion of the insect cadaver. Comparison of the *P. luminescens* genome to *Y. pestis* revealed 2107 orthologous genes, including some that are not shared with *E. coli*. The genome contains a large number of IS elements (about 195) and an astounding 711 inverted repeats of

the enterobacterial repetitive intergenic consensus (ERIC) sequence that has only 21 copies in *E. coli* K-12.

Genome sequences have been completed for five bacterial endosymbionts of insects, three from the genus *Buchnera* (aphid host), *Blochmannia floridanus* (carpenter ant host), and *Wigglesworthia glossinidia* (tsetse fly host). These symbionts reside in specialized host cells called bacteriocytes, and the bacteria are transmitted maternally during insect development. The symbioses all seem to involve insect hosts that feed on nutrient sources lacking in key metabolites; amino acids in the case of phloem-feeding aphids and carpenter ants, which are believed to subsist primarily on the honeydew of phloem-feeding insects, and vitamins in the case of tsetse flies that feed on vertebrate blood. The bacteria are believed to supplement their hosts with these key metabolites in exchange for metabolic products that the bacteria require. An obvious consequence of these long-term associations, and the stabile environments provided by hosts, is an extreme reduction in genome size in all of these lineages. Genomes from all three genera have lost many typical enterobacterial genes such as those involved in transport and transcriptional regulation activities. Each of the three endosymbiont lineages has retained a somewhat different subset of genes from the ancestral genome that likely reflect the differing nutritional needs of their host species. *Wigglesworthia,* for example, contains many genes involved with cofactor biosynthesis that are absent in *Buchnera* that contains more genes involved in amino acid biosynthesis. A less obvious consequence of this endosymbiont lifestyle is an increase in the rate of nucleotide substitution and reduction in the rate of genome rearrangement compared with other free-living enterobacteria. This apparent paradox can be explained by the types of sequences missing from these genomes that include most of the repetitive elements, such as IS, multiple rRNA operons, and recombination-associated genes that are associated with genome rearrangements, and genes encoding repair and stress-response functions that are associated with maintaining sequence integrity. Interestingly, some of the genes maintained in *Blochmannia* and *Wigglesworthia* encode components of the flagellar machinery and cell wall constituents that suggest more complex interactions with their hosts than currently known.

9. Beyond genome sequencing

As new sequences accumulate, existing genomes of enterobacteria are being put to work in a multitude of ways. Applications of the sequences include hypothesis-driven research like experimental verification of roles of putative new virulence factors identified in the genomes. The distribution of individual islands identified through comparative genomics in populations of pathogens has led to correlations of epidemiologically relevant characteristics with genomic markers. Other types of polymorphisms identified by analyzing and comparing genomes provide additional means for subtyping isolates of pathogens. High-throughput multilocus PCR assays and microarray technologies are beginning to provide genome-scale subtyping alternatives. *E. coli* K-12, with its large complement of experimental data, is a choice target for modeling cell growth, metabolism, and regulation, giving rise

to novel systems-based approaches to understanding the relationship between genotypes and phenotypes.

References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M and Aksoy S (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nature Genetics*, **32**(3), 402–407.

Bell KS, Sebaihia M, Pritchard L, Holden MT, Hyman LJ, Holeva MC, Thomson NR, Bentley SD, Churcher LJ, Mungall K, *et al*. (2004) Genome sequence of the enterobacterial phytopathogen Erwinia carotovora subsp. atroseptica and characterization of virulence factors. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(30), 11105–11110.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al*. (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**(5331), 1453–1474.

Boyd EF, Porwollik S, Blackmer F and McClelland M (2003) Differences in gene content among Salmonella enterica serovar typhi isolates. *Journal of Clinical Microbiology*, **41**(8), 3823–3828.

Brubaker RR (2004) The recent emergence of plague: a process of felonious evolution. *Microbiology Ecol* **47**(3), 293–299.

Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, *et al*. (2004) Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(38), 13826–13831.

Darling AC, Mau B, Blattner FR and Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, **14**(7), 1394–1403.

Daubin V, Moran NA and Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**(5634), 829–832.

Deng W, Burland V, Plunkett G III, Boutin, A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, *et al*. (2002) Genome sequence of Yersinia pestis KIM. *Journal of Bacteriology*, **184**(16), 4601–4611.

Deng W, Liou SR, Plunkett G III, Mayhew GF, Rose DJ, Burland V, Kodoyianni V, Schwartz DC and Blattner FR (2003) Comparative genomics of salmonella enterica serovar Typhi strains Ty2 and CT18. *Journal of Bacteriology*, **185**, 2330–2337.

Duchaud E, Rusniok C, Frangeul L, Buchrieser C, Givaudan A, Taourit S, Bocs S, Boursaux-Eude C, Chandler M, Charles JF, *et al*. (2003) The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nature Biotechnology*, **21**(11), 1307–1313.

Eisen JA, Heidelberg JF, White O and Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology*, **1**(6), RESEARCH0011.

Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Holldobler B, *et al*. (2003) The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(16), 9388–9393.

Glasner JD, Liss P, Plunkett G III, Darling A, Prasad T, Rusch M, Byrnes A, Gilson M, Biehl B, Blattner FR, *et al*. (2003) ASAP, a systematic annotation package for community analysis of genomics. *Nucleic Acids Research*, **31**, 147–151.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, *et al*. (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research*, **8**(1), 11–22.

Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, Nichols RA, Oyston PC, Hinds J, Titball RW and Wren BW, (2003) Application of DNA microarrays to study the

evolutionary genomics of Yersinia pestis and Yersinia pseudotuberculosis. *Genome Research* **13**(9), 2018–2029.

Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, *et al*. (2002) Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Research*, **30**(20), 4432–4441.

Klasson L and Andersson SG (2004) Evolution of minimal-gene-sets in host-dependent bacteria. *Trends in Microbiology*, **12**(1), 37–43.

Lan R and Reeves PR (2002) Escherichia coli in disguise: molecular origins of Shigella. *Microbes and Infection*, **4**(11), 1125–1132.

Lerat E, Daubin V and Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biology*, **1**(1), E19.

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, *et al*. (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, **413**(6858), 852–856.

Medigue C, Rouxel T, Vigier P, Henaut A and Danchin A (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. *Journal of Molecular Biology*, **222**(4), 851–856.

Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology*, **6**(5), 512–518.

Moran NA and Mira A (2001) The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biology*, **2**(12), RESEARCH0054.

Morschhauser J, Vetter V, Emody L and Hacker J (1994) Adhesin regulatory genes within large, unstable DNA regions of pathogenic Escherichia coli: cross-talk between different adhesin gene clusters. *Molecular Microbiology*, **11**(3), 555–566.

Ohnishi M, Kurokawa K and Hayashi T (2001) Diversification of Escherichia coli genomes: are bacteriophages the major contributors? *Trends in Microbiology*, **9**(10), 481–485.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, *et al*. (2001a) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature*, **413**(6858), 848–852.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, *et al*. (2001b) Genome sequence of Yersinia pestis, the causative agent of plague. *Nature*, **413**(6855), 523–527.

Perna NT, Glasner JD, Burland V and Plunkett G III (2002) *Escherichia Coli: Virulence Mechanism of a Versatile Pathogen*, Donnenberg MS (Ed.), Academic Press, Elsevier Science: San Diego, pp. 3–53.

Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, *et al*. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, **409**(6819), 529–533.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK and Whittam TS (2000) Parallel evolution of virulence in pathogenic Escherichia coli. *Nature*, **406**(6791), 64–67.

Ren CP, Chaudhuri RR, Fivian A, Bailey CM, Antonio M, Barnes WM and Pallen MJ (2004) The ETT2 gene cluster, encoding a second type III secretion system from Escherichia coli, is present in the majority of strains but has undergone widespread mutational attrition. *Journal of Bacteriology*, **186**(11), 3547–3560.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y and Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature*, **407**(6800), 81–86.

Smith JM, Smith NH, O'Rourke M and Spratt BG (1993) How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, **90**(10), 4384–4388.

Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H, *et al*. (2004) Complete genome sequence of Yersinia pestis strain 91001, an isolate avirulent to humans. *DNA Research*, **11**(3), 179–197.

Stevens MP, Roe AJ, Vlisidou I, van Diemen PM, La Ragione RM, Best A, Woodward MJ, Gally DL and Wallis TS (2004) Mutation of toxB and a truncated version of the efa-1 gene in Escherichia coli O157:H7 influences the expression and secretion of locus of enterocyte effacement-encoded proteins but not intestinal colonization in calves or sheep. *Infection and Immunity*, **72**(9), 5402–5411.

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA and Andersson SG (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, **296**(5577), 2376–2379.

Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, Mayhew GF, Plunkett G III, Rose DJ and Darling A (2003) Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. *Infection and Immunity*, **71**(5), 2775–2786.

Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, *et al*. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(26), 17020–17024.

Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, **3**(11), 850–861.

Wertz JE, Goldstone C, Gordon DM and Riley MA (2003) A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. *Journal of Evolutionary Biology*, **16**(6), 1236–1248.

Whittam TS and Bumbaugh AC (2002) Inferences from whole-genome sequences of bacterial pathogens. *Current Opinion in genetics & development*, **12**(6), 719–725.

Zhou D, Han Y, Dai E, Song Y, Pei D, Zhai J, Du Z, Wang J, Guo Z and Yang R (2004a) Defining the genome content of live plague vaccines by use of whole-genome DNA microarray. *Vaccine* **22**(25–26), 3367–3374.

Zhou D, Han Y, Song Y, Tong Z, Wang J, Guo Z, Pei D, Pang X, Zhai J, Li M, *et al*. (2004b) DNA microarray analysis of genome dynamics in Yersinia pestis: insights into bacterial genome microevolution and niche adaptation. *Journal Bacteriol* **186**(15), 5138–5146.

Specialist Review

Genomics of the *Mycobacterium tuberculosis* complex and *Mycobacterium leprae*

Stephen V. Gordon
Veterinary Laboratories Agency, New Haw, UK

Roland Brosch
Institut Pasteur, Paris, France

1. Introduction

Mycobacterial infectious diseases are among the most ancient diseases known to man, yet they continue to cause a staggering burden of death and morbidity in the modern world. This is in spite of the fact that the causative agents of tuberculosis and leprosy, *Mycobacterium tuberculosis* and *Mycobacterium leprae* respectively, were among the first pathogens ever to be identified, with Hansen identifying *M. leprae* in 1874 and Koch describing the tubercle bacillus in 1882. Our inability to eradicate these infections is due to a combination of factors, not least of which is the slow pace of research on these organisms, primarily due to the slow growth of *M. tuberculosis* (3 weeks to form visible colonies), and the resistance of *M. leprae* to all attempts at axenic culture.

The *M. tuberculosis* complex is a group of highly related bacteria that are the etiological agents of tuberculosis (TB). The type strain of the complex is *M. tuberculosis*, the agent of human tuberculosis that causes 8 million new infections and kills over 2 million humans a year. The other main member is *Mycobacterium bovis*, the cause of tuberculosis in a range of domesticated and wild mammals, and man, that is responsible for annual losses of approximately $3 billion to world agriculture. At the beginning of the twentieth century, Calmette and Guérin at the Pasteur Institute of Lille used *M. bovis* as the starting point in their search for a vaccine against tuberculosis, and after 13 years of in vitro passage, produced the *M. bovis* BCG strain, currently the most widely used vaccine in the world. The remaining members of the complex are less well known: *M. africanum* and *M. canettii* cause limited numbers of human tuberculosis cases in Africa; *M. caprae* causes tuberculosis in goats and other mammals; *M. pinnipedii* has been found in seals with tuberculosis. *M. microti* is the agent of tuberculosis in wild rodents and voles, but most strains are attenuated for man; this feature led to its use as a live-attenuated vaccine against tuberculosis. However, the *M. tuberculosis* complex are not the only significant pathogens found in the genus.

Leprosy is characterized by peripheral nerve damage and skin lesions, ultimately leading to progressive debilitation. Although the WHO had aimed to eliminate leprosy by 2000, meaning a reduction in prevalence to below 1 case per 10 000 of the population, this target was not achieved. Instead, leprosy is still a major cause of suffering in many parts of the world, with the number of new cases actually increasing.

The third most common mycobacterial disease of immunocompetent people after tuberculosis and leprosy is Buruli ulcer, characterized by debilitating chronic skin lesions that are caused by *Mycobacterium ulcerans*. The prevalence of Buruli ulcer throughout West Africa appears to have increased dramatically since the late 1980s, but the disease also occurs in other parts of the world. Unlike other mycobacterial pathogens, *M. ulcerans* is found mainly at an extracellular location during infection and produces a toxin, mycolactone (George *et al*., 1999). Treatment of the disease is usually by surgical excision of infected and surrounding tissue, as the organism in situ is unresponsive to drug therapy. The mode of transmission of *M. ulcerans* has not yet been completely elucidated, but a transmission chain involving aquatic insects seems likely (Marsollier *et al*., 2002).

Clearly, there is a desperate need for new control strategies if we are to reduce the burden exacted by the mycobacterial pathogens. It was against this backdrop that the genome sequencing projects of *M. tuberculosis*, *M. leprae*, *M. bovis*, *M. bovis* BCG Pasteur, and *M. ulcerans* were initiated. This chapter shall deal with the insights we have gained into the biology of these pathogens through exploration of their genome sequences, and highlight the new research that is developing promising leads into disease control. As a stylistic point, we shall use the *M. tuberculosis* H37Rv genome sequence as the reference, with "Rv" denoting genes from this sequence, while "Mb" indicates *M. bovis* and "ML" indicates *M. leprae* genes. At the time of writing, the genome sequence of *M. ulcerans* is not yet published, but initial sequence analysis has shown that the capacity of *M. ulcerans* to produce its very potent macrolide toxin is linked to the presence of a large 174-kb plasmid harboring the genes involved in the synthesis of mycolactone (Stinear *et al*., 2004). This is the first example of a plasmid-encoded mycobacterial virulence factor.

2. The *M. tuberculosis* complex: genome highlights

The complete genome sequences of two strains of *M. tuberculosis* (H37Rv and CDC1551) and one of *M. bovis* (AF2122/97) are available at the time of writing (Garnier *et al*., 2003; Cole *et al*., 1998; Fleischmann *et al*., 2002). Examination of the genomes underlined the high level of genetic similarity across the members of the complex, with *M. bovis* proving >99.9% identical to the *M. tuberculosis* strains (Garnier *et al*., 2003). Incomplete sequence data is also available for the "210 Beijing" strain of *M. tuberculosis* (**http://www.tigr.org/tdb/mdb/mdbinprogress.html**), as well as for *M. bovis* BCG Pasteur and *M. microti* (**http://www.sanger.ac.uk/Projects/Microbes/**). The exploitation of these data is providing new insights in the evolution, physiology, and virulence of the *M. tuberculosis* complex.

The original genome publications and subsequent reviews have dealt in detail with the initial findings from the completed genome sequences (Brosch *et al*., 2000a; Brosch *et al*., 2000b; Cole *et al*., 1998; Gordon *et al*., 2002; Garnier *et al*., 2003; Fleischmann *et al*., 2002), and it is not our wish to go over this territory again in detail. Rather, we will briefly summarize common features across the genomes of the *M. tuberculosis* complex, highlighting significant differences and new work that provides insight into the biology of the complex.

2.1. Genome structure

As the type strain of the complex, *M. tuberculosis* H37Rv was the first member to have its genome sequenced, revealing a single circular chromosome, ∼4.4 Mb in size with a G+C content of 65.6% (Cole *et al*., 1998). The *M. tuberculosis* CDC1551 sequence showed a similar genome size, with no evidence of extensive translocations, transversions, or duplications relative to the H37Rv strain (Fleischmann *et al*., 2002). An intriguing difference was noted in the ratio of synonymous (silent) to nonsynonymous substitutions (Ks/Ka) between the two strains, with a value of 1.6 (Fleischmann *et al*., 2002). Compared to other bacteria, where a range between ∼8 and 26 as a result of purifying selection acting against amino acid changes is usual, this was surprising. The *M. bovis* sequence revealed a similar overabundance of nonsynonymous substitutions (Garnier *et al*., 2003), suggesting that among these members of the complex, purifying selection had not had time to act, underlining the relatively recent divergence of these strains.

The *M. bovis* genome showed colinearity with the *M. tuberculosis* strains, and no evidence of extensive translocations, duplications, or inversions. Prior to the availability of the *M. bovis* genome sequence, comparative analyses with members of the *M. tuberculosis* complex had been performed using hybridization-based methods, exploiting this high degree of sequence identity (Gordon *et al*., 1999a; Behr *et al*., 1999). This revealed 11 deletions from the genome of *M. bovis*, ranging in size from ∼1 to 12.7 kb, and these were confirmed by the sequence data. Surprisingly, the *M. bovis* sequence contains only one locus in *M. bovis*, termed TbD1 for *M. tuberculosis* specific deleted region 1, which is absent from the majority of extant *M. tuberculosis* strains (Garnier *et al*., 2003). Therefore, at a gross level, deletion has been a major mechanism in shaping the *M. bovis* genome.

2.2. The PE and PPE proteins

Analysis of the *M. tuberculosis* H37Rv genome revealed a number of unexpected findings, not least of which was the presence of two families of repetitive sequences that encode the PE and PPE proteins (Cole *et al*., 1998). The PE family is so called after the presence of the motif proline-glutamic acid (PE) at positions 8 and 9 in a conserved N-terminal region of approximately 110 amino acids. The family contains 99 members that can be subdivided into the PE and PE-PGRS subfamilies, the latter group containing multiple repeats of a glycine-glycine-alanine or glycine-glycine-asparagine. Members of the PPE family contain the

motif proline-proline-glutamic acid (PPE) at positions 7–9, followed by a variable C-terminal region. This family contains 68 members, which can be grouped into three subfamilies. The first of these families contains the major polymorphic tandem repeat (MPTR) sequences that are characterized by asparagine-rich repeats.

The PGRS and MPTR sequences had originally been described as genetically hypervariable loci (Poulet and Cole, 1995; Hermans *et al*., 1992). Use of PGRS-based probes revealed a high degree of genetic differences between tubercle bacilli, allowing the loci to be exploited as epidemiological markers for typing of *M. tuberculosis* complex strains. The availability of genome data has permitted comparative analyses to identify alleles of the PPE proteins that vary across strains. For example, between *M. bovis* AF2122/97 and *M. tuberculosis* H37Rv, there are blocks of sequence variation in genes encoding 29 different PE-PGRS and 28 PPE proteins. While the majority of proteins are identical between the human and bovine bacilli, ∼60% of the PE and PPE proteins differ across these two pathogens, a feature that is clearly at odds with the rest of the genome. This may indicate that the PE and PPE genes can support extensive sequence polymorphism that may provide a source of variation for selective pressures to act upon.

In this light, it is intriguing that there is now a considerable body of evidence to suggest that at least some of the PE and PPE proteins are surface exposed and may play a role in adhesion or immune evasion. Delogu and colleagues have shown that the PE-PGRS Rv1818c is surface exposed, with the PE domain involved in subcellular localization and the PGRS domain influencing cell shape (Delogu *et al*., 2004). Similarly, Banu *et al*. (2002) have shown that polyclonal antibodies raised against the PE-PGRS Rv1411c could be used to surface-stain *M. tuberculosis*. Interestingly, two studies have found that the PPE Rv1753c is essential for growth in vitro (Lamichhane *et al*., 2003; Sassetti *et al*., 2003). However, considering that there are over 160 genes encoding PE and PPE proteins, genome-wide mutagenesis studies have identified only a few members whose disruption leads to attenuation. Hence, transposon insertion into the PE and PPE genes Rv3018, Rv3872, and Rv3873 attenuated the mutant, but this may have been due to polar effects on nearby regions that code for secretion of ESAT-6 family members (Camacho *et al*., 1999; Sassetti and Rubin, 2003). These findings suggest functional redundancy across the PE and PPE proteins. Indeed, the expression of PE and PPE genes has been shown to be controlled by a variety of independent systems, and they show little evidence for global coregulation (Voskuil *et al*., 2004).

2.3. Cell envelope and antigenic variation

Cell walls of pathogenic bacteria are known to show variation in protein sequences and macromolecular composition, reflecting selective pressures on these structures; the PE and PPE proteins offer a case in point. However, this holds true for other cell wall associated proteins, with the greatest degree of sequence variation between the human and bovine bacilli found in genes encoding cell wall and secreted proteins. Comparison of the *M. bovis* and *M. tuberculosis* sequences revealed variation in genes encoding lipoproteins, with *lppO*, *lpqT*, *lpqG*, and *lprM* deleted or frameshifted in *M. bovis* (Garnier *et al*., 2003). Similarly, the *M. bovis rpfA*

gene, one of a five-membered family encoding secreted proteins that promote the resuscitation of dormant or nongrowing bacilli, shows an in-frame deletion of 240 bp that leads to the synthesis of a shorter protein. Whether this affects the function of the protein, or again reflects antigenic variation, is unclear.

A group of known antigens affected by deletions from *M. bovis* is the ESAT-6 family. The ESAT-6 protein was originally described as a potent T-cell antigen secreted by *M. tuberculosis*, and belongs to a >20-membered family that contains other T-cell antigens such as CFP-10 and CFP-7. The demonstration of an interaction between ESAT-6 and CFP-10 suggested that other members of the family may also act in pairs, and this appears to be the case (Okkels and Andersen, 2004; Renshaw *et al*., 2002). Six ESAT-6 proteins, encoded by Rv2346c, Rv2347c, Rv3619c, Rv3620c, Rv3890c (Mb3919c), and Rv3905c (Mb3935c) in *M. tuberculosis*, are missing or altered in *M. bovis*. The consequences of their loss are difficult to predict, although they may impact on antigen load either singly or in combination.

Differences are also seen between *M. tuberculosis* and *M. bovis* in genes encoding the synthesis (*pks*), or transport (*mmpSL*) of polyketides and complex lipids with polyketide moieties. These lipids are major factors in inducing host pathologies that create more favorable environments for the pathogens. The genes *pks1, mmpL13*, and Mb1695c (a putative macrolide transporter adjacent to the *pks10/7/8/17/9/11* cluster) could be translated to functional products in *M. bovis* but are disrupted in *M. tuberculosis*. The opposite is the case (i.e., disrupted in *M. bovis*) for the linked *pks6* and *mmpL1* genes, and *mmpL9*. It has been shown functionally that *pks1* codes for the biosynthesis of the major phenolic glycolipid (PGL) of *M. bovis* and *M. canettii* as in strains where *pks1* is disrupted, such as *M. tuberculosis* H37Rv, where no PGL is produced (Constant *et al*., 2002). Other genes from the flanking regions of the *pks1* locus, such as Rv2958c encoding a putative glycosyl transferase, are involved in the modification of the sugar component of the PGL molecules (Perez *et al*., 2004). As the predicted amino acid sequence for Rv2958c and its *M. bovis* ortholog Mb2982c are only 83% identical, it seems possible that this difference could alter the PGL sugar modifications in *M. bovis* and hence lead to antigenic variability.

The TbD1 locus, containing the gene *mmpS6* and the 5′ region of *mmpL6*, is absent from the majority of *M. tuberculosis* strains but intact in *M. bovis* strains (Brosch *et al*., 2002). Deletion of TbD1 may prevent trafficking of specific lipids to the cell wall of *M. tuberculosis*. However, according to the sequence characteristics, the deletion of the TbD1 region resulted in the fusion of the remaining *mmpS6* and *mmpL6* fragments. It has not yet been determined whether the MmpS/MmpL6 hybrid protein is expressed in TbD1 deleted strains or has a specific function in them (Figure 1). Furthermore, a deletion in *M. bovis* of 808 bp is proximal to the TbD1 region and truncates the *treY* gene. As *treY* encodes a maltooligosyltrehalose synthase, an enzyme in a pathway for trehalose production (two other pathways are intact), its deletion in *M. bovis* may have an effect on the range of trehalose-based glycolipids that are produced.

2.4. *Metabolic insight*

The mycobacterial cell wall has been the target of in-depth studies for many decades, revealing a complex repertoire of unusual lipids that give the structure its

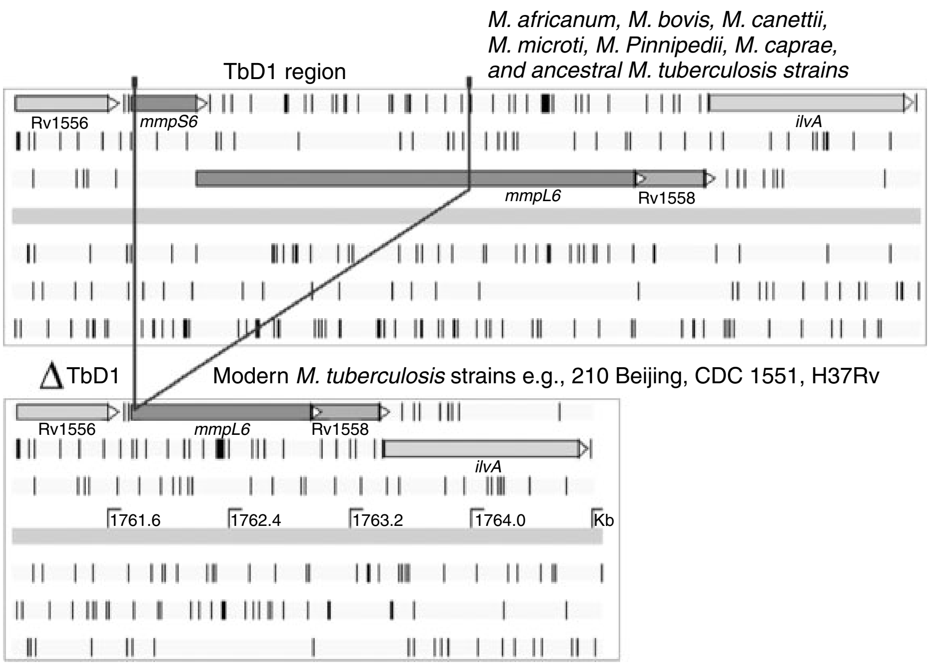


Figure 1 The TbD1 deletion locus. The figure shows a schematic of the TbD1 locus in *M. tuberculosis* (ΔTbD1) and in other strains of the complex where the region is intact. The six reading frames are shown, with CDSs represented as pointed boxes showing the direction of transcription, and stop codons shown as small vertical bars. Gene designations are as described on the TubercuList database (**http://genolist.pasteur.fr/TubercuList/**)

unique architecture. Hence, it could have been expected that the mycobacterial genome would encode sophisticated machinery for lipid metabolism; however, it was still unexpected that over 9% of the genome coding capacity of H37Rv would be dedicated to lipids. A striking level of redundancy in lipid metabolic genes was apparent in the *M. tuberculosis* H37Rv genome with 36 *fadD* alleles encoding acyl-CoA synthase, 36 *fadE* genes encoding acyl-CoA dehydrogenase, and 21 *echA* genes for enoyl-CoA hydratase/isomerase (Cole *et al*., 1998). The apparent redundancy in lipid enzymes may however need to be readdressed in the light of recent work with the *fadD* genes. Gokhale and colleagues have shown that some of the *fadD* alleles do not encode fatty acyl-CoA ligases but instead code for a new class of fatty acyl-AMP ligases that are linked to a proximal *pks* gene and are dedicated to the synthesis of unique polyketides (Trivedi *et al*., 2004). It is therefore possible that the apparent redundancy in lipid enzymes hides novel enzyme activities.

An example of the power of comparative genomics is evidenced by the elucidation of the reason for one of the key in vitro differences between *M. bovis* and *M. tuberculosis*. The bovine bacillus requires pyruvate to be added to media where glycerol is the sole carbon source, presumably reflecting a defect in the metabolism of glycerol by *M. bovis*. Trawls of the *M. bovis* genome sequence revealed a point mutation in the gene for pyruvate kinase (PK), the enzyme that catalyzes the final irreversible step in glycolysis, the dephosphorylation of phosphoenolpyruvate to

Strain	Sequence (codons 215-225)	PK activity
M. tuberculosis H37Rv	GTG ATC GCC AAG CTG GAG AAG CCG GAA GCC ATC	Yes
M. bovis 2122/97	GTG ATC GCC AAG CTG GAT AAG CCG GAA GCC ATC	No
M. bovis 1307/01	GTG ATC GCC AAG CTG GAT AAG CCG GAA GCC ATC	No
M. bovis BCG	GTG ATC GCC AAG CTG GAG AAG CCG GAA GCC ATC	Yes
M. bovis AN5	GTG ATC GCC AAG CTG GAG AAG CCG GAA GCC ATC	Yes
	V I N K L E/D K P E A I	

(a)

(b)

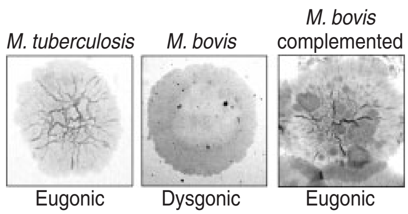


Figure 2 Pyruvate kinase analysis across *M. bovis* and *M. tuberculosis*. (a) A sequence alignment of codons 215–225 of the *pykA* gene is shown from *M. tuberculosis* H37Rv, *M. bovis* 2122/97, *M. bovis* 1307/01 (a recent field isolate), *M. bovis* BCG Pasteur, and *M. bovis* AN5 (used for production of bovine tuberculin). Pyruvate kinase is shown as active (yes) or inactive (no). The latter two strains were adapted for growth on glycerol, a process that was selected for an active PK. (b) Growth of strains on glycerol-containing medium. Strains are described as eugonic (abundant growth in the presence of glycerol, dry, crumbly, and raised colonies) or dysgonic (sparse growth on solid medium containing glycerol, colonies moist, glossy, and flat). The picture shows the colony morphology of *M. tuberculosis* H37Rv, *M. bovis* wild-type, *pykA*$^+$ complemented *M. bovis*, and *M. bovis*/plasmid control, after three-week growth on Middlebrook 7H11 agar plates with 0.05% glycerol and antibiotics as appropriate

pyruvate. The mutation was predicted to affect binding of the Mg^{++} cofactor to PK, and enzyme analysis showed that *M. bovis* lacked pyruvate kinase activity (Figure 2a). Hence, in *M. bovis*, glycolytic intermediates are blocked from feeding into oxidative metabolism, meaning that in vivo *M. bovis* must rely on amino acids or fatty acids as a carbon source for energy metabolism. Complementation of *M. bovis* with the *pykA* allele from *M. tuberculosis* H37Rv permitted abundant growth of the recombinant *M. bovis* on glycerol-containing media. However, the complemented strain also displayed the characteristic "eugonic" colony morphology that is classically associated with *M. tuberculosis* strains; *M. bovis* strains normally display "dysgonic" growth on media containing glycerol (Keating *et al*., submitted; Figure 2b). Hence, it appears that the presence of a functioning PK enzyme is intimately linked to the surface features of the bacillus.

Although the tubercle bacilli are classified as aerobic organisms, the genome data revealed the potential for microaerophilic and anaerobic respiration. An operon, *narGHJI*, is present, encoding a nitrate reductase that allows utilization of nitrate as a terminal electron acceptor. Investigating the role of nitrate reductase, Bange and colleagues generated a *narG* mutant of *M. bovis* BCG (Weber *et al*., 2000). Immunodeficient mice infected with the *narG* mutant developed smaller granulomas than those infected with the wild type. Furthermore, mice infected with the mutant presented no clinical signs of disease after more than 200 days. It, therefore, appears that the ability to respire anaerobically contributes to virulence. It is also noteworthy that one of the classical microbiological methods to differentiate *M. bovis* from *M. tuberculosis* is based on nitrate reductase activity; *M. tuberculosis* reduces nitrate to nitrite while *M. bovis* performs this reduction very poorly. Bange

and colleagues have also shown that this defect in nitrate reductase activity is due to a point mutation in the promoter of the *M. bovis narGHIJ* cluster (Stermann *et al*., 2004).

2.5. Repetitive DNA

In contrast to the high degree of nucleotide identity across the *M. tuberculosis* complex, repetitive DNA acts as a substrate for the generation of genetic diversity, for example, through recombination between repeats or transposition of IS elements. Analysis of the *M. tuberculosis* H37Rv genome revealed 56 loci with similarity to IS elements, occupying approximately 77 kb (Gordon *et al*., 1999b). These elements could be classified into the major IS families such as IS*3*, IS*5*, or IS*21*, with the most abundant IS in the genome being IS*6110,* an IS*3* family member that varies from 0 to >25 copies across strains. This divergence in copy number between *M. tuberculosis* isolates is the basis of the IS*6110*-based molecular typing system. What was originally thought to be a novel IS grouping, the IS*1535* family, was identified that contained six intact elements and one pseudogene. However, with the increase in genome sequence from other bacteria, it is now apparent that the IS*1535* group forms part of the IS*605* family (as defined in the IS Finder database, **http://www-is.biotoul.fr/is.html**).

A novel class of repeats was uncovered by Supply and colleagues and designated "Mycobacterial Interspersed Repetitive Units", or MIRUs (Supply *et al*., 1997). These elements are dispersed throughout the genomes of the *Mycobacterium* complex strains and *M. leprae.* They display no significant homology to other bacterial repetitive sequences, and in contrast to many other repeat elements, they do not contain obvious palindromic structures. Intriguingly, many MIRUs are located intergenically, overlapping termination and initiation codons of adjacent genes, and they contain small CDSs oriented in the same translational direction as the contiguous genes, an arrangement that strongly suggests translational coupling. MIRUs have also been developed as a powerful typing system for the *M. tuberculosis* complex, and Supply and colleagues have used data from 12 MIRU loci across a bank of *M. tuberculosis* strains to show that the population structure of *M. tuberculosis* was clonal (Supply *et al*., 2003).

The so-called direct repeat (DR) region harbors numerous direct repeat units of 36 bp that are interspersed with 27- to 41-bp segments of unique sequence. Recombination events between the direct repeats and/or copies of IS*6110* elements inserted in this genomic region create polymorphism across strains of the *M. tuberculosis* complex that can be visualized and distinguished by a technique called spoligotyping (Kamerbeek *et al*., 1997). Comparative analysis of information from spoligotype with analysis of genome deletions, point mutations, and MIRU variation has shown that combinations of these molecular characteristics appear to be strictly correlated in strains of the *M. tuberculosis* complex, allowing the development of rapid strain identification and typing strategies (Banu *et al*., 2004). Figure 3 shows an example of how these data can be used to group strains and define certain strain families, such as the so-called *M. tuberculosis* Beijing family, which is a particularly successful clonal group.

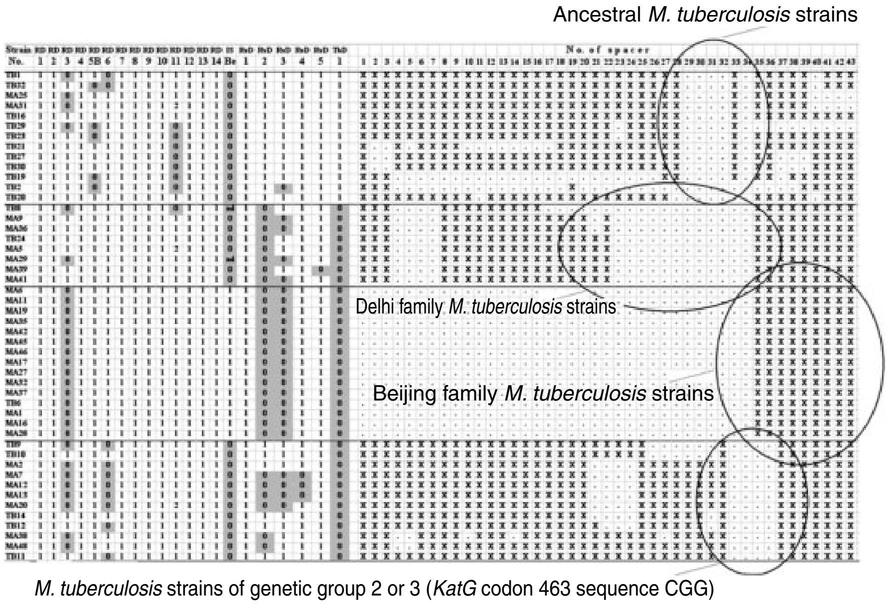


Figure 3 Genetic characterization of *M. tuberculosis* clonal groups. A range of molecular markers were screened against the panel of *M. tuberculosis* strains described in Banu *et al*. (2004). PCR results for the "regions of difference", or RD, are shown, with 0 = region not present; 1 = region present; 2 = both internal and flanking primers gave a product. "Be" designates the presence of the IS *6110* element in the *dnaA-N* locus, which is indicative of the Beijing family of *M. tuberculosis*. Spoligotype results are also shown against the standard set of 43 spacers, where an "X" shows that the spacer is present, while "-" designates a deleted spacer. Clonal groups of *M. tuberculosis* strains have clear combinations of markers; ND, not determined

Two prophage-like elements, phiRv1 and phiRv2, are present on the *M. tuberculosis* H37Rv genome. The presence and position of these prophages is variable across sequenced members of the *M. tuberculosis* complex, with phiRv2 not present in *M. bovis* AF2122/97, *M. bovis* BCG missing both prophages, and with *M. tuberculosis* CDC1551 having phiRv2 integrated at a different locus than H37Rv. The variation in integration sites of phiRv2 is due to its attachment site being contained in at least four copies of the REP13E12 repeats, a family of seven repeats that range in size from 1.3 to 1.4 kb and so called after the annotation of the first element on cosmid MTCY13E12 (Cole *et al*., 1998).

2.6. *Evolution*

Comparative studies of *M. bovis* BCG, *M. bovis*, and *M. tuberculosis* H37Rv identified several regions of difference (RD) absent from BCG (Mahairas *et al*., 1996; Gordon *et al*., 1999a; Behr *et al*., 1999). From close inspection of these regions, it was apparent that several segments of conserved genes were missing from BCG and *M. bovis* that were still intact in *M. tuberculosis*, arguing against the possibility that the RD regions resulted from the insertion of genomic

material into *M. tuberculosis,* but rather represent deletions (Brosch *et al*., 2001). Together with the finding that the *M. tuberculosis* complex has a clonal population structure, with no evidence of horizontal transfer (recombination) between strains (Supply *et al*., 2003), deletion events can be used as unidirectional markers for phylogenetic reconstruction. Using deletion analysis, we were able to generate a novel evolutionary scenario for the *M. tuberculosis* complex, based on the successive loss of DNA from certain lineages (Brosch *et al*., 2002) that contradicted previous theories. It had been assumed that *M. tuberculosis* arose at the time of the domestication of cattle, approximately 10 000–15 000 years ago, when the bovine tubercle bacillus was transmitted to the human population. This was based on the observation that extant *M. bovis* strains have a wide host range, infecting wild and domesticated mammals as well as man, while *M. tuberculosis* appears restricted to humans. The new scenario refutes the notion that modern strains of *M. bovis* should lie closer to the common ancestor of the complex, and in fact places the common ancestor closer to *M. tuberculosis* and/or *M. canettii*. Since then, this phylogeny has been backed up by other studies (Gutacker *et al*., 2002; Mostowy *et al*., 2002) and the *M. bovis* genome sequencing project, which found no unique genes per se in *M. bovis* relative to *M. tuberculosis* (Garnier *et al*., 2003).

Deletion analysis has also provided fundamental insight into the evolutionary history and transmission patterns of *M. tuberculosis*. Using microarray analysis of 100 epidemiologically well-defined *M. tuberculosis* isolates from San Francisco, Small and colleagues were able to show that strains show close association with their host populations over time, so much so that a patient's region of birth could be used as a predictor of strain carriage (Hirsh *et al*., 2004). Their analysis confirmed the lack of any significant horizontal gene transfer in *M. tuberculosis*, and also showed that deletions were grouped in the genome, revealing regions of the genome that are prone to deletions (Hirsh *et al*., 2004; Tsolaki *et al*., 2004). The functional implications of deletions were less clear cut; on balance, it appeared that most deletion events were slightly deleterious, with a minority, such as loss of the *katG* region that imparts resistance to the front-line drug isonizaid, offering any obvious selective advantage (Tsolaki *et al*., 2004).

3. Genomics of the leprosy bacillus

3.1. Genome downsizing

The genome of *M. leprae* consists of a single circular chromosome of 3 268 203 bp (Cole *et al*., 2001a). The most remarkable feature of the genome is the wholesale loss of genetic information, with a coding capacity of less than 50% compared to 90% in *M. tuberculosis*. This reduction in coding capacity is due both to the deletion of chromosomal regions and the accumulation of pseudogenes. These decayed gene remnants are abundant and presumably reflect the removal of functions superfluous to the in vivo growth of the bacillus; alternatively, they may indicate an accumulation of deleterious mutations as the organism goes through bottlenecks in its secluded niche.

Potentially, the most interesting genes from the point of view of understanding *M. leprae* are the 165 genes that have no counterpart in other sequenced mycobacteria. While a putative function can be ascribed to 29 of these genes, 136 have no similarity to other genes currently in sequence databases, and this suggests that their functions may be specific to the leprosy bacillus. From a clinical point of view, these genes are attractive candidates for specific diagnostic reagents.

Perhaps the most striking difference in gene families between the leprosy and tubercle bacilli is the case of the PE and PPE proteins. While these two families account for 167 genes in *M. tuberculosis*, only nine intact PE or PPE genes could be identified in *M. leprae*, with complete loss of the PE–PGRS subfamily. This most likely reflects both downsizing in *M. leprae* and expansion in *M. tuberculosis*. Interestingly, some of the intact PE and PPE genes in *M. leprae*, such as ML1828, ML1182, or ML0411, are in gene-poor regions or surrounded by pseudogenes (Cole *et al*., 2001a). Hence, while neighboring loci were deleted or accumulated mutations, these PE and PPE genes were maintained, suggesting a requirement for at least a minimal set of functional PE and PPE proteins.

Indeed, there are some parallels with gene loss between *M. bovis* and *M. leprae*, with many of the common genes either deleted or inactivated in the two organisms. For example, genes involved in transport and cell surface structures *(pstB*, *ugpA, mce3A-F, lppO, lpqG, lprM, pks6, mmpL1, mmpL9,* Rv1510, Rv1508, Rv1371), fatty acid metabolism (*fadE22, echA1),* cofactor biosynthesis *(moaE*, *moaC2)*, detoxification (*ephA*, *ephF, alkA*), and intermediary metabolism (*epiA, gmdA*) are pseudogenes or deleted in both bacilli (Cole *et al*., 2001a; Garnier *et al*., 2003). Similarly, *M. leprae* and *M. bovis* have lost the AtsA system for recycling sulfate. AtsA is an arylsulfatase that catalyzes the hydrolysis of sulfate esters to release inorganic sulfate. Loss of this function may reflect the lack of sulfated glycolipid in these two mycobacteria. Furthermore, *recBCD* are deleted in *M. leprae*, while *recB* is frameshifted in *M. bovis*.

3.2. *Metabolic insight*

While the genome of *M. tuberculosis* encodes a broad metabolic potential, *M. leprae* appears to have streamlined its genome to the minimum required. This is perhaps best seen in lipid metabolism, where the multiplicity of *fadD*, *fadE*, and *echA* enzymes in *M. tuberculosis* is reduced to 8 *fadD*, 4 *fadE*, and 2 *echA* in *M. leprae*. Of the 8 *fadD*, 4 are of the novel fatty acyl-AMP ligase family (*fadD26*, *fadD28*, *fadD29*, *fadD32*,) and each has their cognate PKS gene intact, indicating the functional linkage of these genes (Trivedi *et al*., 2004). Similarly, while there are 22 putative lipases in *M. tuberculosis*, only 2 are present in *M. leprae*; the consequence of this may be to reduce the spectrum of lipid substrates that can be exploited by the bacillus.

Mycobacterium leprae has lost genes involved in microaerophilic and anaerobic metabolism, such as fumarate reductase and nitrate reductase. This is somewhat surprising in that it is predicted that the in vivo environment is oxygen limited, and may be reflected in the slow growth of the bacillus in the host. Most strikingly, *M. leprae* has deleted the majority of the operon encoding the membrane spanning

NADH oxidase involved in recycling NADH. Hence, the bacillus will be limited both in its capacity to recycle NADH and to generate energy from NADH produced from the TCA cycle. As the ability to recycle NADH is essential, it is possible that enzymes such as malate dehydrogenase, lactate dehydrogenase, or the putative NADH dehydrogenase encoded by ML2061, may function to regenerate NAD.

While catabolic functions have suffered heavily through mutation, the anabolic systems of the bacillus appear replete (Wheeler, 2001). Indeed, the coding commitment to cell wall biosynthesis is very similar to that in *M. tuberculosis*, stressing the complexity of this structure. Surprisingly for an obligate intracellular pathogen, all genes necessary for the biosynthesis of purines and pyrimidines are present. Similarly, the only lesions in amino acid biosynthesis are in the pathway to methionine. This would suggest that the in vivo niche occupied by the organism is nutrient poor, necessitating de novo synthesis. This may also explain the loss of most amino acid permeases. However, it is remarkable in an organism with such extensive streamlining that the asparagine permease is duplicated. Considering that asparagine is the preferred nitrogen source of tubercle bacilli, this may reflect a common metabolic preference.

3.3. Repetitive DNA

The specialized niche occupied by *M. leprae* limits its opportunity for contact with other organisms. Horizontal transfer of genetic material, such as insertion sequences (IS), would therefore be a rare event. This is borne out by the evidence in the genome, where there are 26 transposase pseudogenes but apparently no functional IS. These elements were most likely acquired prior to sequestration of the bacillus in the in vivo niche, with subsequent inactivation by mutation.

Tandem repeats are found in the genome of *M. leprae*, and some of these are proving useful as markers for molecular epidemiological analysis. Matsuoka *et al*. (2004) identified a TTC microsatellite repeat that varied from 8 to 29 copies across *M. leprae* isolates. They used the variability at this locus to show that there were multiple sources of infection in a local village setting, with infected family members showing different genotypes. MIRU minisatellite repeats are also found in *M. leprae*, but a scan of their variability across 14 *M. leprae* isolates from a wide geographical distribution did not reveal any evidence of polymorphism (Cole *et al*., 2001b), contrary to what is seen in *M. tuberculosis* and ruling out their utility as an epidemiological tool.

The genome contains at least 4 families of repetitive DNA contributing approximately 2% to the total genome size (Cole *et al*., 2001b). These sequences are designated RLEP (37 copies), REPLEP (15 copies), LEPREP (8 copies) and LEPRPT (5 copies), and are specific to *M. leprae*. The LEPREP sequence contains pseudogenes with similarity to transposases and the maturases of class II introns, enzymes that catalyze DNA transposition. This suggests a once-functional mechanism for replication of the sequence through the genome. Recombination between repetitive loci appears to have been a central shaper of the *M. leprae* genome. Comparison with *M. tuberculosis* reveals that repetitive elements occupy the junctions of mosaic segments. It is probable that replication, or homologous

recombination, of repetitive elements led to looping out and excision of intervening DNA segments, catalyzed by the once-functional RecBCD complex.

3.4. Failure of axenic culture

The genome sequence does not reveal a specific reason for the inability to culture *M. leprae* axenically. It is most likely that *M. leprae* is so specialized to the host niche that in vitro culture will require a detailed knowledge of the carbon sources and metabolites available to the bacillus in vivo. Moreover, the extensive loss of genes for catabolic functions suggests a limited range of carbon sources would support growth. Metabolic streamlining will also affect the flux of carbon through metabolism, again suggesting that a specific set of carbon substrates will be needed to maintain equilibrium. Computer-aided metabolic reconstruction, with particular reference to energy sources available in vivo, offers one route to the identification of the likely carbon substrates. However, as the doubling rate of *M. leprae* in vivo is 14 days, it would take over 1 year for a colony to form even if in vitro growth was achieved.

4. Conclusions

In contrast to the situation in the 1970/1980s when mycobacterial genetics lagged behind the strides being made in other bacteria, today the mycobacteria represent a genomically well-characterized genus, for which numerous genetic tools exist. Apart from the four published complete genome sequences of *M. tuberculosis*, *M. bovis*, and *M. leprae*, a range of other mycobacterial genome sequencing projects are at different stages of completion, including the vaccine strains BCG and *M. microti*, as well as the fish pathogen *M. marinum*, several strains from the *M. avium-intracellulare* complex, and the environmental organism *M. smegmatis*. (For an overview, see **http://www.pasteur.fr/recherche/unites/Lgmb/Overview-Genome-Projects.html**.) This information provides a knowledge base that is catalyzing research into new disease-control strategies in diagnosis (see Cockle *et al*., 2002), vaccine development (e.g., De Groot and Martin, 2003), and drug-targets (see Smith and Sacchettini, 2003) that are desperately needed to cope with the burden of mycobacterial disease.

Acknowledgments

The work described herein was funded by The Wellcome Trust, The Institute Pasteur, the Association Française Raoul Follereau, ILEP, the New York Community Trust, and the UK Department of Food, Environment and Rural Affairs (DEFRA).

References

Banu S, Gordon SV, Palmer S, Islam MR, Ahmed S, Alam KM, Cole ST and Brosch R (2004) Genotypic analysis of *Mycobacterium tuberculosis* in Bangladesh and prevalence of the Beijing strain. *Journal of Clinical Microbiology*, **42**, 674–682.

Banu S, Honore N, Saint-Joanis B, Philpott D, Prevost MC and Cole ST (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular Microbiology*, **44**, 9–19.

Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S and Small PM (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science*, **284**, 1520–1523.

Brosch R, Gordon SV, Eiglmeier K, Garnier T and Cole ST (2000a) Comparative genomics of the leprosy and tubercle bacilli. *Research in Microbiology*, **151**, 135–142.

Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, *et al*. (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3684–3689.

Brosch R, Gordon SV, Pym A, Eiglmeier K, Garnier T and Cole ST (2000b) Comparative genomics of the mycobacteria. *International Journal of Medical Microbiology*, **290**, 143–152.

Brosch R, Pym AS, Gordon SV and Cole ST (2001) The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends in Microbiology*, **9**, 452–458.

Camacho LR, Ensergueix D, Perez E, Gicquel B and Guilhot C (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Molecular Microbiology*, **34**, 257–267.

Cockle PJ, Gordon SV, Lalvani A, Buddle BM, Hewinson RG and Vordermeier HM (2002) Identification of novel *Mycobacterium tuberculosis* antigens with potential as diagnostic reagents or subunit vaccine candidates by comparative genomics. *Infection and Immunity*, **70**, 6996–7003.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, *et al*. (2001a) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.

Cole ST, Supply P and Honore N (2001b) Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. *Leprosy Review*, **72**, 449–461.

Constant P, Perez E, Malaga W, Laneelle MA, Saurel O, Daffe M and Guilhot C (2002) Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the M. tuberculosis complex: Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbour a frameshift mutation in the pks15/1 gene. *The Journal of Biological Chemistry*, **277**, 38148–38158.

De Groot AS and Martin W (2003) From immunome to vaccine: epitope mapping and vaccine design tools. *Novartis Foundation Symposium*, **254**, 57–72, discussion 72–6, 98–101, 250–2.

Delogu G, Pusceddu C, Bua A, Fadda G, Brennan MJ and Zanetti S (2004) Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Molecular Microbiology*, **52**, 725–733.

Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, *et al*. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *Journal of Bacteriology*, **184**, 5479–5490.

Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, *et al*. (2003) The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7877–7882.

George KM, Chatterjee D, Gunawardana G, Welty D, Hayman J, Lee R and Small PL (1999) Mycolactone: a polyketide toxin from *Mycobacterium ulcerans* required for virulence. *Science*, **283**, 854–857.

Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K and Cole ST (1999a) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Molecular Microbiology*, **32**, 643–655.

Gordon SV, Brosch R, Eiglmeier K, Garnier T, Hewinson RG and Cole ST (2002) Royal Society of Tropical Medicine and Hygiene Meeting at Manson House, London, 18th January 2001.

Pathogen genomes and human health. Mycobacterial genomics. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **96**, 1–6.

Gordon SV, Heym B, Parkhill J, Barrell B and Cole ST (1999b) New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, **145**(Pt 4), 881–892.

Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN and Musser JM (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics*, **162**, 1533–1543.

Hermans PW, van Soolingen D and van Embden JD (1992) Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *Journal of Bacteriology*, **174**, 4157–4165.

Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW and Small PM (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4871–4876.

Kamerbeek J, Schouls L, Kolk A, vanAgterveld M, vanSoolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, *et al*. (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology*, **35**, 907–914.

Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Grosset J, Broman KW and Bishai WR (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7213–7218.

Mahairas GG, Sabo PJ, Hickey MJ, Singh DC and Stover CK (1996) Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *Journal of Bacteriology*, **178**, 1274–1282.

Marsollier L, Robert R, Aubry J, Saint Andre JP, Kouakou H, Legras P, Manceau AL, Mahaza C and Carbonnelle B (2002) Aquatic insects as a vector for Mycobacterium ulcerans. *Applied and Environmental Microbiology*, **68**, 4623–4628.

Matsuoka M, Zhang L, Budiawan T, Saeki K and Izumi S (2004) Genotyping of *Mycobacterium leprae* on the basis of the polymorphism of TTC repeats for analysis of leprosy transmission. *Journal of Clinical Microbiology*, **42**, 741–745.

Mostowy S, Cousins D, Brinkman J, Aranaz A and Behr MA (2002) Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *The Journal of Infectious Diseases*, **186**, 74–80.

Okkels LM and Andersen P (2004) Protein-protein interactions of proteins from the ESAT-6 family of *Mycobacterium tuberculosis*. *Journal of Bacteriology*, **186**, 2487–2491.

Perez E, Constant P, Lemassu A, Laval F, Daffe M and Guilhot C (2004) Characterization of three glycosyltransferases involved in the biosynthesis of the phenolic glycolipid antigens from the *Mycobacterium tuberculosis* complex. *The Journal of Biological Chemistry*. **279**, 42584–42592.

Poulet S and Cole ST (1995) Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Archives of Microbiology*, **163**, 87–95.

Renshaw PS, Panagiotidou P, Whelan A, Gordon SV, Hewinson RG, Williamson RA and Carr MD (2002) Conclusive evidence that the major T-cell antigens of the *Mycobacterium tuberculosis* complex ESAT-6 and CFP-10 form a tight, 1:1 complex and characterization of

the structural properties of ESAT-6, CFP-10, and the ESAT-6*CFP-10 complex. Implications for pathogenesis and virulence. *The Journal of Biological Chemistry*, **277**, 21598–21603.

Sassetti CM, Boyd DH and Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, **48**, 77–84.

Sassetti CM and Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12989–12994.

Smith CV and Sacchettini JC (2003) *Mycobacterium tuberculosis*: a model system for structural genomics. *Current Opinion in Structural Biology*, **13**, 658–664.

Stermann M, Sedlacek L, Maass S and Bange FC (2004) A promoter mutation causes differential nitrate reductase activity of *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *Journal of Bacteriology*, **186**, 2856–2861.

Stinear TP, Mve-Obiang A, Small PL, Frigui W, Pryor MJ, Brosch R, Jenkin GA, Johnson PD, Davies JK, Lee RE, *et al*. (2004) Giant plasmid-encoded polyketide synthases produce the macrolide toxin of *Mycobacterium ulcerans*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 1345–1349.

Supply P, Magdalena J, Himpens S and Locht C (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Molecular Microbiology*, **26**, 991–1003.

Supply P, Warren RM, Banuls AL, Lesjean S, Van Der Spuy GD, Lewis LA, Tibayrenc M, Van Helden PD and Locht C (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Molecular Microbiology*, **47**, 529–538.

Trivedi OA, Arora P, Sridharan V, Tickoo R, Mohanty D and Gokhale RS (2004) Enzymic activation and transfer of fatty acids as acyl-adenylates in mycobacteria. *Nature*, **428**, 441–445.

Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M and Small PM (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4865–4870.

Voskuil MI, Schnappinger D, Rutherford R, Liu Y and Schoolnik GK (2004) Regulation of the *Mycobacterium tuberculosis* PE/PPE genes. *Tuberculosis (Edinb)*, **84**, 256–262.

Weber I, Fritz C, Ruttkowski S, Kreft A and Bange FC (2000) Anaerobic nitrate reductase (narGHJI) activity of *Mycobacterium bovis* BCG in vitro and its contribution to virulence in immunodeficient mice. *Molecular Microbiology*, **35**, 1017–1025.

Wheeler PR (2001) The microbial physiologist's guide to the leprosy genome. *Leprosy Review*, **72**, 399–407.

Specialist Review

The Mycoplasmas – a congruent path toward minimal life functions

Leka Papazisi and Scott N. Peterson
The Institute for Genomic Research, Rockville, MD, USA

1. The mycoplasma lifestyle

The *Mycoplasmas* are a diverse group of bacteria that arose within the low-G+C gram-positive branch approximately 600 million years ago (Maniloff, 1996; Woese *et al*., 1980). Within this group, genomes range in size between 580 Kb (*M. genitalium*) and 1700 Kb. Despite their ancestral relationship to gram-positive bacteria, *Mycoplasmas* lack a cell wall and the genes encoding cell wall components. Human and animal infections with *Mycoplasmas* appear to affect mostly mucosal tissues in the respiratory and genito-urinary tract, joints, or mammary glands. Generally, infections are not fatal to the host, but instead result in chronic disease sequelae. *Mycoplasmas* have long been considered as surface pathogens, however, more recently, an increasing number of reports suggest an intracellular existence (Lo *et al*., 1989; Lo, 1992; Dallo and Baseman, 2000; Baseman *et al*., 1995; Winner *et al*., 2000; Much *et al*., 2002). It is not yet clear if these invasive *Mycoplasma* species can replicate inside the host cell, but the ability to establish an intracellular lifestyle helps to explain the persistence of *Mycoplasma* infection in the face of host immune response and antibiotic treatment. Readers wishing to learn more about the biology of these bacteria are directed to several excellent reviews (Dybvig and Voelker, 1996, Razin *et al*., 1998, Razin and Herrmann, 2002, Maniloff *et al*., 1992).

As we implement comparative genomic characterization of various bacterial species, it is becoming increasingly clear that bacterial genomes are dynamic and that many genes are more appropriately viewed as being in flux rather than permanent residents. The parasitic lifestyle of the *Mycoplasmas* has allowed the loss of numerous genes encoding proteins representing complete or nearly complete metabolic pathways. *Mycoplasma* genomes are essentially devoid of enzymes involved in the biosynthesis of any amino acids, nucleotides, or lipids. The energy-producing pathways are also minimal. *Mycoplasmas* lack enzymes associated with the tricarboxylic acid cycle and cytochromes. Given the unifying theme of parasitism and the pathway of reductive evolution, it is remarkable that such a diversity

of genomic solutions has arisen for the purpose of energy production. Fermentative *Mycoplasmas* utilize carbohydrates via glycolysis, whereas nonfermentative cells carry out the hydrolysis of arginine. The *Ureaplasmas*, whose favored niche is the urogenital tract, generate 90% of their energy requirements through urea hydrolysis (Smith *et al*., 1993; Razin *et al*., 1998). When one applies clustering analysis to diverse bacterial genomes on the basis of the presence or absence of genes, the resulting clusters group the bacteria bearing minimal genomes together despite their sometimes distant phylogenetic relationships (Hutchison III and Montague, 2002). This result suggests that reductive evolution is, to a large extent, convergent. A comprehensive analysis of the presumed metabolic capacities of the *Mycoplasmas* by Pollack resulted in the identification of a "*Mycoplasma* consensus metabolism" (Pollack, 2002a,b), which, as expected, underscores a "genomic economy". It is likely that as we improve our comparative capabilities, we will see other examples that display the diversity of environmental-specific solutions to common cellular demands. At first glance, it would seem that bacteria have found many ways of achieving fitness that depend on a very large number of genes present on earth.

2. The "minimal cell"

Chromosomal segments within bacterial genomes are both acquired and lost at frequencies far greater than previously appreciated. The rate at which genomes evolve is related, in part, to the type and magnitude of selective pressures the microbe is forced to contend with. Some gene-acquisition events are special in that the inherited genes allow the microbe to occupy a new environment. A new environment generally involves new selective pressures. Therefore, the selective value of each gene in a genome is redefined in accordance with the new environment. Genes possessing a selective value below a particular threshold are likely to be lost from the genome. The transition to an intracellular existence may represent a situation wherein the selective pressures are reduced dramatically, especially for certain types of genes. This mode of genome sampling and clearing of nonadvantageous genes is not unique to the minimal bacteria but rather thought to represent a generalized means of adaptation in the bacterial world. The *Mycoplasmas* have simply taken things to an extreme, because their environment allows them to. It would appear that the uniformity of pathogenic (parasitic) character among the *Mycoplasmas* and a small genome is not coincidental. Some genomes, like those of the *Mycoplasmas*, have come close to achieving an environmentally defined genome minimum. We remain ignorant as to just how close they have come.

A cell bearing a truly minimal genome does not exist in nature, since the formation of such a genome could only occur in an environment that is free of selective pressure. Mushegian and Koonin (1996) were the first to apply a computationally based comparative analysis of completely sequenced microbial genomes as a means of identifying a minimal gene set. Through this analysis, a core set of conserved genes was identified that encodes the basic machinery required for translation, DNA replication, minimal transcription, anaerobic metabolism (glycolysis and substrate level phosphorylation), and a minimal metabolite transport system (Mushegian and

Koonin, 1996). The genomic era has provided impetus for the increase in experimental reports intended to globally identify dispensable genes in genomes, in order to infer which genes are essential (Hutchison *et al*., 1999; Gerdes *et al*., 2003; Kobayashi *et al*., 2003; Ji *et al*., 2001; Jardine *et al*., 2002). Each of these studies defines a similar but distinct set of core functions representing highly conserved genes similar to those defined by the computational approach. It is interesting that each of these studies also identifies a substantial number of essential genes that are lineage specific and poorly conserved over even short phylogenetic distances. For example, the essentiality of the fimbrial gene *fim*A has been speculated to correlate with Zinc transport (Akerley *et al*., 2002). In *Bacillus subtilis*, most genes involved in the Embden–Meyerhof–Parnas pathway (not considered essential) were found to be essential indeed, which led investigators to speculate that besides their (primary) annotated functions, they may play additional roles in this bacterium (Kobayashi *et al*., 2003).

3. Mycoplasma genomes

A wealth of whole-genome sequencing data pertaining to *Mycoplasmas* has allowed many new insights to be made and has furthered our understanding of genome properties common to minimal genomes. The low-G+C nucleotide composition is a signature, not only of Mycoplasma genomes but also of other obligate intracellular pathogens or endo-symbionts. A positive correlation exists between genome size and G+C content in eubacteria. Comparative analysis indicates that genes encoding many DNA repair functions are absent among bacteria with minimal genomes. For example, the gene encoding for uracil *N*-glycosylase (*ung*), an enzyme mediating the removal of uracil from DNA, is absent in some species. Interestingly, those *Mycoplasma* genomes retaining uracil *N*-glycosylase activity possess enzymes with reduced activity, compared to the *Escherichia coli* encoded enzyme (Razin *et al*., 1998; Zou and Dybvig, 2002). The significance of an AT-rich genome and its relationship to minimal genomes is not yet clear, but it has been proposed to be a means of conserving energy (Rocha and Danchin, 2002) and/or a mechanism for evading the innate immunity in multicellular eukaryotes since methylated bacterial CpG strings are known to induce a proinflammatory response via toll-like receptors (Hemmi *et al*., 2000).

Whole-genome sequencing has revealed that genes in many bacterial genomes are encoded asymmetrically on the two DNA strands. Although a precise explanation for this phenomenon remains unclear, the occurrence of extreme gene asymmetry is a characteristic of the low-G+C bacteria (Rocha *et al*., 2000; Rocha, 2002). Comparative analysis of the *M. genitalium* and *M. pneumoniae* genomes suggests that maintenance of strand bias is important and under positive selection (Himmelreich *et al*., 1997). Among the genomic rearrangements that distinguish these genomes, the degree of strand bias remains essentially unaltered. The authors speculated that DNA replication and gene transcription might be coupled. The biased representation of genes on the leading strand may reduce the occurrence of collisions between the DNA replication and RNA transcription machinery. On the basis of computational methods, Rocha and Danchin (2003) concluded that the leading strand is biased in favor of essential genes.

Mycoplasma evolution has resulted in a reduction of noncoding DNA (intergenic) in these genomes (range 8–10%). The average noncoding DNA in low-G+C gram-positive and other prokaryotic genomes is higher (15% and 13% respectively) (Rogozin *et al*., 2002). By contrast, the average gene length in *Mycoplasma* genomes is larger when comparing orthologous genes from other bacteria (Wong and Houry, 2004; Oliver and Marin, 1996; Skovgaard *et al*., 2001). Oliver and Marin (1996) observed a correlation between the average gene length and a genome's G+C composition. It is known that the average length of conserved genes is longer than nonconserved genes (Lipman *et al*., 2002). While it is true that *Mycoplasma* genomes contain a larger portion of conserved genes and this may thereby account for the observed gene length increase, it may also highlight gene fusion events, wherein variable portions of two coding sequences merge into one, with expanded functional capacity. Two-thirds of the *M. genitalium* gene products are predicted to contain at least two domains (Teichmann *et al*., 1998; Teichmann *et al*., 2001). Putative intraoperonic spacers (distance between two unidirectional gene pairs) are the lowest among all bacteria in Mycoplasma genomes (Fukuda *et al*., 1999; Rogozin *et al*., 2002). Using the method of Ermolaeva *et al*. (2001), we found that the number of genes in presumed operons is higher in *Mycoplasmas* compared to other gram positive bacteria or obligatory intracellular pathogens (Buchnera omitted).

4. Repetitive DNA

The occurrence of repetitive DNA is conserved among the *Mycoplasmas* and constitutes a surprisingly high percentage of their genomes, ranging from 4.2% in *M. genitalium* to 28% in *M. mycoides*. *Mycoplasmas* are rich in both large and short repetitive elements. Large repeats consist mostly of insertion sequences, pseudogenes, and paralogous gene families, generally encoding membrane proteins. It is common for individual members of paralogous gene families to exhibit length polymorphisms in the encoded proteins. Their role as a reservoir of DNA for generating antigenic variation and/or epitope masking has been well documented (Razin *et al*., 1998; Rosengarten *et al*., 2001). Short repeats are found both in coding and noncoding sequences. In addition to their presence within genes encoding surface proteins, short repeats are frequently found within genes whose products participate in recombination, repair, and transcription. Karlin *et al*. (1997) and Rocha and Blanchard (2002) examined the frequency and location of short repeats within coding DNA sequences and provided novel insights into the role of these repeats on species fitness and evolution. Repetitive DNA in *Mycoplasma* genomes exhibits the same strong propensity to be oriented in the same localized direction as neighboring transcription units, suggesting that insertion of such elements in the opposite polarity are poorly tolerated.

5. Transcription in the mycoplasmas

Detailed information pertaining to transcription and its regulation in the *Mycoplasmas* is somewhat limited, however, some themes appear to be emerging. The

core structure of *Mycoplasma* promoters resembles those from other eubacteria and is composed of both −10 and −35 regions. The Pribnow Box (−10 region) sequence has been found to be conserved, whereas the −35 region appears more divergent (Weiner *et al*., 2000; Muto and Ushida, 2002). Other cis-acting regulatory sequences have not been reported and appear refractory to identification. An exception to this is that some *Mycoplasma* genomes contain CIRCE-like elements upstream of several heat-shock genes.

The *vlh*A gene family in *M. gallisepticum* encodes surface lipoprotein hemagglutinins. The expression of *vlh*A genes is controlled by a simple repeat sequence (GAA) located in upstream promoter regions (Markham *et al*., 1993; Glew *et al*., 1995; Glew *et al*., 2000). Alterations in the copy number of this tri-nucleotide repeat result in gene expression changes that reach a maximum when the repeat number is 12. In *M. hyorhinins*, the sequence between the −10 and −35 promoter region of variable lipoprotein (*vlp*) family genes contains of a polyA tract. Contraction and expansion of this spacer region alters promoter topology, thus affecting the efficiency of transcription of downstream genes (Citti and Wise, 1995). It is noteworthy that both of these examples involve the regulation of genes known to contribute to establishing antigenic variation (Baumberg *et al*., 1995). Antigenic variation is advantageous but the selection acts at the level of the population, not the cell, since the genotypic changes giving rise to antigenic variation occur at a very low frequency.

Transcription termination and its regulation are poorly understood in the *Mycoplasmas*. The gene encoding the terminator protein Rho is absent in the *Mycoplasmas*. Short inverted repeats following a gene, whether they fit a consensus terminator motif or not, are indicative of a potentially energetically favorable RNA hairpin loop formation capable of facilitating transcription termination. Neither *M. pneumoniae* nor *M. genitalium* appears to rely on hairpin formation for transcription termination, since such sequences are not present downstream of coding sequences at any appreciable frequency (Washio *et al*., 1998). The genes encoding the transcription termination factors, NusA, NusB, NusG, or GreA, are found alone or in various combinations among *Mycoplasma* genomes. These proteins are known to be involved in RNA transcript termination, antitermination, and/or attenuation.

All *Mycoplasmas* possess the gene encoding the vegetative sigma factor (σ^{70}) but appear devoid of genes encoding alternative sigma factors. The *Mycoplasmas* are essentially devoid of genes encoding cellular signal transduction (two-component regulators). One known exception to this lies in the genome of *M. penetrans*, which does possess such regulators (Sasaki *et al*., 2002). Protein sequence motifs commonly found in DNA binding and regulatory proteins are absent in some Mycoplasmas such as *M. genitalium,* and are otherwise scarce throughout this bacterial group (Table 1) (Razin *et al*.).

The lack of alternative sigma factors, and/or recognizable transcription factors in *M. genitalium*, is consistent with our experimental findings using DNA microarrays to monitor global gene expression. We have failed to identify any RNA abundance alteration for any gene in response to either serum starvation or nutrient deprivation, which entailed a comparison of transcript abundance in cells growing in mid-logarithmic growth to cells in stationary and late stationary phase growth. Recently,

Table 1 Frequency of selected Prosite motifs per unit (1 Mb) genome among various prokaryotes found through PEDANT

Motifs	"Pneumoniae clade" ($n = 5$)	All Mycoplasmas ($n = 8$)	Symbionts ($n = 3$)	Intracellular pathogens ($n = 3$)	Gram positives ($n = 10$)	Gram negatives ($n = 8$)	Other free-living bacteria ($n = 3$)	Prokaryotes other than *Mycoplasmas* ($n = 27$)
Proteases	9	9	6	6	7	6	5	6
Chaperones	12	10	6	6	5	7	3	7
Signal peptidases I + II	1	1	2	2	2	1	1	2
Helix-turn-helix + helix-loop-helix	3	2	9	9	12	12	10	9

an investigation of the *M. pneumoniae* transcriptional response to cold- and heat-shock by microarray analysis indicated that gene expression changes greater than twofold were not detected (Weiner *et al*., 2000).

It remains unclear whether transcriptional regulation and termination in the *Mycoplasmas* occurs through a novel mechanism, or whether their reduced repertoire of genes devoted to this function represents a minimal solution to regulating transcription. It is conceivable that as a genome approaches a minimum size and the fraction of genes encoding essential functions gets sufficiently high, the organism may benefit from dispensing with or relaxing transcriptional regulation. Although the experimental evidence is still limited, protein-level analysis and predictive computational analyses suggest that posttranslational modification occurs routinely. Wasinger *et al*. (2000) investigated the *M. genitalium* proteome during exponential and stationary phase growth and revealed a substantial, but not unusually high, occurrence of posttranslationally modified proteins. The abundance of proteins correlated with codon adaptation indices to a large extent (Wasinger *et al*., 2000). It has been established by Krebes *et al*. (1995) that one of the HMW proteins that plays a role in the formation of the cytadherence organelle undergoes phosphorylation. In *M. gallisepticum* and *M. fermentans*, another form of regulation involving differential cleavage of membrane proteins has been documented (Davis and Wise, 2002; Gorton and Geary, 1997). Specific proteolytic cleavage of cytadherence-related proteins in various *Mycoplasma* species such as *M. pneumoniae, M. hyopneumoniae*, and *M. gallisepticum* is known to occur (Layh-Schmitt and Harkenthal, 1999; Djordjevic *et al*., 2004; Steven J. Geary, University of Connecticut, US, personal communication).

The genes encoding the highly conserved chaperones such as the GroEL/GroES proteins involved in intermediate protein folding are found sporadically in *Mycoplasma* genomes. It has been speculated that a broadened functional capability of early step folding carried out by proteins such as Tig and DnaK, or by the proteases Clp and/or Lon may compensate for this genomic loss (Wong and Houry, 2004). For reasons that are not yet clear, *Mycoplasmas* have established a fundamental shift toward processes featuring protein processing (proteolysis, chaperonins, modification, degradation) as a primary means of protein-level regulation. In support of this model, we find an inverse correlation between the frequency of the helix-turn/loop-helix and chaperones or protease motifs per unit length of genomic DNA (Table 1).

6. A reductive evolution model: The loss of transcriptional regulators

Bacterial genes are typically regulated at the level of transcription by both positive (activators) and negative (repressors) regulators. In general, these regulators function through interaction with specific regulatory sequences within promoters. It is interesting to consider the consequences of losing a gene encoding for a transcriptional activator. The immediate impact to the regulator's target genes is clear. Transcription of target genes may either be dampened or extinguished

completely. The complete loss of expression of target genes can only be tolerated if the genes in question encode dispensable functions, and their loss in subsequent genomic reductions will also be tolerated. There are genes within all genomes, like the highly conserved gene, *rec*A, that while dispensable, provide selective value to the cell and its overall fitness. When genes of this type fail to be expressed, the cell is under greater pressure to find alternative means of reestablishing their expression. Given what we know about *M. genitalium* gene expression, it is possible that the cell was able to exploit its genomic "weakness" regarding poor recognition of transcriptional termination and turn it into an advantageous phenotype. Essential genes or dispensable genes of selective value, left with no means for expression, may be rescued by transcriptional read-through of termination signals from neighboring genes (Figure 1).

One model pertaining to genome evolution, referred to as the "selfish operon" model, starts by assuming that horizontal acquisition of chromosomal segments has been common during bacterial evolution (Lawrence and Roth, 1996). According to the selfish operon model, the inheritance of novel, advantageous, multigenic traits is more probable if all required genes can be acquired in a single DNA transfer event. Most genes within horizontally acquired DNA provide no selective advantage to the host and are lost over time. The sequential loss of useless DNA serves to increase the physical linkage of coselected genes. The increased physical linkage between coselected genes, in turn, increases the probability that they will be coacquired in

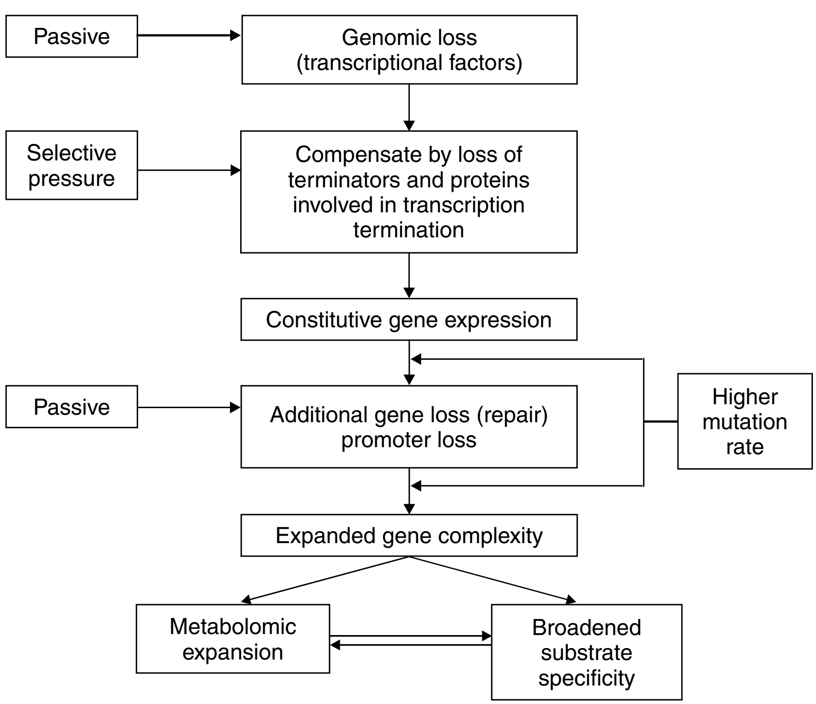


Figure 1 Reductive evolution model

a single event, and so on. Theoretically, this self-perpetuating process reaches an end point when coselected genes achieve maximal linkage (a coregulated operon). The fact that bacterial operons contain "like-minded" genes with respect to their function is a natural consequence of the fact they were coselected through a process of reductive evolution.

The scenario described above with regard to the loss of a transcriptional activator contains several intriguing analogies to the selfish operon model. First, the rescue of advantageous genes with no means of expression by read-through transcription also depends on the physical linkage of selectively advantageous genes to active promoters. Deletions that minimize the distance and boundaries separating transcription units containing essential genes may be under strong positive selection. Like the selfish operon model, a ratcheting effect is also implicit. As the selective pressure to reduce the efficiency of transcriptional terminator recognition is successfully accomplished, through the loss of proteins such as Rho, the probability is that subsequent loss of transcriptional regulators controlling essential genes will be tolerated. As regulators are lost, the cell becomes progressively more dependent on constitutive promoters. It is clear that the *Mycoplasmas* have become highly dependent on σ^{70}. If correct, this model implies that the loss of efficient transcriptional termination was perhaps a crucial step that enabled the substantial reductive evolution of the *Mycoplasmas.* By contrast to the selfish operon model, the forces acting to drive genome reduction act at short distances, perhaps limited by the processivity of RNA polymerases. The extant *M. genitalium* genome and its unique features support the basic tenants of this model, however, without specific knowledge of the temporal order of loss in this minimal genome, it is difficult to establish stronger support for this model of reductive evolution. We are currently pursuing experimental validation of various aspects of this model.

Visual inspection of the gene organization of the *M. genitalium* genome, while difficult to quantify, suggests that it is reminiscent of a viral genome. The asymmetry of gene orientation to be consistent with the direction of DNA replication is extreme and accounts for nearly 85% of the annotated genes. Operons are conspicuously long and compactly organized, with genes often overlapping by a few nucleotides. There are 35 regions on the chromosome-harboring genes oriented counter to the majority. This means that there are 35 places in the genome where neighboring genes are oriented in a head-to-head arrangement (3′ end to 3′ end). In 25 of these regions (Figure 2), the break point between properly oriented and misoriented regions is defined by a perfect head-to-head arrangement (little or no intervening sequence). At the opposite end of the region, we observe more heterogeneity in structure, suggesting that sequences upstream of genes are less able to adopt such compactness. Among the 10 cases (of 35) where perfect head-to-head gene arrangement do not exist, the break point in four of these lie immediately upstream of tRNAs or adhesin gene repetitive DNA. The remainder are flanked by rare voids in the annotated genome. These voids may represent genes undergoing decay and are perhaps destined to be lost from the genome. The tight packaging of genes in operons together with the differential behavior or 5′ and 3′ noncoding regions, the latter being more susceptible to removal of all intergenic sequence, may reflect an active mechanism used to vigilantly remove DNA sequences lacking functional significance.

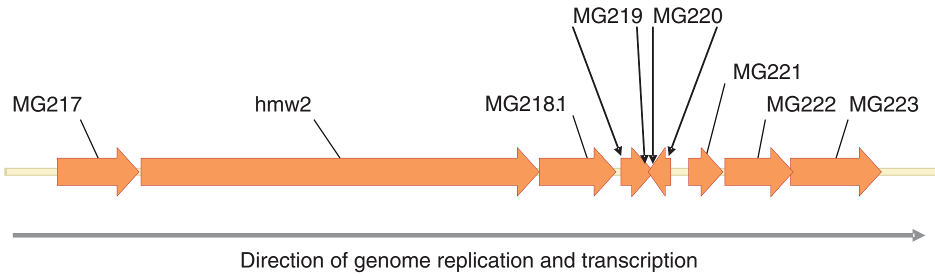


Figure 2 MG220 is in a nonmajority orientation. Arrows indicate perfect head-to-head arrangement of MG219 and MG220, and the more variable tail-to-tail arrangement of MG220 and MG221

7. Coping with a minimal genome

Genome reduction in the *Mycoplasmas* has generated a streamlined recombination and repair system. The lack of many DNA repair functions undoubtedly contributes to the high mutation rate observed in *Mycoplasma* genomes. *Mycoplasma* gene mutation rate is the highest among the prokaryotes and may be as high as 10^{-2} to 10^{-4} per generation (Razin *et al*., 1998; Ochman *et al*., 1999). This high mutation rate is significant and not necessarily a bad phenotype for a minimal cell to possess. It would be reasonable to assume that the Mycoplasmas were able to compensate for the loss of various biosynthetic capabilities through a modest expansion of their transporter repertoire. It has been reported that the number of transporters encoded by a bacterium is proportional to genome size. Mycoplasma genomes do not represent an exception (Fraser *et al*., 2000). While the number of proteins predicted to be involved in transport is relatively limited, *Mycoplasmas* may have alternatively evolved transport systems with broadened substrate specificity (Razin *et al*., 1998; Fraser *et al*., 2000; Maniloff, 1996; Saurin and Dassa, 1996). *Mycoplasma* transporter sequences show a significant divergence compared to orthologs in closely related genomes.

Similar evolution has been noted among metabolic enzymes. Cordwell *et al*. (1997) reported the unique activity of *M. genitalium* lactate dehydrogenase (LDH). Their analysis indicated that this 2-ketoacid dehydrogenase class enzyme may also confer malate dehydrogenase (MDH) activity. *Mycoplasmas* lack nucleoside phosphate transport. In this case, a solution relying on membrane-bound broad-affinity enzymes with 5′-nucleotidase activity that converts nontransportable nucleotides into transportable nucleosides (Pollack, 2002b). Purine nucleoside phosphorylase (PNP) has been shown to have equal activity for nucleobases and nucleosides (McElwain *et al*., 1988). The "patchwork" model of metabolic evolution suggests that the evolution of metabolic pathway has been driven by the expansion or alteration of enzyme substrate specificity. In other words, the novel preexisting enzymes with extended substrate specificity may be recruited to perform the same chemical conversions on a larger number of substrates (Jensen, 1976; Lazcano and Miller, 1999; Copley, 2000). It is interesting in this regard that the *M. pneumoniae* genome contains numerous tandem arrays of genes that define small paralogous families. As mentioned previously, these duplicated genes display

substantial length and sequence polymorphisms. It may well be that it is through mechanisms like these that such broadened substrate specificity is achieved.

The path of genomic minimalism undertaken by the *Mycoplasmas* comes with the price of driving the species evolution toward niche speciation. In this regard, *Mycoplasmas* rarely cross (phylogenetically distant) host species (Razin *et al*., 1998; Peterson and Fraser, 2001; Himmelreich *et al*., 1997). Reductive evolution represents one path among potentially countless others. It is possible that by gaining further insights into the forces causing reductive evolution and the mechanisms acting to achieve a minimal genome, we will be provided insights into a general mechanism that has acted in all bacterial genomes to continually ensure that genes of selective value are maintained and those that do not are lost. If genome evolution is in fact driven by horizontal transfer and acquisition of novel genes and functional capabilities, it seems likely that bacteria have developed efficient mechanisms for processing useless DNA for removal. It may be that the future evolutionary fate of the *Mycoplasmas* in this regard is not entirely in their own hands but, like that of any obligate intracellular pathogen, inextricably linked to the evolutionary fate of their hosts. On the other hand, we should not be too quick to underestimate the resourcefulness of the minimalist bacteria, as they do seem more than capable of finding ways to manage, even with a reduced set of tools at their disposal.

References

Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N and Mekalanos JJ (2002) A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 966–971.

Baseman JB, Lange M, Criscimagna NL, Giron JA and Thomas CA (1995) Interplay between mycoplasmas and host target cells. *Microbial Pathogenesis*, **19**, 105–116.

Baumberg S, Young JPW, Wellington EMH and Saunders JR (Eds.) (1995) *Population Genetics of Bacteria*, The Society for General Microbiology: Cambridge.

Citti C and Wise KS (1995) Mycoplasma hyorhinis vlp gene transcription: critical role in phase variation and expression of surface lipoproteins. *Molecular Microbiology*, **18**, 649–660.

Copley SD (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends in Biochemical Sciences*, **25**, 261–265.

Cordwell SJ, Basseal DJ, Pollack JD and Humphery-Smith I (1997) Malate/lactate dehydrogenase in mollicutes: evidence for a multienzyme protein. *Gene*, **195**, 113–120.

Dallo SF and Baseman JB (2000) Intracellular DNA replication and long-term survival of pathogenic mycoplasmas. *Microbial Pathogenesis*, **29**, 301–309.

Davis KL and Wise KS (2002) Site-specific proteolysis of the MALP-404 lipoprotein determines the release of a soluble selective lipoprotein-associated motif-containing fragment and alteration of the surface phenotype of Mycoplasma fermentans. *Infection and Immunity*, **70**, 1129–1135.

Djordjevic SP, Cordwell SJ, Djordjevic MA, Wilton J and Minion FC (2004) Proteolytic processing of the Mycoplasma hyopneumoniae cilium adhesin. *Infection and Immunity*, **72**, 2791–2802.

Dybvig K and Voelker LL (1996) Molecular biology of Mycoplasmas. *Annual Review of Microbiology*, **50**, 25–57.

Ermolaeva MD, White O and Salzberg SL (2001) Prediction of operons in microbial genomes. *Nucleic Acids Research*, **29**, 1216–1221.

Fraser CM, Eisen J, Fleischmann RD, Ketchum KA and Peterson S (2000) Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases*, **6**, 505–512.

Fukuda Y, Washio T and Tomita M (1999) Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. *Nucleic Acids Research*, **27**, 1847–1853.

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, *et al*. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of Bacteriology*, **185**, 5673–5684.

Glew MD, Browning GF, Markham PF and Walker ID (2000) pMGA phenotypic variation in Mycoplasma gallisepticum occurs in vivo and is mediated by trinucleotide repeat length variation. *Infection and Immunity*, **68**, 6027–6033.

Glew MD, Markham PF, Browning GF and Walker ID (1995) Expression studies on four members of the pMGA multigene family in Mycoplasma gallisepticum S6. *Microbiology*, **141**, 3005–3014.

Gorton TS and Geary SJ (1997) Antibody-mediated selection of a Mycoplasma gallisepticum phenotype expressing variable proteins. *FEMS Microbiology Letters*, **155**, 31–38.

Hemmi H, Takeuchi O, Kawai T, Kaisho T, Sato S, Sanjo H, Matsumoto M, Hoshino K, Wagner H, Takeda K, *et al*. (2000) A Toll-like receptor recognizes bacterial DNA. *Nature*, **408**, 740–745.

Himmelreich R, Plagens H, Hilbert H, Reiner B and Herrmann R (1997) Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. *Nucleic Acids Research*, **25**, 701–712.

Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO and Venter JC (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. *Science*, **286**, 2165–2169.

Hutchison CA III and Montague MG (2002) In *Molecular Biology and Pathogenicity of Mycoplasmas*, Razin S and Herrmann R (Eds.), Kluwer Academic/Plemum Publishers: New York, pp. 221–253.

Jardine O, Gough J, Chothia C and Teichmann SA (2002) Comparison of the small molecule metabolic enzymes of Escherichia coli and Saccharomyces cerevisiae. *Genome Research*, **12**, 916–929.

Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annual Review of Microbiology*, **30**, 409–425.

Ji Y, Zhang B, Van Horn SF, Warren P, Woodnutt G, Burnham MK and Rosenberg M (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, **293**, 2266–2269.

Karlin S, Mrazek J and Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, **179**, 3899–3913.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, *et al*. (2003) Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4678–4683.

Krebes KA, Dirksen LB and Krause DC (1995) Phosphorylation of Mycoplasma pneumoniae cytadherence-accessory proteins in cell extracts. *Journal of Bacteriology*, **177**, 4571–4574.

Lawrence JG and Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.

Layh-Schmitt G and Harkenthal M (1999) The 40- and 90-kDa membrane proteins (ORF6 gene product) of Mycoplasma pneumoniae are responsible for the tip structure formation and P1 (adhesin) association with the Triton shell. *FEMS Microbiology Letters*, **174**, 143–149.

Lazcano A and Miller SL (1999) On the origin of metabolic pathways. *Journal of Molecular Evolution*, **49**, 424–431.

Lipman DJ, Souvorov A, b Koonin EV, Panchenko AR and Tatusova TA (2002) The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, **2**, 20.

Lo, SC (1992) In *Mycoplasmas, Molecular Biology and Pathogenesis*, Maniloff J, McElhney RN, Finch LR and Baseman JB (Eds.), A.S.M.: Washington, pp. 525–545.

Lo SC, Dawson MS, Wong DM, Newton PB III, Sonoda MA, Engler WF, Wang RY, Shih JW, Alter HJ and Wear DJ (1989) Identification of Mycoplasma incognitus infection in patients with

AIDS: an immunohistochemical, in situ hybridization and ultrastructural study. *The American journal of tropical medicine and hygiene*, **41**, 601–616.

Maniloff J (1996) The minimal cell genome: "on being the right size". *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10004–10006.

Maniloff J, McElhney RN, Finch LR and Baseman JB (Eds.) (1992) *Mycoplasmas, Molecular Biology and Pathogenesis*, A.S.M.: Washington.

Markham PF, Glew MD, Whithear KG and Walker ID (1993) Molecular cloning of a member of the gene family that encodes pMGA, a hemagglutinin of Mycoplasma gallisepticum. *Infection and Immunity*, **61**, 903–909.

McElwain MC, Williams MV and Pollack JD (1988) Acholeplasma laidlawii B-PG9 adenine-specific purine nucleoside phosphorylase that accepts ribose-1-phosphate, deoxyribose-1-phosphate, and xylose-1-phosphate. *Journal of Bacteriology*, **170**, 564–567.

Much P, Winner F, Stipkovits L, Rosengarten R and Citti C (2002) Mycoplasma gallisepticum: Influence of cell invasiveness on the outcome of experimental infection in chickens. *FEMS Immunology and Medical Microbiology*, **34**, 181–186.

Mushegian AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10268–10273.

Muto A and Ushida C (Eds.) (2002) *Transcription and Translation*, Kluwer Academic/Plenum Publishers: Totowa.

Ochman H, Elwyn S and Moran NA (1999) Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 12638–12643.

Oliver JL and Marin A (1996) A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*, **43**, 216–223.

Peterson SN and Fraser CM (2001) The complexity of simplicity. *Genome Biology*, **2**, 1–7, COMMENT2002.

Pollack D (2002a) Central carbohydrate pathways: metabolic flexibility and extrarole of some "housekeeping" enzymes, In *Molecular Biology and Pathogenicity of Mycoplasmas*, Razin S and Herrmann R (Eds.) Kluwer Academic/Plenum Publishers: Totowa, pp. 163–199.

Pollack JD (2002b) The necessity of combining genomic and enzymatic data to infer metabolic function and pathways in the smallest bacteria: amino acid, purine and pyrimidine metabolism in Mollicutes. *Front Bioscience*, **7**, d1762–d1781.

Razin S and Herrmann R (Eds.) (2002) *Molecular Biology and Pathogenicity of Mycoplasmas*, Kluwer Academic/Plemum Publishers: New York.

Razin S, Yogev D and Naot Y (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiology and Molecular Biology Reviews: MMBR*, **62**, 1094–1156.

Rocha E (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends in Microbiology*, **10**, 393–395.

Rocha EP and Blanchard A (2002) Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. *Nucleic Acids Research*, **30**, 2031–2042.

Rocha EP and Danchin A (2002) Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, **18**, 291–294.

Rocha EP and Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genetics*, **34**, 377–378.

Rocha EP, Guerdoux-Jamet P, Moszer I, Viari A and Danchin A (2000) Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *Journal of Biotechnology*, **78**, 209–219.

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J and Koonin EV (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Research*, **30**, 4264–4271.

Rosengarten R, Citti C, Much P, Spergser J, Droesse M and Hewicker-Trautwein M (2001) The changing image of mycoplasmas: from innocent bystanders to emerging and reemerging pathogens in human and animal diseases. *Contributions to Microbiology*, **8**, 166–185.

Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, Furuya K, Yoshino C, Horino A, Shiba T, Sasaki T, *et al*. (2002) The complete genomic sequence of Mycoplasma penetrans, an intracellular bacterial pathogen in humans. *Nucleic Acids Research*, **30**, 5293–5300.

Saurin W and Dassa E (1996) In the search of Mycoplasma genitalium lost substrate binding proteins: sequence diveregence could be the result of a broader substrate specificity. MicroCorrespondence. *Molecular Microbiology*, **22**, 389–391.

Skovgaard M, Jensen LJ, Brunak S, Ussery D and Krogh A (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, **17**, 425–428.

Smith DG, Russell WC, Ingledew WJ and Thirkell D (1993) Hydrolysis of urea by Ureaplasma urealyticum generates a transmembrane potential with resultant ATP synthesis. *Journal of Bacteriology*, **175**, 3253–3258.

Teichmann SA, Murzin AG and Chothia C (2001) Determination of protein function, evolution and interactions by structural genomics. *Current Opinion In Structural Biology*, **11**, 354–363.

Teichmann SA, Park J and Chothia C (1998) Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14658–14663.

Washio T, Sasayama J and Tomita M (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Research*, **26**, 5456–5463.

Wasinger VC, Pollack JD and Humphery-Smith I (2000) The proteome of Mycoplasma genitalium. Chaps-soluble component. *European Journal of Biochemistry / FEBS*, **267**, 1571–1582.

Weiner J III, Herrmann R and Browning GF (2000) Transcription in Mycoplasma pneumoniae. *Nucleic Acids Research*, **28**, 4488–4496.

Winner F, Rosengarten R and Citti C (2000) In vitro cell invasion of Mycoplasma gallisepticum. *Infection and Immunity*, **68**, 4238–4244.

Woese CR, Maniloff J and Zablen LB (1980) Phylogenetic analysis of the mycoplasmas. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 494–498.

Wong P and Houry WA (2004) Chaperone networks in bacteria: analysis of protein homeostasis in minimal cells. *Journal of Structural Biology*, **146**, 79–89.

Zou N and Dybvig K (2002) In *Molecular Biology and Pathogenicity of Mycoplasmas*, Razin S and Herrmann R (Eds.), Kluwer Academic/Plemum Publishers: New York, pp. 303–321.

Specialist Review

The nuclear genome of apicomplexan parasites

James W. Ajioka and Elizabeth T. Brooke-Powell
University of Cambridge, Cambridge, UK

Kiew-Lian Wan
Universiti Kebangsaan Malaysia, Bangi, Selangor DE, Malaysia

1. Introduction and background

The phylum Apicomplexa represents a unique opportunity to explore how evolution and natural selection have generated what is arguably the most successful and important group of eukaryotic parasitic pathogens, accounting for morbidity and mortality figures in the hundreds of millions per annum. The phylum consists of a highly diverse group of unicellular organisms that are obligate intracellular parasites of metazoans. It is represented by about 4600 described species, with the possibility of an order of magnitude more remaining undiscovered (Ellis *et al*., 1998). The current lack of complete molecular data for representative members of the Apicomplexa preclude a detailed and reliable phylogenetic reconstruction within the phylum, but recent studies indicate that distinct lineages will emerge (see for example Leander *et al*., 2003; Aravind *et al*., 2003). Although particular lineage rankings and relationships are currently the subject of some debate, the molecular data are generally consistent with the traditional taxonomic classifications based on morphology and life cycles such that the relationships between species of medical and veterinary importance are sufficiently robust for most comparative purposes (see Table 3).

The defining characteristic of the Apicomplexa, the "apical complex", is found on the parasites' asexual forms and mediates attachment and invasion of host cells. It consists of microtubule-based features known as the conoid, polar rings, and subpellicular microtubules. Through this structure, the associated secretory organelles known as the rhoptries and micronemes are able to release their contents (Blackman and Bannister, 2001; Dubey *et al*., 1998). Other features likely to be shared by most of the phylum members are a single mitochondrion and the plastid-like organelle known as the "apicoplast", acquired via a secondary endosymbiotic event between a eukaryotic cell and a photosynthetic algae (known exceptions are *Cryptosporida* spp.; Zhu *et al*., 2000; Abrahamsen *et al*., 2004), leaving a single nucleus as the sole chromosomal compartment.

The apicomplexan nuclear genome is a mosaic likely consisting of nuclear genes from both secondary endosymbionts, genes originating from the mitochondrial genome and from the apicoplast genome. Several representative species within the phylum have (nuclear) genomic sequencing and related projects completed or under way, so comparative genetics, genomics, and downstream methods should provide a wealth of information toward both a basic biological understanding of these organisms and ways to combat the diseases they cause (see Tables 1 and 2). Since the representative species are quite distantly related compared to mammalian model organisms, a thoughtful notion of homology is the primary caveat emptor for using these data and guiding the interplay between computational and experimental studies (see, for example, Barta, 1997).

Although the sexual cycle defines the primary host (and hence the range of potential secondary host(s)), asexual reproduction of apicomplexan parasites is the main contributor to human and animal disease. The life cycle of a particular apicomplexan may involve a wide variety of hosts, but sexual reproduction is restricted to a single vertebrate species/species group and cognate arthropod host as appropriate (see Table 3, Figure 1). Apicomplexans exist as nominally haploid (N) cells for the vast majority of their life cycle and bona fide diploid (2N) cells are only associated with sexual reproduction (see Figure 1). Diploid zygote formation followed by a conventional meiosis and sporozoite formation has been established in *Plasmodium* spp., *Eimeria* spp., and *Toxoplasma gondii* (Walliker *et al*., 1976; Walliker *et al*., 1975; Pfefferkorn and Pfefferkorn, 1980; Jeffers, 1976), and is therefore inferred for the other species in the phylum.

Shared (and likely homologous) processes such as schizogony, a process by which several rounds of nuclear and organellar replication occur before the re-formation of individual parasite cells, and gamete formation may be most easily studied in a particular species. So, comparative inference will be a powerful tool for understanding these fundamental properties across the phylum. The control of asexual reproduction appears to be critical for parasite development/differentiation, where basic properties of the cell cycle may be common features in developmental decisions as disparate as gametogenesis, sporogenesis, and tissue cyst formation. Gametogenesis in *Plasmodium* spp. may be linked to the control of asexual replication as it has been viewed as “opting out” of the cell cycle (Dyer and Day, 2000). *Eimeria* spp. produce large schizonts with a set number of asexual replication cycles prior to the onset of the sexual cycle and sporogeny (see, for example, McDonald and Rose, 1987; McDonald and Shirley, 1987). Infection with *T. gondii* sporozoites results in a transformation to tachyzoites that have a limit of about 20 asexual divisions before they develop into the cyst form bradyzoites (Jerome *et al*., 1998). A detailed analysis of asexual reproduction and cell cycle control will likely shed considerable light on mechanisms of development in apicomplexans.

2. Nuclear genomes and genetics

Defining the sexual cycle in the apicomplexans has allowed genetic linkage analysis of the nuclear genome, underpinned population genetic studies, and provides a foundation for genomic analysis (see Tables 1 and 2). The development of genetic

Table 1 Comparative genome statistics

Species	Genome size (Mb)	Chromosome number	Range of chromosome size (Mb)	GC content (%)	Host species	References
Babesia bovis	9.4	4	1.4–3.2	44[a]	Cattle	Jones *et al*. (1997)
Babesia canis canis	14.5	5	0.8–6.0	45–50[b]	Dog	Depoix *et al*. (2002)
Babesia canis rossi	16	5	0.9–6.0	45–50[b]	Dog	Depoix *et al*. (2002)
Cryptosporidium parvum	9.6–10.4	8	1.04–1.54	30–40	Human	Abrahamsen *et al*. (2004)
Eimeria tenella	60	14	1–>6	53	Chicken	Piper *et al*. (1998) Blunt *et al*. (1997) Shirley (1994) **http://www.sanger.ac.uk/Projects/E_tenella/**
Neospora caninum	60[a]	No data	No data	No data	Dog/cattle[c]	
Plasmodium berghei	25–27	14	0.6–3.8	24[a]	Mouse/mosquito	**http://www.sanger.ac.uk/Projects/P_berghei/**
Plasmodium chabaudi	25–30	14	0.7–3.0	20	Human/mosquito	**http://www.ncbi.nlm.nih.gov/projects/Malaria/Rodent/chabaudi.html**
Plasmodium falciparum	22.9	14	0.7–3.4	20	Human/mosquito	Gardner *et al*. (2002)
						Carlton *et al*. (1999)
Plasmodium vivax	30	14	1.1–3.4	45	Human/mosquito	Carlton *et al*. (1999)
Plasmodium yoelii	23	14	No data	32	Human/mosquito	Carlton *et al*. (2002); Carlton *et al*. (1999)
Sarcocystis neurona	No data	No data	No data	No data	Opossum/horse[c]	
Theileria annulata	10	4	1.8–4.5	32[a]	Cattle/tick	**http://www.sanger.ac.uk/Projects/T_annulata/**
Theileria parva	10	4	2.2–3.2	31	Cattle/tick	Nene *et al*. (2000) Allsopp and Allsopp (1988)
Toxoplasma gondii	65	13–14	1.8–7.4	53	Cat	**http://www.toxodb.org/ToxoDB.shtm** **http://www.toxomap.wustl.edu/linkage_map.html** Sibley and Boothroyd (1992a)

[a] A. Pain, personal communication.
[b] B. Carcy, personal communication.
[c] D. Howe, personal communication.

Table 2 Genome weblinks

Species	Data type(s)	URL address
Babesia bovis	EST	**http://www.sanger.ac.uk/Projects/B_bovis/**
Cryptosporidium parvum	Multiple	**http://CryptoDB.org**
	Genomic	**http://www.cbc.umn.edu/ResearchProjects/AGAC/Cp/index.htm**
		http://www.parvum.mic.vcu.edu/
	GSS	**http://medsfgh.ucsf.edu/id/CpDemoProj/**
Eimeria tenella	Genomic	**http://www.sanger.ac.uk/Projects/E_tenella/**
	EST	**http://www.cbil.upenn.edu/paradbs-servlet/index.html**
		http://www.genome.wustl.edu/est/index.php?eimeria=1
Neospora caninum	EST	**http://www.cbil.upenn.edu/paradbs-servlet/index.html**
		http://www.genome.wustl.edu/est/index.php?neospora=1
Plasmodium	Multiple	**http://plasmoDB.org**
Species all		**http://www.ncbi.nlm.nih.gov/projects/Malaria/**
P. berghei	Genomic	**http://www.sanger.ac.uk/Projects/P_berghei**
		http://parasite.vetmed.ufl.edu
		http://www.tigr.org/tdb/tgi/pbgi
		http://www.GeneDB.org
P. chabaudi	Genomic	**http://www.sanger.ac.uk/Projects/P_chabaudi**
		http://www.GeneDB.org
P. falciparum	Genomic	**http://sequence-www.stanford.edu/group/malaria/**
		http://www.tigr.org/tdb/e2k1/pfa1/
		http://www.sanger.ac.uk/Projects/P_falciparum/
		http://www.GeneDB.org
	EST	**http://fullmal.ims.u-tokyo.ac.jp/**
		http://parasite.vetmed.ufl.edu/falc.htm
		http://www.cbil.upenn.edu/paradbs-servlet/index.html
	GSS	**http://parasite.vetmed.ufl.edu/**
	Microsatellite map	**http://www.ncbi.nih.gov/projects/Malaria/Mapsmarkers/PfGMap/pfgmap.html**
	Optical map	**http://www.lmcg.wisc.edu/research/research.html#plasmodium**
	Oligonucleotide array	**http://malaria.ucsf.edu/**
	Affymetrix array	**http://www.scripps.edu/cb/winzeler/malariatext.html**
P. knowlesi	Genomic	**http://www.sanger.ac.uk/Projects/P_knowlesi/**
P. reichenowi	Genomic	**http://www.sanger.ac.uk/Projects/P_reichenowi/**
P. vivax	Genomic	**http://www.tigr.org/tdb/e2k1/pva1/intro.shtml**
		http://parasite.vetmed.ufl.edu
	GSS	**http://parasite.vetmed.ufl.edu/viva.htm**
	YAC	**http://www.sanger.ac.uk/Projects/P_vivax/**
P. yoelii	Genomic	**http://www.tigr.org/tdb/e2k1/pya1/**
	EST	**http://www.tigr.org/tdb/tgi/pygi/**
Sarcocystis neurona	EST	**http://www.cbil.upenn.edu/paradbs-servlet/index.html**
		http://www.genome.wustl.edu/est/index.php?sarcocystis=1
Theileria annulata	Genomic	**http://www.sanger.ac.uk/Projects/T_annulata/**
Theileria parva	Genomic	**http://www.tigr.org/tdb/e2k1/tpa1/**
Toxoplasma gondii	Multiple	**http://ToxoDB.org**
	Genomic	**http://www.tigr.org/tdb/t_gondii/**
	EST	**http://www.cbil.upenn.edu/paradbs-servlet/index.html**
		http://www.genome.wustl.edu/est/index.php?toxoplasma=1
	BAC-end	**http://www.sanger.ac.uk/Projects/T_gondii/**
	Genome map	**http://www.toxomap.wustl.edu**

Table 3 Comparative life cycles

Subclass (suborder)	Genus	Meiosis & sporulation	Host tissue/ cell infected	Asexual tissue cycle	Trophozoite & schizont/ merozoite cycle	Gametogenesis	Microgamete exflagella-tion	Gametic fusion/ zygote formation
Coccidia (Haemospo-rina)	*Plasmodium*	Mosquito gut wall	Human hepatocyte, erythrocyte	None	Human hepatocyte, erythrocyte	Human erythrocyte	Mosquito gut	Mosquito gut
Coccidia (Eimeriina)	*Eimeria*	Avian feces environment	Avian gut epithelia	None	Avian gut epithelia	Avian gut epithelia	Avian gut	Avian gut epithelia
Coccida (Eimeriina)	*Toxoplasma*	Cat feces environment	Cat gut epithelia	Any nucleated cell; cat or secondary host	Cat gut epithelia	Cat gut epithelia	Cat gut	Cat gut epithelia
Coccida (Eimeriina)[a]	*Cryptosporidia*	Mammalian gut epithelia cell membrane	Mammalian gut epithelia cell membrane	None	Mammalian gut epithelia	Mammalian gut lumen	None	Mammalian gut lumen
Piroplasmia	*Theileria*	Tick salivary gland	Mammalian lymphocyte	None	Macroschizonts make infected lymphocytes divide and ultimately become microschizonts; no merozoite reinfection of new cells	Products of microschizonts invade mammalian erythrocytes	Tick gut lumen	Tick gut lumen

[a]Recent phylogenetic analyses suggest that *Cryptosporidia* may be a deep branching lineage and not a Coccidian (see, for example, Leander *et al*., 2003).

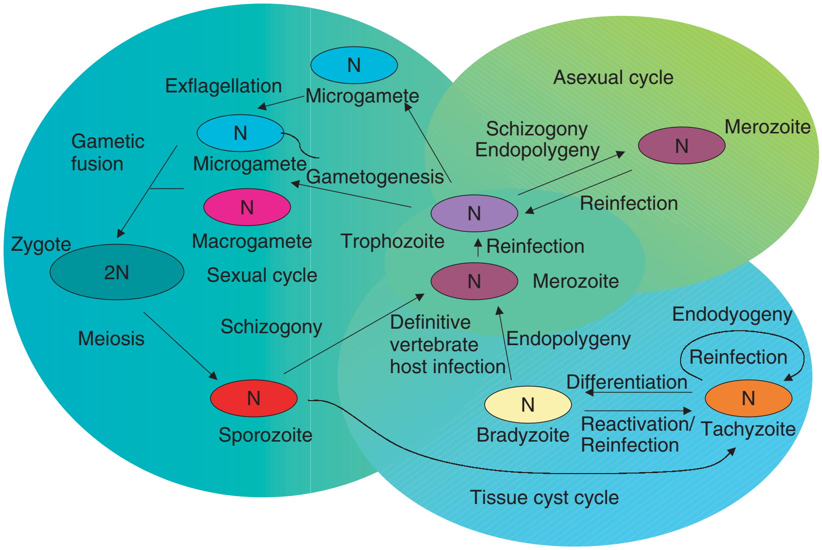


Figure 1 Apicomplexan reproduction. Modes of Apicomplexan reproduction vary, but all known species maintain the sexual and asexual cycles with tissue cyst cycle confined to *T. gondii* and its close relatives. The sexual cycle determines the definitive vertebrate host. Sporozoites infect the definitive vertebrate host, replicate via schizogony. In Coccidians, the newly formed merozoites reinfect host cells and enter the asexual cycle, continuing to amplify in numbers via schizogony/reinfection cycle. In contrast, *Theileria* spp. schizogony do not use host cell lysis/reinfection for amplification, but stimulate infected lymphocyte cell division to increase the number of host cells within which to replicate. In a species-specific manner, after several rounds of asexual reproduction, a proportion of the resulting merozoites go through gametogenesis to form micro- and macrogametocytes. In *Plasmodium* spp. the gametocytes are taken up by the mosquito in a blood meal from the mammalian host. *Theileria* spp. require the products of the schizont to infect erythrocytes to facilitate gametogenesis. The gametes may undergo further changes such as exflagellation of microgametocytes before zygote formation, meiosis, and sporulation. In *Cryptosporidium* spp., exflagellation does not occur but rather the differentiated microgamont goes directly into zygote formation. A small number of Coccidians use tissue cyst formation to exploit carnivorous consumption of bradyzoites for reinfection of the definitive host, and *T. gondii* further uses this mechanism to escape sexual reproduction altogether by transmission between secondary hosts. N refers to haploidy and 2N refers to diploidy

markers has allowed investigations into the inheritance of important disease-related phenotypes such as drug resistance and virulence traits. These genetic markers have also supported studies to understand the population structure and geographical distribution of virulence-associated alleles.

2.1. *Plasmodium spp*

Plasmodium spp. life cycle requires both a primary vertebrate host and a mosquito vector. The major mode of asexual reproduction, schizogony, occurs first in vertebrate host hepatocytes and then in erythrocytes. The sexual phase begins with gametogenesis in the vertebrate host followed by uptake by the mosquito, where zygote formation, meiosis, and the generation of infective sporozoites occur (see Figure 1).

Although the primary focus of plasmodium genetics has been the human malarial parasite *Plasmodium falciparum*, the first demonstrations of genetic recombination were crosses in the rodent species *P. yoelli* and *P. chaubaudi* (Walliker *et al*., 1971; Walliker *et al*., 1975; Walliker *et al*., 1976). Deliberate mixtures of genetically marked clones fed to mosquitoes revealed inheritance patterns consistent with a conventional chromosomal meiosis in the nuclear genome. The proportions of progeny, those from parental selfing and cross-fertilization, were within Hardy–Weinberg expectations.

The development of genetic markers and analysis of crosses has been used to investigate *P. falciparum* virulence-related traits such as cytoadherance, host cell invasion, transmission (Walliker *et al*., 1987; Vaidya *et al*., 1995; Guinet *et al*., 1996; Day *et al*., 1993; Wellems *et al*., 1987), and drug resistance as both heritable phenotypes and as a tool to track resistance phenotypes geographically (Peterson *et al*., 1988; Wellems *et al*., 1990; Cowman and Karcz, 1993; Goldberg *et al*., 1997; Su *et al*., 1997). Since *P. falciparum* genetic crosses are difficult and generally have been used in multiple studies, one such cross between HB3 and Dd2 parents was extended and used to produce a genetic map, defining recombination parameters (Su *et al*., 1999). Thirty-five progeny were analyzed with 901 RFLP and microsatellite markers showing 14 linkage groups totaling 1556 centimorgans (cM). The average length of the 326 mapped segments is estimated to be about 80 kb, giving an average map unit size of 17 kb per cM. Moreover, this figure varies little across and between chromosomes, indicating that the crossing-over frequency is not only relatively high, but uniform as well. However, a separate analysis of this cross indicated that the subtelomeric *var* genes showed a higher than expected recombination frequency due to gene conversion events between heterologous chromosomes (Freitas-Junior *et al*., 2000). This result provides a mechanism that explains the high diversity within the *var* gene family. Most of the markers showed approximately equal numbers of parental alleles but a few linkage groups displayed uniparental bias. Linkage groups on chromosome 2 and terminal regions of chromosomes 9 and 13 showed excess Dd2 markers compared to linkage groups on chromosomes 3 and 8 that maintained an HB3 bias (see **http://www.ncbi.nlm.nih.gov/projects/Malaria/Mapsmarkers/PfSegData/segdata.html**). Overall, the underlying mechanisms responsible for inheritance bias are unclear but, the *var* gene conversion/recombination events can produce inheritance bias, as evidenced by 10 of 13 progeny inheriting HB3var10-1 via biased gene conversion onto chromosome 9 (Freitas-Junior *et al*., 2000). Uniparental bias in inheritance of organelles is also evident from the previous studies of HB3xDd2 recombinant progeny that showed that the Dd2 parent did not form microgametocytes efficiently (Vaidya *et al*., 1993). The excess of Dd2 macrogametocytes (hence cytoplasm and organelles) resulted in recombinant progeny with exclusively Dd2 cytoplasm and organelles. Bias in cytoplasmic inheritance was also observed in an HD3x3D7 cross, but bias in gametocyte production could not explain this result, so unidirectional gametocyte incompatibility was raised as an explanation (Vaidya *et al*., 1993). Although the sexual cycle is obligatory in *Plasmodium* spp., selfing probably occurs most often, so investigations into both recombination and cytoplasmic inheritance biases may reveal genetic incompatibilities that reflect natural

selection for particular combinations of alleles and organelles, as well as elucidating mechanisms underlying biased gene conversion in terminal chromosomal regions.

The genomic sequence is effectively complete for *P. falciparum* (clone 3D7) and *P. yoelli* (17XNL clone) (Gardner *et al*., 2002; Carlton *et al*., 2002). For *P. falciparum* 3D7, sequence from chromosomal shotgun and selected yeast artificial chromosome genomic clones were assembled, and sequence tag sites, microsatellite markers, HAPPY mapping, and optical restriction maps were used to place, join, orient, and confirm contigs. Within the Apicomplexa, *Plasmodium* species are highly represented in genomic sequencing efforts (see Table 2), and some excellent comparative analyses have been published (see for example, (Aravind *et al*., 2003)). Consistent with the genetic map, the nuclear genome consists of 22.9 Mb distributed across 14 chromosomes ranging in size from about 0.7 to 3.4 Mb. The overall A+T content is just over 80% where regions with lower A+T content generally mark protein-coding sequences. Despite a genome nearly twice the size of the fission yeast *Schizosaccharomyces pombe*, a similar number of genes (~5300) and proportion of genes containing introns (54%) were identified. The average coding sequence length is 2.6 kb, which is larger than that of other unicellular organisms. A recent analysis of *P. falciparum* genes showed that this is partly explained by enrichment with stretches that encode low-complexity regions composed of homopolymeric runs of 10 to >100 asparagine residues, generally between predicted globular domains (Aravind *et al*., 2003). An analysis of GTPases with known structure showed that the insertion of the homopolymeric runs mapped to loops between secondary structural elements and distant from functional P-loop and Walker B domains. The function of the low-complexity inserts remains unclear considering their size, frequency, and interspecific differences. However, their presence in several species of *Plasmodium* suggests that they may be maintained by natural selection, raising the intriguing possibility that common cross-reactive epitopes such as asparagine-rich peptides may impair a useful host immunological response (Anders *et al*., 1986; Aravind *et al*., 2003).

Following the completion of the *P. falciparum* genome, microarray analysis has come to complement classical cloning and sequence comparison studies for the analysis of gene expression. Initial studies focused on expression changes throughout the life cycle of *P. falciparum* (Bozdech *et al*., 2003; Le Roch *et al*., 2003). Asexual intraerythrocytic development of *P. falciparum* monitored over 46 time points showed that approximately 60% of the genome is transcriptionally active during schizogony (Bozdech *et al*., 2003). Transcription appears to be on an "as-needed" basis, gradually changing over time as shown by the peak in expression of all factors associated with DNA replication and synthesis around 30 h, early schizogony. This peak coincides with the transition out of the ring stage, a low point of gene expression, and the initiation of concerted replication of the genome (Bozdech *et al*., 2003).

2.2. *Toxoplasma gondii*

T. gondii has the most promiscuous life cycle of the major disease-causing apicomplexans. Along with the sexual and asexual cycles, *T. gondii* can replicate

in virtually any nucleated cell from any warm-blooded animal and form tissue cysts that can be passed between secondary hosts via carnivory (Hill and Dubey, 2002; see Figure 1). *T. gondii* and its close relatives usually enter the sexual cycle via consumption of encysted bradyzoites and require asexual reproduction via a specialized schizogony called *endopolygeny*. The tissue cyst cycle, unique to *T. gondii* and its close relatives, can be initiated by the ingestion of either a sporozoite or bradyzoite cyst. These parasites reactivate into the rapidly dividing tachyzoite form for further reinfection to virtually any cell type. Tachyzoites multiply via a binary dividing process known as *endodyogeny*, generating large numbers of parasites, and causing acute disease through cellular destruction and host inflammatory response. Asymptomatic chronic infection follows by the differentiation of tachyzoites into bradyzoites that form a tissue cyst that may remain indefinitely in host tissue (Dubey *et al*., 1998).

Several lines of evidence suggest that the cell cycle of these various asexual reproductive forms differ from yeast and mammalian cells. Cell cycle analysis of asexual replication of synchronized *T. gondii* tachyzoites showed that G1 (vast majority of cells = N) occupies about 60% of the cell cycle where the nuclear DNA replication (S-phase) is biphasic (minor fraction of cells in early $S = 1 - 1.7N$; major fraction of cells late $S = 1.8N$) and accounts for about 30% of the cell cycle (Radke *et al*., 2001). S-phase is quickly followed by mitosis such that G2 is very short or nonexistent. Moreover, in stained *P. falciparum* and *Theileria* nuclei in schizogony, a similarly short/nonexistent G2 may be inferred by the lack of 2N nuclei (Irvin *et al*., 1982; Jacobberger *et al*., 1992). It is speculated that the relatively lengthy late S-phase may replace G2 as a premitotic checkpoint for replication in apicomplexans (Radke *et al*., 2001).

The genetic linkage analysis of *T. gondii* represents the most systematic attempt amongst apicomplexan species to develop a genetic map as a general framework for understanding the inheritance of virulence-related phenotypes and for positional cloning of candidate loci. After showing that drug resistance markers on two independent strains of *T. gondii* segregate with Mendelian inheritance in a genetic cross (Pfefferkorn and Pfefferkorn, 1980), a similar cross was used to generate a genetic map using randomly generated restriction fragment length polymorphic markers (RFLP; Sibley *et al*., 1992). Recent refined genetic mapping, in combination with physical mapping/genomic sequencing, extends the number of chromosomes to a total of 13–14 varying in size from 1.8 to 7.4 Mb (D. Sibley, personal communication; **http://toxomap.wustl.edu**).

In contrast to *Plasmodium* spp., the sexual cycle is not obligatory, a property unique to *T. gondii*, which allows asexual reproduction through a tissue cyst/carnivorous cycle between secondary hosts (see Figure 1). This property may explain why studies on the population structure of *T. gondii* show that there are three main clonal lineages (Sibley and Boothroyd, 1992b; Howe and Sibley, 1995) into which the great majority of isolates can be placed (Su *et al*., 2003). These lineages are closely related and appear to be progeny from a single genetic cross (Grigg *et al*., 2001). The vast majority of single nucleotide polymorphisms (SNP) can be classified as one of two parental alleles where all of the isolates within a clonal lineage will share the same allele at every SNP position. The remaining SNPs are so rare that an analysis of neutral SNPs (e.g., those found in introns) suggests

that a rapid global expansion of *T. gondii* occurred very recently, probably within the last 10 000 years (Su *et al*., 2003). Although the clonal lineages are very closely related, there are phenotypic differences including virulence in the mouse model and the ability to form tissue cysts (Sibley and Boothroyd, 1992b; Hill and Dubey, 2002). A cross between the virulent Type I GT-1 ($FUDR^R$; murine $LD_{100} = 1$) and the nonvirulent Type III CTG ($AraA^R$, SNF^R; murine $LD_{100} = 10^4$) strains revealed a major locus for murine virulence on chromosome VII and a minor locus on chromosome IV (Su *et al*., 2002). The genomic sequence assembly across these regions should facilitate the identification of candidate loci for the murine virulence phenotype.

In contrast to the relatively compact A+T-rich *Plasmodium* genomes, the *T. gondii* genome is currently estimated to be 65 Mb and 53% G+C (I. Paulsen, M. Berriman, D. Roos, and J. Ajioka, personal observation). The difference in genome size may be partly explained by more repetitive DNA, an increased number and size of introns combined with a somewhat lower gene density. Of these possibilities, only repetitive DNA has been systematically studied. The gene(s) encoding the B1 antigen are arranged as a single tandem array of 35 elements (Burg *et al*., 1988; Burg *et al*., 1989). Dispersed repetitive elements include the mitochondria-like REP family (Ossorio *et al*., 1991), simple 2–6 nucleotide microsatellite repeats (Ajzenberg *et al*., 2002; Blackston *et al*., 2001), and repeats that are likely to be subtelomeric as represented in the ABTg collection (Matrajt *et al*., 1999). What appear to be bona fide telomere repeats, sharing the same sequence motif with *Plasmodium* spp. and *Eimeria* spp., TTTAGGG, have also been identified (M. Berriman and J. Ajioka, personal observation).

Although gene expression in *T. gondii* appears to be transcriptionally regulated, conventional cis-acting eukaryotic promoters such as the TATA box or SP1 motif are notably absent (see, for example, Soldati and Boothroyd, 1995; Mercier *et al*., 1996; Nakaar *et al*., 1998; M. Berriman and J. Ajioka, personal observation). In searches for putative promoter elements, upstream sequence analysis of several genes has revealed short repeats with a highly conserved consensus T/AGAGACG heptanucleotide core element that qualitatively act like SP1 elements (Soldati and Boothroyd, 1995; Mercier *et al*., 1996; Nakaar *et al*., 1998). As with *Plasmodium* spp., general mechanisms controlling gene expression remain elusive and large-scale analysis of gene expression should define expression patterns that may guide further investigations.

While the *T. gondii* genome has been largely sequenced, there has been slow progress in gene finding and annotation. Despite this, researchers are using printed cDNA microarrays to address fundamental questions in the development of the parasite and its interaction with a host (Blader *et al*., 2001; Cleary *et al*., 2002). The present generation of microarrays available is made from variant and stage-specific ESTs relating to the asexual tissue cyst cycle unique to *T. gondii* and its close relatives. Several published studies have used a cDNA microarray generated from the in vivo bradyzoite EST library (Manger *et al*., 1998a). The original study characterized in vitro expression changes during high pH shock of infected tachyzoites, a treatment previously shown to result in tachyzoite-to-bradyzoite differentiation (Soete *et al*., 1994; Weiss *et al*., 1995). The majority of genes showed little expression difference between bradyzoites and tachyzoites but

revealed a new surface antigen, regulatory and metabolic enzymes, and secretory organelle proteins that showed clear expression differences (Cleary *et al*., 2002). To extend this study, mutants were selected by both chemical and insertional mutagenesis to lack in vitro differentiation (Matrajt *et al*., 2002; Singh *et al*., 2002). The microarray analysis of all mutants showed decreased expression of those genes identified in the original study and generally showed tachyzoite-like expression profiles when cultured under "bradyzoite conditions". A hierarchy of genes associated with bradyzoite formation was identified, suggesting a "cascade" model for controlling transcript levels (Singh *et al*., 2002).

2.3. *Eimeria spp*

In contrast to *T. gondii*, *Eimeria* spp. have a simple direct life cycle through a single host that is composed of sequential phases of asexual reproduction (schizogony) in the intestinal tract followed by a final sexual phase where gametogenesis, zygote formation, and meiosis result in the fecal shedding of a vast number of highly infective sporozoites (Hammond, 1982; see Figure 1). In avian *Eimeria* spp., several lines have been selected to complete the life cycle in a fewer number of asexual cycles, hence limiting the number of parasites produced and damage to the gut (Jeffers, 1975). In an effort to understand the genetic basis for the "precocious" development phenotype, a genetic map was established in *E. tenella* using a variety of DNA-based markers (Shirley and Harvey, 1996; Shirley and Harvey, 2000). In a cross between a "precocious" parent and a drug-resistant parent, 443 markers were mapped using 22 recombinant progeny, resulting in 16 linkage groups that defined 12 chromosomes. A linkage group on chromosome 2 showed significant association with precocious development and a linkage group on chromosome 1 showed significant association with resistance to the anticoccidial drug Arprinocid (Merck Research Laboratories).

In order to exploit the genetic map for positional cloning and gene identification in general, the sequencing of the nuclear genome of *E. tenella* H strain is being carried out using the whole-genome shotgun approach (Shirley *et al*., 2004). Previous and current evidence suggests that the genome is approximately 60 Mb, distributed amongst 14 chromosomes ranging in size from 1 Mb to >6 Mb (Shirley, 1994; **http://www.sanger.ac.uk/Projects/E_tenella/**). Complementary to this whole-genome sequencing effort, the complete sequence of the two smallest chromosomes of *E. tenella* H strain is being generated (Shirley *et al*., 2004). Chromosome 1 (∼1.0 Mb) and 2 (∼1.2 Mb) have been implicated via genetic linkage mapping, with resistance to the anticoccidial drug aprinocid and accelerated parasite growth, respectively. Preliminary analysis reveals that while the *E. tenella* genome is highly enriched in repetitive sequences, the distribution of these elements is markedly different between the two smallest chromosomes. Analysis of the chromosome 1 sequence reveals the presence of the repetitive heptamer TTTAGGG, which commonly characterizes the telomeres of other protozoan parasites including *P. falciparum*. A previously identified low-complexity repeat, GCA/TGC, has been found in arrays up to ∼20 triplets interspersed in both coding and noncoding sequence (Jenkins, 1988; Shirley, 1994; Shirley *et al*., 2004). This pattern of

low-complexity regions in coding sequence does not appear to follow phylogenetic lines as they do not appear to be present in *T. gondii*.

2.4. *Theileria spp*

The life cycle of *Theileria* spp. parallels that of *Plasmodium* spp., with some unusual twists. Asexual reproduction begins with sporozoite infection of the vertebrate host lymphocytes where schizogony occurs in synchrony with cell division (for review, see Norval *et al*., 1992). In certain species, the lymphocytes exhibit phenotypic traits reminiscent of transformed tumor cells. A proportion of these go on to infect erythrocytes within which they form gamonts that are ingested by the tick. The gamonts transform into micro- and macrogametocytes that fuse to form a zygote that undergoes meiosis, resulting in sporozoite formation (see Figure 1). Since the major disease caused by *Theileria* is in bovines, *T. parva*, the parasite responsible for the African cattle disease "East Coast Fever" was chosen for genome analysis (Nene *et al*., 1998).

The nuclear genomes of *Theileria* spp. and specifically *T. parva* are amongst the smallest analyzed thus far amongst apicomplexans. The genome consists of four chromosomes ranging in size from about 2.2 to 3.2 Mb for a total of approximately 10 Mb (Nene *et al*., 1998; Nene *et al*., 2000). The genome has an average G+C content of 31% and appears to have very little dispersed repetitive DNA. The paucity of repetitive DNA, apparent high gene density, with introns that tend to be few and short compared to other apicomplexans, contributes to making this a very compact genome.

2.5. *Cryptosporidia spp*

Cryptosporidium spp. infect a wide range of mammalian species with similar pathologies (for recent review, see Fayer *et al*., 1997). *C. parvum* is a major source of water and food-borne contamination, causing a self-limiting diarrhea in normal patients, but it can result in a severe life-threatening disease in the immunocompromised. There are currently no effective therapies for human infection. The life cycle of *Cryptosporidium* spp. generally resembles that of a Coccidian except that it invades only the intestinal cell membrane and does not enter the cytoplasm. Also, microgametes do not exflagellate, and sporulation resulting in four sporozoites occurs in the gut allowing both fecal shedding and autoinfection.

C. parvum is unusual in that it appears to lack both a mitochondrion and an apicoplast with a highly compact nuclear genome consisting of eight chromosomes ranging in size from 1.04 to 1.54 Mb (Blunt *et al*., 1997; Caccio *et al*., 1998; Piper *et al*., 1998). In contrast to other apicomplexan genomes, for the *C. parvum* genome, a complete HAPPY map provided a detailed physical description of the genome prior to genomic sequence assembly (Piper *et al*., 1998). This map, in conjunction with a random shotgun genomic sequence to ~13X coverage, revealed a 9.1-Mb genome distributed across eight chromosomes with a G+C content of 70% (see Table 1). Just over 3800 genes are predicted, with an average coding

sequence of about 1.8 kb, and of which only 5% are estimated to contain introns. Unlike *Plasmodium* spp. and *E. tenella*, *C. parvum* coding sequences are not enriched with low-complexity sequences. The absence of organelles and lack of obvious surface protein families accounts at least in part for the relative paucity of genes compared to the ~5300 estimated in *P. falciparum*. Although some putative mitochondrial coding sequences were identified, the absence of genes encoding proteins critical to electron transport and the Krebs cycle indicate that these parasites do not use oxidative phosphorylation but likely rely on glycolysis for ATP production. Other metabolic processes such as fatty acid and nucleotide synthesis are also comparatively limited. The lack of key enzymes in these pathways suggests that *C. parvum* relies heavily on scavenging and may also explain why some conventional drugs are ineffective. Nevertheless, the complete genome sequence has identified some traditional and new candidates for chemotherapeutic intervention.

3. Bioinformatics

Over the last decade, species-specific and nonspecific databases for Apicomplexans have been developed as repositories for sequence data, with some including associated nonsequence-based data (see Table 2). At present, the available resources vary tremendously between organisms, but the data is rapidly changing and gaining in amount as well as quality. Malaria databases are the most advanced following the completion of *P. falciparum* genome sequence almost 2 years ago, covering a wide variety of data types including genomic, proteomic, and expression data.

Parasite-associated databases have many different styles of user interface often depending on where they were constructed, but the underlying architecture may be either flat-file, relational/object-oriented, or a combination of the two. Early databases were flat-file project-specific databases, mostly from the EST sequencing projects and the initial genome sequencing efforts. Many of these flat-file databases are still in use today as they provide an easy way to share information, and their information can be easily opened using a wide variety of spreadsheet or word-processing software, for example, the *T. gondii* clustered EST database at the University of Pennsylvania (**http://paradb.cis.upenn.edu/toxo1/index.html**). This database was originally assembled using the cap2 program (Huang, 1996) and has a basic user interface with BLAST, text searching, and total database download tools incorporated. However, the limited scope of flat-file databases has led to the more recent development of complex general databases based on a combination of flat-file and relational data structures. Examples include the *Plasmodium* genome resource PlasmoDB (**http://www.plasmoDB.org**), the *Toxoplasma* genome resource ToxoDB (**http://www.toxodb.org**), and the *Cryptosporidium* genome resource CryptoDB (**http://cryptodb.org**).

PlasmoDB for example, is able to incorporate completed sequence information for *P. falciparum*, as well as data from other related projects. These include alternative species sequencing efforts, RNA (EST, SAGE, and microarray) and protein expression profiling, genetic organization, and population structure studies. Many tools have been built into the user interface for the researcher to find, relate, and understand data, including BLAST (with several user-defined options),

text searching for sequence retrieval, XCluster (for expression data), protein identification, and prediction tools as well as various motif-searching tools (for further details, see **http://www.plasmodb.org/restricted/Tools.shtml**). Since very few *P. falciparum* gene/proteins have been characterized, these tools combined with the database structure now allow a researcher to find a gene and retrieve information on its expression changes throughout a process, for example, parasite invasion of a red blood cell.

The ability to integrate data from many different sources has allowed scientists to ask questions that were previously impossible to address. This has shifted the focus to the interpretation and validation of genomic organization, predicted protein structure, and putative protein function. An example was the application of new computer-based analysis methods for cross-species gene discovery (Ajioka *et al*., 1998). Using the data generated further, bench-based analysis was carried out and published (Manger *et al*., 1998b).

Over the next few years, the sequences for several other phylum members are likely to be completed and then the task of data integration and database building begins in earnest.

Acknowledgments

We are grateful to David Ferguson, David Walliker, and Michael White for providing useful insight into apicomplexan life cycles, apicomplexan genetics, and the cell cycle respectively. Funding for this work was provided by the BBSRC (JWA)

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, *et al*. (2004) Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science*, **304**, 441–445.

Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, *et al*. (1998) Gene discovery by EST sequencing in Toxoplasma gondii reveals sequences restricted to the Apicomplexa. *Genome Research*, **8**, 18–28.

Ajzenberg D, Banuls AL, Tibayrenc M and Darde ML (2002) Microsatellite analysis of Toxoplasma gondii shows considerable polymorphism structured into two main clonal groups. *International Journal for Parasitology*, **32**, 27–38.

Allsopp BA and Allsopp MT (1988) Theileria parva: genomic DNA studies reveal intra-specific sequence diversity. *Molecular and Biochemical Parasitology*, **28**, 77–83.

Anders RF, Shi PT, Scanlon DB, Leach SJ, Coppel RL, Brown GV, Stahl HD and Kemp DJ (1986) Antigenic repeat structures in proteins of Plasmodium falciparum. *Ciba Foundation Symposium*, **119**, 164–183.

Aravind L, Iyer LM, Wellems TE and Miller LH (2003) Plasmodium biology: genomic gleanings. *Cell*, **115**, 771–785.

Barta JR (1997) Investigating phylogenetic relationships within the Apicomplexa using sequence data: the search for homology. *Methods*, **13**, 81–88.

Blackman MJ and Bannister LH (2001) Apical organelles of Apicomplexa: biology and isolation by subcellular fractionation. *Molecular and Biochemical Parasitology*, **117**, 11–25.

Blackston CR, Dubey JP, Dotson E, Su C, Thulliez P, Sibley D and Lehmann T (2001) High-resolution typing of Toxoplasma gondii using microsatellite loci. *The Journal of Parasitology*, **87**, 1472–1475.

Blader IJ, Manger ID and Boothroyd JC (2001) Microarray analysis reveals previously unknown changes in Toxoplasma gondii-infected human cells. *The Journal of Biological Chemistry*, **276**, 24223–24231.

Blunt DS, Khramtsov NV, Upton SJ and Montelone BA (1997) Molecular karyotype analysis of Cryptosporidium parvum: evidence for eight chromosomes and a low-molecular-size molecule. *Clinical and Diagnostic Laboratory Immunology*, **4**, 11–13.

Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J and DeRisi JL (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS Biology*, **1**, E5.

Burg JL, Grover CM, Pouletty P and Boothroyd JC (1989) Direct and sensitive detection of a pathogenic protozoan, Toxoplasma gondii, by polymerase chain reaction. *Journal of Clinical Microbiology*, **27**, 1787–1792.

Burg JL, Perelman D, Kasper LH, Ware PL and Boothroyd JC (1988) Molecular analysis of the gene encoding the major surface antigen of Toxoplasma gondii. *Journal of Immunology*, **141**, 3584–3591.

Caccio S, Camilli R, La Rosa G and Pozio E (1998) Establishing the Cryptosporidium parvum karyotype by NotI and SfiI restriction analysis and Southern hybridization. *Gene*, **219**, 73–79.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, **419**, 512–519.

Carlton JM, Galinski MR, Barnwell JW and Dame JB (1999) Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Molecular and Biochemical Parasitology*, **101**, 23–32.

Cleary MD, Singh U, Blader IJ, Brewer JL and Boothroyd JC (2002) Toxoplasma gondii asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. *Eukaryotic Cell*, **1**, 329–340.

Cowman AF and Karcz S (1993) Drug resistance and the P-glycoprotein homologues of Plasmodium falciparum. *Seminars in Cell Biology*, **4**, 29–35.

Day KP, Karamalis F, Thompson J, Barnes DA, Peterson C, Brown H, Brown GV and Kemp DJ (1993) Genes necessary for expression of a virulence determinant and for transmission of Plasmodium falciparum are located on a 0.3-megabase region of chromosome 9. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 8292–8296.

Depoix D, Carcy B, Jumas-Bilak E, Pages M, Precigout E, Schetters TP, Ravel C and Gorenflot A (2002) Chromosome number, genome size and polymorphism of European and South African isolates of large Babesia parasites that infect dogs. *Parasitology*, **125**, 313–321.

Dubey JP, Lindsay DS and Speer CA (1998) Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts. *Clinical Microbiology Reviews*, **11**, 267–299.

Dyer M and Day KP (2000) Commitment to gametocytogenesis in Plasmodium falciparum. *Parasitology Today*, **16**, 102–107.

Ellis JT, Morrison DA and Jefferies AC (1998) *The Phylum Apicomplexa: an Update on the Molecular Phylogeny*, Kluwer: Boston.

Fayer R, Speer CA and Dubey JP (1997) *The General Biology of Cryptosporidium*, CRC: Boca Raton.

Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE and Scherf A (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature*, **407**, 1018–1022.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.

Goldberg DE, Sharma V, Oksman A, Gluzman IY, Wellems TE and Piwnica-Worms D (1997) Probing the chloroquine resistance locus of Plasmodium falciparum with a novel class of multidentate metal(III) coordination complexes. *The Journal of Biological Chemistry*, **272**, 6567–6572.

Grigg ME, Bonnefoy S, Hehl AB, Suzuki Y and Boothroyd JC (2001) Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries. *Science*, **294**, 161–165.

Guinet F, Dvorak JA, Fujioka H, Keister DB, Muratova O, Kaslow DC, Aikawa M, Vaidya AB and Wellems TE (1996) A developmental defect in Plasmodium falciparum male gametogenesis. *The Journal of Cell Biology*, **135**, 269–278.

Hammond DM (1982) *Life Cycles and Development of Coccidia*, University Park Press: Baltimore.

Hill D and Dubey JP (2002) Toxoplasma gondii: transmission, diagnosis and prevention. *Clinical Microbiology and Infection*, **8**, 634–640.

Howe DK and Sibley LD (1995) Toxoplasma gondii comprises three clonal lineages: correlation of parasite genotype with human disease. *The Journal of Infectious Diseases*, **172**, 1561–1566.

Huang X (1996) An improved sequence assembly program. *Genomics*, **33**, 21–31.

Irvin AD, Ocama JG and Spooner PR (1982) Cycle of bovine lymphoblastoid cells parasitised by Theileria parva. *Research in Veterinary Science*, **33**, 298–304.

Jacobberger JW, Horan PK and Hare JD (1992) Cell cycle analysis of asexual stages of erythrocytic malaria parasites. *Cell Proliferation*, **25**, 431–445.

Jeffers TK (1975) Attenuation of Eimeria tenella through selection for precociousness. *The Journal of Parasitology*, **61**, 1083–1090.

Jeffers TK (1976) Genetic recombination of precociousness and anticoccidial drug resistance in Eimeria tenella. *Zeitschrift Fur Parasitenkunde*, **50**, 251–255.

Jenkins MC (1988) A cDNA encoding a merozoite surface protein of the protozoan Eimeria acervulina contains tandem-repeated sequences. *Nucleic Acids Research*, **16**, 9863.

Jerome ME, Radke JR, Bohne W, Roos DS and White MW (1998) Toxoplasma gondii bradyzoites form spontaneously during sporozoite-initiated development. *Infection and Immunity*, **66**, 4838–4844.

Jones SH, Lew AE, Jorgensen WK and Barker SC (1997) Babesia bovis: genome size, number of chromosomes and telomeric probe hybridisation. *International Journal for Parasitology*, **27**, 1569–1573.

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, *et al*. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.

Leander BS, Clopton RE and Keeling PJ (2003) Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. *International Journal of Systematic and Evolutionary Microbiology*, **53**, 345–354.

Manger ID, Hehl A, Parmley S, Sibley LD, Marra M, Hillier L, Waterston R and Boothroyd JC (1998a) Expressed sequence tag analysis of the bradyzoite stage of Toxoplasma gondii: identification of developmentally regulated genes. *Infection and Immunity*, **66**, 1632–1637.

Manger ID, Hehl AB and Boothroyd JC (1998b) The surface of Toxoplasma tachyzoites is dominated by a family of glycosylphosphatidylinositol-anchored antigens related to SAG1. *Infection and Immunity*, **66**, 2237–2244.

Matrajt M, Angel SO, Pszenny V, Guarnera E, Roos DS and Garberi JC (1999) Arrays of repetitive DNA elements in the largest chromosomes of Toxoplasma gondii. *Genome*, **42**, 265–269.

Matrajt M, Donald RG, Singh U and Roos DS (2002) Identification and characterization of differentiation mutants in the protozoan parasite Toxoplasma gondii. *Molecular Microbiology*, **44**, 735–747.

McDonald V and Rose ME (1987) Eimeria tenella and E. necatrix: a third generation of schizogony is an obligatory part of the developmental cycle. *The Journal of Parasitology*, **73**, 617–622.

McDonald V and Shirley MW (1987) The endogenous development of virulent strains and attenuated precocious lines of Eimeria tenella and E. necatrix. *The Journal of Parasitology*, **73**, 993–997.

Mercier C, Lefebvre-Van Hende S, Garber GE, Lecordier L, Capron A and Cesbron-Delauw MF (1996) Common cis-acting elements critical for the expression of several genes of Toxoplasma gondii. *Molecular Microbiology*, **21**, 421–428.

Nakaar V, Bermudes D, Peck KR and Joiner KA (1998) Upstream elements required for expression of nucleoside triphosphate hydrolase genes of Toxoplasma gondii. *Molecular and Biochemical Parasitology*, **92**, 229–239.

Nene V, Bishop R, Morzaria S, Gardner MJ, Sugimoto C, ole-MoiYoi OK, Fraser CM and Irvin A (2000) Theileria parva genomics reveals an atypical apicomplexan genome. *International Journal for Parasitology*, **30**, 465–474.

Nene V, Morzaria S and Bishop R (1998) Organisation and informational content of the Theileria parva genome. *Molecular and Biochemical Parasitology*, **95**, 1–8.

Norval RAI, Perry BD and Young AS (1992) *The Epidemiology of Theileriosis in Africa*, Academic Press: Orlando.

Ossorio PN, Sibley LD and Boothroyd JC (1991) Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of Toxoplasma gondii. *Journal of Molecular Biology*, **222**, 525–536.

Peterson DS, Walliker D and Wellems TE (1988) Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 9114–9118.

Pfefferkorn LC and Pfefferkorn ER (1980) Toxoplasma gondii: genetic recombination between drug resistant mutants. *Experimental Parasitology*, **50**, 305–316.

Piper MB, Bankier AT and Dear PH (1998) A HAPPY map of Cryptosporidium parvum. *Genome Research*, **8**, 1299–1307.

Radke JR, Striepen B, Guerini MN, Jerome ME, Roos DS and White MW (2001) Defining the cell cycle for the tachyzoite stage of Toxoplasma gondii. *Molecular and Biochemical Parasitology*, **115**, 165–175.

Shirley MW (1994) The genome of Eimeria tenella: further studies on its molecular organisation. *Parasitology Research*, **80**, 366–373.

Shirley MW and Harvey DA (1996) Eimeria tenella: genetic recombination of markers for precocious development and arprinocid resistance. *Applied Parasitology*, **37**, 293–299.

Shirley MW and Harvey DA (2000) A genetic linkage map of the apicomplexan protozoan parasite Eimeria tenella. *Genome Research*, **10**, 1587–1593.

Shirley MW, Ivens A, Gruber A, Madeira AM, Wan KL, Dear PH and Tomley FM (2004) The Eimeria genome projects: a sequence of events. *Trends in Parasitology*, **20**, 199–201.

Sibley LD and Boothroyd JC (1992a) Construction of a molecular karyotype for Toxoplasma gondii. *Molecular and Biochemical Parasitology*, **51**, 291–300.

Sibley LD and Boothroyd JC (1992b) Virulent strains of Toxoplasma gondii comprise a single clonal lineage. *Nature*, **359**, 82–85.

Sibley LD, LeBlanc AJ, Pfefferkorn ER and Boothroyd JC (1992) Generation of a restriction fragment length polymorphism linkage map for Toxoplasma gondii. *Genetics*, **132**, 1003–1015.

Singh U, Brewer JL and Boothroyd JC (2002) Genetic analysis of tachyzoite to bradyzoite differentiation mutants in Toxoplasma gondii reveals a hierarchy of gene induction. *Molecular Microbiology*, **44**, 721–733.

Soete M, Camus D and Dubremetz JF (1994) Experimental induction of bradyzoite-specific antigen expression and cyst formation by the RH strain of Toxoplasma gondii in vitro. *Experimental Parasitology*, **78**, 361–370.

Soldati D and Boothroyd JC (1995) A selector of transcription initiation in the protozoan parasite Toxoplasma gondii. *Molecular and Cellular Biology*, **15**, 87–93.

Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW and Sibley LD (2003) Recent expansion of Toxoplasma through enhanced oral transmission. *Science*, **299**, 414–416.

Su X, Ferdig MT, Huang Y, Huynh CQ, Liu A, You J, Wootton JC and Wellems TE (1999) A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. *Science*, **286**, 1351–1353.

Su C, Howe DK, Dubey JP, Ajioka JW and Sibley LD (2002) Identification of quantitative trait loci controlling acute virulence in Toxoplasma gondii. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10753–10758.

Su X, Kirkman LA, Fujioka H and Wellems TE (1997) Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa. *Cell*, **91**, 593–603.

Vaidya AB, Morrisey J, Plowe CV, Kaslow DC and Wellems TE (1993) Unidirectional dominance of cytoplasmic inheritance in two genetic crosses of Plasmodium falciparum. *Molecular and Cellular Biology*, **13**, 7349–7357.

Vaidya AB, Muratova O, Guinet F, Keister D, Wellems TE and Kaslow DC (1995) A genetic locus on Plasmodium falciparum chromosome 12 linked to a defect in mosquito-infectivity and male gametogenesis. *Molecular and Biochemical Parasitology*, **69**, 65–71.

Walliker D, Carter R and Morgan S (1971) Genetic recombination in malaria parasites. *Nature*, **232**, 561–562.

Walliker D, Carter R and Sanderson A (1975) Genetic studies on Plasmodium chabaudi: recombination between enzyme markers. *Parasitology*, **70**, 19–24.

Walliker D, Sanderson A, Yoeli M and Hargreaves BJ (1976) A genetic investigation of virulence in a rodent malaria parasite. *Parasitology*, **72**, 183–194.

Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, Corcoran LM, Burkot TR and Carter R (1987) Genetic analysis of the human malaria parasite Plasmodium falciparum. *Science*, **236**, 1661–1666.

Weiss LM, Laplace D, Takvorian PM, Tanowitz HB, Cali A and Wittner M (1995) A cell culture system for study of the development of Toxoplasma gondii bradyzoites. *The Journal of Eukaryotic Microbiology*, **42**, 150–157.

Wellems TE, Panton LJ, Gluzman IY, do Rosario VE, Gwadz RW, Walker-Jonah A and Krogstad DJ (1990) Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. *Nature*, **345**, 253–255.

Wellems TE, Walliker D, Smith CL, do Rosario VE, Maloy WL, Howard RJ, Carter R and McCutchan TF (1987) A histidine-rich protein gene marks a linkage group favored strongly in a genetic cross of Plasmodium falciparum. *Cell*, **49**, 633–642.

Zhu G, Marchewka MJ and Keithly JS (2000) Cryptosporidium parvum appears to lack a plastid genome. *Microbiology*, **146**(Pt 2), 315–321.

Specialist Review

Reverse vaccinology: a critical analysis

Guido Grandi
Chiron Vaccines, Siena, Italy

1. Introduction

When attacked for the first time by a new pathogen, our organism responds through the activation of the two immune defense pathways known as innate immune response and adaptive immune response. While the innate response has the peculiarity of being nonspecific but very rapid and has the role of greatly attenuating the potentially devastating effects of pathogen invasion, the adaptive immune response is pathogen–specific. It takes a few weeks to be fully activated, and not only is it usually capable of eliminating the pathogen but it also protects our organism from subsequent aggressions. The adaptive immune response works through the recognition of few specific pathogen components that become the targets of cell- and/or antibody-mediated responses that ultimately kill the pathogen. The identification of these components and their administration before primary infection prevent the outbreak of the disease and represent the tasks of modern vaccinology. In this respect, vaccinology can be seen as a "search-for-the-needle-in-the-haystack" type of undertaking, as it requires the identification of the very few protective antigens among several hundred pathogen components.

With their pioneering work published in Science in 2000, Pizza *et al*. (2000) proposed a revolutionary approach to vaccine discovery. The approach, named "reverse vaccinology" (Rappuoli, 2000), stems from the simple and straightforward consideration that, if in a given pathogen protein antigens with protective immunological properties exist, their coding genes must be sitting somewhere in the pathogen genome. Therefore, by providing the complete list of proteins, the knowledge of genome sequence offers the opportunity to systematically analyze each protein until the ones having the desired properties are unveiled. In its most classical application, reverse vaccinology can be outlined as follows: (1) genome sequencing of the pathogen of interest, (2) gene selection by in silico analysis of the genome, (3) high-throughput cloning and expression of selected genes, (4) purification of recombinant proteins, and (5) identification of potential vaccine candidates by systematic analysis of all purified proteins using appropriate in vitro and/or in vivo assays. Subsequently, it became clear that gene selection could be optimized by applying, in addition to the in silico analysis, other criteria that make the entire process more efficient (Grandi, 2004). These criteria include DNA microarray and proteomics analyses.

The present work will first review published as well as unpublished examples of reverse vaccinology and then, on the basis of the results presented, it will attempt a critical analysis of the technology with the aim of facilitating future vaccine discovery projects.

2. Examples of reverse vaccines

2.1. Group B *Neisseria meningitidis*

Meningococcal meningitis and sepsis are caused by *Neisseria meningitidis*, a gram-negative, capsulated bacterium, classified into five major pathogenic serogroups (A, B, C, Y, and W135) on the basis of the chemical composition of their capsular polysaccharides (Gotschlich *et al*., 1969a,b). Very effective vaccines based on the capsular polysaccharides against Meningococcus C are already on the market, and anti-Meningococcus A/C/Y/W polyvalent vaccines are expected to be launched in the near future. However, because of the poor immunogenicity of its capsular polysaccharide, no vaccines are available yet against MenB, the meningococcal serotype responsible for a large proportion (from 32 to 80%) of all meningococcal infections in industrialized countries (Scholten *et al*., 1993). Because of the usefulness of the capsular polysaccharide, a few vaccines based on surface-exposed proteins have been tested, although some membrane-associated proteins have been shown to elicit protective bactericidal antibodies (Poolman, 1995; Martin *et al*., 1997). However, many of the major surface protein antigens in MenB show sequence and antigenic variability, thus failing to confer protection against many heterologous strains. Therefore, the challenge for anti-MenB vaccine research is the identification of highly conserved antigens eliciting protective immune responses (bactericidal antibodies) against a broad range of MenB isolates. To achieve this goal, a new approach called reverse vaccinology was developed and applied for the first time in Chiron Vaccines (Pizza *et al*., 2000).

The MenB genome sequence (Tettelin *et al*., 2000) was submitted to computer analysis to identify genes potentially encoding surface-exposed or exported proteins. Of the 650 proteins thus predicted, approximately 50% were successfully expressed in *Escherichia coli*. The recombinant proteins were purified and used to immunize mice and the immune sera were tested for bactericidal activity, an assay that strongly correlates with protection in humans (Goldscheider *et al*., 1969). Twenty-eight sera turned out to be bactericidal. To analyze sequence conservation of the protective antigens, the nucleotide sequences of the corresponding genes from a large panel of *N. meningitidis* clinical isolates (>250) were compared. This analysis led to the identification of five highly conserved antigens whose combination induced antibodies capable of killing most of the meningococcal strains so far tested in the complement-mediated bactericidal assay. Phase I clinical studies are about to be completed and will soon establish the ability of these antigens to induce bactericidal antibodies in humans.

2.2. *Streptococcus pneumoniae*

Streptococcus pneumoniae is the most common cause of fatal community-acquired pneumonia in the elderly, and is also one of the most common causes of middle ear infections and meningitis in children. Penicillin resistance is global in *S. pneumoniae* and the incidence of resistance to several other antibiotics is becoming a serious medical concern. Although a heptavalent glycoconjugate vaccine is on the market and is highly effective against 80% of the *S. pneumoniae* isolates in the United States (Obaro, 2002), the vaccine only covers 60% and 40% of the strains in Europe and in the rest of the world, respectively. Furthermore, considering the adaptive capacity of *Pneumococcus*, the selective pressure exerted by population-wide vaccination may result in the emergence of strains that are not included in the current pneumococcal vaccine.

Taking advantage of the availability of the genome sequence of a clinical isolate of *S. pneumoniae* (Tettelin *et al*., 2001), researchers at MedImmune selected 130 genes with sequence motifs common to secreted proteins and virulence factors. Of the 130 proteins, 108 were expressed, purified, and used to immunize mice. Six of the proteins were shown to confer protection against a disseminated mouse model of *S. pneumoniae* infection (Adamou *et al*., 2001). Although no data were reported on the conservation of these protective antigens among the plethora of *S. pneumoniae* subtypes, these results clearly show, as is the case for MenB, that protein antigens can become important components of new generation anti-*S. pneumoniae* vaccines.

2.3. *Porphyromonas gingivalis*

Porphyromonas gingivalis is a gram-negative bacterium that grows and colonizes the human oral cavity and has been implicated in the etiology of chronic adult peridontitis (The American Academy of Periodontology, 1999). By following an approach very similar to the one described for *N. meningitidis*, 120 genes were selected on the basis of the predicted localization of their coded proteins on the surface of the bacterial membrane. The selected genes were expressed in *E. coli* and tested for their capacity to be recognized by a panel of antisera against *P. gingivalis*. The subset of proteins positive to the immunological analysis was subsequently used for immunization of mice, and these were challenged with live bacteria in a subcutaneous abscess model. Two of these proteins, both homologous to the *Pseudomonas* sp. OprP proteins, demonstrated significant protection in the animal model and, therefore, were proposed as promising candidates for an anti-peridontitis vaccine (Ross *et al*., 2001).

2.4. *Group B streptococcus*

Group B streptococcus (GBS) is the major cause of neonatal sepsis in the industrialized world, accounting for 0.5–3.0 deaths/1000 live births. Eighty percent

of the GBS infections in newborns occur within the first 24–48 h after delivery (Schuchat, 1998). This group is known as early onset disease, and is generally caused by direct transmission of the bacteria from the mother to the baby during labor. A second peak of infections, which begins a week after birth and continues through the first month of life, is known as late onset disease and is usually nosocomial. The elderly are also susceptible to GBS infections, and, in the last few years, the incidence of such infections is growing to a level of particular concern.

Protection in humans against invasive GBS disease correlates with high titers of anticapsule antibodies. Since these antibodies can pass through the placenta, children born from mothers with high titers of anti-GBS antibodies have a negligible risk of being infected by GBS in the first months of life.

Experiments in mice have demonstrated that glycoconjugates of capsular polysaccharide with tetanus toxoid carrier protein can induce an immune response in pregnant females, which can confer protection in the pups against lethal GBS challenge (Paoletti *et al*., 1994). These data suggest that immunizing women before they become pregnant could effectively prevent the majority of invasive GBS disease in newborns. Unfortunately, there are at least nine capsular serotypes, and antibodies against any one of these fail to confer protection against the other serotypes (Berg *et al*., 2000; Davies *et al*., 2001; Hickman *et al*., 1999; Lin *et al*., 1998; Suara *et al*., 1998). An alternative approach would be to identify a few conserved protein antigens eliciting protective immunity against most, preferably all, GBS serotypes. In line with the reverse vaccinology strategy, researchers at Chiron Vaccines used a series of computer programs to identify, among all genes of the GBS genome (Tettelin *et al*., 2002), those encoding proteins carrying signal peptides (PSORT, SignalP), transmembrane spanning regions (TMPRED), lipoproteins and cell-wall anchored proteins (Motifs), and proteins with homology to known surface proteins in other bacteria (FastA). This analysis ultimately led to the selection of 473 genes that were subjected to the high-throughput expression purification procedure. Overall, 357 recombinant proteins were successfully purified and tested in the maternal active immunization assay. According to this assay (Paoletti *et al*., 1994), female mice are first immunized with the recombinant antigens, then mated, and the resulting offspring are challenged with a lethal dose of GBS within the first 48 hours of life. For the pups to be protected, immunization has to induce in the mothers sufficiently high levels of antibodies, which can cross the placenta and reach the mice in utero. Using this model, four new antigens were found to confer a statistically significant protection (survival rate >30% over the background, with a $P_{\mathrm{val}} < 0.05$) against at least one of the GBS strains used for challenge. Interestingly, antigen combinations were capable of protecting up to 100% of the animals from lethal doses of different GBS strains, indicating the additive, if not synergistic, effect of antigen coadministration. A particular combination of these antigens is currently in the development phase.

2.5. *Group A streptococcus*

Group A streptococcus (GAS) can colonize human throat and skin, causing, in general, relatively mild diseases but, nevertheless, very costly in terms of health

care visits, workdays lost by parents, and number of schooldays lost (Cunningham, 2000). Like GBS, GAS can also cause severe invasive diseases including scarlet fever, a frequently lethal toxic shock syndrome, and necrotizing fasciitis. In this latter respect, GAS is also known as the flesh-eating bacterium since it can turn a small wound into massive necrosis, which necessitates emergency measures including extensive surgical intervention and tissue reconstruction. Perhaps of more importance are rheumatic fever (RF), rheumatic heart disease (RHD), and glomerulonephritis, the autoimmune sequelae that can follow, in some countries at high frequencies, throat and skin infection, and scarlet fever. Overall, it has been estimated that more than 600 million people are annually infected by GAS worldwide, 500 thousand of whom die because of GAS invasive disease and RHD.

There is no vaccine available for GAS, and attempts to identify protective antigens have so far been unsuccessful (Dale, 1999). A major immunodominant antigen, the M-protein, has shown type-specific protection in both humans and animal models. However, there are over 124 known serotypes of this protein; therefore, although M-protein-based vaccines are being attempted (Hu *et al*., 2002), their efficacy still awaits confirmation in the clinics, and they are unlikely to provide broad coverage against GAS infections.

Starting from the available GAS genome sequences (Ferretti *et al*., 2001), researchers at Chiron used in silico analysis and DNA microarray technology to identify highly expressed, membrane-associated proteins. In total, 285 genes were selected and successfully expressed in *E. coli* as either His-tagged proteins or GST fusions. An adult mouse model of invasive infection based on the intraperitoneal challenge of CD1 mice with a virulent M1 serotype strain (LD_{50} = 10 CFUs) was then used for vaccine candidate selection. Protection in this model implied the elicitation of circulating opsonic/bactericidal antibodies capable of preventing systemic infection in the animals. Six antigens have so far been identified showing a statistically significant protective activity. Particularly promising is one antigen, GAS40, which conferred a survival rate above 50% and elicited opsonic antibodies. The antigen is also highly conserved (homology $>$98.5%) among a large panel of GAS clinical isolates belonging to different M-serotypes.

2.6. *Chlamydia trachomatis*

Like all obligate intracellular pathogens, for its survival and propagation, the gram-negative bacterium *Chlamydia trachomatis* must accomplish several essential tasks, which include adhering to and entering host cells, creating an intracellular niche for replication, exiting host cells for subsequent invasion of neighboring cells, and also avoiding host defense mechanisms (Stephens, 1999). To carry out all these functions, *C. trachomatis* has developed a unique biphasic life cycle involving two developmental forms, a sporelike infectious form (elementary bodies, EB), and an intracellular replicative form (reticulate bodies, RB). Adhesion, host cell colonization capabilities, and ability to cope with the host defense mechanisms when outside the cell presumably rely in large part on EB surface organization.

Its unique life cycle renders *C. trachomatis* very successful in avoiding host immune responses and establishing a chronic infection, often leading to serious

diseases (Stephens, 1999). Chronic infection of the ocular mucosa can result in blindness, whereas, in the female, infection of the upper genital tract can lead to pelvic inflammatory disease, ectopic pregnancy, and sterility. Indeed, *C. trachomatis* infection is one of the most serious causes of both male and female sterility in industrialized countries. Sexually transmitted disease (STD) induced by *C. trachomatis* have also been implicated as a risk factor for the sexual transmission of other serious pathogens such as the human immunodeficiency virus (HIV) (Ho *et al*., 1995).

In spite of years of efforts by several research groups around the world, a vaccine against human chlamydial infection is still unavailable. This may be attributed to several reasons, among which are the difficulty in culturing large quantities of the pathogen (limiting the purification of antigens to be tested in vaccine studies) and the inability to carry out any kind of genetic analysis. As a consequence of these limitations, vaccine studies have been restricted to very few chlamydial antigens, mostly tested in the mouse, intravaginally challenged with either human or mouse-adapted chlamydial isolates. From these studies, as well as from epidemiological data and vaccine trials in humans, it has been established that protection against chlamydial infection most likely correlates with both the elicitation of a CD4+ T cell-specific cytotoxic activity and a neutralizing antibody response. The data also indicate that none of the antigens so far tested is capable of conferring a consistent, robust protection in the mouse, the only efficacious vaccination being represented by the intravaginal administration of live *Chlamydia*, which protects the mouse against subsequent *C. trachomatis* challenges (Ramsey *et al*., 1999).

With this background, researchers at Chiron scanned the *C. trachomatis* genome in search of genes encoding putative surface-exposed antigens. Ninety-three genes were selected, cloned in *E. coli,* and the corresponding recombinant proteins purified. Each purified protein was injected into mice and the immune sera were used in two types of assays: FACS analysis of EBs to confirm their surface exposure (Montigiani *et al*., 2002) and in vitro neutralization of infection (Finco *et al*., 2005). Forty-eight proteins out of 93 were positive to the FACS assay, and 13 proteins elicited antibodies with neutralizing activity in vitro. These antigens, never described before as being capable of eliciting neutralizing antibodies, represent potential vaccine candidates and are currently under analysis for their protective activity in vivo in the mouse model of infection.

3. A critical analysis of reverse vaccinology

The available examples of reverse vaccinology offer the opportunity to critically analyze the technology and to discuss a few take-home lessons that might help to improve and optimize its future applications.

The first important lesson is that the selection of the correlate-of-protection assay used for the high-throughput screening of the antigens plays a crucial role for the final success of the technology. It is intuitive that, in the absence of a robust assay that correlates with protection in humans, the labor-intensive process, which goes from genome sequence to gene selection and high-throughput expression and screening of antigens, results in a useless effort. In general, more than one animal

model of infection has been described for the same pathogen. However, not all of them are supported by convincing data that demonstrate their correlation with the human system. Typical examples are the mouse models used for GAS. In GAS, the protective capacity of vaccine candidates are investigated using either mucosal immunization followed by intranasal infection (Schulze *et al*., 2003; Hall *et al*., 2004), or systemic immunization followed by intraperitoneal challenge (McMillan *et al*., 2004; Kawabata *et al*., 2001). According to the first model, protection is largely mediated by the elicitation of mucosal IgA, whereas in the second model, mice survival only occurs if the injected antigen is capable of inducing bactericidal/opsonic-circulating antibodies. While it is still not clear whether in humans the presence of bactericidal antibodies is strictly necessary to prevent streptococcal infection, it is obvious that the selection of one model or the other has important consequences on the probability that the antigens found to be protective in the mouse can work in humans as well.

The second important take-home message is that the "Holy Grail" in vaccinology, namely, the antigen that alone elicits neutralizing immune responses against all the clinical isolates of a given pathogenic species, can exist (examples are the tetanus and diptheria toxins) but is very rare. Even with reverse vaccinology that has the power to scan the protective activity of most of the proteins belonging to a given pathogen, the number of such universal antigens that are being identified is limited. More realistically, protective antigens are found that altogether have the potential to provide broad cross-protection by virtue of the fact that each of them is conserved among specific pathogen subtypes. This implies that for a successful application of reverse vaccinology, the availability of the genome sequence of only one isolate may be largely insufficient. This statement is supported by the discovery that genomic sequences from different isolates of the same pathogen have shown numerous genetic differences. This is true for *Mycobacterium tuberculosis* (two genomes), *Helicobacter pylori* (two genomes), *Chlamydia pneumoniae* (four genomes), *Staphylococcus aureus* (five genomes), *Streptococcus pyogenes* (five genomes), *Yersinia pestis* (three genomes), *Escherichia coli* (four genomes), and *Group B Streptococcus* (seven genomes, unpublished).

Undertaking a reverse vaccinology project with the availability of a single genome sequence has two important implications. First, once a vaccine candidate has been identified, its similarity among as many worldwide clinical isolates as possible needs to be investigated. A similar approach has been followed in the reverse vaccinology of *N. meningitidis,* whereby the sequence similarity of the five vaccine candidates was determined in over 250 isolates before moving to the development phase. Second, protective antigens only expressed in isolates not belonging to the subtype of the sequenced strain will be missed. This limits, quite substantially, the probability of finding a sufficiently large group of antigens whose combinations could ultimately lead to an effective vaccine formulation.

The third important lesson learnt is that the conservation of a protective antigen among a group of isolates does not necessarily translate into the capacity of such a conserved antigen to confer protection against all the isolates in which the antigen is conserved. Quite a few MenB, GBS, and GAS selected antigens that were expected to be cross-protective on the basis of their gene sequence conservation, in fact, failed to protect mice against heterologous challenge. This apparent paradox

was experimentally solved by demonstrating that even if conserved, the relative abundance of an antigen on the bacterial surface can vary quite substantially from strain to strain. This fluctuation may be due to different levels of expression, different stability, and different expression of other bacterial components that ultimately mask the protective antigen (this is particularly true in the case of capsulated bacteria where, depending upon environmental conditions, variation in capsule expression is well-known to occur among strains, and even within the same strain). Now, since the level of expression is approximately proportional to the probability that the antigen will be recognized by the effector functions of the immune system, pathogens having a poorly expressed antigen are less exposed to antigen-specific immune responses. This observation implies that the extent of cross-protection conferred by a protective antigen cannot be extrapolated only by its degree of gene conservation but rather must be experimentally verified. In practical terms, if the assay for antigen selection is based on animal immunization followed by animal challenge with the pathogen, multistrain challenge is necessary to demonstrate the cross-protection efficacy of the candidate.

One last comment deserves some attention and future homework. One question that is still open in vaccinology is whether or not all protective antigens fall into the category of those antigens that are naturally highly immunogenic. If we look at the list of subunit-based vaccines that are currently on the market, the answer seems to be, definitely yes. Tetanus toxin, diphtheria toxin, *Bordetella pertussis* vaccine components, and capsular polysaccharides, to give some examples, are all highly immunogenic and elicit strong antibody responses in humans during infection. Obviously, a conclusive answer to this question has important implications on the future strategies of vaccine discovery. Reverse vaccinology does not take into account the immunogenic properties of antigens, and the selection of proteins eligible for the high-throughput screening phase is based on their location in the cell (surface and membrane-associated antigens) and their known or predicted function (virulence factors, toxins, etc.). These selection criteria impose a substantial amount of work in the subsequent gene cloning, protein expression and purification, and candidate-identification steps, and may miss some good candidates. Should only the naturally immunogenic antigens be the relevant ones for vaccine applications, strategies based on (1) immunogenic antigen selection and (2) antigen expression purification and screening are expected to be more efficient. One such strategy has been recently described (Etz *et al*., 2002). *Staphylococcus aureus* genomic DNA was fragmented and the derived fragments were cloned in the *E. coli* LamB and FhuA expression systems. In so doing, *E. coli* libraries with *S. aureus* protein domains expressed on the surface of the cells were generated. These libraries were screened with human sera with high anti-*S. aureus* antibody titers and opsonic activity, and positive clones that specifically reacted with the sera were identified. Finally, the *S. aureus* antigens expressed on the surface of the positive clones were tested for their capacity to confer protection in a mouse model of infection, and four antigens were reported to be protective. These data indicate that approaches based on the preselection of immunogenic antigens can be very effective in vaccine candidate identification. However, if protective antigens exist that are not naturally immunogenic, these antigens are necessarily lost.

To establish, once and for all, whether all protective antigens are immunogenic, it would be sufficient to take the same pathogen, search for protective antigens using both reverse vaccinology and the immunogenic protein selection approach, and see whether or not the two strategies identify the same vaccine candidates.

References

Adamou JE, Heinrichs JH, Erwin AL, Walsh W, Gayle T, Dormitzer M, Dagan R, Brewah YA, Barren P Lathigra R, *et al*. (2001) Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. *Infection and Immunity*, **69**, 949–958.

Berg S, Trollfors B, Lagergard T, Zackrisson G and Claesson BA (2000) Serotypes and clinical manifestations of group B streptococcal infections in western Sweden. *Clinical Microbiology and Infection*, **6**, 9–13.

Cunningham MW (2000) Pathogenesis of group A streptococcal infections. *Clinical Microbiology Reviews*, **13**, 470–511.

Dale JB (1999) Group A streptococcal vaccines. *Infectious Disease Clinics of North America*, **13**, 227–243, viii.

Davies HD, Raj S, Adair C, Robinson J and McGeer A (2001) Population-based active surveillance for neonatal group B streptococcal infections in Alberta, Canada: implications for vaccine formulation. *The Pediatric Infectious Disease Journal*, **20**, 879–884.

Etz H, Minh DB, Henics T, Dryla A, Winkler B, Triska C, Boyd AP, Söllner J, Schmidt W, von Ahsen U, *et al*. (2002) Identification of *in vivo* expressed vaccine candidate antigens from *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6573–6578.

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, *et al*. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4658–4663.

Finco O, Bonci A, Agnusdei M, Scarselli M, Petracca R, Norais N, Ferrari G, Garaguso I, Donati M, Sambri V, *et al*. (2005) Identification of new potential vaccine candidates against *Chlamydia pneumoniae* by multiple screenings. *Vaccine*, **23**, 1178–1188.

Goldscheider I, Gotschlich EC and Artenstein MS (1969) Human immunity to meningococcus. I. The role of humoral antibodies. *The Journal of Experimental Medicine*, **129**, 1307–1326.

Gotschlich EC, Goldschneider I and Artenstein MS (1969a) Human immunity to the meningococcus. IV. immunogenicity of group A and group C meningococcal polysaccharides in human volunteers. *The Journal of Experimental Medicine*, **129**, 1367–1384.

Gotschlich EC, Liu TY and Artenstein MS (1969b) Human immunity to the Meningococcus. 3. preparation and immunochemical properties of the group A, group B and group C meningococcal polysaccharides. *The Journal of Experimental Medicine*, **129**, 1349–1365.

Grandi G (2004) Bioinformatics, DNA microarrays and proteomics in vaccine discovery: competing or complementary technologies? In *Genomics, Proteomics and Vaccines*, Grandi G (Ed.) John Wiley & Sons.

Hall MA, Stroop SD, Hu MC, Walls MA, Reddish MA, Burt DS, Lowell GH and Dale JB (2004) Intranasal immunization with multivalent group A streptococcal vaccines protects mice against intranasal challenge infections. *Infection and Immunity*, **72**, 2507–2012.

Hickman ME, Rench MA, Ferrieri P and Baker CJ (1999) Changing epidemiology of group B streptococcal colonization. *Pediatrics*, **104**, 203–209.

Ho JL, He S, Hu A, Geng J, Basile FG, Almeida MG, Saito AY, Laurence J and Johnson WD Jr (1995) Neutrophils from human immunodeficiency virus (HIV)-seronegative donors induce HIV replication from HIV-infected patients' mononuclear cells and cell lines: an *in vitro* model of HIV transmission facilitated by *Chlamydia trachomatis*. *The Journal of Experimental Medicine*, **181**, 1493–1505.

Hu MC, Walls MA, Stroop SD, Reddish MA, Beall B and Dale JB (2002) Immunogenicity of a 26-valent group A streptococcal vaccine. *Infection and Immunity*, **70**, 2171–2177.

Kawabata S, Kunitomo E, Terao Y, Nakagawa I, Kikuchi K, Totsuka K and Hamada S (2001) Systemic and mucosal immunizations with fibronectin-binding protein FBP54 induce protective immune responses against *Streptococcus pyogenes* challenge in mice. *Infection and Immunity*, **69**, 924–930.

Lin FY, Clemens JD, Azimi PH, Regan JA, Weisman LE, Philips JB III, Rhoads GG, Clark P, Brenner RA, Ferrieri P, *et al*. (1998) Capsular polysaccharide types of group B streptococcal isolates from neonates with early-onset systemic infection. *The Journal of Infectious Diseases*, **177**, 790–792.

Martin D, Cadieux N, Hamel J and Brodeur BR (1997) Highly conserved *Neisseria meningitidis* surface protein confers protection against experimental infection. *The Journal of Experimental Medicine*, **185**, 1173.

McMillan DJ, Davies MR, Good MF and Sriprakash KS (2004) Immune response to superoxide dismutase in group A streptococcal infection. *Immunology and Medical Microbiology*, **40**, 249–256.

Montigiani S, Falugi F, Scarselli M, Finco O, Petracca R, Galli G, Mariani M, Manetti R, Agnusdei M, Cevenini R, *et al*. (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infection and Immunity*, **70**, 368–379.

Obaro SK (2002) The new pneumococcal vaccine. *Clinical Microbiology and Infection*, **8**, 623–633.

Paoletti LC, Wessels MR, Rodewald AK, Shroff AA, Jennings HJ and Kasper DL (1994) Neonatal mouse protection against infection with multiple group B streptococcal (GBS) serotypes by maternal immunization with a tetravalent GBS polysaccharide-tetanus toxoid conjugate vaccine. *Infection and Immunity*, **62**, 3236–3243.

Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, *et al*. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, **287**, 1816–1820.

Poolman JT (1995) Development of a meningococcal vaccine. *Infectious Agents and Disease*, **4**, 13–28.

Ramsey KH, Cotter TW, Salyer RD, Miranpuri GS, Yanez MA, Poulsen CE, DeWolfe JL and Byrne GI (1999) Prior genital tract infection with a murine or human biovar of *Chlamydia trachomatis* protects mice against heterotypic challenge infection. *Infection and Immunity*, **67**, 3019–3025.

Rappuoli R (2000) Reverse vaccinology. *Current Opinion in Microbiology*, **3**, 445–450.

Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, Patterson M, Agius C, Camuglia S, Reynolds E, *et al*. (2001) Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine*, **19**, 4135–4142.

Scholten RJ, Bijlmer HA, Poolman JT, Kuipers B, Caugant DA, Van Alphen L, Dankert J and Valkenburg HA (1993) Meningococcal disease in the Netherlands, 1958-1990: a steady increase in the incidence since 1982 partially caused by new serotypes and subtypes of *Neisseria meningitidis*. *Clinical Infectious Diseases*, **16**, 237–246.

Schuchat A (1998) Epidemiology of group B streptococcal disease in the United States: shifting paradigms. *Clinical Microbiology Reviews*, **11**, 497–513.

Schulze K, Medina E, Chhatwal GS and Guzman CA (2003) Stimulation of long-lasting protection against *Streptococcus pyogenes* after intranasal vaccination with non adjuvanted fibronectin-binding domain of the SfbI protein. *Vaccine*, **21**, 1958–1964.

Stephens RS (Ed.) (1999) *Chlamydia. Intracellular Biology, Pathogenesis, and Immunity*, ASM Press: Washington.

Suara RO, Adegbola RA, Mulholland EK, Greenwood BM and Baker CJ (1998) Seroprevalence of antibodies to group B streptococcal polysaccharides in Gambian mothers and their newborns. *Journal of the National Medical Association*, **90**, 109–114.

Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, *et al*. (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12391–12396.

Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, *et al*. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.

Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, *et al*. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.

The American Academy of Periodontology (1999) The pathogenesis of periodontal diseases. *Journal of Periodontology*, **70**, 457–470.

Short Specialist Review

The staphylococci

Steven R. Gill
The Institute for Genomic Research, Rockville, MD, USA

1. Introduction

The staphylococci are a major cause of nosocomial and community-acquired infections, ranging in severity from minor skin inflammation to life-threatening systemic infections that are increasingly resistant to antibiotics. The two major staphylococcal pathogens, *Staphylococcus aureus* and *Staphylococcus epidermidis*, are found on 30–70% (Peacock *et al*., 2001) and 100% (von Eiff *et al*., 2002) of the human population, respectively, where they live as commensals on the skin and mucous membranes. Infections develop when the staphylococci contaminate a breach in the host cutaneous system that occurs as a result of trauma or cuts. Nosocomial *S. aureus* infections are caused by a small number of successful epidemic clones (Lindsay and Holden, 2004; Johnson *et al*., 2001) and occur in hospitalized patients who are often subjected to treatments with needles or catheters and who have comprised immune systems. Community-associated *S. aureus* (CASA), previously associated with common skin infections (Stevens, 2003), have increasingly been linked to severe infections and lethal hemolytic pneumonia in children (Lindsay and Holden, 2004; Gillet *et al*., 2002), likely as a result of the acquisition of the Panton–Valentine leukocidin gene (*PV-luk*) (Dufour *et al*., 2002; Herold *et al*., 1998). In contrast, infections caused by the less-aggressive pathogen, *S. epidermidis*, primarily remain associated with implanted medical devices (von Eiff *et al*., 2002).

Acquisition of resistance to most classes of antimicrobial agents has made control of staphylococcal infections increasingly difficult. Widespread treatment of staphylococcal infections in the 1960s led to the emergence of methicillin-resistant *S. aureus* (MRSA) and *S. epidermidis* (MRSE), which continue to persist in both the health care and community environments (Stevens, 2003). In the United States and Japan, ~60% of nosocomial *S. aureus* isolates are resistant to methicillin and some strains have developed resistance to more than 20 different antimicrobial agents (Paulsen *et al*., 1997). The glycopeptide antibiotic, vancomycin, has been viewed as the last-resort therapy against most strains of multidrug-resistant staphylococci (Walsh, 1999). However, its effectiveness has been limited by the emergence in 1997 (CDC, 1997) of *S. aureus* with intermediate levels of resistance to vancomycin (VISA or vancomycin intermediate *S. aureus*) and the most recent emergence of *S. aureus* with high levels of resistance to vancomycin (VRSA or vancomycin resistant *S. aureus*) (CDC, 2002).

The virulence of these two pathogens is multifactorial and mediated by a wide array of extracellular toxins and surface structures. Of the two organisms, *S. aureus* produces the largest number of potential virulence factors, including hemolysins, enterotoxins, exfoliative toxins, proteases, leukocidins, and toxic shock syndrome toxin (Projan and Novick, 1997), which contribute to *S. aureus*'s aggressiveness and success as one of the most successful opportunistic human pathogens. In contrast, *S. epidermidis* produces far fewer potential virulence factors, the exception being an expanded family of phenol soluble modulins (Gill *et al*., 2005). In both species, virulence is thought to be regulated by global gene regulators, such as *agr* and *sar* (Novick, 2003), which respond to environmental signals, thereby allowing the bacteria to differentially express selected virulence factors in response to environmental or host signals.

2. Staphylococcal genomes

Seven *S. aureus* genomes (Gill *et al*., 2005; Holden *et al*., 2004; Baba *et al*., 2002; Kuroda *et al*., 2001; or available on-line at **http://www.genome.ou.edu/staph**) and two *S. epidermidis* genomes (Gill *et al*., 2005; Zhang *et al*., 2003) have been sequenced and are publicly available (Table 1). Whole-genome analysis of these genomes demonstrates that they are syntenic throughout a well-conserved core region, with differences being the result of genomic elements including genome islands (νSa, νSe, SSCmec, and SSC-like elements), integrated prophage, IS elements, composite transposons, and integrated plasmids, which are associated with disease and virulence (Table 1). Depending upon the isolate, these genomic elements make up approximately 10–20% of the *S. aureus* genomes and ∼10% of the *S. epidermidis* genomes; a proportion that is similar to that found in other gram-positive pathogens, such as group A Streptococcus (∼10%) (Beres *et al*., 2002) and *Enterococcus faecalis* (25%) (Paulsen *et al*., 2003). The core genome is composed of genes associated with central metabolism, housekeeping functions and surface proteins required for growth and survival in the host. Genes outside of the core genome are frequently species-specific genes or virulence factors carried on one of the six pathogenicity genomic islands (νSa) identified in the sequenced *S. aureus* genomes or two (νSe) in the sequenced *S. epidermidis* genomes. The νSa islands carry approximately one-half of *S. aureus* virulence factors, and the presence or absence of individual νSa determines the pathogenic potential of isolates within this species. For example, *S. aureus* MRSA252 (Holden *et al*., 2004) and MW2 (Baba *et al*., 2002) contain novel islands, SaPI4 and νSa3 respectively, that likely contribute to the virulence of these isolates. A genome island in *S. epidermidis*, νSeγ, encodes multiple members of the phenol soluble modulin (psm) family, which is likely a key virulence factor in this species (Gill *et al*., 2005). Overall, the paucity of pathogenicity genomic islands in *S. epidermidis* when compared to *S. aureus* is a direct reflection of the greater pathogenic potential of *S. aureus*. The two sequenced *S. epidermidis* differ in their ability to form a biofilm, the key factor in their role as pathogens. A comparison of RP62a (a biofilm producer) and ATCC12228 (a biofilm nonproducer) revealed that the key differences are the presence of the cell wall–associated biofilm protein (Bap) or

Table 1 Major genomic islands, bacteriophage, and associated virulence factors in sequenced staphylococcal genomes[a]

	Staphylococcus aureus						Staphylococcus epidermidis	
Strain	COL	N315	Mu50	MW2	MSSA476	MRSA252	RP62a	ATCC12228
Chromosome length (bp)	2 809 422	2 813 641	2 878 084	2 820 462	2 799 802	2 902 619	2 616 530	2 499 279
Number of ORFs	2721	2797	3028	2849	2565	2671	2553	2381
Background	HA-MRSA[b]	HA-MRSA[b]	HA-MRSA[b]	CA-MRSA[c]	CA-MSSA[d]	HA-MRSA[b]	HA-MRSE[e]	Typing strain
Reference	Gill *et al*.	Kuroda *et al*.	Kuroda *et al*.	Baba *et al*.	Holden *et al*.	Holden *et al*.	Gill *et al*.	Zhang *et al*.
GenBank accession	CP000046	BA000018	BA000017	BA000033	BX571857	BX571856	CP000028	AE015929
Genomic Islands								
νSaα	*set spl*, *lukDE*	*Set spl*, *lukDE,*	*Set spl*, *lukDE,*	*Set spl*, *lukDE,*	*Set spl*, *lukDE,*	*Set spl*, *hysA,*	–	–
νSaβ		enterotoxin	enterotoxin	*bsa*	*bsa*	enterotoxin	–	–
νSaγ	*set, eta, psmβ*	*set, eta, psmβ*	*set, eta, psmβ*	*set, eta, psmβ*	*set, eta, psmβ*	*set, eta, psmβ*	–	–
νSeγ	–	–	–	–	–	–	*psmβ*	*psmβ*
νSe1	–	–	–	–	–	–	*cadCD*	Unknown ORF
νSe2	–	–	–	–	–	–	–	*srtA*, LPXTG
Pathogenicity islands								
νSa1(SaPI1, SaPI3)	*seb, ear, sek, sei*	–	–	–	–	–	–	–
νSa3 (SaPI3)	–	–	*fhuD*	*ear, sel2, sec4*	–	–	–	–
νSa4 (SaPI2)	–	*sel, sec3, tsst*	*sel, sec3, tsst*	–	–	–	–	–
SaPI4	–	–	–	–	–	A	–	–
Bacteriophage								
ϕSa1	–	–	A	–	–	–	–	–
ϕSa2	–	–	–	*lukSF-PV*	–	A	–	–
ϕSa3	–	*sak, sep*	*sak, sea*	*sak, sea, seg2,*	*sak, sea, seg2,*	*sak, sea*	–	–
				sek2	*sek2*		–	–
ϕSa4	–	–	–	–	A	–	–	–
ϕCOL	A	–	–	–	–	–	–	–
SPβ	–	–	–	–	–	–	A	–

A: genome elements are present, but do not encode virulence or drug resistance genes.

Abbreviations: *bsa*, bacteriocin biosynthesis genes; *cadCD*, cadmium resistance genes; *ear*, putative β-lactamase protein; *eta*, exfoliative toxin A-like protein; *fhuD,* siderophore transporter; *geh*, lipase; *hysA*, hyaluronate lyase; LPXTG, cell surface protein containing LPXTG motif; *lukDE*, two components of the leukocidin DE toxins; *lukSF-PV*, two components of the Panton–Valentine leukocidin toxin; psmβ, phenol soluble modulin; *sak,* staphylokinase; *sea*, enterotoxin A, *seb*, enterotoxin B; *sec3*, enterotoxin C3; *seg2*, enterotoxin G2; *sei*, enterotoxin I; *sek*, nterotoxin K; *sel*, enterotoxin L; *set,* staphylococcal exotoxins; *spl*, staphylococcal serine proteases; *srtA*, sortase A; *yeeE*, putative transport system permease.

[a]The genome sequence of NCTC8325 has not been published but is available at **http://www.genome.ou.edu/staph**.

[b]HA-MRSA: Hospital-acquired MRSA (methicillin-resistant *S. aureus*).

[c]CA-MRSA: Community-acquired MRSA.

[d]CA-MSSA: Community-acquired MSSA (methicillin-sensitive *S. aureus*).

[e]HA-MRSE: Hospital-acquired MSRE (methicillin-resistant *S. epidermidis*).

Bap homologous protein (Bhp) and of the intercellular adhesin locus (*icaA,B,C,D*) that encodes the polysaccharide intercellular adhesin that participates in biofilm formation (Gill *et al*., 2005). Acquisition of virulence factors and formation of genome islands in staphylococci likely occur as a result of gene movement of plasmids and phage among other family members of the low-GC gram-positive bacteria. One example of this occurrence is the presence in *S. epidermidis* of a *capA,B,C* operon, which was first identified in *Bacillus anthracis* where it encodes the polyglutamate capsule, a major virulence factor. Similarly, a *Bacillus subtilis* ϕSPβ-like bacteriophage found in *S. epidermidis* RP62a has been modified to encode a nuclease and a species and RP62a specific LPXTG cell surface binding protein. Finally, VRSA occurred as a result of conjugative transfer of *Tn1546vanA* from *Enterococcus faecalis* to *S. aureus* (Weigel *et al*., 2003).

3. Concluding remarks

The emergence of CASA (Stevens, 2003) and VRSA (Weigel *et al*., 2003) demonstrates the dynamic nature of the staphylococcal genome and the role gene transfer and acquisition of virulence factors has played in the evolution of this genus. The identification of a novel pathogenicity island carrying staphylococcal enterotoxin C (*sec*) in a *S. epidermidis* isolate (S. Gill, unpublished data) illustrates the progression of this species toward a bona fide pathogen. Genome-wide comparisons of additional staphylococcal genomes, such as *Staphylococcus haemolyticus* and *Staphylococcus carnosus* that are being sequenced, will likely lead to insights into evolution of the staphylococci and novel therapeutic approaches for control of nosocomial and community-acquired infections.

References

Baba T, Takeuchi F, Kuroda M, Yuzawa H, Aoki K, Oguchi A, Nagai Y, Iwama N, Asano K, Naimi T, *et al*. (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359**, 1819–1827.

Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, Liu MY, Smoot JC, Porcella SF, Parkins LD, *et al*. (2002) Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10078–10083.

CDC (1997) *Staphylococcus aureus* with reduced susceptibility to vancomycin–United States, 1997. *Morbidity and Mortality Weekly Report*, **46**, 765–766.

CDC (2002) Vancomycin-resistant *Staphylococcus aureus*-Pennsylvania, 2002. *Morbidity and Mortality Weekly Report*, **51**, 902.

Dufour P, Gillet Y, Bes M, Lina G, Vandenesch F, Floret D, Etienne J and Richet H (2002) Community-acquired methicillin-resistant *Staphylococcus aureus* infections in France: emergence of a single clone that produces Panton-Valentine leukocidin. *Clinical Infectious Diseases*, **35**, 819–824.

Gill SR, Fouts DE, Archer GL, Mongodin EF, DeBoy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, *et al*. (2005) Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin resistant *Staphylococcus aureus* and a biofilm

producing methicillin resistant *Staphylococcus epidermidis* strain. *Journal of. Bacteriology*, **187**, 2426–2438.

Gillet Y, Issartel B, Vanhems P, Fournet JC, Lina G, Bes M, Vandenesch F, Piemont Y, Brousse N, Floret D, *et al*. (2002) Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet*, **359**, 753–759.

Herold BC, Immergluck LC, Maranan MC, Lauderdale DS, Gaskin RE, Boyle-Vavra S, Leitch CD and Daum RS (1998) Community-acquired methicillin-resistant *Staphylococcus aureus* in children with no identified predisposing risk. *JAMA*, **279**, 593–598.

Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, *et al*. (2004) Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9786–9791.

Johnson AP, Aucken HM, Cavendish S, Ganner M, Wale MC, Warner M, Livermore DM and Cookson BD (2001) Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European antimicrobial resistance surveillance system (EARSS). *The Journal of Antimicrobial Chemotherapy*, **48**, 143–144.

Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, *et al*. (2001) Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–1240.

Lindsay JA and Holden MT (2004) *Staphylococcus aureus*: superbug, super genome? *Trends in Microbiology*, **12**, 378–385.

Novick RP (2003) Autoinduction and signal transduction in the regulation of staphylococcal virulence. *Journal of Molecular Microbiology and Biotechnology*, **48**, 1429–1449.

Paulsen IT, Firth M and Skurray RA (1997) Resistance to antimicrobial agents other than β-lactams. In *The Staphylococci in Human Disease*, Crossley KB and Archer GL (Eds.), Churchill Livingstone: New York, pp. 175–212.

Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR, Heidelberg JF, *et al*. (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*, **299**, 2071–2074.

Peacock SJ, de Silva I and Lowy FD (2001) What determines nasal carriage of *Staphylococcus aureus*? *Trends in Microbiology*, **9**, 605–610.

Projan SJ and Novick RP (1997) The molecular basis of pathogenicity. In *The staphylococci in human disease*, Crossley KB and Archer GL (Ed.), Churchill Livingstone: New York, pp. 55–82.

Stevens DL (2003) Community-acquired *Staphylococcus aureus* infections: Increasing virulence and emerging methicillin resistance in the new millennium. *Current Opinion in Infectious Diseases*, **16**, 189–191.

von Eiff C, Peters G and Heilmann C (2002) Pathogenesis of infections due to coagulase-negative staphylococci. *The Lancet Infectious Diseases*, **2**, 677–685.

Walsh C (1999) Deconstructing vancomycin. *Science*, **284**, 442–443.

Weigel LM, Clewell DB, Gill SR, Clark NC, McDougal LK, Flannagan SE, Kolonay JF, Shetty J, Killgore GE and Tenover FC (2003) Genetic analysis of a high-level vancomycin-resistant isolate of *Staphylococcus aureus*. *Science*, **302**, 1569–1571.

Zhang YQ, Ren SX, Li HL, Wang YX, Fu G, Yang J, Qin ZQ, Miao YG, Wang WY, Chen RS, *et al*. (2003) Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). *Molecular Microbiology*, **49**, 1577–1593.

Short Specialist Review

Genome-wide analysis of group A *Streptococcus*

Nicole M. Green

Baylor College of Medicine, Houston, TX, USA

University of California Davis, Davis, CA, USA

James M. Musser

Baylor College of Medicine, Houston, TX, USA

1. Introduction

Infections caused by the human pathogen group A *Streptococcus* (GAS) likely were described by Hippocrates in the fifth century B.C. Historically, GAS has caused outbreaks of pharyngitis (strep throat), scarlet fever, rheumatic fever, and puerperal sepsis or "childbed fever". Although GAS commonly causes pharyngitis and skin infections, the organism recently has received considerable attention because of its ability to cause necrotizing fasciitis, a devastating infection sometimes referred to as the "flesh-eating" syndrome (Cunningham, 2000). Since the late 1980s, there has been an unexplained resurgence in several forms of severe invasive disease including necrotizing fasciitis, streptococcal toxic shock syndrome, and septicemia. It is now known that GAS causes ~10 000–15 000 cases annually of invasive infections in the United States, with a mortality rate exceeding 50% in some reports.

For more than 50 years, GAS have been classified on the basis of serological diversity in M protein, a major surface antigen and virulence factor (Cunningham, 2000). However, sequencing of the hypervariable part of the *emm* gene encoding M protein largely has replaced serological typing of strains. More than 125 *emm* types are recognized and this allelic diversity has been useful for categorizing strains for epidemiological studies. Although no one *emm* type is solely responsible for any single GAS infection type, strains expressing certain M types have been associated repeatedly with specific diseases. For example, serotype M1 and M3 strains are the leading causes of invasive infections in many western countries, and M18 strains have been commonly associated with rheumatic fever (Cunningham, 2000). Because strains of relatively few M types cause a disproportionate amount of infections, strains of these M types have been the subject of many studies. However, a classification scheme of GAS based on diversity in a single surface antigen does

not accurately reflect the extensive chromosomal and allelic diversity that exists within the species (Reid *et al*., 2001). Recombination events, many involving bacteriophages and genes encoding a wide assortment of virulence factors play an important role in generating diversity (Banks *et al*., 2002). Knowledge of the population structure of GAS and the level of naturally occurring genetic variation may help explain epidemiological observations such as rapid shifts in the predominant M serotypes causing GAS disease. Geographic and temporal differences in serotype distribution, disease frequency, and disease character are likely related to changes in GAS gene content, clonal selection, and fitness of particular strains.

2. A wealth of genome sequences

The increase in GAS disease frequency and severity has sparked renewed interest in understanding the molecular mechanisms of pathogenesis, bacterial population genetics, and vaccine development. Although host factors are undoubtedly a key factor determining infection and its outcome, a basic understanding of GAS molecular pathogenesis is needed to understand how this bacterium causes a wide range of disease types. To facilitate research and discovery, the genomes of seven GAS strains (serotypes M1, M3, M5, M6, M18, and M28) commonly causing pharyngitis and invasive disease have been sequenced recently, and additional strains are under study (Ferretti *et al*., 2001; Smoot *et al*., 2002; Beres *et al*., 2002; Nakagawa *et al*., 2003; Banks *et al*., 2004; Green *et al*., 2005; **http://www.sanger.ac.uk/Projects/S_pyogenes/**). The availability of genome sequences of M protein serotypes causing distinct diseases has yielded a tremendous amount of new information important in pathogenesis and other research.

The genome sequences have provided new insight into the extent of strain variation within and between serotypes. All strains studied thus far have a genome size of ∼1.8–1.9 Mb, similar G+C content (38.5%), six highly conserved rRNA operons, and a core group of proven and putative virulence genes. Importantly, all strains are polylysogenic, that is, contain multiple prophages or prophage-like elements that encode one or more virulence factors such as toxins. Approximately 90% of gene content is shared among strains and constitutes the core GAS genome. All strains differ in insertion sequences, small indels, and single-nucleotide polymorphisms, but the majority of variation in gene content between strains is caused by prophages or prophage-like elements. Twenty-three distinct prophages or prophage-like elements have been described thus far. As noted, an extremely important feature of these elements is that the great majority encode one or two proven or putative extracellular virulence factors such as pyrogenic toxin superantigens (PTSAgs), DNAses, a novel phospholipase A_2 (SlaA), macrolide efflux pump resistance genes, and a novel cell-wall anchored protein hypothesized to be an adhesin. Many prophage-associated virulence factor genes were unknown prior to their discovery by GAS genome sequencing projects. Other new discoveries include two-component regulators, variation in chromosomal arrangement, secreted and cell-wall anchored proteins, lipoproteins, and single-nucleotide polymorphisms between strains.

3. Prophages and variation in gene content among GAS strains

Although the existence of GAS phages was first demonstrated in the 1920s, only recently has the extent of polylysogeny and implications for several areas of GAS biology been appreciated. GAS strains vary in their prophage content, and a second level of diversity is contributed by modular recombination involving parts of prophages.

Alignment of the prophage sequences present in six available GAS genomes has revealed extensive mosaicism, presumably caused by modular recombination. Phylogenetic comparisons of individual phage genes and entire phage genomes indicate a lack of congruency, suggesting that recombination has been an important contributor to generating diversity among GAS prophages. In addition, sequencing of prophage genes such as those encoding PTSAgs (SpeA, SpeC) and SlaA from GAS strains representing distinct genetic backgrounds has identified many allelic variants, some of which may have a biologically relevant effect in pathogenesis. For example, one of the variants of SpeA (SpeA3) has been reported to have significantly more superantigen activity than the SpeA1 variant (Reid *et al*., 2001).

The genome sequences have made possible genome-wide comparisons of GAS strains. Genome comparisons by DNA–DNA microarray analysis and whole-genome PCR scanning have been used to detect differences in total gene content among strains of the same serotype (Smoot *et al*., 2002; Banks *et al*., 2004; Beres *et al*., 2004). For example, using clinical isolates from defined disease types, these methods have clearly shown that not all GAS isolates of the same M protein serotype are genetically equivalent, nor even recent clonal derivatives. Although many strains of the same serotype have a closely similar shared ("core") chromosomal gene content, strains can be highly variable in their prophage content. Inasmuch as natural competence has not been reported in GAS, it is probable that phage transduction is a primary generator of genetic diversity in this pathogen.

Two examples highlight the extent to which prophages serve as the main source of variation in gene content among strains of the same serotype (M18 and M3) (Smoot *et al*., 2002; Beres *et al*., 2002, 2004). DNA microarray analysis was performed on 36 serotype M18 strains collected over 50 years (1948–2000) from various geographic locations (Smoot *et al*., 2002). The analysis showed that prophages were the primary source of variation in gene content. Conversely, with the exception of prophages and prophage-like elements, very little variation in gene content was detected among these strains. Similarly, identity in prophage-related gene content was evident among strains collected during certain time periods and among strains involved in specific GAS outbreaks that occurred in specific geographic locations. Analogous observations were made in a recent population-based study of 255 serotype M3 strains causing two epidemics of invasive disease episodes between 1992 and 2002 in Ontario, Canada (Beres *et al*., 2004). All differences in gene content between strains were due to variation in prophage content.

4. Use of genome-wide studies to understand molecular events underlying GAS epidemics

Molecular factors that contribute to the emergence of new virulent bacterial subclones and epidemics are poorly understood. We recently studied this topic by analysis of a population-based strain sample of serotype M3 GAS recovered from patients with invasive disease (Beres *et al*., 2004). Serotype M3 strains are a leading cause of invasive GAS infections as shown by population-based surveillance studies in the United States and Canada. Patients infected with serotype M3 strains are more likely to have severe infections and die. To gain insight into molecular factors contributing to GAS subclone emergence and bacterial epidemics, a genome-wide investigative approach was used to study 255 contemporary clinical M3 strains obtained from an 11-year population-based surveillance study of invasive disease in Ontario, Canada. Genetic diversity in these serotype M3 strains was investigated by several methods including pulsed-field gel electrophoresis, DNA microarray, whole-genome PCR scanning, and PCR-based prophage genotyping to determine prophage content, *emm* gene sequencing, and single-nucleotide polymorphism (SNP) analysis (Beres *et al*., 2004). The results revealed the presence of nine distinct prophage genotypes that matched with the pulse-field gel electrophoresis patterns. However, the majority of the strains had the same prophage content as the sequenced M3-MGAS315 genome and these strains were abundant in two peaks of infection. Both DNA microarray and whole-genome PCR scanning analysis of selected strains identified virtually no differences in gene content unrelated to prophage, that is, the core genome. Sequence analysis of the *emm* gene and SNP analysis were used to further classify genetic relationships among the M3 strains. The data were used to determine that the strains were pauciclonal, with a limited number of subclone groups identified. Interestingly, statistically significant associations were present between certain prophage genotypes and GAS disease types. By applying molecular genetic techniques to epidemiological observations, temporal changes in M3 subclone gene content were discovered. Acquisition or loss of prophages, allelic variation in chromosomal genes, expansion of subclone populations, and introduction of new subclone variants were shown to contribute to peaks of infection and different infection types (Beres *et al*., 2004).

5. Insight into the emergence of drug-resistant strains: an outbreak of pharyngitis caused by macrolide resistant strains

GAS infections commonly are treated with penicillin or a related β-lactam antibiotic. Although treatment failures occur, fortunately all strains remain exquisitely susceptible in vitro to this class of antibiotics. Erythromycin and related macrolide antibiotics sometimes are used as an alternative treatment for GAS pharyngitis. In contrast to β-lactams, resistance of GAS to macrolide antibiotics was described in the late 1950s, and has increased dramatically worldwide in the last 10 years.

Recently, an outbreak of pharyngitis caused by erythromycin-resistant serotype M6 GAS strains was reported among schoolchildren in Pittsburgh, Pennsylvania (Martin *et al*., 2002). This outbreak was of particular concern because within a few months the frequency of macrolide resistant GAS increased rapidly and drug-resistant strains spread to surrounding communities. Molecular studies revealed that serotype M6 strains of a single clone were responsible and resistance was due to the presence of the *mefA* gene encoding a macrolide efflux pump.

Inasmuch as serotype M6 strains are one of the more common causes of pharyngitis and invasive infections, and no information was available about the molecular mechanism of acquisition of the *mefA* gene in the Pittsburgh case clone, we chose to sequence the genome of a genetically representative strain. Several important findings were revealed (Banks *et al*., 2003, 2004). Most importantly, we discovered that the *mefA* gene was encoded by a 58.8-kb foreign genetic element with characteristics of both a transposon and a prophage. This chimeric element was inducible under in vitro conditions and was present in all serotype M6 pharyngeal isolates tested from the Pittsburgh outbreak, as well as across multiple other serotypes from various geographic locations. These observations indicate that acquisition of the *mefA* element by horizontal gene transfer, followed by clonal expansion, were key contributors to the outbreak.

6. Expression microarray analyses

The molecular basis underlying bacterial responses to host signals during natural infections is poorly understood. To begin to address this deficit, several expression microarray studies have been done with GAS, and extensive new information has been obtained (Smoot *et al*., 2001; Graham *et al*., 2002; Voyich *et al*., 2003, 2004; Graham *et al*., 2005). Owing to space constraints, here we summarize the results from only one of these studies.

During the transition from a throat or skin infection to an invasive infection, GAS must adapt to changing environments and host factors. Recently, we used transcript profiling and functional analysis to investigate the transcriptome of a wild-type serotype M1 GAS strain in human blood (Graham *et al*., 2005). This was a particularly important investigation because GAS sepsis is a devastating infection with a high morbidity and mortality rate. Hence, insight into molecular events transpiring in blood was crucial. Using a custom-made high-density oligonucleotide array, we discovered that global changes in GAS gene expression occur rapidly in response to human blood exposure. Increased transcription was identified for many genes that are likely to enhance bacterial survival, including those encoding superantigens and host-evasion proteins. For example, upon blood exposure, we observed increased expression of GAS genes that interact with host cell surfaces (adhesions such as M1 protein, collagen-binding proteins, and capsule), and that contribute to the evasion of the host innate defenses (Sic, Mac, and SpeA). The analysis also provided new evidence that the CovR–CovS two-component gene-regulatory system functions to coordinate bacterial fitness attributes during disseminated host infections. This study provided crucial insights into strategies

used by a bacterial pathogen to thwart host defenses and survive in human blood, and suggested new vaccine and therapeutic strategies.

7. Summary

Genomic analysis of GAS has progressed substantially in the last three years, making it one of the more thoroughly analyzed human bacterial pathogens. Comparative genome sequence analysis has shown that prophages are the major source of variation in gene content among GAS strains. Genome sequencing also has identified many previously unknown virulence factors and genetic elements encoding drug resistance. Expression microarray analyses has revealed hitherto unknown regulatory pathways contributing to survival in response to temperature change and human polymorphonuclear leukocytes. Genome-wide analyses have significantly enhanced our understanding of the molecular basis underlying the emergence of new, highly successful GAS clones. Taken together, the results of these studies indicate that genomic analyses of GAS is an area of considerable interest and promises to yield additional contributions to our understanding of host–pathogen interactions and bacterial population genetics.

References

Banks DJ, Beres SB and Musser JM (2002) The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends in Microbiology*, **10**, 515–521.

Banks DJ, Porcella SF, Barbian KD, Beres SB, Philips LE, Voyich JM, DeLeo FR, Martin JM, Somerville GA and Musser JM (2004) Progress toward characterization of the group A *Streptococcus* metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. *Journal of Infectious Diseases*, **190**, 727–738.

Banks DJ, Porcella SF, Barbian KD, Martin JM and Musser JM (2003) Structure and distribution of an unusual chimeric genetic element encoding macrolide resistance in phylogenetically diverse clones of group A *Streptococcus*. *Journal of Infectious Diseases*, **188**, 1898–1908.

Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, Liu MY, Smoot JC, Porcella SF, Parkins LD, *et al*. (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10078–10083.

Beres SB, Sylva GL, Sturdevant DE, Granville CN, Liu M, Ricklefs SM, Whitney AR, Parkins LD, Hoe NP, Adams GJ, *et al*. (2004) Genome-wide molecular dissection of serotype M3 group A *Streptococcus* strains causing two epidemics of invasive infection. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11833–11888.

Cunningham M (2000) Pathogenesis of group A streptococcal infections. *Clinical Microbiology Reviews*, **13**, 470–511.

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, *et al*. (2001) Complete genomes sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4658–4663.

Graham MR, Smoot LM, Lux Migliaccio CA, Sturdevant DE, Porcella SF, Federle MJ, Scott JR and Musser JM (2002) Group A *Streptococcus* global virulence network delineated by

genome-wide transcript profiling. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13855–13860.

Graham MR, Virtaneva K, Porcella SF, Barry WT, Gowen BB, Johnson CR, Wright FA and Musser JM (2005) Group A *Streptococcu*s transcriptome dynamics during growth in human blood reveals bacterial adaptive and survival strategies. *American Journal of Pathology*, **166**, 455–465.

Green NM, Zhang S, Porcella SF, Barbian KD, Beres SB, LeFebvre RB and Musser JM (2005) Genome sequence of a serotype M28 strain of group A *Streptococcus*: new insights into puerperal sepsis and bacterial disease specificity. *Journal of Infectious Diseases*, In Press.

Martin JM, Green M, Barbadora KA and Wald ER (2002) Erythromycin-resistant group A streptococci in children in Pittsburgh. *New England Journal of Medicine*, **346**, 1200–1206.

Nakagawa I, Kurokawa K, Yamashita A, Nakata M, Tomiyasu Y, Okahashi N, Kawabata S, Yamazaki K, Shiba T, Yasunaga T, *et al*. (2003) Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Research*, **13**, 1042–1045.

Reid SD, Hoe NP, Smoot LM and Musser JM (2001) Group A *Streptococcus*: allelic variation, population genetics, and host-pathogen interactions. *Journal of Clinical Investigation*, **107**, 393–399.

Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, Sturdevant DE, Ricklefs SM, Porcella SF, Parkins LD, *et al*. (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4668–4673.

Smoot LM, Smoot JC, Graham MR, Somerville GA, Sturdevant DE, Lux Migliaccio CA, Sylva GL and Musser JM (2001) Global differential gene expression in response to growth temperature alteration in group A *Streptococcus*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10416–10421.

Voyich JM, Braughton KR, Sturdevant DE, Vuong C, Kobayashi SD, Porcella SF, Otto M, Musser JM and DeLeo FR (2004) Engagement of the pathogen survival response used by group A *Streptococcus* to avert destruction by innate host defense. *Journal of Immunology*, **173**, 1194–1201.

Voyich JM, Sturdevant DE, Braughton KR, Kobayashi SD, Lei B, Virtaneva K, Dorward DL, Musser JM and DeLeo FR (2003) Genome-wide protective response used by Group A *Streptococcus* to evade destruction by human polymorphonuclear leukocytes. *Proceedings of the National Academy of Science USA*, **100**, 1996–2001.

Short Specialist Review

Yersinia

Nicholas R. Thomson
The Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

The genus *Yersinia* is composed of 11 species: *Y. pseudotuberculosis, Y, enterocolitica, Y. pestis, Y. intermedia, Y. kristensenii, Y. frederiksenii, Y. aldovae, Y. rohdei, Y. mollaretii, Y. bercovieri,* and *Y. ruckeri* (Sulakvelidze, 2000). Of the 11 species, only *Y. pseudotuberculosis, Y, enterocolitica*, and *Y. pestis* are pathogenic. There are several theories as to the evolution of these pathogenic strains but the favored notion is that they evolved from a nonpathogenic ancestor by the accretion of plasmids and chromosomally encoded genetic determinants.

Multilocus sequence analysis and DNA–DNA hybridization studies have shown that the pathogenic *Yersinia* are closely related: *Y. enterocolitica* and *Y. pseudotuberculosis* are thought to have diverged within the last 200 Myr and *Y. pestis* is a clone that has split from *Y. pseudotuberculosis* as recently as 1500 years ago and could in fact constitute a *Y. pseudotuberculosis* subspecies (Achtman *et al*., 1999; Bercovier *et al*., 1980).

Although the pathogenic *Yersinia* are genetically highly related, clinically they are clearly divided: *Y. pseudotuberculosis* and *Y. enterocolitica* are enteropathogens causing generally self-limiting gastroenteritis and infecting by the fecal-oral route. On the other hand, *Y. pestis* is primarily a rodent pathogen and is usually transmitted subcutaneously by the bite of an infected flea (principally *Xenopsylla cheopis*) and causes the often fatal bubonic plague, or upon infection of the lungs, causes pneumonic plague, in humans.

Thus, *Y. pestis* appears to be a species that has rapidly adapted from being a mammalian enteropathogen to an obligate bloodborne pathogen of mammals, which can also parasitize insects and utilize them as vectors for onward dissemination of infection (Achtman *et al*., 1999; Pepe and Miller, 1993). Historically, *Y. pestis* is thought to have been responsible for three human pandemics, the most infamous of which is The Black Death (fourteenth to nineteenth centuries). However, there is a current pandemic of plague that has seen recent outbreaks in Madagascar and Algeria and that claims an average of ∼2000 lives per annum (Titball and Williamson, 2001; WHO figures **http://www.who.int**).

Genetically it is clear that key evolutionary leaps in the emergence of these pathogenic *Yersinia* were made by the acquisition of a selection of virulence plasmids. All of the highly pathogenic yersiniae carry the 70-kb *Yersinia* virulence plasmid,

Table 1 [a]The completely sequenced *Yersinia* genomes and plasmids

	Y. pestis CO92	*Y. pestis* KIM	pMT1[b]	pMT1[c]	pMT1[c]	pCD1[b]	pCD13	pCD1[c]	pYVe227[d]	pPCP1[b]	pPCP1[c]
Size bp	4 653 728	4 600 755	96 210	10 0990	10 0984	70 305	70 559	70 504	69 673	9612	9610
Overall % G+C	47.64	47.64	50.96	50.16	50.16	44.84	44.81	44.81	44.21	45.27	45.28
% G+C of coding regions	48.9	48.86	50.95	51.02	51.59	46.28	44.88	46.19	45.54	57.03	44.41
Total no CDS	4016	4090	103	115	78	97	76	70	69	9	5
Coding percentage	80.3	83.3	82.8	89.5	68.4	76.5	64.6	64.8	69.5	57.2	44.1
RNA operons	6	7	0	0	0	0	0	0	0	0	0
tRNAs	70	73	0	0	0	0	0	0	0	0	0
Pseudo genes	149	54	3	0	0	8	0	6 (1)[e]	0	0	0
Accession number	AL590842	AE009952	AL117211	AF074611	AF053947	AL117189	AF074612	AF053946	AL102990	AL109969	AF053945

[a]Data taken from the EMBL database sequence files, see bottom row for accession numbers.
[b]Plasmid isolated from *Y. pestis* CO92.
[c]Plasmid isolated from *Y. pestis* KIM.
[d]Plasmid isolated from *Y. enterocolitica*.
[e]Number in parentheses represents partial sequence.

collectively known as the low calcium response (LCR) plasmid. Uniquely, *Y. pestis* also carries two other plasmids, pMT1 and pCD1 (discussed below; Table 1).

2. The *Yersinia* plasmids

The LCR plasmids from *Y. enterocolitica* (pYV), *Y. pestis* (pCD1), and *Y. pseudotuberculosis* (pIB1) all carry the highly characterized Yop virulon, which directs the production of a type III secretion system (TTSS), translocators, and secreted effector proteins, known as Yops (*Yersinia* outer proteins). The Yop virulon is essential for virulence of all the highly pathogenic yersinae in a variety of its hosts (for a review, see Brubaker, 1991 and Cornelis, 2002). The externally exposed portion of the TTSS apparatus appears as a needlelike structure, which penetrates the host cell membrane and facilitates the direct injection of the Yops from the bacterium into the eukaryotic host cell (Hoiczyk and Blobel, 2001). Once injected, Yops inhibit phagocytosis and block any proinflammatory response mounted by the host's immune cells. The organization of the Yop virulon is highly conserved amongst all the LCR plasmids, although there are some differences in the gross structure and organization of these plasmids, which can largely be explained by recombination and transposition associated with insertion sequence (IS) elements (Prentice *et al*., 2001; Hu *et al*., 1998).

The LCR plasmids also encode the *yadA* gene, which encodes an adhesin involved in mucus and epithelial cell attachment and invasion (Pepe and Miller, 1993; reviewed by El Tahir and Skurnik, 2001). The *yadA* gene is intact in the enteropathogenic *Yersinia,* but carries a frameshift mutation in *Y. pestis*. Moreover, complementation of the YadA phenotype in *Y. pestis* has been shown to result in a significant decrease in virulence by the subcutaneous infection route (Rosqvist *et al*., 1988).

Of the *Y. pestis* specific plasmids, three pMT1 plasmids (for murine toxin; also known as pFra (Fraction 1 antigen)) have been sequenced, ranging in size between 96 and 110 kb and predicted to encode between 78 and 115 genes (Parkhill *et al*., 2001b; Lindler *et al*., 1998; Hu *et al*., 1998) (Table 1). Like pYV, the variability in the size and gene order in different pMT1 plasmids is mainly a consequence of IS mediated recombination and/or deletion (Filippov *et al*., 1990; Prentice *et al*., 2001). Comparisons of pMT1 with other plasmids revealed that they share extensive regions of similarity with the cryptic *Salmonella enterica* serovar Typhi (*S*. Typhi) plasmid pHCM2 (>50% of pFra sharing >90% nucleotide identity with pHCM2; Prentice *et al*., 2001; Figure 1). This abnormally high level of sequence conservation between plasmids of *Y. pestis* and *S*. Typhi was thought to be indicative of recent horizontal exchange and led to the current notion that the immediate predecessor of *Y. pestis* acquired pMT1 following the coinfection of a common host, perhaps a rodent, with another enteropathogen, such as *S*. Typhi (Prentice *et al*., 2001).

Plasmid pMT1 carries several important virulence factors, the fraction 1 protective antigen and the *ymt* gene encoding the murine toxin. The fraction 1 antigen, while being strongly immunoprotective, is not a virulence factor (Friedlander *et al*.,

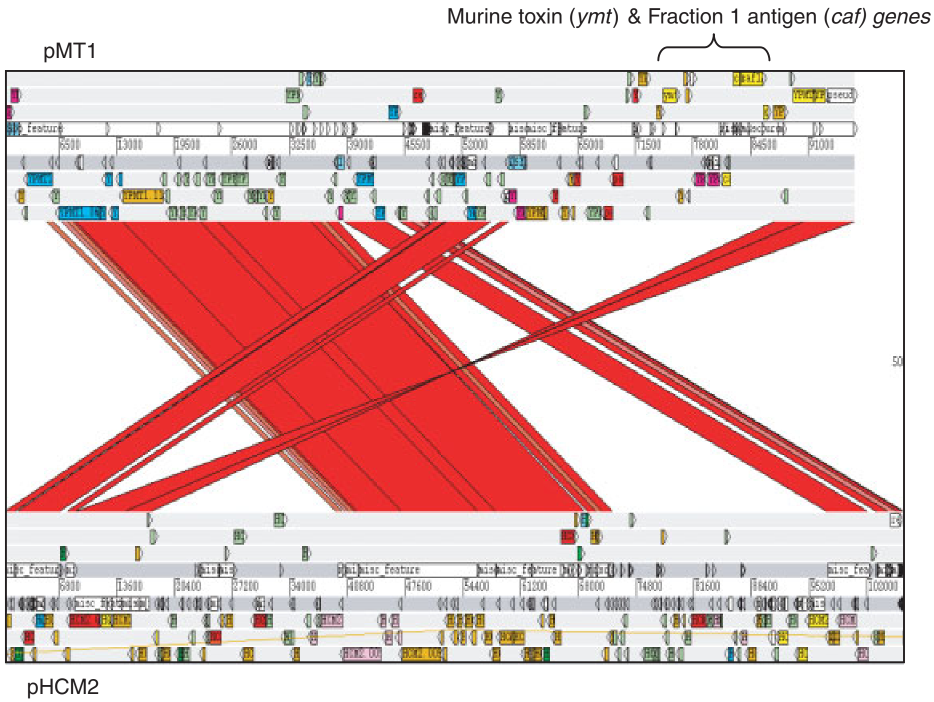


Figure 1 Global comparison between *Y. pestis* CO92 plasmid pMT1 and *S.* Typhi CT18 plasmid pHCM2. The figure shows DNA:DNA matches (computed using BLASTN and displayed using ACT **http://www.sanger.ac.uk/software/ACT**). The gray bars between the genomes represent individual BLASTN matches. Some of the shorter and weaker BLASTN matches have been removed to show the overall structure of the comparison

1995). Conversely, the *ymt* gene product is highly toxic in the mouse and rat models and was thought to be important for the high lethality of the plague in these hosts (Brown and Montie, 1977). However, isogenic *ymt* mutants do not significantly affect LD_{50} in mice. More recently, the gene product of *ymt* has been shown to have motifs in common with phospholipases and to be essential for *Y. pestis* to successfully colonize and survive within the flea midgut (Hinnebusch *et al*., 2000; Hinnebusch *et al*., 2002b).

At 9.6 kb, plasmid pPCP1 (also known as pPla and pPst) is the smallest of the three *Y. pestis* specific plasmids and predicted to encode 5/9 protein products (Parkhill *et al*., 2001b; Hu *et al*., 1998) (Table 1). These include pesticin (a pore-forming colicin) and pesticin immunity factor (Vollmer *et al*., 1997) as well as the plasminogen activator protein Pla. Pla is important for the systemic spread of plague within the mammalian host. It is a multifunctional protein, which gets its name from the ability to promote the maturation of the mammalian proenzyme plasminogen into plasmin (Beesley *et al*., 1967; Lahteenmaki *et al*., 1998; Lahteenmaki *et al*., 2001). The consequence of this is the bacterial-induced cleavage of host fibrin and extracellular matrices that limit the systemic dissemination of *Y. pestis* within the mammalian host. It has been suggested that pPCP1 was acquired after pMT1 and, since the midgut of the flea provides excellent opportunities for the acquisition of

horizontally transferred DNA (Hinnebusch *et al*., 2002a), this is where pPCP1 may have been acquired.

It has been known for sometime that while the pYV virulence plasmid of *Y. enterocolitica* is required for full virulence, there are other chromosomally located virulence factors that are also important (Heesemann and Laufs, 1984; Heesemann *et al*., 1984; Heesemann and Laufs, 1983). Some of these chromosomally encoded virulence determinants have been extensively studied such as *yst*, *inv*, and the hemin storage locus (see Revell and Miller, 2001 and references therein). However, compared to the depth of knowledge held for the plasmid encoded pathogenicity determinants, relatively little was known about the chromosome as a whole. This has been the impetus for the completed and ongoing whole genome sequencing projects.

3. The *Yersinia pestis* chromosome

Currently, two complete genome sequences are available: *Yersinia pestis* biovars Orientalis strain CO92 (CO92) (isolated from a fatal human case of primary pneumonic plague; Parkhill *et al*., 2001b) and Mediaevalis strain KIM10+ (KIM), (a genetically amenable laboratory strain; Deng *et al*., 2002). However, there are also sequencing projects for another biovar of *Y. pestis* biovar Mediaevalis strain 91001 (The Institute of Microbiology and Epidemiology [China] AE017042), as well as *Y. pseudotuberculosis* strain IP32953 (Lawrence Livermore (USA)/Institute Pasteur (France)) and *Y. enterocolitica* strain 8081 (Sanger Institute (UK)).

The genomes of CO92 and KIM are very similar in size consisting of a 4 600 755-bp and a 4 653 728-bp chromosome, respectively (Table 1). Whole-genome analysis of the two sequenced *Y. pestis* genomes revealed extensive evidence of recent intragenomic rearrangements, which clearly distinguished the two different *Y. pestis* biovars (Figure 2; Deng *et al*., 2002; Parkhill *et al*., 2001b).

Comparisons with more distant enterobacterial relatives revealed that much of the apparent colinearity preserved between, for example, *Escherichia coli* K12 and *S.* Typhi had been lost by these two yersiniae (Figure 2). However, if the relative distances of orthologous genes from the origins in *E. coli* and *Y. pestis* are compared, then there is a surprisingly high level of conservation. On the basis of these data, it was estimated that almost 50% of the core genes of KIM had been subject to interreplichore inversions during the evolution of the species (Deng *et al*., 2003). Many of these inversions are thought to be associated with recombination between IS elements. Although they are likely to have been important, the precise effects of these rearrangements on the biology and pathogenicity of *Y. pestis* remains unclear.

4. Genome contents

The genome contents of KIM and CO92 are also highly similar. Of the total genes predicted for each genome (4090 for KIM and 4016 for CO92), 3672 were common to both. The most significant differences in gene content between these genomes

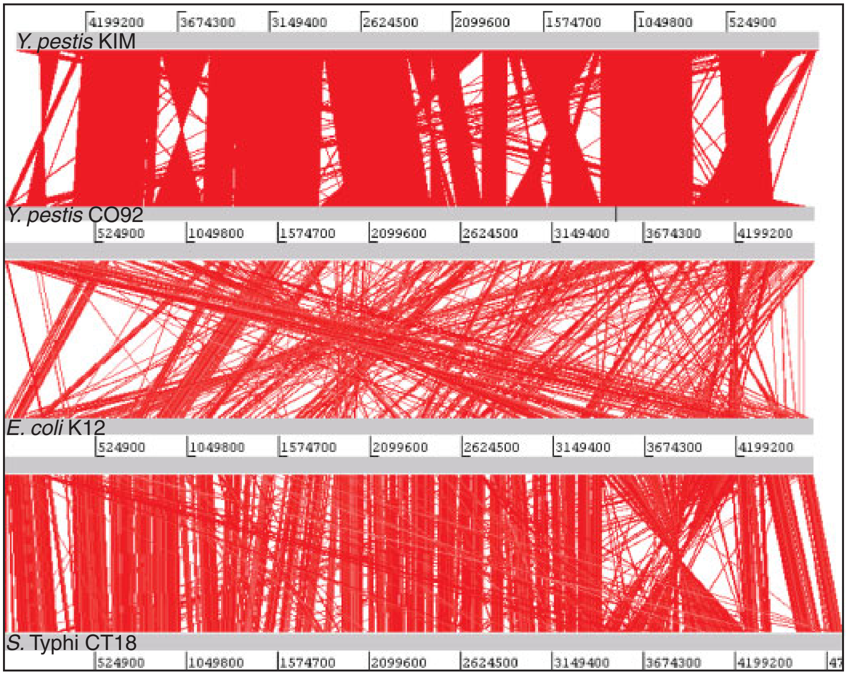


Figure 2 Global comparison between *Y. pestis* KIM, *Y. pestis* CO92, *E. coli* K12, and *S.* Typhi CT18. The figure shows six frame translated DNA:translated DNA matches (computed using TBLASTX and displayed using ACT **http://www.sanger.ac.uk/software/ACT**) pairwise between the four enterobacterial genomes. The genomes are, from the top down, *Y. pestis* KIM, *Y. pestis* CO92, *E. coli* K12, and *S.* Typhi CT18. The red bars between the genomes represent individual TBLASTX matches. Some of the shorter and weaker TBLASTX matches have been removed to show the overall structure of the comparison

included an extra rRNA operon in KIM and the loss of many flagella-related genes from KIM, many of which remain intact in CO92 (>80 genes). Other genomic differences were represented by an expanded population of IS elements in CO92 (140 vs. 122 complete or partial IS elements for CO92 and KIM, respectively) and an integrated prophage absent from KIM.

5. Pathogenicity islands/laterally acquired DNA

Pathogenicity Islands (PAI), first described by Hacker *et al.* (1997), are a phenomenon common to many enteric pathogens and the yersiniae are no exception. The genome sequences of the two *Y. pestis* biovars revealed a large number of regions displaying many of the characteristics of PAIs. These included those encoding genes that were involved in iron uptake as well as those encoding homologs of high–molecular weight insecticidal toxins and viral enhancins (Waterfield *et al.*, 2001). However, some of the *Y. pestis* insect toxin genes were found to carry frameshift mutations and so it is not clear whether these genes are simply vestiges of a former pathogenic association with an insect host.

In addition to the highly characterized plasmid-borne type III secretion system (Yop; see above), a second novel TTSS was discovered within a potential PAI on the *Y. pestis* genome. This novel system was remarkably similar to the TTSS located on *Salmonella* pathogenicity island 2 (SPI-2) and, like that of SPI-2, may act at different stages in the infection process to the plasmid-borne system.

6. Gene decay and genetic streamlining

The genomes of *Y. pestis* carry a large number of pseudogenes amounting to ∼4% of the total gene complement. The presence of significant numbers of pseudogenes and expansion of IS elements have been seen in other recently emerged pathogens (Cole *et al*., 2001; Andersson *et al*., 1998; Parkhill *et al*., 2001a; Parkhill *et al*., 2003) where genes required solely for the former lifestyle are lost in the process of phenotypic streamlining.

Most of the *Y. pestis* pseudogenes were associated with pathogenicity and/or predicted to encode surface-exposed proteins. Examples include the *iucA*, a gene essential for the production of aerobactin, several flagella biosynthetic genes, and many genes involved in lipopolysaccharide (LPS) biosynthesis important in *Y. enterocolitica* for resistance to complement-mediated and phagocyte killing (Darwin and Miller, 1999).

Mutations in energy metabolism and central and intermediary metabolism were underrepresented in the *Y. pestis* genome. Examples of pseudogenes in this class include those involved in glycerol fermentation. Genes *glpD*, *glpK*, and *glpX* are intact in biovar Medievalis (KIM) but have been subject to deletion events in biovar Orientalis (CO92), and this is the genetic explanation for the phenotype that is used to distinguish the two biovars (Parkhill *et al*., 2001b; Deng *et al*., 2002).

What is clear from the genomes of *Y. pestis* is that there has been significant gene loss as well as gain and that in combination with the extensive genome rearrangements observed, they are all likely to have been significant in the evolution of this acute pathogen. Combined with what we already know about the chromosome and extrachromosomal elements of the other *Yersinia*, it is clear that many of these determinants have been acquired incrementally during the evolution of this pathogen, possibly as an antecedent to speciation. The completion of the ongoing *Yersinia* genome projects should allow for a more complete analysis of the changes that occurred preceding the emergence of these pathogens, new and old.

The genomes of *Y. pestis* strain 91001 and *Y. pseudotuberculosis* strain IP32953 have now been published and offer further fascinating insights into the yersiniae (Chain *et al*., 2004 and Song *et al*., 2004).

References

Achtman M, Zurth K, Morelli C, Torrea G, Guiyoule A and Carniel E (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14043–14048.

Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson AS, Winkler HH and Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.

Beesley ED, Brubaker RR, Janssen WA and Surgalla MJ (1967) Pesticins .3. Expression of Coagulase and Mechanism of Fibrinolysis. *Journal of Bacteriology*, **94**, 19–26.

Bercovier H, Mollaret HH, Alonso JM, Brault J, Fanning GR, Steigerwalt AG and Brenner DJ (1980) Intraspecies and Interspecies Relatedness of *Yersinia-pestis* by DNA hybridization and its relationship to *Yersinia-pseudotuberculosis*. *Current Microbiology*, **4**, 225–229.

Brown SD and Montie TC (1977) Beta-Adrenergic Blocking Activity of Yersinia-Pestis Murine Toxin. *Infection and Immunity*, **18**, 85–93.

Brubaker RR (1991) Factors Promoting Acute and Chronic Diseases Caused by yersiniae. *Clinical Microbiology Reviews*, **4**, 309–324.

Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, *et al*. (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA* **101** 13826–13831.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.

Cornelis GR (2002) *Yersinia* type III secretion: send in the effectors. *Journal of Cell Biology*, **158**, 401–408.

Darwin AJ and Miller VL (1999) Identification of *Yersinia enterocolitica* genes affecting survival in an animal host using signature-tagged transposon mutagenesis. *Molecular Microbiology*, **32**, 51–62.

Deng W, Burland V, Plunkett G, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou SG, *et al*. (2002) Genome sequence of *Yersinia pestis* KIM. *Journal of Bacteriology*, **184**, 4601–4611.

Deng W, Liou SR, Plunkett G, Mayhew GF, Rose DJ, Burland V, Kodoyianni V, Schwartz DC and Blattner FR (2003) Comparative genomics of *Salmonella enterica* serovar *typhi* strains Ty2 and CT18. *Journal of Bacteriology*, **185**, 2330–2337.

El Tahir Y and Skurnik M (2001) YadA, the multifaceted *Yersinia adhesin*. *International Journal of Medical Microbiology*, **291**, 209–218.

Filippov AA, Solodovnikov NS, Kookleva LM and Protsenko OA (1990) Plasmid content in *Yersinia-pestis* strains of different origin. *Fems Microbiology Letters*, **67**, 45–48.

Friedlander AM, Welkos SL, Worsham PL, Andrews GP, Heath DG, Anderson GW, Pitt MLM, Estep J and Davis K (1995) Relationship between virulence and immunity as revealed in recent studies of the F1 capsule of *Yersinia-pestis*. *Clinical Infectious Diseases*, **21**, S178–S181.

Hacker J, BlumOehler G, Muhldorfer I and Tschape H (1997) Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. *Molecular Microbiology*, **23**, 1089–1097.

Heesemann J, Algermissen B and Laufs R (1984) Genetically manipulated virulence of *Yersinia-enterocolitica*. *Infection and Immunity*, **46**, 105–110.

Heesemann J and Laufs R (1983) Plasmid-Mediated Antigens of Human Pathogenic Yersinia-Enterocolitica Strains. *Zentralblatt Fur Bakteriologie Mikrobiologie Und Hygiene Series a-Medical Microbiology Infectious Diseases Virology Parasitology*, **253**, 428–429.

Heesemann J and Laufs R (1984) Genetic Manipulation of Virulence of Yersinia-*enterocolitica* and *Yersinia-pseudotuberculosis*. *Zentralblatt Fur Bakteriologie Mikrobiologie Und Hygiene Series a-Medical Microbiology Infectious Diseases Virology Parasitology*, **256**, 416–417.

Hinnebusch BJ, Rosso ML, Schwan TG and Carniel E (2002a) High-frequency conjugative transfer of antibiotic resistance genes to *Yersinia pestis* in the flea midgut. *Molecular Microbiology*, **46**, 349–354.

Hinnebusch BJ, Rudolph AE, Cherepanov P, Dixon JE, Schwan TG and Forsberg A (2002b) Role of Yersinia murine toxin in survival of *Yersinia pestis* in the midgut of the midgut of the flea vector. *Science*, **296**, 733–735.

Hinnebusch J, Cherepanov P, Du Y, Rudolph A, Dixon JD, Schwan T and Forsberg A (2000) Murine toxin of *Yersinia-pestis* shows phospholipase D activity but is not required for virulence in mice. *International Journal of Medical Microbiology*, **290**, 483–487.

Hoiczyk E and Blobel G (2001) Polymerization of a single protein of the pathogen *Yersinia enterocolitica* into needles punctures eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4669–4674.

Hu P, Elliott J, McCready P, Skowronski E, Garnes J, Kobayashi A, Brubaker RR and Garcia E (1998) Structural organization of virulence-associated plasmids of *Yersinia pestis*. *Journal of Bacteriology*, **180**, 5192–5202.

Lahteenmaki K, Kuusela P and Korhonen TK (2001) Bacterial plasminogen activators and receptors. *Fems Microbiology Reviews*, **25**, 531–552.

Lahteenmaki K, Virkola R, Saren A, Emody L and Korhonen TK (1998) Expression of plasminogen activator Pla of *Yersinia pestis* enhances bacterial attachment to the mammalian extracellurlar matrix. *Infection and Immunity*, **66**, 5755–5762.

Lindler LE, Plano GV, Burland V, Mayhew GF and Blattner FR (1998) Complete DNA sequence and detailed analysis of the *Yersinia pestis* KIM5 plasmid encoding murine toxin and capsular antigen. *Infection and Immunity*, **66**, 5731–5742.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MTG, *et al*. (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**, 848–852.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MTG, Churcher CM, Bentley SD, Mungall KL, *et al*. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genetics*, **35**, 32–40.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MTG, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, *et al*. (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.

Pepe JC and Miller VL (1993) *Yersinia-enterocolitica* Invasin - a Primary Role in the Initiation of Infection. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 6473–6477.

Prentice MB, James KD, Parkhill J, Baker SG, Stevens K, Simmonds MN, Mungall KL, Churcher C, Oyston PCF, Titball RW, *et al*. (2001) *Yersinia pestis* pFra shows biovar-specific differences and recent common ancestry with a *Salmonella enterica* serovar *typhi* plasmid. *Journal of Bacteriology*, **183**, 2586–2594.

Revell PA and Miller VL (2001) *Yersinia* virulence: more than a plasmid. *Fems Microbiology Letters*, **205**, 159–164.

Rosqvist R, Skurnik M and Wolfwatz H (1988) Increased Virulence of *Yersinia-pseudotuberculosis* by 2 Independent Mutations. *Nature*, **334**, 522–525.

Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H, *et al*. (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* **11**, 179–197.

Sulakvelidze A (2000) Yersiniae other than Y-enterocolitica, *Y. pseudotuberculosis*, and *Y. pestis*: the ignored species. *Microbes and Infection*, **2**, 497–513.

Titball RW and Williamson ED (2001) Vaccination against bubonic and pneumonic plague. *Vaccine*, **19**, 4175–4184.

Vollmer W, Pilsl H, Hantke K, Holtje JV and Braun V (1997) Pesticin displays muramidase activity. *Journal of Bacteriology*, **179**, 1580–1583.

Waterfield NR, Bowen DJ, Fetherston JD, Perry RD and ffrench-Constant RH (2001) The *tc* genes of *Photorhabdus*: a growing family. *Trends in Microbiology*, **9**, 185–191.

Short Specialist Review

Chlamydiae

Timothy D. Read
Biological Defense Research Directorate, Naval Medical Research Center, Rockville, MD, USA

1. Introduction

The chlamydiae are a distinct group of gram-negative bacteria restricted to growth within a specialized vacuole of eukaryotic cells. Chlamydiae have a biphasic life cycle, with a metabolically dormant, infectious "elementary body" (EB) and a replicative form (reticulate body; RB) that is restricted to an intracellular vacuole (Rockey and Matsumoto, 1999). Members of the genus *Chlamydiaceae* include important human and animal pathogens, notably *Chlamydia trachomatis*, a common agent of ocular trachoma and sexually transmitted disease, and *Chlamydia pneumoniae*, frequently a cause of community acquired lung infections in humans, which has also been linked to atherosclerosis. It is becoming apparent from 16 S RNA sequencing studies (*see* Article 6, **The genetic structure of human pathogens**, Volume 1) that the *Chlamydiaceae* are members of a very diverse class of organisms, the Chlamydiales, ubiquitous in nature (Ossewaarde and Meijer, 1999). Owing to the difficulties inherent in experimental work on an organism that must be cultured within eukaryotic cells and has no convenient system for genetics, the advent of genome sequencing has had an immense impact on understanding chlamydiae biology.

At present, there are six published complete chlamydiae sequence (Table 1). The three *C. pneumoniae* genomes are almost identical, with only a few hundred polymorphisms over 1.23 Mb genomes (Daugaard *et al*., 2001), reflecting the probable very recent worldwide clonal spread of the organism. *Chlamydia pneumoniae* and *Chlamydia caviae* are members of the *Chlamydophila* branch of Chlamydiaceae (Everett *et al*., 1999), approximately 200 kb larger than the Chlamydia genomes (Table 1). A conserved small plasmid (Thomas *et al*., 1997) was present in all genomes sequenced but *C. pneumoniae*. *Chlamydia pneumoniae* AR39 contained the replicative form of a bacteriophage of the microvirus family (Read *et al*., 2000).

2. Genome architecture

Chlamydiae genomes are reduced in size compared to those of free-living bacteria, which are typically > 2.5 Mb. However, there is no evidence of the recent genome

Table 1 Features of published chlamydiaceae genomes sequences (April 2004)

	C. trachomatis (serovar D)	*C. muridarum* (MoPn)	*C. pneumoniae* (AR39)[a]	*C. caviae* (GPIC)
Chromosome (nt)	1 042 519	1 072 950	1 229 858	1 173 390
Plasmids/*Phage (nt)	7493	7501	*4524	7996
%GC	41.3	40.3	40.6	39.2
Total annotated genes	895	921	1130	1009
tRNAs	37	37	38	38
rRNA operons	2	2	1	1
Accession numbers	AE001273	AE002160, AE002162	AR39: AE002161; J138: BA000008; CWL029: AE001363	AE015925.1, AE015925.1,
References	(Stephens *et al*., 1998)	(Read *et al*., 2000)	(Kalman *et al*., 1999; Read *et al*., 2000; Shirai *et al*., 2000)	(Read *et al*., 2003)

[a]Figures for strain AR39 are nearly identical to J138 and CWL029.

degradation seen as the accumulation of genomic repeats and pseudogenes in other obligate pathogens such as *Mycobacterium leprae* (*see* Article 52, **Genomics of the *Mycobacterium tuberculosis* complex and *Mycobacterium leprae***, Volume 4) or *Rickettsia prowazaki*. The completed chlamydiae genomes lack mobile genetics elements such as prophages and insertion sequences. There are no regions in the genome with unusually high or low percent GC sequence or atypical dinucleotide composition (*see* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4), which suggest the presence of pathogenicity islands (*see* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4), or other recent acquisition of genes by horizontal transfer. Differences between species appear to be localized to the region of the genome near the predicted DNA replication termination locus that has been termed the *plasticity zone*. While there are no true prophages, it does appear that portions of microvirus sequence have been integrated in this area of the *C. caviae* and *C. pneumoniae* genomes, presumably by illegitimate recombination. As in other bacteria, the replication origin and termination regions appear to be the focus of large symmetrical inversions (Suyama and Bork, 2001). Using these inversions as a molecular clock, it seems that there is an unusually high degree of nucleotide sequence divergence compared to other groups of bacteria. This may be due to an elevated mutation rate, although chlamydiae genomes appear to contain most common bacteria DNA repair systems.

3. Common functions in chlamydia genomes

Genome sequencing was a bounding leap forward in basic understanding of the chlamydial cell. Chlamydiae were found not to be strict “energy parasites” (Moulder, 1991), but instead encoded most of the enzymes for aerobic metabolism via glycolysis and tricarboxylic acid cycle (TCA) cycle (McClarty, 1999; Stephens *et al*., 1998). The partial TCA cycle can be rescued by uptake of glutamate from

the intracellular environment (McClarty, 1999). There are also genes encoding respiratory chain components for the reoxidation of reduced cofactors produced by the TCA cycle. The genomes contains V-ATPase complex enzyme genes found in eukaryotes and plastids rather than the usual F-type ATPases seen in other bacteria (McClarty, 1999). Chlamydiae are able to import ATP from the host cell via an ADP/ATP transporter also found in Rickettsiae and plant genomes. All genomes also contain a paralagous gene that acts a general NTP importer. Chlamydiae contain all the genes necessary for glycogen synthesis, which may be an important energy store for the dormant EB phase. Like several other intracellular pathogens, Chlamydiae have a reduced component of membrane transport proteins compared to free-living organisms (Paulsen *et al*., 2001). There is a preponderance of transporters involved in peptide and amino-acid import, reflecting the reduced capacity for de novo biosynthesis of these molecules.

Type III secretion system (*see* Article 49, **Bacterial pathogens of man**, Volume 4) genes are conserved in all chlamydiae genomes, reflecting their importance in interactions with the host. Aside from the previously identified major outer membrane protein, genome sequencing also revealed a complex multigene family of polymorphic outer membrane proteins (Pmps) that vary in number from 9 (*C. trachomatis*) to 21 (*C. pneumoniae*). These proteins have been localized to the surface of the RB and may play roles in adhesion and cell signaling.

Chlamydiae tend to be deeply separated from other bacteria in phylogenetic trees constructed with common conserved proteins. This reflects an enduring ecological and genetic isolation as an intracellular pathogen over many millions of years. The fact that many chlamydial proteins group with cyanobacterial, plastid, or plant sequences in evolutionary reconstructions suggests that the chlamydiae may have had their most common bacterial ancestor with the organisms that went on to form endosymbiotic relationships with plant cells (Brinkman *et al*., 2002).

4. Species and strain-specific genes

Despite their relative genome simplicity the chlamydiae cause a wide variety of different conditions over a broad range of vertebrate hosts. This may be surprising, considering two-thirds to three-quarters of the genes may be necessary for basic cellular propagation since they are conserved in all genomes (Read *et al*., 2003). The genes that are specific to individual species and strains (many of which are located in the plasticity zone) (Read *et al*., 2000) may offer clues about how tropisms occur. One constant theme in the sequenced genomes is a differing capacity for overcoming the almost complete lack of nucleotide biosynthetic genes by salvaging precursors from the cell environment. In this regard, there is strong evidence of lateral gene transfer of nucleotide salvage genes between the *Chlamydia* and *Chlamydophila* branches of the Chlamydiaceae (Read *et al*., 2003) Another theme is differences in capacity to catabolize the amino acid tryptophan: *C. caviae* has genes encoding an almost complete biosynthetic pathway, whereas *C. pneumoniae*, a highly successful human pathogen, has none (Read *et al*., 2000, 2003). The proinflammatory cytokine γ-interferon can lead to restriction of intracellular tryptophan levels. McClarty and colleagues have shown how even

the presence of a partial tryptophan biosynthesis cluster in *C. trachomatis* (Wood *et al*., 2003) may aid the cell in avoiding the effects of γ-interferon expression. Other notable strain-specific genes include a large toxin-like determinant, related in amino-acid sequence to an *Escherichia coli* O157:H7 protein (*see* Article 51, **Genomics of enterobacteriaceae**, Volume 4) (Read *et al*., 2000), found in *C. caviae* and *C. muridarum,* and a gene in *C. caviae* that encodes an invasin-like protein (Read *et al*., 2003). Some difference between strains may be effected by changes as subtle as gene duplication or single nucleotide polymorphisms (Belland *et al*., 2001).

5. Future directions

There are several chlamydiae genome projects ongoing, both within the *Chlamydiaceae* and also *Simkania* and *Parachlamydia*, genera that represent the vast undersampled diversity of the group. Sequencing enables functional genomics, for example, studying gene expression during cell differentiation using whole-genome microarrays (Nicholson *et al*., 2003) (*see* Article 94, **Expression and localization of proteins in mammalian cells**, Volume 4). Genome sequencing will likely be a driving force for discovery in the continuing absence of genetic methods in the group, and the future of chlamydiae research will be increasingly driven by questions discovered through various type of genomic analysis.

Related articles

Article 2, **Genome sequencing of microbial species**, Volume 3; Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4; Article 13, **Prokaryotic gene identification *in silico***, Volume 7; Article 45, **Phylogenomics for studies of microbial evolution**, Volume 7

References

Belland RJ, Scidmore MA, Crane DD, Hogan DM, Whitmire W, McClarty G and Caldwell HD (2001) *Chlamydia trachomatis* cytotoxicity associated with complete and partial cytotoxin genes. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13984–13989.

Brinkman FSL, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC, Finlay BB, Otto SP, Ouellette BFF, Keeling PJ, *et al*. (2002) Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between *Chlamydiaceae*, cyanobacteria, and the chloroplast. *Genome Research*, **12**, 1159–1167.

Daugaard L, Christiansen G and Birkelund S (2001) Characterization of a hypervariable region in the genome of *Chlamydophila pneumoniae*. *FEMS Microbiology Letters*, **203**, 241–248.

Everett KD, Bush RM and Andersen AA (1999) Emended description of the order Chlamydiales, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and

five new species, and standards for the identification of organisms. *International Journal of Systematic Bacteriology*, **49** (Pt 2), 415–440.

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW and Stephens RS (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genetics*, **21**, 385–389.

McClarty G (1999) Chlamydial metabolism as inferred from the complete genome sequence. In *Chlamydia: Intracellular Biology, Pathogenesis and Immunity*, Stephens RS (Ed.), American Society for Microbiology: Washington, pp. 69–100.

Moulder JW (1991) Interaction of chlamydiae and host cells in vitro. *Microbiological Reviews*, **55**, 143–190.

Nicholson TL, Olinger L, Chong K, Schoolnik G and Stephens RS (2003) Global stage-specific gene regulation during the developmental cycle of *Chlamydia trachomatis*. *Journal of Bacteriology*, **185**, 3179–3189.

Ossewaarde JM and Meijer A (1999) Molecular evidence for the existence of additional members of the order Chlamydiales. *Microbiology*, **145**, 411–417.

Paulsen IT, Chen J, Nelson KE and Saier MH Jr (2001) Comparative genomics of microbial drug efflux systems. *Journal of Molecular Microbiology and Biotechnology*, **3**, 145–150.

Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, *et al*. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research*, **28**, 1397–1406.

Read TD, Myers GS, Brunham RC, Nelson WC, Paulsen IT, Heidelberg J, Holtzapple E, Khouri H, Federova NB, Carty HA, *et al*. (2003) Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Research*, **31**, 2134–2147.

Rockey DD and Matsumoto A (1999) The chlamydial developmental cycle. In *Prokaryotic Development*, Brun YV and Shimkets LJ (Eds.), ASM Press: Washington, DC, pp. 403–425.

Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, *et al*. (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Research*, **28**, 2311–2314.

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, *et al*. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754–759.

Suyama M and Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in Genetics*, **17**, 10–13.

Thomas NS, Lusher M, Storey CC and Clarke IN (1997) Plasmid diversity in Chlamydia. *Microbiology*, **143**(Pt 6), 1847–1854.

Wood H, Fehlner-Gardner C, Berry J, Fischer E, Graham B, Hackstadt T, Roshick C and McClarty G (2003) Regulation of tryptophan synthase gene expression in *Chlamydia trachomatis*. *Molecular Microbiology*, **49**, 1347–1359.

Short Specialist Review

Spirochete genomes

George Weinstock
Baylor College of Medicine, Houston, TX, USA

Spirochetes are bacteria originally defined as a distinct group based on their distinct morphology. Spirochetes appear as long threadlike spiral-shaped organisms in the microscope. Subsequent classification based on sequence typing with rRNA sequences confirmed this grouping, making the spirochetes the only bacteria to be correctly classified on the basis of morphological criteria. Spirochetes can be highly motile but their flagellar apparatus is located in the periplasm, unlike most other bacteria that have externally located flagella. Their body shape is believed to help them maintain high motility even in viscous environments, where they can "corkscrew" through the medium. Spirochetes are ubiquitous in the environment as well as within hosts, both as commensals as well as pathogenic invaders. In humans, commensal spirochetes are readily found in the mouth, for example. Some of the more well known spirochetes are *Treponema pallidum*, causative agent of syphilis, and *Borrelia burgdorferi*, the bacterium that causes Lyme disease (*see* Article 49, **Bacterial pathogens of man**, Volume 4).

A number of spirochete genomes have been sequenced (*see* Article 6, **The genetic structure of human pathogens**, Volume 1) including *T. pallidum* (1.1 Mb) (Fraser *et al*., 1998), *B. burgdorferi* (1.2 Mb) (Fraser *et al*., 1997), *Treponema denticola* (2.8 Mb) (Seshadri *et al*., 2004), and *Leptospira interrogans* (4.6 Mb) (Ren *et al*., 2003; Nascimento *et al*., 2004), while numerous other spirochete genome projects are in progress. Spirochetes are a discrete branch of the Eubacterial tree (*see* Article 40, **The domains of life and their evolutionary implications**, Volume 7, Article 44, **Phylogenomic approaches to bacterial phylogeny**, Volume 7, and Article 45, **Phylogenomics for studies of microbial evolution**, Volume 7), and are distinct from many of the other disease-causing organisms that have been targets for genome sequencing, such as bacteria from the Proteobacteria or gram-positive bacteria branches. In addition, some spirochetes, such as *T. pallidum*, live in very specific niches and do not appear to have undergone much genetic exchange with other bacteria. Because of this separation, spirochete genomes tend to contain spirochete-specific genes (e.g., for virulence factors) in addition to core genes found in all bacteria, such as genes for the flow of genetic information or for metabolic pathways that are broadly distributed (e.g., glycolysis).

T. pallidum was one of the first genome projects undertaken, starting in 1991 with the publication of the sequence in 1998 (Fraser *et al*., 1998). Because this organism is one of the last of the major pathogens that cannot be cultured in

the laboratory (it must be grown in rabbit testes), the genome sequence was deemed the best way to understand the molecular biology of syphilis, and the impact of the genome project has been significant. The *T. pallidum* genome is a single circular DNA molecule. It does not contain any recognizable prophages, translocatable elements, plasmids, or other remnants of horizontal gene transfer. Of the 1031 predicted protein-coding genes, about 40% either have no database match or only match sequences of unknown function from other organisms. However, a number of treponeme-specific gene families were discovered from the genome sequence, and these appear to include genes coding for virulence factors of this organism. Moreover, the sequence showed that *T. pallidum* has a limited repertoire of metabolic pathways, a consequence of the limited biological niche in which it lives (it only lives in humans with no external reservoir), and which may contribute to the difficulty in culturing the organism outside of its host. As a follow-up to the sequencing of this genome, each predicted protein-coding gene was individually cloned in *Escherichia coli* to produce each protein in this surrogate host (McKevitt *et al*., 2003). This clone set was used to identify all antigenic proteins following infections of rabbits or humans. A total of 106 antigenic proteins were found, about 10% of the proteome, and these were highly enriched for proteins that were predicted to be exported from the cytoplasm.

The *T. pallidum* strain that was sequenced was the Nichols strain of *T. pallidum* subsp. *pallidum*, the syphilis causing subspecies of *T. pallidum*. Several other subspecies exist that cause distinct diseases. Like the syphilis strain, none of these other subspecies can be cultured in the laboratory, and DNA–DNA hybridization showed the genomes to be more than 95% identical. One of these, *T. pallidum* subsp. *pertenue* (Gauthier strain), causes the tropical disease yaws and has been compared to the syphilis strain in detail. While yaws is distinct from syphilis (it is an invasive skin disease that is not sexually transmitted), the differences in gene content of the two genomes is limited to only a few genes, with no major differences such as the presence or absence of pathogenicity islands. In addition, there are less than 300 single nucleotide polymorphisms, out of over one million bases, so the differences between these organisms, leading to their different disease phenotypes, appear to be quite subtle.

A second treponeme whose genome has been sequenced is *T. denticola* (Seshadri *et al*., 2004). *T. denticola* is a resident of the oral cavity and is associated with periodontal and gingival diseases. It is a component of the biofilm that forms on teeth and interacts with other genera of bacteria in the mouth. It is one of a number of spirochetes that colonize the mouth and appear to be a part of the normal human flora. *T. denticola*, while in the same genus as *T. pallidum*, is quite different at both the genotypic and phenotypic levels. Its genome size is nearly three times that of *T. pallidum* and its G+C content is much lower. *T. denticola* has a more substantial metabolic capability than *T. pallidum* and can be cultured in the laboratory. Its genome also shows considerable evidence of horizontal gene transfer and contains a high content of repeated sequences. Most notable about *T. denticola* is that it has the largest number of predicted transporters of any bacterial genome. This, plus its metabolic capabilities, presumably allows it to compete favorably with other organisms in a range of environments.

B. burgdorferi has one of the most unusual prokaryotic genomes with its major chromosome being a linear molecule complete with telomeric ends, as well as dozens of linear and circular plasmids, some of which contain genes to enhance growth and infection and which are therefore not simply accessories (Fraser *et al*., 1997; Casjens *et al*., 2000). The number of these plasmids varies somewhat from strain to strain. This remarkable genome must have a curious origin and a novel selection for its maintenance. One notable observation is the presence of homologs of the *E. coli recBCD* recombination genes in *B. burgdorferi*. This system has not been widely found among bacterial genomes that have been sequenced, although in *E. coli* and its relatives it is a major system for DNA repair and homologous recombination. The presence of this recombination system may imply a unique DNA metabolism in *B. burgdorferi* compared to other spirochetes.

Leptospira interrogans is a spirochete that is broadly found in the environment and in animals, including humans. It is the cause of the zoonotic infection leptospirosis. It is carried by a number of animals, and frequently colonizes rats without causing disease. Two strains of *L. interrogans* have been sequenced (Ren *et al*., 2003; Nascimento *et al*., 2004) and the reveal a complex genome as expected for an organism with such a broad biological niche.

References

Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ, *et al*. (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete Borrelia burgdorferi. *Molecular Microbiology*, **35**(3), 490–516.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, *et al*. (1997) Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. *Nature*, **390**(6660), 580–586.

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, *et al*. (1998) Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science*, **281**(5375), 375–388.

McKevitt M, Patel K, Smajs D, Marsh M, McLoughlin M, Norris SJ, Weinstock GM and Palzkill T (2003) Systematic cloning of Treponema pallidum open reading frames for protein expression and antigen discovery. *Genome Research*, **13**(7), 1665–1674.

Nascimento AL, Ko AI, Martins EA, Monteiro-Vitorello CB, Ho PL, Haake DA, Verjovski-Almeida S, Hartskeerl RA, Marques MV, Oliveira MC, *et al*. (2004) Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis. *Journal of Bacteriology*, **186**(7), 2164–2172.

Ren SX, Fu G, Jiang XG, Zeng R, Miao YG, Xu H, Zhang YX, Xiong H, Lu G, Lu LF, *et al*. (2003) Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature*, **422**(6934), 888–893.

Seshadri R, Myers GS, Tettelin H, Eisen JA, Heidelberg JF, Dodson RJ, Davidsen TM, DeBoy RT, Fouts DE, Haft DH, *et al*. (2004) Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(15), 5646–5651.

Short Specialist Review

Comparative genomics of the ε-proteobacterial human pathogens *Helicobacter pylori* and *Campylobacter jejuni*

Nick Dorrell and Brendan Wren
London School of Hygiene & Tropical Medicine, London, UK

1. Introduction

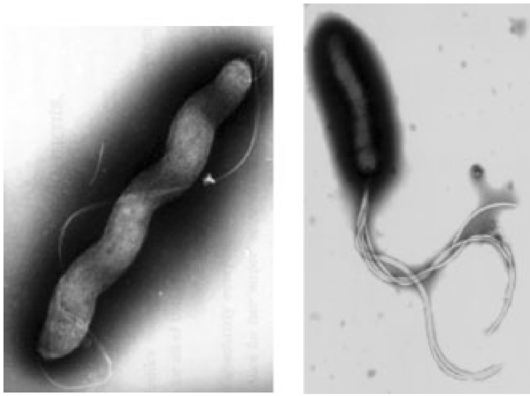
Helicobacter pylori and *Campylobacter jejuni* are among the most common bacterial pathogens encountered by humans. Infection with *H. pylori* can cause gastritis and ulceration, and it can be found in the stomach of half the world's population, whereas *C. jejuni* is the most frequently identified gastrointestinal pathogen and is one of the most common causes of diarrhoeal disease. Both pathogens can cause serious postinfection sequalae, with *H. pylori* being responsible for some forms of gastric cancer and *C. jejuni* causing neuromuscular diseases such as Guillian–Barré syndrome. Another curiosity is that despite the prevalence and importance of the diseases caused by *H. pylori* and *C. jejuni*, they have only been recognized as human pathogens within the last 25 years.

At the genetic, metabolic, and morphological levels, they share many common characteristics. *H. pylori* and *C. jejuni* are both members of the ε-proteobacteria subdivision of eubacteria. They are microaerophilic, spiral shaped, and motile, residing in the mucosa of their hosts (Figure 1). In fact, when *H. pylori* was first identified in 1983, it was classified in the *Campylobacter* genus. However, closer inspection confirms that they are from distinct genuses and have different primary hosts. *H. pylori* is a human-specific pathogen, whereas humans are an accidental part of the life cycle of *C. jejuni*, which are normally found as a commensal in the avian crop. *Campylobacter jejuni* can also survive in aquatic environments.

How can the availability of the genome sequences and subsequent genomic studies help explain the genetic, ecological, and pathogenic similarities and differences between these two major pathogens?

2. Common genome characteristics

H. pylori (26695) and *C. jejuni* (NCTC11168) were among the first microorganisms to be fully sequenced (Parkhill *et al.*, 2000; Tomb *et al.*, 1997), and



Campylobacter jejuni *Helicobacter pylori*

Figure 1 The rogues gallery – electron microscope pictures of the two organisms

indeed *H. pylori* was the first bacterium for which the genome of a second strain was determined (Alm *et al*., 1999). The 26695 and NCTC11168 genomes are of relatively small size and have similar GC content (1.67 Mb and 1.64 Mb and 39% and 30.6% respectively). Forty-eight percent of their predicted protein-coding sequences (CDSs) are orthologous (>30% similarity). These include CDSs involved in general housekeeping functions, metabolism, respiration, chemotaxis, and motility. Other similarities were the general lack of regulatory genes, including only three sigma factors and few two component regulatory systems. The CDSs are frequently found in linked groups transcribed in the same direction (Figure 2), but CDSs of an expected related function often appear to be scattered across the genome, with a general lack of operon structure (e.g., genes involved in flagella biogenesis). Close inspection of the nucleotide sequences of both genomes identified dozens of homopolymeric tracts or dinucleotide repeats characteristic of regions of slip strand mispairing. This results in phase variable proteins often involved in the biosynthesis of surface structures. Such repeats may be considered as a primordial mechanism that some mucosal pathogens with small genomes use to hugely vary their surface structure. Subsequent studies have confirmed that these phase variable genes are an important feature for the respective life cycles of both *H. pylori* and *C. jejuni* (Appelmelk *et al*., 1998; Linton *et al*., 2000).

3. What is in and what is out

The complete genome sequences of *H. pylori* and *C. jejuni* allow direct comparison of their genome complements and the opportunity to relate the presence and absence of sets of genes to the respective lifestyles of these pathogens. In the case of *H. pylori*, it is a human-specific pathogen that resides in the acidic environment of the stomach and interacts directly with gastric epithelial cells. These characteristics are reflected by the presence of the urease locus that allows the organism to survive at low pH and the *cag* pathogenicity island, encoding a type IV secretion system

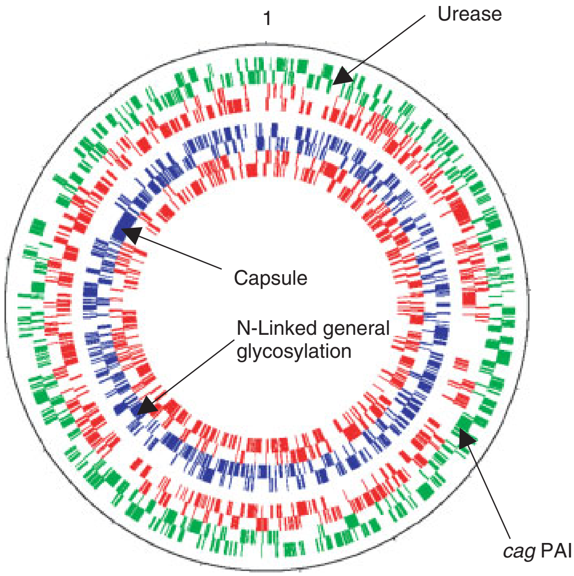


Figure 2 Genome comparison of *Helicobacter pylori* 26695 and *Campylobacter jejuni* NCTC11168. The outer four circles represent the *H. pylori*-predicted protein-coding sequences on both the plus and minus strands; green indicates *H. pylori* unique genes and red indicates genes shared with *C. jejuni*. The urease locus and *cag* pathogenicity island are highlighted. The inner four circles represent the *C. jejuni*-predicted protein-coding sequences on both the plus and minus strands; blue indicates *C. jejuni* unique genes and red indicates genes shared with *H. pylori*. The capsule biosynthesis locus and *N*-linked general glycosylation locus are highlighted

that is involved in interactions between *H. pylori* and gastric epithelial cells. These loci are absent from the *C. jejuni* NCTC11168 genome (Figure 2). By contrast, the *C. jejuni* genome has a capsule locus and an *N*-linked general glycosylation pathway, both features absent in *H. pylori* 26695 (Figure 2) and unknown prior to the commencement of the genome project. Subsequently, a noncapsulated mutant was shown to have reduced ability to adhere to and invade intestinal epithelial cells and also reduced virulence in the ferret model of diarrhoeal disease (Bacon *et al*., 2001). The *N*-linked general glycosylation pathway modifies over 30 *C. jejuni* proteins, with a heptasaccharide (Young *et al*., 2002). The purpose of this modification is unknown, but it is speculated that it may be important in suppressing the immune response in avians to maintain its commensal status. *Campylobacter jejuni* appears to possess a greater metabolic and regulatory versatility than *H. pylori*, probably reflecting the very restricted niche of the human stomach where the latter is found and the more diverse environments in which the former can grow (Kelly, 2001).

4. Postgenome studies

Because *H. pylori* and *C. jejuni* were among the first organisms to be sequenced, they have been well studied in the postgenome **era**. Further representatives of these

species have been, or are being sequenced. Whole-genome comparison studies using microarrays (Article 98, **Bacterial genome organization: comparative expression profiling, operons, regulons, and beyond**, Volume 4) have shown both species to be diverse with less than 80% of the respective genomes representing functional core or species-specific genes (Dorrell *et al*., 2001; Salama *et al*., 2000). Several transcriptome studies (*see* Article 93, **Microarray CGH**, Volume 4) have been performed examining the organisms under a range of stress- and host-related environments. Proteome maps are available for both organisms, and a comprehensive protein–protein interaction map has been published for *H. pylori* (Rain *et al*., 2001).

5. Further sequenced ε-proteobacteria

More recently, further members of the ε-proteobacteria have been sequenced. *Helicobacter hepaticus* is most similar to *H. pylori,* but crucially lacks the *Cag* pathogenicity island and the urease operon, as may be expected, as *H. hepaticus* does not reside in the human stomach but in the small intestine (Suerbaum *et al*., 2003). In fact, some *H. hepaticus* metabolic genes most closely resemble those in *C. jejuni*, suggesting some form of environmental adaptation as both organisms reside in the intestine. *Wollinella succinogenes* represents a halfway house between *H. pylori* and *C. jejuni*, with some features from each of the organisms (Baar *et al*., 2003). By contrast to *H. pylori* and *C. jejuni*, *W. succinogenes* has an enormous range of regulatory genes, with the highest percentage of genes in any genome coding for histidine kinases or their cognate regulators. This feature may reflect the broad range of environments in which this bacterium is known to survive.

6. Conclusions

Ironically, although *H. pylori* and *C. jejuni* are relatively recently identified human pathogens, they are now among the most comprehensively studied. Helicobacteriologists and campylobacteriologists have truly benefited from the postgenome bonanza and this has increased our understanding of the dynamic and complex interplay between these pathogens and their respective hosts. The emphasis now must be to convert this new-found knowledge into appropriate intervention strategies to reduce the burden of these two problematic pathogens on human health.

References

Alm RA, Ling LSL, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, *et al*. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.

Appelmelk BJ, Shiberu B, Trinks C, Tapsi N, Zheng PY, Verboom T, Maaskant J, Hokke CH, Schiphorst WECM, Blanchard D, *et al*. (1998) Phase variation in *Helicobacter pylori* lipopolysaccharide. *Infection and Immunity*, **66**, 70–76.

Baar C, Eppinger M, Raddatz G, Simon J, Lanz C, Klimmek O, Nandakumar R, Gross R, Rosinus A, Keller H, *et al*. (2003) Complete genome sequence and analysis of *Wolinella succinogenes*.

Proceedings of the National Academy of Sciences of the United States of America, **100**, 11690–11695.

Bacon DJ, Szymanski CM, Burr DH, Silver RP, Alm RA and Guerry P (2001) A phase-variable capsule is involved in virulence of *Campylobacter jejuni* 81-176. *Molecular Microbiology*, **40**, 769–777.

Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV, *et al*. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Research*, **11**, 1706–1715.

Kelly DJ (2001) The physiology and metabolism of *Campylobacter jejuni* and *Helicobacter pylori*. *Journal of Applied Microbiology*, **90**, 16 S–24 S.

Linton D, Gilbert M, Hitchen PG, Dell A, Morris HR, Wakarchuk WW, Gregson NA and Wren BW (2000) Phase variation of a beta-1, 3 galactosyltransferase involved in generation of the ganglioside GM1-like lipo-oligosaccharide of *Campylobacter jejuni*. *Molecular Microbiology*, **37**, 501–514.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, *et al*. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.

Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L and Falkow S (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14668–14673.

Suerbaum S, Josenhans C, Sterzenbach T, Drescher B, Brandt P, Bell M, Droge M, Fartmann B, Fischer HP, Ge Z, *et al*. (2003) The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7901–7906.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, *et al*. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.

Young NM, Brisson JR, Kelly J, Watson DC, Tessier L, Lanthier PH, Jarrell HC, Cadotte N, St Michael F, Aberg E, *et al*. (2002) Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, *Campylobacter jejuni*. *Journal of Biological Chemistry*, **277**, 42530–42539.

Short Specialist Review

The neisserial genomes: what they reveal about the diversity and behavior of these species

Nigel J. Saunders and Lori A. S. Snyder
University of Oxford, Oxford, UK

The human pathogens *Neisseria meningitidis* and *Neisseria gonorrhoeae* are closely related organisms, with high sequence identity (typically greater than 95%) between their common genes, yet the diseases they cause are vastly different. *N. meningitidis* normally exists as a harmless commensal in the human nasopharynx, but when it becomes invasive it causes bacterial meningitis and severe septicaemia, which are both life threatening even with prompt and appropriate antibiotic treatment. *N. gonorrhoeae*, on the other hand, almost always causes disease following colonization, although gonorrhoea infections can be asymptomatic, particularly in women. Gonococcal infection is normally limited to epithelial surfaces but can ascend the female reproductive tract, leading to fertility-threatening pelvic inflammatory disease, and can also cause blindness following eye infection of vaginally delivered infants. *N. gonorrhoeae* can, untypically, cause disseminated infections including arthritis and a septicaemic infection, but these are much less severe than the disseminated meningococcal infections.

Since a genome sequence is a single time-point snapshot of a subculture, of a strain, of a species of bacteria, it is inherently an example of a bacterial system rather than a representative of the whole species. Fortunately, as more related bacterial genomes are sequenced, this provides a wider context in which to interpret these individual snapshots. With multiple genome sequences, comparative analyses are possible, which can be usefully linked to existing experimentally derived information for the species. In this way, a much deeper and more complete picture of a species' evolutionary and functional behavior can be perceived. To date, there are two complete and published (Parkhill *et al*., 2000; Tettelin *et al*., 2000) and one complete and publicly available (Sanger Institute) *N. meningitidis* genome sequences, and one *N. gonorrhoeae* genome sequence (currently unpublished from ACGT-OU), which together are far more informative than any single sequence could be alone.

Because of their medical importance, the *Neisseria* spp. have been studied intensely since the discovery of the gonococcus toward the end of the nineteenth Century (Neisser, 1879). The results of these studies can be drawn upon, tested against, and reconsidered in the light of the complete genome sequences. The genome

sequences, therefore, in the context of the other information available, provide an important framework for understanding their biology from a whole-system perspective, and are key to the design and construction of future experiments.

Considering the available information as a whole, and the features that are apparent from study of the genome sequences, the most striking general feature of these species is that, above all, they are capable of great change and flexibility. Although it may initially seem counterintuitive, the "static" genome sequences provide many indications of the "dynamic" nature of these species. They are models of a very different paradigm of bacterial evolution and function to those presented by the "classical model bacterial systems" such as *Escherichia coli* and *Bacillus subtilis* in that they have evolved a very different strategy for adapting to changing environmental conditions founded upon rapid evolutionary principals rather than upon the stable maintenance of highly specialized, tightly regulated, systems. However, this should not lead to a conclusion that they are untypical of bacterial systems as a whole. Indeed, they may be usefully considered representative of many other species with similar evolutionary and adaptive strategies, including, for example, *Haemophilus*, *Helicobacter*, *Campylobacter*, and *Bordetella* species groups.

The *Neisseria* spp. are known to have relatively panmictic population structures owing to frequent intraspecies recombination events (Smith *et al*., 1993; Feil *et al*., 1999; Holmes *et al*., 1999) facilitated by a common uptake signal sequence (Goodman and Scocca, 1988; Elkins *et al*., 1991). Despite this context, genome-level comparisons of the gene complements generated some initially surprising results. The strain MC58 and Z2491 meningococcal genome sequences are approximately as divergent from one another, as the meningococcal strain MC58 is from gonococcal strain FA1090. This variability in the gene complements of the genome sequences makes the task of identifying the genes that differentiate the meningococci and their behavior and host interactions from the gonococci a large and complex task.

On the basis of the pool of currently annotated features, each neisserial genome contains between 67 and 183 unique genes that are not present in the other three genome sequences. Comparison of the gene complements of all of the meningococcal genome sequences with the single gonococcal sequence identifies 645 differences in the presence of genes, but some of these differences are specific to the gonococcal strain used in genome sequencing, as shown by comparisons with the other main experimental strains using microarray-based comparative genome hybridization (Snyder *et al*., 2004).

Probably the longest recognized differentiating characteristic between *N. meningitidis* and *N. gonorrhoeae* is the presence of the meningococcal capsule. The genes responsible for the possession of a polysaccharide capsule by the meningococci are believed to have been horizontally acquired either after or as part of the speciation split between these pathogens. Differences between the type of capsule produced are due to gene complement differences at the capsule gene locus, a locus that is devoid of all capsule-associated genes in the gonococcus, in the nonpathogenic *N. lactamica*, and some noninvasive strains of *N. meningitidis* (Claus *et al*., 2002; Dolan-Livengood *et al*., 2003). However, the differences between these species are clearly far deeper and more complex, and it should not be assumed that the gonococcus would behave similarly to the meningococcus, simply following acquisition

of a capsule, or that an acapsulate meningococcus behaves in a gonococcal fashion. Getting to the heart of this subject is likely to depend upon extensive comparative studies addressing many strains, and should be greatly assisted by the other neisserial sequencing projects currently underway, particularly that of *N. lactamica* (Sanger Institute).

Some conserved genes flank other genes, or groups of genes, that differ between the sequenced strains. These genes have been recognized to be mobile, and exist in different combinations in different strains. These locations have been defined as "Minimal Mobile Elements" (MMEs) and are characterized as sites in which strain-specific genes are preferentially located between flanking genes with conserved sequence and chromosomal organization, such that these flanking regions can serve as substrates for homologous recombination following natural transformation. The first of the regions to be studied in an extended set of strains, between the *pheS* and *pheT* genes of the *Neisseria* spp., contains seven different genes or combinations of genes between the conserved flanking sequences (Saunders and Snyder, 2002). There are many more such regions that are currently under investigation and this will certainly lead to the identification of new neisserial genes, and to a better picture of strain differentiating characteristics. This is a good example of how the study of genomes from a limited number of strains can be extended into analysis of the wider bacterial population.

The capacity for adaptive change mediated by changes at the DNA level in these species resides within individual strains as well as being a product of genetic exchange between different organisms. The neisserial silent and expressed pilus system is well known and has been a teaching paradigm for many years (Haas *et al*., 1992). Two additional silent and expressed cassette systems have been proposed following assessments of the genome sequences, which reveal the way in which the study of whole-genome sequences can give rapid new insights into bacterial processes. These new systems are proposed to affect the *mafB* and *fhaB* genes, although these are still subject to experimental investigation for functional confirmation (Klee *et al*., 2000; Parkhill *et al*., 2000).

Another source of flexibility is achieved through the variable expression of the many phase-variable genes in these species (Saunders *et al*., 2000; Snyder *et al*., 2001). Phase variation is typically associated with genes involved in environmental change (Salaün *et al*., 2003; Saunders, 2003), and the preponderance of these genes in these species are associated with mediating direct interactions with the host, and immune evasion. The analysis of the complete genome sequences led to the prediction of the complete repertoires of phase variable genes in these species, which are amongst the largest so far seen in any species, providing the capacity to generate vast numbers of combinations of expressed and unexpressed genes. An additional insight gained from using a comparative genome analyses methodology for the identification of these genes was that not only are the genes' phase variably expressed but also the presence of the genes frequently differs between strains, as does their potential to phase-vary. So, while two strains may possess the same gene, this gene's expression can be mediated by phase variation in one but not the other, and in a third strain the gene may be missing entirely. A very similar pattern is seen in *H. pylori*, in which species this has been explored in more detail (Salaün *et al*., 2004).

The short-motif simple sequence repeats mediating phase variation have been recognized for some time to generate diversity within the population through changes in length in this species. Comparisons of the genes containing longer repeats for which changes in length would be associated with changes in the composition of the genes containing them, rather than leading to ON–OFF switching, suggests that this is also important in generating diversity in this species. Comparison of the genomes was used to identify a number of genes containing coding tandem repeats, which were then pursued in a diverse strain collection to determine whether these were a source of diversity. In total, 22 such genes were identified within the *Neisseria* spp., of which 16 were demonstrated to display different numbers of repeats between different strains of the same species (Jordan *et al*., 2003). This is another example of the way in which genome analysis can be extended from "index" sequenced strains to the wider population.

Differences between the strains are not limited to coding sequences, although the functional consequences of other differences are usually far harder to interpret. Two intergenic sequences that appear to be unique to the *Neisseria* spp., although not restricted to the pathogenic species, are the neisserial uptake signal sequence (Goodman and Scocca, 1988; Elkins *et al*., 1991) and the Correia repeat and associated Correia Repeat Enclosed Elements (Correia *et al*., 1986; Correia *et al*., 1988; Liu *et al*., 2002). While a single complete neisserial genome allows the identification of the locations of these sorts of elements, the comparison of multiple genomes has allowed the differences in their locations and functions to be assessed as well (Liu *et al*., 2002; Snyder *et al*., 2003).

While genomes are now an essential context for the identification of new candidate genes for classical gene-by-gene investigations and for the design of constructs and experiments, they also provide the basis for whole-system studies of bacterial behavior and function. The strain MC58 genome sequencing project was unusual in that it represented a highly interactive three-way collaboration between a company (Chiron), an academic research group (in Oxford University), and a sequencing centre (TIGR). The results of this interaction were reflected not only in the publication of the genome sequence itself (Tettelin *et al*., 2000), but also of an extensive study of all of the predicted surface proteins as vaccine candidates (Pizza *et al*., 2000). This group, as well as several others, have developed microarray tools for the investigation of cellular functions and behavior under different infection-related conditions (Grifantini *et al*., 2002a,b; Dietrich *et al*., 2003; Grifantini *et al*., 2003; Kurz *et al*., 2003). Progressively, these tools are becoming routine components of the study of these pathogens, just as they are of others, and the essential platform that the genome sequences and high-quality annotations represent to these endeavors has to be considered for a proper appreciation of the impact of the genome sequences in this field. Similarly, as proteomic approaches to these pathogens develop, this too will depend upon the genome sequences and their annotations.

References

Claus H, Maiden MC, Maag R, Frosch M and Vogel U (2002) Many carried meningococci lack the genes required for capsule synthesis and transport. *Microbiology (Reading, England)*, **148**, 1813–1819.

Correia FF, Inouye S and Inouye M (1986) A 26-base-pair repetitive sequence specific for *Neisseria gonorrhoeae* and *Neisseria meningitidis* genomic DNA. *Journal of Bacteriology*, **167**, 1009–1015.

Correia FF, Inouye S and Inouye M (1988) A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *The Journal of Biological Chemistry*, **263**, 12194–12198.

Dietrich G, Kurz S, Hubner C, Aepinus C, Theiss S, Guckenberger M, Panzner U, Weber J and Frosch M (2003) Transcriptome analysis of *Neisseria meningitidis* during infection. *Journal of Bacteriology*, **185**, 155–164.

Dolan-Livengood JM, Miller YK, Martin LE, Urwin R and Stephens DS (2003) Genetic basis for nongroupable *Neisseria meningitidis*. *The Journal of Infectious Diseases*, **187**, 1616–1628.

Elkins C, Thomas CE, Seifert HS and Sparling PF (1991) Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. *Journal of Bacteriology*, **173**, 3911–3913.

Feil EJ, Maiden MC, Achtman M and Spratt BG (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution*, **16**, 1496–1502.

Goodman SD and Scocca JJ (1988) Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 6982–6986.

Grifantini R, Bartolini E, Muzzi A, Draghi M, Frigimelica E, Berger J, Randazzo F and Grandi G (2002a) Gene expression profile in *Neisseria meningitidis* and *Neisseria lactamica* upon host-cell contact: from basic research to vaccine development. *Annals of the New York Academy of Sciences*, **975**, 202–216.

Grifantini R, Bartolini E, Muzzi A, Draghi M, Frigimelica E, Berger J, Ratti G, Petracca R, Galli G, Agnusdei M, *et al*. (2002b) Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. *Nature Biotechnology*, **20**, 914–921.

Grifantini R, Sebastian S, Frigimelica E, Draghi M, Bartolini E, Muzzi A, Rappuoli R, Grandi G and Genco CA (2003) Identification of iron-activated and -repressed fur-dependent genes by transcriptome analysis of *Neisseria meningitidis* group B. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9542–9547.

Haas R, Veit S and Meyer TF (1992) Silent pilin genes of *Neisseria gonorrhoeae* MS11 and the occurrence of related hypervariant sequences among other gonococcal isolates. *Molecular Microbiology*, **6**, 197–208.

Holmes EC, Urwin R and Maiden MC (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Molecular Biology and Evolution*, **16**, 741–749.

Jordan P, Snyder LAS and Saunders NJ (2003) Diversity in coding tandem repeats in related *Neisseria* spp. *BMC Microbiology*, **3**, 23.

Klee SR, Nassif X, Kusecek B, Merker P, Beretti JL, Achtman M and Tinsley CR (2000) Molecular and biological analysis of eight genetic islands that distinguish *Neisseria meningitidis* from the closely related pathogen *Neisseria gonorrhoeae*. *Infection and Immunity*, **68**, 2082–2095.

Kurz S, Hubner C, Aepinus C, Theiss S, Guckenberger M, Panzner U, Weber J, Frosch M and Dietrich G (2003) Transcriptome-based antigen identification for *Neisseria meningitidis*. *Vaccine*, **21**, 768–775.

Liu SV, Saunders NJ, Jeffries A and Rest RF (2002) Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic *Neisseria*. *Journal of Bacteriology*, **184**, 6163–6173.

Neisser A (1879) Ueber eine der gonorrhoe eigentümliche micrcoccenform. *Centralblatt Für die medicinischen Wissenschaften*, **17**, 497–500.

Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, *et al*. (2000) Complete DNA sequence of a serogroup a strain of *Neisseria meningitidis* Z2491. *Nature*, **404**, 502–506.

Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, *et al*. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, **287**, 1816–1820.

Salaün L, Linz B, Suerbaum S and Saunders NJ (2004) The diversity within an expanded and re-defined repertoire of phase variable genes in *Helicobacter pylori*. *Microbiology (Reading, England)*, **150**, 817–830.

Salaün L, Snyder LAS and Saunders NJ (2003) Adaptation by phase variation in pathogenic bacteria. *Advances in Applied Microbiology*, **52**, 263–301.

Saunders NJ (2003) Evasion of antibody responses: Bacterial phase variation. *Advances in Molecular and Cellular Microbiology*, **2**, 103–124.

Saunders NJ, Jeffries AC, Peden JF, Hood DW, Tettelin H, Rappuoli R and Moxon ER (2000) Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Molecular Microbiology*, **37**, 207–215.

Saunders NJ and Snyder LAS (2002) The minimal mobile element. *Microbiology (Reading, England)*, **148**, 3756–3760.

Smith JM, Smith NH, O'Rourke M and Spratt BG (1993) How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 4384–4388.

Snyder LAS, Butcher SA and Saunders NJ (2001) Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp. *Microbiology (Reading, England)*, **147**, 2321–2332.

Snyder LAS, Davies JK and Saunders NJ (2004) Microarray genomotyping of key experimental strains of *Neisseria gonorrhoeae* reveals gene complement diversity and five new neisserial genes associated with minimal mobile elements. *BMC Genomics*, **5**, 23.

Snyder LAS, Shafer WM and Saunders NJ (2003) Divergence and transcriptional analysis of the division cell wall (*dcw*) gene cluster in *Neisseria* spp. *Molecular Microbiology*, **47**, 431–442.

Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, *et al*. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.

Short Specialist Review

Kinetoplastid genomics

Chris Peacock and Christiane Hertz-Fowler
The Wellcome Trust Sanger Institute, Cambridge, UK

The Kinetoplastida are a group of flagellated protozoa characterized by the kinetoplast, a specialized extranuclear DNA contained within the single mitochondrion. As an order, they are very versatile at adapting to their environment, existing both as free-living organisms through to obligate parasites of plants, insects, fish, reptiles, birds, mammals, and humans. Kinetoplastida are divided morphologically into two distinct suborders. The Bodonidae contain parasitic, ectocommensal, and free-living species, while the members of the Trypanosomatidae suborder are all obligate parasites. Research focuses predominately on members of the Trypanosomatidae, in particular, the human and veterinary pathogens within the genera *Leishmania* and *Trypanosoma* and the "model" Kinetoplastida within the genera *Crithidia* and *Leptomonas*. The diverse pathology caused by these parasites causes morbidity and mortality to millions of people and domestic livestock each year, with hundreds of millions at risk in approximately 90 countries (WHO reports).

The Kinetoplastida are one of the earliest diverging eukaryotic organisms (Sogin *et al*., 1986), encoding genes of bacterial and possible plant origin within their genomes (Hannaert *et al*., 2003b; Couvreur *et al*., 2002; Krepinsky *et al*., 2001; Sinha *et al*., 1999). They exhibit many classical eukaryotic pathways, yet have adapted the means to flourish in their differing host environments, using some processes that are either unique to the Kinetoplastida or that were initially elucidated in this group and subsequently discovered in other eukaryotes (e.g., RNA editing, trans-splicing of mRNAs, glycosylphosphatidylinositol anchoring of proteins (Ferguson, 1999), compartmentalization of energy metabolism (Hannaert *et al*., 2003a)). Although *Trypanosoma* and *Leishmania* spp share many aspects of "basic biology", they differ fundamentally in terms of their pathology and the hostile niches they inhabit within their hosts and vectors.

Three large-scale genome projects, initiated in the mid-1990s under the auspice of multicenter genome networks, have resulted in the near complete sequence and annotation of the 36 chromosomes of *Leishmania major* (reference strain MHOM/IL/80/Friedlin) genome, the 11-Mb chromosomes of *Trypanosoma brucei* (TREU927/4) and up to 20 chromosomal bands of *Trypanosoma cruzi* (CL Brener strain) by an international consortium (see Table 1 for further details). The genome projects have been underpinned by detailed karyotype analyses as well as genetic and physical maps (Tait *et al*., 2002; Melville *et al*., 2000; Ivens *et al*., 1998; Santos *et al*., 1997; Henriksson *et al*., 1995). They have also greatly benefited from

the large volume of other genomic data publicly available (Table 1). This varies from individually characterized genes submitted to the public databases, through to more extensive projects sequencing expressed sequence tags (ESTs), genome survey sequences (GSS), or cosmid and BAC clones (Aguero *et al*., 2000; Verdun *et al*., 1998; Levick *et al*., 1996; El-Sayed and Donelson, 1997; El-Sayed *et al*., 1995). More recently, whole genome shotgun projects of additional *Leishmania* and *Trypanosoma* species have been undertaken (Table 1).

Initial studies have shown significant physical differences between the genome architecture of the three major pathogenic groups sequenced so far. Whereas *L. major* exhibits little chromosome length polymorphisms between the homologs of each of the diploid chromosomes (Ivens *et al*., 1998), *T. brucei* (Melville *et al*., 2000) and *T. cruzi* (Porcile *et al*., 2003) show extensive homolog variations. As a consequence of such polymorphisms, haploid genome contents have been challenging to assemble, particularly in *T. cruzi*. Such genome plasticity is thought to be a result of not only the existence of mobile genetic elements in both *T. brucei* and *T. cruzi* but also of the expansion of both gene families, such as the retrotransposon hot spot proteins in the subtelomeric regions of chromosomes, and the presence of repetitive sequences (Bringaud *et al*., 2004; Hall *et al*., 2003; Wickstead *et al*., 2003; Bhattacharya *et al*., 2002; Bringaud *et al*., 2002a; Bringaud *et al*., 2002b). In the case of *T. brucei*, it has recently emerged that RNA interference, a mechanism apparently absent in both *L. major* and *T. cruzi*, is involved in the regulation of retroposon transcript abundance and thus of genome integrity (Ullu *et al*., 2004; Shi *et al*., 2004).

Even prior to their completion, the individual sequencing projects have not only reaffirmed previous experimental observations on a larger scale but revealed insights into interesting kinetoplastid biology. The data published from the three genomes so far (McDonagh *et al*., 2000; Worthey *et al*., 2003; Hall *et al*., 2003; El-Sayed *et al*., 2003; Andersson *et al*., 1998; Ghedin *et al*., 2004) and, in the case of *L. major*, confirmed by nuclear run-on analyses (Monnerat *et al*., 2004; Martinez-Calvillo *et al*., 2003; Martinez-Calvillo *et al*., 2004) are concurrent with our understanding of transcription and translation in these organisms. Genes are arranged in highly compact blocks on the same strand with small intergenic regions separating one gene from the next. Transcription takes place in long polycistronic units with gene regulation being predominantly controlled posttranscriptionally (reviewed by Campbell *et al*., 2003; Clayton, 2002). However, there is little evidence of genes being clustered on the basis of either related function or similar expression levels (Hall *et al*., 2003; El-Sayed *et al*., 2003).

Trans-splicing of a highly conserved 39 nucleotide sequence – the spliced leader – onto the 5′ end of all mRNAs (Parsons *et al*., 1986; Kooter *et al*., 1984) occurs cotranscriptionally with the simultaneous polyadenylation of the upstream transcript (Sutton and Boothroyd, 1986; LeBowitz *et al*., 1993), generating monocistronic transcripts. The spliced leader sequence is encoded in a single array (Roberts *et al*., 1996), the true extent of which will be unraveled by the genome projects. Cis-splicing is rare, with a single instance of an intron in both *T. brucei* and *T. cruzi* having so far been published (Mair *et al*., 2000). Despite the biochemical characterization of the three classical RNA polymerases in Trypanosomatids, no sequences with RNA polymerase II promoter activity transcribing protein coding

Table 1 List of kinetoplastid resources available via the Web

Organism	Resource	Website
Leishmania spp	GeneDB database	**www.genedb.org/genedb/leish**
	WTSI *L. major* ftp site	**ftp://ftp.sanger.ac.uk/pub/databases/L.major_sequences**
	WTSI *L. infantum* ftp site	**ftp://ftp.sanger.ac.uk/pub/pathogens/L_infantum/**
	SBRI sequencing project page	**http://apps.sbri.org/genome/lmjf/Lmjf.aspx**
	WTSI *L. major* project pages	**http://www.sanger.ac.uk/Projects/L_major/index.shtml**
	WTSI *L. infantum* project pages	**http://www.sanger.ac.uk/Projects/L_infantum/**
	Leishmania genome network	**http://www.ebi.ac.uk/parasites/leish.html**
	Genome survey sequences (GSS)	**http://genome.wustl.edu/est/index.php?leishmania=1**
	TIGR gene index (EST database)	**http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=leishmania**
	Minicircle database	**http://www.ebi.ac.uk/parasites/kDNA/P2aleish.html**
	L. major proteomics database (2D gel)	**http://www.cri.crchul.ulaval.ca/proteome/Proteome.htm**
Trypanosoma brucei	GeneDB database	**http://www.genedb.org/genedb/tryp/index.jsp**
	TIGR TbGAD database	**http://www.tigr.org/tdb/e2k1/tba1/index.shtml**
	TIGR ftp site	**http://www.tigr.org/tigr-scripts/license/new.pl?genre=euk**
	WTSI ftp site	**ftp://ftp.sanger.ac.uk/pub/databases/T.brucei_sequences/**
	TIGR *T. brucei* project pages	**http://www.tigr.org/tdb/e2k1/tba1/**
	WTSI *T. brucei* project pages	**http://www.sanger.ac.uk/Projects/T_brucei/**
	Trypanosoma brucei genome network	**http://parsun1.path.cam.ac.uk/**
	TIGR Gene index (EST database)	**http://www.tigr.org/tdb/tgi/tbgi/**
	U-insertion/deletion Edited Sequence db	**http://164.67.60.203/trypanosome/database.html**
	Guide RNA database	**http://biosun.bio.tu-darmstadt.de/goringer/gRNA/gRNA.html**
	Minicircle database	**http://www.ebi.ac.uk/parasites/kDNA/P2atryp.html**
	Trypanosome VSG database	**http://leishman.cent.gla.ac.uk/pward001/vsgdb/about.html**
	TrypanoFAN functional genomics	**http://www.trypanofan.org**
Trypanosoma vivax	GeneDB database	**http://www.genedb.org/genedb/tvivax/index.jsp**
	WTSI ftp site	**ftp://ftp.sanger.ac.uk/pub/databases/T.vivax_sequences/**
	WTSI *T. vivax* project pages	**http://www.sanger.ac.uk/Projects/T_vivax/**
Trypanosoma congolense	WTSI ftp site	**ftp://ftp.sanger.ac.uk/pub/databases/T.congolense_sequences/**
	WTSI *T. congolense* project pages	**http://www.sanger.ac.uk/Projects/T_congolense/**
Trypanosoma cruzi	GeneDB database	**http://www.genedb.org/genedb/tcruzi/index.jsp**
	TIGR *T. cruzi* database	**http://www.tigr.org/tdb/e2k1/tca1/index.shtml**
	TcruziDB database	**http://tcruzidb.org/**
	TIGR ftp site	**http://www.tigr.org/tigr-scripts/license/new.pl?genre=euk**
	Karolinska Institute project pages	**http://web.cgb.ki.se/**

(*continued overleaf*)

Table 1 (*continued*)

Organism	Resource	Website
	SBRI project pages	**http://apps.sbri.org/genome/Tcruzi/TCruziIndex.aspx**
	TIGR project pages	**http://www.tigr.org/tdb/e2k1/tca1/**
	T.cruzi genome initiative (FioCruz)	**http://www.dbbm.fiocruz.br/TcruziDB/index.html**
	TIGR gene index (EST database)	**http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=t_cruzi**
	T. cruzi EST project	**http://www.genpat.uu.se/tryp/tryp.html**
General links	Structural genomics data resource	**http://depts.washington.edu/sgpp/**
	EMBL nucleotide sequence database	**http://www.ebi.ac.uk/embl/**
	GenBank nucleotide sequence database	**http://www.ncbi.nlm.nih.gov/Genbank/index.html**
	Uniprot protein database	**http://www.ebi.uniprot.org/index.shtml**
	Protein structure definition and functional annotation at EOL	**http://www.eolproject.org:8080/index.jsp**

Abbreviations: SBRI: Seattle Biomedical Institute; TIGR: The Institute for Genomic Research; WTSI: Wellcome Trust Sanger Institute.

genes have been either experimentally characterized or identified computationally – although there is some evidence that the strand switch region between two polycistronic regions appears to be essential at least in *L. major* (Dubessay *et al*., 2002).

Extensive research has also focused on the content and organization of the telomeric and subtelomeric regions, particularly in *T. cruzi* and *T. brucei*, where genes mediating host immune system evasion and modulation are encoded. *T. brucei* periodically switches expression of variant surface glycoproteins (VSGs), a process termed *antigenic variation* (recently reviewed by Barry *et al*., 2003; McCulloch, 2004). VSGs are encoded on all three classes of chromosomes and transcribed from telomeric expression sites located on megabase and intermediate sized chromosomes. Monoallelic expression of one VSG gene is ensured by recruitment of a single expression site into a subnuclear body (Navarro and Gull, 2001). It is now emerging that VSGs are present as silent arrays in chromosome-internal locations, apparently predominantly as pseudogenes (El-Sayed *et al*., 2003) and require recombination events to form functional VSG genes. The structure of up to six expression sites and the organization of a small proportion of VSG gene repertoire have been described (Berriman *et al*., 2002). However, the completed genome project and ensuing work will provide further insight into how VSG diversity is maintained and activated during antigenic variation.

T. cruzi, despite being an obligate intracellular parasite, also expresses families of surface antigens, some of which are encoded at the telomeres (Chiurillo *et al*., 1999).

In contrast, *Leishmania* utilizes the host's complement pathway to gain entry to, and reside in, cells of the monocyte lineage, a process that requires stage specific differential expression of surface expressed genes (reviewed by Matlashewski, 2001).

In contrast to the nuclear genome, deciphering the mitochondrial DNA, consisting of a few dozen maxicircle and thousands of minicircle structures, has not been part of the genome projects. The sequences of these minicircles and maxicircles were determined in an effort to elucidate the mechanism of uridine insertion/deletion RNA editing of the mitochondrial mRNA transcripts, a process originally described in Trypanosomatidae (reviewed by Simpson *et al*., 2003; Estevez and Simpson, 1999).

The sequencing of the genomes of these organisms is only the start of a long process of using this wealth of data to complement what is already known about their complex biology, eventually aiding in the development of new ways to combat the effects of the pathogenic members of this group. It is very noticeable that almost all the frontline drugs used against these organisms have not only been in use for decades (leading to drug resistance) but also have detrimental toxic effects on the patients. There are as yet no commercial vaccines for any of these parasites and vector and reservoir control programs are at best maintaining the status quo. Comparative genomics will help to identify genes and regulatory sequences that are unique to Kinetoplastida and unique to individual species. Selecting species for sequencing that differ in specific aspects of pathogenicity, host tropism, or survival strategy allows candidate genes to be identified that may encode those specific differences. These genes may provide leads for further functional experimentation. Already the striking conservation of gene function and order is becoming apparent within this group, as are the regions of divergence (Ghedin *et al*., 2004; Bringaud *et al*., 1998).

Large-scale functional genomics projects are essential for turning sequence and annotation into information with practical implications. Data are already publicly available from large-scale 2D gel proteomic (Drummelsmith *et al*., 2003) and RNAi (TrypanoFAN, **http://trypanofan.path.cam.ac.uk/cgi-bin/WebObjects/trypanofan**) projects, complemented by emerging information from array technologies investigating host/pathogen interaction (Mukherjee *et al*., 2003) as well as expression patterns of the pathogen alone (Akopyants *et al*., 2004; Almeida *et al*., 2004; Saxena *et al*., 2003; Diehl *et al*., 2002; Minning *et al*., 2003). Ultimately, the availability of so much genomic data in the public domain should hasten the identification of new candidates for drug targets, vaccines, and diagnostic tools.

References

Aguero F, Verdun RE, Frasch AC and Sanchez DO (2000) A random sequencing approach for the analysis of the Trypanosoma cruzi genome: general structure, large gene and repetitive DNA families, and gene discovery. *Genome Research*, **10**, 1996–2005.

Akopyants NS, Matlib RS, Bukanova EN, Smeds MR, Brownstein BH, Stormo GD and Beverley SM (2004) Expression profiling using random genomic DNA microarrays identifies differentially expressed genes associated with three major developmental stages of the protozoan parasite Leishmania major. *Molecular and Biochemical Parasitology*, **136**, 71–86.

Almeida R, Gilmartin BJ, McCann SH, Norrish A, Ivens AC, Lawson D, Levick MP, Smith DF, Dyall SD, Vetrie D, *et al*. (2004) Expression profiling of the Leishmania life cycle: cDNA arrays identify developmentally regulated genes present but not annotated in the genome. *Molecular and Biochemical Parasitology*, **136**, 87–100.

Andersson B, Aslund L, Tammi M, Tran AN, Hoheisel JD and Pettersson U (1998) Complete sequence of a 93.4-kb contig from chromosome 3 of Trypanosoma cruzi containing a strand-switch region. *Genome Research*, **8**, 809–816.

Barry JD, Ginger ML, Burton P and McCulloch R (2003) Why are parasite contingency genes often associated with telomeres? *International Journal for Parasitology*, **33**, 29–45.

Berriman M, Hall N, Sheader K, Bringaud F, Tiwari B, Isobe T, Bowman S, Corton C, Clark L, Cross GA, *et al*. (2002) The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, **122**, 131–140.

Bhattacharya S, Bakre A and Bhattacharya A (2002) Mobile genetic elements in protozoan parasites. *Journal of Genetics*, **81**, 73–86.

Bringaud F, Biteau N, Melville SE, Hez S, El-Sayed NM, Leech V, Berriman M, Hall N, Donelson JE and Baltz T (2002a) A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of Trypanosoma brucei. *Eukaryotic Cell*, **1**, 137–151.

Bringaud F, Biteau N, Zuiderwijk E, Berriman M, El-Sayed NM, Ghedin E, Melville SE, Hall N and Baltz T (2004) The ingi and RIME non-LTR retrotransposons are not randomly distributed in the genome of Trypanosoma brucei. *Molecular Biology and Evolution*, **21**, 520–528.

Bringaud F, Garcia-Perez JL, Heras SR, Ghedin E, El-Sayed NM, Andersson B, Baltz T and Lopez MC (2002b) Identification of non-autonomous non-LTR retrotransposons in the genome of Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, **124**, 73–78.

Bringaud F, Vedrenne C, Cuvillier A, Parzy D, Baltz D, Tetaud E, Pays E, Venegas J, Merlin G and Baltz T (1998) Conserved organization of genes in trypanosomatids. *Molecular and Biochemical Parasitology*, **94**, 249–264.

Campbell DA, Thomas S and Sturm NR (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes and Infection / Institut Pasteur*, **5**, 1231–1240.

Chiurillo MA, Cano I, Da Silveira JF and Ramirez JL (1999) Organization of telomeric and subtelomeric regions of chromosomes from the protozoan parasite Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, **100**, 173–183.

Clayton CE (2002) Life without transcriptional control? From fly to man and back agin. *The EMBO Journal*, **21**, 1881–1888.

Couvreur B, Wattiez R, Bollen A, Falmagne P, Le Ray D and Dujardin JC (2002) Eubacterial HslV and HslU subunits homologs in primordial eukaryotes. *Molecular Biology and Evolution*, **19**, 2110–2117.

Diehl S, Diehl F, El-Sayed NM, Clayton C and Hoheisel JD (2002) Analysis of stage-specific gene expression in the bloodstream and the procyclic form of Trypanosoma brucei using a genomic DNA-microarray. *Molecular and Biochemical Parasitology*, **123**, 115–123.

Drummelsmith J, Brochu V, Girard I, Messier N and Ouellette M (2003) Proteome mapping of the protozoan parasite leishmania and application to the study of drug targets and resistance mechanisms. *Molecular and Cellular Proteomics*, **2**, 146–155.

Dubessay P, Ravel C, Bastien P, Crobu L, Dedet JP, Pages M and Blaineau C (2002) The switch region on Leishmania major chromosome 1 is not required for mitotich stability or gene expression, but appears to be essential. *Nucleic Acids Research*, **30**, 3692–3697.

El-Sayed NM, Alarcon CM, Beck JC, Sheffield VC and Donelson JE (1995) cDNA expressed sequence tags of Trypanosoma brucei rhodesiense provide new insights into the biology of the parasite. *Molecular and Biochemical Parasitology*, **73**, 75–90.

El-Sayed NM and Donelson JE (1997) A survey of the Trypanosoma brucei rhodesiense genome using shotgun sequencing. *Molecular and Biochemical Parasitology*, **84**, 167–178.

El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, Wanless D, Peterson J, Hou L, Taylor S, *et al*. (2003) The sequence and analysis of Trypanosoma brucei chromosome II. *Nucleic Acids Research*, **31**, 4856–4863.

Estevez AM and Simpson L (1999) Uridine insertion/deletion RNA editing in trypanosome mitochondria-a review. *Gene*, **240**, 247–260.

Ferguson MA (1999) The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. *Journal of Cell Science*, **112**(Pt 17), 2799–2809.

Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, Andersson B, Bontempi E, Eisen J, Angiuoli S, *et al*. (2004) Gene synteny and evolution of genome architecture in trypanosomatids. *Molecular and Biochemical Parasitology*, **134**, 183–191.

Hall N, Berriman M, Lennard NJ, Harris BR, Hertz-Fowler C, Bart-Delabesse EN, Gerrard CS, Atkin RJ, Barron AJ, Bowman S, *et al*. (2003) The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism. *Nucleic Acids Research*, **31**, 4864–4873.

Hannaert V, Bringaud F, Opperdoes FR and Michels PA (2003a) Evolution of energy metabolism and its compartmentation in Kinetoplastida. *Kinetoplastid Biology and Disease*, **2**, 11.

Hannaert V, Saavedra E, Duffieux F, Szikora JP, Rigden DJ, Michels PA and Opperdoes FR (2003b) Plant-like traits associated with metabolism of Trypanosoma parasites. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1067–1071.

Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, Cazzulo JJ, Frasch AC and Pettersson U (1995) Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, **73**, 63–74.

Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chan HM and Smith DF (1998) A physical map of the Leishmania major Friedlin genome. *Genome Research*, **8**, 135–145.

Kooter JM, De Lange T and Borst P (1984) Discontinuous synthesis of mRNA in trypanosomes. *The EMBO Journal*, **3**, 2387–2392.

Krepinsky K, Plaumann M, Martin W and Schnarrenberger C (2001) Purification and cloning of chloroplast 6-phosphogluconate dehydrogenase from spinach. Cyanobacterial genes for chloroplast and cytosolic isoenzymes encoded in eukaryotic chromosomes. *European Journal of Biochemistry / FEBS*, **268**, 2678–2686.

LeBowitz JH, Smith HQ, Rusche L and Beverley SM (1993) Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes and Development*, **7**, 996–1007.

Levick MP, Blackwell JM, Connor V, Coulson RM, Miles A, Smith HE, Wan KL and Ajioka JW (1996) An expressed sequence tag analysis of a full-length, spliced-leader cDNA library from Leishmania major promastigotes. *Molecular and Biochemical Parasitology*, **76**, 345–348.

Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, *et al*. (2000) A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA*, **6**, 163–169.

Martinez-Calvillo S, Nguyen D, Stuart K and Myler PJ (2004) Transcription initiation and termination on Leishmania major chromosome 3. *Eukaryotic Cell*, **3**, 506–517.

Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K and Myler PJ (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Molecular Cell*, **11**, 1291–1299.

Matlashewski G (2001) Leishmania infection and virulence. *Medical Microbiology and Immunology (Berlin)*, **190**, 37–42.

McCulloch R (2004) Antigenic variation in African trypanosomes: monitoring progress. *Trends in Parasitology*, **20**, 117–121.

McDonagh PD, Myler PJ and Stuart K (2000) The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes. *Nucleic Acids Research*, **28**, 2800–2803.

Melville SE, Leech V, Navarro M and Cross GA (2000) The molecular karyotype of the megabase chromosomes of Trypanosoma brucei stock 427. *Molecular and Biochemical Parasitology*, **111**, 261–273.

Minning TA, Bua J, Garcia GA, McGraw RA and Tarleton RL (2003) Microarray profiling of gene expression during trypomastigote to amastigote transition in Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, **131**, 55–64.

Monnerat S, Martinez-Calvillo S, Worthey E, Myler PJ, Stuart KD and Fasel N (2004) Genomic organization and gene expression in a chromosomal region of Leishmania major. *Molecular and Biochemical Parasitology*, **134**, 233–243.

Mukherjee S, Belbin TJ, Spray DC, Iacobas DA, Weiss LM, Kitsis RN, Wittner M, Jelicks LA, Scherer PE, Ding A, *et al*. (2003) Microarray analysis of changes in gene expression in a murine model of chronic chagasic cardiomyopathy. *Parasitology Research*, **91**, 187–196.

Navarro M and Gull K (2001) A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei. *Nature*, **414**, 759–763.

Parsons M, Nelson RG and Agabian N (1986) The trypanosome spliced leader small RNA gene family: stage-specific modification of one of several similar dispersed genes. *Nucleic Acids Research*, **14**, 1703–1718.

Porcile PE, Santos MR, Souza RT, Verbisck NV, Brandao A, Urmenyi T, Silva R, Rondinelli E, Lorenzi H, Levin MJ, *et al*. (2003) A refined molecular karyotype for the reference strain of the Trypanosoma cruzi genome project (clone CL Brener) by assignment of chromosome markers. *Gene*, **308**, 53–65.

Roberts TG, Dungan JM, Watkins KP and Agabian N (1996) The SLA RNA gene of Trypanosoma brucei is organized in a tandem array which encodes several small RNAs. *Molecular and Biochemical Parasitology*, **83**, 163–174.

Santos MR, Cano MI, Schijman A, Lorenzi H, Vazquez M, Levin MJ, Ramirez JL, Brandao A, Degrave WM and da Silveira JF (1997) The Trypanosoma cruzi genome project: nuclear karyotype and gene mapping of clone CL Brener. *Memorias do Instituto Oswaldo Cruz*, **92**, 821–828.

Saxena A, Worthey EA, Yan S, Leland A, Stuart KD and Myler PJ (2003) Evaluation of differential gene expression in Leishmania major Friedlin procyclics and metacyclics using DNA microarray analysis. *Molecular and Biochemical Parasitology*, **129**, 103–114.

Shi H, Djikeng A, Tschudi C and Ullu E (2004) Argonaute protein in the early divergent eukaryote Trypanosome brucei: control of small interfering RNA accumulation and retroposon transcript abundance. *Molecular and Cellular Biology*, **24**, 420–427.

Simpson L, Sbicego S and Aphasizhev R (2003) Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business. *RNA*, **9**, 265–276.

Sinha KM, Ghosh M, Das I and Datta AK (1999) Molecular cloning and expression of adenosine kinase from Leishmania donovani: identification of unconventional P-loop motif. *The Biochemical Journal*, **339**(Pt 3), 667–673.

Sogin ML, Elwood HJ and Gunderson JH (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 1383–1387.

Sutton RE and Boothroyd JC (1986) Evidence for trans splicing in trypanosomes. *Cell*, **47**, 527–535.

Tait A, Masiga D, Ouma J, MacLeod A, Sasse J, Melville S, Lindegard G, McIntosh A and Turner M (2002) Genetic analysis of phenotype in Trypanosoma brucei: a classical approach to potentially complex traits. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **357**, 89–99.

Ullu E, Tschudi C and Chakraborty T (2004) RNA interference in protozoan parasites. *Cellular Microbiology*, **6**, 509–519.

Verdun RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC and Sanchez DO (1998) Gene discovery through expressed sequence Tag sequencing in Trypanosoma cruzi. *Infection and Immunity*, **66**, 5393–5398.

Wickstead B, Ersfeld K and Gull K (2003) Repetitive elements in genomes of parasitic protozoa. *Microbiology and Molecular Biology Reviews*, **67**, 360–375. Table of contents.

Worthey EA, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, Fazelinia G, Fong C, Fu G, Hassebrock M, Hixson G, *et al*. (2003) Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene. *Nucleic Acids Research*, **31**, 4201–4210.

World Health Organization websites:
http://www.who.int/mediacentre/factsheets/fs259/en/, 2001
http://www.who.int/health_topics/chagas_disease/en/,
http://www.who.int/health_topics/leishmaniasis/en/.

Short Specialist Review

The organelles of apicomplexan parasites

James W. Ajioka and Elizabeth T. Brooke-Powell
University of Cambridge, Cambridge, UK

Kiew-Lian Wan
Universiti Kebangsaan Malaysia, Bangi, Selangor DE, Malaysia

1. Introduction

All eukaryotic cells are chimeric insofar as two major organelles, the mitochondrion and chloroplast, are products of primary endosymbiotic events with an alpha-proteobacterium and cyanobacterium respectively. The evolutionary history of the apicomplexa represents a further degree of complexity as a result of a secondary endosymbiotic event between a eukaryotic cell and a photosynthetic algae (Fast *et al*., 2001; Foth and McFadden, 2003). The mitochondrial and apicoplast genomes are very small, 6 kb linear and 35 kb circular respectively, compared to their counterparts in other species, each encoding a minimum of proteins (Feagin, 1992; Roos *et al*., 2002). The reduced nature of these genomes may be accounted for by both loss of genes and transfer of genes to the nuclear genome (*see* Article 54, **The nuclear genome of apicomplexan parasites**, Volume 4; Huang *et al*., 2004; Foth and McFadden, 2003).

2. The mitochondrial genome

With the exception of *Cryptosporidium* spp. (Abrahamsen *et al*., 2004), apicomplexans carry a single mitochondrion that maintains a membrane potential and retains some electron transport proteins (Srivastava *et al*., 1997; Vercesi *et al*., 1998). Evidence varies between species as to its ability to produce ATP via oxidative phosphorylation (see, for example, Srivastava *et al*., 1997; Vercesi *et al*., 1998). Alternatively, the organelle's main function may be to remove electrons generated by dihyrdroorotate dehydrogenase in the de novo synthesis of pyrimidines (Gero *et al*., 1984; Prapunwattana *et al*., 1988). Although only the mitochondrial genomes of *Plasmodium* species have been well characterized, DNA sequence

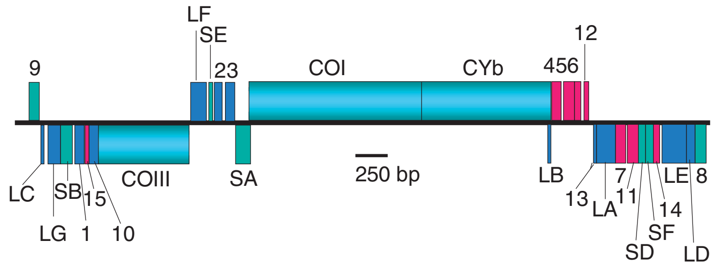


Figure 1 Single unit of the tandemly repeated *P. falciparum* mitochondrial genome map. Green boxes correspond to small subunit rRNA fragments, whereas blue boxes correspond to large subunit rRNA. The red boxes correspond to transcripts with characteristics like the rRNA fragments, which cannot be specifically placed in small or large subunit rRNA (known as misc RNA in Table 1). Details of the mapping coordinates can be found in Table 1 (Reproduced with the kind permission of J. E. Feagin)

data suggest that it is conserved through the apicomplexa (Vaidya *et al*., 1989; Feagin, 1992; McFadden *et al*., 2000; Feagin, 1994); see Figure 1 and Table 1). The 6-kb mitochondrial genome is arranged as a linear concatemer of about 20 unit copies per cell, replicates via a rolling circle mechanism, and appears to be uniparentally inherited from the macrogametocyte (Preiser *et al*., 1996; Creasey *et al*., 1993). It is one of the smallest known mitochondrial genomes, encoding only three proteins in the electron transport chain and small fragmented rRNA genes (Feagin, 1992; Ji *et al*., 1996; Feagin *et al*., 1997). The rRNAs range in size between 40 and 200 nt, and the genes are interspersed between the genes encoding the cytochrome *c* oxidase subunits I and III (COI and COIII) and the apocytochrome *b* (CYb). The isolation and characterization of apocytochrome *b* in both *Plasmodium falciparum* and *Toxoplasma gondii* demonstrate that these genes encode functional proteins (Vaidya *et al*., 1993; Srivastava *et al*., 1999; McFadden and Boothroyd, 1999). The detection of nearly whole-genome sized transcripts suggest that the genome is polycistronically transcribed, but the standing RNA pools contain mostly processed single transcripts (Ji *et al*., 1996). Transcript mapping studies revealed that the genes do not overlap but are very closely packed to the extent that some of the rRNA genes do not have any intervening sequence between them (Feagin *et al*., 1997; Rehkopf *et al*., 2000). Moreover, the polyadenylation of the processed transcripts results in very short or absent tails where the pattern is gene specific. From these findings, it is speculated that the control of RNA abundance is through precise cleavage of the polycistronic RNA and stability of the processed transcripts. The functional constraints imposed by this system may limit possible variations such that apicomplexan mitochondrial genomes evolve more slowly than their counterparts in other organisms.

3. The apicoplast genome

An elegant synthesis of molecular evidence with previous microscopic observations led to the rediscovery of an organelle now known as the apicoplast (for recent

Table 1 *Plasmodium falciparum* mitochondrial genome annotation

(GenBank Reference M76611)

Key	Synonym	Map name	Strand	5′ coord	3′ coord
rRNA	RNA9	9	Top	100	165
rRNA	LSUC	LC	Bottom	221	206
rRNA	LSUG	LG	Bottom	389	283
rRNA	SSUB	SB	Bottom	502	390
rRNA	RNA1	1	Bottom	606[a]	506[a]
rRNA	RNA15[a]	15[a]	Bottom[a]	624[a]	594[a]
rRNA	RNA10	10	Bottom	724	625
CDS	Cytochrome *c* oxidase subunit 3	COIII	Bottom	1487[a]	725[a]
Misc feature			Bottom	1488	1474
rRNA	LSUF	LF	Top	1501	1630
rRNA	SSUE	SE	Top	1650	1688
rRNA	RNA2	2	Top	1697	1763
rRNA	RNA3	3	Top	1831	1910
rRNA	SSUA	SA	Bottom	2023	1916
Misc feature			Top	2036	2050
CDS	Cytochrome *c* oxidase subunit 1	COI	Top	2037[a]	3479[a]
Misc feature			Top	3478	3493
CDS	Apocytochrome *b*	CYb	Top	3480[a]	4624[a]
rRNA	LSUB	LB	Bottom	4618	4594
Misc RNA	RNA4	4	Top	4625	4696
Misc RNA	RNA5	5	Top	4717	4802
Misc RNA	RNA6	6	Top	4803	4865
Misc RNA	RNA12	12	Top	4887	4945
rRNA	RNA13	13	Bottom	5025	4996
rRNA	LSUA	LA	Bottom	5201	5026
Misc RNA	RNA7	7	Bottom	5283	5202
Misc RNA	RNA11	11	Bottom	5378	5284
rRNA	SSUD	SD	Bottom	5446	5379
rRNA	SSUF	SF	Bottom	5507	5447
Misc RNA	RNA14	14	Bottom	5562	5508
rRNA	LSUE	LE	Bottom	5771	5577
rRNA	LSUD	LD	Bottom	5854	5772
rRNA	RNA8	8	Bottom	5955	5855

[a]J. E. Feagin (personal communication).

reviews, see Archibald and Keeling, 2002; Foth and McFadden, 2003). Amongst the morphological descriptions of a variety of apicomplexans were reports of a multimembranous organelle, variously called the "Hohlzylinder" and "Golgi-adjunct" (Siddall, 1992), but despite repeated observations, the organelle's function remained a mystery. Electron microscopic studies of *Plasmodium lophura* showed a circular DNA thought to be mitochondrial (Kilejian, 1975). This observation was confirmed by the analysis of density-gradient fractionated whole genomic

P. knowlesi DNA that revealed a band that represented an A-T-rich circular 35-kb DNA (Williamson *et al*., 1985). Subsequent reports on studies of *P. falciparum* suggested that this 35-kb circular DNA was of plastid origin and might be associated with a nonmitochondrial organelle (see Figure 2 and Table 2). Observations including the inverted repeat structure of the genes encoding the large and small subunit ribosomal RNAs (rRNAs) and split RNA polymerase *rpoC1* and *rpoC2* genes provided strong evidence that the 35-kb circular DNA is a degenerate plastid genome (Gardner *et al*., 1991a,b; Gardner *et al*., 1993). The localization of small subunit rRNA transcripts with an antisense probe to *T. gondii* thin sections provided the first direct evidence that the multimembranous organelle is a four-membrane-bound degenerate plastid (Kohler *et al*., 1997). The organelle's four-membrane composition suggests that the apicoplast is derived from a secondary endosymbiotic event with a photosynthetic eukaryote and may pre-date the origin of the phylum, and hence be a shared characteristic with all alveolates (Williams and Keeling, 2003). Although there is an ongoing debate as to whether the apicoplast is of green or red algal origin, the balance of arguments favor a rhodophyte (red) alga (Williams and Keeling, 2003).

The 35-kb apicoplast genomes characterized in *Plasmodium* spp., *T. gondii*, and *Eimeria tenella* are nearly identical structurally and, despite the differences in nuclear genome G+C contents, they are all very A+T rich; *P. falciparum* (86%), *T. gondii* (78.5%), *E. tenella* (79.4%) (see Figure 2 and Table 2; Cai *et al*., 2003). The apicoplast genomes have some of the major characteristics of chloroplast genomes but appear to encode only a fraction of the genes observed in the chloroplast (Foth and McFadden, 2003). For example, the *T. gondii* apicoplast genome maintains 33 tRNAs capable of translating all codons and 28 predicted coding sequences (CDS) encoding 17 ribosomal proteins, *tuf*A, *clp, rpo*B, *rpo*C1, *rpo*C2, ORF470 (*ycf* 24), and five CDSs of unknown function (see Figure 2 and Table 2). This limited coding capacity suggested that the vast majority of the organelle's proteins are encoded in the nucleus. Plastids carry out many other functions aside from photosynthesis, and plastid-related fatty acid, isoprenoid, and heme biosynthesis appear to be retained in the apicoplast. The Type II fatty acid synthesis protein, acyl carrier protein (ACP) showed putative orthologs in both *P. falciparum* and *T. gondii* databases and the *T. gondii* version was demonstrably targeted to the apicoplast (Waller *et al*., 1998). Importantly, this study showed that ACP and other apicoplast-targeted proteins shared a long N-terminal extension that functions as a bipartite signal required for the secretion and entry into the organelle. Synthesis of isoprenoids such as sterols and ubiquinones require isopentenyl diphosphate as a precursor, and apicomplexans depend on the mevalonate-independent 1-deoxy-D-xylulose 5-phosphate (DOXP) pathway for this function. The *P. falciparum* DOXP synthase and DOXP reductoisomerase orthologs can be functionally inhibited and target the apicoplast via a bipartite N-terminal extension (Jomaa *et al*., 1999). A systematic search of the *P. falciparum* genome for this signature bipartite signal sequence estimates that 500–600 proteins target to the apicoplast (Gardner *et al*., 2002).

Table 2 *Toxoplasma gondii* and *Plasmodium falciparum* apicoplast annotation

(GenBank References U87145; X95275 and X95276)

Key	Map name	*T. gondii*	Strand	5′ coordinate	3′ coordinate	*P. falciparum*
tRNA	I	tRNA-Ile	Bottom	79	8	tRNA-Ile
rRNA	SSU rRNA	Small subunit ribosomal RNA	Bottom	1745	246	Small subunit ribosomal RNA
tRNA	A	tRNA-Ala	Bottom	1867	1795	tRNA-Ala
tRNA	N	tRNA-Asn	Top	1908	1980	tRNA-Asn
tRNA	L	tRNA-Leu	Bottom	2078	2003	tRNA-Leu
tRNA	R	tRNA-Arg	Bottom	2165	2093	tRNA-Arg
tRNA	V	tRNA-Val	Bottom	2246	2175	tRNA-Val
tRNA	R	tRNA-Arg	Top	2282	2354	tRNA-Arg
tRNA	M	tRNA-Met	Top	2380	2453	tRNA-Met
rRNA	LSU rRNA	Large subunit ribosomal RNA	Top	2512	5198	Large subunit ribosomal RNA
tRNA	T	tRNA-Thr	Top	5199	5270	tRNA-Thr
CDS	rps4	Ribosomal protein S4	Top	5312	5908	rps4
tRNA	H	tRNA-His	Top	5936	6008	tRNA-His
tRNA	C	tRNA-Cys	Top	6017	6090	tRNA-Cys
tRNA	L	tRNA-Leu	Top	6099	6134	tRNA-Leu
Intron	L	Interrupts anticodon of tRNA-Leu	Top	6135	6322	tRNA-Leu
tRNA	L	tRNA-Leu	Top	6323	6362	tRNA-Leu
tRNA	M	tRNA-Met	Top	6386	6461	tRNA-Met
tRNA	Y	tRNA-Tyr	Top	6493	6575	tRNA-Tyr
tRNA	S	tRNA-Ser	Top	6610	6699	tRNA-Ser
tRNA	D	tRNA-Asp	Top	6720	6793	tRNA-Asp
tRNA	K	tRNA-Lys	Bottom/top	6896	6825	tRNA-Lys
tRNA	E	tRNA-Glu	Top	6924	6996	tRNA-Glu
tRNA	P	tRNA-Pro	Top	7001	7074	tRNA-Pro
CDS	rpl4	Ribosomal protein L4	Top	7117	7752	rpl4
CDS	rpl23	N/A	Pf top	N/A	N/A	rpl23
CDS	rpl2	Ribosomal protein L2	Top	7774	8577	rpl2
CDS	rps19	Ribosomal protein S19	Top	8597	8809	rps19
CDS	rps3	Ribsomal Protein S3	Top	8870	9544	rps3
CDS	rpl16	Ribsomal Protein L16	Top	9566	9955	rpl16
CDS	rps17	Ribosomal protein S17	Top	9960	10 181	rps17
CDS	rpl14	Ribosomal protein L14	Top	10 198	10 563	rpl14
Misc_ feature	rps8		Top	10 615	10 968	rps8
CDS	rpl16	Ribosomal protein L6	Top	11 008	11 556	rpl16
CDS	rps5	Ribosomal protein S5	Top	11 582	12 388	rps5
CDC	ORF-A	N/A	Pf top	N/A	N/A	ORF91
CDS	rpl36	Ribosomal protein L36	Top	12 405	12 518	rpl36
CDS	rps11	Ribosomal protein S11	Top	12 522	12 929	rps11
CDS	rps12	Ribosomal protein S12	Top	12 936	13 301	rps12

(*continued overleaf*)

Table 2 (*continued*)

(GenBank References U87145; X95275 and X95276)						
Key	Map name	*T. gondii*	Strand	5′ coordinate	3′ coordinate	*P. falciparum*
CDS	rps7	Ribosomal protein S7	Top	13 322	13 732	rps7
CDS	tufA	Elongation factor-Tu	Top	13 791	14 996	tufA
CDS	ORF-B	ORF-B	Top	15 007	15 138	ORF78
tRNA	F	tRNA-Phe	Bottom	15 222	15 151	tRNA-Phe
tRNA	Q	tRNA-Gln	Top	15 241	15 312	tRNA-Gln
tRNA	G	N/A	Pf top	N/A	N/A	tRNA-Gly
tRNA	W	tRNA-Trp	Top	15 341	15 411	tRNA-Trp
CDS	rpl11	Ribosomal protein L11	Top	15 451	15 846	ORF129
CDS	ORF-F	ORF-F	Top	15 855	16 031	Rearranged
CDS	ORF-E	ORF-E	Top	16 035	16 352	Rearranged
CDS	clpC	clp	Top	16 395	18 692	clp C
tRNA	G	tRNA-Gly	Top	18 683	18 755	tRNA-Gly
CDS	ORF-C	ORF-C	Top	18 806	19 015	ORF79
tRNA	S	tRNA-Ser	Top	19 023	19 106	tRNA-Ser
CDS	ORF-D	ORF-D	Top	19 158	19 382	ORF105
CDS	rps2	Ribosomal protein S2	Bottom	20 073	19 372	rps2
Misc_ feature	rpoC2		Bottom	23 381	20 103	rpoD
CDS	rpoC1	RNA Polymerase C1	Bottom	25 083	23 386	rpoC
CDS	rpoB	RNA Polymerase B	Bottom	28 258	25 103	rpoB
CDS	ORF-E	Rearranged	Pf top	N/A	N/A	ORF101
CDS	ORF-F	Rearranged	Pf top	N/A	N/A	ORF51
CDS	ORF-G	ycf24 Homolog	Bottom	29 686	28 289	ORF470
tRNA	T	tRNA-Thr	Bottom	29 798	29 728	Missing
rRNA	LSU rRNA	Large subunit ribosomal RNA	Bottom	32 485	29 799	Large subunit ribosomal RNA
tRNA	M	tRNA-Met	Bottom	32 617	32 544	tRNA-Met
tRNA	R	tRNA-Arg	Bottom	32 715	32 643	tRNA-Arg
tRNA	V	tRNA-Val	Top	32 751	32 822	tRNA-Val
tRNA	R	tRNA-Arg	Top	32 837	32 909	tRNA-Arg
tRNA	L	tRNA-Leu	Top	32 919	32 994	tRNA-Leu
tRNA	N	tRNA-Asn	Bottom	33 089	33 017	tRNA-Asn
tRNA	A	tRNA-Ala	Top	33 130	33 202	tRNA-Ala
rRNA	SSU rRNA	Small subunit ribosomal RNA	Top	33 252	34 751	Small subunit ribosomal RNA
tRNA	I	tRNA-Ile	Top	34 918	34 989	tRNA-Ile

Note: The *Plasmodium falciparum* coordinates do not correlate with the *Toxoplasma gondii* coordinates because the Pf GenBank accession numbers are split.

Compared to photosynthetic plastids, the apicoplast function is very limited, with current evidence restricted to Type II fatty acid and isoprenoid biosynthesis. Although the 35-kb genome encodes transcriptional and translational machinery, the vast majority of proteins associated with apicoplast function are encoded in the nucleus and imported via a bipartite signal sequence.

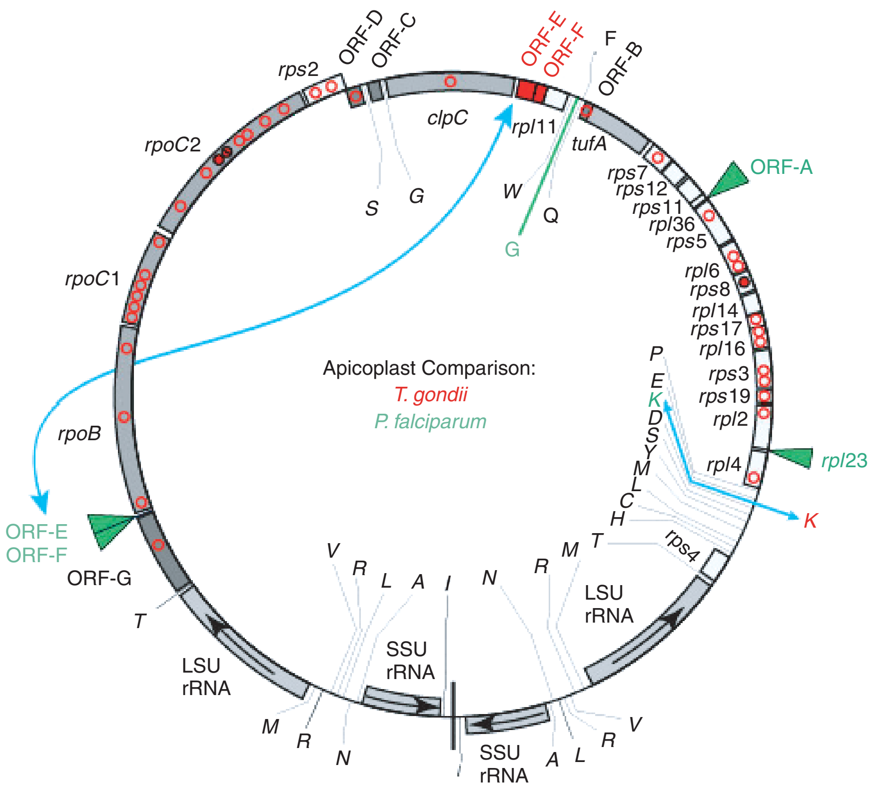


Figure 2 Combined map of *T. gondii* and *P. falciparum* apicoplast genome. The *E. tenella* apicoplast sequence has been completed and is identical to the *T. gondii* map shown (GenBank Reference AY217738). Red text indicates features that are present only in the *T. gondii* and *E. tenella*, whereas green text indicates *P. falciparum* specific features. Red open circles indicate the presence of in-frame UGA codons predicted to encode tryptophan, and filled red circles represent the presence of an in-frame stop codon (UAA and UAG; represented as misc_feature in Table 2). (Reproduced with the kind permission of J. Kissinger and D. Roos)

Acknowledgments

We would like to thank Jessie Kissinger, David Roos, and Jean Feagin for their help on the apicoplast and mitochondria figures and annotation. Funding for this work was provided by the BBSRC (JWA).

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, *et al.* (2004) Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science*, **304**, 441–445.

Archibald JM and Keeling PJ (2002) Recycled plastids: a 'green movement' in eukaryotic evolution. *Trends in Genetics*, **18**, 577–584.

Cai X, Fuller AL, McDougald LR and Zhu G (2003) Apicoplast genome of the coccidian Eimeria tenella. *Gene*, **321**, 39–46.

Creasey AM, Ranford-Cartwright LC, Moore DJ, Williamson DH, Wilson RJ, Walliker D and Carter R (1993) Uniparental inheritance of the mitochondrial gene cytochrome b in Plasmodium falciparum. *Current Genetics*, **23**, 360–364.

Fast NM, Kissinger JC, Roos DS and Keeling PJ (2001) Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Molecular Biology and Evolution*, **18**, 418–426.

Feagin JE (1992) The 6-kb element of Plasmodium falciparum encodes mitochondrial cytochrome genes. *Molecular and Biochemical Parasitology*, **52**, 145–148.

Feagin JE (1994) The extrachromosomal DNAs of apicomplexan parasites. *Annual Review of Microbiology*, **48**, 81–104.

Feagin JE, Mericle BL, Werner E and Morris M (1997) Identification of additional rRNA fragments encoded by the Plasmodium falciparum 6 kb element. *Nucleic Acids Research*, **25**, 438–446.

Foth BJ and McFadden GI (2003) The apicoplast: a plastid in Plasmodium falciparum and other Apicomplexan parasites. *International Review of Cytology*, **224**, 57–110.

Gardner MJ, Feagin JE, Moore DJ, Rangachari K, Williamson DH and Wilson RJ (1993) Sequence and organization of large subunit rRNA genes from the extrachromosomal 35 kb circular DNA of the malaria parasite Plasmodium falciparum. *Nucleic Acids Research*, **21**, 1067–1071.

Gardner MJ, Feagin JE, Moore DJ, Spencer DF, Gray MW, Williamson DH and Wilson RJ (1991a) Organisation and expression of small subunit ribosomal RNA genes encoded by a 35-kilobase circular DNA in Plasmodium falciparum. *Molecular and Biochemical Parasitology*, **48**, 77–88.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.

Gardner MJ, Williamson DH and Wilson RJ (1991b) A circular DNA in malaria parasites encodes an RNA polymerase like that of prokaryotes and chloroplasts. *Molecular and Biochemical Parasitology*, **44**, 115–123.

Gero AM, Brown GV and O'Sullivan WJ (1984) Pyrimidine de novo synthesis during the life cycle of the intraerythrocytic stage of Plasmodium falciparum. *The Journal of Parasitology*, **70**, 536–541.

Huang J, Mullapudi N, Sicheritz-Ponten T and Kissinger JC (2004) A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *International Journal for Parasitology*, **34**, 265–274.

Ji YE, Mericle BL, Rehkopf DH, Anderson JD and Feagin JE (1996) The Plasmodium falciparum 6 kb element is polycistronically transcribed. *Molecular and Biochemical Parasitology*, **81**, 211–223.

Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M, Turbachova I, Eberl M, Zeidler J, Lichtenthaler HK, *et al*. (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, **285**, 1573–1576.

Kilejian A (1975) Circular mitochondrial DNA from the avian malarial parasite Plasmodium lophurae. *Biochimica Et Biophysica Acta*, **390**, 276–284.

Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD and Roos DS (1997) A plastid of probable green algal origin in Apicomplexan parasites. *Science*, **275**, 1485–1489.

McFadden DC and Boothroyd JC (1999) Cytochrome b mutation identified in a decoquinate-resistant mutant of Toxoplasma gondii. *The Journal of Eukaryotic Microbiology*, **46**, 81S–82S.

McFadden DC, Tomavo S, Berry EA and Boothroyd JC (2000) Characterization of cytochrome b from Toxoplasma gondii and Q(o) domain mutations as a mechanism of atovaquone-resistance. *Molecular and Biochemical Parasitology*, **108**, 1–12.

Prapunwattana P, O'Sullivan WJ and Yuthavong Y (1988) Depression of Plasmodium falciparum dihydroorotate dehydrogenase activity in in vitro culture by tetracycline. *Molecular and Biochemical Parasitology*, **27**, 119–124.

Preiser PR, Wilson RJ, Moore PW, McCready S, Hajibagheri MA, Blight KJ, Strath M and Williamson DH (1996) Recombination associated with replication of malarial mitochondrial DNA. *The EMBO Journal*, **15**, 684–693.

Rehkopf DH, Gillespie DE, Harrell MI and Feagin JE (2000) Transcriptional mapping and RNA processing of the Plasmodium falciparum mitochondrial mRNAs. *Molecular and Biochemical Parasitology*, **105**, 91–103.

Roos DS, Crawford MJ, Donald RG, Fraunholz M, Harb OS, He CY, Kissinger JC, Shaw MK and Striepen B (2002) Mining the Plasmodium genome database to define organellar function: what does the apicoplast do?. *Philosophical Transactions of The Royal Society of London. Series B, Biological Sciences*, **357**, 35–46.

Siddall ME (1992) Hohlzylinders. *Parasitology Today*, **8**, 90–91.

Srivastava IK, Morrisey JM, Darrouzet E, Daldal F and Vaidya AB (1999) Resistance mutations reveal the atovaquone-binding domain of cytochrome b in malaria parasites. *Molecular Microbiology*, **33**, 704–711.

Srivastava IK, Rottenberg H and Vaidya AB (1997) Atovaquone, a broad spectrum antiparasitic drug, collapses mitochondrial membrane potential in a malarial parasite. *The Journal of Biological Chemistry*, **272**, 3961–3966.

Vaidya AB, Akella R and Suplick K (1989) Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite. *Molecular and Biochemical Parasitology*, **35**, 97–107.

Vaidya AB, Lashgari MS, Pologe LG and Morrisey J (1993) Structural features of Plasmodium cytochrome b that may underlie susceptibility to 8-aminoquinolines and hydroxynaphthoquinones. *Molecular and Biochemical Parasitology*, **58**, 33–42.

Vercesi AE, Rodrigues CO, Uyemura SA, Zhong L and Moreno SN (1998) Respiration and oxidative phosphorylation in the apicomplexan parasite Toxoplasma gondii. *The Journal of Biological Chemistry*, **273**, 31040–31047.

Waller RF, Keeling PJ, Donald RG, Striepen B, Handman E, Lang-Unnasch N, Cowman AF, Besra GS, Roos DS and McFadden GI (1998) Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 12352–12357.

Williams BA and Keeling PJ (2003) Cryptic organelles in parasitic protists and fungi. *Advances in Parasitology*, **54**, 9–68.

Williamson DH, Wilson RJ, Bates PA, McCready S, Perler F and Qiang BU (1985) Nuclear and mitochondrial DNA of the primate malarial parasite Plasmodium knowlesi. *Molecular and Biochemical Parasitology*, **14**, 199–209.

Short Specialist Review

Environmental shotgun sequencing

Gene W. Tyson
University of California, Berkeley, CA, USA

Philip Hugenholtz
Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

1. Introduction

Genome sequencing has revolutionized the study of microorganisms. Determining the complete genetic makeup of an organism, in principle, lays bare its metabolic potential and evolutionary history. However, microbial genomes sequenced to date lack environmental context. They are derived from microorganisms maintained in pure culture on artificial growth media and are unlikely to be representative of the population or community from which they were obtained. This limitation can be bypassed by direct sequencing of microbial communities from the environment. Two decades ago, Norman Pace and colleagues outlined an approach to use ribosomal RNAs (rRNAs) as phylogenetic markers for the organisms present in an environmental sample by extracting DNA, PCR-amplifying, cloning, and sequencing rRNA genes directly from the sample (Pace *et al*., 1986). This approach revealed and continues to reveal an extraordinary diversity of organisms in the microbial world, orders of magnitude greater than had been appreciated by culture-dependent techniques (Head *et al*., 1998; Hugenholtz, 2002). While markers such as 16S rRNA give an indication of which organisms are present in the environment, it reveals little about what those organisms might be doing. The natural progression of the rRNA-based work, therefore, was the direct cloning and sequencing of genomic DNA from environmental samples into large insert vectors such as BACs and fosmids (Beja *et al*., 2000a; Rondon *et al*., 2000). These genomic libraries are typically screened for rRNA genes (and other conserved markers) and clones are fully sequenced. Because of the size of the genomic inserts, dozens of protein-coding sequences associated with the rRNA gene can be identified. Using this approach, Ed DeLong and colleagues discovered genes for light-driven proton pumps (proteorhodopsins) in ocean waters belonging to an uncultivated lineage of marine *Gammaproteobacteria*, strongly suggesting a major role for phototrophy in marine ecosystems (Beja *et al*., 2000b; Beja *et al*., 2001). This strategy, however, reveals only a fraction of the metabolic potential of a community. Therefore,

when shotgun sequencing was demonstrated as a viable approach to quickly obtain complete organism genome sequences (Fleischmann *et al*., 1995; *see also* Article 2, **Genome sequencing of microbial species**, Volume 3), it was only a matter of time before it was applied to microbial communities. The first communities to be studied using an environmental shotgun sequencing approach were an acid mine drainage (AMD) biofilm (Tyson *et al*., 2004) and samples of the Sargasso Sea (Venter *et al*., 2004). The AMD biofilm is a low-diversity community nominally comprising three bacterial and three archaeal members by 16S rRNA analysis, and the Sargasso Sea samples are moderately complex comprising several hundred 16S rRNA phylotypes. Large genomic fragments of the dominant community members could be reconstructed by assembly of shotgun sequence reads in both studies, verifying the feasibility of the approach.

2. The impact of environmental shotgun sequencing on microbial ecology

Direct sampling of microbial communities provides insight into the metabolism of uncultured organisms and an overview of community function. For example, the bacterial members of the AMD biofilm were inferred to be the primary producers of fixed carbon and nitrogen in the community, and the archaeal members appear to be adapted to scavenge these nutrients mainly in the form of amino acids (Tyson *et al*., 2004). In the Sargasso Sea study, a large number of proteorhodopsin homologs were identified, suggesting that phototrophy is a widespread strategy for energy production in the ocean (Venter *et al*., 2004).

Environmental sequence data can also provide clues for the cultivation of uncultured microorganisms. For instance, only one set of nitrogen fixing genes was identified in the AMD genome data belonging to an uncultured bacterial member of the community. This provided the basis for a directed isolation strategy using nitrogen-free media, which resulted in a pure culture of the targeted bacterium (Tyson *et al*., in prep).

One long recognized drawback of PCR-based molecular surveys is the reliance of the method on broad-specificity primers to amplify the gene of interest from all organisms in an environmental sample. Therefore, microorganisms with mismatches to these primers may be overlooked. Shotgun sequencing bypasses this problem as it does not rely on PCR. A striking example of this was the unexpected identification of a novel euryarchaeota in the AMD biofilm, which was missed during an exhaustive analysis of a 16S rDNA PCR clone library because it has three mismatches to one of the conserved primers (Baker *et al*., in prep).

One of the most exciting aspects of environmental shotgun sequencing is the ability to resolve population structure of cohabiting and coevolving strains and species. This is possible because each shotgun read likely originates from a different individual within a population, giving an overview of the genomic variation within that population. To date, population genetics has largely relied on comparison of isolates, usually pathogens, from different habitats (Spratt *et al*., 2001). One intriguing observation from the AMD data is that for one archaeal species population at least,

individual genomes are recombinant mosaics of closely related strains (Tyson *et al*., 2004). This suggests that, contrary to current opinion, genetic exchange akin to that in sexual organisms may be the cohesive force holding microbial species together. Since the frequency of homologous recombination decreases exponentially with genome divergence (Vulic *et al*., 1997), some microbial species may be naturally defined by their ability to recombine. However, genomic mosaicism is an isolated observation in an extreme habitat that needs to be confirmed with other sympatric microbial populations.

3. Hurdles and caveats

Environmental shotgun sequencing presents a number of technical challenges, both experimental and computational. Central to the success of any sequencing project is the extraction of high-quality DNA. Since environmental samples contain multiple species, a second objective is to obtain DNA from all community members quantitatively representative of each species in the sample. Microbial ecologists conducting 16S rRNA-based molecular surveys have recognized this for many years and devoted much attention to optimizing DNA extraction procedures for a range of habitats. The extraction step is even more critical for direct cloning of environmental DNA since PCR is not available as a buffer to provide high-quality DNA. For example, DNA extracted from AMD samples suitable for PCR amplification has proven difficult to clone into large insert vectors (unpublished observations). Shotgun libraries (~3-kb inserts) may be the only viable approach for some environmental samples where obtaining high-purity DNA suitable for direct cloning results in low-molecular-weight DNA.

Assembly of environmental shotgun sequence data presents a number of challenges. In contrast to sequencing a microbial isolate where all reads are derived from a single clonal genome, environmental sequences sample the genomes of multiple strains and species. Standard assembly methods can easily reconstitute the genomes of different species from a mixed pool of sequence reads but genetically distinct strains are often assembled into single composite genomic fragments (Tyson *et al*., 2004; Venter *et al*., 2004). This has the benefit of highlighting single nucleotide polymorphisms within a population but is problematic when trying to apply standard population genetics methods in which strain separation is essential (Hartl, 1997). It may be possible to resolve fragments of individual strain genomes by increasing assembly stringency and sequencing coverage; however, novel methods will likely need to be developed to analyze composite genomic fragments. In highly complex environments, such as soil, assembly may not be feasible (Figure 1) without a vast amount of sequencing, which is currently impractical given the cost of DNA sequencing (Tringe *et al*., in prep). Therefore, it will be important to develop methods that can utilize single reads for comparative analysis.

Environmental shotgun sequencing introduces a new element to genome data analysis. Since a microbial community comprises multiple species, the genomic fragments obtained from the community need to be binned or classified into their respective populations. This was achieved using a combination of GC content, read depth, and similarity to isolate genomes for both the AMD and Sargasso Sea

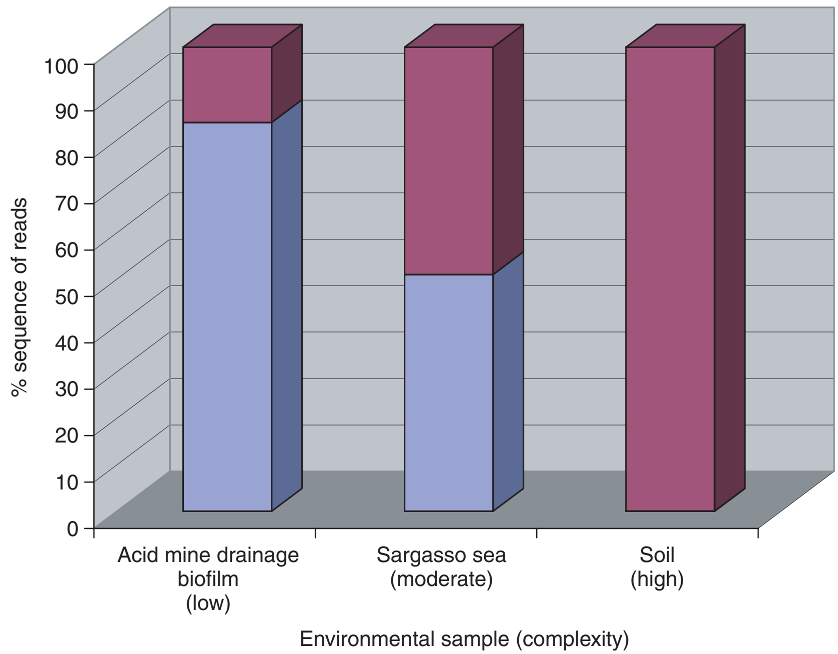


Figure 1 Effect of community complexity on the ability to assemble environmental shotgun sequence data. As community complexity increases from low (acid mine drainage biofilm; Tyson *et al*., 2004) to moderate (Sargasso Sea; Venter *et al*., 2004) to high (soil; Tringe *et al*., in prep.), the ability to assemble the data decreases. All datasets were normalized for amount of sequence. Blue: assembled reads; Red: unassembled reads

studies. Each of these criteria have limitations; for example, different species may have indistinguishable GC contents, read depth varies within a given genome, and the accuracy of similarity-based binning is dependent on having sequenced genomes of closely related phylogenetic neighbors. Our current sampling of the tree of life is highly skewed owing to a cultivation bias (Hugenholtz, 2002), meaning that many lineages are poorly sampled for representative genomes. Isolates obtained from the environmental sample under study provide the best reference points for binning since they are representatives of populations in the community. For example, an isolate of one of the archaeal populations in the AMD biofilm was invaluable for separating two AMD populations indistinguishable by GC content and read depth (Tyson *et al*., 2004). Genome signatures, such as oligonucleotide frequencies, hold promise for finer scale binning of environmental genome fragments because they are distinctive even for closely related organisms (Teeling *et al*., 2004).

Ecologists have shown that many natural communities follow a lognormal species abundance distribution characterized by a few dominant and many rare species (Magurran and Henderson, 2003). This distribution was observed in the environmental genomic data at both the species and the strain level (Tyson *et al*., 2004) and resulted in sampling of only the numerically dominant populations with little or no sampling of rare populations. This caveat of the method is easily overlooked in the deluge of sequence data, and can have important implications

for assembly, binning, and annotation of the data since sampling unevenness can deleteriously affect all three. Methods that could be used or adapted to access rare populations in a community include normalized libraries (Patanjali *et al*., 1991) or physical isolation of cells prior to DNA extraction and library preparation (Zengler *et al*., 2002).

An emerging issue with environmental sequence data is data handling. For example, the Sargasso Sea study generated 1.045 Gb of nonredundant genome sequence that temporarily swamped the NCBI database until it and the AMD data were placed in a separate environmental sequence database. Furthermore, basic genome analyses such as all versus all comparisons scale quadratically and pose a real computational challenge for large environmental sequence datasets.

4. The promise of environmental shotgun sequencing for microbial ecology and evolution

Environmental shotgun data enabled by emerging technologies such as microarrays and proteomics hold great promise for the field of microbial ecology. In the next decade, these technological advancements should place organisms in the context of their community and environment, reveal how communities function as a whole, provide one or more definitions of a microbial species, explain how differentiation arises within a sympatric population, and reveal the importance of natural selection in the evolution of microbial species.

Acknowledgments

We thank Susannah Green Tringe for helpful comments on the article.

References

Baker BJ, Tyson GW, Webb RI, Hugenholtz P and Banfield JF A. novel, acidophilic ultra-small archaeon revealed by community genome sequencing. *Science* in review.

Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M, Garrigues C, Jovanovich SB, *et al*. (2000a) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology*, **2**, 516–529.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, *et al*. (2000b) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**, 1902–1906.

Beja O, Spudich EN, Spudich JL, Leclerc M and DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. *Nature*, **411**, 786–789.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Hartl DL (1997) *Principles of Population Genetics*, Third Edition, Sinauer Associates, Publishers: Sunderland.

Head IM, Saunders JR and Pickup RW (1998) Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology*, **35**, 1–21.

Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biology*, **3**, reviews 0003.1–0003.8.

Magurran AE and Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714–716.

Pace NR, Stahl DA, Lane DJ and Olsen GJ (1986) The analysis of natural microbial-populations by ribosomal-RNA sequences. *Advances in Microbial Ecology*, **9**, 1–55.

Patanjali SR, Parimoo S and Weissman SM (1991) Construction of a uniform-abundance (normalized) cDNA library. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 1943–1947.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, *et al*. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology*, **66**, 2541–2547.

Spratt BG, Hanage WP and Feil EJ (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology*, **4**, 602–606.

Teeling H, Meyerdierks A, Bauer M, Amann R and Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, **6**, 938–947.

Tringe SG, von Mering C, Kobayashi A, Salamov A, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al*. Comparative metagenomics of microbial communities. In prep.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Tyson GW, Lo I, Baker B, Allen EE, Hugenholtz P and Banfield JF Genome directed isolation of the key nitrogen fixer in acid mine drainage communities. In prep.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al*. (2004) Environmental genome shotgun sequencing of the sargasso sea. *Science*, **304**, 66–74.

Vulic M, Dionisio F, Taddei F and Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9763–9767.

Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM and Keller M (2002) Cultivating the uncultured. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15681–15686.

Short Specialist Review

Methods for detecting horizontal transfer of genes

Jeffrey G. Lawrence
University of Pittsburgh, Pittsburgh, PA, USA

1. Introduction

Bacteria and Archaea are well known to reproduce by binary fission, whereby the genetic material contained in the parental cell is typically replicated and passed to daughter cells unfettered except for the action of mutational processes. Yet, it has long been recognized that superimposed upon this fundamental biology lies the process of gene exchange, whereby cells can receive genetic material from a nonmaternal parent (Koonin *et al*., 2001; Ochman *et al*., 2000). Despite careful exploration of the mechanisms of gene exchange since the early twentieth century, well-documented cases of gene exchange among both closely and distantly related organisms, the identification – by independent methods – of numerous genes recently introduced into bacterial genomes by gene transfer, and dramatic phenotypic differences among closely related strains that can be directly attributed to gene change (Lawrence and Ochman, 2002), the role of this process in the evolution of microbial genomes remains a contentious, hotly debated issue (Gogarten *et al*., 2002; Kurland, 2000). While no one doubts that gene transfer occurs, valid questions remain as to what the overall impact of the process has been.

Herein, I will discriminate between two classes of gene transfer, both of which may be properly labeled as "horizontal" or "lateral." First, genes may be mobilized among genomes of closely related strains, typically via conjugation, transformation, or bacteriophage-mediated transduction. After introduction into the cytoplasm, the introduced DNA is recombined into a stable replicon by enzymes that perform homologous recombination. This process results in allelic change, or gene replacement. Its effects were initially measured by isozyme analysis, then multilocus enzyme electrophoresis, DNA sequence analysis, and now large-scale MultiLocas Sequence Typing (MLST) approaches. In all cases, the loss of linkage disequilibrium between distinct loci of different genes is taken as a measure for the rate of DNA exchange by this route. The phenomenon of intragroup gene exchange lies beyond the scope of this review. Instead, I will focus on methods for detecting DNA exchange among distantly related organisms; such transfer events usually do not rely upon homologous recombination to introduce the incoming DNA into a stable replicon, and gene acquisition – which can alter dramatically the

physiological capabilities of the recipient organism – often results (Lawrence, 1999, 2002). It is this sort of process, for example, that can lead to the rapid adaptation of certain strains of *Escherichia coli* as pathogenic organisms (Welch *et al*., 2002).

At the center of assessing the impact of gene transfer comes the identification of foreign genes themselves. Numerous methods have been employed, and cogent reviews of their methodology appear elsewhere. Herein, I review the two major classes of methods employed to identify genes introduced into the genome from a foreign source, with an emphasis on how and why different methods detect different sets of genes in the same genome. That is, it is critical to recognize two features of any approach for the identification of alien genes: (1) what one finds depends strongly on the method employed and typically reflects a fundamentally different null hypothesis being tested (Lawrence and Hendrickson, 2003) and (2) any method can provide at best a probability that a gene has been introduced into the chromosome from a foreign source; no method can provide a straightforward yes-or-no answer as to whether a gene is "native" or "foreign." Considering that the number of genes shared among free-living bacteria has been estimated at <200, and among all prokaryotes at <80, it is highly likely that no gene has not experienced horizontal transfer at some point during its tenure within a lineage (Ragan and Charlebois, 2002). While genes may be transferred with different efficiencies (Jain *et al*., 1999), no gene appears immune to HGT (horizontal gene transfer). Genes encoding core metabolic functions, conserved biosynthetic pathways, components of the transcription and translation apparatus, and even ribosomal RNA have been subject to HGT (Gogarten *et al*., 2002). Robust identification of the genes most likely to be subject to transfer, the phylogenetic scope of their potential recipients with respect to their donor, and other issues lie at the heart of placing HGT within the context of the multitude of factors influencing microbial evolution (Lawrence and Hendrickson, 2003).

2. Phylogenetic methods

HGT can occur between phylogenetically distantly related organisms, for example, between bacteria and Eukarya (Garcia-Vallve *et al*., 2000). As a result, phylogenetic inferences derived from different molecules from the same set of taxa are only rarely completely congruent; when they are, the genes are typically found in all taxa represented in the data set (Daubin *et al*., 2003); among genes found only in a subset of taxa, phylogenetic congruence is rarely observed (Clarke *et al*., 2002; Ragan and Charlebois, 2002). Therefore, the effects of HGT in confounding phylogenetic reconstruction have been taken as evidence for its existence, although alternative explanations cannot be excluded (Doolittle, 1999a,b; Koonin *et al*., 2001; Nelson *et al*., 1999). Phylogenetic methods rely upon gene relationship to infer gene transfer, although formal creation of phylogenetic trees is often not performed. Methods often attempt to discern atypical distributions of genes across genomes, and may include the identification of (1) genes present in isolated taxa but absent from closely related species (implying gene gain), (2) genes with an unusually high degree of similarity to genes found in otherwise unrelated taxa (Clarke *et al*., 2002), and (3) genes whose phylogenetic relationships are not

congruent with the relationships inferred from other genes in their respective genomes (Doolittle, 1999a,b, 2000; Olendzenski *et al*., 2000).

Caveats to all phylogenetic approaches include the confounding effects of convergent evolutionary forces, the amplification of gene families that prevent the robust identification of orthologues, and phylogenetic artifacts such as long-branch attraction (Philippe and Forterre, 1999). As a result, even apparent robust conclusions for widespread gene transfer based on phylogenetic methods (Raymond *et al*., 2002) may be called into question based on questions of methodology (Daubin *et al*., 2003; Daubin and Ochman, 2004).

3. Parametric measures

In contrast, phylogeny-independent methods identify genes that appear atypical in their current genomic context. Here, discordance is interpreted as reflecting long-term evolution in genomes with different mutational biases. Mutational biases result naturally during the DNA replication process as the mutational proclivities of DNA polymerases, the specific dNTP and tRNA contents of the cell, lineage-specific mismatch-repair systems, and other variables contribute to directional mutation pressures (Sueoka, 1988). Such parametric methods examine nucleotide and dinucleotide frequencies (Karlin and Burge, 1995; Lawrence and Ochman, 1997, 1998), codon usage biases (Karlin *et al*., 1998; Moszer *et al*., 1999), or patterns uncovered by Markov chain analyses (Hayes and Borodovsky, 1998) or minimum entropy analyses (Azad *et al*., 2002).

Early applications of parametric methods suffered from poor characterization of "typical" chromosomal genes, which prevented robust identification of "atypical" genes (Koski *et al*., 2001). More advanced methods employ more rigorous criteria for gene length (Lawrence and Ochman, 2002), increase the number of parameters being analyzed, and often employ iterative and floating baseline techniques to quantify "typical" genic parameters with a higher degree of accuracy (Lawrence, unpublished data). Moreover, new techniques will combine multiple analytical approaches, thereby preventing the foibles of any one method from leading the investigators to ambiguous or misleading conclusions (Hsiao *et al*., 2003).

4. No one answer

Each of the methods described above has been used to infer that substantial portions of different genomes have arisen by HGT. However, each method examines different properties of the genes or genomes; as a result, they naturally identify different subsets of genes as being "foreign." In each case, what kinds of genes obtain this label is a function of the method employed. For genes identified as atypical, only those recently acquired may be identified, since the process of gene amelioration will erase the mutation bias signal of the donor genome over time (Lawrence and Ochman, 1997, 1998). In contrast, phylogenetic methods may detect ancient transfer events (Woese *et al*., 2000) but cannot identify genes that were acquired from relatively closely related organisms. Therefore, the different methods

are appropriate for testing different sorts of hypotheses (Lawrence and Ochman, 2002; Ragan, 2001a,b).

References

Azad RK, Rao JS, Li W and Ramaswamy R (2002) Simplifying the mosaic description of DNA sequences. *Physical Review E: Statistical, Nonlinear and Soft Matter Physics*, **66**, Epub 031913.

Clarke GD, Beiko RG, Ragan MA and Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *Journal of Bacteriology*, **184**, 2072–2080.

Daubin V and Ochman H (2004) Quartet mapping and the extent of lateral transfer in bacterial genomes. *Molecular Biology and Evolution*, **21**, 86–89.

Daubin V, Moran NA and Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**, 829–832.

Doolittle WF (1999a) Lateral genomics. *Trends in Cell Biology*, **9**, M5–M8.

Doolittle WF (1999b) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129.

Doolittle WF (2000) The nature of the universal ancestor and the evolution of the proteome. *Current Opinion in Structural Biology*, **10**, 355–358.

Garcia-Vallve S, Romeu A and Palau J (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Molecular Biology and Evolution*, **17**, 352–361.

Gogarten JP, Doolittle WF and Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, **19**, 2226–2238.

Hayes WS and Borodovsky M (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Research*, **8**, 1154–1171.

Hsiao W, Wan I, Jones SJ and Brinkman FS (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.

Jain R, Rivera MC and Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 3801–3806.

Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, **11**, 283–290.

Karlin S, Mrazek J and Campbell AM (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Molecular Microbiology*, **29**, 1341–1355.

Koonin EV, Makarova KS and Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*, **55**, 709–742.

Koski LB, Morton RA and Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Molecular Biology and Evolution*, **18**, 404–412.

Kurland CG (2000) Something for everyone. *EMBO Reports*, **1**, 92–95.

Lawrence JG (1999) Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology*, **2**, 519–523.

Lawrence JG (2002) Gene transfer in bacteria: speciation without species? *Theoretical Population Biology*, **61**, 449–460.

Lawrence JG and Hendrickson H (2003) Lateral gene transfer: when will adolescence end? *Molecular Microbiology*, **50**, 739–749.

Lawrence JG and Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution*, **44**, 383–397.

Lawrence JG and Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 9413–9417.

Lawrence JG and Ochman H (2002) Reconciling the many faces of gene transfer. *Trends in Microbiology*, **10**, 1–4.

Moszer I, Rocha EP and Danchin A (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Current Opinion in Microbiology*, **2**, 524–528.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.

Ochman H, Lawrence JG and Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG and Gogarten JP (2000) Horizontal transfer of Archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. *Journal of Molecular Evolution*, **51**, 587–599.

Philippe H and Forterre P (1999) The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, **49**, 509–523.

Ragan MA (2001a) Detection of lateral gene transfer among microbial genomes. *Current Opinion in Genetics & Development*, **11**, 620–626.

Ragan MA (2001b) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiology Letters*, **201**, 187–191.

Ragan MA and Charlebois RL (2002) Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *International Journal of Systematic and Evolutionary Microbiology*, **52**, 777–787.

Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY and Blankenship RE (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science*, **298**, 1616–1620.

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2653–2657.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, *et al*. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 17020–17024.

Woese CR, Olsen GJ, Ibba M and Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, **64**, 202–236.

Short Specialist Review

Genomics of *Rickettsia*

Hiroyuki Ogata
Structural & Genomic Information Laboratory, UPR 2589, IBSM, CNRS, Parc Scientifique de Luminy, 13288 Marseille Cedex 9, France

Didier Raoult
Unité des Rickettsies, UMR 6020, IFR 48, CNRS, Faculté de Médecine, 13385 Marseille Cedex 5, France

1. *Rickettsia*

Rickettsia are classified in the α subgroup of proteobacteria (Raoult and Roux, 1997). In nature, rickettsiae live in arthropods such as ticks, fleas, lice, or mites. In arthropods, most rickettsiae are vertically transmitted from mother to their progeny by transovarial transmission. Human beings are incidental hosts infected by arthropod bite or by autoinoculation through infected feces. Rickettsioses are mild to severe diseases presenting usually with fever, headaches, and a rash. In humans, rickettsiae multiply in vascular endothelial cells or in monocytes (e.g., *R. akari* and *Orientia tsutsugumushi*). Due to their obligate association with hosts, rickettsiae have few opportunities to exchange genes with other bacteria (*see* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4), although recent studies provide evidences for rather ancient gene exchanges with bacteria living in amoeba. Rickettsiae need to regulate their metabolism to synchronize their life cycle to that of the host. For example, rickettsiae are quiescent in ticks when the hosts are resting. When the tick temperature increases after feeding, the rickettsiae are reactivated and start to multiply. Metabolic activity of the hosts provides various nutrients accessible to rickettsiae, and makes many de novo biosynthetic pathways of rickettsiae nonessential for their life. This allowed the degradation of unnecessary genes and caused extensive genome reduction.

2. Genome organization

Genome sequences have been determined and published for five *Rickettsia* species: *R. prowazekii* (Andersson *et al*., 1998) and *R. typhi* (McLeod *et al*., 2004) of the typhus group (TG), *R. conorii* (Ogata *et al*., 2001) and *R. felis* (Ogata *et al*., 2005a) of the spotted fever group (SFG) and the earliest diverging species, *R. bellii* (Ogata *et al*., 2006). *Rickettsia* genomes are reduced in size and rich in adenine/thymine bases. Most of the sequenced genomes are comprised of a single

circular chromosome. However, they still exhibit substantial intraspecies variations in terms of size, gene composition; repeat content as well as the presence of extra chromosomal DNA elements. The size of the sequenced genomes varies between 1.1 and 1.5 Mb. The number of predicted protein-coding genes varies from 834 (for *R. prowazekii*) to 1512 (for *R. felis*). Repeated DNA sequences (*see* Article 9, **Repeatfinding**, Volume 7) are rare (0.3%) in the genomes of TG *Rickettsia*, while relatively abundant in the SFG *Rickettsia* (1.6–4.3%). In addition to the main chromosomal DNA, *R. felis* exhibits circular plasmids.

3. Genome content

Three axes characterize the strategies that *Rickettsia* adopted for their parasitic intracellular life: (1) the ability of invading the hosts, (2) the ability of up taking nutrients, and (3) the ability of synchronizing their replication with that of the host cells.

3.1. Host invasion

Through horizontal transmission involving a mammal–arthropod cycle, *Rickettsia* propagate in the host populations (Azad and Beard, 1998). *Rickettsia* enter nonphagocytic cells such as intestinal epithelial cells by inducing the formation of a phagocytic vacuole, and rapidly escape from the membrane-bound vacuole into the host cytoplasm (Gouin *et al*., 2005). Genome analysis allowed the identification of five genes with potential membrane-lytic activities: a hemolysin (*tlyA*, *tlyC*), a phospholipase D (*pld*), and a patatin-like phospholipase A (*pat1*, *pat2*). Recent experimental data suggests that *tlyC* and *pld* have a role in the process of escape from the phagocytic vacuoles (Whitworth *et al*., 2005). After being internalized in the cytoplasm, *Rickettsia* start replication. SFG *Rickettsia*, *R. bellii* as well as *R. typhi* induce actin polymerization at their surface and move around intra- and intercellularly (Gouin *et al*., 2005). The genome sequencing of *R. conorii* revealed a candidate gene responsible for this actin-based motility (Ogata *et al*., 2001). The product of the gene, named as *RickA*, was later demonstrated to have a critical role in the induction of actin polymerization (Jeng *et al*., 2004; Gouin *et al*., 2004). Two outer membrane proteins, rOmpA and rOmpB, have been suggested to have a role in adhesion to host cells (Li and Walker, 1998; Uchiyama, 2003). These two proteins are members of a large paralogous gene family, the “surface cell antigens (Sca)”. The analysis of the *Rickettsia* genomes revealed 17 members of the Sca family (Blanc *et al*., 2005). Interestingly, the genes for the Sca proteins exhibit remarkably diverse patterns of presence/absence across different species. The *N*-terminal parts of these proteins, which are exported to the cell surface, are highly variable between orthologs. It is suggested that accelerated amino acid changes and differential gene degradation of Sca paralogs contributed to the intraspecies variation of those cell-surface proteins and adaptation to different environments. All *Rickettsia* genomes contain genes encoding proteins with ankyrin repeats and tetratricopeptide repeats. *R. felis* and *R. bellii* genomes are especially enriched with

those genes. These repeats are protein–protein interaction motifs found in diverse organisms, and involved in different cellular activities including cell division. Their roles in the adaptation of parasites to their hosts have been suggested (Mosavi *et al.*, 2002; Rubtsov and Lopina, 2000; Blatch and Lassle, 1999). Rickettsiae may alter the intracellular environment of their hosts for survival, by targeting those proteins to the hosts. Finally, *R. felis* possesses additional genes potentially involved in the host invasion process such as a hyaluronidase gene, a chitinase gene, and a gene for a homolog of ecotin.

3.2. Uptake of nutrients

Rickettsia genomes lack genes for the glycolytic pathway and exhibit a restricted set of genes for the biosynthesis of small molecules such as amino acids, nucleotides, and cofactors. To compensate for the lack of genes for the biosynthesis of small molecules, *Rickettsia* exploit the host cells by different transporters. All *Rickettsia* exhibit five paralogs for the ATP/ADP translocases, hallmark enzymes for the energy parasitism of rickettsiae and chlamydiae. All *Rickettsia* possess a gene for the transporter of *S*-adenosylmethionine (SAM) (Tucker *et al.*, 2003), which is an essential methyl donor in various methyltransferase reactions. This contrasts with the degradation of the SAM synthetase gene (*metK*) in most *Rickettsia* species (Andersson and Andersson, 1999; Driskell *et al.*, 2005). *Rickettsia* genomes encode a number of paralogs for the ATP-binding cassette transporters and for the major facilitator superfamily transporters (*see* Article 81, **Transporter protein families**, Volume 6). Some of those transporters are predicted to be specific to amino acids and nucleotides. However, the substrate specificities of these transporters are mostly unknown.

3.3. Synchronization of replication

Most *Rickettsia* species are stably associated with their arthropod hosts for a long period of time. *Rickettsia* are thus thought to have developed a regulatory mechanism to synchronize their replication with that of the hosts. Genome analyses suggest that *spoT* and toxin–antitoxin (TA) systems could be the major players in this process (Ogata *et al.*, 2005a). Despite their reduced genome size, all the sequenced *Rickettsia* possess paralogous *spoT* genes. SpoT and RelA control the concentration of alarmone, (p)ppGpp (guanosine tetra- and pentaphosphates) in response to starvation in enterobacteria (Chatterji and Ojha, 2001). The alarmone in turn acts as an effector of transcription and changes the global cellular metabolism (i.e., "stringent response"). *Rickettsia* genomes contain many fragmented *spoT* paralogs (4–14 copies), some of which contain ppGpp hydrolase domain and others, which contain the ppGpp synthetase domain. In *R. felis*, all the 14 *spoT* genes were demonstrated to be transcribed (Ogata *et al.*, 2005a). In *R. conorii*, the transcription of five *spoT* genes are regulated depending on the stress or the type of infected hosts (Rovery *et al.*, 2005). The chromosomes of SFG *Rickettsia* and *R. bellii* harbor genes for TA systems (Ogata *et al.*, 2005a, 2006). Genes for TA systems were originally identified in bacterial plasmids. TA systems are

composed of toxin and antitoxin gene pairs, and ensure stable plasmid inheritance by a mechanism known as *postsegregational killing* when they are encoded in plasmids. Recent genome surveys showed that TA systems are also abundant in the chromosomes of bacteria (Gerdes *et al*., 2005). It is suggested that TA systems encoded in the bacterial chromosomes participate in the cascade of the stringent response pathway. Thus, the rickettsial TA systems might contribute to the global regulation of metabolism, eventually leading to selective killing (a primitive form of bacterial apoptosis) or reversible stasis of bacterial subpopulations during period of starvation or other stress (Gerdes, 2000; Engelberg-Kulka and Glaser, 1999). Alternatively, the rickettsial TA systems could be targeted to the host cells, and help the bacteria to persist in the host.

4. Plasmids and conjugation

Genome sequencing of *R. felis* were identified, besides a circular chromosome, two circular plasmids of 39 and 63 kb, respectively (Ogata *et al*., 2005a). The larger plasmid exhibits a cluster of several genes for conjugative DNA transfer. Conjugative plasmids have a role in the exchange of genes with other bacteria. Indeed, the *R. felis* genome bears traces of ancient gene acquisitions from nonrickettsial bacteria, along with gene transfers back-and-forth between the chromosome and the plasmids. The sequencing of the earliest diverging *R. bellii* later revealed a more complete gene cluster for conjugative DNA transfer (Ogata *et al*., 2006). By electron microscopy, sex pili-like surface appendages physically connecting rickettsial cells (i.e., mating pairs) were observed for both *R. felis* and *R. bellii*. The lack of genetic transformation tools for rickettsiae has hindered progress in the molecular characterization of the bacteria. The rickettsial plasmids and the conjugal DNA transfer genes are expected to provide a molecular basis for the development of a genetic transformation tool to study rickettsiae.

5. Genome evolution

Besides the massive amplification of transposase genes in the genomes of *R. bellii* and *R. felis*, few cases of recent lateral gene acquisition and gene duplication have been identified in *Rickettsia*. This is consistent with their protected and restricted intracellular niches. Hence, the majority of the variations between different *Rickettsia* genomes originate in the genome reduction that has degraded different sets of genes across different *Rickettsia* lineages. Except *R. bellii*, the genomes of *Rickettsia* exhibit remarkably similar overall gene orders. The long-range genome colinearity allowed precise identifications of orthologous loci between the genomes, and helped detailed analyses of gene degradation process. Comparison of the *R. prowazekii* and *R. conorii* genomes allowed the recognition of many gene remnants in the noncoding regions of the genomes, and suggested that most of the genome differences between the two genomes were derived by genome degradation (Ogata *et al*., 2001). Furthermore, the comparison led to the identification of various intermediate stages of gene deterioration process, including

fragmented genes and "split genes". The detection of mRNAs from those split genes suggests that some, if not all, of them may still retain their functions.

All *Rickettsia* genomes harbor palindromic sequences of about 100–150 bp, known as *Rickettsia palindromic elements* (RPEs) (Ogata *et al*., 2000, 2002). Remarkably, several families of RPEs exhibit a capability of inserting themselves within protein-coding genes and generating new peptide segments (30–50 aa) in the gene products. Recent analyses showed similar phenomena in *Wolbachia* (Ogata *et al*., 2005b) and *Methanocaldococcus jannaschii* (Suyama *et al*., 2005). It is hypothesized that such insertions of palindromic elements into genes might have contributed to the divergence of protein sequences in the course of bacterial evolution (Claverie and Ogata, 2003).

Despite the lack of evidence for recent gene transfers, *Rickettsia* genomes exhibit traces of ancient gene acquisitions especially from bacteria living in amoeba (Ogata *et al*., 2006). Phylogenetic analyses suggest that some of the genes associated with intracellular parasitism have been acquired from bacteria living in amoeba. For example, several authors suggest that the ancestral gene for the ATP/ADP translocase was transferred from *Parachlamydia* to *Rickettsia* (Greub and Raoult, 2003; Schmitz-Esser *et al*., 2004). The conjugal DNA transfer genes are most similar to those found in *Parachlamydia*. Furthermore, comparative sequence analyses revealed the abundance of rickettsial genes exhibiting a high level of sequence similarity to the homologs in amoeba-associated bacteria, including *Legionellaceae*, *Burkholderiaceae*, *Pseudomonadaceae*, *Parachlamydiaceae,* and *Coxiella*. The ancestor of *Rickettsia* might have lived within amoeba, given the capability of *R. bellii* to survive within amoeba and the finding of amebal symbonts related to *Rickettsia* (Fritsche *et al*., 1999). Since many amoeba-associated bacteria or their relatives are pathogens of humans, amoebae have been suggested to act as evolutionary "training grounds" for bacteria to prepare the ability to infect the cells of higher eukaryotes (Molmeret *et al*., 2005; Barker and Brown, 1994). Gene exchanges between these bacteria may have significantly contributed to their evolution, by conferring an immediate selective advantage in the adaptation to the intracellular environment of eukaryotic cells.

6. Conclusion

Genomics of *Rickettsia* has revealed remarkable diversities of their genomes, as recently illustrated by the discovery of the putative conjugative plasmid in *R. felis*. Further determinations of new *Rickettsia* genome sequences will provide more comprehensive pictures of the molecular and evolutionary diversity of these medically important intracellular bacteria.

7. Acknowledgements

We thank Prof. Jean-Michel Claverie (head of UPR 2589) for laboratory space and support.

References

Andersson JO and Andersson SG (1999) Genome degradation is an ongoing process in Rickettsia. *Molecular Biology and Evolution*, **16**, 1178–1191.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH and Kurland CG (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, **396**, 133–140.

Azad AF and Beard CB (1998) Rickettsial pathogens and their arthropod vectors. *Emerging Infectious Diseases*, **4**, 179–186.

Barker J and Brown MR (1994) Trojan horses of the microbial world: protozoa and the survival of bacterial pathogens in the environment. *Microbiology*, **140**(Pt 6), 1253–1259.

Blanc G, Ngwamidiba M, Ogata H, Fournier PE, Claverie JM and Raoult D (2005) Molecular evolution of Rickettsia surface antigens: evidence of positive selection. *Molecular Biology and Evolution*, **22**, 2073–2083.

Blatch GL and Lassle M (1999) The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *BioEssays*, **21**, 932–939.

Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology*, **4**, 160–165.

Claverie JM and Ogata H (2003) The insertion of palindromic repeats in the evolution of proteins. *Trends in Biochemical Sciences*, **28**, 75–80.

Driskell LO, Tucker AM, Winkler HH and Wood DO (2005) Rickettsial metK-encoded methionine adenosyltransferase expression in an Escherichia coli metK deletion strain. *Journal of Bacteriology*, **187**, 5719–5722.

Engelberg-Kulka H and Glaser G (1999) Addiction modules and programmed cell death and antideath in bacterial cultures. *Annual Review of Microbiology*, **53**, 43–70.

Fritsche TR, Horn M, Seyedirashti S, Gautom RK, Schleifer KH and Wagner M (1999) In situ detection of novel bacterial endosymbionts of Acanthamoeba spp. phylogenetically related to members of the order Rickettsiales. *Applied and Environmental Microbiology*, **65**, 206–212.

Gerdes K (2000) Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. *Journal of Bacteriology*, **182**, 561–572.

Gerdes K, Christensen SK and Lobner-Olesen A (2005) Prokaryotic toxin-antitoxin stress response loci. *Nature Reviews. Microbiology*, **3**, 371–382.

Gouin E, Egile C, Dehoux P, Villiers V, Adams J, Gertler F, Li R and Cossart P (2004) The RickA protein of Rickettsia conorii activates the Arp2/3 complex. *Nature*, **427**, 457–461.

Gouin E, Welch MD and Cossart P (2005) Actin-based motility of intracellular pathogens. *Current Opinion in Microbiology*, **8**, 35–45.

Greub G and Raoult D (2003) History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. *Applied and Environmental Microbiology*, **69**, 5530–5535.

Jeng RL, Goley ED, D'alessio JA, Chaga OY, Svitkina TM, Borisy GG, Heinzen RA and Welch MD (2004) A Rickettsia WASP-like protein activates the Arp2/3 complex and mediates actin-based motility. *Cellular Microbiology*, **6**, 761–769.

Li H and Walker DH (1998) rOmpA is a critical protein for the adhesion of Rickettsia rickettsii to host cells. *Microbial Pathogenesis*, **24**, 289–298.

McLeod MP, Qin X, Karpathy SE, Gioia J, Highlander SK, Fox GE, McNeill TZ, Jiang H, Muzny D, Jacob LS, *et al*. (2004) Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae. *Journal of Bacteriology*, **186**, 5842–5855.

Molmeret M, Horn M, Wagner M, Santic M and Abu Kwaik Y (2005) Amoebae as training grounds for intracellular bacterial pathogens. *Applied and Environmental Microbiology*, **71**, 20–28.

Mosavi LK, Minor DL and Peng Jr ZY (2002) Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16029–16034.

Ogata H, Audic S, Abergel C, Fournier PE and Claverie JM (2002) Protein coding palindromes are a unique but recurrent feature in Rickettsia. *Genome Research*, **12**, 808–816.

Ogata H, Audic S, Barbe V, Artiguenave F, Fournier P-E, Raoult D and Claverie J-M (2000) Selfish DNA in protein-coding genes of Rickettsia. *Science*, **290**, 347–350.

Ogata H, Audic S, Renesto-Audiffren P, Fournier P-E, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie J-M, *et al*. (2001) Mechanisms of evolution in Rickettsia conorii and R. prowazekii. *Science*, **293**, 2093–2098.

Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, Fournier PE, Claverie JM and Raoult D (2006) Genome sequence of Rickettsia bellii illuminates the role of amoeba in gene exchanges between intracellular pathogens. *PLoS Genetics* (submitted).

Ogata H, Renesto P, Audic S, Robert C, Blanc G, Fournier PE, Parinello H, Claverie JM and Raoult D (2005a) The genome sequence of Rickettsia felis identifies the first putative conjugative plasmid in an obligate intracellular parasite. *PLoS Biology*, **3**, e248.

Ogata H, Suhre K and Claverie JM (2005b) Discovery of protein-coding palindromic repeats in Wolbachia. *Trends in Microbiology*, **13**, 253–255.

Raoult D and Roux V (1997) Rickettsioses as paradigms of new or emerging infectious diseases. *Clinical Microbiology Reviews*, **10**, 694–719.

Rovery, C, Renesto, P, Crapoulet, N, Matsumoto, K, Parola, P, Ogata, H and Raoult, D (2005) Transcriptional response of Rickettsia conorii exposed to temperature variation and stress starvation. *Research in Microbiology*, **156**, 211–218.

Rubtsov AM and Lopina OD (2000) Ankyrins. *FEBS Letters*, **482**, 1–5.

Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M and Horn M (2004) ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae. *Journal of Bacteriology*, **186**, 683–691.

Suyama M, Lathe III WC and Bork P (2005) Palindromic repetitive DNA elements with coding potential in Methanocaldococcus jannaschii. *FEBS Letters*, **579**, 5281–5286.

Tucker AM, Winkler HH, Driskell LO and Wood DO (2003) S-adenosylmethionine transport in Rickettsia prowazekii. *Journal of Bacteriology*, **185**, 3031–3035.

Uchiyama T (2003) Adherence to and invasion of Vero cells by recombinant Escherichia coli expressing the outer membrane protein rOmpB of Rickettsia japonica. *Annals of the New York Academy of Sciences*, **990**, 585–590.

Whitworth T, Popov VL, Yu XJ, Walker DH and Bouyer DH (2005) Expression of the Rickettsia prowazekii pld or tlyC gene in Salmonella enterica serovar Typhimurium mediates phagosomal escape. *Infection and Immunity*, **73**, 6668–6673.

Short Specialist Review

Listeriae

Carmen Buchrieser and Philippe Glaser
Unité de Génomique des Microorganismes Pathogènes and CNRS URA 2171
Institut Pasteur, 75724 Paris Cedex 15, France

1. Introduction

Listeriae are gram-positive rods of low G + C content, which are present in a variety of animals and niches, including processed food. *Listeriae* are resistant to extreme conditions, such as high salt concentrations (10% NaCl), a broad pH range (from 4.5 to 9.0), and a wide temperature range (4–45°C), demonstrating a great adaptability to different environments. The genus *Listeria* contains six species, two of which are pathogenic: *Listeria monocytogenes*, the food-borne human pathogen responsible for listeriosis, on which this chapter is focused, and *Listeria ivanovii*, an animal pathogen (Vazquez-Boland *et al*., 2001).

L. monocytogenes is the causative agent of listeriosis, a systemic bacterial infection that causes miscarriage in pregnant women and that is often fatal to immunocompromised individuals (Farber and Peterkin, 1991; Jurado *et al*., 1993). In addition, *L. monocytogenes* is a leading cause of meningitis in neonates and the elderly. The mortality rates can be as high as 30%, and the fetal death rate caused by *Listeria* is most likely higher due to an unknown number of unexplained abortions attributable to *Listeria*. Recently *L. monocytogenes* was also shown to cause gastroenteritis (Aureli *et al*., 2000; Dalton *et al*., 1997; Riedo *et al*., 1994).

The ability of *Listeria* to grow at low temperatures and under high salt conditions allows its spread by contaminated food, such as soft cheese and meat products, even when they have been properly stored. Several large outbreaks of listeriosis have been traced back to food production and packaging plants, demonstrating that this organism is well suited to exploit weaknesses of the food distribution system.

Listeria disseminates from the intestinal lumen to the central nervous system and the feto-placental unit. In the infected host, *Listeria* survives to host defenses owing mainly to its resistance to phagocyte killing and to its capacity to invade nonphagocytic cells where it can replicate. The cell biology of the infectious process has been investigated in detail and *L. monocytogenes* has become a paradigm to study intracellular parasitism (Cossart and Lecuit, 1998).

Despite the elucidation of several players necessary for entry and intracellular replication of *L. monocytogenes* (Vazquez-Boland *et al*., 2001) many open

questions remain, like “How does *L. monocytogenes* cross the blood-brain barrier?” or “What is the genetic basis of virulence differences among different *L. monocytogenes* strains?” or “What are the specific features allowing *Listeria* to grow at low temperatures and thus to replicate in refrigerated food?”

With the aim to answer these questions a comparative genomics approach was applied by a consortium of 10 European laboratories, by sequencing the genome of the pathogen *L. monocytogenes* (strain EGDe) and the closely related nonpathogenic species *Listeria innocua* (strain CLIP11626) (Glaser *et al*., 2001) (**http://genolist.pasteur.fr/ListiList/**). This comparison revealed that all known virulence factors are absent from the apathogenic, closely related *L. innocua* and pointed to new putative virulence genes, which were analyzed functionally during the last years. Additional *Listeria* genome sequences were obtained three years later by TIGR in collaboration with the FDA (Nelson *et al*., 2004), which now allow the comparison of several different isolates from the same species. At present there are three complete and two partial, published *Listeria* sequences available (Table 1). *Listeriae* contain one circular chromosome of about 2.9 Mb and an average G + C content of 39%. Two of the sequenced strains (*L. monocytogenes* H7858 and *L. innocua*) contain a plasmid of about 80 kb in size that resemble each other and the plasmid pOX2 identified in *Bacillus anthracis* (Read *et al*., 2002).

Table 1 General features of published *Listeria* genome sequences

	L. monocytogenes EGDe (1/2a)	*L. monocytogenes* F6854 (1/2a)[c]	*L. monocytogenes* F2365 (4b)	*L. monocytogenes* H7858 (4b)[c]	*L. innocua* CLIP11262 (6a)
Size of the chromosome (bp)	2 944 528	~2 953 211	2 905 310	~2 893 921	3 011 209
G + C content (%)	38	37.8	38	38	37.4
G + C content (%)	38.4	38.5	38.5	38.4	38.0
Total number of CDS[a]	2853	2973	2847	3024	2970
Percentage coding	89.2%	90.3%	88.4%	89.5%	89.1%
Number of prophage regions	1	3	2	2	5
Monocins	1	1	1	1	1
Plasmid	–	–	–	1 (94 CDS)	1 (79 CDS)
Number of strain specific genes[b]	61	97[c]	51	69[c]	78
Number of transposons	1 (Tn916 like)	–	–	–	–
Number of rRNA operons	6	6	6	6	6
Number of tRNA genes	67	67[c]	67	65[c]	66

[a]CDS = coding sequence.
[b]Except prophage genes.
[c]Draft genome sequence (8-fold coverage without gap closure).

2. Genome organization and comparative genomics

The *Listeria* genomes have a conserved, colinear genome organization with around 60–100 regions of one to several Kbs, specific for each species, scattered around the chromosome and one to five (*L. innocua*) prophage regions. *L. innocua* and all *L. monocytogenes* except the serotype 4b strain (F2365) contain a phage of the A118 family (Glaser *et al*., 2001; Loessner *et al*., 2000) inserted in the *comK* gene. Furthermore each sequenced strain carries a monocin region. However, phages do not seem to play a role in virulence acquisition in *Listeria* as they are not carrying known or putative virulence genes and are not conserved among strains.

In spite of this mosaic structure, the three completely sequenced *Listeria* genomes show a very strong conservation in genome organization with no inversions or shifts of large genome segments (Figure 1) (Buchrieser *et al*., 2003). This conserved genome organization may be related to the low occurrence of

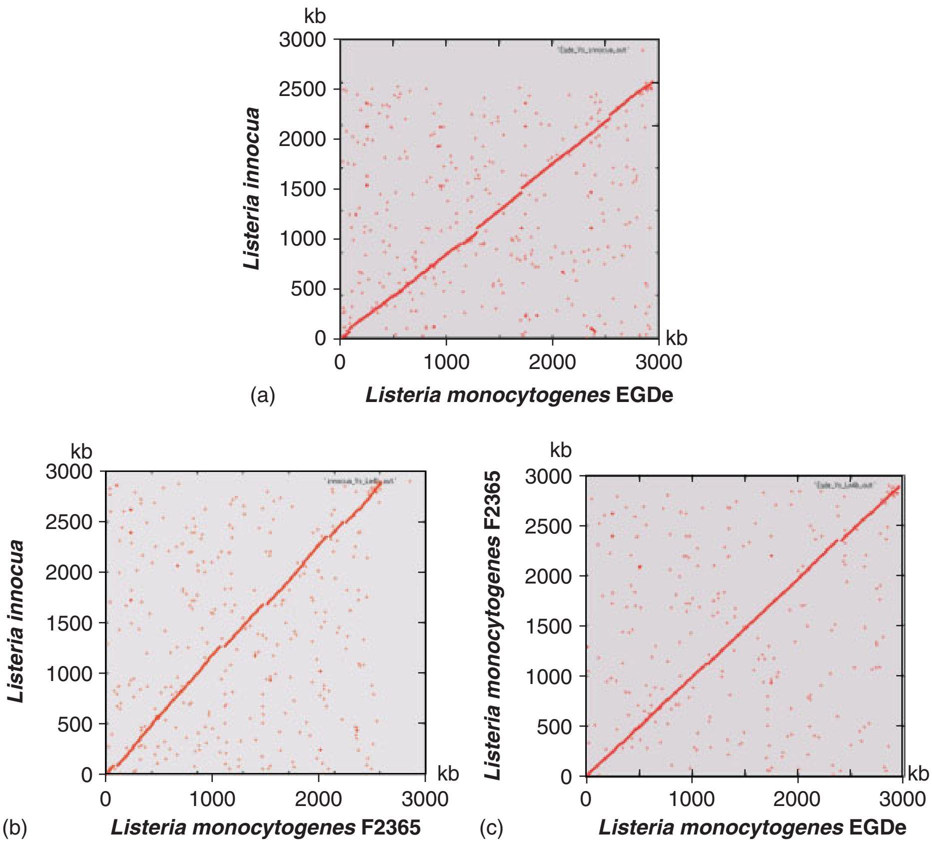


Figure 1 Synteny between (a) *Listeria monocytogenes* EGDe and *Listeria innocua* CLIP11262, (b) *L. innocua* CLIP11262 and *L. monocytogenes* F2365, (c) *L. monocytogenes* F2365 and *L. monocytogenes* EGDe. Orthologous genes were defined by bidirectional best hits based on Blastp comparisons

insertion sequences (IS) elements, suggesting that IS transposition or IS mediated deletions are not key evolutionary mechanisms in *Listeria*. The chromosome of the serotype 4b strains (F2365 and H7858) lack intact IS but do contain four transposases of the IS3 family that are present in homologous locations in both strains. The serotype 1/2a strains (F6854 and EGDe) each contain three while *L. innocua* contains four of these elements. Furthermore, the serotype 1/2a strains contain an intact IS element named ISLmo1, two are present in strain F6854 and three, one of which is not intact, are present in EGDe. ISLmo1 is missing in the serotype 4b and the *L. innocua* strains. The *Listeria* genomes contain putative DNA uptake genes, homologous to *Bacillus subtilis* competence genes. Thus competence may play a role in the evolution of *Listeria*.

3. Specific features of the *Listeria* genomes

The most striking features of the *Listeria* genomes are an exceptionally large number of surface proteins, an abundance of transport proteins, in particular, proteins, dedicated to carbohydrate transport, and an extensive regulatory repertoire. These are characteristics that are undoubtedly conferring on *Listeria* their adaptability to many diverse environments, like soil, water, effluents, a large variety of foods, and eukaryotic cells.

1. *Surface proteins*: Surface proteins have important roles in the interactions of microorganisms with their environments, in particular, during host infection. The *Listeria* genomes encode many such proteins, for example, 4.7% of all predicted genes of *L. monocytogenes* EGDe. The largest surface protein family are lipoproteins and the second largest family are LPXTG proteins including the internalin family. Major virulence factors of *L. monocytogenes* belong to the Internalin family, like Internalin (InlA) and InlB necessary to enter eukaryotic cells. Although the major known virulence factors are conserved, there is a pronounced diversity within the surface proteins of the different strains. For example, when comparing *L. monocytogenes* EGDe (serovar 1/2a) to *L. monocytogenes* F2635 (serotype 4b) five lipoproteins, nine LPXTG proteins and two autolysins are specific to the serotype 4b strain. As another example, among the 41 LPXTG proteins identified in *L. monocytogenes* EGDe, 11 are absent from *L. innocua* CLIP11262. *L. innocua* CLIP11262 codes for 34 LPXTG proteins, 14 of which are absent from *L. monocytogenes* EGDe (Cabanes *et al*., 2002; Glaser *et al*., 2001). This pronounced diversity among LPXTG proteins was further substantiated by a comparative genomics study using DNA/DNA array hybridization. The distribution of 55 genes coding for putative surface proteins belonging to two sequenced *L. monocytogenes* genomes (serotype 1/2a and 4b) and *L. innocua* CLIP11262 was investigated among 93 *L. monocytogenes* and 20 *Listeria* sp. strains. The analysis identified 25 surface protein-coding genes as specific for the species *L. monocytogenes* including *inlAB*. Others are specific for certain subgroups of strains or for *L. innocua* (Doumith *et al*., 2004). The fact that different subgroups

of *L. monocytogenes* strains contain different sets of surface proteins may reflect their different potential to cause disease or to multiply in different niches.

2. *Transporters*: The *Listeria* genomes encode an abundance of transport proteins (e.g., 11.6% of all predicted genes of *L. monocytogenes* EGDe). These comprise, in particular, proteins dedicated to carbohydrate transport conferring on *Listeria* probably in part its ability to colonize a broad range of ecosystems. The overall array of sugar transporters is similar in all *Listeria* genomes, in particular, among the four sequenced *L. monocytogenes* strains, but also in *L. innocua*. *Listeria* are predicted to transport and metabolize many simple as well as complex sugars including fructose, rhamnose, rhamnulose, glucose, mannose, chitin, sucrose, cellulose, pullan, trehalose, and tagatose. These sugars are largely associated with the environments where *Listeriae* are found. As in most bacterial genomes the predominant class corresponds to ABC transporters. Interestingly, most of the carbohydrate transport proteins belong to phosphoenolpyruvate-dependent phosphotransferase system (PTS) mediated carbohydrate transport genes. The PTS allows the use of different carbon sources and in many bacteria studied so far, the PTS is a crucial link between metabolism and regulation of catabolic operons (Barabote and Saier, 2005; Kotrba *et al*., 2001). The *Listeria* genomes contain an unusually large number of PTS loci (e.g., nearly twice as many as *Escherichia coli* and nearly three times as many as *B. subtilis*). Most of these PTS systems are conserved in the different sequenced genomes, however, subtle differences can be observed, probably allowing niche specific adaptation. An example is the family of β-glucoside-specific PTSs, of which eight are present in *L. monocytogenes* serotype 1/2a; two of those are missing in the *L. monocytogenes* serotype 4b strains and five are missing from *L. innocua*. As one of these β-glucoside-specific PTS systems, named *BvrABC*, was shown to be implicated in virulence of *L. monocytogenes* (Brehm *et al*., 1996) these differences might play a role in virulence differences among strains.
3. *Regulators*: Given the fact that *L. monocytogenes* is a ubiquitous, opportunistic pathogen that needs a variety of combinatorial pathways to adapt its metabolism to a given niche, an extensive regulatory repertoire is needed. Indeed, a little more than 7% of the *Listeria* genes predicted in the genomes are dedicated to regulatory proteins. Listeria have almost twice as many regulators than *Staphylococcus aureus* despite the similar genome size. Only *Pseudomonas aeruginosa* (Stover *et al*., 2000) another ubiquitous, opportunistic pathogen encodes more, with over 8% of its predicted genes being regulatory proteins. Diversity among the *Listeria* genomes with respect to regulatory genes is not as pronounced as the differences identified in the surface protein repertoire, suggesting they function primarily in specific niches common to the lifestyle of *Listeriae* outside a mammalian host. The most studied regulatory gene of *L. monocytogenes* is *prfA*, encoding the master regulator of virulence. In line with its function in regulating the expression of genes encoding proteins necessary for the entry and for intracellular multiplication of *L. monocytogenes*, PrfA is absent form *L. innocua* but conserved in all *L. monocytogenes* strains.

4. Concluding remarks and future directions

Sequencing the genomes of four *L. monocytogenes* strains and one *L. innocua* strain has brought a wealth of new data and insight into this fascinating pathogen. Now different genomics approaches are applied and provide for the first time a global view and a more and more complete knowledge on gene distribution and the genetic content present in the gene pool of the genus *Listeria*. This information represents a fundamental basis for functional studies to better understand phenotypic and virulence differences between *L. monocytogenes* strains and to gain knowledge on the evolution of the pathogen. The ongoing sequencing projects aimed at determining the complete genome sequence of one representative of each species of the genus *Listeria* (*L. ivanovii*, *L. seeligeri*, *L. welshimeri* and *L. grayi* strains) by the Institut Pasteur and the German PathoGenomiK network (**http://www.pasteur.fr/recherche/unites/gmp/**; **http://www.genomik.uni-wuerzburg.de/seq.htm**) and the determination of the complete genome sequence of an additional 19 *Listeria* strains by the Broad Institute (**http://www.broad.mit.edu/seq/msc/**) will be the driving force for understanding the function of the many factors encoded by the genome, whether involved or not in virulence, and to understand strain specific differences in niche adaptation and virulence.

References

Aureli P, Fiorucci GC, Caroli D, Marchiaro G, Novara O, Leone L and Salmaso S (2000) An outbreak of febrile gastroenteritis associated with corn contaminated by Listeria monocytogenes. *The New England Journal of Medicine*, **342**, 1236–1241.

Barabote RD and Saier MH (2005) Comparative genomic analyses of the bacterial phosphotransferase system. *Microbiology and Molecular Biology Reviews*, **69**, 608–634.

Brehm K, Kreft J, Ripio MT and Vazquez-Boland JA (1996) Regulation of virulence gene expression in pathogenic *Listeria*. *Microbiologia*, **12**, 219–236.

Buchrieser C, Rusniok C, Kunst F, Cossart P and Glaser P (2003) Comparison of the genome sequences of *Listeria monocytogenes* and *Listeria innocua*: clues for evolution and pathogenicity. *FEMS Immunology and Medical Microbiology*, **35**, 207–213.

Cabanes D, Dehoux P, Dussurget O, Frangeul L and Cossart P (2002) Surface proteins and the pathogenic potential of *Listeria monocytogenes*. *Trends in Microbiology*, **5**, 238–245.

Cossart P and Lecuit M (1998) Interactions of *Listeria monocytogenes* with mammalian cells during entry and actin-based movement: bacterial factors, cellular ligands and signaling. *The EMBO Journal*, **17**, 3797–3806.

Dalton CB, Austyin CC, Sobel J, Hayes PS, Bibb WF, Graves LM, Swaminathan B, Proctor ME and Griffin PM (1997) An outbreak of gastroenteritis and fever due to *Listeria monocytogenes* in milk. *The New England Journal of Medicine*, **336**, 100–105.

Doumith M, Cazalet C, Simoes N, Frangeul L, Jaquet C, Kunst F, Martin P, Cossart P, Glaser P and Buchrieser C (2004) New aspects regarding evolution and virulence of *Listeria monocytogenes* revealed by comparative genomics. *Infection and Immunity*, **72**, 1072–1083.

Farber JM and Peterkin PI (1991) *Listeria monocytogenes*, a food-borne pathogen. *Microbiological Reviews*, **55**, 476–511.

Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, *et al*. (2001) Comparative genomics of *Listeria* species. *Science*, **294**, 849–852.

Jurado RL, Farley MM, Pereira E, Harvey RC, Schuchat A, Wenger JD and Stephens DS (1993) Increased risk of meningitis and bacteremia due to *Listeria monocytogenes* in patients with human immunodeficiency virus infection. *Clinical Infectious Diseases*, **17**, 224–227.

Kotrba P, Inui M and Yukawa H (2001) Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism. *Journal of Bioscience and Bioengineering*, **92**, 502–517.

Loessner MJ, Inman RB, Lauer P and Calendar R (2000) Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of *Listeria monocytogenes*: implications for phage evolution. *Molecular Microbiology*, **2**, 324–340.

Nelson KE, Fouts D, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, *et al*. (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. *Nucleic Acids Research*, **32**, 2386–2395.

Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, *et al*. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033 [Epub 2002 May 2029].

Riedo FX, Pinner RW, Tosca ML, Cartter ML, Graves LM, Reeves MW, Weaver RE, Plikaytis BD and Broome CV (1994) A point-source foodborne listeriosis outbreak: documented incubation period and possible mild illness. *The Journal of Infectious Diseases*, **170**, 693–696.

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, *et al*. (2000) Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.

Vazquez-Boland JA, Kuhn M, Berche P, Chakraborty T, Dominguez-Bernal G, Goebel W, Gonzalez-Zorn B, Wehland J and Kreft J (2001) *Listeria* pathogenesis and molecular virulence determinants. *Clinical Microbiology Reviews*, **14**, 1–57.

Introductory Review

History of genetic mapping

Newton E. Morton
University of Southampton, Southampton, UK

1. Genetic and physical maps

Examples of linkage were reported in the first decade of the last century, but it was not until 1913 that Sturtevant elaborated the concept of linear arrangement in a *linkage map.* Recognition of double crossing-over and interference led to the understanding that linear arrangement implies not merely order of loci but the additivity of their distances. In 1919, Haldane introduced the Morgan (=100 centimorgans) as the length of a chromatid that on the average has experienced one crossover event per meiosis. This concept is the basis of genetic maps that have been adopted for all organisms. Over nearly a century, geneticists have become familiar with genetic distances and chromosome lengths in centimorgans (cM), and have reached substantial consensus about methods and results. On the contrary, a second type of *genetic map* with the essential properties of linearity and additivity is a recent invention still in rapid evolution (Maniatis *et al*., 2002). A *linkage disequilibrium map* determines the distance between pairs of loci not from recombination but from allelic association, also called linkage disequilibrium (LD), which is measured in linkage disequilibrium units (LDU). One LDU corresponds to the length of a chromatid in which on the average one crossover event has taken place in t generations, and so the resolution is t times as great as for the linkage map. This greatly increases the power to localize a gene within a candidate region at the expense of typing a larger number of markers and using methods that at this early stage are less transparent to many geneticists.

These two genetic maps (linkage and LD) do not exhaust the set of useful maps that are linear with additive distances. *Physical maps* (q.v.) depend not on crossing-over but on distance in the DNA sequence, ideally measured in base pairs (bp) or kilobases (kb). The closest approach to this ideal is by the DNA sequence nominally finished in 2003, although many relatively small errors remain, and of course, polymorphisms affecting the DNA sequence are omitted. The utility of *radiation hybrid maps*, based not on homologous crossing-over but on chromosome breakage, is limited to organisms without a finished DNA sequence. Finally, *chromosome bands* provide mutually exclusive and jointly exhaustive projections of genetic or cytogenetic candidate regions on the physical map as the first step in identification of genes that affect a phenotype of interest but are of unknown sequence and function. Their position must be refined on the genetic map before functional and physical studies become feasible.

2. Linkage mapping

The theory and practice of genetic mapping were established in experimental organisms before Bernstein (1931) realized that linkage could be detected in human pedigrees without knowing haplotypes. Geneticists who understood the importance of this insight made various modifications of his method, culminating in the elegant maximum likelihood u scores of Fisher (1935) and Finney (1940). Although fully efficient in the limit for loose linkage, only the asymptotic theory is known and it can be misleading in small samples. The scores are difficult to calculate except in simple two-generation pedigrees, and do not accommodate matings of known phase. They are inefficient for close linkage and do not estimate the recombination fraction accurately. It was not recognized that the low prior probability for linkage of two random loci implies an unusually stringent significant level to support a convincing claim of linkage.

In the early fifties, u scores were replaced by logarithms of odds (lods) based on the likelihood ratio of Neyman and Pearson (1928), as developed by Barnard (1949). The first application to human linkage was made by Haldane and Smith (1947) and extended by Smith (1953). Lods are simple, efficient, exact even in small samples, relatively easy to compute in large pedigrees, additive over samples, and provide a maximum likelihood estimate of recombination. It was inevitable that lods would replace u scores, but the question was whether they would take the Bayesian direction advocated by Barnard and Smith, which requires choice of a prior distribution and sacrifices additivity of scores. Wald (1947) provided a simpler approach that was applied to human linkage in the Ph.D. thesis of Morton (1955), who had experienced the limitations of u scores while analyzing pedigrees in Japan. A probability distribution for recombination was deduced that determined an appropriate significance level, an approach now used to control the false discovery rate (FDR) for both genome scans and functional tests (Storey and Tibshirani, 2003). The first application of lods showed that dominant elliptocytosis is caused by different mutations in different pedigrees (Morton, 1956). This encouraged the construction of genetic maps based initially on blood groups and isozymes as markers to localize an exponentially increasing number of disease loci. Sex-specific recombination was recognized (Smith, 1954; Renwick, 1968) and the power of heterogeneity tests was increased (Smith, 1963). Algorithms to consider multiple markers simultaneously and to reduce computational time in complex pedigrees were introduced (reviewed in Thompson, 2001). Theory that had no practical consequences before the DNA revolution became valuable as the number of markers increased. For example, Smith (1953) derived likelihoods for individuals homozygous for a rare recessive allele causing disease. This *autozygosity mapping*, a variant of linkage mapping in which the allele is not merely homozygous but identical by descent (loosely called homozygosity mapping), was not exploited until rediscovered by Lander and Botstein (1987) and shown to be a powerful method to localize such genes. DNA typing revealed great heterogeneity in most diseases and became the driving force for continued development of genetic mapping and ultimately for the physical map of the human genome.

Until near the end of the last century, most genes localized by genetic mapping were of large effect with high penetrance, low gene frequency, and dominance or

recessivity that could be reliably inferred by segregation analysis. The few attempts to detect oligogenes that are more common and of lower penetrance used methods that had been developed when the number of polymorphic markers was small (Penrose, 1935; Haseman and Elston, 1972). The models are weakly parametric, with the multiple parameters for gene frequency, penetrance, and interaction summarized by a smaller number of variance components. Ascertainment is often complex, segregation analysis inconclusive, and claims of linkage unconfirmed. The candidate region typically covers many centimorgans and therefore many loci. The relatively small number of confirmed locus identifications is a tribute both to the tirelessness of investigators and their subsequent use of LD and functional tests to identify a causal gene within a coarse candidate region suggested by linkage.

3. Association mapping

Linkage mapping depends on inference of recombination between two haplotypes in a single generation, interpreting pedigrees by a linkage map. *Association mapping* depends on the pattern of LD in single haplotypes, interpreted by an LD map. Currently, both approaches take as their prime objective the identification of a gene by determining a candidate region in the genome in the absence of information about the structure or function of the gene product. Such a candidate region may be determined by linkage, LD, or cytogenetic observation of a chromosomal aberration in probands. At that point, linkage has too little resolution to localize a gene, leaving LD as the method of choice until the candidate region is reduced to a few loci distinguished by functional tests (especially for a candidate locus), which share with LD the problem of distinguishing between association and causation.

During most of the last century, LD was a theoretical problem unrelated to association mapping, which even the most creative geneticists had not imagined. Robbins (1918) developed an evolutionary theory for LD between pairs of neutral loci, later extended to estimation of haplotypes from diplotypes (Hill, 1974). Sved (1971) developed a neutral theory conditional on identity by descent for pairs of diallelic markers, following Malecot (1948). The long interval between Robbins and Hill was dominated by the belief that LD is maintained through selection operating on all polymorphisms, of which only a few were then known (Ford, 1940). On this assumption, no useful theory of polymorphism could be developed without imagining the effects of selection. This impasse was broken by the discovery of many molecular polymorphisms, for most of which the neutral theory of Kimura (1968) was appropriate. Chakravarti (1984) used restriction fragment length polymorphisms (RFLPs), the first fruit of the DNA revolution, to infer a hot spot of recombination from an LD cold spot in the beta hemoglobin locus. Kerem *et al*. (1989) used LD to localize *CFTR*, the gene for cystic fibrosis, which stimulated Terwilliger (1995) to introduce composite likelihood (the product of dependent probabilities) as a mapping tool. In the absence of a relevant evolutionary theory, he was able to locate the *CFTR* gene, although the most frequent causal marker (ΔF508) lay outside his support interval. Devlin and Risch (1995) recognized that ascertainment bias should be allowed for the case-control design that is appropriate to rare genes. Their δ method is not applicable to more common

genes, because it assumes that the marker allele frequency approaches zero and the enrichment probability approaches ∞. The time was ripe for a more general solution. Collins and Morton (1998) introduced the association metric ρ, initially applied to rare major genes. Later extensions showed that the 20 years it took to localize the hemochromatosis gene (*HFE*) was a consequence of an unrecognized LD hot spot complementary to the LD cold spot noted earlier by Chakravarti, but understanding of blocks and steps came later (*see* Article 73, **Creating LD maps of the genome**, Volume 4). Daly *et al*. (2001) showed that most haplotypes do not recombine at LD cold spots, where the number of inferred recombinants is highly variable. Consequently, there are no natural haplosets within which each haplotype is delimited by recombination at the same pair of steps. A reasonable conjecture is that an unknown proportion of steps represent recombination hot spots dependent on DNA sequence, while the remainder occur independently and at random. Distinguishing rigorously between the two modes of origin is impossible without other evidence, which Jeffreys *et al*. (2001) provided by meiotic recombination. The large steps in their study represent small intervals with high crossing-over. A beginning has been made in identifying sequences that predispose to recombination (Tapper *et al*., 2003).

As soon as the LD fine structure was recognized, it gave birth to maps in additive LDU constructed from the association metric ρ that provides a basis for association mapping, in which they play the same role as linkage maps but at much greater resolution (Maniatis *et al*., 2002). LD maps are most reliable at high density, vary among populations, and increase in length with gene frequency and the number of generations since the last population bottleneck (Lonjou *et al*., 2003). Whereas microsatellites are currently most used for linkage analysis because of their high heterozygosity, single nucleotide polymorphisms (SNPs) are more useful for association analysis because of their much greater number and ease of typing. For brevity, it is customary to pool duplications, deletions, and insertions with the much larger number of SNPs. An international effort called HapMap is being made to create an SNP database that will be useful for many purposes, including creation of LD maps and choice of markers for association mapping (Couzin, 2002). A topic so young that it is missing from every book on genetics, bioinformatics, genomics, and statistics is being pursued by hundreds of investigators who may disagree about the means but not the importance of association mapping.

4. The future of genetic mapping

Linkage has such a long history that it is difficult to see what lies ahead. Meiotic recombination of SNPs in small candidate regions promises to recognize recombinogenic sequences, but extension to a whole chromosome would at present be a technical tour de force. It would identify recombination hot spots and give a linkage map at much greater precision than current maps, but this would not increase their resolution for gene localization in the absence of meiotic expression. Combined segregation and linkage analysis with credible allowance for ascertainment bias is so far restricted to major loci and single markers (Shields *et al*., 1994), but extension to multiple markers is possible and would be more

useful than continued modification of weakly parametric models that subsume gene frequency and other genetic parameters in a variance component.

Association mapping is so young that it has many more directions to expand. Composite likelihood remains a serious burden, but coalescence models have not provided the power that Monte Carlo simulation gives to genome scans by linkage. Nevertheless, genome scans by allelic association have much greater resolution than linkage, albeit at the cost of massive automation beyond the reach of most academics. It is an open question how long conventional genome scanning with microsatellites can compete with SNPs, which already dominate association mapping once a candidate region is identified. At present two-dimensional color graphics representing haplotypes, of untested utility for association mapping, are in competition with the additive distances of maps. These graphics depend entirely on association between adjacent SNPs and are therefore more sensitive to sample size and SNP density than LD maps based on covariances with neighboring SNPs (Ke *et al*., 2004). Undoubtedly, the suboptimal algorithms to construct LD maps will be improved, but it is not clear how haplotype graphs can be significantly improved. HapMap, faced with the problem of providing SNP data of general utility, has not yet decided whether to present its data as maps. Meanwhile, map determinants are being studied. Unlike linkage maps that have not been shown to vary among populations, LD maps are sensitive to demographic factors and especially to time since the last major bottleneck. Most but not all of the information in a local map can be extracted by fitting the Malecot model to a cosmopolitan map obtained by pooling samples from a larger region, even globally. A local map at low resolution, such as might be constructed for a genome scan, has intervals (called "holes") where the length is indeterminate and is therefore less reliable than a cosmopolitan map. Conversely, a local map at higher resolution, such as would be used to identify causal SNPs within a candidate region, may be substantially more reliable than a cosmopolitan map.

Construction of LD maps and their use for association mapping of disease susceptibility have the excitement of linkage mapping nearly a century ago. In the last stage of association mapping, and perhaps earlier, it will be complemented by advances in functional analysis aimed at causal SNPs. The haplotypes that carry a causal SNP are confounding variables that provide no additional information for its recognition. Once a causal SNP is identified, the haplotypes that carry it are of even less interest. In the last stage of association mapping, LD is superseded by functional tests in which presumptive causal SNPs are discriminated with allowance for association. These expression tests are, by definition, directed not at haplotypes but at causal SNPs, posing exactly the same disjunction between association and causation. The time is approaching when linkage, LD, and expression will be inseparable in the history of genetic mapping.

Further reading

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between loci of linked factors. *Journal of Genetics*, **8**, 299–309.

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *The Journal of Experimental Zoology*, **14**, 43–59.

References

Barnard GA (1949) Statistical inference. *Journal of the Royal Statistical Society*, **B11**, 115–130.

Bernstein F (1931) Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen mit besonderer Berucksichtigung der Blutgruppen. *Z. Induktive Abstammungs. u. Vererbungslehre*, **57**, 113–138.

Chakravarti A, Buetow KH, Antonarakis SE, Waller PG, Boehm CD and Kazazian HH (1984) Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, **36**, 1239.

Collins A and Morton NE (1998) Mapping a disease gene by allelic association. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1741–1745.

Couzin J (2002) Human genome HapMap launched with pledges of $100 million. *Science*, **298**, 941–942.

Daly M (2001) High resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Devlin D and Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–322.

Finney DJ (1940) The detection of linkage. *Annals of Eugenics*, **10**, 171–214.

Fisher RA (1935) The detection of linkage with "dominant" abnormalities. *Annals of Eugenics*, **6**, 187–201.

Ford EB (1940) Polymorphism and taxonomy. In *The New Systematics*, Huxley J (Ed.), Clarendon Press, Oxford, pp. 493–513.

Haldane JBS and Smith CAB (1947) A new estimate of the linkage between the genes for haemophilia and colour-blindness in man. *Annals of Eugenics*, **14**, 10–31.

Haseman JK and Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, **2**, 3–19.

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**, 217–222.

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Wittacker P, Collins A, Morris AP, Bentley D, *et al*. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, **13**, 577–588.

Kerem B, Rommers JM, Buchanan JA, Markiewicz CTK, Chakravarti A, Buchwald M and Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, 1073–1080.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Lander ES and Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Lonjou C, Zhang W, Collins A, Tapper W, Elahi E, Maniatis N and Morton NE (2003) Linkage disequilibrium in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6069–6074.

Malecot G (1948) *Les Mathématiques de l'hérédité, Masson & Cie, Paris; The impact of this work is reviewed in Modern Developments in Theoretical Population Genetics*, Slatkin M. and Veuille M. (Eds.), (2002), Oxford University Press Oxford.

Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X and Morton NE (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2228–2233.

Morton NE (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics*, **7**, 277–318.

Morton NE (1956) The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *American Journal of Human Genetics*, **8**, 80–96.

Neyman J and Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical reference. *Biometrika*, **20A**, 175–240, 263–294.

Penrose L (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics*, **6**, 133–138.

Renwick JH (1968) Ratio of female to male recombination fraction in man. *Bulletin of the European Society of Human Genetics*, **2**, 7–12.

Robbins RB (1918) Some applications of mathematics to breeding problems *III*. *Genetics*, **3**, 375–389.

Shields DC, Ratanachaiyavong S, McGregor AM, Collins A and Morton NE (1994) Combined segregation and linkage analysis of Graves disease with a thyroid autoantibody diathesis. *American Journal of Human Genetics*, **55**, 540–554.

Smith CAB (1953) The detection of linkage in human genetics (with discussion). *Journal of the Royal Statistical Society*, **B14**, 153–192.

Smith CAB (1954) The separation of the sexes of parents in the detection of human linkage. *Annals of Eugenics*, **18**, 278–301.

Smith CAB (1963) Testing for heterogeneity of recombination fraction in human pedigrees. *Annals of Human Biology*, **27**, 175–182.

Storey JD and Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, **2**, 125–141.

Tapper WJ, Maniatis N, Morton NE and Collins A (2003) A metric linkage disequilibrium map of a human chromosome. *Annals of Human Genetics*, **67**, 1–8.

Terwilliger J (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, **56**, 777–787.

Thompson EA (2001) Linkage analysis In *Handbook of Statistical Genetics*, Balding DJ, Bishop M and Cannings C (Eds.), John Wiley & Sons, Chichester, pp. 544–563.

Wald A (1947) *Sequential Analysis*, Wiley, New York.

Introductory Review

Normal DNA sequence variations in humans

Kenneth K. Kidd

Yale University, New Haven, CT, USA

1. Introduction

For the past century, researchers have been identifying normal genetic variation and studying that variation in diverse human populations. However, population geneticists had relatively few genetic markers to use to study normal populations globally – in many parts of the world the markers were either not useful (polymorphic) or had not been studied. A breakthrough in human population genetics occurred with the demonstration in 1966 by Harry Harris that many red cell enzymes and serum proteins showed electrophoretic variation and could be used for population genetic studies (Harris, 1966). Human populations from around the world have subsequently been studied for many of those normal polymorphisms, and a large compendium of the data collected over the ensuing quarter century has been published (Cavalli-Sforza *et al*., 1994).

Those markers are now generally considered collectively as the "classical" polymorphisms because, starting in 1978, a new surge in identification of polymorphisms, this time directly in the DNA, began with the discovery of a DNA sequence polymorphism near the beta-globin gene (Kan and Dozy, 1978). By 1980, it was obvious that enough DNA-based polymorphisms would be available to generate a linkage map of *Homo sapiens* (Botstein *et al*., 1980). It was also clear that the limitations in human population genetics that resulted from relatively few classical markers being highly variable in all parts of the world would be overcome by the large numbers of DNA markers. There has been a near exponential increase in the numbers of DNA polymorphisms since then such that now more than 10 million are cataloged in dbSNP (Sherry *et al*., 2001), of which over 5 million are validated by two or more independent discoveries (build 123). Other databases with subsets of the information in dbSNP and other types of information that researchers find useful also catalog DNA sequence variation: HGVBase (Fredman *et al*., 2002) and JSNP (Hirakawa *et al*., 2002). One database, ALFRED, the ALlele FREquency Database (Osier *et al*., 2002; Rajeevan *et al*., 2003), is specifically devoted to cataloging allele frequencies for DNA-based polymorphisms in defined populations. None of those sources of information catalogs all of the polymorphisms of any one type that are currently known and described somewhere in the scientific literature,

but any one of them will identify a very large number of polymorphisms. Now the challenge is to understand this variation: Grand Challenge I-3 formulated by the National Human Genome Research Institute is "Develop a detailed understanding of the heritable variation in the human genome" (Collins *et al*., 2003). Understanding that variation has many components. Here, we focus on the description of the types of DNA sequence variation and on the pattern of variation around the world.

2. Types of variation

Two groups of DNA sequence variants can be distinguished – those that are by their nature diallelic and those that tend to be multiallelic. Among the diallelic polymorphisms are single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels). The multiallelic polymorphisms include micro- and minisatellites and haplotyped loci. Haplotyped systems consist of two or more polymorphisms within a small segment of DNA. In addition, sequence variants on mitochondrial DNA (mtDNA) (*see* Article 5, **Studies of human genetic history using mtDNA variation**, Volume 1) and the nonrecombining part of the Y chromosome (NRY) (*see* Article 4, **Studies of human genetic history using the Y chromosome**, Volume 1) have uniparental inheritance with no recombination, and so each is intrinsically a single haplotypable locus in inheritance and population genetics. This brief introduction will not deal with those loci, but focus on variation in the autosomal genome.

2.1. *Diallelic polymorphisms (SNPs and Indels)*

Single-nucleotide polymorphisms (SNPs), the occurrence of different nucleotides at a specific place in the genome, are the most common type of human DNA sequence variation, occurring on average 1 per 500 to1000 bp on a randomly selected chromosome. As noted later, this number varies somewhat depending on the population studied. Insertion/deletion polymorphisms (Indels), the occurrence of some more or fewer nucleotides at one position of the genome, are less common but do occur frequently. The inserted/deleted sequence can be a few nucleotides long (Weber *et al*., 2002) up to several hundred nucleotides long, as is the case with the transposable Alu elements (Carter *et al*., 2004). Some very large polymorphic duplications, hundreds of kilobasepairs long, have also been identified (Sebat *et al*., 2004; Iafrate *et al*., 2004). Because these polymorphisms are diallelic, they are less informative than microsatellites; however, SNPs are more common in the genome and more amenable to automation and DNA chip technology. Most of the RFLPs (restriction fragment length polymorphisms) defined between the late 1970s and the late 1990s were SNPs in a restriction site but some were Indels occurring between restriction sites. Both SNPs and Indels are generally mutationally more stable than microsatellites; they can be considered as one-off events in evolution. That is, they are caused by a single mutational event that occurred once in the history of a species. We do not expect to see recurrent mutations at the same site

except as extremely rare events. For the polymorphic Alu elements, we know that the ancestral state is absence and the insertion of the Alu is the derived state. For the other markers, we cannot tell from the human polymorphism which allele is ancestral and which is derived. However, we can determine the ancestral state in almost all of those cases simply by determining the genotype of our nearest relatives, the other great apes, following the logic in Iyengar *et al*. (1998) (see also Hacia *et al*., 1999). In most cases, humans share a single allele with the other apes and, by inference, this is the ancestral or original allele in humans and the other allele is the derived allele.

2.2. Multiallelic polymorphisms (STRPs and VNTRs)

Microsatellites consist of approximately 10–50 tandemly repeated copies of particular DNA sequence motifs ranging from 1 to 10 (most commonly 2–4) nucleotide base pairs. These repeat sequences, discovered in 1989 (Weber and May, 1989; Litt and Luty, 1989), occur frequently and randomly across the human genome. When the repeat number is polymorphic, microsatellites are also called *short tandem repeat polymorphisms* (*STRPs*). (Other acronyms have also been used, e.g., SSLP.) STRP loci usually have multiple alleles and can have high levels of variation, that is, high heterozygosity. They rapidly replaced RFLP markers in gene mapping (especially disease genes) studies owing to these features and to their ease of typing, including the small amount of template DNA required. Remember that heterozygosity can never exceed 50% for a diallelic marker, whereas STRPs can easily have heterozygosities >75%. Because of this high heterozygosity, STRPs are the markers on which the most detailed human linkage maps are based (Dib *et al*., 1996; Kong *et al*., 2002; Jorgenson *et al*., 2005). STRPs have also become the standard for forensics and paternity testing (Budowle *et al*., 2001). The larger minisatellite arrays (also referred to as VNTRs, variable number of tandem repeats) are also highly polymorphic and powerful markers in forensic and paternity studies (DNA fingerprinting) (Jeffreys *et al*., 1985; Armour *et al*., 1996). However, they are less common than STRPs and are not evenly distributed throughout the genome. Their larger sequence motifs make them less amenable to PCR technology and use in genomic screening analyses. STRPs and VNTRs, on the other hand, tend to be much more dynamic than the diallelic SNPs and Indels. The mutation rates are higher, there are more alleles, and it is usually impossible to determine which is the ancestral allele.

2.3. Haplotypes

Polymorphisms of any type can be combined into groups of two or more tightly clustered on the chromosome – *haplotypes*. If two or more polymorphisms exist in close proximity in the DNA (e.g., up to a few thousands of bases apart), recombination will be extremely infrequent and they can be studied jointly as a haplotype. A haplotype is the *haploid genotype* of the alleles at polymorphisms along a rather short stretch of the chromosome. When the recombination rate

between any two markers is low, the history of the mutational events and random genetic drift result in a nonrandom occurrence of the alleles on chromosomes in the populations. This nonrandomness is called *linkage disequilibrium* (*LD*). Because recombination is very infrequent in molecularly very short regions, each of the possible haplotypes can be considered for analytic purposes as a distinct allele exactly analogous to alleles at any other polymorphism (SNP, Indel, etc.). We can estimate haplotype frequencies, haplotype heterozygosity, and so on. In samples of unrelated individuals, haplotypes require statistical inference because the individual sites are generally typed separately and 2^{n-1} genotypes are possible for n heterozygous sites. Several algorithms are used to estimate the actual phase, that is, which alleles at the heterozygous sites are on the same chromosome (Excoffier and Slatkin, 1995; Stephens *et al*., 2001b; among others). Within families, Mendelian rules will usually (but not always) allow the specific haplotypes to be inferred. Molecular methods can also determine phase, but they are tedious and usually provide little extra information over statistical inference (Tishkoff *et al*., 2000).

3. Global patterns of variation

Over the past century, researchers have identified normal genetic variation in antigens and proteins and studied that variation in diverse human populations to determine the amounts and distributions of that variation. With the advent of DNA-based markers in the last quarter century, these studies have accelerated. That information is being used to develop an understanding of the demographic histories of the different populations and the species as a whole, among other studies. These data support an "Out of Africa" hypothesis for human dispersal around the world, and are beginning to refine our understanding of population structures and genetic relationships. The historical-demographic explanation of the pattern is supported by the pattern of heterozygosity for STRPs and haplotypes, and by the pattern of LD in addition to the similarities in allele frequencies among populations.

3.1. Variation in the amount of intrapopulation variation: global distribution of heterozygosity

Most SNPs and diallelic markers that have been studied in multiple populations were initially ascertained in a non-African, usually European, population. This ascertainment bias results in the heterozygosity usually being quite high in European populations. Then, in most other populations it can only be lower since the maximum is 0.5. In the future, this bias will be less because more recent large-scale searches for SNPs are using samples with a significant non-European ancestry, especially ancestry from Africa. The evidence that is accumulating argues that there are more SNPs in African populations. STRPs that have been studied in multiple populations were generally ascertained in European populations or selected for additional study because of high heterozygosity in European populations. However, since additional alleles could be found in other populations, the heterozygosity is not limited to the values seen in European populations. Indeed, in all of the studies

done, the average heterozygosity of STRPs is higher in African than in European populations. Heterozygosity is generally lower in East Asian populations and even lower in Native American populations (Calafell *et al*., 1998; Jorde *et al*., 2000; Bowcock *et al*., 1994).

Haplotypes also tend to have high heterozygosity in African populations even though the individual SNPs are subject to the European bias. LD usually limits the allelic combinations seen in Europe. The consistent pattern of less LD in Africa results in more combinations occurring and the individual haplotypes common in Europe are usually less common in Africa. As with any generalization, there are exceptions, but the rule is well supported (e.g., Gabriel *et al*., 2002; Kidd *et al*., 2004). Other than that generalization, different haplotypes show a variety of different global patterns.

3.2. Variation in the amount of interpopulation variation

Baseline information for comparing findings at different loci is accumulating to aid in the identification of loci subject, now or in the past, to selection (directional or balancing). One standard measure of gene frequency variation among populations is the statistic F_{st} (Wright, 1969). This statistic can be related theoretically to random genetic drift as a function of time and effective population size, but can more simply be considered as a measure of the relative amounts of variation among populations shown by different genetic polymorphisms. The distribution of the F_{st} statistic for 369 individual SNPs studied in 38 populations representing all major continental regions shows a skewed distribution with Mean = 0.138 and a range of 0.042 to 0.380 (Kidd *et al*., 2004). Each population is represented by its own sample of individuals, averaging about 50 individuals, that is the basis for allele frequency calculations for all of the SNPs. None of the SNPs has any known or likely functional relevance. Some SNPs show little variation in allele frequencies around the world – a low F_{st} – while others show a great deal of variation in allele frequencies – a high F_{st}. Such variation in F_{st} is expected for loci independently subject to random genetic drift through the identical historical demography. However, it is not immediately obvious why this distribution is skewed. Other studies of multiple loci in smaller numbers of ethnic groups also show a similar skewed distribution (Stephens *et al*., 2001a; Bamshad and Wooding, 2003). This distribution may have been affected by the complex, biased ascertainment of the SNPs, as noted above. That is an area to be investigated. The global F_{st} for STRPs is noticeably smaller (Rosenberg *et al*., 2002), both because the higher mutation rates will tend to damper the effects of random genetic drift and because the higher heterozygosities of STRPs impose algebraic limits on F_{st}.

3.3. Global pattern of linkage disequilibrium

LD is the term usually used to describe nonrandom combinations of alleles at multiple sites on chromosomes in the population. LD is not an all-or-nothing phenomenon but ranges from nearly random combinations of alleles to complete

correlation such that a given allele at one site always occurs with a specific allele at another site and vice versa. A variety of statistics can be used to quantify the level of nonrandomness in any segment of DNA (Devin and Risch, 1995; Zhao *et al*., 1999). As one can imagine from differences in haplotype frequencies, these statistical measures, which are based on those haplotype frequencies, show differences in the magnitude of LD among populations from different parts of the world. The general pattern is random or nearly random association of alleles in populations within Africa, higher values and significantly nonrandom allelic associations in Eurasian populations, and even more nonrandomness in Native American populations (Tishkoff and Kidd, 2004). A more precise summary is difficult to formulate because recombination rates differ by orders of magnitude across segments of DNA such that loci differ considerably in strength and pattern of LD and recent demography of individual populations can have a significant effect.

3.4. Population similarities based on DNA sequence variation

On a global basis, there are many methodologies available for studying genetic similarity among populations. All of them show that populations that are geographically closer have allele and haplotype frequencies that are more similar. One study of allele frequencies of 377 STRPs in 52 populations (Rosenberg *et al*., 2002) showed geographic clustering of the populations but the Central Asian populations showed a gradient between Europe and East Asia. Other analyses of these data have emphasized more strongly the clinal aspects of the allele frequency data (Bamshad *et al*., 2004; Serre and Paabo, 2004). Another study used primarily SNPs and haplotypes thereof to define >600 alleles in 37 populations (Tishkoff and Kidd, 2004). These data were used to generate a tree diagram of the populations (Figure 1) that shows both geographic clustering and a clinal pattern extending from Africa through Europe and Central Asia and then extending in two separate branches, one to East Asia and the Pacific and the other to the Americas. The large bootstrap values basically define four strong clusters: the African populations (including African Americans), the European populations (including European Americans) and middle eastern populations along with one population from northwestern Siberia (Komi), the East Asian populations, and the Native American populations. Several populations are definitely outside of those four clusters: the Ethiopian Jews, the Khanty from northwestern Siberia, the Yakut from northeastern Siberia, the Micronesians, and the Nasioi Melanesians. Were there more geographically intermediate populations, the clustering would not be so evident, as seen in such analyses as Bamshad *et al*. (2004).

The out of Africa model first supported molecularly by studies of mitochondrial DNA (Cann *et al*., 1987) is supported by most of these data except the heterozygosity data on SNPs. The model can be described as follows. Genetic variation had already accumulated in anatomically modern humans in Africa between 150 000 years BP and 100 000 years BP. That variation was not evenly distributed across the continent, as expected in an isolation-by-distance model, but considerable randomness among closely linked sites had accumulated. That

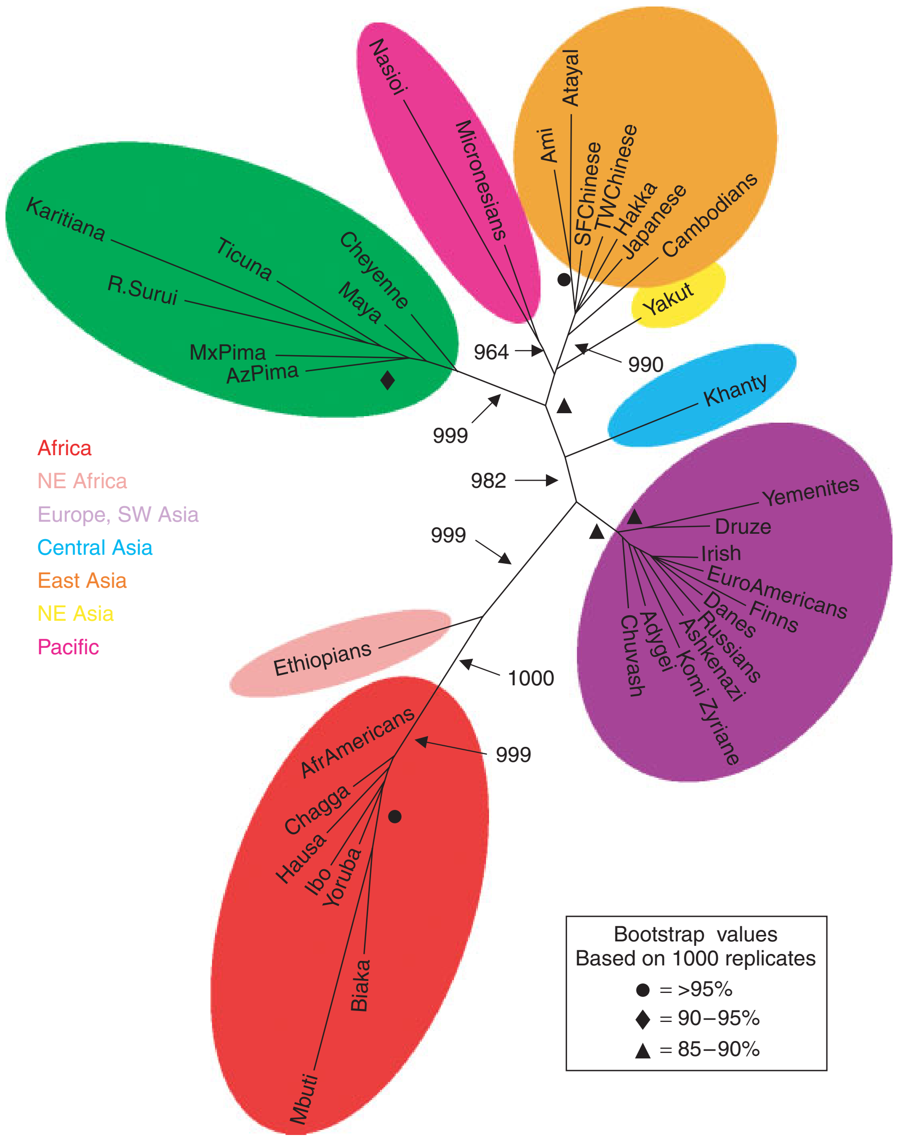


Figure 1 A least-squares tree structure representing the genetic distances among 37 populations. Eighty independent loci (41 as multiallelic haplotypes, 36 biallelic, and 3 STRPs) with about 600 statistically independent alleles were used to calculate the genetic distances. While the assumptions underlying the calculations and representation do not allow this to be interpreted as a precise representation of evolutionary history, the main structure of the tree corresponds to the recent African origin model with increasing genetic distance as early populations migrated away from Africa. SF: San Francisco; TW: Taiwan; MX: Mexico; AZ: Arizona; R: Rondonian. The arrows with actual bootstrap values (out of 1000) indicate segments along the backbone of the tree with high consistent support among the genetic loci. Other large bootstrap values are indicated with symbols: circles, >95%; diamonds, 90–95%; triangles, 85–95%. (Reproduced from Tishkoff and Kidd (2004). © Nature Publishing Group)

variation and low levels of LD still exist in most modern African populations. About 100 000 years BP, some people from Northeast Africa migrated into Southwest Asia. Since the people who migrated originated from the populations of Northeast Africa, they sampled from that already partially diverged gene pool, and the sampling error (founder effect) of the migration accentuated the loss of variation. Only a fraction of the genetic variation in Africa as a whole was represented in that initial "non-African" population. It was that population in Southwest Asia that then increased in numbers and spread geographically to occupy all of Eurasia and Australo-Melanesia by about 40 000 years BP with progressive loss of variation in the populations through accumulating genetic drift as they spread eastward and eventually reached far East Asia. At some time more recent than 40 000 years BP, some of the populations from Siberia migrated to the Americas and expanded to occupy first North and then South America. Additional variation was lost during that colonization but the effect was less than that associated with the initial expansion out of Africa. At all of those stages where variation was lost, nonrandomness (LD) increased among the remaining variants in small segments of DNA (Figure 2). Thus, much of the LD seen in non-African populations is the result of the founder effect associated with the expansion out of Africa. An abstract, artistic rendition of this model can be found at **http://info.med.yale.edu/genetics/kkidd/point.html**.

During most of the century since the beginning of scientific study of genetic variation in humans, there was no clear explanation for the differences among populations. Indeed, the global pattern was not clear. The huge numbers of DNA sequence variants now being studied have changed that. The global pattern of DNA sequence variation we see illustrated in the tree analyses (Figure 1) and in principal components analyses (Kidd *et al*., 2004) is a reflection of the early history of modern humans expanding out of Africa and occupying the rest of the world. Our early history as a species is responsible for the broad-stroke distribution of DNA sequence variation.

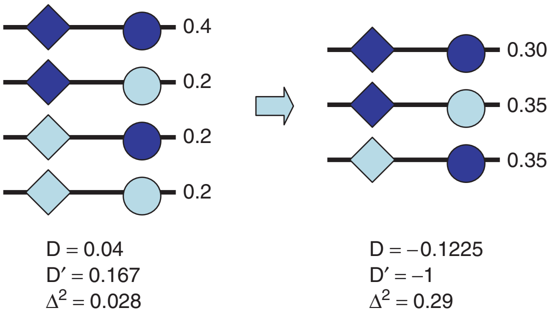


Figure 2 Consider the diamond and circle loci with light and dark alleles at each and haplotype frequencies as indicated. If random drift causes one haplotype to be lost, as indicated by the arrow, it is possible for the individual allele frequencies to change very little (dark alleles increased in frequency from 0.6 to 0.65) while the nonrandomness, indicated by several standard disequilibrium statistics (Devin and Risch, 1995), has increased greatly

Related articles

Article 2, **Modeling human genetic history**, Volume 1; Article 7, **Genetic signatures of natural selection**, Volume 1; Article 71, **SNPs and human history**, Volume 4

Further reading

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, *et al*. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, **235**, 1616–1622.

References

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, Bertranpetit J, Paabo S and Jeffreys AJ (1996) Minisatellite diversity supports a recent African origin of modern humans. *Nature Genetics*, **13**, 154–160.

Bamshad M and Wooding SP (2003) Signatures of natural selection in the human genome. *Nature Reviews Genetics*, **4**, 99–111.

Bamshad M, Wooding SP, Salisbury BA and Stephens JC (2004) Deconstructing the relationship between genetics and race. *Nature Reviews Genetics*, **5**, 598–608.

Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High Resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Budowle B, Masibay A, Anderson SJ, Barna C, Biega L, Brenneke S, Brown BL, Cramer J, DeGroot GA, Douglas D, *et al*. (2001) STR primer concordance study. *Forensic Science International*, **124**, 47–54.

Calafell F, Shuster A, Speed WC, Kidd JR and Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. *European Journal of Human Genetics*, **6**, 38–49.

Cann R, Stoneking M and Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Carter AB, Salem AH, Hedges DJ, Deegan CN, Kimball B, Walker JA, Watkins WS, Jorde LB and Batzer MA (2004) Genome-wide analysis of the human Alu Yb-lineage. *Human Genomics*, **1**, 167–178.

Cavalli-Sforza LL, Menozzi P and Piazza A (1994) *The History and Geography of Human Genes*, Princeton University Press: Princeton.

Collins FS, Green ED, Guttmacher AE and Guyer MS (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.

Devin B and Risch N (1995) A comparison of linkage disequilibrium measures for fine scale mapping. *Genomics*, **29**, 311–322.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Missasseau P, Marc S, Hazan J, Seboun E, *et al*. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**(Suppl A1–A138), 152–154.

Excoffier L and Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–927.

Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H and Brookes AJ (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research*, **30**, 387–391.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel Bhiggins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, *et al*. (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, **22**, 164–167.

Harris H (1966) Enzyme polymorphisms in man. *Proceedings of the Royal Society of London B. Biological Sciences*, **22**, 298–310.

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T and Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Research*, **30**, 158–162.

Iafrate AJ, Feuk L, Rivera Mn, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **9**, 949–951.

Iyengar S, Seaman M, Deinard AS, Rosenbaum HC, Sirugo G, Castiglione CM, Kidd JR and Kidd KK (1998) Analyses of cross-species polymerase chain reaction products to infer the ancestral state of human polymorphisms. *DNA Sequence*, **8**, 317–327.

Jeffreys AJ, Wilson V and Thein SL (1985) Hypervariable "minisatellite" regions in human DNA. *Nature*, **314**, 67–83.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT and Batzer MA (2000) The distribution of human genetic diversity: a comparison of Mitochondrial, Autosomal, and Y chromosome data. *American Journal of Human Genetics*, **66**, 979–988.

Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E, *et al*. (2005) Ethnicity and human genetic linkage maps. *American Journal of Human Genetics*, **76**, 276–290.

Kan YW and Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proceedings of the National Academy of Science U S A*, **75**, 5631–5635.

Kidd KK, Pakstis AJ, Speed WC and Kidd JR (2004) Understanding human DNA sequence variation. *Journal of Heredity*, **95**, 406–420.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, **31**, 241–247.

Litt M and Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, **44**, 397–401.

Osier MV, Cheung K-H, Kidd JR, Pakstis AJ, Miller PL and Kidd KK (2002) ALFRED: an allele frequency database for anthropology. *American Journal of Physical Anthropology*, **119**, 77–83.

Rajeevan H, Osier MV, Cheung H-K, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, *et al*. (2003) ALFRED: the ALlele FREquency Database. Update. *Nucleic Acids Research*, **31**, 270–271.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**(5683), 525–528.

Serre D and Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, **14**, 1679–1685.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotin K (2001) DbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308–311.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al*. (2001a) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.

Stephens M, Smith NJ and Donnelly P (2001b) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Tishkoff SA and Kidd KK (2004) Implications of biogeography of human populations for “race” and medicine. *Nature Genetics*, **36**(11), S21–S27.

Tishkoff SA, Pakstis AJ, Ruano G and Kidd KK (2000) The accuracy of statistical methods for estimating haplotype frequencies: an example from the CD4 locus. *American Journal of Human Genetics*, **67**, 518–522.

Weber JL, David D, Heil J, Fan Y, Xhao C and Marth G (2002) Human diallelic insertion/deletion polymorphisms. *American Journal of Human Genetics*, **71**, 854–862.

Weber JL and May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, **44**, 388–396.

Wright S (1969) *Evolution and the Genetics of Populations*, *Volume 2: The Theory of Gene Frequencies*. University of Chicago Press: Chicago, p 511.

Zhao H, Pakstis AJ, Kidd JR and Kidd KK (1999) Assessing linkage disequilibrium in a complex genetic system I. Overall deviation from random association. *Annals of Human Genetics*, **63**, 167–179.

Specialist Review

Reliability and utility of single nucleotide polymorphisms for genetic association studies

C. Leigh Pearce
USC/Keck School of Medicine, Norris Comprehensive Cancer Center, Los Angeles, CA, USA

Joel N. Hirschhorn
Children's Hospital, Boston, MA, USA
Harvard Medical School, Boston, MA, USA
Broad Institute of Harvard and MIT, Cambridge, MA, USA

1. Introduction

Most common diseases and disease-related quantitative traits (such as body weight or blood pressure) are complex genetic traits (*see* Article 58, **Concept of complex trait genetics**, Volume 2), under the influence of multiple genetic and nongenetic factors. Efforts to understand the contribution of genetic variation to complex traits have increased exponentially since the late 1990s, as the completion of the Human Genome Project (*see* Article 23, **The technology tour de force of the Human Genome Project**, Volume 3 and Article 24, **The Human Genome Project**, Volume 3) and related efforts have greatly increased the feasibility of these pursuits. Single-nucleotide polymorphisms (SNPs) are by far the most common type of genetic variation in the human genome (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4), with estimates of 11 million SNPs with a frequency of 1% or greater (Reich *et al*., 2003; Kruglyak and Nickerson, 2001). The first comprehensive assessment of coding region variation revealed that a typical gene has two missense and two silent SNPs, and the overall density of SNPs was consistent with current expectations of a common SNP every 300 bp (Halushka *et al*., 1999; Cargill *et al*., 1999). This was followed shortly by the formation of The SNP Consortium (TSC), which by 2001 had submitted more than 1.4 million SNPs located throughout the genome to the public dbSNP database (**http://www.ncbi.nlm.nih.gov/projects/SNP/**) (Sachidanandam *et al*., 2001). The efforts of the Human Genome Project and The International Haplotype Map Project

(HapMap) (**www.hapmap.org**) have added to the availability of SNP data, and now there are more than 10 million in the public database with genotype frequency available for a substantial proportion (The International HapMap Consortium, 2003). Furthermore, studies have shown that most SNPs are strongly correlated with other neighboring SNPs, meaning that the overwhelming majority of undiscovered common SNPs are reflected by those currently in dbSNP.

Association studies, which are discussed in more detail below, use SNPs as genetic markers, and test whether an allele is more common in cases than in controls (or in individuals with higher trait values vs. lower trait values). However, results of these studies have been inconsistent (Hirschhorn *et al*., 2002). To help understand how SNPs can be effectively used for association studies, we consider two aspects of reliability of SNP-based association studies. First, we consider whether the SNPs in the dbSNP database are in fact valid and useful for association studies. Second, we detail the reasons why association studies with these SNPs have been so inconsistent, and the possible routes to minimize the inconsistency.

2. Comprehensiveness, validity, and frequencies of SNPs in the public dbSNP database

Genetic association studies are a potentially powerful tool to identify variants that contribute to common disease susceptibility or other disease-related traits (Risch and Merikangas, 1996; Hirschhorn and Daly, 2005; *see also* Article 59, **The common disease common variant concept**, Volume 2). There are two major approaches for using SNPs in association studies (discussed in more detail below): a direct approach to association studies, in which putative functional SNPs are tested for association with the phenotype, and an indirect approach, which exploits correlations (linkage disequilibrium, LD) between nearby SNPs and uses a subset of SNPs that capture the vast majority of the common variation in a region (*see* Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3 and Article 74, **Finding and using haplotype blocks in candidate gene association studies**, Volume 4). Regardless of the approach used, the validity and comprehensiveness of SNPs from the public SNP database (dbSNP) is critical in order to ensure that resources, both labor and money, are used wisely for genotyping and disease-associated regions are not missed because of the relevant SNPs being absent from the database.

2.1. Reliability of SNPs in dbSNP

The public database (dbSNP) currently contains over 10 million SNPs. Although essentially all of the SNPs were discovered by resequencing the same stretch of DNA in different chromosomes, the details of their ascertainment varies. The number of chromosomes used in the SNP discovery process has ranged from two to hundreds, and individuals from a variety of different ethnic groups have been used. These differences in ascertainment affect SNP frequencies because of human

history. Until recently, the human population was actually quite small (about 10 000 individuals), so population genetics (*see* Article 1, **Population genomics: patterns of genetic variation within populations**, Volume 1) predicts that most SNPs identified by comparing two chromosomes will be common (allele frequencies well over 1%). Indeed, this prediction has been verified (Gabriel *et al*., 2002). SNPs seen twice out of four chromosomes ("double hit") are even more common on average. By contrast, studies that resequence a large number of chromosomes to identify SNPs (deep resequencing) will identify not only the common SNPs but also rare SNPs. Indeed, population genetics also predicts that SNPs seen once in a study that has examined hundreds of chromosomes will have a true frequency of well under 1/1000.

The ethnicity of the samples used for SNP discovery also affects their allele frequencies in different populations (*see* Article 2, **Modeling human genetic history**, Volume 1). In particular, because non-African populations are derived from an ancestral African population (Tishkoff and Williams, 2002), populations with recent African ancestry have more diversity than other populations, so SNPs identified in these populations will sometimes be monomorphic in other populations. Finally, because none of the algorithms used to identify SNPs from sequence data are perfect, some apparent SNPs will in fact turn out to be sequencing artifacts. Although the overall reliability of these 10 million SNPs is high, some are false-positives, some are vanishingly rare, and others are private to specific populations.

A number of groups have reported the rate of "validation" of SNPs – that is, when SNPs from dbSNP are genotyped in a particular population, how often are they observed to be polymorphic? A true estimate of the false-positive rate requires genotyping the individual(s) in whom the SNP was discovered. If the individuals used for SNP discovery are not assayed, rare SNPs (or SNPs that are private to particular populations) may appear to be false-positives because they are monomorphic in the genotyped population. The International SNP Map Working Group was the first to provide a dense set of markers across the human genome, with an average density of one SNP ever 1.9 kb (Sachidanandam *et al*., 2001). This group validated a subset of 1500 SNPs that they discovered by genotyping them in the same population in which they were identified. They found that 4% of their SNPs were false-positives (Sachidanandam *et al*., 2001). Similarly, Reich *et al*. (2003) found through a deep resequencing effort that 98% of the TSC SNPs were validated in the population in which they were discovered.

When 3738 of these SNPs were genotyped in four other populations, 89% were polymorphic in at least one population (Gabriel *et al*., 2002), confirming that these SNPs, which were discovered largely by comparing two chromosomes, are generally common enough to be detected at appreciable frequencies. By contrast, other studies (Carlson *et al*., 2003; Cutler *et al*., 2001) have reported that SNPs from dbSNP are polymorphic at much lower rates. The lower rates result from a combination of including SNPs ascertained by different methods (and hence with different frequencies), plus most likely some rare SNPs were falsely thought to be monomorphic (Carlson *et al*., 2003).

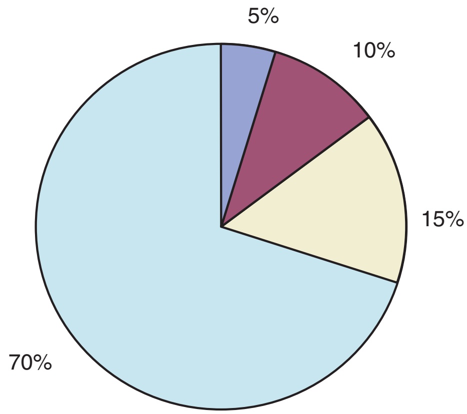


Figure 1 Approximate fractions of SNPs in dbSNP (2004) that are false-positive (blue), rare in all populations (purple), or common (>1%) in some (beige) or all populations (green). Estimates are based on Gabriel *et al*., (2002), Reich *et al*. (2003), and unpublished observations by our groups and others (D. Altshuler, S. Gabriel, and M. J. Daly, personal communication)

Most of the SNPs in dbSNP were discovered by an automated algorithm identical or similar to the method used by the International SNP Map Working Group (Sachidanandam *et al*., 2001), suggesting that the false-positive rate in dbSNP is likely to be near 4%. This relatively low false-positive rate is encouraging, however, it is important to remember that the rate at which true positive SNPs will be polymorphic in any particular population depends on the method by which the SNPs were ascertained, and the ethnicity and size of the population(s) tested (Cavalli-Sforza *et al*., 1994). False-positives aside, how often will SNPs be common across populations, given the known differences in allele frequencies across populations (Cavalli-Sforza *et al*., 1994)? Gabriel *et al*., (2002) found that only 59% of the SNPs they tested were variable in all four of their populations. Only 52% of the common SNPs assayed by Carlson *et al*. (2003) were common in both of the populations they studied. As expected, most of the SNPs that were found only in one major population group were specific to populations with recent African ancestry. Figure 1, which is derived from these and similar data, displays the approximate current status of the SNPs in dbSNP.

Efforts can be made to improve the likelihood that a SNP genotyped as part of an association study is polymorphic. Importantly, many of the SNPs in dbSNP are being computationally or experimentally validated, and allele frequencies will be available for a substantial fraction (such frequency data is readily available from dbSNP and the HapMap). Also, utilizing SNPs that have been identified by more than one investigator also greatly increases the likelihood of those SNPs being polymorphic in the population under study – since rare SNPs are unlikely to have been seen twice and the identical sequencing artifact leading to a false-positive SNP is unlikely to be replicated in multiple labs. For example, in the study by Carlson *et al*. (2003), approximately 85% of SNPs that were reported by more than one investigator were polymorphic. Identifiers representing the number of times a SNP has been discovered and its validation status are available in dbSNP.

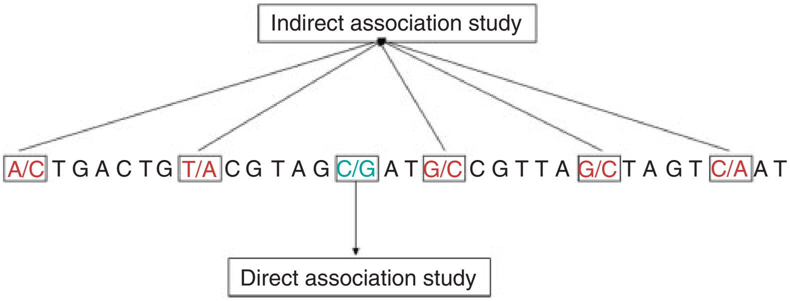


Figure 2 The figure shows six SNPs across a genomic region, five in red without functional consequence to the protein and one in blue that is functional. In a direct association study only the SNP in blue would be tested, whereas in an indirect association study, some of the five SNPs in red would be tested in an effort to mark the functional SNP

2.2. *SNPs for indirect and direct association studies*

As mentioned above, SNPs can be used for association studies in two complementary manners, as shown in Figure 2. They can be used indirectly as genetic markers, which will be effective if there is sufficient correlation (LD) between the typed SNP and a nearby causal variant, or if the typed SNP is itself causal. Alternatively, efforts can be made to try to directly identify and genotype those SNPs that are likely to be functional and therefore causal, without having to rely on LD.

The underlying assumption of an indirect association study is that the SNPs that are genotyped serve as effective proxies for other unmeasured, potentially causal SNPs. For this approach, the key question is then, "Can SNPs in the public database capture enough of the underlying diversity across the genome"? The simple answer, for common genetic variation, is that the vast majority of variation can probably be captured using this approach (Hirschhorn and Daly, 2005). Recent resequencing efforts surrounding the encyclopedia of DNA elements (ENCODE) and HapMap projects have suggested that, in populations with recent European ancestry, 80–90% of the estimated 11 million common SNPs in the human genome can be well captured by the SNPs that are already in public databases, suggesting that if the LD relationships between SNPs are known in a population, indirect association studies can be carried out effectively (*see* Article 12, **Haplotype mapping**, Volume 3 and Article 73, **Creating LD maps of the genome**, Volume 4). Indeed, a significant majority of SNPs with frequencies above 10% are already in the database (D. Richter, S. Gabriel, D. Altshuler, personal communication); rarer SNPs (particularly those below 5%) are less well represented, although long-range haplotypes of tag SNPs may help capture these rare SNPs (Lin *et al*., 2004). How large a set of "tag SNPs" is required to efficiently capture the remaining common variation has yet to be determined, but will depend on the population being studied – populations with recent African ancestry will require more SNPs because they have shorter average extents of LD and greater haplotype diversity within regions of LD (Gabriel *et al*., 2002; Reich *et al*., 2002; Crawford *et al*., 2004). The HapMap is designed to provide extensive SNP frequency data on multiple ethnic groups in an effort to

fill in many of these gaps (The International HapMap Consortium, 2003). The current HapMap goals are for 3–5 million SNPs to be typed in four populations representing three major continental groups: Han Chinese and Japanese, Yoruba from Nigeria, and European-Americans. Because of this fairly broad representation in the study, tag SNPs chosen on the basis of data from these populations may well do a good job of representing a diverse set of populations worldwide, a hypothesis supported by our group's unpublished observations. However, the general utility of tag SNPs selected on the basis of HapMap data needs to be assessed for each of the major population group that is used in association studies.

In the direct association approach, certain classes of SNPs are hypothesized to have a higher prior probability of influencing disease risk. These types of SNPs include nonsynonymous SNPs (nsSNPs) and SNPs that fall in evolutionary conserved regions (ECRs).

Nonsynonymous SNPs are both less abundant and have lower average allele frequencies than other SNPs, because many potential changes to the coding sequence of genes are evolutionarily deleterious (Cargill *et al*., 1999). The abundance of nsSNPs is 30% less than would be expected under a neutral model, and only 41% of nsSNPs have a frequency of 5% or greater compared with 61% of noncoding variants (Cargill *et al*., 1999). Because they are on average rarer than typical SNPs, nsSNPs are relatively underrepresented in dbSNP, and identifying all common (>1%) nsSNPs would require a substantial focused effort.

Because nsSNPs alter the encoded amino acid, they are believed to be more likely to alter the function of the gene product in important ways. It has been suggested that association studies should focus on these types of variants (Botstein and Risch, 2003). This argument is based on the allelic spectrum of single gene disorders, although there is reason to suggest that regulatory variants may play an equally important role in common polygenic diseases and quantitative traits (Reich and Lander, 2001; Hirschhorn and Daly, 2005). In addition, not all nsSNPs are created equal, and a number of mechanisms have been put in place to facilitate identifying nsSNPs that have a higher probability of influencing the structure or function of a gene (Ng and Henikoff, 2002; Ng and Henikoff, 2003; Conde *et al*., 2004). One such method is SIFT (Sorting Intolerant From Tolerant), which takes user-supplied data on the nsSNPs and then links to a series of publicly available databases to predict whether the nsSNP is tolerant (less likely to be damaging to the protein) or intolerant (more likely to be damaging to the protein) (Ng and Henikoff, 2002; Ng and Henikoff, 2003). Zhu *et al*. (2004) used SIFT to retrospectively assign 46 candidate cancer nsSNPs in 39 genes to a tolerance index. The investigators then compared the odds ratios associated with the nsSNP and the cancer outcome to the tolerance index. Zhu *et al*. (2004) found a strong relationship between the tolerance index and the magnitude of the cancer-nsSNP odds ratio ($p = 0.002$). However, the false-positive and false-negative rates of this and related algorithms remain unknown.

SNPs that fall in noncoding genomic regions conserved across species may also be more likely to be functionally relevant (Nobrega *et al*., 2003; Frazer *et al*., 2004). A number of tools have been put in place to identify these ECRs, such as the ECR Browser (**http://ecrbrowser.dcode.org/**) and VISTA (**http://www-gsd.lbl.gov/vista/index.shtml**), both of which can align sequences across multiple

species. Selecting SNPs that fall into these regions may increase the probability of finding an association with disease risk, but it is much more difficult within ECRs than it is within coding regions to distinguish those variants that are functionally important.

In summary, abundant SNPs are present in the dbSNP database. Most of these are "real" SNPs, although their frequency will vary by population studied. There are sufficient SNPs to embark on indirect association approaches, and the HapMap is generating data to allow selection of sets of SNPs for this approach. Direct tests of putative functional SNPs are also possible, but more work would be required to build a catalog of potentially functional SNPs for use in association studies. The reliability of association studies is discussed in the next section.

3. Application of SNPs for association-based studies

Now that the public database contains enough SNPs to carry out comprehensive and meaningful genetic association studies, what are the challenges in the interpretation of these types of studies? Replication of association study results has proved challenging; a recent review of association studies showed that only 6 of 162 associations were very consistently replicated (Hirschhorn *et al*., 2002). There are three main classes of reasons why association studies fail to replicate: true variability in the populations studied (heterogeneity), false-positives, and false-negatives (see Figure 3). Each of these is discussed below.

3.1. Heterogeneity

A frequently invoked but rarely proven explanation for inconsistency in genetic studies is heterogeneity between populations. In other words, it is proposed that the populations under study are different in important ways that make the association much stronger in some of the populations (Figure 3). For example, there could be environmental factors (such as diet) that both vary between populations and modify the association between the genetic variant and disease susceptibility. Similarly, there could in theory be genetic modifiers that vary substantially between

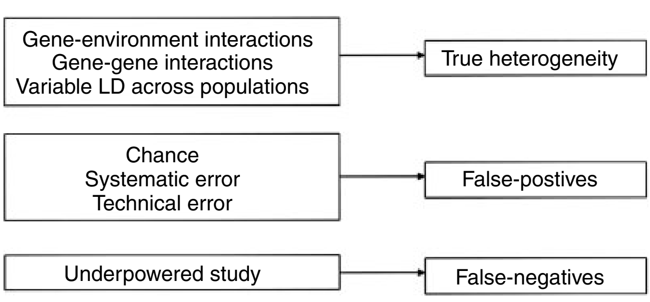


Figure 3 The figure shows the three situations leading to inconsistency in genetic association studies on the right-hand side, and possible causes on the left. See text for details

populations and lead to variability in association results. Heterogeneity could arise if the variant that is genotyped is in LD with the causal allele, and the strength of the LD varies across populations. Variable LD is in theory easy to address by comprehensive surveys of variation in genes showing evidence of association. However, demonstrating gene–environment and gene–gene interactions requires knowledge of the modifiers that differ between populations, and extremely large sample sizes will be usually required to prove convincingly that gene–gene and gene–environment interactions are important for a given association. Finally, phenotypic heterogeneity could also account for differences in findings across populations. If disease classification changes over time, or if diagnostic criteria are inconsistent or subjective, then differences in association results could be observed.

3.2. False-positive associations

Recent surveys have estimated that at least 70 to 80% of reported associations are false-positives (Ioannidis *et al*., 2001; Lohmueller *et al*., 2003). False-positives can occur as a result of chance fluctuations that are enhanced by inadequate statistical control and interpretation, systematic biases in study design, or systematic bias due to technical issues, each of which can result in the expenditure of valuable resources in terms of both money and time.

3.2.1. Chance

Chance fluctuations are by far the major source of false-positive associations, mainly because of the criteria for declaring a "positive" association. The standard measure for determining whether any given finding is the result of chance is the p-value. By definition, a p-value is the probability of observing the study result or one more extreme than that observed when the null hypothesis is in fact true. Historically, associations between a SNP and a phenotype are considered statistically significant if the p-value is 0.05 or less, regardless of the number of SNPs that were tested. A p-value of 0.05 means, however, that one out of 20 tests will be positive by chance alone. At the extreme, if one tested the entire range of 10 million common SNPs throughout the human genome, a p-value of 0.05 would mean that 500 000 associations would be observed simply due to chance. Multiple phenotypic tests would further increase the number of p-values below 0.05. Thus, the threshold of 0.05, although standard throughout the scientific literature, is inadequate given the number of hypotheses that could be tested in an association study, compared with the number of variants in the genome that are actually causal, which is likely a very small number. The false-positive results will dwarf any truly causal associations, therefore making all of the results uninterpretable. What then is the appropriate statistical threshold for determining if an association is simply due to chance because so many hypotheses were tested?

The issue of multiple hypothesis testing and related adjustment of the p-value for association has been considered by many investigators (Thomas and Clayton, 2004; Wacholder *et al*., 2004; Cardon and Bell, 2001; Dudbridge and Koeleman, 2004; Colhoun *et al*., 2003). One approach to reduce type I errors is a Bonferroni correction that is simply computed by dividing the standard statistical

significance level of 0.05 by the number of tests performed. The Bonferroni correction is somewhat overly conservative, because it assumes that all of the tests are independent (Thomas and Clayton, 2004), but quite often, many of the tests are correlated (as when multiple SNPs that are in LD are tested). To estimate a Bonferroni correction for surveying the entire genome, Risch and Merikangas (1996) advocated a p-value threshold of 5×10^{-8} (which is the equivalent of a p-value of 0.05 after adjusting for the assumed one million tests). Calculating the exact number of tests is problematic, but one can still estimate the sample sizes required to meet this conservative threshold. For alleles with strong effects on disease risk, even this stringent adjustment does not demand unreasonably large sample sizes. For example, assuming a log additive mode of inheritance, a minor allele frequency of 20%, 80% power, and an odds ratio of 3.0, only 165 cases and 165 controls would be required. But, much larger sample sizes are required for alleles with more modest effects. If the odds ratio were only 1.3, more than 3000 cases and 3000 controls would be required. Similarly, rarer alleles require larger sample sizes as well. Because many alleles that affect complex traits are likely to have odds ratios of 1.5 or less (Lohmueller *et al*., 2003), very large sample sizes will be needed to achieve statistical significance that withstands a genome-wide correction. However, it is important to note that several associations have already achieved these stringent levels of significance (Hirschhorn and Daly, 2005).

Other approaches, such as permutation testing, that can take into account non-independence of different tests, may thus be valuable (Dudbridge and Koeleman, 2004). Regardless of the method used, however, it is helpful for any particular study to include an accurate assessment of the likelihood of achieving the observed p-value by chance, given the tests that were performed in that study. A Bayesian method can also be used to help interpret how much a particular study strengthens support for an association. This approach is informed by our understanding of the disease process and the existing evidence of a disease-SNP association, that is, what is the prior probability of the SNP being associated with the phenotype? Wacholder *et al*. (2004) described the "false-positive report probability" (FPRP) as the probability that there is no true association despite a statistically significant finding. The FPRP is determined on the basis of the size of the observed p-value, the power of the study, and the prior probability of an SNP-disease relationship. This approach is extremely appealing because each observed association can be evaluated on its own merits, however, the determination of an acceptable FPRP is rather arbitrary and specifying the prior probability for each SNP-disease relationship is also challenging. Regardless of these issues, the FPRP can be determined for a range of priors allowing the investigator to provide this information for others to evaluate. Sample sizes needed to achieve a low FPRP will be comparable to those needed for a Bonferroni correction for having tested a sizable fraction of the genome. Thus, regardless of the method used to evince convincing statistical support for an association, large sample sizes will usually be required.

3.2.2. Systematic bias

The above section considers the false-positive rates for association studies in the absence of any systematic bias toward positive results. However, if such a

bias exists, the false-positive problem will be exacerbated. The primary potential source of systematic bias toward false-positives in genetic association studies using unrelated subjects is population substructure (*see* Article 75, **Avoiding stratification in association studies**, Volume 4). Although some researchers believe that population substructure is not likely to be a major problem (Wacholder *et al*., 2002; Wacholder *et al*., 2000), others recognize its potential role in false-positive association study results (Thomas and Witte, 2002; Gorroochurn *et al*., 2004). Population substructure refers simply to confounding by ethnicity, a situation in which both allele frequencies and disease incidence are related to ethnicity and a spurious association between genotype and disease can therefore be observed (Lander and Schork, 1994; Ewens and Spielman, 1995). Three situations are believed to give rise to this type of bias in association studies: gross population stratification, cryptic relatedness, and population admixture (Pritchard *et al*., 2000a; Pritchard and Donnelly, 2001).

Although the terms stratification, cryptic relatedness, and admixture are often used to describe similar situations, in this context we will refer to gross population stratification as the existence of multiple ethnically heterogeneous groups all of which are grossly categorized together in the context of a disease that varies in its frequency by ethnic group. In the most basic sense, this would include ignoring broad ethnic categories such as White and Asian, and grouping them together without consideration for ethnicity. This situation is addressed easily by utilizing self-reported ethnicity as a means for classification. Any allele that varied in its frequency between these two ethnic groups would appear associated with disease, however, simple adjustment for ethnicity in any statistical analysis would eliminate this type of confounding.

More controversial is the situation of cryptic relatedness (often called cryptic or mild stratification) in which there are potential differences in allele frequencies between subpopulations within a grossly homogeneous population that shares the same ethnic self-identification (Pritchard *et al*., 2000a; Pritchard and Donnelly, 2001). For example, cases may be more closely related to each other than are controls, because they by definition share many of the genetic factors that contribute to their disease status. Because even homogeneous populations may in fact have gradients of allele frequencies (Helgason *et al*., 2005), cryptic relatedness may result in uneven distributions between cases and controls not only of causal alleles but also of other noncausal alleles, thereby leading to false-positive associations. Mild stratification can also occur when cryptic differences in ethnicity track with differences in exposure to nongenetic susceptibility factors, with the end result that one subgroup is over represented in the cases. If there are genetic markers that are more common in the subgroup, false-positive associations could ensue. Empirical studies have shown that mild stratification cannot be ruled out, even in studies of apparently homogeneous populations (Freedman *et al*., 2004), and we have recently demonstrated a false-positive association due to mild stratification in a self-described Caucasian US population (Campbell *et al*., 2005). However, it is not clear how serious a problem stratification will be in practice.

On the other hand, recent admixture, which occurs when two or more ethnically diverse groups mix, has been shown to cause false-positive associations. In this case, the recent admixture results in a group with varying levels of ancestry from

each of the two ancestral groups. For example, the proportion of European ancestry among African-Americans, an admixed population, is estimated to range from 11.6 to 22.5 depending on the geographic region of the United States. (Parra *et al*., 1998). If, for example, the risk of disease increases with increasing European ancestry, individuals with higher European ancestry will be overrepresented in cases, and any of the many alleles that are more common in European populations than in African populations would appear to be associated with disease.

Two methods, genomic control (Devlin and Roeder, 1999; Devlin *et al*., 2001; Reich and Goldstein, 2001) and structured association (SA) (Pritchard *et al*., 2000a; Pritchard and Rosenberg, 1999), have been developed to address the issue of population substructure in genetic association studies by using unlinked markers across the genome. Genomic control is based on the assumption that random markers throughout the genome that are unlinked to disease loci should not be associated with disease in the situation of no population substructure (Devlin and Roeder, 1999; Reich and Goldstein, 2001). If population substructure is present, however, the chi-square test statistics for these unlinked markers will be higher than the expected null distribution. The amount of this increase can then be quantified and converted into an inflation factor (λ) for the chi-square test statistic of interest, namely, that obtained from testing the association between disease and a candidate SNP. SA is based on using unlinked markers to categorize study subjects into subpopulations (Pritchard *et al*., 2000b; Pritchard and Donnelly, 2001; Pritchard and Rosenberg, 1999). The assumption of this approach is that although there is heterogeneity within a population, the subgroups can be identified and categorized into discrete groups based on genotype data. Originally, this method only provided a method to determine whether or not population substructure was present, but with no mechanism to deal with the information (Pritchard and Rosenberg, 1999). This was extended to allow for the estimate of genotype effect after adjusting for population substructure (Pritchard *et al*., 2000a; Hoggart *et al*., 2003). Publicly available software can be used to implement the population structure categorization algorithm (**http://pritch.bsd.uchicago.edu/**). Freedman *et al*., (2004) tested for population substructure among the 11 epidemiologic studies they evaluated using the SA method as well. On the basis of their original sample size and number of unlinked loci tested, they found no statistically significant evidence of population substructure, despite λ values that could be as high as 4.0, which can greatly increase the chances of false-positive association (Marchini *et al*., 2004).

The lack of significant evidence of substructure in these empirical studies could mean that no such substructure exists, or that not enough markers were typed to detect the substructure. The number of unlinked loci that need to be tested to confidently determine λ or to assign individuals to a population subgroup is unclear. Estimates of the number of loci vary widely, but are likely to be well over 100, depending on the precision with which assignment to a population group is made or λ is estimated, as well as the allele frequency differences across populations, the number of subpopulations in the study group, and the ability of markers to distinguish between the subpopulations (Pritchard *et al*., 2000a; Pritchard and Rosenberg, 1999; Hoggart *et al*., 2003; Turakulov and Easteal, 2003; Falush *et al*., 2003; Hao *et al*., 2004; Rosenberg *et al*., 2003). As genotyping costs are reduced, genotyping 100 or more loci in any genetic association study of

unrelated individuals will become routinely feasible and should be conducted to minimize the likelihood of a false-positive finding.

3.2.3. Technical errors

Errors in genotyping are another potential source of false-positives in association studies, especially family-based studies. Genotyping error is considered to be random and therefore should result in bias toward the null, but Mitchell *et al*. (2003) have shown that in the context of the transmission/disequilibrium test (TDT) this is not necessarily the case. These investigators found that genotyping error that goes undetected will result in apparent overtransmission of the common allele leading to incorrect associations (Mitchell *et al*., 2003). This result is in concert with those of Gordon *et al*. 2001, who also found that genotyping error is a reason for false-positives in TDT-based SNP studies. Missing data can also increase the false-positive rate in family-based studies (Hirschhorn and Daly, 2005). By contrast, in case–control studies using SNPs, undetected random error or missing data will result in a loss of power to detect an association (Gordon and Ott, 2001).

A number of other nonrandom laboratory-based errors are also of concern, but are easier to control. For example, if cases and controls are genotyped separately, rather than intermingled on genotyping plates, then any increased genotypic dropout or error that is plate-specific will be nonrandom with respect to disease status – leading to a biased association finding. This problem can be avoided by simple planning of plate layouts and quality control procedures.

3.3. False-negatives

Finally, false-negative studies also are an important contributor to the inconsistency in association reports (Figure 3). Recent meta-analyses of association studies have shown that most validated associations have modest odds ratios of 2 or less (indicating that carriage of the risk allele increases the risk of disease by less than twofold) (Ioannidis *et al*., 2001; Lohmueller *et al*., 2003). Modest odds ratios such as these require large sample sizes (in the thousands) to achieve p-values below 0.05; even larger samples are required to meet more stringent statistical thresholds. Because most association studies have not even approached the required sample sizes, most have not had power to achieve even nominal significance ($p < 0.05$), so false-negative studies have been rampant for most validated associations (Lohmueller *et al*., 2003).

The problem of replicating valid associations is complicated by the fact that the first "significant" report of an association between an SNP and phenotype almost invariably has an inflated odds ratio as a result of the "winners curse". Because of this phenomenon, a description of which appears in Lohmueller *et al*. (2003), actual odds ratios are even more modest than they first appear, so replication studies will require even larger sample sizes than would be estimated based on the first published odds ratio. These large sample sizes can be achieved either in individual studies, or, with appropriate care, through meta-analysis. Thus, just as

rigorous statistical thresholds and large sample sizes offer a way to minimize false-positives, large sample sizes can reduce the contribution of false-negative studies to the inconsistency of association studies.

Although there are considerable challenges for conducting SNP-based association studies, sound study design and interpretation can lead to reliable results. False-positive results due to either chance, population substructure, or systematic technical errors can be minimized by large sample sizes combined with appropriate statistical thresholds, consideration of the possibility of admixture, and rigorous laboratory protocols. False-negatives, particularly in the replication stage of an association, can also be greatly reduced by large sample sizes and careful interpretation. However, the true reliability of association studies will only become apparent after the genome has been surveyed for association in large samples (Hirschhorn and Daly, 2005).

4. Summary

SNP databases currently contain enough reliable, validated, and useful SNPs to adequately assess common genetic variation for a role in disease. Establishing a complete catalog of potentially functional SNPs will still require additional effort, as will assembling the tools to survey rare variation. The rapidly accumulating data in the human Haplotype Map will permit LD-based association studies of common variants. The reliability of these association studies will be improved by increased statistical rigor and increased sample sizes.

References

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl), 228–237.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman MF, Groop LC, Altshuler D, Ardlie KG and Hirschhorn JN (2005) Demonstrating stratification in a European-American population. *Nature Genetics*, in press.

Cardon LR and Bell JI (2001) Association study designs for complex diseases. *Nature Reviews. Genetics*, **2**(2), 91–99.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, *et al* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, **22**(3), 231–238.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**(4), 518–521.

Cavalli-Sforza LL, Menozzi P and Piazza A (1994) *The History and Geography of Human Genes*, Princeton University Press: Princeton.

Colhoun HM, McKeigue PM and Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. *Lancet*, **361**(9360), 865–872.

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M and Dopazo J (2004) PupaSNP finder: A web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Research*, **32**(Web Server issue), W242–W248.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L and Nickerson DA (2004) Haplotype diversity across 100 candidate genes for inflammation,

lipid metabolism, and blood pressure regulation in two populations. *American Journal of Human Genetics*, **74**(4), 610–622.

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, *et al*. (2001) High-throughput variation detection and genotyping using microarrays. *Genome Research*, **11**(11), 1913–1925.

Devlin B and Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

Devlin B, Roeder K and Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, **60**(3), 155–166.

Dudbridge F and Koeleman BP (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics*, **75**(3), 424–435.

Ewens WJ and Spielman RS (1995) The transmission/disequilibrium test: History, subdivision, and admixture. *American Journal of Human Genetics*, **57**(2), 455–464.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**(4), 1567–1587.

Frazer KA, Tao H, Osoegawa K, de Jong PJ, Chen X, Doherty MF and Cox DR (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Research*, **14**(3), 367–372.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, *et al*. (2004) Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**(4), 388–393.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**(5576), 2225–2229.

Gordon D and Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pacific Symposium on Biocomputing*, 18–29.

Gordon D, Heath SC, Liu X and Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics*, **69**(2), 371–380.

Gorroochurn P, Hodge SE, Heiman G and Greenberg DA (2004) Effect of population stratification on case-control association studies. II. False–positive rates and their limiting behavior as number of subpopulations increases. *Human Heredity*, **58**(1), 40–48.

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R and Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, **22**(3), 239–247.

Hao K, Li C, Rosenow C and Wong WH (2004) Detect and adjust for population stratification in population-based association study using genomic control markers: An application of Affymetrix Genechip Human Mapping 10K array. *European Journal of Human Genetics*, **12**(12), 1001–1006.

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J and Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. *Nature Genetics*, **37**(1), 90–95.

Hirschhorn JN and Daly MJ (2005) Genome-wide association studies for common disease and complex traits. *Nature Reviews. Genetics*, **6**(2), 95–108.

Hirschhorn JN, Lohmueller K, Byrne E and Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genetics in Medicine*, **4**(2), 45–61.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG and McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics*, **72**(6), 1492–1504.

Ioannidis JP, Ntzani EE, Trikalinos TA and Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nature Genetics*, **29**(3), 306–309.

Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**(3), 234–236.

Lander ES and Schork NJ (1994) Genetic dissection of complex traits. *Science*, **265**(5181), 2037–2048.

Lin S, Chakravarti A and Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics*, **36**(11), 1181–1188.

Lohmueller KE, Pearce CL, Pike M, Lander ES and Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**(2), 177–182.

Marchini J, Cardon LR, Phillips MS and Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**(5), 512–517.

Mitchell AA, Cutler DJ and Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics*, **72**(3), 598–610.

Ng PC and Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Research*, **12**(3), 436–446.

Ng PC and Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**(13), 3812–3814.

Nobrega MA, Ovcharenko I, Afzal V and Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**(5644), 413.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, *et al*. (1998) Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, **63**(6), 1839–1851.

Pritchard JK and Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*, **60**(3), 227–237.

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, **65**(1), 220–228.

Pritchard JK, Stephens M and Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959.

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P (2000b) Association mapping in structured populations. *American Journal of Human Genetics*, **67**(1), 170–181.

Reich DE and Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, **20**(1), 4–16.

Reich DE and Lander ES (2001) On the allelic spectrum of human disease. *Trends in Genetics*, **17**(9), 502–510.

Reich DE, Gabriel SB and Altshuler D (2003) Quality and completeness of SNP databases. *Nature Genetics*, **33**(4), 457–458.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES and Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics*, **32**(1), 135–142.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Rosenberg NA, Li LM, Ward R and Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**(6), 1402–1422.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**(6822), 928–933.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**(6968), 789–796.

Thomas DC and Clayton DG (2004) Betting odds and genetic associations. *Journal of the National Cancer Institute*, **96**(6), 421–423.

Thomas DC and Witte JS (2002) Point: Population stratification: A problem for case-control studies of candidate-gene associations? *Cancer Epidemiology, Biomarkers and Prevention*, **11**(6), 505–512.

Tishkoff SA and Williams SM (2002) Genetic analysis of African populations: Human evolution and complex disease. *Nature Reviews. Genetics*, **3**(8), 611–621.

Turakulov R and Easteal S (2003) Number of SNPS loci needed to detect population structure. *Human Heredity*, **55**(1), 37–45.

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L and Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, **96**(6), 434–442.

Wacholder S, Rothman N and Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *Journal of the National Cancer Institute*, **92**(14), 1151–1158.

Wacholder S, Rothman N and Caporaso N (2002) Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiology, Biomarkers and Prevention*, **11**(6), 513–520.

Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB and Wu X (2004) An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. *Cancer Research*, **64**(6), 2251–2257.

Specialist Review

Pharmacogenetics and the future of medicine

Alun D. McCarthy, James L. Kennedy and Lefkos T. Middleton

Genetics Research, GlaxoSmithKline Research & Development, Uxbridge, UK

1. Introduction: a historical perspective

The unprecedented international research efforts over the last 20 years in identifying polymorphisms that are either causative of, or affect the susceptibility to, human disease, are beginning to have a visible and significant impact on medical care. First, the identification of causative mutations of monogenic diseases and subsequent phenotype–genotype correlation studies in the late 1980s and 1990s have resulted in the definition of their nosological boundaries (i.e., disease classification). The subclassification for many of these diseases has been clarified, and knowledge of their heterogeneity and pathogenesis has been realized (Weatherall, 2001). Prevention programs are already widely used around the world, and therapeutic attempts have become possible for diseases such as Huntington's disease, thalassemias, and muscular dystrophies. The efforts to sequence the human genome have resulted in the identification of most of the DNA sequence. This information, coupled with the more recent availability of high-density single-nucleotide polymorphism (SNP) maps and other tools in genomics, bioinformatics, RNA expression, and proteomics, have accelerated the application of these new tools to complex diseases and interindividual variability to drug response. For the latter, the impact is upon both drug efficacy and drug safety. The promise of the "new genetics" lies in its potential to provide insight into each individual's genetic makeup, infer susceptibility to not only a given disease but also specific kinds of symptoms (e.g., vas deferens pathology in cystic fibrosis), and to enable better therapeutic decision making. Overall, this knowledge of DNA variations can contribute to well-informed medical action. Furthermore, the recent availability of a number of imaging and molecular translational technologies allows for a more accurate diagnosis of the individual's physiological and phenotypic status. These technologies include imaging techniques that go far beyond anatomical structure, to real-time physiological processes in vivo revealing correlations with behavior (such as functional MRI; Hariri and Weinberger, 2003). What was recently remarkable is now routine – pharmacological events can be observed in vivo using MR spectroscopy, PET, and SPECT. Human biological processes can be further illuminated by an increasing

armamentarium of molecular tools exploring transcriptomic expression, proteomic and metabolomic phenomena (Campbell and Ghazal, 2004; Plumb *et al*., 2003).

The application of these physiologic and genomic technologies will provide greater insight into the heterogeneity that clearly exists within our current diagnostic nosology and a more accurate understanding of individual patients to increase the probability of successful treatment.

It is important that the term “personalized medicine (PM)” is not misunderstood. Taken literally, it could mean the development of “tailored” medicines to suit individual patients – clearly impractical. Instead, it is more appropriate to consider the application of these emerging technologies to a greater use of “personal” medical information to increase the efficiency of both diagnosis and treatment. Even considering the contribution of confounders such as environmental variables (stress, diet, and other medications) that will not allow absolute prediction, there will remain a major improvement in the probabilities of finding the right drug and dosage for a given patient. The result will be a knowledge-based reduction in the trial-and-error work of the physician, and the patient will benefit from more assured efficacy and fewer side effects.

Among the new technologies mentioned above, this paper will focus on the application of pharmacogenetics (PGx), one of the most promising biomarker technologies for addressing the challenge of variable response to medicine. The more commonly used term “pharmacogenetics” refers to the study of interindividual variations in DNA sequence related to drug response, efficacy, and toxicity (CPMP Position Paper on Terminology in Pharmacogenetics). “Pharmacogenomics” (PGm) is usually employed in a broader scope that includes genome-wide variations and potential complex interactions as well as alterations in gene expression and posttranslational modifications (proteomics) that correlate with drug response. A number of the issues associated with the evolution of PGx are applicable across a variety of other technologies; hence PGx is used as an exemplar for how the practice of medicine will be changed in future. Pharmacogenomics – and in particular, the use of gene expression patterns – is a rapidly emerging technology with a very exciting potential. The scope of this technology is being broadened by observations that lymphocyte expression profiles in easy-to-access peripheral lymphocytes appear to correlate with response in organ systems not traditionally thought to be reflected by these immune-related cells (Chon *et al*., 2004).

PGx is approaching a key milestone – it is almost 50 years since the term was first coined to capture the inheritance of altered drug metabolism (Vogel, 1959). In the intervening time, the development of the science has been impressive, but particularly so in the last few years (Roses, 2002; Goldstein *et al*., 2003): the whole area of genetics research has been transformed by technology developments underpinning the beginning of a qualitative leap in our ability to understand the basis of variable medicine response.

For much of its 50-year history, PGx has been focused on the impact of DNA variants, mainly in the drug-metabolizing system, on a medicine’s pharmacokinetic profile. *Pharmacokinetics* refers to the phenomena involved in the ushering of the active molecule into the bloodstream, its transport into the target organ, and subsequent metabolism and excretion. The impact at the target and subsequent signaling and the cascade of events that characterize the therapeutic effect are all

components of the medicine's *pharmacodynamics*. There are practical reasons why pharmacokinetics would be the first to be explored: the phenotype – concentration measures of the drug and its metabolites in blood or urine – is easily and accurately quantifiable, the number of genes involved is relatively limited, the amount of genotyping required is small (certainly by today's standards), and the impact of the genetic variation is usually high. Robust analysis of more complex phenotypes was not possible as such studies required more complex genetic analysis beyond the reach of these first technologies.

However, key technology developments have been responsible for making a wider range of efficacy and safety pharmacogenetics studies possible. First, the completion of the human genome sequence has been central to the exponential expansion of all human genetic activities, and PGx is no exception (Subramanian *et al*., 2001). Second, the advent of SNP-based high-throughput genotyping methods and platforms has fueled a dramatic increase in genotyping capacity and speed, coupled with a significant drop in costs. The increased capacity to genotype SNPs using reliable high-density maps (**http://snp.cshl.org/**) has made genome-wide scan (WGS) association studies a reality. These tools are critical in PGx, where family-based linkage analyses are both impractical and rarely possible. The most recent – and still evolving – addition to the technology armamentarium is the HapMap, which will facilitate the clustering of SNPs according to their linkage disequilibrium relationships. This will allow the selection of the most informative SNPs for genotyping, so that redundant or noninformative assays are avoided. However, such massive data-generating technologies and platforms also generate new problems in data management and massive multiple testing, requiring modified statistical analysis techniques and the application of new data mining and pattern-recognition methodologies.

2. PGx applications in enhancing the efficiency of drug discovery research and development

The last decade has witnessed astounding technological developments in many sectors of biology. However, drug discovery research has benefited from only a few novel developments despite considerable resources. In fact, the pharmaceutical industry as a whole has submitted 50% fewer new drug applications to the FDA in 2002–2003 compared to 1997–1998, while investment in biomedical research has increased almost 2.5-fold in the same period (Lesko and Woodcock, 2004). The challenge for the industry is to introduce novel strategies aimed at enhancing early stage discovery as well as the efficiency of decision making at the various stages of the pipeline (Roses, 2004).

2.1. The early stages of drug discovery: from gene to target to candidate selection

The widely held expectation, in the late 1990s, that the Human Genome Project would result in thousands of new targets led to an avalanche of ventures in

genetic research for gene and target identification for a variety of mostly common diseases, mainly in the relatively new (at the time) biotechnology sector. The majority of these initiatives were founded on a somewhat naïve extrapolation from early successes in monogenic diseases that significantly underestimated the inherent complexities of common diseases. Furthermore, the difficulties in moving from a genome sequence to disease-relevant, "tractable" (i.e., amenable to chemical screening) targets and subsequent preclinical drug discovery (Roses, 2004) were nearly universally overlooked. However, the number of known tractable targets has increased more than twofold in the last few years, from an approximate number of 500. This was mainly the result of new knowledge from the Human Genome Project, coupled with the application of novel technologies such as bioinformatics that permit more rapid exploration of gene pathways and characteristic coding regions that bear similarities to screenable gene classes (DeFife and Wong-Staal, 2002; Searls, 2003; Debouck and Metcalf, 2000). During the same period, our understanding of the molecular mechanisms underlying disease heterogeneity, gene–gene, and gene–environment interactions is steadily increasing as evidenced by the number of peer review publications in the international scientific literature, fueled by the increasing armamentarium of high-throughput genotyping, data analysis, and related technologies. Novel approaches for high-throughput experiments to discern associations between disease and disease traits with large numbers of tractable drug targets are now available (Roses, 2005).

2.2. From selecting a candidate to the clinic: the critical path of drug development

The decision as to which molecules – having been shown to have a proven effect on targets – should progress into preclinical development for subsequent use in man is a pivotal point in the pharmaceutical pipeline that has tremendous implications in financial costs, time, and effort. This "critical path" is defined as the path from candidate selection to an optimized drug through a series of (mostly animal) experiments to accumulate sufficient evidence of therapeutic effect and safety (preclinical phase). The "successful" molecule that manages to clear these hurdles is subsequently used in human (healthy and affected) volunteers in *clinical trials* designed to accurately determine its efficacy and safety.

Phase I trials comprise the first exposure in humans of a candidate medicine, and are intended to explore pharmacokinetic variables, and to ensure that there are no unacceptable safety or tolerability issues. Typical phenotypes that are studied in this phase are PK parameters such as bioavailability, clearance, and drug–drug interactions.

Phase II studies are conducted in patients aimed at evaluating a first indication of efficacy of the compound, and are sufficiently large that some safety signals may be apparent, especially "nuisance" or quality of life (QOL) side effects such as reversible changes to liver function tests. Achieving key efficacy (and safety) milestones is a critical part of the phase II studies.

Phase III studies are large trials, costing tens to hundreds of millions of dollars, providing the most convincing evidence of the efficacy and safety to support a regulatory submission. Aimed at minimizing bias and variability, they are designed as “randomized controlled trials” (RCT). The double-blind RCT model, in which neither the physician or the research subject knows what arm of the study (active drug or placebo) the patient is assigned, is the phase III “gold standard”. Owing to the size of these studies, less-common adverse events (AEs) may become apparent.

Once a medicine is registered, it can be used by a much wider population of patients. It is at this stage (*Phase IV*) that rare AEs may be identified, which could not be discovered during clinical development: even the large pivotal phase III studies lack the power to detect AEs occurring at rates less than 0.1%.

Clinical trials are, by their nature, based on a frequentist statistical model, aimed at detecting convincing evidence of efficacy and safety through the use of the drug in large numbers of patients. These numbers are needed to overcome many issues related to disease heterogeneity and the partial understanding of underlying disease mechanisms, the variability to drug response, the placebo effect, and so on. Currently, the failure rate of potential products in development is more than 80%, mainly due to poor efficacy and/or suboptimal safety. This “pipeline attrition” has a tremendous financial cost and time wastage, as the average cost to develop a market product is estimated in excess of $800 million (DiMasi *et al*., 2003). The average time to market varies between 8 and 15 years. The large-scale phase III trials consume the “lion’s share” in both time and finances and also represent approximately 50% of the overall attrition (Gilbert *et al*., 2003). There is a widespread recognition that to reduce attrition, novel approaches and tools such as PGx may enable exploration of the pathophysiological mechanisms underlying differences in drug response (Lesko *et al*., 2003). Pharmacogenetics can be applied retrospectively, looking back over the results of clinical trials, with genotypes in hand, to respond to questions and issues such as the kinetic and dynamic properties of drugs, efficacy and AEs. *Prospective PGx* would allow proactive identification of patient subgroups (e.g., disease subtypes) that would be correlated with, and be predictive of, response to a drug. If such data were available before or between Phase IIa and IIb trials, this would significantly shorten and simplify Phase III and increase their probability of success (Lesko and Woodcock, 2004). Furthermore, the ability to segment patients by therapeutic response prospectively during early Phase II development would permit the progression of multiple compounds that can treat overlapping groups of patients having the same disease label (Roses, 2004).

It is important to differentiate between safety and efficacy applications of PGx. It is now clear that the efficacy of a medicine is strongly affected by genetic variation, and is also affected by other factors such as environmental influences and placebo response. This means that efficacy PGx is unlikely to be deterministic at the individual level, but will have a critical role in significantly increasing the probability of effective response for the identified subgroups of patients. In contrast, safety PGx is focused on the individual, with specific decisions about whether to

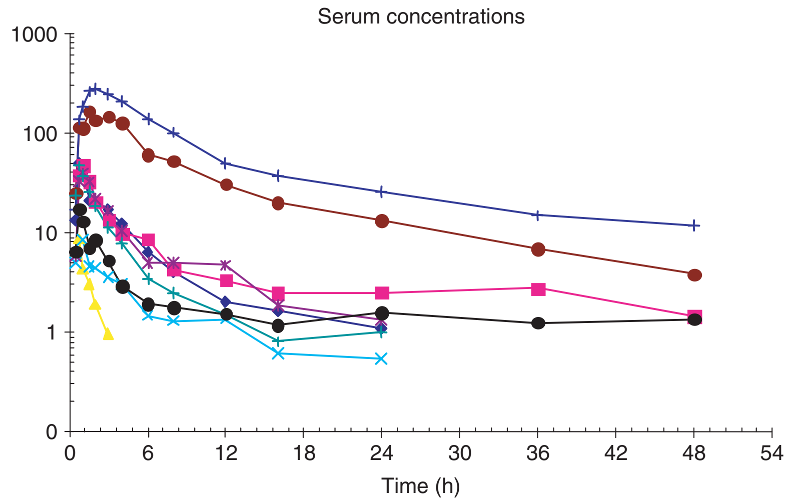


Figure 1 Typical results showing serum concentrations as a function of time for a new medicine dosed to individual human volunteers

prescribe a medicine on the basis of genetic information predicting rare dangerous events and/or common adverse effects.

2.3. PGx and pharmacokinetics: aiming to predict the right dose

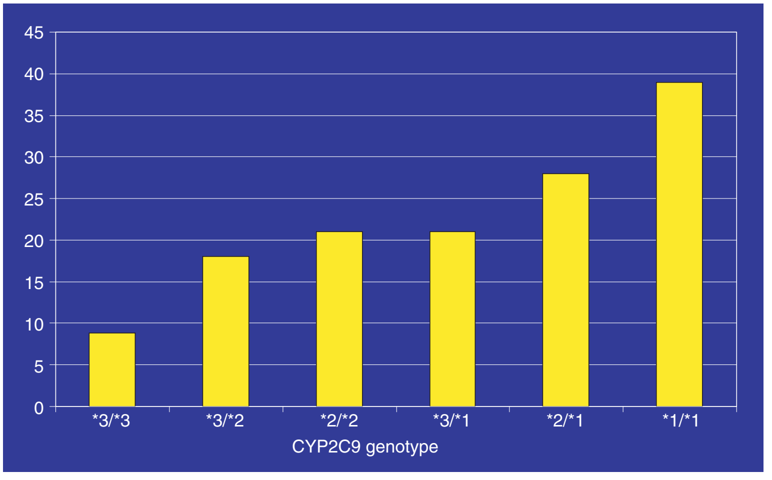
As mentioned earlier, the first use of PGx from the 1950s onward was to explore genetic variants that affected pharmacokinetics – especially drug metabolism (Daly, 2003). Excessive exposure in patients to medicines resulting in toxicity was the first PGx phenotype.

Figure 1 gives an example of the type of variable pharmacokinetics that can be seen in phase I studies. Each line represents the serum concentration over time for an individual subject. While most subjects cluster with similar peak concentrations and time course, there are unusually high exposures in a couple of patients, and low exposure in others. There are a number of reasons for such variability, and polymorphic variation in the drug-metabolizing enzymes is a fundamental variable to consider.

The principles underpinning such PGx experiments have remained essentially the same for many years, although current analyses are inevitably more extensive. For example, we are much more aware of the range of enzymes and transporters involved in drug disposition and clearance. Absorption and distribution (not just metabolism of drugs) are better understood and the relevant alleles are well characterized as well as their distribution in different ethnic groups (Lin *et al*., 1996; Lin *et al*., 2001). Box 1 shows an example of the quantitative impact that CYP polymorphisms can have on drug clearance. This example shows the effect of CYP2C9 variant on the maintenance dose of S-warfarin required by patients needing anticoagulation treatment.

Box 1: Effect of CYP2C9 genetic polymorphisms on warfarin maintenance dose

The chart shows the maintenance dose (in mg/week) of S-warfarin necessary to maintain adequate anticoagulant activity in patients with different CYP2C9 polymorphisms. While nongenetic factors also contribute to interindividual variability, CYP2C9 genetic polymorphisms have a significant impact on dose requirements (Reprinted from Clinical Pharmacology & Therapeutics, 72, Scordo MG *et al*., Influence of CYP2C9 and CYP2C19 genetic polymorphisms on warfarin maintenance dose and metabolic disease, 702–710. ©2002, Americal Society for Clinical Pharmacology & Therapeutics).



Such studies have established that at least 40% of drug metabolism is via polymorphic CYP450 enzymes (Ingelman-Sundberg, 2004). While these data appear straightforward, there are complexities that may impact the interpretation. For example, environmental confounders (such as smoking) can affect CYP expression and change the metabolic route. In particular, the whole area of drug–drug interactions shows the complexities that can be observed. It has been estimated that 6% of patients on two medications experience ADRs, whereas ADRs are reported by 50% of those on five medications, and nearly 100% of those on 10 medications. A major cause of these reactions is the changes produced by one drug in the metabolism of another through P450 pathways. Guzey *et al*. (2002) describe a case wherein a CYP2D6 extensive metabolizer became a phenotypical PM during treatment with the potent 2D6 inhibitor bupropion, demonstrating the complexity of assessing the exact causes of ADRs.

Despite almost 50 years of research, the application of PGx has yet to significantly impact clinical practice. For example, it is well established that CYP2D6 poor metabolizers have increased biological exposure and increased risk of AEs to a variety of commonly prescribed medicines. However, these observations have not as yet resulted in CYP2D6 testing being a routine part of the prescribing practice.

While the research literature is extensive, there is little in the way of clinical outcome studies, making it difficult for physicians to apply this technology in their prescribing decisions. Suitable prospective studies (e.g., looking at the effect of genotyping CYP2C9 on Warfarin dosing and bleeding events) are, however, now being initiated, and the results from these studies will be critical if PGx is to make significant inroads into medical practice.

Although prescribing practices are not yet significantly changed, PGx information is starting to appear in the written material (drug label) provided when the drug is dispensed. In the United States, labels for atomoxetine and 6-mercaptopurine provide the physician with information regarding metabolism by polymorphic enzymes (CYP2D6 and thiopurinemethyl transferase respectively). Physicians are alerted that tests are available to identify poor metabolizers, but are not required to carry out these tests prior to prescribing these medicines (Lesko *et al*., 2003).

The application of PGx to study pharmacokinetics is nonetheless a particular focus of attention at the moment, as drug regulatory authorities are actively exploring PGx tools to better understand drug exposure (and in particular, toxicity) and make this information available to physicians. For example, the US FDA and Japanese MHLW have already released draft guidelines relating to genomic data submission (Lesko *et al*., 2003), and the CPMP in Europe has an Expert Group on PGx to address these issues.

2.4. *Efficacy PGx – impact on attrition in development*

In addition to its direct clinical impact, variable efficacy is also an important issue for drug development. Failure to show efficacy in phase II studies is the most common reason for terminating the development of medicines. As noted above, large phase III studies to confirm safety and efficacy are very expensive, therefore if more information could be extracted from earlier, smaller Phase II studies to establish more clearly the efficacy of a candidate medicine, then valuable time and resources would be saved during Phase III. In fact, the variable efficacy of medicines, even in apparently homogeneous patient groups recruited in phase II studies, can obscure true and significant efficacy in a subset of patients, thus leading to inappropriate termination of the compound. One challenge in efficacy PGx is that in contrast to the more "single gene" character of the pharmacokinetic PGx described above, the efficacy phenotype is likely to be multigenic and thus requires more research to be fully clinically applicable.

Nonetheless, efficacy prediction is a very exciting area for PGx (see Table 1). By using genetic and/or other biomarkers to identify appropriately responding subgroups in phase II studies, compounds that are effective in patient subgroups may be further progressed, significantly increasing the delivery of new medicines to meet unmet patient needs, and increasing the productivity of pharmaceutical R&D. The critical issue is whether these phase II studies are appropriate to generate robust PGx data that can be used to impact further development of compounds. Although published data are scarce, some initial findings seem promising.

The example of Herceptin^{TM} (trastuzumab) highlights how a pharmacogenetic test can be used to progress medicines through the R&D pipeline. Overexpression

Table 1 Pharmacodynamic (drug target) polymorphisms associated with variation in medication response

Gene	Medication	Phenotype change	References
Angiotensin converting enzyme (ACE)	ACE inhibitors (imidapril, enalapril)	Blood pressure	Ohmichi *et al*. (1997)
		Kidney damage reduction	Jacobsen *et al*. (1998)
			Penno *et al*. (1998)
		Left ventricular hypertrophy reduction	Kohno *et al*. (1999)
		Blood vessel stenosis	Okamura *et al*. (1999)
Arachidonate 5 lipoxygenase (ALOX)	Antiasthmatics (leukotriene inhibitors)	Forced Epiratory Volume (FEV-1) improvement	Drazen *et al*. (2003)
Beta-2 adrenergic receptor (ADRBR2)	Beta-2 agonists (albuterol)	Vascular reactivity	Cockcroft *et al*. (2000) Dishy *et al*. (2001)
		Bronchodilation	Martinez *et al*. (1997) Lima *et al*. (1999) Israel *et al*. (2001)
Corticotrophin releasing hormone receptor 1	Inhaled corticosteroids	Improved lung function (FEV-1)	Tantisira *et al*. (2004)
Dopamine D3	Traditional antipsychotics (chlorpromazine, haloperidol)	Abnormal involuntary muscle movements (tardive dyskinesia) Akathisia	Steen *et al*. (1997) Basile *et al*. (1999) Lerer *et al*. (2002)
Dopamine D2	Risperidone (antipsychotic)	Response of schizophrenia symptoms	Yamanouchi *et al*. (2003)
Growth hormone receptor	Growth hormone	Increased responsiveness to growth hormone	Dos Santos *et al*. (2004)
Serotonin transporter	Antidepressants	Mood improvement	Smeraldi *et al*. (1998) Serretti *et al*. (2002)
		Side effects	Murphy *et al*. (2004) Mundo *et al*. (2001)

of the *ErbB2* gene is associated with increased tumor aggressiveness. Herceptin – a humanized monoclonal antibody against the ErbB2 receptor – is now approved for the treatment of breast cancer (Noble *et al*., 2004; Vogel and Franco, 2003). Retrospective examination of the clinical trials of Herceptin showed that a positive response was more likely in patients with tumors overexpressing ErbB2. So the measurement of ErbB2 overexpression can be used to assess whether treatment with Herceptin is appropriate. The availability of a test for a subgroup with a better probability of responding to treatment with Herceptin allowed this drug to progress through further studies to approval. The same paradigm has recently been applied to understand the response of lung cancer patients to gefitinib, where positive response is closely associated with the presence of activating mutations in the drug target (EGFR) in the tumor (Lynch *et al*., 2004). These striking results have had an immediate effect both on the way clinicians assess the role of gefitinib in cancer treatment and also on the questions that regulatory authorities require to be answered during drug development.

In another example in a recent phase II study of a GSK antiobesity compound, analysis of the whole patient groups showed efficacy less than that reported in the literature for the current "gold standard" therapies. However, PGx analysis based on candidate genes around the compound's target and presumed mechanism of action

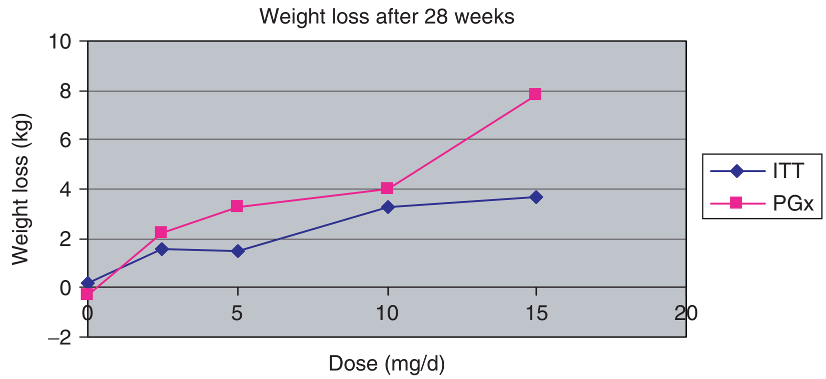


Figure 2 Dose response data showing weight loss in the whole treated population (ITT) and the subgroup identified by PGx Analysis

showed association between three genetic markers and weight loss. Using the presence of any one of these alleles to identify a subgroup, 36% of the patients could be clustered to show significantly greater weight loss. The dose response in the whole ("ITT") patient group and the PGx-defined subgroup ("PGx") is shown in Figure 2.

Analysis of a subsequent phase II study with a different antiobesity compound showed a histogram of patient numbers versus response (Figure 3). Although the average response in the placebo group was 0% change, there was considerable variation even in 8 weeks. The bulk of treated patients show some benefit, while there are some who show much greater effect ("super-responders") and also some that show very little benefit, gaining weight during the study. This visualization of the data indicates clearly the opportunities for PGx to provide insights into the variable response seen in different patient subgroups.

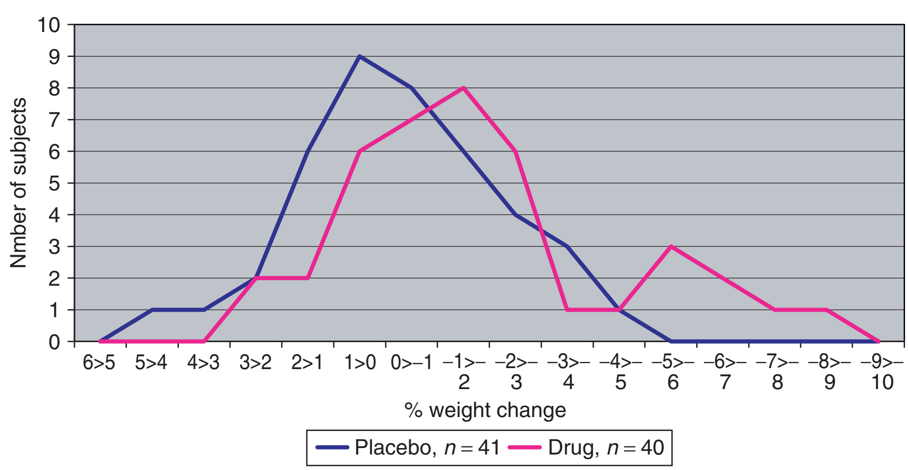


Figure 3 Variation of weight loss in placebo- and drug-treated cohorts after 8 weeks dosing

Data such as these show that phase II studies are sufficiently sized to generate PGx hypotheses that can influence subsequent development of candidate medicines. The data can be replicated and further refined in subsequent phase IIb or phase III studies.

3. PGx and safety

While there is ample evidence that medicines provide significant benefit in terms of mortality, morbidity, and cost-effectiveness, it is inevitable that ADRs are observed. Lazarou *et al*. (1998) estimated that ADRs caused approximately 106 000 deaths each year in the United States. There is increasing evidence to show that genetic variations can predispose individuals to such ADRs. For example, Rau *et al*. (2004) demonstrated that when patients were treated with antidepressants, nearly 30% of those with ADRs were CYP2D6 poor metabolizers, a group that is only 7% of Caucasians. A key objective of clinical development is to define the potential safety issues that might be associated with a new medicine, so that the risk/benefit of the medicine can be assessed.

4. Mitigating risk in development

In addition to establishing efficacy, clinical development studies must also define the safety parameters for a medicine. Although the full safety profile of a medicine cannot be established until it is in widespread use after launch, potential safety signals can be apparent in early phase II studies, and can have a significant impact on the risk for further development. For example, if reversible changes in liver function tests are seen in a small subset of patients in a phase II study, it can be difficult to assess the importance of this. Many valuable and effective medicines have a small impact on liver function, but on the other hand, a number of medicines have failed either in late development or after launch due to a subset of patients exhibiting these liver function changes, subsequently developing severe liver failure. If high-risk patients could be identified with inexpensive genetic screening before starting the drug, the overall safety would increase considerably, and abrupt termination of a drug that has progressed to late stages could be avoided.

PGx is expected to greatly illuminate the basis of liver function changes in drug development. In recent clinical studies on Tranilast (a product intended to reduce restenosis after coronary angioplasty), some 8% of individuals showed an increase in unconjugated bilirubin, which resolved on termination of drug treatment. This phenotype showed some similarity to Gilbert's syndrome, a well-recognized condition characterized by episodic increases in unconjugated bilirubin but not associated with any long-term impact on liver functioning.

Genetic analysis of Gilbert's syndrome patients has established a strong genetic susceptibility marker in the promoter of the gene *UGT-1A*, where a variable TA repeat is located. "Wild-type" activity is associated with six copies of the TA repeat, whereas seven copies of the TA repeat is associated with reduced expression of UGT-1A, and increased propensity to Gilbert's syndrome.

Figure 4 shows the result of an association analysis studying bilirubin levels as phenotype in the Tranilast-treated subjects. There is a highly significant association between the TA repeat genotype and the likelihood of developing raised bilirubin after treatment with Tranilast. This strongly supports the hypothesis that the observed hyperbilirubinemia can be described as a Tranilast-induced Gilbert's syndrome.

As the elevated bilirubin could be ascribed to a well-known, benign syndrome, with no evidence for progression to serious liver disease, the likelihood that this safety signal with Tranilast could lead to serious liver complications was significantly reduced. A similar clinical observation (i.e., elevated unconjugated bilirubin in a subset of patients) has also been seen in studies with atazanavir (BMS) and the phenotype was in turn strongly associated with the UGT-1A promoter polymorphism, suggesting a similar basis for the clinical observation.

In summary, these results not only show how PGx can provide insights into safety signals apparent at phase II but also underscore the role of PGx in providing information to help R&D decision making. This use of PGx data will be critical in facilitating the development of new medicines by pharmaceutical companies.

As significant numbers of samples were available from phase III studies (where elevated bilirubin was also seen in a small percentage of subjects), this data set has provided significant information on the power of genetic datasets to identify PGx signals and to explore other aspects of experimental design. For example, greater power can be obtained by increasing the number of controls. Perhaps more surprising is that epidemiological controls from an unrelated population can be as effective as the matched controls from the phase III study sites. This is particularly

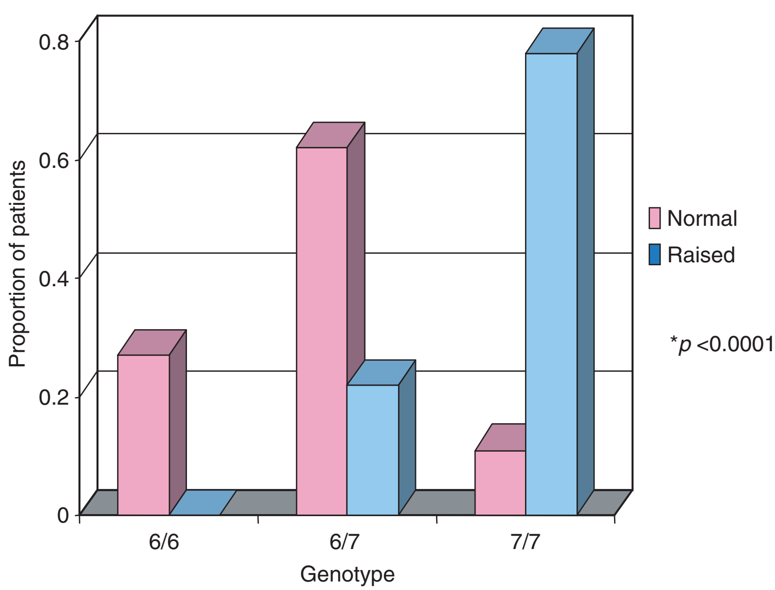


Figure 4 Effect of TA repeat number in the UGT-1A1 promoter region on the proportion of patients with normal or raised bilirubin levels after treatment with tranilast

Table 2 Impact of varying number of cases of elevated bilirubin after tranilast treatment on statistical significance of association between SNPs in the promoter region of UGT-1A1

Approximate number of cases	Approximate number of random controls	SNP poly ID			
		4082379	3729885	3730948	3737550
10	3000	0.10392	0.01542	0.04623	0.00644
20	3000	0.00143	4.37E-6	0.00014	9.96E-8
30	3000	3.93E-6	2.91E-7	4.14E-5	5.59E-9
50	3000	8.69E-8	7.39E-08	2.47E-5	1.32E-10
100	3000	1.80E-10	3.87E-13	1.24E-8	9.12E-16
120	3000	9.21E-11	1.91E-15	3.26E-10	2.21E-18
146	3000	2.56E-13	2.70E-18	6.10E-13	4.53E-23

true if genomic control measurements (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999) are performed to document and correct for stratification in the samples. In addition, it is clear that in this case, as few as 10 cases of hyperbilirubinemia would be needed to establish a positive association with the UGT-1A promoter polymorphism, using a sufficient number of controls – findings that will help immensely in the design of future PGx experiments (Danoff *et al*., 2004). Table 2 shows the PGx association p-value for increasing number of cases using 3000 population controls. Significance can be seen even with 10 cases, and with 20 cases the finding would be significant even when correcting for multiple testing in a broader candidate gene study.

5. PGx and adverse events – postlaunch

Much PGx discussion has centered on the contribution of PGx to understanding rare AEs. Because there are fewer cases of such AEs, it is unlikely that the approaches described above, that is, the selection of cases from a clinical trial population using the remainder of the cohort as controls, is viable; sufficient cases may only be available once a medicine is registered and more widely available. In this case, retrospective collection of both cases and controls may be required, and various approaches are being investigated to expedite efficient collection of such cases and controls to allow PGx experimentation.

The studies on hypersensitivity reaction (HSR) to abacavir (ABC) have been used as proof of concept on the feasibility of retrospective case/control approaches wherein the number of cases (rare AEs) is small, and genome-wide association is applied. HSR affects ∼4% of subjects with HIV who are treated with ABC, and while the clinical sequelae can be serious, tight clinical management of HIV subjects has prevented significant mortality due to this AE. DNA was collected from subjects reporting HSR after exposure to ABC, and from matched controls with no HSR symptoms after ABC treatment. Genotyping included both candidate gene and genome-wide SNP association approaches.

Candidate gene analysis identified a site in the human leukocyte antigen (HLA) of chromosome 6 with strong association with HSR (Hetherington *et al*., 2002). This finding has been confirmed by at least two other laboratories. Work to analyze

the genome-wide association data is being finalized and will be reported soon. The association between markers in the HLA region is strong in Caucasians, weaker in Hispanics, and cannot be detected in Blacks.

Studies have also been reported on genetic associations with Stevens Johnson Syndrome (SJS), a rare, serious cutaneous reaction to some medicines (Chung *et al*., 2004). The finding that the marker HLA-B*1502 is strongly associated with carbamazepine-induced SJS is another indication that this technology can provide significant new insights into the basis of such rare AEs.

There are now an increasing number of studies – HSR to ABC, SJS in response to carbamazepine, hyperbilirubinemia in response to Tranilast – that have established PGx associations with only a few tens of AE cases, possibly an indication of the genetics "truism" that the rarer an event, the more likely there is to be a strong genetic component.

These studies have led to several significant learnings:

- Retrospective case/control PGx studies on AEs are certainly feasible, although particular care must be taken in defining the phenotype with great accuracy. Phenotype collection is arguably the most critical part of any genetics study, and it may be difficult to collect sufficient detail for AEs observed away from the controlled, monitored clinical trial environment.
- Findings in one ethnic group are not necessarily applicable to other ethnic groups. It is unclear at this stage whether this is due to different genetic architecture at the same locus, or different aetiology of the AE, but resolution of this issue is important for moving forward.

Taking the physicians' point of view, each day they face the same dilemma over and over again: what is the right drug and the right starting dose for this patient? Medical training and clinical experience has taught the physician some of the considerations important in this decision, including the patient's age, sex, race, compliance, and level of anxiety. However, as every physician knows, prescribing medications is a combination of the clinical arts and inexact sciences, leading to an often bewildering variety of outcomes across a group of patients. At times, the physician may be overwhelmed by the complexity and unpredictability of a given patient's response to a medication. The way out of this trial and error prescribing lies in the increased data collection and sophistication of our knowledge regarding the causes of variable response and side effects. A cultural shift will need to occur in the way physicians approach prescribing drugs. In the coming years, automated and progressively more inexpensive genotyping will provide the physician with increasing amounts of information regarding how a given patient will react to a particular medication. However, most of this information will arrive in their hands as probabilities, particularly in predictions of efficacy, because of the interactive complexities of multiple genetic and environmental determinants. Medical schools may need to develop new curriculum in the area of pharmacogenetics, with input not only from the mainstream areas of pharmacology and genetics, but also from statistics, epidemiology, and genetic counseling, perhaps building on the experience of other parts of medicine – such as CV risk assessment – where estimation and communication of risk has been central for a number of years.

The examples given above are a clear indication of the potential of PGx to have a major impact on the development of new medicines as well as their use by prescribing physicians. Once the technology is mature, integrated into drug development processes and incorporated into clinical practice, the expectation is that:

- the delivery of new medicines through pharmaceutical R&D will increase, both through a greater understanding of nonserious safety issues, and progressing compounds effective in a genetically defined patient group;
- the clinical effectiveness of these medicines will be enhanced through an understanding of the patient subgroups and their different responses;
- the mortality and morbidity associated with AEs will be reduced as high-risk individuals can be directed toward alternative therapies.

The transformation of medical practice resulting from these developments will make significant difference to the lives of patients, as well as making the use of resources more rational, both by health care providers and pharmaceutical companies.

The development of PGx and other technologies applicable to "personalization" of medicines is proceeding at such a pace that it is not a matter of "if" these technologies impact medicine prescribing, but "when" – and to what extent. It is critical that the development of this approach is not unnecessarily hampered and that the balance between scientific progress and safeguarding patient rights is appropriate. Potential threats to the development of this technology include not only the natural resistance to new medical methods but also "genetic exceptionalism" – the view that genetic information is inherently more sensitive than other biomarkers and hence requires extreme regulation over its generation and use. This view is most common in people who do not understand genetics well and thus develop a fear of the unknown or whose views are dominated by the determinism that characterizes the single gene disorders. Despite these obstacles, the emerging data in support of PGx are compelling and warrant time and effort to develop their utility and to bring these applications to clinical fruition.

Further reading

Basile VS, Masellis M, Potkin SG and Kennedy JL (2002) Pharmacogenomics in schizophrenia; the quest for individualized therapy. *Human Molecular Genetics*, **11**(20), 2517–2530.

Malhotra AK, Buchanan RW, Kim S, Kestler L, Breier A, Pickar D and Goldman D (1999) Allelic variation in the promoter region of the dopamine D2 receptor gene and clozapine response. *Schizophrenia Research*, **36**, 92–93.

Malhotra AK, Murphy GM and Kennedy JL (2004) Pharmacogenetics of psychotropic drug response. *The American Journal of Psychiatry*, **161**, 780–796.

Schafer M, Rujescu D, Giegling I, Guntermann A, Erfurth A, Bondy B and Moller H-J (2001) Association of short-term response to haloperidol treatment with a polymorphism in the dopamine D2 receptor gene. *The American Journal of Psychiatry*, **158**, 802–804.

Spear BB, *et al*., on behalf of the Pharmacogenetics Working Group (2001) Terminology for sample collection in clinical genetic studies. *The Pharmacogenomics Journal*, **1**, 101–103.

Zanardi R, Benedetti F, De Bella D, Catalano M and Smeraldi E (2000) Efficacy of paroxetine in depression is influenced by a functional polymorphism within the promoter of the serotonin transporter gene. *Journal of Clinical Psychopharmacology*, **20**, 105–107.

References

Basile VS, Masellis M, Badri F, Paterson AD, Meltzer HY, Lieberman JA, Potkin SG, Macciardi F and Kennedy JL (1999) Association of the MscI polymorphism of the dopamine D3 receptor gene with tardive dyskinesia in schizophrenia. *Neuropsychopharmacology*, **21**(1), 17–27.

Campbell CJ and Ghazal P (2004) Molecular signatures for diagnosis of infection: application of microarray technology. *Journal of Applied Microbiology*, **96**, 18–23.

Chon H, Gaillard CA, van der Meijden BB, Dijstelbloem HM, Kraaijenhagen RJ, van Leenen D, Holstege FC, Joles JA, Bluyssen HA, *et al*. (2004) Broadly altered gene expression in blood leukocytes in essential hypertension is absent during treatment. *Hypertension*, **43**, 947–951.

Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, Ho HC, Wu JY and Chen YT (2004) Medical genetics: a marker for Stevens-Johnson syndrome. *Nature*, **428**, 486.

Cockcroft JR, Gazis AG, Cross DJ, Wheatley A, Dewar J, Hall IP and Noon JP (2000) Beta(2)-adrenoceptor polymorphism determines vascular reactivity in humans. *Hypertension*, **36**(3), 371–375.

Daly AK (2003) Pharmacogenetics of the major polymorphic metabolizing enzymes. *Fundamental & Clinical Pharmacology*, **17**(1), 27–41.

Danoff TM, Campbell DA, McCarthy LC, Lewis KF, Repasch MH, Saunders AM, Spurr NK, Purvis IJ, Roses AD, *et al*. (2004) A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. *The Pharmacogenomics Journal*, **4**, 49–53.

Debouck C and Metcalf B (2000) The impact of genomics on drug discovery. *Annual Review of Pharmacology and Toxicology*, **40**, 193–207.

Devlin B and Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

DeFife KM and Wong-Staal F (2002) Integrated approaches to therapeutic target gene discovery. *Current Opinion in Drug Discovery & Development*, **5**, 683–689.

DiMasi J, Hansen RW and Grabowski HG (2003) The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, **22**, 151–185.

Dishy V, Sofowora GG, Xie HG, Kim RB, Byrne DW, Stein CM and Wood AJ (2001) The effect of common polymorphisms of the beta2-adrenergic receptor on agonist-mediated vascular desensitization. *The New England Journal of Medicine*, **345**(14), 1030–1035.

Dos Santos C, Essioux L, Teinturier C, Tauber M, Goffin V and Bougneres P (2004) A common polymorphism of the growth hormone receptor is associated with increased responsiveness to growth hormone. *Nature Genetics*, **36**, 720–724.

Drazen JM, Yandava CN, Dube L, Szczerback N, Hippensteel R, Pillari A, Israel E, Schork N, Silverman ES, *et al*. (2003) Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. *Nature Genetics*, **22**, 168–170.

Gilbert J, Henske P and Singh A (2003) Rebuilding Big Pharma's Business Model. In Vivo: *The Business & Medicine Report*, **21**(10), 73.

Goldstein DB, Tate SK and Sisodiya SM (2003) Pharmacogenetics goes genomic. *Nature Reviews. Genetics*, **4**, 937–947.

Guzey C, Norstrom A and Spigset O (2002) Change from the CYP2D6 extensive metabolizer to the poor metabolizer phenotype during treatment with bupropion. *Therapeutic Drug Monitoring*, **24**, 436–437.

Hariri AR and Weinberger DR (2003) Functional neuroimaging of genetic variation in serotonergic neurotransmission. *Genes, Brain, and Behavior*, **2**, 341–349.

Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Fling ME, *et al*. (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet*, **359**, 1121–1122.

Ingelman-Sundberg M (2004) Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Naunyn-Schmiedebergs Archives of Pharmacology*, **369**, 89–104.

Israel E, Drazen JM, Liggett SB, Boushey HA, Cherniack RM, Chinchilli VM, Cooper DM, Fahy JV, Fish JE, Ford JG, *et al*. (2001) National Heart, Lung, and Blood Institute's Asthma Clinical Research Network. Effect of polymorphism of the beta(2)-adrenergic receptor on response to

regular use of albuterol in asthma. *International Archives of Allergy Immunology*, **124**(1–3), 183–186.

Jacobsen P, Rossing K, Rossing P, Tarnow L, Mallet C, Poirier O, Cambien F and Parving HH (1998) Angiotensin converting enzyme gene polymorphism and ACE inhibition in diabetic nephropathy. *Kidney International*, **53**(4), 1002–1006.

Kohno M, Yokokawa K, Minami M, Kano H, Yasunari K, Hanehira T and Yoshikawa J (1999) Association between angiotensin-converting enzyme gene polymorphisms and regression of left ventricular hypertrophy in patients treated with angiotensin-converting enzyme inhibitors. *The American Journal of Medicine*, **106**(5), 544–549.

Lazarou J, Pomeranz BH and Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Journal of the American Medical Association*, **279**, 1200–1205.

Lerer B, Segman RH, Fangerau H, Daly AK, Basile VS, Cavallaro R, Aschauer HN, McCreadie RG, Ohlraun S, Ferrier N, *et al*. (2002) Pharmacogenetics of tardive dyskinesia: combined analysis of 780 patients support association with dopamine D3 receptor gene Ser9Gly polymorphism (A Multi-Center Study). *Neuropsychopharmacology*, **27**, 105–119.

Lesko LJ, Salerno RA, Spear BB, Anderson DC, Anderson T, Brazell C, Collins J, Dorner A, Essayan D, Gomez-Mancilla B, *et al*. (2003) Pharmacogenetics and pharmacogenomics in drug development and regulatory decision making: report of the first FDA-PWG-PhRMA-DruSafe Workshop. *Journal of Clinical Pharmacology*, **43**, 342–358.

Lesko LJ and Woodcock J (2004) Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective. *Nature Reviews. Drug Discovery*, **3**, 763–769.

Lima JJ, Thomason DB, Mohamed MH, Eberle LV, Self TH and Johnson JA (1999) Impact of genetic polymorphisms of the beta2-adrenergic receptor on albuterol bronchodilator pharmacodynamics. *Clinical Pharmacology Therapeutics*, **65**(5), 519–525.

Lin KM, Poland RE, Wan YY, Smith M and Lesser IM (1996) The evolving science of pharmacogenetics: clinical and ethnic perspectives. *Psychopharmacology Bulletin*, **32**, 205–217.

Lin KM, Smith MW and Ortiz V (2001) Culture and psychopharmacology. *The Psychiatric Clinics of North America*, **24**, 523–537.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto R, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, *et al*. (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, **350**, 2129–2139.

Martinez FD, Graves PE, Baldini M, Solomon S and Erickson R (1997) Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *Clinical Investigation*, **100**(12), 3184–3188.

Mundo E, Walker M, Cate T, Macciardi FM and Kennedy JL (2001) The role of serotonin transporter protein gene in antidepressant-induced mania in bipolar disorder. Preliminary findings. *Archives of General Psychiatry*, **58**, 539–544.

Murphy DL, Lerner A, Rudnick G and Lesch KP (2004) Serotonin transporter: gene, genetic disorders, and pharmacogenetics. *Molecular Interventions*, **4**(2), 109–123.

Noble ME, Endicott JA and Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. *Science*, **303**, 1800–1805.

Ohmichi N, Iwai N, Uchida Y, Shichiri G, Nakamura Y and Kinoshita M (1997) Relationship between the response to the angiotensin converting enzyme inhibitor imidapril and the angiotensin converting enzyme genotype. *American Journal of Hypertension*, **10**(8), 951–955.

Okamura A, Ohishi M, Rakugi H, Katsuya T, Yanagitani Y, Takiuchi S, Taniyama Y, Moriguchi K, Ito H, Higashino Y, *et al*. (1999) Pharmacogenetic analysis of the effect of angiotensin-converting enzyme inhibitor on restenosis after percutaneous transluminal coronary angioplasty. *Angiology*, **50**(10), 811–822.

Penno G, Chaturvedi N, Talmud PJ, Cotroneo P, Manto A, Nannipieri M, Luong LA and Fuller JH (1998) Effect of angiotensin-converting enzyme (ACE) gene polymorphism on progression of renal disease and the influence of ACE inhibition in IDDM patients: findings from the EUCLID

Randomized Controlled Trial. *EURODIAB Controlled Trial of Lisinopril in IDDM. Diabetes*, **47**(9), 1507–1511.

Plumb RS, Stumpf CL, Granger JH, Castro-Perez J, Haselden JN and Dear GJ (2003) Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Communications in Mass Spectrometry*, **17**, 2632–2638.

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics*, **65**, 220–228.

Rau T, Wohlleben G, Wuttke H, Thuerauf N, Lunkenheimer J, Lanczik M and Eschenhagen T (2004) CYP2D6 genotype: impact on adverse effects and nonresponse during treatment with antidepressants-a pilot study. *Clinical Pharmacology and Therapeutics*, **75**(5), 386–393.

Roses AD (2002) Genome-based pharmacogenetics and the pharmaceutical industry. *Nature Reviews. Drug Discovery*, **1**, 541–554.

Roses AD (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nature Reviews. Drug Discovery*, **3**, 645–656.

Roses AD, Burns DK, Chissoe S, Middleton L and St Jean P (2005) Disease-specific target selection: a critical first step down the right road. *Drug Discovery Today*. **10**(3), 177–189.

Scordo MG, Pengo V, Spina E, Dahl ML, Gusella M and Padrini R (2002) Influence of CYP2 C9 and CYP2 C19 genetic polymorphisms on warfarin maintenance dose and metabolic clearance. *Clinical Pharmacology and Therapeutics*, **72**, 702–710.

Searls DB (2003) Pharmacophylogenomics: genes, evolution & drug targets. *Nature Reviews. Drug Discovery*, **2**, 613–623.

Serretti A, Lilli R and Smeraldi E (2002) Pharmacogenetics in affective disorders. *European Journal of Pharmacology*, **438**(3), 117–128.

Smeraldi E, Zanardi R, Benedetti F, De Bella D, Perez J and Catalano M (1998) Polymorphism within the promoter of the serotonin transporter gene and antidepressant efficacy of fluvoxamine. *Molecular Psychiatry*, **3**, 508–511.

Steen VM, Lovlie R, MacEwan T and McCreadie RG (1997) Dopamine D3-receptor gene variant and susceptibility to tardive dyskinesia in schizophrenic patients. *Molecular Psychiatry*, **2**(2), 139–145.

Subramanian G, Adams MD, Ventnre JC and Broder S (2001) Implications of the human genome for understanding human biology and medicine. *Journal of the American Medical Association*, **286**, 2296–2307.

Tantisira KG, Lake S, Silverman ES, Palmer LJ, Lazarus R, Silverman EK, Liggett SB, Gelfand EW, Rosenwasser LJ, Richter B, *et al*. (2004) Corticosteroid pharmacogenetics: association of sequence variants in *CRHR1* with improved lung function in asthmatics treated with inhaled corticosteroids. *Human Molecular Genetics*, **13**, 1353–1359.

Vogel F (1959) Moderne probleme der Humangenetik. *Ergebnisse der Inneren Medizin und Kinderheilkunde*, **12**, 52, 125.

Vogel CL and Franco SX (2003) Clinical experience with trastuzumab (herceptin). *The Breast Journal*, **9**, 452–462.

Weatherall DJ (2001) Towards molecular medicine; reminiscences of the haemoglobin field, 1960-2000. *British Journal of Haematology*, **115**, 729–738.

Yamanouchi Y, Iwata N, Suzuki T, Kitajima T, Ikeda M and Ozaki N (2003) Effect of DRD2, 5-HT2A, and COMT genes on antipsychotic response to risperidone. *The Pharmacogenomics Journal*, **3**(6), 356–361.

Specialist Review

SNPs and human history

Jeffrey D. Wall

The University of Southern California, Los Angeles, CA, USA

1. Introduction

Both physical anthropologists and evolutionary biologists have long been fascinated by the evolution of our own species. Researchers have tended to focus on either the demographic history, such as population structure, population expansions/bottlenecks, and the colonization of new islands/continents, or the selective history such as the genetic basis of human-specific adaptations (e.g., language) or the evolution of population-specific local adaptations (e.g., lactose tolerance). Single-nucleotide polymorphisms (SNPs) are informative about these questions, especially SNPs that are obtained from resequencing studies. By carefully analyzing patterns of SNP variation within and between populations, we can learn a lot about human evolution.

Early studies of human SNP variation centered around mitochondrial DNA (mtDNA) and the nonrecombining portion of the Y chromosome (e.g., Vigilant *et al*., 1991; Hammer, 1995). These two regions are easier to study because of their uniparental mode of inheritance, but they represent a minute fraction of the total information contained in our genomes. With the development of new SNP typing and sequencing technology, we have witnessed an explosion of SNP data over the last several years. Researchers have gathered SNP data from markers throughout the genome (e.g., Akey *et al*., 2002; Gabriel *et al*., 2002; IHC, 2003) as well as resequencing data from genes (Reich *et al*., 2001; Stephens *et al*., 2001; Carlson *et al*., 2003), intergenic regions (Frisse *et al*., 2001), and even whole chromosomes (Patil *et al*., 2001). These and other studies have revealed three major patterns: (1) overall levels of variation are quite low relative to other great ape species, (2) there is a tremendous amount of variability across regions in overall patterns of SNP variation, including differences in the level of nucleotide diversity, the strength of linkage disequilibrium, and the allele frequencies of SNPs, and (3) there are systematic differences in levels and patterns of polymorphism among human populations. The connections between these observations and explicit models of human evolution are not always clear.

2. Small effective population size

Studies of human variation have consistently found low levels of polymorphism, corresponding to a long-term "effective population size" (N_e) of roughly 10 000–

15 000 (Ptak *et al*., 2004; Yu *et al*., 2004). In other words, levels of human genetic variation are equal to what would be expected in a selectively neutral, random-mating population of 10 000–15 000 individuals. On the surface, this observation is surprising, since hominins have occupied much of the Old World for more than 1 million years. In contrast, great ape species have had much smaller ranges; yet, three out of four species (all but the bonobo) have effective population sizes substantially larger than the human effective population size (Kaessmann *et al*., 2001; Yu *et al*., 2004). The small human N_e has been taken as evidence against the multiregional model of modern human origins (Harpending *et al*., 1998). The multiregional model, which posits that modern humans evolved simultaneously on multiple continents, generally assumes that the fossil hominins that cover much of Europe, Asia, and Africa were our direct ancestors (Wolpoff *et al*., 1984). How could such a widely dispersed successful species have such a small N_e? One possible explanation is that many of the humanlike fossils that have been discovered (e.g., the Neanderthals in Europe) are not our direct ancestors. This is the main premise of the Recent African Origin model, which posits that modern humans evolved in a small area of Africa roughly 130 000–200 000 years ago (Stringer and Andrews, 1988). Under this model, modern humans subsequently expanded and replaced all other hominine populations. However, it is not clear whether $N_e = 10\,000–15\,000$ is really incompatible with long-term continuous occupation of the Old World as census sizes can be substantially larger than effective population sizes. In fact, recent studies of historical human populations have suggested that census sizes can be hundreds of times larger than effective population sizes (e.g., Helgason *et al*., 2003).

3. Heterogeneity across regions

As more and more data have become available, it has become apparent that patterns of SNP variation are not uniform across the genome. For example, while many regions have an excess of low-frequency variants, others do not. Some regions show extensive linkage disequilibrium between markers that are hundreds of kilobases apart, while other regions have a complete breakdown of LD over just a few kilobases. Some regions show extensive differentiation between different populations, while others show no differences across populations. Much of this heterogeneity is due to the inherent stochasticity of the evolutionary process, but part is attributable to variation in the underlying biological parameters. Both the neutral mutation rate and the recombination rate are not constant across the genome; the recombination rate, for example, can vary by several orders of magnitude over fine scales (Crawford *et al*., 2004; McVean *et al*., 2004). Natural selection can also lead to unusual patterns of variation; the *FOXP2* gene shows an excess of rare alleles due to a recent selective sweep (Enard *et al*., 2002), while other genes show increased differentiation due to adaptation to specific environments (Hamblin *et al*., 2002).

This heterogeneity across regions makes the interpretation of the patterns of variation at any particular locus extremely challenging. For example, an excess of rare polymorphisms at a locus may be evidence for recent population growth (Wall and

Przeworski, 2000), or the signature of recent adaptive evolution (Braverman *et al*., 1995), or just the chance effects of genetic drift. Without additional information, it may be impossible to distinguish among these three possibilities. It is for this reason that studies of a single linkage group, such as the plethora of evolutionary studies on mitochondrial SNPs, are inherently limited.

The effects of natural selection on patterns of variation tend to be localized to small regions (i.e., sites linked to those under selection) owing to recombination, while the effects of demography (e.g., subdivision, admixture, changes in population size, etc.) are visible throughout the whole genome. Thus, to adequately address questions of demography, it is necessary to gather polymorphism data from many evolutionarily independent regions, preferably noncoding regions that are less likely to have been affected by natural selection. Once enough data from putatively neutral loci have been gathered, one can use them as a baseline when looking for loci that have recently been affected by natural selection. Selected loci may show up as outliers in the empirical distribution of some summary of the data (e.g., Hudson *et al*., 1987; Akey *et al*., 2002).

4. Differences across populations

Despite the heterogeneity across regions in patterns of SNP variation, there are systematic differences in the patterns of variation of different human populations. On the basis of SNP data from hundreds of loci, we know that (many but perhaps not all) sub-Saharan African and African-American populations have more genetic variation (Frisse *et al*., 2001; Carlson *et al*., 2003), harbor more rare alleles (Ptak and Przeworski, 2002), and have lower levels of linkage disequilibrium (Frisse *et al*., 2001; Reich *et al*., 2001; Wall and Pritchard, 2003) than other populations. These systematic differences suggest differences in the underlying histories of sub-Saharan African (and African-American) versus other populations. We note in passing that these observations are supported by other marker systems such as microsatellites (e.g., Bowcock *et al*., 1994).

Some have claimed that all non-African populations have undergone a recent population bottleneck (Tishkoff *et al*., 1996; Reich *et al*., 2001), as predicted by the Recent African Origin model. Indeed, recent bottlenecks lead to reduced variation, increased linkage disequilibrium and a shift toward more common alleles (e.g., Reich *et al*., 2001), so some form of bottleneck is a likely and appealing explanation. There still remain some unresolved questions though. Little work has been done to estimate the timing and strength of a putative bottleneck, to ascertain whether all non-African populations show evidence of a recent bottleneck and all sub-Saharan African populations (and populations with a substantial sub-Saharan African component) do not show evidence of a bottleneck, or to rule out other possible demographic explanations (e.g., Relethford and Jorde, 1999).

5. Models and data

What do the patterns of SNP variation tell us about the various models of human evolution? It has been difficult to come to definitive conclusions (e.g., on the validity of the Recent African Origin or multiregional models) for several reasons. First

of all, anthropological genetics models are verbal models that are not necessarily easily defined or quantified. The multiregional model is particularly vague, with a definition that seems to have shifted over the years (Relethford, 2001). Even the simpler Recent African Origin model is rarely made specific. Only explicitly formulated models can be rigorously tested using genetic data.

Take for example the debate on modern human origins. In a genetic sense, we can ask what contribution various "archaic" human populations (e.g., the Neanderthals in Europe or *Homo erectus* in East Asia) have made to the contemporary human gene pool. The multiregional model predicts that this contribution would be substantial, while the Recent African Origin model predicts that this contribution is negligible. Other models predict intermediate contributions of archaic populations to the modern gene pool (e.g., Bräuer, 1984). When the debate is phrased in this way, we have a well-defined genetic question and a set of mutually exclusive predictions that might allow us to discriminate between models. This well-defined question is not particularly easy to answer, because the predictions of the different models differ only slightly and because the theoretical tools for distinguishing between these small differences have not yet been developed. There is hope, though, that we will be able to solve this question in the next few years.

Given only contemporary human sequences, it is difficult to say much about admixture levels 30 000–100 000 years ago since genetic drift in the intervening years would have mostly obscured the pattern of admixture. What we would really like is to have a direct comparison of DNA sequences from both archaic and modern populations. This is technically challenging but in a recent breakthrough, researchers have managed to sequence fragments of Neanderthal mtDNA from fossil bones (e.g., Krings *et al*., 1997; Serre *et al*., 2004). All of the Neanderthal mtDNA sequences were found to be highly diverged from contemporary human mtDNA sequences, implying that Neanderthals made no contribution to the contemporary mtDNA gene pool. However, we still do not know whether Neanderthals made any contribution to the nuclear gene pool. Recent studies suggest that even if Neanderthal genes make up as much as 25% of the current gene pool, Neanderthal mtDNA could still have been lost from the current mtDNA gene pool by chance (Nordborg, 1998; Serre *et al*., 2004).

It is unlikely that any appreciable amount of nuclear DNA sequence will be recovered from Neanderthal remains, so further progress will probably come from the analysis of contemporary human DNA. Given a simple admixture model and polymorphism data from a sufficiently large number of loci, it should be possible to estimate admixture proportions (Wall, 2000). The question remains open in part because the requisite experimental data have not yet been gathered.

6. Future directions: data

Despite the presence of a huge amount of SNP data gathered by the genetic mapping community, much of these data are not directly useful for evolutionary studies. Genotype data on SNPs from the public databases (e.g., Gabriel *et al*., 2002; IHC, 2003) do not provide information on the levels of variation and are difficult to analyze because of ascertainment bias. Public database SNPs are

not a random collection of SNPs. Most were ascertained by sequencing a small number of (predominantly European) reference chromosomes. SNPs discovered in this manner are often typed in a larger group of individuals, and the patterns of linkage disequilibrium and population frequencies in such data are affected by the process by which the SNPs were originally identified (Nielsen and Signorovitch, 2003). While, in principle, analytical methods can correct for ascertainment bias, this presupposes that the process by which SNPs were first ascertained is completely known; often this is not the case. In fact, even the number of reference chromosomes is often unknown! So, it is far simpler for analysis and interpretation to gather resequencing data.

Ideal data for studies of human population history might be resequencing data from many independent regions in a worldwide sample of humans. To reduce the confounding effects of natural selection, the regions should be single-copy noncoding regions in areas of low gene density and high recombination. To help distinguish various demographic effects (e.g., population structure and population growth), it is best to collect data with a consistent sampling scheme and to have a sizable sample size from each population studied (Ptak and Przeworski, 2002).

While there are some large-scale resequencing projects in progress (e.g., Frisse *et al*., 2001, or the sequencing of ENCODE regions for the HapMap project), these projects have generally considered a limited sample of human populations: a European population, an East Asian population, and a West African population. It is still an open question whether such limited sampling adequately represents the full spectrum of human SNP diversity. If recent studies of microsatellite variability (e.g., Rosenberg *et al*., 2002) are any indication, the answer is no. Our understanding of human demographic history will be richer if we gather sequence data from multiple sub-Saharan African populations, and from populations from the Middle East, South Asia, Melanesia, and the Americas.

For questions regarding selective history, many of the same suggestions apply. Resequencing data from a worldwide sample of humans provide the ideal data for understanding the effects of natural selection on the human genome. The only difference would be a focus on functional regions (e.g., genes or promoter regions) rather than on putatively nonfunctional ones. Ideally, all population genetic studies (whether the focus were on selection or demography) would use a common set of DNA samples. That way it would be easier for the studies of natural selection to control for the confounding effects of demography. The Human Genome Diversity Project cell line panel housed at the Foundation Jean Dausset (CEPH) in Paris is one obvious possibility (Cann *et al*., 2002).

7. Future directions: models

From the modeling side, more effort needs to be spent formulating specific models, estimating parameters, and developing methodology for discriminating between models. Most models are based on the coalescent (Hudson, 1990), which provides a simple yet powerful method for simulating population histories. So far, methods have been developed for looking at population growth (e.g., Pluzhnikov *et al*., 2002), population structure (e.g., Hey *et al*., 2004), ancient admixture (Wall, 2000),

recent natural selection (e.g., Przeworski, 2003), and variation in recombination rates (Crawford *et al*., 2004; McVean *et al*., 2004). In reality, these models are too simple. An accurate representation of human demography requires both population structure and changes in population size, while studies of natural selection cannot ignore the underlying demography. The challenge lies in formulating models that are detailed enough to capture the fundamental biological complexity but simple enough to be computationally tractable. These methods will need to be tailored toward handling the flood of SNP data that are already becoming available.

8. Conclusions

The field of human evolutionary genetics is in a state of flux. The answers to crucial questions that have been open for decades seem to be tantalizingly close; in most cases, we have partial answers but no definitive conclusions. Given the recent advances in sequencing technology, there will soon be a flood of sequencing data gathered specifically to answer human evolutionary questions. The methods for analyzing these data will soon follow. This is an exciting time – one where we will reach the limit of what we can learn about our own history from our own DNA sequences.

References

Akey JM, Zhang G, Zhang K, Jin L and Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Bräuer G (1984) The Afro-European sapiens hypothesis and hominid evolution in East Asia during the late middle and upper Pleistocene. *Courier Forschungsinstitut Senckenberg*, **69**, 145–165.

Braverman JM, Hudson RR, Kaplan NL, Langley CH and Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783–796.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al*. (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L and Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**, 518–521.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA and Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, **36**, 700–706.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP and Pääbo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J and Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, **69**, 831–843.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Hamblin MT, Thompson EE and Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. *American Journal of Human Genetics*, **70**, 369–383.

Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature*, **378**, 376–378.

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR and Sherry ST (1998) Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1961–1967.

Helgason A, Hrafnkelsson B, Gulcher JR, Ward R and Stefansson K (2003) A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *American Journal of Human Genetics*, **72**, 1370–1388.

Hey J, Won YJ, Sivasundar A, Nielsen R and Markert JA (2004) Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Molecular Ecology*, **13**, 909–919.

Hudson RR (1990) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, Vol. 7, Harvey PH and Partridge L (Eds.), Oxford University Press: New York, pp. 1–44.

Hudson RR, Kreitman M and Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796.

Kaessmann H, Wiebe V, Weiss G and Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics*, **27**, 155–156.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M and Pääbo S (1997) Neanderthal DNA sequences and the origin of modern humans. *Cell*, **90**, 19–30.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR and Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Nielsen R and Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, **63**, 245–255.

Nordborg M (1998) On the probability of Neanderthal ancestry. *American Journal of Human Genetics*, **63**, 1237–1240.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Pluzhnikov A, Di Rienzo A and Hudson RR (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, **161**, 1209–1218.

Przeworski M (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.

Ptak SE and Przeworski M (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, **18**, 559–563.

Ptak SE, Voelpel K and Przeworski M (2004) Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics*, **167**, 387–397.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Relethford JH (2001) *Genetics and the Search for Modern Human Origins*, Wiley-Liss: New York.

Relethford JH and Jorde LB (1999) Genetic evidence for larger African population size during recent human evolution. *American Journal of Physical Anthropology*, **108**, 251–260.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G and Pääbo S (2004) No evidence of Neanderthal mtDNA contribution to early modern humans. *PLoS Biology*, **2**, E57.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al*. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.

Stringer CB and Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science*, **239**, 1263–1268.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P and Krings M (1996) Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. *Science*, **271**, 1380–1387.

Vigilant L, Stoneking M, Harpending H, Hawkes K and Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–1507.

Wall JD (2000) Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, **154**, 1271–1279.

Wall JD and Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**, 587–597.

Wall JD and Przeworski M (2000) When did the human population size start increasing? *Genetics*, **155**, 1865–1874.

Wolpoff MH, Wu X and Thorne AG (1984) Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia. In *The Origins of Modern Humans: A World Survey of the Fossil Evidence*, Smith F and Spencer F (Eds.), Liss: New York, pp. 411–483.

Yu N, Jensen-Seaman MI, Chemnick L, Ryder O and Li WH (2004) Nucleotide diversity in gorillas. *Genetics*, **166**, 1375–1383.

Short Specialist Review

Evolutionary modeling in haplotype analysis

Peter Donnelly
University of Oxford, Oxford, UK

Continuing advances in high-throughput genotyping technologies mean that data sets documenting genetic variation at very closely spaced markers are becoming routine, even on genomic scales. The International Haplotype Map project (HapMap), which by mid-2005 will provide genotypes at a density of more than 1 SNP marker per kilobase for 270 individuals from four population samples, is just one public resource (**www.hapmap.org**). With the advent of genome-wide association studies for complex human diseases, genetic variation data at hundreds of thousands of markers will be combined with phenotype information from disease cases and from control individuals.

The patterns observed in such variation data result from a complex interaction of different effects: the genetic forces of mutation, recombination, and possibly natural selection; the demographic history of the populations carrying ancestors of the sampled chromosomes; and a whole series of chance events. In addition, the genomes of individuals affected by a disease with a genetic component are more likely than those of control individuals to carry the genetic mutations that increase the relative risk of suffering from the disease. In principle, then, variation data is informative about each of these effects, and different studies may focus on different aspects. For example, several studies have recently utilized variation data to elucidate the fine-scale structure of recombination rate variation in the human genome (e.g., McVean *et al*., 2004; Crawford *et al*., 2004), to detect regions possibly under natural selection (Sabeti *et al*., 2002), and to study the histories of human populations (e.g., Reich *et al*., 2002).

In human disease studies and for other many other purposes, the evolutionary history of the sample and the genetic forces that shape it are not the primary interest. Nonetheless, even in these contexts, it can be helpful to bear in mind these evolutionary effects. At an informal level, they allow useful insights into the structure of the data. Of more practical relevance, it is becoming clear that analytical methods that directly or indirectly model evolutionary history are often more powerful than existing approaches, even for questions not directly related to that history. In summary, if we want to use genetic variation data from population samples to learn about evolutionary forces (recombination rates, selective pressures), we must use some kind of model that relates the forces of interest to aspects of the data. But even if we want to use the data for other purposes,

including disease mapping, we will often be able to get more information from the data if we exploit knowledge of the evolutionary forces that shaped it.

We illustrate by briefly considering one particular problem: inference of haplotype phase from genotype data. Most genotyping methods give the diploid genotypes at each of the markers tested independently. Of course, in any individual, variant alleles at nearby markers are arranged along the individual's two chromosomes. If we were able to read each chromosome separately, we would learn the two *haplotypes* carried by that individual, that is, the sequence of variants along each of the two chromosomes. In practice, genotyping (and DNA sequencing) methods routinely provide unphased information: we learn which variants the individual has at each marker, but not the "joined up" information about which variant at one marker is on the same chromosome as the variants at a nearby marker. For many purposes, it is the haplotypes themselves that are of primary interest as these are the natural units by which genetic information is passed from generation to generation. There are experimental methods for directly determining haplotype phase, and much of the phase information can be recovered if parents or offspring are genotyped in addition to the individuals of primary interest. But these approaches involve considerable additional cost.

An alternative is to use statistical methods to infer, or estimate, phase from the original genotype data, and it turns out that the best such statistical methods perform remarkably well. This may initially appear paradoxical. Where is the information coming from to learn about phase? Evolutionary models provide the insights. First, although there is an enormous number (2^{L-1} for L SNP loci) of different possible haplotypes, it turns out that models predict that there will be relatively few different haplotypes actually present in a population sample, since the haplotypes present share a common ancestor relatively recently (e.g., Donnelly and Tavaré, 1995). Were it not for mutation and recombination, all sampled haplotypes would be the same as their most recent common ancestor. Mutation and recombination do create new haplotypes, but our understanding of evolutionary models shows that there is just not enough time to create too many, and certainly nowhere near the number of all possible haplotypes. Empirical studies in humans and other species have verified this theoretical prediction. Methods for inferring phase from genotype data directly or indirectly take advantage of this observation, in effect looking for collections of haplotypes that are consistent with the genotype data in which a small number of haplotypes recur in many different individuals. (Some approaches explicitly try to minimize the number of observed haplotypes, although this is a hard computational problem for large data sets. In other approaches, including, for example, the use of maximum likelihood via the EM algorithm, it happens indirectly.)

Phase inference is thus a simple example where an understanding of evolutionary models is helpful. In fact, there is another lesson from that example. The most accurate phase inference methods such as PHASE (Stephens *et al*., 2001; Stephens and Donnelly, 2003) do better than other approaches because they exploit this understanding more deeply. Not only will there be relatively few haplotypes in a sample, but exactly because they share a recent common ancestor, the haplotypes that do occur will be relatively similar. PHASE embodies the intuition that when we see a haplotype that has not been previously observed, it is likely to be closely

related to haplotypes already seen in the sample, in the sense that it can be created from them by a small number of mutation and recombination events.

What do the evolutionary models used in these applications look like? There is a natural family of models based around the *coalescent*. In its simplest form, the coalescent is a probabilistic description of the ancestral trees that link sampled haplotypes back to their most recent common ancestor, in a large randomly mating population of constant size. There are extensions to this basic model to incorporate much more complicated demographic scenarios, the effects of recombination, and of natural selection (see Donnelly and Tavaré, 1995, or Nordborg, 2001, and references therein). Empirical data documenting genetic variation has a complicated correlation structure: the variants carried on a particular chromosome are correlated as one moves along the chromosome (this is known as linkage disequilibrium). In addition, these are correlated across chromosomes because of the shared evolutionary ancestry. When it comes to interpreting genetic variation data, we can think of the coalescent as providing a model for the randomness generating the data, and in particular for the correlations in the data.

Although coalescent models are well understood, and very useful for simulating data under various scenarios, their use in empirical data analysis is not without its challenges. In effect, the complications of the model make its direct use difficult. There has been considerable recent research activity in this area into approaches that try to utilize all of the information in the data (e.g., Stephens, 2001; Fearnhead and Donnelly, 2001, and references therein). But these are extremely computationally intensive, and currently impracticable for all but small data sets. A common alternative approach has been to summarize the information in the data, and to base inference on these summaries. More recently, various authors have developed approximations to the full coalescent model. The challenge in this approach is to retain enough of the features of the ordinary coalescent to keep much of the power of coalescent-based approaches, yet to simplify things sufficiently so as to make inference practicable for the sizes of data set currently being generated.

One interesting application of this last strategy has been in methods for studying fine-scale variation in recombination rates, and in detecting recombination hotspots, from population data. It has long been known from pedigree studies that recombination rates vary over chromosomal scales in the human genome. However, these approaches do not have the resolution to estimate recombination rates over kilobase scales (since recombination occurs on average once per 100 Mb in human meiosis). More recently, sperm studies have characterized recombination hotspots (in males) as regions of 1–2 kb in which recombination rates are elevated considerably compared to the background rate. Sperm studies are restricted to male recombination, and are not practicable over genomic scales. Population variation data is shaped by historical recombination events, among many other things, and so in principle carries information about recombination rates, provided these can be teased out from the other effects. Several statistical methods have been developed, each employing a different approximation to the coalescent model (McVean *et al*., 2004; Crawford *et al*., 2004; Fearnhead *et al*., 2004). These methods have revealed extensive fine-scale variation in historical recombination rates, and have shown that recombination hotspots are a ubiquitous and common feature of the human genome, providing valuable insights not available from any other approach. (There

is growing evidence that this fine-scale recombination landscape is evolving very quickly. If this is the case, population-based methods, which estimate sex-averaged recombination rates averaged over long periods of time, will be estimating fundamentally different quantities from sperm studies, which estimate current male recombination rates in the individuals studied. For many purposes, including disease mapping and most population genetics applications, it is the historical rates that are of primary interest. In understanding the molecular drivers of recombination, it is the current rates that are of interest.)

The coalescent models that underlie current methods make very simple assumptions about population demography, and concerns are often expressed about how unrealistic these are for actual human population histories. Given this obvious mismatch between the modeling assumptions and reality, how can coalescent approaches be useful at all? Statisticians have an aphorism that "all models are false, but some are useful". One should not ask whether a model captures all the relevant features of the real world. What matters is whether the model captures *enough* important features to make it useful, and there is now substantial evidence that coalescent approaches do indeed capture enough of the structure of empirical data to make them extremely valuable. The two applications we have outlined above, namely, phase inference and detection of recombination hotspots, are good examples of this.

In summary, an understanding of the evolutionary processes that lead to the empirical data being generated in many studies can be extremely helpful in guiding its interpretation. More quantitatively, the explicit use of evolutionary models, in particular, the coalescent and approximations to it, has been shown to lead to statistical approaches that are often much more powerful than existing approaches. This may not be surprising for studies that focus directly on the effects shaping the data. But it remains true even when these are not of primary interest. Phase inference is one example. Another context of considerable current interest is disease mapping. Coalescent methods have been successfully developed for fine mapping. It seems likely that they will soon be available, and offer improvements in power, for analyses of genome-wide association studies, or in the choice of tagging SNPs.

References

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA and Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, **36**, 700–706.

Donnelly P and Tavaré S (1995) Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, **29**, 401–421.

Fearnhead P and Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Fearnhead P, Harding RM, Schneider JA, Myers S and Donnelly P (2004) Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, **167**, 2067–2081.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR and Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Nordborg M (2001) Coalescent theory. In *Handbook of Statistical Genetics*, Balding DJ, Bishop M and Cannings C (Eds.), John Wiley & Sons: England, pp. 179–212.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES and Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics*, **32**, 135–142.

Sabeti P, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Stephens M (2001) Inference under the coalescent. In *Handbook of Statistical Genetics*, Balding DJ, Bishop M and Cannings C (Eds.), John Wiley & Sons: England, pp. 213–238.

Stephens M and Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Smith N and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Short Specialist Review

Creating LD maps of the genome

Andrew Collins and Sarah Ennis
University of Southampton, Southampton, UK

1. Linkage and linkage disequilibrium maps

Linkage mapping has been tremendously successful in the localization of many genes involved in severe genetic disorders (major genes). The localization of a disease-influencing polymorphism is achieved by analyzing families to track the coinheritance of disease phenotype and marker polymorphisms. This enables polymorphisms that are linked (found on the same familial haplotype as the disease gene) to be identified. Once linkage is established, various approaches are employed to fine map the candidate gene region (which may span several megabases) until the causal gene is itself identified. Through this process, known as "positional cloning", disease genes are localized without having prior knowledge of their function. Linkage disequilibrium mapping exploits allelic association in a similar way but has the outstanding advantage of significantly higher resolution.

Recombination events do not occur uniformly across the chromosome, which results in some chromosome regions having intense recombination (hot spots) and other regions having low recombination (cold spots). In contrast to physical maps, which give locations for markers in megabases (Mb) or kilobases (kb), genetic linkage maps are expressed in centimorgans (cM). Although a good rule of thumb is that 1 cM is equal to 1 Mb, patterns of recombination are such that recombination hot spots show a much higher cM/Mb ratio and recombination cold spots show the converse. Understanding the variation in recombination rates is vital for disease mapping. Making the assumption that a small genetic distance implies close proximity of a disease gene can be very misleading. In the Hemochromatosis gene region, the cM/Mb ratio is only 0.16, indicating a recombination cold part of the genome. Positional cloning of this gene was delayed because applying the 1 cM$\sim$1 Mb rule suggested that the disease gene was within $\sim$750 kb of a marker, whereas in actual fact, the gene was 4.6 Mb away (Lonjou *et al*., 1998).

Although linkage maps are essential for multipoint mapping of major genes, they are of relatively low resolution because rather small numbers of meioses are accumulated in the families studied. A linkage disequilibrium map, however, reflects the historical accumulation of recombination events over many generations. LD maps therefore offer much higher resolution, suitable for mapping genes for common conditions, such as asthma and heart disease. The relationship between LD patterns and recombination is not a simple one, however, because of the confounding effects of evolutionary processes such as drift, mutation, and selection.

2. Linkage disequilibrium maps – theory and application

The pattern of LD across genomic regions shows domains or blocks of elevated LD interrupted by discrete regions of LD breakdown. These blocks correspond to extended regions of diminished haplotype diversity (Daly *et al*., 2001), while the areas of LD breakdown are coincident with narrow zones of elevated recombination as revealed by direct sperm-typing analysis (Jeffreys *et al*., 2001). To represent these variations, LD maps track fluctuating levels of LD across a genomic segment and represent these data in an additive (or metric) manner when plotted against physical distance in kilobases. An approach to constructing such maps adopts the rho (ρ) metric for association (Collins *et al*., 1999), between pairs of single nucleotide polymorphisms (SNPs), and models the decline of association with distance by the "Malecot" equation: $\rho = \mathrm{L} + (1 - \mathrm{M})\mathrm{Me}^{-\varepsilon \mathrm{d}}$. This equation has the same form as that described by Malecot for isolation by distance, but has different parameters (Morton, 2002). The population genetics theory, which provides the justification for the use of this equation to describe LD, is given by Morton *et al*. (2001). Each of the parameters has a biological interpretation, which allows comparisons of populations or chromosome regions; L accounts for the bias introduced by residual association at large distance; M reflects association at zero distance with values close to one indicating monophyletic origin and values less than one indicating polyphyletic inheritance; ε reflects the exponential decline of association with distance, dominated by recombination (Collins and Morton, 1998). The inverse of ε known as the "swept radius" is a more intuitive interpretation of ε as the distance at which LD has declined to approximately one-third of its magnitude at zero distance. Optimality of the ρ metric for pairwise LD has been demonstrated (Morton *et al*., 2001). The ρ metric was shown to be the most robust to variations in allele frequency when compared with a wide range of alternative pairwise measures of LD. Estimates are also consistent, whether calculated using phase-unknown genotypes (diplotypes) or from haplotype data.

Maniatis *et al*. (2002) applied the ρ metric to create LD maps from pairwise association data by estimating ε for each interval between adjacent SNPs in a high-density map. This is achieved by fitting the Malecot model for all informative pairs that contain the interval. The model weights pairwise data by a function of the distance between each pair so that pairs of SNPs in close proximity contribute the most information and pairs at large distance are correspondingly down-weighted. The product of ε for the ith interval and the width of the interval in kilobases ($\varepsilon_i d_i$) corresponds to a distance in linkage disequilibrium units (LDU). Summation of LDUs over successive intervals in a region generates map locations, which, when plotted against the kb scale, delimit regions of extensive LD and LD breakdown. LDU map distances are known to be additive where there is sufficient SNP coverage of a region (Ke *et al*., 2004). Where the LD pattern is fully characterized, the characteristic "block/step" pattern is revealed, with steeper steps denoting more intense recombination and low LD and the plateaus identifying regions with negligible recombination and extensive LD (Figure 1). A useful property of the maps arises because one LDU is equal to one swept radius. One swept radius defines, in effect, a discrete segment on the chromosome within which LD exists above "background" levels. The width of these units (in kilobases) varies

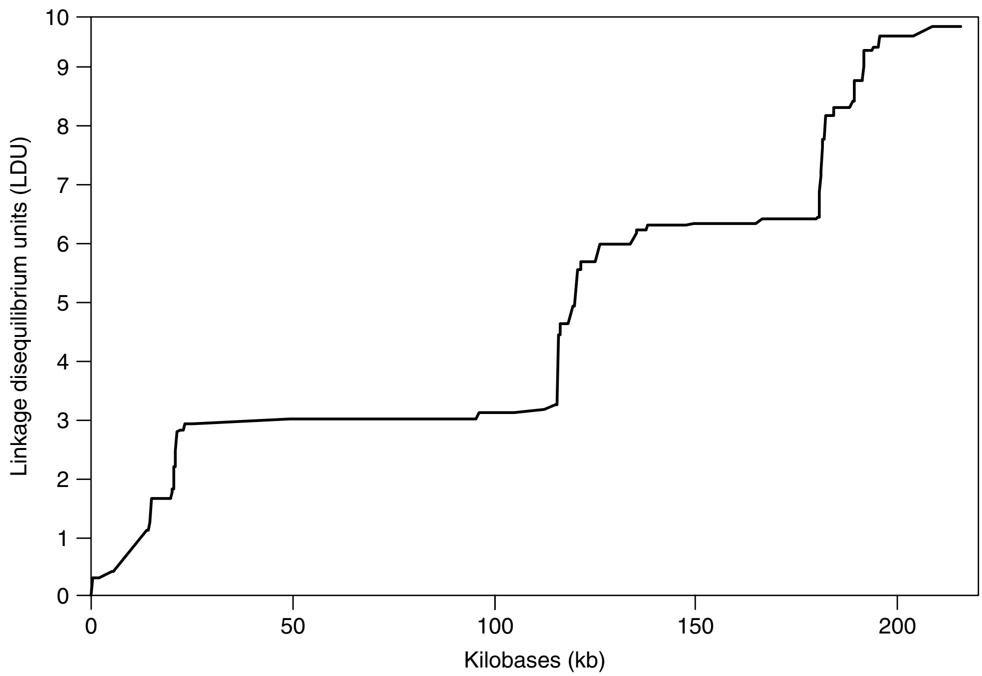


Figure 1 LD map showing characteristic "block/step" pattern of the class II HLA region

extensively across the genome, reflecting the extreme variations in the extent of LD, which is characterized by the LD map. Therefore, effective screening of a candidate region for disease polymorphisms requires even spacing of SNPs on the LDU scale with every LD unit covered. The relationship between power for disease mapping and the numbers and frequencies of SNPs required in each LDU has not yet been established.

There is a high correlation between LDU and cM maps (Figure 2), suggesting that recombination is a major determinant of LD patterns (Tapper *et al*., 2003). The pattern of recombination in different populations is of considerable interest. Most populations (excluding sub-Saharan Africans) are thought to be derived from a relatively small number of founders who moved out of Africa approximately 100 000 years ago. The effect of such population bottlenecks is to create LD, which gradually decays as recombination events are accumulated through successive generations. However, further population bottlenecks (such as those created by famine, war, disease, and migration) will recreate LD in the same way, and so there may be a partly cumulative effect of bottlenecks. Therefore, the observed LD patterns of today reflect a complex interaction between population history, recombination, and other processes. Sub-Saharan African populations, not subjected to the major out of Africa bottleneck, have an LD structure, which reflects more accumulated recombination events. However, because the majority of recombination is concentrated in narrow hot spot regions, which may be between 1 and 2 kb (Jeffreys *et al*., 2000), LD maps differ in length but not pattern (Figure 3, De La Vega

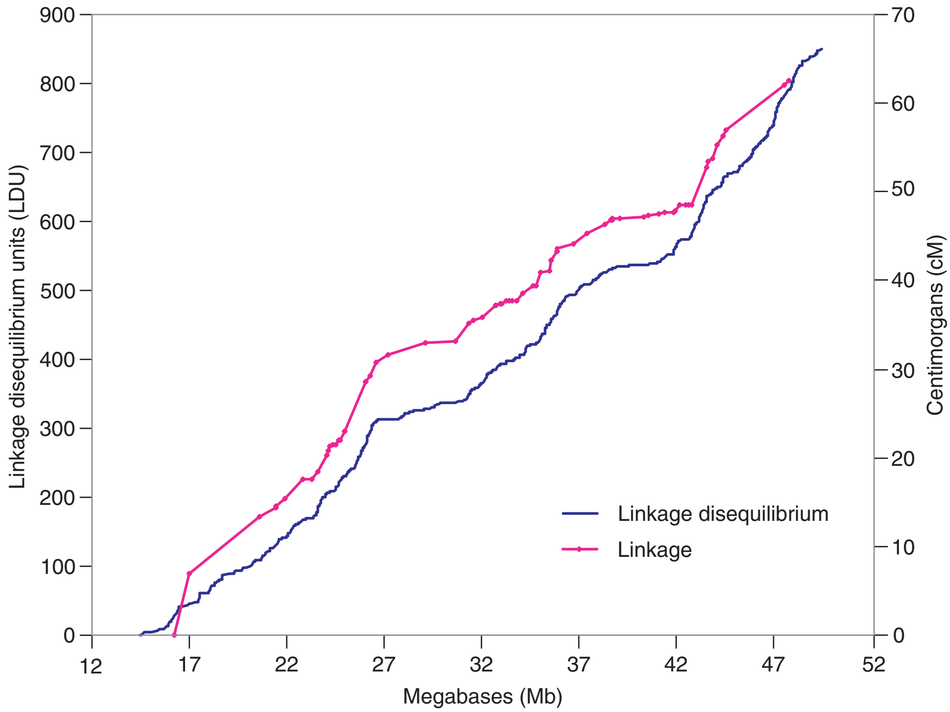


Figure 2 Comparison of the (sex averaged, sequence-based) genetic linkage map to the LD map of chromosome 22 (based on Caucasian samples)

et al., 2003). The increased length of LD maps from populations with African ancestry, reflecting the longer duration, is a consequence of longer steps, reflecting more accumulated recombination events, which are mostly concentrated in hot spots. For these older populations, additional marker typing is required to achieve coverage of all of the LD units in the map. Lonjou *et al*. (2003) found that the contours of different maps are so highly correlated that they suggest the idea of a cosmopolitan map, which may be (linearly) scaled by population-specific "scaling factors". The authors showed that scaling of a map built on data combined across ethnic groups recaptured 95% of the information contained in individual population-specific maps. There is, therefore, potential to develop a unified cosmopolitan LD map of the genome that is useful for all populations irrespective of their underlying haplotype structure (which is known to be very diverse, Jeffreys *et al*., 2000).

3. Linkage disequilibrium maps and positional cloning

The abundance of SNPs and the density at which they may be typed across the genome proffers huge potential for LD mapping of genes for common diseases. Recent estimates predict up to 15 million SNPs in the human genome (Botstein and Risch, 2003). Only a very small proportion of these are likely to have a role in

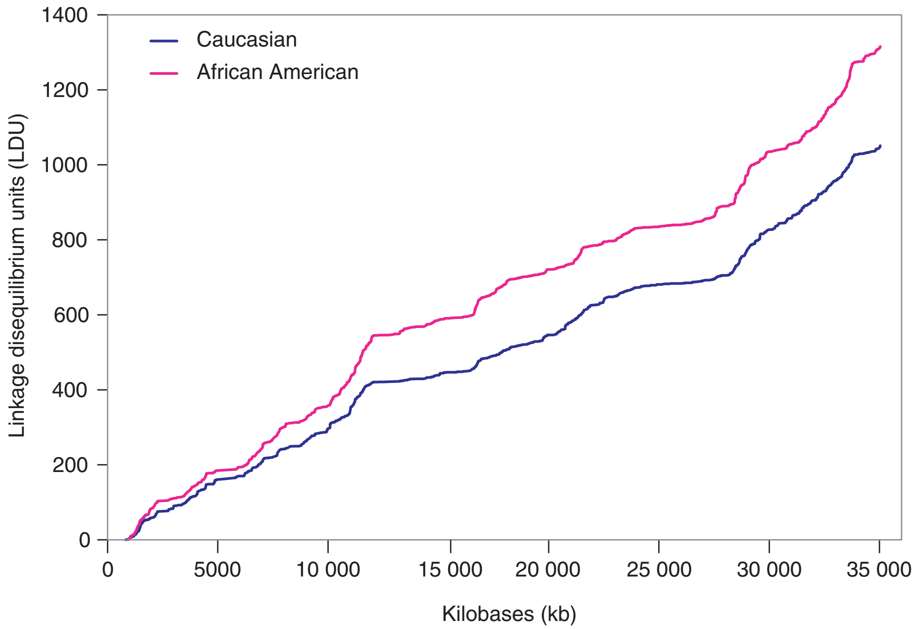


Figure 3 Comparative patterns of linkage disequilibrium across chromosome 22q for Caucasian and African American populations

common human disease. For this reason, LD mapping of disease genes relies on the analysis of multiple SNP markers in the hope of finding a disease association within a narrow genomic region. Simulation studies offer a route to examine the power of these analyses under different scenarios. Maniatis *et al*. (2004) examined the efficiency of LD mapping in a simulation study using a multiple pairwise analysis by the Malecot model, modified to predict the location of a disease-causing variant. The greatest power is achieved when the disease-causing variant is localized on an underlying LDU map, rather than a kb map. The superiority of the LDU scale is particularly evident when the candidate region shows intense recombination hot spots. The next step, which is likely to offer greater power, is to exploit haplotype (rather than multiple pairwise) analysis for localization of a disease polymorphism on an LDU map.

4. Conclusions

There is an increasing body of evidence to suggest that LD patterns are dominated by recombination to the extent that a whole-genome LD analog of the linkage map is useful. One possible route is through the construction of a scale with additive LD units. The colocalization of recombination hot spots in different populations points towards a cosmopolitan "standard" map applicable to all populations, given suitable scaling factors. Simulation studies suggest that there is higher power for multilocus gene localization on the LDU scale. As a minimum, a metric LD map of the genome

will compliment ongoing efforts to represent human haplotype diversity. Strategies that exploit both LD maps and haplotype analysis have the potential to localize genes involved in common diseases with high power.

Related articles

Article 11, **Mapping complex disease phenotypes**, Volume 3; Article 15, **Linkage mapping**, Volume 3; Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3; Article 24, **The Human Genome Project**, Volume 3; Article 67, **History of genetic mapping**, Volume 4; Article 71, **SNPs and human history**, Volume 4

References

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl 3), 228–237.

Collins A, Lonjou C and Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 15173–15177.

Collins A and Morton NE (1998) Mapping a disease locus by allelic association. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1741–1745.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

De La Vega FM, Avi-Itzhak H, Halldorsson B, Scafe C, Istrail S, Gilbert DA and Spier EG (2003) Distribution, sharing, and ancestry of common haplotypes in African-American, Caucasian, and Asian populations. *American Journal of Human Genetics*, **73**, 215.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**, 217–222.

Jeffreys AJ, Ritchie A and Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human Molecular Genetics*, **9**, 725–733.

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, *et al*. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, **13**(6), 577–588.

Lonjou C, Collins A, Ajioka R, Jorde L, Kushner J and Morton NE (1998) Allelic association under map error and recombinational heterogeneity: A tale of two sites. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 11366–11370.

Lonjou C, Zhang W, Collins A, Tapper W, Elahi E, Maniatis N and Morton NE (2003) Linkage disequilibrium in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6069–6074.

Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S and Morton NE (2002) The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2228–2233.

Maniatis N, Collins A, Gibson J, Zhang W, Tapper W and Morton NE (2004) Positional cloning by linkage disequilibrium. *American Journal of Human Genetics*, **74**, 846–855.

Morton NE (2002) Applications and extensions of Malecot's work in human genetics. In *Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malecot*, Montgomery S and Michel V (Ed.), Oxford University Press Inc: New York, pp. 20–33.

Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y and Collins A (2001) The optimal measure of allelic association. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5217–5221.

Tapper W, Maniatis N, Morton NE and Collins A (2003) A metric linkage disequilibrium map of a human chromosome. *Annals of Human Genetics*, **67**, 487–494.

Short Specialist Review

Finding and using haplotype blocks in candidate gene association studies

Daniel O. Stram
University of Southern California, Los Angeles, CA, USA

1. Introduction

It is well understood that there are limitations to the use of traditional family-based linkage analysis in locating disease genes influencing the risk of complex diseases where no highly penetrant variant with a clearly Mendelian pattern of familial inheritance is evident (Risch and Merikangas, 1996). For complex diseases, large numbers of subjects with disease are required for the statistical analysis of risk factors, whether genetic or not. Particularly for diseases with late age-at-onset, it may be difficult to develop family pedigrees with enough cases to perform traditional linkage analysis. Therefore, studies of either completely unrelated subjects or of parent–offspring trios are increasingly becoming important epidemiological designs for genetic analysis. For the purpose of gene hunting, whether among candidate genes or (as is becoming increasingly possible) whole-genome scans, these studies rely upon linkage disequilibrium (LD) in order to identify the genes or regions involved in disease susceptibility. Key to these studies are the current large-scale investments being made in the mapping of single-nucleotide polymorphisms (SNP) haplotype structure of, initially, particular sets of genes (Cambien *et al*., 1999; Reich *et al*., 2001; Bonnen *et al*., 2002; Crawford *et al*., 2004), and now the entire human genome in the International Haplotype Map (HapMap) Project (The International HapMap Consortium, 2003; *see also* Article 67, **History of genetic mapping**, Volume 4).

Underlying the use of LD-based genetic association studies are four primary assumptions

1. that an important fraction of susceptibility to a given disease may be explained by relatively modest effects of a small number of relatively common variants;
2. that many of these common variants either are SNPs themselves or at least may be thought to arise or mutate at rates no greater than the rate that new SNPs appear;
3. that rates at which SNPs or other variants arise and become common are slow enough that many or most of the carriers today of a given common variant inherited this variant from a single ancestor;

4. and finally that recombination rates are low enough that carriers of a common disease-related variant will also tend to carry a pattern of SNPs that reflects the SNPs that appeared on the ancestral chromosomal locus near to the causal variant.

There are varying degrees of evidence regarding the truth of each one of these assumptions and particular controversy regarding assumption 1, which is often termed the common disease–common variant (CDCV) hypothesis (Terwilliger and Weiss, 1998; Weiss and Clark, 2002). These assumptions plus the recognition that there are millions of common SNPs throughout the human genome, provide a motivation for the study of SNPs in association with human disease (Risch and Merikangas, 1996; Botstein and Risch, 2003). Assumption 4 implies that it is not necessary to genotype the actual causal variant in an association study, since genotyping markers that also fell on the original ancestral locus nearby to the causal variant may pick up some part of the signal of the true causal variant. Once a signal is detected (and found to be replicable in later studies), then further genotyping and/or deep sequencing of high-risk DNA, as well as functional analysis of all variants found, is required to follow-up the association.

Recent discoveries (Daly *et al*., 2001; Jeffreys *et al*., 2001; Gabriel *et al*., 2002) that there is considerable heterogeneity in the amount of LD between SNP markers over the genome and the implication that recombination rates themselves may also be considerably heterogeneous (McVean *et al*., 2004) not only adds to the motivation for conducting association-based testing but also suggests specific analysis strategies and affects the details of the design of such studies. In the remainder of this article, we focus on the implications on association studies of the existence of well-defined "haplotype blocks" covering much of the genome over which there is little evidence for recombination and within which there is limited haplotype diversity (*see* Article 73, **Creating LD maps of the genome**, Volume 4.

2. Haplotype block definition and haplotype diversity

In the manner of Gabriel *et al*. (2002), haplotype blocks are regions in which the standard measure of pairwise linkage disequilibrium (D′) is consistent (or nearly consistent) with no recombination ($|D'| = 1$) for all (or nearly all) pairs of markers in the region, based upon approximate confidence intervals for the statistic. An emerging picture of the human genome is that most of the SNPs in the genome can be defined as falling within haplotype blocks and that between haplotype blocks there are shorter regions in which there is a relatively high level of recombination. Gabriel *et al*. (2002) made an empirical observation that within blocks the number of haplotypes was typically quite limited, with 4–6 haplotypes per block making up approximately 80% or more of all chromosomes. In particular, as more and more common SNPs are genotyped within a block, these will tend to be confined to existing haplotypes rather than splitting existing haplotypes into ever more complex entities. Thus, the number of observed haplotypes (4–6) is expected to be much smaller than the theoretical maximum of $(n + 1)$ nonrecombinant haplotypes among n SNPs. This empirical observation is completely consistent with population

genetics in that, absent recombination, only a fraction of haplotypes arising as new mutations occur are expected to withstand random drift toward extinction. The very high-density SNP discovery and genotyping of 10 regions each of approximately 500 kb in extent (the "ENCODE" regions) as a part of the HapMap project further reinforces the general observations of Gabriel *et al*., with dozens of nearby SNPs often falling upon very few haplotypes.

3. Implications of limited haplotype diversity for association studies

Under assumption 2 above (that common disease-causing variants "behave" like SNPs), it follows that a common causal variant occurring within a haplotype block will fall upon one or more of the common haplotypes, so that a method of predicting common haplotypes and assessing each one (and/or each combination) of them for risk should be sensitive to the presence of an unmeasured causal variant. In its idealized form, each block is separated by regions of high recombination between haplotypes so that it is possible to estimate the haplotype-specific risk associated with haplotypes in each block as if they are independent between blocks. This approach has formed the basis of the haplotype-specific risk estimation methods used by several groups (Haiman *et al*., 2003; Karamohamed *et al*., 2003; Prince *et al*., 2003; Sai *et al*., 2003), which seek to detect and localize to a particular block the signal of an unmeasured variant. In the following, we outline the following three steps in performing these association studies.

1. SNP and Haplotype discovery
2. Picking of haplotype-tagging SNPs (ht SNPs) for large-scale genotyping (e.g., in a case-control study)
3. Testing and estimation of haplotype-specific risk in a large-scale case–control study.

4. SNP and haplotype discovery

In the initial stages of an association study, a large number of SNPs are identified and genotyped in a "haplotype discovery panel" comprising a relatively small number of subjects. Some groups (Carlson *et al*., 2003; Chapman *et al*., 2003) have relied upon deep resequencing of DNA for the subjects that make up the haplotype discovery panel. Other studies rely upon the use of public or private databases to determine which SNPs may be available. Even with the public availability of large numbers of SNPs (nearly 7 million being deposited in dbSNP **http://www.ncbi.nlm.nih.gov/SNP/** as of this writing), there has, up to this point, been a considerable amount of work required in assembling a suitably dense panel of SNP markers for which assays can be developed for any specific genotyping platform and which are frequent enough so as to be useful in the pursuit of common causal variants, and intensive efforts to discover SNPs by deep resequencing of candidate genes for various diseases are currently underway (Crawford *et al*., 2004).

Currently, several important projects (Olden and Wilson, 2000; The International HapMap Consortium, 2003), the largest of which is the International HapMap Project, are considerably reducing the difficulty of determining which SNPs are suitably polymorphic, and are providing a dense network of SNPs for haplotype discovery.

Haplotype discovery consists of genotyping the selected markers in the DNA for the subjects that comprise the haplotype discovery panel and then assessing haplotype structure. Two key issues are how large the haplotype discovery panel (how many subjects) should be and how dense a network of SNPs is needed to be fully informative about haplotype structure. Up to the advent of the HapMap, this was a problem that was dealt with on a study-by-study or group-by-group basis. It is expected that the DNA genotyped by the HapMap (from 45 to 90 individuals depending upon ethnic group) will essentially play the role of the haplotype discovery panel used by most association studies in the future, with an SNP density now expected to vary between 1 SNP every 2.5–3 kb throughout the entire genome.

5. Haplotype block structure and ht SNP selection

The most popular criteria for defining haplotype blocks appears to be the Gabriel *et al*. (2002) method, which bases its block definition upon upper and lower confidence limits for $|D'|$. Other approaches toward block definition either rely upon slightly different LD measurements (e.g., the 4-gamete test) to define regions of limited recombination, or focus instead directly upon finding regions of limited haplotype diversity (Patil *et al*., 2001; Zhang *et al*., 2002), irrespective of recombination. A helpful review of these procedures is provided by Wall and Pritchard (2003a).

A number of criteria have been chosen for ht SNP selection, these may broadly be broken into methods that seek to optimize SNP prediction (Chapman *et al*., 2003; Meng *et al*., 2003; Weale *et al*., 2003; Carlson *et al*., 2004) versus prediction of haplotypes (Ke and Cardon, 2003; Sebastiani *et al*., 2003; Stram *et al*., 2003a). The first group of methods focuses upon the statistical prediction of the specific SNPs that were genotyped in the haplotype discovery panel. That is, they seek a set of ht SNPs that when genotyped can be used to predict the remaining ungenotyped SNPs. These methods are most appropriate when the candidate loci in question have been fully resequenced during SNP discovery in a large enough number of subjects so that the variants found may be thought to constitute a complete census of all common mutations in these genes. Efforts to provide such a census of all polymorphisms, for genes considered most important in specific diseases is continuing (Crawford *et al*., 2004); however, this is a daunting task in general. When the SNPs found in the discovery phase cannot be considered to be an entire census of all variation, it may be more relevant to concentrate on the prediction of common haplotypes, with the presumption that common unmeasured causal variants must fall upon common haplotypes, although statistical assessment of haplotype prediction versus SNP prediction has not yet been completely explored.

Within haplotype blocks, it is relatively easy to compute the predictability either of SNPs or of common haplotypes, based upon a given set of SNPs, which will

be chosen to be measured as ht SNPs in a latter association study. A standard approach is provided by Stram *et al*. (2003a). This consists of (1) the estimation of haplotype frequencies within the block and (2) the computation of the squared correlation between measured ht SNPs genotypes and haplotypes or non–ht SNPs. In both cases, the best predictor of the number of copies $\delta(H) = 0$, 1, or 2 of a given haplotype or unmeasured SNP is computed as the conditional expected value $E\{\delta(H)|G_{ht}\}$ of $\delta\ (H)$ given any ht SNP genotype, based upon the haplotype frequency estimates. The squared correlation between the true and predicted value of $\delta(H)$ is then computed by enumeration over all possible ht SNP genotypes.

This calculation forms the basis of the ht SNP selection methods of Stram *et al*. (2003a). SNPs are selected as ht SNPs depending upon the degree to which their inclusion increases this R^2 measure. A program (tagSNPs) that maximizes the minimum R^2 for either the common haplotypes or for unmeasured SNPs using a modified forward selection algorithm is available (**http://www-rcf.usc.edu/~stram/tagSNPs.html**).

The value of any ht SNP selection method is dependent upon there being restricted haplotype diversity in the region of interest. The block definitions of Gabriel *et al*. (2002) provide boundaries of regions wherein it is expected that haplotype diversity is indeed quite restricted. Outside the regions of restricted haplotype diversity, ht SNP selection methods will fail to adequately capture the variability of SNPs and haplotypes. The decision of whether to include SNPs that lie within regions of low LD depends on whether they are of special interest in and of themselves (e.g., missense SNPs) or if they are a part of a complete census of common variation (e.g., from deep resequencing studies) of the locus of interest. Of course, as the HapMap and other projects reach very high densities of SNP genotyping, it will be increasingly reasonable to regard the SNPs that have been characterized as representing, if not a complete census, then a least a considerable fraction of all genomic variation that may be functionally related to disease.

6. Testing of haplotype-specific risk

We now consider the statistical testing and estimation of haplotype-specific risks of disease for each common haplotype within each haplotype block, with our interest focused upon the use of the case–control study data. The methods described here are based on those described in Schaid *et al*. (2002), Zaykin *et al*. (2002), and Stram *et al*. (2003b). After genotyping ht SNPs for the case–control study subjects, the ht SNP genotype data for the cases and controls are combined to estimate ht SNP haplotype frequencies using the same EM (or PLEM) algorithm that was used in the discovery panel. Upon convergence of the EM algorithm, the estimated haplotype type counts $E\{\delta_h(H)|G_{ht}\}$ are computed for each individual i as

$$E\{\delta_h(H)|G_{ht,i}) = \frac{\sum\limits_{H\sim G_{ht,i}} \delta_h(H) p_{h_1} p_{h_2}}{\sum\limits_{H\sim G_{ht,i}} p_{h_1} p_{h_2}} \qquad (1)$$

where $H \sim G_{ht,\mathrm{i}}$ refers to the ht SNP haplotype pairs (h_1, h_2), which are compatible with the ht SNP genotype data $G_{ht,\mathrm{i}}$ for the individual in question. These estimated haplotype counts are computed for each individual and each common haplotype, and these are merged with the outcome data (disease status) and other covariates (age, sex, etc.), which may be required in the analysis. Finally, logistic regression analysis is used to analyze the data with the $E\{\delta_h(H)|G_{ht}\}$ treated as if they are equivalent to the true haplotype counts in the analysis. In particular, Zaykin *et al.* (2002) describe the formation of unbiased score and likelihood ratio tests for the null hypothesis that the haplotypes are not associated with increased risk of disease against a log-additive alternative. These tests are available in any standard logistic regression program, and these programs allow considerable flexibility in the inclusion of other covariates (age, ethnicity, known risk factors, etc.) into the analysis.

The score or likelihood ratio tests mentioned above are closely related to the optimal score statistics described from likelihood theory for generalized linear models by Schaid *et al.* (2002), and as such they represent efficient tests of the null hypothesis that there is no association between haplotypes and disease status. They can be conducted as single degree of freedom tests in which the per-copy risk associated with carrying one copy of a given haplotype is compared to carrying zero copies of the same haplotype (so that the baseline category becomes "all other haplotypes"), or can be used to test the global null hypothesis that no haplotypes are related to risk, with the most common haplotype generally nominated to serve as the baseline category.

7. Estimation of haplotype type–specific risk

In the above, we have concentrated upon the testing of null hypotheses (of no haplotype-specific risk) rather than on estimation of risk. For testing purposes, the analyses given in Zaykin *et al.* (2002) indicate that we are justified in ignoring the uncertainty of haplotype type count estimation (see also Tosteson and Ware, 1990; Stram and Kopecky, 2003, for similar results in other settings). On the other hand, estimation of haplotype-specific risks is affected by haplotype count uncertainty. Such uncertainty affects confidence interval estimation, comparisons of nonzero haplotype-specific risks between two or more haplotypes, and evaluation of the significance of gene–environment interactions. These biases arise for two basic reasons (Stram *et al.*, 2003b), the first is that the haplotype frequency estimates, p_h, from the EM algorithm, used to calculate the haplotype count variables in equation (1) are subject to sampling variation, which has been ignored, and second, that case–control selection implies that high-risk haplotypes are overrepresented in case-control studies leading to biased estimates of the haplotype frequencies used in that calculation. Some headway has been made on dealing with these two issues directly (Epstein and Satten, 2003; Stram *et al.*, 2003; Zhao *et al.*, 2003). However, for large case–control studies using a block-based approach, these biases are generally quite minor. Limited recombination between the SNPs within a block implies little inherent uncertainty in ht SNP haplotype prediction, and the large sample sizes needed in the case–control study minimizes sampling variability in the haplotype frequency estimates for the ht SNPs. Only when very

high haplotype-specific risks are present, does overselection of high-risk haplotypes lead to perceptible bias in either haplotype risk or confidence limit estimation. This observation also implies that the logistic regression methods may be applied to other problems as well, such as the testing and estimation of haplotype by environment or haplotype by haplotype interactions (Kraft *et al*., 2004), greatly enhancing the attractiveness of the logistic regression approach over methods which simply seek to quantify differences in haplotype frequency between cases and controls.

8. Other issues

Other important practical considerations must be encountered in every analysis. These include:

1. What to do with uncommon haplotypes?
2. What if a log-additive model is not of primary interest?
3. What should be done about SNPs that do not fall within blocks?

8.1. Uncommon haplotypes

Approaches used with uncommon haplotypes have included the pooling of all the uncommon to form a separate category, which is then treated as a single "haplotype" in the logistic regression or probably more reasonably, to pool a given uncommon haplotype with the common haplotype that is most similar to it (Thomas *et al*., 2003). Within a haplotype block the uncommon haplotypes (those of frequency less than 5–10%) are expected to make up no more than about 10–20% of all the haplotypes, so that this is less of a problem within blocks than without (see below).

8.2. Other penetrance models

The efficient score test for other penetrance models than the log additive can be computed in the same manner as described above, except that the expected haplotype count variables are replaced with expected indicator variables, for example to test a purely dominant effect of a particular haplotype one would modify equation (1) to compute

$$E\{I_h^D(H)|G_{ht,i}) = \frac{\displaystyle\sum_{H\sim G_{ht,i}} I(\delta_h(H) > 0)\, p_{h_1} p_{h_2}}{\displaystyle\sum_{H\sim G_{ht,i}} p_{h_1} p_{h_2}} \qquad (2)$$

for each member of the case–control study, with this variable then merged with the case–control status and used as before. More generally, codominant models may be tested for and estimated by computing the expectations of indicator variables for $\{\delta_h(H) = 1\}$ and $\{\delta_h(H) = 2\}$ separately, and again merged with the

case–control data. As with the log-linear model, tests of the null hypothesis of no haplotype-specific risks remain valid despite uncertainty in estimation of the indicator functions, and within haplotype blocks the precision of estimation of these functions is very high.

8.3. Models for quantitative phenotypes

The same general approaches outlined above can be applied towards quantitative outcomes so that, for example, standard linear regression software can be used to analyze the relationship between haplotypes and mean response of measured outcomes. The same principles apply, that is, both univariate and global tests of any relationship between haplotypes and outcomes are unbiased under the null hypothesis, and within regions of restricted recombination, confidence intervals for either haplotype, main effects, or their interactions with other variables are generally quite reliable.

8.4. Multiple testing

In studies in which a large number of haplotype blocks are being evaluated (e.g., in a complex gene or large set of genes), the nominal p-value for any given block-specific test of the relationship between haplotypes and disease may be a very poor reflection of the overall statistical significance of the study as a whole. When haplotypes within blocks are independent, then a simple Bonferroni correction of each of the block-specific p-values (i.e., multiplying the global p-value within each block by the number of blocks considered in the analysis) is a simple and often quite accurate approach toward correcting for multiple testing. The extent of recombination between haplotypes in nearby blocks varies widely; however, in many cases, haplotypes may be very highly correlated across neighboring blocks. In this, the Bonferroni test may be overly conservative since fewer genuinely independent tests are really being performed.

One attractive approach for dealing with this issue is permutation testing. In this procedure, the permutation distribution of the smallest block-specific p-value is developed by randomly switching case-control status a large number of times for each individual in the study (keeping the total number of cases and controls the same in each switch). The observed smallest p-value is then compared to the permutation distribution and if it falls within the extreme tail of the distribution then the result is considered significant, that is, if the observed value is smaller than α percent of the permuted values, then the test is significant at the $1-\alpha$ level of confidence. This procedure is complicated when the logistic regression is to be corrected for the influence of other covariates in the analysis (since case–control status should be permuted only between otherwise similar subjects) and can be time consuming for large studies. An alternative is simply to merge blocks that are very highly correlated with each other prior to the Bonferroni test or even to put all haplotypes in all blocks into a single model and perform a single global test.

9. The extent of coverage of haplotype blocks

Looked upon on a gene-by-gene or region-by-region basis, it is clear that there is considerable local variation in haplotype block structure (Wall and Pritchard, 2003a). Candidate regions in which little or no LD is seen between nearby markers are not, currently, good prospects for SNP or SNP-haplotype-based association testing. Unless a very complete survey of SNPs is available for such regions, there is little hope that the SNPs selected will capture signals from causal variants. Even when such a survey is available, the costs of genotyping every common SNP in candidate gene regions where there is little LD may be prohibitive – although genotyping costs are now dropping rapidly.

The most important issues in the implementation of haplotype block-based association studies are those regarding the number, size, and coverage of blocks over typical candidate gene regions or the genome as a whole. Recent surveys (Wall and Pritchard, 2003b; Crawford *et al*., 2004) have indicated that apparent haplotype block structure becomes more complicated as the density of SNPs increases; specifically, as the density increases, the average length of haplotype blocks decreases, because regions that initially did not appear to have been in blocks are found to show high levels of LD over short ranges as more SNPs are genotyped.

Until now, it was difficult to judge the fraction of all SNPs that are contained within haplotype blocks over the entire human genome, or the number of ht SNPs that would be required for well powered association-based genome-wide scans. However, with the advent of the HapMap project, the resources for just such a survey are rapidly becoming available, first for populations of European origin and soon for other ethnic groups. It now seems likely that considerably more ht SNPs will be needed for most studies than it appeared when the first reports of the existence of haplotype block structure were published. Advances in high-throughput genotyping imply, however, that even if very dense networks of ht SNPs are required for candidate gene or genome-wide association-based studies, this will not be the ultimate limiting factor determining their feasibility. As a detailed haplotype structure is developed by the HapMap and assuming that genotyping costs continue to drop rapidly, it may soon be possible to consider an optimized whole-genome-wide scan in which ht SNPs are selected within blocks supplemented by all known common SNPs in low LD regions. This would allow a hybrid approach towards case-control analysis, exploiting haplotype structure over most of the genome while performing individual tests on the remaining low LD SNPs. Very large case–control studies will be needed in order to exploit this approach while controlling adequately for false-positive associations.

References

Bonnen PE, Wang PJ, Kimmel M, Chakraborty R and Nelson DL (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Research*, **12**(12), 1846–1853.

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(Suppl), 228–237.

Cambien F, Poirier O, Nicaud V, Herrmann SM, Mallet C, Ricard S, Behague I, Hallet V, Blanc H, Loukaci V, *et al*. (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. *American Journal of Human Genetics*, **65**(1), 183–191.

Carlson C, Eberle M, Reider M, Smith J, Kruglyak L and Nickerson D (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, **33**(4), 518–521.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L and Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, **74**(1), 106–120.

Chapman JM, Cooper JD, Todd JA and Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity*, **56**, 18–32.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L and Nickerson DA (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *American Journal of Human Genetics*, **74**(4), 610–622.

Daly MJ, Rioux J, Schaffner S, Hudson T and Lander E (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–232.

Epstein MP and Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*, **73**(6), 1316–1329.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**(5576), 2225–2229.

Haiman CA, Stram DO, Pike MC, Kolonel LN, Burtt NP, Altshuler D, Hirschhorn J and Henderson BE (2003) A comprehensive Haplotype analysis of CYP19 and breast cancer risk: the multiethnic cohort study. *Human Molecular Genetics*, **12**(20), 2679–2692.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**(2), 217–222.

Karamohamed S, Demissie S, Volcjak J, Liu C, Heard-Costa N, Liu J, Shoemaker CM, Panhuysen CI Meigs JB Wilson P, *et al*. (2003) Polymorphisms in the insulin-degrading enzyme gene are associated with type 2 diabetes in men from the NHLBI Framingham Heart Study. *Diabetes*, **52**(6), 1562–1567.

Kraft P, Cox DG, Paynter RA, Hunter D and De Vivol I (2004) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genetic Epidemiology*, in press.

Ke X and Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**(2), 287–288.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR and Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**(5670), 581–584.

Meng Z, Zaykin DV, Xu CF, Wagner M and Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *American Journal of Human Genetics*, **73**(1), 115–130.

Olden K and Wilson S (2000) Environmental health and genomics: visions and implications. *Nature Reviews Genetics*, **1**(2), 149–153.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**(5547), 1719–1723.

Prince JA, Feuk L, Gu HF, Johansson B, Gatz M, Blennow K and Brookes AJ (2003) Genetic variation in a haplotype block spanning IDE influences Alzheimer disease. *Human Mutation*, **22**(5), 363–371.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**(6834), 199–204.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Sai K, Kaniwa N, Itoda M, Saito Y, Hasegawa R, Komamura K, Ueno K, Kamakura S, Kitakaze M, Shirao K, *et al*. (2003) Haplotype analysis of ABCB1/MDR1 blocks in a Japanese population reveals genotype-dependent renal clearance of irinotecan. *Pharmacogenetics*, **13**(12), 741–757.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM and Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, **70**(2), 425–434.

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS and Ramoni MF (2003) Minimal haplotype tagging. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(17), 9900–9905.

Stram D, Haiman C, Hirschhorn JN, Altshuler D, Kolonel L, Henderson B and Pike M (2003a) Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human Heredity*, **55**(27–36).

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE and Thomas DC (2003b) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity*, **55**, 179–190.

Stram D and Kopecky K (2003) Power and uncertainty analysis of epidemiological studies of radiation-related disease risk where dose estimates are based upon a complex dosimetry system; some observations. *Radiation Research*, **160**, 408–417.

Terwilliger JD and Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology*, **9**(6), 578–594.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Thomas DC, Stram DO, Conti D, Molitor J and Marjoram P (2003) Bayesian spatial modeling of haplotype associations. *Human Heredity*, **56**(1–3), 32–40.

Tosteson T and Ware J (1990) Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika*, **77**, 11–21.

Wall JD and Pritchard JK (2003a) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**(8), 587–597.

Wall JD and Pritchard JK (2003b) Assessing the performance of the haplotype block model of linkage disequilibrium. *American Journal of Human Genetics*, **73**(3), 502–515.

Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW and Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *American Journal of Human Genetics*, **73**(3), 551–565.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, **18**(1), 19–24.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ and Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, **53**(2), 79–91.

Zhang K, Deng M, Chen T, Waterman MS and Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(11), 7335–7339.

Zhao LP, Li SS and Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*, **72**(5), 1231–1250.

Short Specialist Review

Avoiding stratification in association studies

Bernie Devlin
University of Pittsburgh, Pittsburgh, PA, USA

Kathryn Roeder
Carnegie Mellon University, Pittsburgh, PA, USA

When hunting for genetic variants (*see* Article 68, **Normal DNA sequence variations in humans**, Volume 4) affecting liability to complex disease, ideal study designs are those amenable to recruitment of large samples, thereby increasing power, and amenable to seamless modeling of both environmental and genetic factors, thereby capturing as many of the sources of variability as possible (*see* Article 69, **Reliability and utility of single nucleotide polymorphisms for genetic association studies**, Volume 4). One solution is provided by association studies (Risch and Merikangas, 1996; Risch, 2000). Association analyses can be conducted by using either family-based samples or population-based samples. The simplest type of population-based association study involves genotyping SNPs in a sample of cases and controls. For each SNP, the strength of the association is assessed using a chi-squared test. In a random mating population, the only SNPs associated with the disease will be those that are tightly linked with a liability allele (LA).

Human populations, however, often exhibit substructure because of nonrandom mating, such as that based on proximity or culture (*see* Article 71, **SNPs and human history**, Volume 4). This substructure, or stratification, creates concern about the validity of association tests (Knowler *et al*., 1988; Lander and Schork, 1994). The problem arises when (1) the allele frequencies vary across subpopulations making up the study population; (2) the prevalence of the disease varies across subpopulations; and (3) these two factors coincide to create associations between phenotypes and genotypes that are not linked to LAs. For example, imagine a worst-case scenario in which affected individuals were selected from subpopulation A and unaffected individuals from subpopulation B; any alleles at higher frequency in A than in B will be spuriously associated with affection status; the greater the differential between A's and B's allele frequencies, the greater the association will appear.

By using the transmission of alleles from parents to offspring, the family-based approach completely removes concerns about population substructure. Since Spielman *et al*.'s (1993) seminal paper, and Ewens and Spielman's (1995) more

rigorous demonstration of Transmission Disequilibrium Test or TDT's robustness to substructure, the number of family-based tests has grown tremendously (Lange and Laird, 2002a,b). A substantial advantage of the family-based design is the robustness to population structure. A substantial disadvantage is that the design can hinder recruitment of large samples. For late-onset diseases, such as the Alzheimer disease and cardio-vascular disease, it is very challenging to collect family-based samples, whereas for childhood-onset diseases such as autism, it is much more natural.

In an effort to discover LAs with small effect, large, population-based samples and large-scale genotyping are being used to evaluate disease/gene associations. To use these study designs, it is desirable to account for the effect of population stratification, thereby avoiding evaluation of many spurious associations.

The effect of stratification is to inflate the test statistic by a multiplicative factor, λ, which is typically greater than 1 if population substructure, admixture, or cryptic relatedness are present in the sample. The magnitude of the effect can be expressed as a linear function of heterogeneity in allele frequencies and disease prevalences among subpopulations and the sample size (Devlin and Roeder, 1999). For instance, in a poorly designed study including equal fractions of two, moderately divergent populations, for which the disease prevalence is twofold higher in one ethnic group, the test is essentially on target for samples of size 100, but 50% inflated for samples of size 1000. Notice that this effect could be removed either by matching cases and controls by ethnic origin or by using a stratified method of analysis (but that assumes the individual populations are themselves homogeneous). However, if the ethnicity was not recorded, these options are not available.

Until recently, few definitive examples of studies in which population substructure has led to spurious association have been identified (Knowler *et al*., 1988; Thomas and Witte, 2002). This may be due, in part, to the limited size of most samples to date. Stratification can have a marked effect on the false-positive rate of tests that are not corrected for this feature in large samples, even if small sample investigations of the same populations show little evidence of structure (Devlin and Roeder, 1999; Devlin *et al*., 2001a). Because association studies are increasingly being performed on much larger samples, the effect of stratification is likely to become more evident (Freedman *et al*., 2004; Marchini *et al*., 2004).

On the basis of the observation that population substructure can cause spurious associations between phenotype and alleles at an unlinked locus, Pritchard and Rosenberg (1999) suggested evaluating a large number of loci unlinked to the candidate gene of interest, which we might call "null loci", to determine if there is evidence of association, indicating substructure a priori. However, because all human populations are stratified to some degree, association will be detected, almost surely, as the sample size or the number of loci tested increases. As in many situations in practice, it will be more fruitful to estimate the size of an effect than to test for its presence. Building on results from evolutionary theory (e.g., Wright (1969); Lewontin and Krakauer (1973)), Devlin and Roeder (1999) demonstrated that the effects of stratification on test statistics of interest are essentially constant across the genome, under certain conditions. On the basis of this finding, they suggested using null loci located across the genome to estimate the effect of stratification and then removing the effect from the association test statistic.

Currently, two methods build on the idea of using the genome to correct for stratification in the population: genomic control (GC) (Devlin and Roeder, 1999; Bacanu *et al*., 2000, 2002; Reich and Goldstein, 2001) and structured association (SA) (Pritchard *et al*., 2000a,b; Satten *et al*., 2001; Falush *et al*., 2003; Hoggart *et al*., 2003). The GC approach exploits the fact that population substructure inflates statistics used to assess association. By testing multiple polymorphisms throughout the genome, only some of which are pertinent to the disease of interest, the degree of overdispersion generated by stratification can be estimated and taken into account. The SA approach assumes that the sampled population, while heterogeneous, is composed of subpopulations that are themselves homogeneous, or is an admixture of such subpopulations. By using multiple null polymorphisms throughout the genome, this "latent class method" estimates the probability that the sampled individuals derive from each of these latent subpopulations. It is then possible to condition on subpopulation and remove the effect of stratification.

The GC approach is exceptionally simple to implement: compute the desired test statistic for each candidate marker and each null marker; then estimate the magnitude of the inflation factor λ and divide the candidate test statistics by this quantity before computing p-values. This general approach can be applied to single-locus tests (Devlin and Roeder, 1999), multilocus linear models (Bacanu *et al*., 2002), tests based on haplotypes (Tzeng *et al*., 2003a,b), and tests based on pooled genotyping (Devlin *et al*., 2001b).

A limitation of the GC approach is that it assumes that the effect of stratification is approximately constant over all loci. This assumption is generally supported both theoretically and empirically (Bacanu *et al*., 2000). It is violated only when the loci under study are under strong selective pressure (Robertson, 1975). Specifically, it is required that the locus under investigation is not under differential selective pressure in different subpopulations. Examples of genes under such selection include those coding for pigmentation, malarial resistance, and lactose tolerance. Notice that for some genes this assumption may be violated in some populations, say African Americans, but not for others, such as Northern Europeans. In addition, the assumption is required only for candidate loci. Null loci can violate this assumption without increasing the false-positive rate.

Rather than correcting for population substructure by estimating the effect of stratification, the general idea behind the simplest SA model is to cluster the samples into groups with similar genetic ancestry and then to condition upon this ancestry covariate. For Pritchard *et al*.'s (2000a,b) method, a Bayesian clustering program is run to determine both the number K of subpopulations within the population and the membership probability vector for each individual. Model choice for K is performed by running a Markov chain separately at different values of K and an approximate method is used to estimate posterior probabilities for each value of K. The program STRUCTURE performs this task (Pritchard *et al*., 2000b). As output, one obtains a vector $q = q_1, \ldots q_K$ that indicates an estimate of the fraction of the individual's genome that originated from each of the K subpopulations.

Given an estimated membership vector for each of the subjects in the study, the next step is to compute a likelihood ratio test based upon computing the likelihood of the data under the null and alternative hypotheses (Pritchard *et al*., 2000a). Under the null, it is assumed that the genotype of the candidate gene is independent of

the case/control status of the subject. Under the alternative, the model is quite flexible, allowing any genotype to be associated with case status; the model does not necessarily restrict the association to be constant across subpopulations. In this second phase of the analysis, both K and the membership vectors are treated as known quantities. To assess significance, a smoothed bootstrap is recommended. The program STRAT uses the output from STRUCTURE as input in conducting a test for association.

STRUCTURE can also be run holding K fixed; however, if the wrong value is chosen, the results can be quite misleading. For instance, if all the individuals are from a single population but K is set at 2, then the program will estimate that all the people are about a 50% mixture of each subpopulation. As a consequence, the power of the association analysis can be greatly diminished. Alternatively, if K is too small, the size of the test will be inflated.

The simplest SA model described above can be extended to allow for admixture. In this case, rather than presuming that each individual is a member of one of the subpopulations making up the population, the admixture model allows each person to have ancestry from multiple subpopulations. Now, the membership vector estimates the fraction of their chromosomal material from each source. In practice, this model is quite similar to one developed for admixture mapping (*see* Article 76, **Mapping by admixture linkage disequilibrium (MALD)**, Volume 4) (McKeigue *et al*., 2000; Hoggart *et al*., 2003). Indeed, some improvements over the 2-stage process of modeling employed by Pritchard and colleagues have been achieved in competing implementations (Satten *et al*., 2001; Zhu *et al*., 2002; Hoggart *et al*., 2003; Zhang *et al*., 2003; Chen *et al*., 2003).

More recently, the admixture models have been extended to allow for correlation due to "admixture linkage", which is defined as the spatial correlation across markers due to ancestry from a common subpopulation. With this additional feature, the methods can handle markers that are fairly densely spaced along the chromosome, but not so dense that the marker alleles are highly dependent at the population level (*see* Article 73, **Creating LD maps of the genome**, Volume 4).

Even with decreasing genotype costs, inclusion of null markers comes at a substantial cost, especially for large studies. Although it is unclear how many null markers are required in any given study, whether the analysis will be conducted using SA or GC (Bacanu *et al*., 2000; Pritchard and Donnelly, 2001; Marchini *et al*., 2004), it is worth considering how to minimize the number of null markers required. In some situations, SA can benefit from the use of markers known to differ substantially between subpopulations (see Satten *et al*., 2001; Hoggart *et al*., 2003). For instance, loci conferring malarial resistance or markers in tight linkage disequilibrium with these loci often have allele frequencies that differ substantially between Africans and Europeans. Thus, when studying an African-American population, such loci should be more informative than randomly selected markers. When studying populations composed of widely divergent ethnic groups originating from different continental groups, a targeted approach can greatly reduce the number of required markers to achieve separation by perhaps an order of magnitude. The Haplotype Map project, which is producing data from a huge number of SNPs in three geographically distinct population samples, will surely provide many SNPs with allele frequencies that differ substantially among these

populations. The usefulness of such attempts, however, is less apparent when the population under investigation has subtle effects of structure. For instance, there are no proven methods for preselecting informative null loci within a Caucasian population. Rosenberg *et al*. (2003) do provide calculations on the informativeness of various marker types.

In contrast, markers under targeted selection are not appropriate for GC, which performs best with the use of representative null markers. However, it is permissible to utilize both candidate and null markers to estimate λ with this approach. This is permissible for the following reasons. Even if several of the candidate genes impart a signal, the signal typically does not extend past a narrow region of linkage disequilibrium. Thus, many markers in the candidate genes will yield a null signal. This information is helpful in estimating the effect of stratification. Provided λ is estimated using a robust estimator, the bias arising from including candidate genes is minor. Consequently, in a study investigating many candidate genes, implementing GC can involve a very minor additional expense; see the original Devlin and Roeder (1999) paper for a Bayesian analytic approach for such data.

Although similar in conception, GC and SA have different strengths and weaknesses. The SA approach is more ambitious, attempting to infer the genetic ancestry of each individual sampled, and may require more null loci to ensure a successful outcome (Pritchard and Donnelly, 2001). When too few null loci are used, K may be underestimated, which can cause the method to be anticonservative. Assuming that the population structure can be reconstructed, however, this approach has the capacity to model different effects in different subpopulations. This extra flexibility leads either to a gain in power or a loss, depending upon whether it is needed to model the data appropriately. Certainly, it is biologically plausible that effects could vary across subpopulations. It is difficult to predict it a priori.

In a sense, the GC approach is more general because it corrects for confounding due to cryptic relatedness as well as population substructure and admixture. The SA approach assumes that the observations are independent, conditional upon the inferred genetic associations due to population substructure and admixture. Consequently, the SA approach is not valid for isolated or inbred populations, which tend to have substantial amounts of cryptic relatedness.

In most published simulation studies, the average rejection rates for GC were close to the nominal values (Bacanu *et al*., 2000; Pritchard and Donnelly, 2001; Devlin *et al*., 2001b). The exception is Marchini *et al*. (2004), who explored the performance of GC over a bigger range of conditions, including the extreme tail of the distribution ($<10^{-4}$). They found that GC does not provide a perfect correction in the extreme tails of the distribution. This feature occurs largely because GC treats λ as a known constant (Reich and Goldstein, 2001) and because the variability in the estimate of λ matters for small α. They also considered studies with a high level of population heterogeneity and greatly varying disease prevalence, which are not taken into account in the experimental design or analysis. These are conditions under which the GC method is not recommended, because it will likely have low power. On the other hand, it is simple to calculate correct tail probabilities if a researcher wishes to use extremely small critical values (Devlin *et al*., 2004).

Owing to the high computational demands, the SA method has not been thoroughly investigated, especially in the tail of the distribution and for large sample sizes. Thus far, it appears to be well calibrated, provided sufficient null loci and/or subjects have been included from which to determine K. The method has an inflated false-positive rate when K is underestimated. For instance, in studies including only individuals of European descent, the method typically yields $K = 1$, in which case no correction is provided for the subtle levels of stratification in the sample.

Regarding power, the best choice of analysis for any particular dataset depends primarily on the level of expected population heterogeneity. With a moderate amount of substructure, both methods can outperform TDT in that the power, per person recruited, is generally greater for both the GC and SA methods than it is for the TDT (Bacanu *et al*., 2000; Pritchard *et al*., 2000a; Pritchard and Donnelly, 2001). The notable exception occurs in admixed populations where the TDT receives an extra boost in power (Ewens and Spielman, 1995). Limited simulations have been performed by Pritchard and Donnelly (2001), comparing SA and GC. From these simulations, it appears that GC is more powerful than SA when the allelic effect is nearly constant across subpopulations; however, if the effect varies strongly across subpopulations, particularly if a different allele is associated with the disease, SA is more powerful than GC.

The SA method is well designed to capture the larger scale heterogeneity found in admixed populations, such as African Americans, or in studies where the subjects are not carefully screened for ethnicity. It is in these heterogeneous settings that the flexible parameterization of allelic effects is likely to pay off. For instance, a marker allele may be in LD with the causal polymorphism in the European subpopulation but not in LD in the African subpopulation. Without the flexible parameterization typically employed in the SA method, the effect can be muted.

In carefully designed studies, with fairly homogeneous populations, however, a large number of null loci will be required before SA can detect unrecorded subpopulations. In addition, in these populations, it is difficult to construct biologically plausible motivation for different allelic effects across subpopulations. Consequently, SA may not be the method of choice in this type of setting.

In contrast, GC was designed to be applied to studies in which careful efforts have been made to obtain a fairly homogeneous sample (Devlin and Roeder, 1999; Bacanu *et al*., 2000). When applying GC, the power of the analysis will be enhanced if efforts are made to match the cases and controls for ethnicity (Lee, 2004), thus further removing the effect of heterogeneity. In application, GC adjusts for the remaining effects of stratification that cannot be removed by careful planning of the study. While the GC approach is approximately valid in highly heterogeneous populations, in this situation, the power of the test may be low because the inflation factor can be large.

Owing to the plummeting costs of SNP genotyping (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4) and the availability of nearly a million validated SNPs from the Haplotype Map project, it seems reasonable to predict that whole-genome association (WGA) scans for LAs are in our future. In fact, to our knowledge, a handful of WGA scans have been performed or are in progress. Consider the following scenario. A research team wishes to perform a WGA scan in which hundreds of thousands of markers are to be genotyped on a

sample of mixed European ancestry, with the sample consisting of a 1000 affected and a 1000 unaffected individuals. How should the researchers design/implement the study? Hinds *et al*. (2004) provide one elegant approach. They propose to use several hundred SNPs, which have been selected because they show subpopulation variation in their allele frequencies, to match case- and control-individuals for genetic ancestry (using, e.g., SA or other clustering algorithms). This first-phase analysis would be performed on more than 2000 case- and control-individuals to optimize matching for the sample to be used for WGA. Now, perform the massive genotyping on the matched sample and analyze the data using standard methods for matched designs. As also shown by Lee (2004), Hinds *et al*. (2004) find that power is greatly enhanced over analyses ignoring population structure.

Is there a way to use the logic of GC to account for any residual impact of structure after WGA using matched samples? Lee (2004) shows how the standard approach to GC achieves this goal for a single locus. For experiments involving a large number of tests of genetic markers, such as WGA, one should analyze the entire distribution of test statistics (Devlin and Roeder, 1999; Tzeng *et al*., 2003a). In addition to the Bayesian methods developed in Devlin and Roeder (1999), we are currently developing new GC methods for this setting.

Further reading

Risch N and Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human disease. I. DNA pooling. *Genome Research*, **8**, 1273–1288.

Wacholder S, Rothman N and Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *Journal of the National Cancer Institute*, **92**, 1151–1158.

References

Bacanu S-A, Devlin B and Roeder K (2000) The power of genomic control. *American Journal of Human Genetics*, **66**, 1933–1944.

Bacanu S-A, Devlin B and Roeder K (2002) Association studies for quantitative traits in structured populations. *Genetic Epidemiology*, **22**, 78–93.

Chen HS, Zhu X, Zhao H and Zhang S (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Annals of Human Genetics*, **67**, 250–264.

Devlin B, Bacanu S-A and Roeder K (2004) Genomic control to the extreme. *Nature Genetics*, **36**, 1129–1130.

Devlin B and Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Devlin B, Roeder K and Bacanu S-A (2001b) Unbiased methods for population-based association studies. *Genetic Epidemiology*, **21**, 273–284.

Devlin B, Roeder K and Wasserman L (2001a) Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, **60**, 155–166.

Ewens WJ and Spielman RS (1995) The transmission/disequilibrium test: history, subdivision and admixture. *American Journal of Human Genetics*, **57**, 455–464.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, *et al*. (2004) Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**, 388–393.

Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR and Ballinger DG (2004) Matching strategies for genetic association studies in structured populations. *American Journal of Human Genetics*, **74**, 317–325.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG and McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics*, **72**, 1492–1504.

Knowler WC, Williams RC, Pettitt DJ and Steinberg AG (1988) Gm3-5,13,14 and type-2 diabetes-mellitus-an association in American-Indians with genetic admixture. *American Journal of Human Genetics*, **43**, 520–526.

Lander ES and Schork N (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.

Lange C and Laird NM (2002a) On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genetic Epidemiology*, **23**, 165–180.

Lange C and Laird NM (2002b) Power calculations for a general class of family-based association tests: dichotomous traits. *American Journal of Human Genetics*, **71**, 575–584.

Lee W-C (2004) Case-control association studies with matching and genomic controlling. *Genetic Epidemiology*, **27**, 1–13.

Lewontin RC and Krakauer J (1973) Distribution of gene frequencies as a test of the theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Marchini J, Cardon LR, Phillips MS and Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512–517.

McKeigue PM, Carpenter JR, Parra EJ and Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics*, **64**, 171–186.

Pritchard JK and Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*, **60**, 227–237.

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, **65**, 220–228.

Pritchard JK, Stephens M and Donnelly P (2000b) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P (2000a) Association mapping in structured populations. *American Journal of Human Genetics*, **67**, 170–181.

Reich DE and Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, **20**, 4–16.

Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science*, **255**, 1516–1517.

Robertson A (1975) Gene frequency distribution as a test of selective neutrality. *Genetics*, **81**, 775–785.

Rosenberg NA, Li LM, Ward R and Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402–1422.

Satten GA, Flanders WD and Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, **68**, 466–477.

Spielman RS, McGinnis RE and Ewens WJ (1993) Transmission test for linkage disequilibrium: The insulin gene region andinsulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, **52**, 506–516.

Thomas DC and Witte JS (2002) Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiology, Biomarkers & Prevention*, **11**, 505–512.

Tzeng JY, Devlin B, Wasserman L and Roeder K (2003a) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics*, **72**, 891–902.

Tzeng J-Y, Wasserman L, Byerley W, Devlin B and Roeder K (2003b) Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the American Statistical Association*, **98**, 236–247.

Wright S (1969) *Evolution and the Genetics of Populations*, *Vol 2: The Theory of Gene Frequencies*, University of Chicago Press: Chicago.

Zhang S, Zhu X and Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology*, **24**, 44–56.

Zhu X, Zhang S, Zhao H and Cooper RS (2002) Association mapping, using a mixture model for complex traits. *Genetic Epidemiology*, **23**, 181–196.

Short Specialist Review

Mapping by admixture linkage disequilibrium (MALD)

Michael William Smith
SAIC-Frederick, National Cancer Institute at Frederick, Frederick, MD, USA

Differentiation of the peoples of the world over the past 100 000 years (*see* Article 6, **The genetic structure of human pathogens**, Volume 1) coupled with prehistoric and more recent mass migrations in the past few hundred years have resulted in different peoples coexisting, intermarrying, and generating populations of mixed origins (Figure 1). Genetically speaking, this is termed *Admixture*, reflecting the more and more common occurrence of individuals from previously separated and differentiated populations producing an intermingled group of offspring that are then a population of mixed ancestry. Since we know that most (85–90%) of the genetic variation in humans is shared across continental groups (Lewontin, 1972; Rosenberg *et al*., 2002), the remainder provides the basis of differentiation between groups and a nonrandom basis to the chromosomes and chromosomal segments in these admixed populations.

During the times since the major continental groups populations were isolated from one another, genetic differentiation naturally occurred through the processes of genetic drift, bottlenecks, and selection. Upon subsequent contact and the concomitant genetic mixing, admixture generates long-range linkage disequilibrium initially on the order of tens of centimorgans due to long stretches of chromosomes that were derived from either contributing population, and was originally described as mapping by admixture linkage disequilibrium (MALD) (Briscoe *et al*., 1994; Chakraborty and Weiss, 1988; Stephens *et al*., 1994; McKeigue 1997). The chromosomal segments from each parental population become smaller and smaller as a function of time since admixture began because of recombination, although with continuous gene flow large unbroken segments are continually added from the parental populations (Figure 2).

Naming has proven controversial, with one group suggesting Admixture Mapping as being more descriptive (Halder and Shriver, 2003; Hoggart *et al*., 2004), yet continuity of the literature and historical precedence supports the original MALD as the general term for the gene discovery method. The MALD name recognizes that the chromosomal structure of ancestral segments present due to admixture linkage disequilibrium (Figure 3) provides a basis for mapping genes with differential risk in ethnic/racial populations, even as the technique evolves and improves on an ongoing basis (Hoggart *et al*., 2004; Patterson *et al*., 2004;

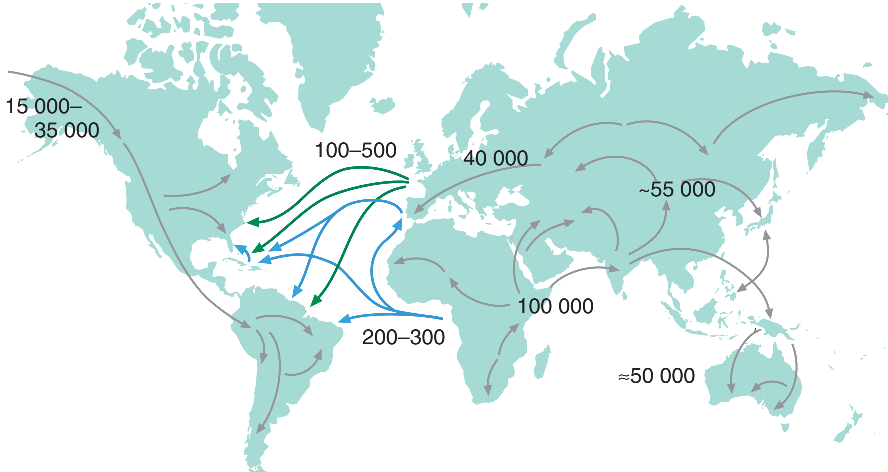


Figure 1 Migrations and diasporas of people that has led to admixed populations primarily in the Americas. (Adapted from several sources: Parra *et al*., 1998; Wiencke, 2004; Cavalli-Sforza and Feldman, 2003)

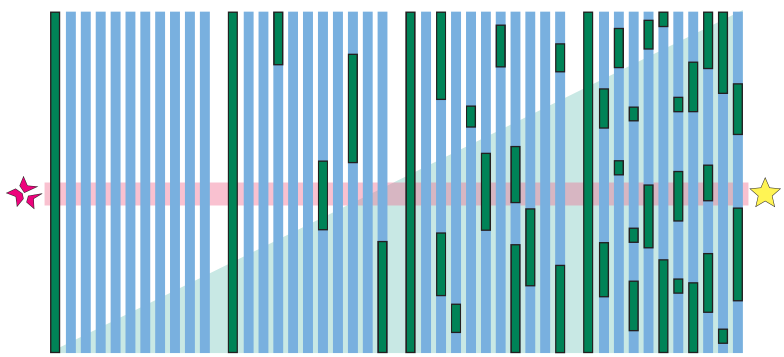


Figure 2 Schematic of chromosomes in the Admixed population with ancestry segments recombining as admixture takes place on an ongoing basis over time

Smith *et al*., 2004). The MALD technique actually only requires a difference in frequency of the various disease-causing variants and haplotypes (Hoggart *et al*., 2004; Patterson *et al*., 2004). An important analytical innovation is the proposed case-only design that uses the remainder of the genome and some individuals from the admixed population as the controls to discover disease genes utilizing the MALD approach (Hoggart *et al*., 2004; Montana and Pritchard, 2004; Patterson *et al*. 2004). The precision of the MALD gene mapping methodology to a few centimorgans (Patterson *et al*., 2004) falls between tens of cM for family studies (*see* Article 15, **Linkage mapping**, Volume 3) and significantly less than 1 cM for haplotype-map-based experimental designs covered elsewhere (*see* Article 12, **Haplotype mapping**, Volume 3 and Article 17, **Linkage disequilibrium and whole-genome association studies**, Volume 3). The intermediate discovery scale is addressed by the now practical method, MALD, for disease gene discovery

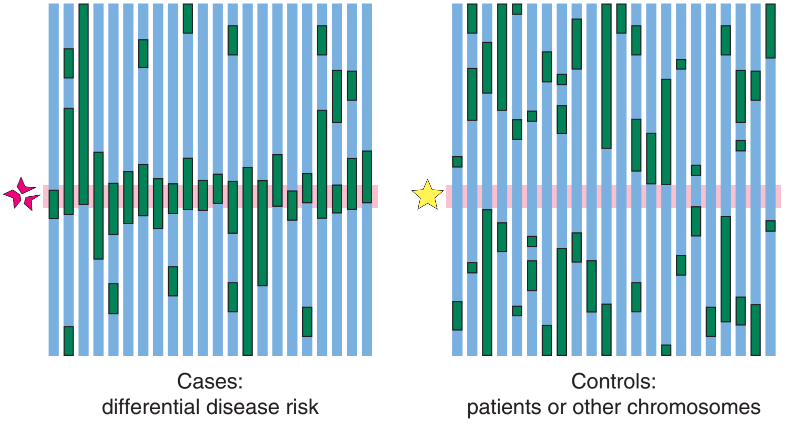


Figure 3 Schematic of experimental design to discover disease genes with differential prevalence in the green contributor to an admixed population

especially in African-Americans, where the necessary dense set of genome spanning markers was recently identified by Smith *et al*. (2004).

The identification of admixture as a novel property of some populations with application to gene identification and localization recognizes the possibility and known genetic mixing of the three major human population groups from Africa, Europe, or Asia (Briscoe *et al*., 1994; Chakraborty and Weiss, 1988; Stephens *et al*., 1994; McKeigue, 1997). Admixed groups include African-Americans, Hispanics, some Pacific Islanders, some South and Central Americans, and other groups of mixed origins (Halder and Shriver, 2003). The genetics of these admixed populations are rather complex as to their origins, timing of mixture, and dynamics have been addressed to a greater or lesser degree (Halder and Shriver, 2003). Utilizing MALD may be possible in all these admixed groups by genotying 10 000 to 15 000 random SNP markers (Montana and Pritchard, 2004), but parental population linkage disequilibrium between markers could prove problematic to this admixture map-less approach. The best understood case is of African-Americans, so this group provides a good starting point for explaining the basis and application of MALD in the real world.

In the African-American population of the United States, the proportion of admixture ranges from about 5 to 25%, depending on the population examined (Parra *et al*., 2001, 1998). In general, admixture is lowest in the southern United States like South Carolina (3–12%) and New Orleans (22%) along with Jamaica (7%), and highest in more northern cities like Pittsburg (25%), and Chicago (16%), Detroit (19%), Philadelphia (14%), New York city (19%) along with western cities like Los Angeles (26%) and Seattle (26%) (Wang, 2003). In parts of central America, a more complicated picture emerges. For instance, in the two populations from Santa Catarina Island in southern Brazil, contributions were estimated as African: 17% and 49%, European: 75% and 45%, and Amerindian 8% and 7% respectively (de Souza *et al*., 2003). Hispanics present a more heterogenous and less-studied picture of admixture with a population in the San Louis Valley of Colorado estimated as 63% European, 34% Native American, and 3% West African

(Bonilla *et al*., 2004). In contrast, Hispanics from California were assessed as 60% European and 40% Amerindian (Collins-Schramm *et al*., 2004). This particular study is consistent with a more general study that grouped Hispanics as Western or Eastern United States with Amerindian contributions of 36 to 58% or 0 to 21%, respectively, reflecting Mexican origins in the west and predominantly Cuban or Puerto Rican contributions in the East (Bertoni *et al*., 2003). Finally, admixture estimates from six Costa Rican populations show a great range in admixture for Europeans from 51 to 66%, Amerindians 27 to 35%, and Amerindians 7 to 14% (Morera *et al*., 2003). Levels of admixture also differ substantially between individuals (Smith *et al*., 2004; Parra *et al*., 2001, 1998). While the admixture levels differ substantially between populations and individuals, in general, the levels of admixture observed that falls between 10 and 90% are ideal for discovering genes that underlie genetic differences in diseases between source populations (Patterson *et al*., 2004).

Disease gene discovery is the reason for using MALD as an investigative tool while fully acknowledging the controversy of investigating the genetics of race and ethnicity (Risch *et al*., 2002; Schwartz, 2001). The merit of disregarding race relies on the fact that the majority of genetic variation is shared between groups, which, for example, predominates in interpretations of drug response (Wilson *et al*., 2001). Some reseachers have taken such simple observations in combination with the history of eugenics to suggest that race is best left alone and unstudied in genetics, while the benefits of examining race for understanding the genetics of disparities between racial groups is acknowledged by others (Risch *et al*., 2002; Schwartz, 2001; *Nature Genetics,* 2000). To that end, a group of diseases has been proposed for MALD-based gene discovery in African-Americans that includes hepatitis C clearance, melanoma, HIV vertical transmission, multiple sclerosis, suicide, HIV progression, lung cancer, stroke, end-stage renal disease, intracranial hemorrhage, focal segmental glomerulosclerosis, prostate cancer, hypertensive heart disease, and myeloma (Smith *et al*., 2004). Utilizing admixture to discover disease genes in admixed populations must be undertaken with caution, but is justified because of the potential benefit to the study subjects and their racial/ethnic group. Many studies are underway to test for genes that underlie these diseases in African-Americans and other groups (Halder and Shriver, 2003).

Discussion of the utilization of MALD to discover genes can be focused on the African-Americans, where the richest genetic resources are available and the nature of the admixture is best characterized. Basically, several requirements must be met to utilize admixture for disease gene discovery. One, the presence of large differences in allele frequencies at causative disease loci between founding populations is required. Since such differences in allele frequency at a disease locus are not knowable in practice, starting with diseases with very different prevalences in the two founding populations has to suffice as a substitute. Two, a set of well-spaced and highly differentiated genetic loci for examination of associations between disease status and chromosomal regions is needed and has been described recently (Smith *et al*., 2004). Three, analysis algorithms that can unravel the genetic complexities of real populations to discover genes must be developed and are becoming better as the ability to collect appropriate data has improved (Hoggart *et al*., 2004; Patterson *et al*., 2004). Finally, massively parallel and multiplexed

genotyping platforms are required to apply these markers to patient samples, and these have come along in their development of technologies for the international haplotype map project (Gibbs *et al*., 2003).

Recently, three papers on MALD were published back-to-back in the American Journal of Human Genetics (Hoggart *et al*., 2004; Patterson, 2004; Smith *et al*., 2004) and the accompanying description of the contents of the journal asked " . . . will it actually be able to find a disease gene?" Discovery of novel regions with disease genes will of course be the true test of the method. However, the validation of the importance of the Human Leukocyte Antigen (HLA) region in multiple sclerosis utilizing MALD analysis provides a promise for its future application (Oksenberg *et al*., 2004). Likewise, the ability to "find" genetic regions like the FY locus initially and many others recently illustrates the power of the technique (Patterson *et al*., 2004; Lautenberger *et al*., 2000). Regardless of these proofs of principle and the availability of a map, the efficacy of MALD will lie in its ability to point the way to discovery of novel disease genes important to human health.

MALD provides an exciting additional gene discovery methodology and tool that can complement both family studies and association approaches. It should be particularly effective in discovering disease with large differences in allele frequencies between racial groups. It should also be a good approach, when an environmental or viral factor is important as a causative agent in the disease process. Because MALD tracks genes through at most tens of generations, rather than the few of most family studies or the hundreds to thousands or more of associated ones, there may very well exist some effects for which MALD gives important insight by discovering genes that cannot be found otherwise. Application of MALD to ongoing disease genetics studies will test the ability of the technique to make robust novel disease genetics discoveries – currently, both the promise and bane of human genetic analysis.

Acknowledgments

I would like to thank Taras Oleksyk, Stephen J. O'Brien, David Reich, Nick Patterson David Altshuler, Josef Coresh, Michael Klag and Michael Dean for discussing important insights into MALD. This publication has been funded with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References

Bertoni B, Budowle B, Sans M, Barton SA and Chakraborty R (2003) Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States. *Human Biology*, **75**, 1–11.

Bonilla C, Parra EJ, Pfaff CL, Dios S, Marshall JA, Hamman RF, Ferrell RE, Hoggart CL, McKeigue PM and Shriven MD (2004) Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. *Annals of Human Genetics*, **68**, 139–153.

Briscoe D, Stephens JC and O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. *Journal of Heredity*, **85**, 59–63.

Cavalli-Sforza LL and Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, **33**(Suppl 3), 266–275.

Chakraborty R and Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 9119–9123.

Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA, Hanson RL, Knowler WC, Silva G and Belmont JW (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Human Genetics*, **114**, 263–271.

de Souza IR, Muniz YC, de M Saldanha G, Alves Junior L, da Rosa FC, Maegawa FA, Susin MF, de S Lipinski M and Petzl-Erler ML (2003) Demographic and genetic structures of two partially isolated communities of Santa Catarina Island, southern Brazil. *Human Biology*, **75**, 241–253.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H-M, Chang LY, Huang W, Liu B and Shen Y (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Halder I and Shriver MD (2003) Measuring and using admixture to study the genetics of complex disease. *Human Genomics*, **1**, 52–62.

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG and McKeigue PM (2004) Design and analysis of admixture mapping studies. *American Journal of Human Genetics*, **74**, 965–978.

Lautenberger JA, Stephens JC, O'Brien SJ and Smith MW (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. *American Journal of Human Genetics*, **66**, 969–978.

Lewontin R (1972) The apportionment of human diversity. *Evolutionary Biology*, **6**, 381–398.

McKeigue PM (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *American Journal of Human Genetics*, **60**, 188–196.

Montana G and Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. *American Journal of Human Genetetics*, **75**, 771–789.

Morera B, Barrantes R and Marin-Rojas R (2003) Gene Admixture in the Costa Rican Population. *Annals of Human Genetics*, **67**, 71–80.

Nature Genetics, (2000) Census, race and science. **24**, 97–98.

Oksenberg JR, Barcellos LF, Cree BA, Baranzini SE, Bugawan TL, Khan O, Lincoln RR, Swerdlin A, Mignot E, Lin L, *et al*. (2004) Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. *American Journal of Human Genetics*, **74**, 160–167.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer M, Cooper R, Forrester T, Allison D, Deka R, Ferrell R, *et al*. (1998) Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, **63**, 1839–1851.

Parra EJ, Kittles R, Argyropoulos G, Pfaff C, Hiester K, Bonilla C, Sylvester N, Parrish-Gause C, Garvey W, Jin L, *et al*. (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *American Journal of Physical Anthropology*, **114**, 18–29.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, *et al*. (2004) Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics*, **74**, 979–1000.

Risch N, Burchard E, Ziv E and Tang H (2002) Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*, **3**, 1–12.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Schwartz RS (2001) Racial profiling in medical research. *The New England Journal of Medicine*, **344**, 1392–1393.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, *et al*. (2004) A high-density admixture map for disease gene discovery in african americans. *American Journal of Human Genetics*, **74**, 1001–1013.

Stephens JC, Briscoe D and O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics*, **55**, 809–824.

Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.

Wiencke JK (2004) Impact of race/ethnicity on molecular pathways in human cancer. *Nature Reviews. Cancer*, **4**, 79–84.

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N and Goldstein DB (2001) Population genetic structure of variable drug response. *Nature Genetics*, **29**, 265–269.

Short Specialist Review

Genotyping technology: the present and the future

Pui-Yan Kwok and Ting-Fung Chan
University of California, San Francisco, CA, USA

1. Introduction

The single nucleotide polymorphism (SNP), the most common form of genetic variation, has become the genetic marker of choice in the mapping of human traits (Schlotterer, 2004). An international consortium has been organized to harness the power of SNP markers and to construct high-density genetic and haplotype maps that will help us understand the genetic factors associated with many common diseases (Gilman *et al*., 2001; The IHC, 2003). Because of the tremendous technical challenge and demand for SNP genotyping, it has recently become the focus of considerable biotechnology entrepreneurship. Over the last decade, many approaches have been developed to reduce reagent and labor costs while improving accuracy, versatility, throughput, and simplicity of operation.

2. The present

Regardless of the principle behind the design, all genotyping methods consist of two stages: sample preparation and analysis. During the sample preparation step, different alleles of a particular SNP can be tagged either by a direct oligonucleotide-based hybridization method or by an enzyme-based method (such as ligation, primer extension, etc.) following probe hybridization. These methods are commonly known as allele-specific labeling (see Figure 1 for examples). As for detection, the product of the labeling reaction is analyzed by a variety of physical methods (such as direct fluorescence, fluorescence resonance energy transfer (FRET), fluorescence polarization (FP), or mass spectrometry, etc.). Many of these technologies have been extensively reviewed elsewhere (Gut, 2001; Syvänen, 2001) and are briefly summarized in Table 1.

3. Limitations

Even with all the genotyping technologies that are currently available, one cannot find a universal method that is suitable for all applications. This is due to the fact

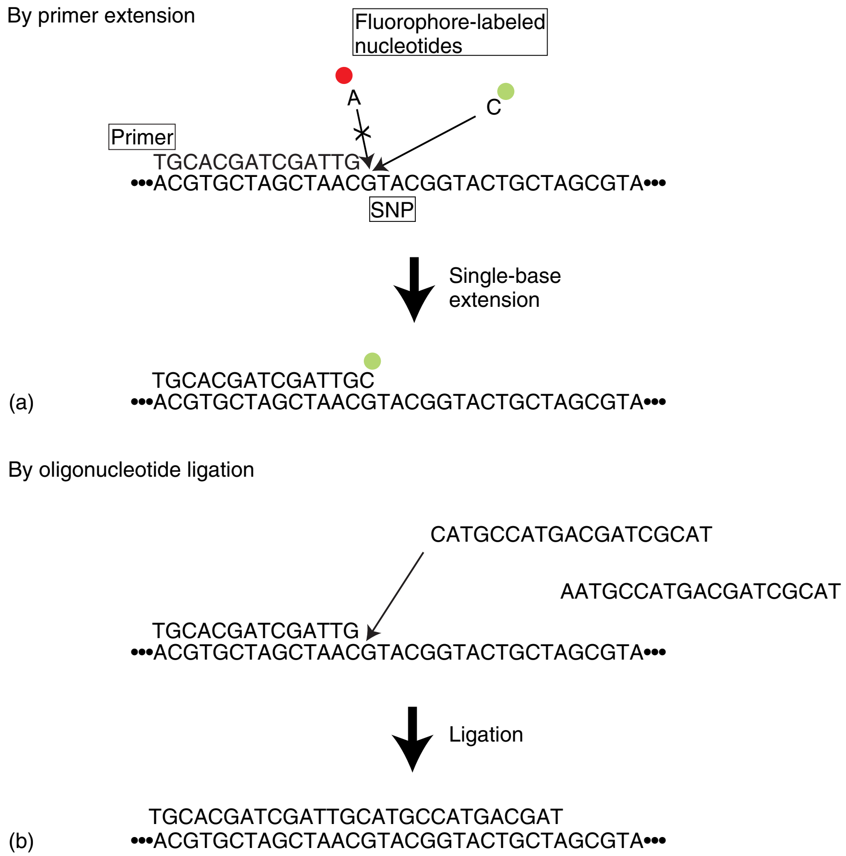


Figure 1 Examples of allele-specific labeling

that none of the current methods are completely versatile: some assays are simple to set up but are impossible to assay many SNPs simultaneously in multiplex reactions; others can be done in a highly multiplex format but require a long assay-development process and expensive, dedicated instruments. In general, the less multiplexed assays are easier to develop and failed assays can be optimized more easily. However, their operation cost is higher per genotype since the reactions are used to assay just one or a handful of SNPs. The highly multiplexed assays are much cheaper to perform once the "master mix" of probes is developed and validated. However, the "master mix" of probes cannot be changed even if it is desirable to do so when new SNPs need to be added to (or redundant SNPs removed from) the mix. Furthermore, any failed assays among the thousands of assays in the multiplex reaction cannot be "redone" since all the assays in multiplex must be assayed all over again. The choice of a technology is therefore heavily dependent on both the cost and the nature of the scientific question asked. Specifically, does a study need to genotype many individuals with a small number of SNPs; many SNPs in a small number of individuals; or many SNPs in many individuals?

Table 1 A selection of current genotyping technologies: examples of methodology

Sample preparation	Sample analysis	Method	Scale vs. flexibility
Hybridization	Fluorescence in array format	GeneChip microarray	High-throughput; low flexibility
	Fluorescence, FRET	Molecular beacons; TaqMan	Multiplex capable but probes are expensive
	Fluorescence	Dynamic allele-specific hybridization (DASH)	Complex probe design rules, high-throughput is possible
	Mass spectrometry	Peptide nucleic acids (PNA)	Multiplex capable but probes are expensive
Allele-specific PCR	Fluorescence, FRET	FRET primers	High-throughput and multiplex capable but primers are expensive; flexible
	FRET	AlphaScreen	High-throughput and multiplex capable but primers are expensive; flexible
Allele-specific primer extension	Chemiluminescence	Pyrosequencing	Difficult to multiplex; flexible
	Fluorescence polarization (FP)	Template-directed incorporation (TDI)	Difficult to multiplex; flexible
	Fluorescence in array format	Illumina coded microsphere	High-throughput; low flexibility
	Mass spectrometry	GOOD assay	Multistep procedure; flexible
Oligonucleotide ligation	Fluorescence in array format	Padlock probe/molecular inversion probe	Multiplex capable but probes are expensive
	Fluorescence	Ligation-rolling circle amplication (L-RCA)	Efficiency not known
Nuclease cleavage	Gel electrophoresis	Restriction site analysis	Not suitable for high throughput
	Mass spectrometry	Invader assay	Requires large amount of DNA, so throughput is questionable

4. The future

The primary concerns for the development of future genotyping technologies are cost, throughput, accuracy, simplicity, and versatility. An added concern is the "SNP content", meaning which sets of SNPs to use for particular studies. Until all the maps and most of the useful SNPs have been found and characterized, it is impossible to nail down a set of SNPs to develop assays around for all applications. In this article, we will therefore concentrate on the more technical issues.

With the emergence of nanotechnology and its interwoven relationships with biology, miniaturization is widely considered to be the next step in the evolution of genotyping technologies (Service, 2002). Microfluidics and microelectronics are the two main themes currently driving the progress to cheaper and faster assaying methods.

Microfluidics in biological assaying techniques can reduce cost significantly by reducing reaction volume, and hence the amount of reagents and enzymes needed for each reaction. Different steps of an assay can be performed in small-enough reaction volumes and be integrated into the so-called lab-on-a-chip format

to increase throughput. In principle, these lab-chip systems can be designed to be fully automated and operate in parallel, thereby reducing both reagent and labor costs, minimizing human errors at the same time (Liu *et al*., 2004; Mitchell, 2001). Microfluidics, however, is not without limitations. The number of genomic DNA copies in these small reaction volumes may be so low that allelic loss may result. Fluid flow obstructed by air bubbles or particulate matter can be a problem. Nonetheless, these problems did not hamper researchers' attempt to place SNP genotyping assay onto a microfluidic system (Russom *et al*., 2003). In time, this technology may mature to a point where routine operation is possible.

Applying microelectronic processes to genetic analysis is not an entirely recent concept. For example, photolithography using expensive chromium masks to fabricate oligonucleotide microarrays for high-throughput genotyping (Barone *et al*., 2001) is being replaced by a digital masking process (Nuwaysir *et al*., 2002), resulting in a faster and cheaper manufacturing process. Electric field microarray is another example of merging microelectronics with genotyping application (Gilles *et al*., 1999). Each microarray contains detection wells that are electronically programmable and can be configured and reconfigured for a particular SNP. An electric field is applied to concentrate the target DNA to the site where the probe is so that the time required for the hybridization process is greatly accelerated. Electronic stringency can also be adjusted at each detection well, which makes multiplexing possible as each SNP locus may require a different hybridization condition.

Microelectronic biosensors have also been developed for genetic analysis (Fritz *et al*., 2002). By the use of a field-electric sensor, which is based on an electrolyte-insulator-silicon (EIS) structure, any slight variation in surface potential by the presence of a charged molecule (i.e., nucleic acid) is detected by a change in conductivity. Because one is monitoring a change in the physical property of the DNA molecules, the need for fluorescent or other labels is eliminated. Although it is still at an early stage of development, the sensor accurately detects single-base mismatches during nucleic acid hybridization. Since as many as a million of such sensor spots can be packed onto a single chip, this is a promising genotyping platform in the future (Uslu *et al*., 2004).

Developing an affordable genotyping platform that is fast and accurate, simple and yet versatile requires continuous research advancement in genomics, nanotechnology, material sciences, and computer technology. This will happen only when people from different areas of expertise – biology, chemistry, physics, mathematics, and engineering – collaborate to tackle the problem. With rapid progress in genome-sequencing technologies during the past decade as precedent, the next wave of breakthroughs in genotyping technology will surely benefit from the age of integrated biology.

References

Barone AD, Beecher JE, Bury PA, Chen C, Doede T, Fidanza JA and McGall GH (2001) Photolithographic synthesis of high-density oligonucleotide probe arrays. *Nucleosides, Nucleotides and Nucleic Acids*, **20**, 525–531.

Fritz J, Cooper EB, Gaudet S, Sorger PK and Manalis SR (2002) Electronic detection of DNA by its intrinsic molecular charge. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14142–14146.

Gilles PN, Wu DJ, Foster CB, Dillon PJ and Chanock SJ (1999) Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. *Nature Biotechnology*, **17**, 365–370.

Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Gut IG (2001) Automation in genotyping of single nucleotide polymorphisms. *Human Mutation*, **17**, 475–492.

Liu RH, Yang J, Lenigk R, Bonanno J and Grodzinski P (2004) Self-contained, fully integrated biochip for sample preparation, polymerase chain reaction amplification, and DNA microarray detection. *Analytical Chemistry*, **76**, 1824–1831.

Mitchell P (2001) Microfluidics–downsizing large-scale biology. *Nature Biotechnology*, **19**, 717–721.

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, *et al*. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Research*, **12**, 1749–1755.

Russom A, Ahmadian A, Andersson H, Nilsson P and Stemme G (2003) Single-nucleotide polymorphism analysis by allele-specific extension of fluorescently labeled nucleotides in a microfluidic flow-through device. *Electrophoresis*, **24**, 158–161.

Schlotterer C (2004) The evolution of molecular markers–just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Service RF (2002) NANOTECHNOLOGY: Biology offers nanotechs a helping hand. *Science*, **298**, 2322–2323.

Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, **2**, 930–942.

The IHC (2003) The international HapMap project. *Nature*, **426**, 789–796.

Uslu F, Ingebrandt S, Mayer D, Bocker-Meffert S, Odenthal M and Offenhausser A (2004) Labelfree fully electronic nucleic acid detection system based on a field-effect transistor device. *Biosensors and Bioelectronics*, **19**, 1723–1731.

Introductory Review

What is an EST?

Winston A. Hide

South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

Andrey Ptitsyn

Pennington Biomedical Research Center, Baton Rouge, LA, USA

1. History of EST development

Technology for the high-throughput capture of DNA sequence was developed in the 1980s and early 1990s (*see* Article 24, **The Human Genome Project**, Volume 3) in preparation for the human genome project. The ability to find genes in genome sequences was limited and dependent upon the arduous task of sequencing a genome fragment. Sequencing technology is constrained to a relatively short read length. Therefore, strategies for sequencing large genome fragments rely upon breaking them up into little pieces. The task involves making a "library" of genomic sequences in a random manner, termed a shotgun library. The fragments in the library are inserted into a sequencing vector and are subsequently sequenced (*see* Article 4, **Sequencing templates – shotgun clone isolation versus amplification approaches**, Volume 3). Each fragment insert is within a size range that allows for a number of such sequence fragments to completely overlap each other and form an alignment or "assembly". The resulting contig that is generated is overlapped with other contigs to make up a genome fragment. The process is expensive and labor intensive, and finding genes within the genome sequence is not an exact science, resulting in a significant challenge (*see* Article 14, **Eukaryotic gene finding**, Volume 7).

Together with his team led by Mark Adams, Craig Venter exploited the shotgun sequencing technology already available to sequence cDNAs instead of shotgun genome fragments. Rather than attempting to sequence complete cDNAs to yield a few hundred gene sequences, Venter's group took a radical step: they made thousands of cDNAs from an mRNA library and simply inserted them into a sequencing vector. Instead of attempting to sequence a handful of full-length cDNAs, they performed shotgun style sequencing on the large collection of cDNAs. The single read sequences that resulted were essentially expressed gene tags. At around the same time as the development of ESTs (expressed sequence tags), drug companies were assessing their drug discovery approaches. George Poste at Smith-Kline was one of the first to recognize the potential of high-throughput information and together with Craig Venter and William Haseltine helped set up genomic

facilities such as Human Genome Sciences (a for profit company) and the Institute for Genome Research (a not-for-profit organization). The result was an immediate development of extensive proprietary EST data collections, and the race to discover all genes in the human genome began. The high-profile actions of Smith-Kline attracted intense competition amongst EST manufacturing centers and companies alike (Marshall, 1996). Entering late into the genomic competition, Merck took a risky strategy of funding "dumping" of ESTs into the public domain, probably in an attempt to reduce the proprietary value of EST tags in the possession of their competitor, Smith-Kline. The resulting database growth provided the foundation upon which modern bioinformatics has developed. 3 618 403 ESTs were present in dbEST, the most well known repository of ESTs, as of October 2004.

Each EST represents only a tiny portion of an entire transcribed gene fragment. Initially, ESTs were intended to merely identify the expressed gene. Mark Adams first published the approach for making ESTs in a widely read article in 1991 (Adams *et al*., 1991). The process by which ESTs are manufactured requires the construction of an mRNA library. Baldo *et al*. (1996) have provided a detailed description of how libraries are constructed and how normalization and library subtraction can be used to increase relative representation of less abundantly transcribed mRNAs. An average size of EST libraries ranges typically from 1000 to 10 000 entries (clones). It is estimated that from 10 to 30 000 different genes are expressed in a particular cell, depending on the cell type, with an average number of approximately 300 000 mRNA molecules per cell. The real number of different transcripts, resulting from alternative splicing, alternative polyadenylation, and posttranscriptional modification is yet to be discovered, but it can be estimated that the number of different transcripts far exceeds the number of genes in genome. Thus, an EST library cannot be regarded as a proper representation of the gene-expression pattern of a tissue (Vingron and Hoheisel, 1999). An EST library only represents a coarse-grained snapshot of the mRNA contents of a certain tissue at a certain moment. Low-expression genes are always underrepresented. As a rich source of discovery, EST libraries immediately attracted the attention of many scientists. During the last decade, hundreds of research projects have employed EST data in some form, either as primary data or as a supplementary source of information.

In order to understand EST sequence information, it is important to understand the means of EST manufacture. ESTs begin with the isolation of mRNAs from the collected biological samples. Each mRNA is then reverse-transcribed into a cDNA. The reverse transcriptase used to manufacture each cDNA in the library will eventually fall off the template (Figure 1), and so will terminate the production of the cDNA. Thus, a series of length-differentiated 3′ delimited cDNA fragments may be produced for each viable template mRNA in the library. The length of the cDNA will vary, and this is an important factor for the development of coverage for each mRNA template of an available gene. Clones are sequenced one at a time, from one or both ends of the DNA insert, using universal primers, which are complementary to the vector at the multiple cloning sites. The M13 forward primer may be located near the 5′ or the 3′ end of the cloned insert, depending on how the inserts were directionally cloned. Consequently, untranslated regions on the ends of the genes tend to be overrepresented by sequence tags. Only a few

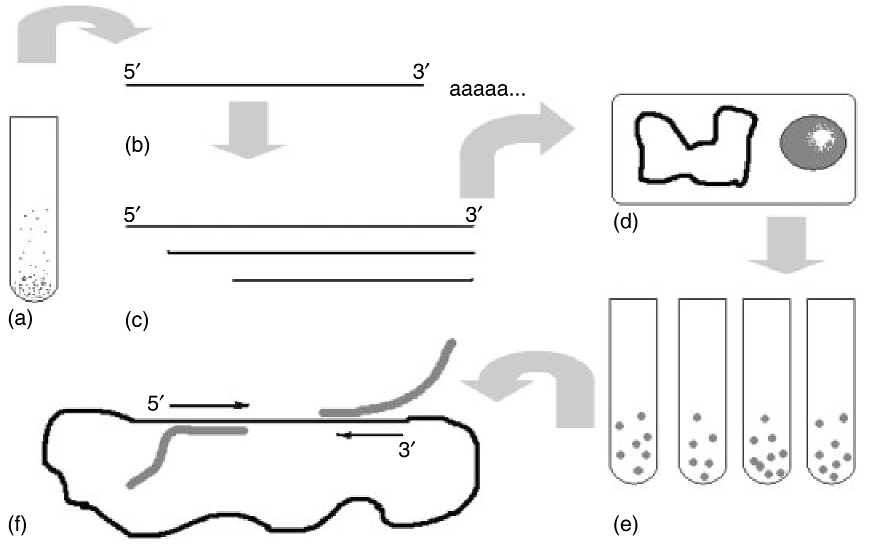


Figure 1 View of an EST manufacture process. After a sample (a) is collected, all transcribed copies of expressed genes (mRNAs) are isolated (b). Each mRNA is reverse-transcribed into a complementary DNA (cDNA) (c). Note that cDNA copies may have different lengths due to the polymerase processivity. The copy number of fragments is increased when they are inserted into a host cell (d). The resulting population of host cells containing cDNA fragments from the sample of interest is called a cDNA library (e). Typically, a few thousand clones are randomly picked from the cDNA library to produce single-pass 3′ reads, sometimes from the 5′ ends, and sometimes from a random location (f). The resulting collection represents fragments of different lengths starting from the polyadenylation site at 3′ end and partially overlapping fragments from 5′ end. Although 3′ and 5′ ESTs rarely overlap, they usually share the same clone ID in the annotation

hundred readable bases are produced from each sequencing read, and yet a full gene transcript may be several thousands of bases long. In publicly available databases, EST length varies from less than 20 to over 7000 bp, with an average length of 360 bp and standard deviation of 120 bp (dbEST, Genbank rel. 104). Not all are true single-read tags, but they are submitted and accepted as such, bringing extra complications to the EST database analysis. There are also countless variations in the EST generation technique. One of the most significant is using random primers, which results in production of fragments without direction, originating from different nonoverlapping parts of the same mRNA (Kapros *et al*., 1994). ESTs thus provide a "tag level" association with an expressed gene sequence, trading quality, and total sequence length for the high quantity of genes partially captured.

EST data quality varies greatly. Generation of EST data results in "low quality" sequence information. A single read is generated for each EST, and as such will contain errors from its generation at each step. These can include clone orientation, associated clone ID chimeras, and missing 3′ and 5′ reads. Because data are single-pass unedited sequences, they are also subject to errors caused by compressions and base-calling problems resulting in frame shifts. Reference to the Washington University website details common aspects of EST error (**http://genome.wustl.edu/est/**

and **http://genome.wustl.edu/est/index.php?std=1&disclaimer=1**). EST sequence has regions of high quality very close to regions of low quality, where quality can be defined as the number of correctly sequenced bases within a known window of reference. It is possible to utilize poor-quality sequence as long as relevant strategies for maximizing their utility are taken.

Futher reading

Jiang J and Jacob HJ (1998) EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Research*, **8**(3), 268–275.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Baldo MF, Lennon G and Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, **6**, 791–806.

Kapros T, Robertson AJ and Waterborg JH (1994) A simple method to make better probes from short DNA fragments. *Molecular Biotechnology*, **2**(1), 95–98.

Marshall E (1996) Genetics: the human gene hunt scales up. *Science*, **274**(5292), 1456.

Vingron M and Hoheisel J (1999) Computational aspects of expression data. *Journal of Molecular Medicine*, **77**(1), 3–7.

Introductory Review

Technologies for systematic analysis of eukaryotic transcriptomes

Robert L. Strausberg
The J. Craig Venter Institute, Rockville, MD, USA

1. Introduction

While much attention during the previous decade was focused on sequencing the genomes of the human (Venter, 2001; Lander, 2001) and other organisms, a revolution was already underway in identifying the coding content of these genomes. Complementary to genomic approaches, advances in transcriptome and proteome technologies provide an unprecedented opportunity to probe the diversity of gene expression and related biological functions. While much attention has been directed to the perceived limited number of human genes compared with previous expectations (Venter, 2001; Collins *et al*., 2004), it has become clear that eukaryotic transcript diversity is indeed very high (Xu and Lee, 2003; Wang *et al*., 2003b; Kapranov, 2002; Kriventseva *et al*., 2003; Hide *et al*., 2001). Indeed, alternative transcription and transcript processing provides new opportunity for discovery, as well as challenges to transcriptome technologies. Discussed in this overview are transcriptome-based technologies based on serial sequencing and hybridization that provide new opportunities for discovery and/or comprehensive high-throughput analysis.

2. Serial sequencing-based transcriptome technologies

The complexity of eukaryotic gene structure provides much opportunity for the generation of a diversity of transcripts from each gene. As shown in Figure 1, alternative transcripts can be generated through the use of alternative promoters, differential exon splicing, as well as the use of alternative polyadenylation sites (Kriventseva *et al*., 2003). Numerous studies have demonstrated that while some alternative transcript processing might reflect biological “noise”, many alternative processing events are correlated with changes in biological function (Xu and Lee, 2003; Yi, 2000; Munro *et al*., 1999).

While presenting an opportunity for discovery, the diversity of transcripts also presents technological challenges. First, transcripts are produced at very different

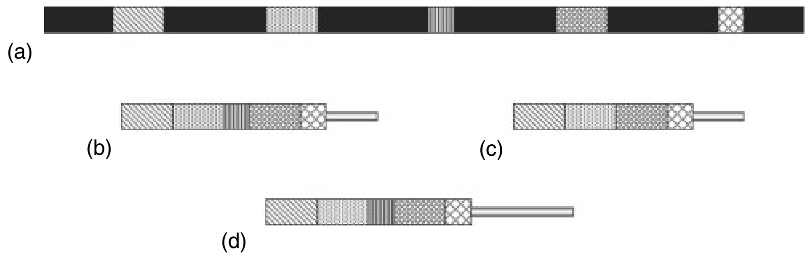


Figure 1 Multiple transcripts from individual genes. The diversity of eukaryotic transcripts is greatly enriched by the production of alternative transcripts from individual genes. For example, the 5′ or 3′ end of the transcripts can be extended by the use of alternative promoters and differential use of polyadenylation sites, respectively. In addition, alternative splicing can result in transcripts with differences in exon content. Illustrative of this phenomenon are three transcripts derived from the hypothetical gene A in which the black bars represent introns and the alternatively filled bars represent exons. Transcript B has a complete complement of the five exons. Transcript C is lacking the third exon and transcript D has an extended 3′ end based on the use of an alternative polyadenylation site

levels in cells, with some expressed only very rarely (Holland, 2002). In addition, transcripts vary greatly in size, with many being more than several thousand nucleotides. Thus, while it is the ultimate goal of transcript biology to gain complete knowledge of transcript diversity, there is no currently available technological approach that fulfills this need. However, through a combination of technological approaches, a more complete knowledge of eukaryotic transcripts is being gleaned, and applied to biomedical research.

3. cDNA libraries and sequencing

At the center of the transcript discovery and categorization are full-length cDNA (Ota, 2004; Imanishi, 2004; Strausberg *et al*., 2002a; Beisel, 2004; Okazaki, 2002) and expressed sequence tag (EST) (Adams, 1991; Brentani, 2003; Riggins and Strausberg, 2001; Scanlan, 2002; Sonstegard, 2002; Strausberg *et al*., 2002; Strausberg *et al*., 2002b; Strausberg and Riggins, 2001; Strausberg *et al*., 2003) efforts. In these approaches, cDNA libraries are prepared starting with mRNA derived from cell lines or tissues, and the cDNAs are single-pass sequenced, usually from the 5′ or 3′ ends (ESTs), or completely. In a novel EST strategy, called ORESTES, a random priming approach is used to generate EST sequences that are derived more frequently from the central coding region of the transcript (Strausberg *et al*., 2002b; Sakabe *et al*., 2003; de Souza, 2000).

Preparation of cDNA libraries can be based on a variety of technological approaches. In part because mRNAs are particularly fragile, the preparation of libraries that are enriched for full-length cDNAs is very challenging. In standard cDNA libraries, the frequency of each cDNA approximately reflects the relative proportion of each type of mRNA in the cell or tissue. Special approaches such as CAP-trapping are employed to capture the 5′ end of the mRNA, thereby enriching for full-length cDNAs (Hirozane-Kishikawa, 2003). Specialized library

normalization strategies (Bonaldo *et al*., 1996) are used to make the relative proportion of each cDNA in the library approximately equal, irrespective of the original differences in transcript proportions. Subtraction technologies (Hirozane-Kishikawa, 2003; Bonaldo *et al*., 1996) are employed to enhance the proportion of certain cDNAs in the library, such as those that have not previously been sampled. Normalization and subtraction are very important for the identification of cDNAs derived from transcripts that are infrequently expressed. These cDNA approaches have been critical for gene discovery, for defining the structure of genes in the human and other genomes, and for identifying alternative transcript processing. In addition to the valuable sequence information, physical collections of cDNAs have been made accessible to the community (Imanishi, 2004; Strausberg *et al*., 2002a,b; Kawai, 2001; Tanaka, 2000; Klein, 2002; Lennon *et al*., 1996; Sharov, 2003). Therefore, they serve as an important platform for all gene expression approaches.

4. Transcript tagging strategies

Although a long-term goal of transcriptome research is to catalog all of the transcript sequences in a cell, as well as to understand variations within networks, at present such an approach is too expensive. Therefore, clever strategies have been employed that allow for surveys of genes that are expressed in various cells. For example, although based on very different technological platforms, both the Serial Analysis of Gene Expression (SAGE) (Weeraratna, 2004; Porter *et al*., 2003; Porter and Polyak, 2003; Fujii, 2002; Velculescu *et al*., 1995; Boon, 2002; Lash, 2000; Velculescu *et al*., 2000; Lal, 1999) and Massively Parallel Signature Sequencing (MPSS) (Brandenberger, 2004; Meyers, 2004; Jongeneel, 2003; Brenner, 2000) provide short sequence tags, generally up to about 20 nucleotides, derived from the 3′ end of transcripts. The SAGE approach has now been employed for a decade and gained popularity because it is quantitative and provides data that is digital (DNA sequence) and is relatively cost-effective for tag generation compared to the cDNA approach because in SAGE the tags are concatenated, and 30 or more gene tags can be derived from each sequencing read. In addition, the digital nature of SAGE data facilitates the development of databases accessible to the community (Boon, 2002; Lash, 2000; Lal, 1999). The MPSS approach is based on a non-gel-based sequencing, together with in vitro cloning of templates on microbeads. MPSS has the advantage of generating several million tags per sample. This is particularly useful in that it extends the proportion of all transcripts that can be identified reliably within a sample (Brandenberger, 2004; Meyers, 2004; Jongeneel, 2003). Both technologies are somewhat limited by their inability to tag a small subset of transcripts because the cDNA lacks the specific restriction site required for tag generation, and because a certain fraction of tags do not uniquely identify the gene from which a transcript is derived.

The SAGE and MPSS approaches have both been employed effectively for identifying genes that are overexpressed in various tissues (for example, tumors compared with their normal counterparts) for biomarker applications, as well as potential targets for therapeutic intervention (Weeraratna, 2004; Porter *et al*., 2003;

Porter and Polyak, 2003; Fujii, 2002; Iacobuzio-Donahue *et al.*, 2003). The application of clever front-end strategies for preparation of specialized samples, together with novel approaches for database mining, has revealed genes that are specifically expressed in very specialized cell populations, such as in tumor-associated endothelium (St Croix, 2000).

5. Microarray-based strategies

Complementing the sequence-based approaches are those based on hybridization technology, using oligonucleotides (Lipshutz *et al.*, 1999; Wang *et al.*, 2003a) or cDNAs (Lapointe, 2004; Murray, 2004; Sorlie, 2003; Chen, 2003; Cheng, 2003; Ramaswamy and Perou, 2003; Alizadeh *et al.*, 2001) as probes to assess the presence and relative amounts of transcripts in cells and tissues. As the density of these chip formats has improved, and the knowledge of transcribed sequences has increased, expression of the complement of the genes in a genome, including the human, can be studied en masse. The high-throughput nature of this approach is a substantial advantage, as is the ability to study the entirety of transcript populations, limited in principle only by the DNA probes displayed on the chip. In cDNA formats, expression of each gene is, in general, assayed compared with an RNA reference, and results for each gene are characterized by the relative level of over- or underexpression (Sorlie, 2003; Chen, 2003; Bohen, 2003; Novoradovskaya, 2003). In the oligonucleotide format of Affymetrix, measurements are made on individual samples, thereby providing a quantitative view of gene expression (Lipshutz *et al.*, 1999). Array technologies have been applied to study a diversity of biological processes and diseases. For example, in cancer research, these technologies have been applied extensively toward molecular classification of tumors (Sorlie, 2003; Ramaswamy and Perou, 2003; Bloom, 2004; Chen, 2002; Chung *et al.*, 2002; Golub, 1999; Grant, 2004; Ramaswamy *et al.*, 2001; Wong, 2003; Ha, 2003; van't Veer, 2002; Perou, 1999).

The choice of probes must be carefully considered to achieve adequate specificity and informativeness. For example, transcripts from related genes may hybridize the same probes. In most cases, oligonucleotide probes have been targeted to the 3′ transcript end. However, recent advances in labeling strategies and chip design have opened opportunities to measure the full complement of transcripts (Yeakley, 2002). In addition, recent efforts to span entire chromosomes and genomes with oligonucleotide probes have facilitated de novo identification of transcripts (Kapranov, 2002).

6. Comparing the output of various gene expression technologies

A number of studies have compared the relative strengths and limitations of transcriptome technologies (Weeraratna, 2004; Bloom, 2004; Yauk *et al.*, 2004; Czechowski *et al.*, 2004; Lu *et al.*, 2004; Kim, 2003; Ye *et al.*, 2002; Toyoda,

2003; Coughlan *et al*., 2004). In general, the results have revealed that the measured expression levels for genes that are produced at moderate to high levels are at least qualitatively similar. However, as mentioned, transcripts are expressed at widely divergent levels, and for those that are produced in very limited quantities, the technologies produce quite different results. In a systematic study of *Arabidopsis* transcription factors by real-time RT-PCR, transcript abundance ranged between 0.001 and 100 copies per cell (Czechowski *et al*., 2004). However, only about half of these genes were detected on oligonucleotide arrays, and many of the transcripts were not present in the EST and MPSS databases. Genes expressed at low levels were especially underrepresented in these databases. This study demonstrates the need to carefully consider the dynamic range of the technology that is utilized.

In general, it is important to verify expression results derived from one technological approach with additional technologies. For example, genes that are measured to be differentially expressed by SAGE technology are often verified through the use of RT-PCR (Weeraratna, 2004; Porter *et al*., 2003; Loging, 2000). In addition, technologies such as tissue microarrays (Bloom, 2004; Mobasheri *et al*., 2004; Hans, 2004; Swierczynski, 2004; Jacquemier, 2003; Kallioniemi *et al*., 2001) offer the opportunity to examine a larger number of biological samples, and to gain perspective on expression and localization of protein derived from specific transcripts. Moreover, additional information can be gleaned by comparing expression data with genomic changes, such as gene copy number alterations (Hyman, 2002; Pollack, 2002).

7. Accessing transcriptome datasets

A variety of Internet-accessible databases (Strausberg *et al*., 2003; Boon, 2002; Lash, 2000; Lal, 1999; Pospisil *et al*., 2002; Ringwald, 2001) are available to facilitate the use of transcriptome data and for integration with additional datasets. For example, the National Center for Biotechnology Information's Gene Expression Omnibus (**http://www.ncbi.nlm.nih.gov/geo/**) displays both microarray and SAGE data submitted by the community. Genome browsers, such as the one developed and maintained at the University of California Santa Cruz (**http://genome.ucsc.edu/**), display a wide variety of transcriptome data, integrated with the human genome sequence and known genetic variation. Other resources, especially the NCI Cancer Genome Anatomy Project (**http://cgap.nci.nih.gov**), display a wealth of cancer-related transcriptome data, integrated with other datasets of interest to the community such as the Mitelman database of chromosome aberrations, as well as genetic variation in cancer-related genes. Together, these and other Web-based sites offer a wealth of data for online viewing and in silico experimentation.

8. Summary

Comprehensive transcriptome technologies offer an unprecedented opportunity to view the interplay of genes and networks within cells and tissues. While each of the technologies is somewhat limited, together they are a powerful new resource for

biological and biomedical research. Continued progress in technology development is required to completely characterize eukaryotic transcriptomes, especially with respect to transcripts that are rarely expressed. The study of alternative splice forms will likely provide much insight to the diversity of biological states. Overall, the progress over the past decade is quite remarkable and sets the stage for future research in which transcriptomes will truly be amenable to comprehensive study.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*. (1991) Complementary-DNA sequencing - expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Alizadeh AA, Ross DT, Perou CM and van de Rijn M (2001) Towards a novel classification of human malignancies based on gene expression patterns. *The Journal of Pathology*, **195**, 41–52.

Beisel KW, Shiraki T, Morris KA, Pompeia C, Kachar B, Arakawa T, Bono H, Kawai J, Hayashizaki Y and Carninci P, (2004) Identification of unique transcripts from a mouse full-length, subtracted inner ear cDNA library. *Genomics*, **83**, 1012–1023.

Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J and Yeatman TJ (2004) Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology*, **164**, 9–16.

Bohen SP, Troyanskaya OG, Alter O, Warnke R, Botstein D, Brown PO and Levy R (2003) Variation in gene expression patterns in follicular lymphoma and the response to rituximab. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1926–1930.

Bonaldo MDF, Lennon G and Soares MB (1996) Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Research*, **6**, 791–806.

Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, *et al*. (2002) An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11287–11292.

Brandenberger R, Khrebtukova I, Thies RS, Miura T, Jingli C, Puri R, Vasicek T, Lebkowski J and Rao M (2004) MPSS profiling of human embryonic stem cells. *BMC Developmental Biology*, **4**, 10.

Brenner S, Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, **18**, 630–634.

Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA, Neto ED, Grivet M, Gruber A, Guimaraes PEM, Hide W, *et al*. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13418–13423.

Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, *et al*. (2002) Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*, **13**, 1929–1939.

Chen X, Leung SY, Yuen ST, Chu KM, Ji JF, Li R, Chan ASY, Law S, Troyanskaya OG, Wong J, *et al*. (2003) Variation in gene expression patterns in human gastric cancers. *Molecular Biology of the Cell*, **14**, 3208–3215.

Cheng L, West RB, Zhu S, Linn SC, Nielsen TO, Goldblum JR, Patel R, Rubin BP, Brown P, Botstein D, *et al*. (2003) Expression profiling of fibromatosis by cDNA gene array analysis. *Modern Pathology*, **16**, 10A–10A.

Chung CH, Bernard PS and Perou CM (2002) Molecular portraits and the family tree of cancer. *Nature Genetics*, **32**(suppl), 533–540.

Collins FS, Lander ES, Rogers J and Waterston RH (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Coughlan SJ, Agrawal V and Meyers B (2004) A comparison of global gene expression measurement technologies in *Arabidopsis thaliana*. *Comparative and Functional Genomics*, **5**, 245–252.

Czechowski T, Bari RP, Stitt M, Scheible WR and Udvardi MK (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: Unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant Journal*, **38**, 366–379.

de Souza SJ, Camargo AA, Briones MRS, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, Andrade LEC, Carrer H, El-Dorry HFA, *et al*. (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12690–12693.

Fujii T, Dracheva T, Player A, Chacko S, Clifford R, Strausberg RL, Buetow K, Azumi N, Travis WD and Jen J (2002) A preliminary transcriptome map of non-small cell lung cancer. *Cancer Research*, **62**, 3340–3346.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grant GM, Fortney A, Gorreta F, Estep M, Del Giacco L, Van Meter A, Christensen A, Appalla L, Naouar C, Jamison C, *et al*. (2004) Microarrays in cancer research. *Anticancer Research*, **24**, 441–448.

Ha PK, Benoit NE, Yochem R, Sciubba J, Zahurak M, Sidransky D, Pevsner J, Westra WH and Califano J (2003) A transcriptional progression model for head and neck cancer. *Clinical Cancer Research*, **9**, 3058–3064.

Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, Muller-Hermelink HK, Campo E, Braziel RM, Jaffe ES, *et al*. (2004) Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, **103**, 275–282.

Hide WA, Babenko VN, van Heusden PA, Seoighe C and Kelso JF (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Research*, **11**, 1848–1853.

Hirozane-Kishikawa T, Shiraki T, Waki K, Nakamura M, Arakawa T, Kawai J, Fagiolini M, Hensch TK, Hayashizaki Y and Carninci P (2003) Subtraction of cap-trapped full-length cDNA libraries to select rare transcripts. *Biotechniques*, **35**, 510–516.

Holland MJ (2002) Transcript abundance in yeast varies over six orders of magnitude. *The Journal of Biological Chemistry*, **277**, 14363–14366.

Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, *et al*. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, **62**, 6240–6245.

Iacobuzio-Donahue CA, Ashfaq R, Maitra A, Adsay NV, Shen-Ong GL, Berg K, Hollingsworth MA, Cameron JL, Yeo CJ, Kern SE, *et al*. (2003) Highly expressed genes in pancreatic ductal adenocarcinomas: A comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Research*, **63**, 8614–8622.

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *Plos Biology*, **2**, 856–875.

Jacquemier J, Ginestier C, Charafe-Jauffret E, Bertucci F, Bege T, Geneix J and Birnbaum D (2003) Small but high throughput: How "tissue-microarrays" became a favorite tool for pathologists and scientists. *Annales de Pathologie*, **23**, 623–632.

Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJ *et al*. (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4702–4705.

Kallioniemi OP, Wagner U, Kononen J and Sauter G (2001) Tissue microarray technology for high-throughput molecular profiling of cancer. *Human Molecular Genetics*, **10**, 657–662.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP and Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, *et al*. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.

Kim HL (2003) Comparison of oligonucleotide-microarray and Serial Analysis of Gene Expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Experimental and Molecular Medicine*, **35**, 460–466.

Klein SL, Strausberg RL, Wagner L, Pontius J, Clifton SW and Richardson P (2002) Genetic and genomic tools for Xenopus research: The NIH Xenopus initiative. *Developmental Dynamics*, **225**, 384–391.

Kriventseva KI, Apweiler R, Vingron M, Bork P, Gelfand MS and Sunyaev S (2003) Increase of functional diversity by alternative splicing. *Trends in Genetics*, **19**, 124–128.

Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, *et al*. (1999) A public database for gene expression in human cancers. *Cancer Research*, **59**, 5403–5407.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 811–816.

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ and Altschul SF (2000) SAGEmap: A public gene expression resource. *Genome Research*, **10**, 1051–1060.

Lennon G, Auffray C, Polymeropoulos M and Soares MB (1996) The IMAGE consortium: An integrated molecular analysis of genomes and their expression. *Genomics*, **33**, 151–152.

Lipshutz RJ, Fodor SPA, Gingeras TR and Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**, 20–24.

Loging WT, Lal A, Siu IM, Loney TL, Wikstrand CJ, Marra MA, Prange C, Bigner DD, Strausberg RL and Riggins GJ (2000) Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Research*, **10**, 1393–1402.

Lu J, Lal A, Merriman B, Nelson S and Riggins G (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, **84**, 631–636.

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H and Decola S (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Research*, **14**, 1641–1653.

Mobasheri A, Airley R, Foster CS, Schulze-Tanzil G and Shakibaei M (2004) Post-genomic applications of tissue microarrays: Basic research, prognostic oncology, clinical genomics and drug discovery. *Histology and Histopathology*, **19**, 325–335.

Munro J, Stott FJ, Vousden KH, Peters G and Parkinson EK (1999) Role of the alternative INK4A proteins in human keratinocyte senescence: Evidence for the specific inactivation of p16INK4A upon immortalization. *Cancer Research*, **59**, 2516–2521.

Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO and Botstein D (2004) Diverse and specific gene expression responses to stresses in cultured human cells. *Molecular Biology of the Cell*, **15**, 2361–2374.

Novoradovskaya N, Perou C, Whitfield ML, Basehore S, Pesich R, Aprelikova O, Fero M, Brown PO, Botstein D and Braman J (2003) Universal human, mouse & rat reference RNA as standards for microarray gene expression analysis. *FASEB Journal*, **17**, A78–A78.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics*, **36**, 40–45.

Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, *et al*. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9212–9217.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL and Brown PO (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12963–12968.

Porter D and Polyak K (2003) Cancer target discovery using SAGE. *Expert Opinion on Therapeutic Targets*, **7**, 759–769.

Porter D, Lahti-Domenici J, Keshaviah A, Bae YK, Argani P, Marks J, Richardson A, Cooper A, Strausberg R, Riggins GJ, *et al*. (2003) Molecular markers in ductal carcinoma in situ of the breast. *Molecular Cancer Research*, **1**, 362–375.

Pospisil H, Herrmann A, Pankow H and Reich JG (2002) A database on alternative splice forms on the Integrated Genetic Map Service (IGMS). *In Silico Biology*, **3**, 0020.

Ramaswamy S and Perou CM (2003) DNA microarrays in breast cancer: The promise of personalised medicine. *Lancet*, **361**, 1576–1577.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 15149–15154.

Riggins GJ and Strausberg RL (2001) Genome and genetic resources from the cancer genome anatomy project. *Human Molecular Genetics*, **10**, 663–667.

Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, Hill DP, Kadin JA and Richardson JE (2001) The Mouse Gene Expression Database (GXD). *Nucleic Acids Research*, **29**, 98–101.

Sakabe NJ, de Souza JE, Galante PA, de Oliveira PS, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, *et al*. (2003) ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *Comptes Rendus Biologies*, **326**, 979–985.

Scanlan MJ, Gordon CM, Williamson B, Lee SY, Chen YT, Stockert E, Jungbluth A, Ritter G, Jager D, Jager E, *et al*. (2002) Identification of cancer/testis genes by database mining and mRNA expression analysis. *International Journal of Cancer. Journal International du Cancer*, **98**, 485–492.

Sharov AA, Piao YL, Matoba R, Dudekula DB, Qian Y, Vanburen V, Falco G, Martin PR, Stagg CA, Bassey UC, *et al*. (2003) Transcriptome analysis of mouse stem cells and early embryos. *Plos Biology*, **1**, 410–419.

Sonstegard TS, Capuco AV, White J, Van Tassell CP, Connor EE, Cho J, Sultana R, Shade L, Wray JE, Wells KD, *et al*. (2002) Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. *Mammalian Genome*, **13**, 373–379.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8418–8423.

St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, Lal A, Riggins GJ, Lengauer C, Vogelstein B, *et al*. (2000) Genes expressed in human tumor endothelium. *Science*, **289**, 1197–1202.

Strausberg RL and Riggins GJ (2001) Navigating the human transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 11837–11838.

Strausberg RL, Buetow KH, Greenhut SF, Grouse LH and Schaefer CF (2002) The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer. *Cancer Investigation*, **20**, 1038–1050.

Strausberg RL, Camargo AA, Riggins GJ, Schaefer CF, de Souza SJ, Grouse LH, Lal A, Buetow KH, Boon K, Greenhut SF, *et al*. (2002a) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16899–16903.

Strausberg RL, Camargo AA, Riggins GJ, Schaefer CF, de Souza SJ, Grouse LH, Lal A, Buetow KH, Boon K, Greenhut SF, *et al*. (2002b) An international database and integrated analysis tools for the study of cancer gene expression. *The Pharmacogenomics Journal*, **2**, 156–164.

Strausberg RL, Simpson AJ and Wooster R (2003) Sequence-based cancer genomics: Progress, lessons and opportunities. *Nature Reviews. Genetics*, **4**, 409–418.

Swierczynski SL, Maitra A, Abraham SC, Iacobuzio-Donahue CA, Ashfaq R, Cameron JL, Schulick RD, Yeo CJ, Rahman A, Hinkle DA, *et al*. (2004) Analysis of novel tumor markers in pancreatic and biliary carcinomas using tissue microarrays. *Human Pathology*, **35**, 357–366.

Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, *et al*. (2000) Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 9127–9132.

Toyoda N, Nagai S, Terashima Y, Motomura K, Haino M, Hashimoto S, Takizawa H and Matsushima K (2003) Analysis of mRNA with microsomal fractionation using a SAGE-based DNA microarray system facilitates identification of the genes encoding secretory proteins. *Genome Research*, **13**, 1728–1736.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Velculescu VE, Vogelstein B and Kinzler KW (2000) Analysing uncharted transcriptomes with SAGE. *Trends in Genetics*, **16**, 423–425.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995) Serial analysis of gene-expression. *Science*, **270**, 484–487.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J and Lee NH (2003a) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology*, **4**, R5.

Wang LH, Yang H, Gere S, Hu Y, Buetow KH and Lee MP (2003b) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research*, **63**, 655–657.

Weeraratna AT, Becker D, Carr KM, Duray PH, Rosenblatt KP, Yang S, Chen YD, Bittner M, Strausberg RL, Riggins GJ, *et al*. (2004) Generation and analysis of melanoma SAGE libraries: SAGE advice on the melanoma transcriptome. *Oncogene*, **23**, 2264–2274.

Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, Lin Y, Liao J, Shih JW, Cheung TH, *et al*. (2003) Expression genomics of cervical cancer: Molecular classification and prediction of radiotherapy response by DNA microarray. *Clinical Cancer Research*, **9**, 5486–5492.

Xu Q and Lee C (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, **31**, 5635–5643.

Yauk CL, Berndt ML, Williams A and Douglas GR (2004) Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*, **32**, e124.

Ye SQ, Lavoie T, Usher DC and Zhang LQ (2002) Microarray, SAGE and their applications to cardiovascular diseases. *Cell Research*, **12**, 105–115.

Yeakley JM, Fan JB, Doucet D, Luo L, Wickham E, Ye Z, Chee MS, and Fu XD (2002) Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology*, **20**, 353–358.

Yi X, White DM, Aisner DL, Baur JA, Wright WE, and Shay JW (2000) An alternate splicing variant of the human telomerase catalytic subunit inhibits telomerase activity. *Neoplasia*, **2**, 433–440.

Specialist Review

EST resources, clone sets, and databases

Janet F. Kelso
University of the Western Cape, Bellville, South Africa

1. Introduction

Gene identification via complete or partial transcript capture has proved to be a valuable and rapid route to gene discovery (Boguski and Schuler, 1995; Schuler *et al.*, 1996). The sequencing of Expressed Sequence Tags (ESTs) has been used in numerous pilot gene identification projects in which it has provided insight into the transcribed genomes of a wide range of organisms.

As a result, many groups, both academic and commercial, have contributed and continue to deposit many thousands of ESTs representing numerous organisms and expression states to public databases. In addition, many also distribute publicly the source clone sets from which these ESTs were generated, providing an experimental resource for expression studies.

2. EST databases

The public EST databases can be divided into (1) those that perform no processing of the incoming data and simply act as data repositories and (2) those that preprocess the data to reduce error and take advantage of sequence redundancy to increase quality.

2.1. Unprocessed EST databases

2.1.1. dbEST http://www.ncbi.nlm.nih.gov/dbEST/

dbEST, the EST division of Genbank, is a public repository for raw EST sequence and annotation information (Boguski *et al.*, 1993; Boguski and Schuler, 1995). More than 20 million ESTs representing in excess of 650 organisms have been deposited in dbEST since its inception in 1991. The most organisms most highly represented in dbEST are listed in Table 1. This EST data is available by anonymous

Table 1 Top 10 organisms represented in dbEST (26 February 2004)

Organism	ESTs
Homo sapiens (human)	5 472 005
Mus musculus + *domesticus* (mouse)	4 056 481
Rattus sp. (rat)	583 841
Triticum aestivum (wheat)	549 926
Ciona intestinalis	492 511
Gallus gallus (chicken)	460 385
Danio rerio (zebrafish)	450 652
Zea mays (maize)	391 417
Xenopus laevis (African clawed frog)	359 901
Hordeum vulgare + subsp. *vulgare* (barley)	352 924

ftp from **ftp://ncbi.nlm.nih.gov/genbank/**. Individual sequences and small batches can be obtained using Entrez (**http://www.ncbi.nlm.nih.gov/entrez/**).

The highly redundant EST data in dbEST is not clustered or assembled, and only a subset is grouped by species of origin. Unrestricted homology searches against dbEST will therefore commonly return numerous sequences that represent the same gene as the query, paralogous genes, and sequences from related species. Both the National Center for Biotechnology Information (NCBI) (**http://www.ncbi.nlm.nih.gov/BLAST/**) and the Swiss Institute of Bioinformatics (SIB) (**http://www.ch.embnet.org/software/aBLAST.html**) offer the ability to search subsets of dbEST restricted by species, with NCBI offering human, mouse, and "other" divisions and SIB offering the ability to select one or more from a large number of divisions including plants, prokaryotes, fungi, invertebrates, zebrafish, human, mouse, and rat.

2.2. *Processed EST databases*

EST data can be organized and presented in such a way as to produce valuable information about gene expression, including details of the location and timing of transcript expression, alternative splicing, and regulation. In general, the data stored in the large public EST databases is largely unorganized, sparsely annotated, and redundant. The sequences themselves are usually short, unprocessed, and error-prone.

However, the sheer volume of EST data generated by large-scale EST sequencing projects means that a significant improvement in reliability can be gained by taking advantage of EST sequence redundancy to reduce error and increase the length of represented transcripts (Jongeneel, 2000).

EST clustering systems that preprocess, cluster, and postprocess EST data to yield higher-quality transcript information aim to construct gene indices; nonredundant catalogs in which all represented transcripts are partitioned into groups (clusters) such that transcripts are placed in the same cluster if they represent the same gene or gene isoform. Gene indices facilitate gene expression studies and novel transcript detection. In addition to clustering, many groups perform

transcript reconstruction, using assembled clusters to build a consensus sequence that provides a longer and more accurate representation of the transcript represented by the cluster.

Homology searching against clustered EST collections such as Unigene will result in a more concise report than searching dbEST. Homology searching against clustered databases that provide contigs and consensus sequences for each cluster is very rapid, though the accuracy of the contig production and consensus sequence generation may affect the quality of the matches obtained.

A number of gene indices have been produced using publicly available EST data. These aim to reduce the error and redundancy present in the raw data and to thereby enhance the useful transcript information that can be gleaned from ESTs. The Institute for Genome Research (TIGR) gene indices (Quackenbush *et al*., 2000; Quackenbush *et al*., 2001) and Unigene (Wheeler *et al*., 2003), produced by NCBI, have focused on providing a reconstruction of the gene complement of various genomes. The STACK database (Christoffels *et al*., 2001; Miller *et al*., 1999) has focused on the detection and visualization of transcript variation and the production of accurate consensus sequences that represent the transcript variation in the context of tissue, developmental stage, and pathological states.

Both the TIGR and STACK databases offer BLASTable gene indices on their respective websites.

2.2.1. Unigene http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

Unigene, based at the National Centre for Biotechnology Information (NCBI), is one of the earliest and most enduring efforts for the automatic production of gene indices from Genbank sequences. Each Unigene cluster contains mRNA and EST sequences that represent a unique gene. Additional information such as the identity of the gene, chromosomal map location, and tissue types in which the gene is expressed (from SAGE and EST data) is also provided. NCBI does not generate contigs and/or consensus sequences for Unigene clusters.

Unigene databases are available for 38 organisms including: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), Cow (*Bos taurus*), Clawed frog (*Xenopus laevis*), Arabidopsis (*Arabidopsis thaliana*), wheat (*Triticum aestivum*), rice (*Oryza sativa*), barley (*Hordeum vulgare*), and maize (*Zea mays*). Databases are updated weekly with new ESTs and bimonthly with newly characterized sequences. All Unigene databases are available for download from **ftp://ncbi.nlm.nih.gov/repository/UniGene/**.

Unigene clusters can be searched by gene name, Unigene cluster ID, chromosomal location, cDNA library, accession number, and text terms. Sequence-based searches against Unigene human, rat, and mouse databases are available from the SIB at **http://www.ch.embnet.org/**.

Unigene has been used for the selection of unique transcripts for the construction of a cDNA microarray for the large-scale analysis of gene expression and as the candidates for the production of a human gene map.

For information on the construction of Unigene, see **http://www.ncbi.nlm.nih.gov/UniGene/build.html**.

2.2.2. TIGR Gene Indices http://www.tigr.org/tdb/tgi.shtml

TIGR produces gene indices for more than 40 organisms including various animal, plant, protist, and fungal species (Quackenbush *et al*., 2001; Quackenbush *et al*., 2000). The TIGR indices incorporate both ESTs sequenced at TIGR, ESTs from dbEST, and mRNAs from Genbank. Each TIGR cluster is represented by a Tentative Consensus sequence (TC, or THC in the case of Tentative Human Consensi), which is a FASTA-formatted sequence with a unique accession as well as additional information including details of the assembly, putative gene identification, and a list of tissues in which the transcript is expressed. Related databases generated by TIGR provide additional information about TCs. The Genomic Maps database provides genomic mapping for a subset of organisms for which TCs are available. The Institute for Genome Research Orthologous Gene Alignment database (TOGA) (Lee *et al*., 2002) provides information about orthologous sequences between TCs for the organisms for which TIGR gene indices have been generated.

The TIGR databases are freely available to researchers at nonprofit organizations at **http://www.tigr.org/tdb/tgi.shtml**. The TIGR Human Gene Index (HGI) is produced annually. The frequency of new releases varies between species and depends on the accumulation of new transcripts. The TIGR gene indices can be searched by nucleotide or protein sequence, EST, transcript or consensus identifiers, tissue, cDNA library name or library identifier, gene product name, or functional classification according to Gene Ontology (GO) terms (Ashburner *et al*., 2000).

2.2.3. STACK http://www.sanbi.ac.za/Dbases.html

The STACK human gene index is generated by clustering EST and mRNA data, and offers human transcript consensus sequences that reflect gene expression forms and alternate expression variants within 15 tissue-based and one disease category (Christoffels *et al*., 2001; Miller *et al*., 1999). This organization of transcript by expression site presents the opportunity to explore transcript expression in specific tissues or subsets such as disease-related sequences.

Each STACK cluster contains alignments, consensus sequences, and assembly information, and is dynamically linked to the Unigene database. Web-based software allows for the visualization of clusters and alignments, and highlights transcript variation. STACK database releases are made available with varying frequency – on average twice a year. STACKdb and the stackPACK toolset used to generate STACK are freely available to academic groups and can be downloaded from **http://www.sanbi.ac.za/CODES**. Sequence-based searching of STACKdb tissue divisions is available at **http://juju.egenetics.com/cgi-bin/stackpack/blast.py**.

STACKdb has been used to support the detection of a novel retinal-specific gene responsible for retinitis pigmentosa. The STACKpack toolset has been used in the production of various gene indices and for the survey of genes in the malarial genome.

2.3. cDNA clone sets

The availability of clones representing partial or full-length transcripts in an organized public collection is a critical resource for the ongoing genomic research. Clone sets have applications in gene discovery, a range of functional genetic analyses, and also as substrates for microarray expression studies.

2.3.1. I.M.A.G.E http://image.llnl.gov/

Recognizing the need for a publicly available clone collection, the Integrated Molecular Analysis of Genomes and their Expression (I.M.A.G.E.) consortium was formed in 1993. The aim of the group was to make cDNA clones representing all known genes, as well as their sequence, map, and expression information, publically available in order to facilitate biological research (Lennon *et al*., 1996). To this end, the group has generated more than 7.5 million clones from over 882 cDNA libraries representing more than 50 human tissues, as well as mouse, rat, zebrafish, rhesus monkey, *Fugu*, and *Xenopus*.

EST projects that are part of the I.M.A.G.E. consortium are listed in Table 2.

The sequences from the I.M.A.G.E. clones are submitted to dbEST, and the clones themselves are generally available royalty-free through a network of distributors in the United States and Europe (**http://image.llnl.gov/image/html/idistributors.shtml**). These groups generally provide added services that allow users to select the most appropriate clones for their research. I.M.A.G.E consortium distributors in Europe and the USA are listed in Table 3.

The I.M.A.G.E. consortium uses and provides IMAGEne, a clustering toolset to cluster the sequences from I.M.A.G.E. cDNA clones (Cariaso *et al*., 1999). Known gene clusters are based on NCBI's RefSeq, and candidate gene clusters are those with no known gene association. Those clones whose sequences do not match any other cDNA are grouped as Singletons. Users of the system are able to query against these cluster sets to obtain a list of available I.M.A.G.E. clones aligned with the corresponding known gene or consensus sequence. By offering the best representative clones that are available for order from the I.M.A.G.E clone set, IMAGEne provides a valuable laboratory research tool. IMAGEne is available via the web at **http://image.llnl.gov/image/imagene/current/bin/search**.

3. Conclusion

The EST databases provide an invaluable view of the transcriptomes of a large and growing number of organisms. The availability of EST data, and the associated annotation information, including details of the tissue source, provides an early expression map of the transcriptomes for these organisms. The public availability of clone sets representing a large number of organisms, tissues, diseases, and developmental stages is a valuable and ongoing resource for expression profiling and functional genomics studies.

Table 2 Some of the major I.M.A.G.E consortium EST projects

Project	Date	URL	Description
Merck/WashU EST Project	1995–1997	**http://genome.wustl.edu/est/index.php?human_merck=1**	A major early contribution to dbEST was the sequencing of 584 000 ESTs by Washington University under a project launched by Merck and Co. The cDNA libraries were constructed by Bento Soares at Columbia University, and arraying for high-throughput processing was performed by Greg Lennon at the Lawrence Livermore National Laboratory (Boguski and Schuler, 1995).
HHMI/WashU Mouse EST Project	1996–1998	**http://genome.wustl.edu/est/index.php?mouse=1**	Approximately, 400 000 mouse ESTs were contributed to dbEST by Washington University under the sponsorship of the Howard Hughes Medical Institute.
Cancer Genome Anatomy Project (CGAP)	1997–(ongoing)	**http://cgap.nci.nih.gov**	Through funding from the National Cancer Institute (NCI), human clones, largely from NCI-CGAP and ORESTES libraries, have been sequenced by Washington University and The NIH Intramural Sequencing Center (NISC). The aim of CGAP is to determine the gene expression profiles of normal, precancer, and cancer cells with a view to improving cancer diagnosis and treatment. Since 1999, CGAP has also contributed libraries and sequences for mouse, rat, Xenopus, and primate.
WashU Zebrafish EST Project	1997–2002	**http://genome.wustl.edu/est/index.php?zebrafish=1**	cDNA libraries produced by the zebrafish research community prior to 2002 have been arrayed by the I.M.A.G.E. consortium and sequenced at Washington University.
NIH Zebrafish Gene Collection	2002–(ongoing)	**http://zgc.nci.nih.gov/**	More than 3400 full-length ORF zebrafish clones have been produced through this initiative sponsored by the National Institutes of Health and are available for public research.
WashU Xenopus EST Project	1999–2002	**http://genome.wustl.edu/est/index.php?xenopus=1**	cDNA libraries produced by the Xenopus research community and CGAP prior to 2002 have been arrayed by the I.M.A.G.E. consortium and sequenced at Washington University.
NIH Xenopus Gene Collection	2002–(ongoing)	**http://www.ncbi.nlm.nih.gov/genome/flcdna/prj.cgi?prjid=15**	More than 600 full-length ORF Xenopus clones have been produced through this initiative sponsored by the National Institutes of Health and are available for public research.

University of Iowa Rat project	1998–(ongoing)	**http://ratest.eng.uiowa.edu/**	Clones from this project are arrayed and sequenced at the University of Iowa, sponsored by the National Institutes of Health. Unique clones are rearrayed and given I.M.A.G.E cloneIDs (located in the comment field of the Genbank entry). 25 000 of these sequence-verified clones are currently available through the I.M.A.G.E. distributors.
The Mammalian Gene Collection (MGC)	1999–(ongoing)	**http://mgc.nci.nih.gov/**	Initiated in 1999 as a collaborative effort between various institutes of the NIH, the Mammalian Gene Collection project aims to provide a catalog of full-length mammalian genes (Strausberg *et al*., 1999). The project has focused initially on the production of full-length cDNAs for human and mouse, and will later extend to include other mammals. Clones produced by the project are prepared from high quality mRNA extracted from cell lines or tissues. Clones are made available through the I.M.A.G.E consortium, while 3′ and 5′ ESTs are generated and released to dbEST. An ongoing informatics challenge is the selection of clones likely to represent full-length transcripts. In the initial phases of the project, clones with inserts of up to 3 to 4 kb were sequenced using techniques such as shotgun sequencing, primer walking, and concatenation. Sequence data is generated to the same standards as those specified by the Human Genome Project – finished sequence is therefore 99.99% accurate. Annotation of the sequence data is also performed. As of January 2002, a nonredundant set of more than 20 000 putative full-length human and mouse clones have been identified and full sequences for 9000 human and 4000 mouse clones have been produced. 75% of the selected clones contain full-length ORFs. Clone library lists, clone lists, and insert sequences in FASTA format are available for download from **http://mgc.nci.nih.gov/**. Sequenced clones can be searched using BLAST at the same site. Additionally, the genes represented by MGC clones can be searched by gene name or keyword at the website.

Table 3 I.M.A.G.E. consortium clone distributors in Europe and the United States of America

Continent	Distributor	URL	Description
Europe	RZPD	**http://www.rzpd.de/products/clones/**	As a nonprofit service facility, the RZPD provides research materials and data including cDNA clone sets. Nonredundant, sequence-verified cDNA clone sets are available for human, mouse, and rat, and full-length and open-reading-frame (ORF) clones are available for more than 30 organisms. More than 35 000 000 clones representing over 1000 cDNA libraries, including the I.M.A.G.E collection, are represented. The linking of various public databases to the available clone sets provides researchers with the ability to search for clones using Gene Data, Clone ID, Unigene Cluster, Chromosome Location, Genbank Accession ID, or Affymetrix Probeset ID.
Europe	MRC geneservice	**http://www.hgmp.mrc.ac.uk/geneservice/ index.shtml**	cDNA clone sets including the I.M.A.G.E, MGC, and NIA mouse cDNA collections, as well as other human, mouse, rat, *Drosophila*, *Fugu*, *Xenopus,* and chicken clones. Various software tools are provided to allow users to interrogate the clone sets.
USA	ATTC	**http://www.atcc.org/**	ATCC is a commercial distributor of the I.M.A.G.E., MGC, and NIA mouse clone sets, as well as variety of clones for human, mouse, rat, and pine. More than 500 000 partially sequenced cDNA clones representing the majority of human genes and more than 300 000 murine cDNAs and 20 000 rat cDNAs are represented in the collection.
USA	Invitrogen	**http://clones.invitrogen.com/index.php**	The clones distributed by Invitrogen are assembled from public resources including I.M.A.G.E. Researchers can search for clones of interest using CloneRanger™ (**http://clones.invitrogen.com/cloneranger.php**). Approximately 10 million clones across a wide range of species. The collection can be searched by clone ID, NCBI accession, Unigene cluster ID, LocusLink ID, keyword, sequence, or plate ID.
USA	Open Biosystems	**http://www.openbiosystems.com/clone_collections.php**	Open Biosystems distributes a large number of clone sets including the I.M.A.G.E. and Incyte sets. Organisms included are human, mouse, rat, dog, pig, *Drosophila, Xenopus, C. elegans*, monkey, and zebrafish. Clones can be searched by Genbank accession, gene name, and clone ID.

4. Further reading

a) Raw EST resources

Raw EST data: dbEST	**http://www.ncbi.nlm.nih.gov/dbEST/**
Download all sequences	**ftp://ncbi.nlm.nih.gov/genbank/**
Download individual sequences and small batches	**http://www.ncbi.nlm.nih.gov/entrez/**
BLAST searchable dbEST	**http://www.ncbi.nlm.nih.gov/BLAST/** **http://www.ch.embnet.org/software/aBLAST.html**
EST Tracefile archives	
Washington University Traces Viewer	**http://genome.wustl.edu/est/estˆsearch/nciˆviewer.html**
NCBI Trace Archive	**http://www.ncbi.nlm.nih.gov/Traces/**
General information about ESTs	
Washington University Genome Sequence Center	**http://genome.wustl.edu/est/**

b) Mining EST data

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. *Briefing in Bioinformatics*, **1**(1), 76–92.

c) Gene indices

Unigene	
Unigene build information	**http://www.ncbi.nlm.nih.gov/UniGene/build.html**
Download Unigene	**ftp://ncbi.nlm.nih.gov/repository/UniGene/**
BLAST searchable Unigene	**http://www.ch.embnet.org/**

Pontius J, Wagner L, Schuler G (2002) Unigene: a unified view of the transcriptome. *NCBI Handbook* http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch21 d1.pdf

TIGR	
TIGR information and download	**http://www.tigr.org/tdb/tgi.shtml**
BLAST searchable TIGR Gene Indices	**http://tigrblast.tigr.org/tgi/**
STACK	
STACK information and download	**http://www.sanbi.ac.za/Dbases.html**
BLAST searchable STACKdb	**http://juju.egenetics.com/stackpack/webblast.html**

d) Gene indices incorporating genome data

Ensembl	**http://www.ensembl.org/**
RefSeq	**http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html**
AllGenes	**http://www.allgenes.org/**

Pruitt KD, Katz KS, Sicotte H and Maglott DR (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends in Genetics*, **16**, 44–47.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

Boguski MS, Lowe TM and Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". *Nature Genetics*, **4**, 332–333.

Boguski MS and Schuler GD (1995) ESTablishing a human transcript map. *Nature Genetics*, **10**, 369–371.

Cariaso M, Folta P, Wagner M, Kuczmarski T and Lennon G (1999) IMAGEne I: clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes. *Bioinformatics*, **15**, 965–973.

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T and Hide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Research*, **29**, 234–238.

Jongeneel CV (2000) The need for a human gene index. *Bioinformatics*, **16**, 1059–1061.

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, *et al*. (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research*, **12**, 493–502.

Lennon G, Auffray C, Polymeropoulos M and Soares MB (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*, **33**, 151–152.

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR and Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Research*, **9**, 1143–1155.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

Quackenbush J, Liang F, Holt I, Pertea G and Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, **28**, 141–145.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, *et al*. (1996) A gene map of the human genome. *Science*, **274**, 540–546.

Strausberg RL, Feingold EA, Klausner RD and Collins FS (1999) The mammalian gene collection. *Science*, **286**, 455–457.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, **31**, 28–33.

Specialist Review

Using ESTs for genome annotation – predicting the transcriptome

Eduardo Eyras
Pompeu Fabra University - IMIM, Barcelona, Spain

1. Introduction

One of the goals of genome projects is to obtain the most complete set of genes from the genomic sequence. Automatic "genebuild" pipelines are designed to produce a set of gene annotations on which validation experiments can be performed. Even though the computational analysis of several eukaryotic genomes has been already carried out, the problem of automatic gene annotation is far from being solved. Different methods exploit a variety of biological data and hypotheses. A standard data source used for genome annotations is expressed sequence tags (ESTs), which from their origin (Adams *et al*., 1991) have played a very important role in the discovery of new genes. In spite of the experimental errors associated with ESTs (*see* Article 78, **What is an EST?**, Volume 4), they were soon recognized as a useful resource for genome annotation (Bailey *et al*., 1998). ESTs represent snapshots of the set of spliced mRNAs in a genome, the transcriptome, and in a given set of cell conditions. It is not clear how much of the transcriptome is represented by the ESTs, but they approximate a random sampling of the expressed genes as they are not necessarily constrained to any particular group of functional families. ESTs, besides enabling the detection of novel gene signals, also give information about possible alternative splicing patterns, polymorphisms, and untranslated regions in genes, complementing usual methods of annotation and ab initio gene prediction. We present here some of the advantages of using ESTs for genomic annotation, the potential problems, and possible solutions to them.

There are major issues regarding the quality of the EST datasets that must be handled before using them for genome annotation (Sorek *et al*., 2003). First of all, EST sequences are unverified single pass reads with an error rate that can be as high as 3% (Hillier *et al*., 1996). Moreover, ESTs represent incomplete transcript information and are subject to vector contamination in the cloning process. There can also be genomic contamination from the same organism, which would give false gene signals or incorrect splicing variations. An EST can also be part of a cDNA cloned from a pre-mRNA, which would give rise to an artifactual intron retention

event. Another problem in the ESTs is possible chimerism: cloned cDNAs may sometimes contain sequences from two different mRNAs, which can also lead to incorrect gene predictions. All these problems need to be addressed in any analysis involving ESTs. Some of the errors can be compensated for by the high redundancy of the data set, by imposing quality constraints in EST genomic alignments, or by using comparative genomics, whereas others require more specific curation.

2. Aligning ESTs to the genome

Before any metazoan genome sequence was available, ESTs were being used for finding genes. Since EST sequences are in general very redundant, a necessary step to look for genes is to cluster the ESTs according to sequence similarity (Adams *et al*., 1991; Schuler, 1997; Burke *et al*., 1998; *see also* Article 88, **EST clustering: a short tutorial**, Volume 4 for details). In these methods, a consensus sequence is derived for each cluster. The aim is to provide a gene index, where each cluster represents a gene in the genome, with information about expression and potential splicing variations (see Bouck *et al*., 1999 for a review). However, there are some problems inherent to these methods. In the process of building the clusters, it is difficult to avoid the placement of two paralog gene sequences in the same cluster, resulting in overclustering. On the other hand, very strict clustering can lead to genes being split into several clusters. Another problem is chimerism, since it can lead to the merging of two genes into one single cluster. Furthermore, owing to the high rate of EST sequencing error, it is sometimes difficult to distinguish an alignment gap from an event of alternative splicing. These problems can be avoided when considering the alignment of the ESTs to the genome. In fact, with the availability of genome sequences, EST self-clustering methods have been expanded to use genomic information, therefore improving the quality of the gene indices.

Here we will consider the approach of first aligning the individual EST sequences to the assembled sequence of a genome, and then, as explained below, clustering these EST alignments according to their exon–intron structure. There are several advantages to this approach, especially, when the sequence and the assembly are of good quality. First of all, it delineates the location of exons and introns in the genomic sequence. This allows the verification of the splice site signals, which enables the assignment of an EST to the correct strand. Low EST sequence quality is reflected as low alignment score and a higher frequency of frameshifts with respect to the genomic sequence. However, using identity thresholds above the expected error rate can avoid many mispredictions. The standard lower bound used for alignment similarity is 97% identity, which is consistent with the error rate mentioned before. Traditionally, the identity threshold is considered over a fixed distance on the EST, usually 50–100 bp. This is done to avoid cases where, for instance, an EST might have several segments aligned at 100% identity and some others at only 60% while still maintaining an average percent identity above threshold.

A standard quality test for an EST alignment is the verification that the splice sites defined by the alignment are canonical. There are three types of canonical introns commonly found in eukaryotes. These are characterized by the dinucleotides

GT-AG, GC-AG, and AT-AC. However, a number of ESTs define splice sites with noncanonical dinucleotides. Some variations might be spurious (Sorek *et al*., 2003) and some might correspond to novel splice sites (Burset *et al*., 2000). In spite of having canonical introns, some alignments can still define exon–intron boundaries with small variations between different ESTs. To avoid this, one can use a consensus approach, taking the most commonly used splice site. In principle, when ESTs from multiple libraries confirm the same splicing event, there is some added reliability to the definition of the intron, whereas an event present in only one EST library might be symptomatic of a sequence error or perhaps some sort of contamination.

For single-exon ESTs, which align as one single block in the genome, there can be ambiguities in the strand to which the exon should be mapped. This can be resolved for some cases using the coding bias of the exon sequence. About 60% of the human ESTs align as single exons on the genomic sequence. Many of these ESTs overlap with the 5′ or 3′ end of a known gene or a spliced EST, providing evidence for possible untranslated regions (UTR). The 3′ ESTs can sometimes align in separate clusters downstream of a gene, indicating possible alternative polyadenilation sites. Other single-exon ESTs can be found in isolation from any other gene features. It is difficult to determine whether these indicate single-exon genes or simply genomic DNA contamination. We would need to test these cases using codon-bias or protein similarity. However, this approach could potentially reject genes with unusual codon-bias that are underrepresented in databases. All these issues make single-exon genes difficult to annotate with ESTs.

Another problem that can be avoided by using genomic alignments is chimerism. Chimeric ESTs would be split into multiple chromosomes or perhaps define a suspiciously long intron on one chromosome. For this reason, it is standard to set an upper bound for allowed intron lengths, usually at around 200 kb, which will avoid chimeric ESTs that split on the same chromosome. Similarly, chimeric ESTs that split into different chromosomes can be avoided by setting a lower bound for the alignment coverage. This is usually chosen to be about 90%. This condition can also avoid possible undetected vector and cross-species contamination, as in these cases the EST would not be expected to align well. On the other hand, by consistently clipping a number of bases from both ends of the ESTs, one can increase the overall coverage of the alignments and keep ESTs that would be otherwise rejected. This clipping would eliminate some of the vector contamination and the poorer quality sequence, which is usually found at the ends. EST sequences in databases can also contain polyA or polyT tails, which should also be clipped prior to the alignment.

The thresholds used in genome alignments are usually higher than those used in EST self-clustering, hence when the genome sequence is of high quality we can gain in specificity over those methods, avoiding the problems derived from the clustering of paralogs. On the other hand, a problem in the genomic approach lies in potential genomic contamination and aberrant splicing. Errors in the splicing machinery could be overrepresented in ESTs owing to the repeated sequencing of the same genes. As pointed out before, although redundancy in the EST data can introduce noise, it can also provide a way to determine by consensus the correct splice sites annotations.

The aim of genebuild pipelines (for more details see Curwen *et al*., 2004) is to delineate all possible functional elements in a genome, in particular, all transcripts encoded in the DNA sequence. This task is usually split into parallel analyses for protein-coding genes, noncoding RNAs, and pseudogenes. ESTs are particularly valuable for detecting UTRs and alternative splicing in coding genes. However, ESTs can detect pseudogene signals as well and misguide the annotation of coding genes. In dbEST (Boguski *et al*., 1993), there are only a few human ESTs reported to be similar to known pseudogenes. However, as ESTs are gene fragments, there is an increased chance of aligning them to pseudogene loci. In the case of processed pseudogenes, one can define some heuristics to identify them. For instance, when a transcript sequence is compared to the entire genome, if the best alignment is spliced, any subsequent unspliced alignment of the same sequence elsewhere in the genome can be considered a candidate for a processed pseudogene. Potential pseudogenes can be further verified by testing for the absence of sequence homology in a closely related genome (Hillier *et al*., 2003). Additionally, one can search for other signals in the genomic sequence such as the presence of a downstream polyA tail, absence of the 5′ end of the transcript, in-frame stop codons and the presence of repeats within introns. All these tests are of course more effective with full-length cDNAs than with fragmented EST data.

Alternatively, one can rely on independent pseudogene annotations to avoid annotating these loci with ESTs. These annotations, however, are not easy to produce for novel genomes, for which there is not so much known gene data, and where ESTs play a role in the first-pass genome annotation effort. ESTs can also identify noncoding RNAs. For these cases, how to recognize and avoid pseudogenes remains a challenge.

ESTs reflect the sequence similarities between gene families. Thus, ESTs can potentially detect multiple gene loci sharing domains or perhaps even UTRs. Using the alignment thresholds given above, taking the best alignment plus any other alignment within 2% of the coverage, and using the above heuristics to avoid pseudogenes, we have found that nearly 10% of the alignable ESTs can align in more than one locus in the human genome, and about 2% align in more than one chromosome. On the other hand, it would make sense to pick only the best-in-genome alignment, as each EST is expected to come from a unique gene sequence. This would also reduce the misassignment of ESTs to gene loci. Accordingly, EST alignments to DNA sequence are most meaningful when performed as a comparison against the entire genomic sequence. This is complicated by the fact that nearby gene loci can sometimes combine in a unique transcript unit (Thomson *et al*., 2000).

The problem of aligning ESTs to a genome becomes a technical challenge when considering the human genome, with a size of about 3 Gb, and the approximately 5 million of ESTs currently available in dbEST (Boguski *et al*., 1993). This requires massive computational power and specialized alignment software that can not only process this amount of data in a reasonable time but which is also able to produce the correct alignment. The traditional approach is hierarchical. First, a fast and approximate alignment is produced that helps reduce the search space. Subsequently, the ESTs are more precisely aligned to this reduced space with a more accurate and computationally intensive method. In the last few

years, a new generation of programs have been produced that can manage this process in one step for an entire genome very quickly: BLAT (Kent, 2002) and Exonerate (Slater, **http://www.ebi.ac.uk/~guy/exonerate/**). The point of such methods is to be very competitive in speed and at the same time use an intron model sophisticated enough to extract an accurate exon–intron structure for the cDNAs and ESTs. A major objective of annotating genomes with ESTs is the delineation of introns and exons correctly on the genome sequence, rather than just the production of a consecutive block of alignments without contextual information.

ESTs are usually used in combination with cDNAs and mRNAs to obtain more complete information. Recent full-length cDNA sequencing projects are therefore very useful for improving our knowledge of protein-coding genes and the transcriptome. However, these projects seem to generate antisense transcripts as well (Kiyosawa *et al*., 2003): These are full-length cDNAs that align perfectly onto the genome sequence but that most often lack any recognizable open reading frame (ORF) and usually align on the opposite strand of a protein-coding gene. Moreover, they may play a regulatory role in eukaryotes (Lehner *et al*., 2002). Accordingly, one may suspect that EST datasets might contain sequences from these antisense transcripts, which might misguide the protein-coding gene search. Using the properties of the antisense RNAs, ESTs could be used to detect antisense transcripts in a genome with a good available annotation. However, in the case that ESTs are used to annotate a genome for which there are few, if any, annotations, it would remain difficult to properly identify antisense RNAs.

3. Expression data

The library of origin of the ESTs (*see* Article 80, **EST resources, clone sets, and databases**, Volume 4) supplies useful information about expression in an organism. Using sequence similarity between gene transcript sequences and ESTs, one can derive information about expression patterns. If, moreover, one uses the exon–intron structure of genes and ESTs, this association with expression information can be done more accurately: an EST is associated to a particular splice form of a gene when the exon–intron structures of both are compatible and in the same genomic locus. In this way, gene predictions can be linked to expression data. Using this method, the eVOC expression vocabularies (Kelso *et al*., 2003), based on normalized EST libraries for human, are linked to the Ensembl predictions (Curwen *et al*., 2004), providing a qualitative expression profile in silico. Likewise, this mapping can be carried out with nonnormalized libraries, providing quantitative information of expression.

Although EST data does not necessarily represent an exhaustive sampling of all cell states, a careful analysis of the expression data allows the identification of splice forms that are potentially specific to particular tissues, and in particular, specific to a disease state (Xie *et al*., 2002; Xu *et al*., 2002; Xu *et al*., 2003; Wang *et al*., 2003). Experimentally, one could scan multiple libraries to cover as many cell conditions as possible. However, every time ESTs are sequenced, we are not any closer to obtaining the complete set of transcripts for that library. In fact, the

same sequence may be read over and over, and lowly expressed genes may remain undetected. As a consequence, while the amount of EST data for human and mouse in dbEST is very large, this high redundancy has not necessarily added any new information. On the contrary, high redundancy can add extra noise and make the annotation of exons and transcripts more difficult. In the case of human and mouse, it is often necessary to perform a very strong filtering of the EST alignments to avoid this noise in the results. Filters can be based simply on duplicated alignments or on higher coverage and identity thresholds. It would be convenient to identify a selective procedure to sequence ESTs from genes that are expressed at very low level, or with a very restrictive pattern of expression.

4. Predicting the transcriptome

Most of the evidence for splicing variation in the human genome observed so far has been based on ESTs (Mironov *et al*., 1999; Brett *et al*., 2000; Modrek *et al*., 2001; Modrek *et al*., 2002; Kan *et al*., 2002). The standard method has been until now to infer local variations of splicing events from the exon–intron structures of the ESTs. Some of the methods are based on variations with respect to mRNA sequences aligned to the genome (Kan *et al*., 2002). Other methods have used the clusters from EST self-clustering and aligned those to the genome.

However, in order to properly define the proteins encoded in the transcriptome, one needs to derive the complete sequences for the alternative forms. This has motivated the development of new methods of transcript prediction from ESTs (Kan *et al*., 2001; Wheeler, 2002; Haas *et al*., 2003; Xing *et al*., 2004; Eyras *et al*., 2004). Genome-based methods base the joining and merging of ESTs on the exon–intron structure. The initial step of sequence similarity is dealt with at the alignment step and the analog of the consensus sequence is here the genomic sequence itself. In this way, the redundancy of the ESTs is translated to splicing structure redundancy. In general, one EST can be part of one or more putative transcripts, as sometimes one cannot uniquely specify the possible full transcript from which it originated. This is the case, for instance, for ESTs that align on constitutive exons. The transcripts produced should not be a combinatorial enumeration of all possible splice site concatenations. Rather, the complete exon–intron structure given by the experimental evidence should be taken into account by the algorithm. This is important in order to reflect the possible global dependencies of the splice events in the final predictions. We are, however, limited by the fragmented nature of the ESTs. By taking a random sampling of tags from known mRNA sequences and maintaining the global dependencies of the simulated tags, genome-based methods can recover the original transcript structures. However, it is still possible that transcripts not present in the original set might be produced. Indeed, no matter the depth of the EST sequencing, short ESTs do not add any knowledge about alternative splicing, and there is an increased uncertainty for every two consecutive sites of alternative splicing that are not covered by at least one single EST (Eyras *et al*., 2004).

In general, we expect many of the predicted transcripts to be partial or even complete coding sequences. Some of them might give rise to new protein sequences,

while others will match known ones. This can also be used as a validation of the predictions. EST-derived transcripts can be searched against databases of known proteins to determine homologies. This, however, can yield a large number of false-positives (Jongeneel, 2000). There are also a number of statistical models that predict the coding regions in ESTs taking into account, among other factors, the transition between UTR and coding sequence. This can also help predict the UTRs in genes, and equivalently detect possible noncoding RNAs. These methods, however, use the bias in coding nucleotide sequences, hence possibly overlooking protein-coding transcripts with unusual codon-bias.

Conservation at the level of sequence implies a functional relevance; hence, we expect many of the expressed sequences to be conserved between closely related species. This can also serve as a validation of the splicing structures and predicted proteins. Alternative splicing has been estimated to be 60% conserved at the splice-junction level between human and mouse (Thanaraj *et al*., 2003). Furthermore, some forms of alternative splicing (*see* Article 78, **What is an EST?**, Volume 4) appear to be more conserved than others (Sugnet *et al*., 2004; Kan *et al*., 2004). We would then expect to find some conservation at the transcript level, keeping in mind that there seems to be no conservation for some classes of genes (Nurtdinov *et al*., 2003). Determination of conservation of the predicted complete transcripts can be carried out at two levels: the sequence level (transcript comparison) and at the splicing structure level (exon–intron comparison between the orthologous genomic regions). The two comparisons together can increase our specificity when assigning conserved structures. This defines a notion of transcript orthology, which has been used to validate transcripts predicted from ESTs in (Eyras *et al*., 2004). Using these methods, one observes that 20% of the transcript variants in human are also found in mouse as predicted from ESTs (data not published). In fact, we expect many more variants to be conserved, as the available EST data is not exhaustive due to the possible differences in sampling between human and mouse.

In general, genome annotation is traditionally based on sequence homology searches and ab initio prediction methods. However, when there is little species specific data or the known gene sequences for closely related genomes are not very well conserved, homology-based methods have a limited predictive power, hence it is necessary to rely on ab initio methods. These methods need, in general, a training set, usually consisting of experimentally verified genes, which for some genomes can be limited. EST sequencing can provide a head start for automatic annotation in these cases. Methods such as ClusterMerge (Eyras *et al*., 2004) can produce transcript predictions from the available ESTs and cDNAs. These transcripts can serve in principle as a first-pass genebuild. As some of the predicted transcripts are bound to be incomplete, the defined transcripts can be used as evidence for gene predictors in order to complete the transcripts according to the contextual genomic sequence (Howe *et al*., 2002), increasing the power of usual gene-prediction methods. Furthermore, EST-based transcripts can also provide spliced structures on which to train the gene-prediction models. This opens up many new possibilities, as gene models could eventually be created according to specific splicing structures and different coding biases. These training methods could be valuable to predict genes according to specific patterns of expression and perhaps also for developing ab initio predictors of genes with alternative splicing. Gene-prediction models that

can incorporate this type of information would be very valuable for the automatic annotation of genomes.

ESTs remain a crucial data source for transcript-discovery tools, and their potential is not yet exhausted. Handling this type of data is not easy, but the amount of potentially useful information is worth its use. Knowing well the pitfalls of EST data and developing methods that take these into account can provide very useful information for genome annotation.

Related articles

Article 78, **What is an EST?**, Volume 4; Article 80, **EST resources, clone sets, and databases**, Volume 4; Article 88, **EST clustering: a short tutorial**, Volume 4

Acknowledgments

I would like to thank C. Ben Dov, R. Castelo, G. Parra, R. Guigo, M. Caccamo, V. Curwen, M. Clamp, J. Cuff, W. Hide, J. Valcarcel, and A. Hatzigeorgiou for useful discussions, and T. Alioto for his valuable comments on the manuscript. I would also like to thank everybody at the Ensembl project and the Informatics Division of the Sanger Institute where I was able to experiment and learn many of the things described in this article.

Further reading

Christofels A, van Gelder A, Greyling G, Miller R, Hide T and Hide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Research*, **31**, 234–238.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi S, Pertea G, Sultana R and White J (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B and Moreno RF (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**(5013), 1651–1656.

Bailey LC Jr, Searls DB and Overton GC (1998) Analysis of EST-driven gene annotation in human genomic sequence. *Genome Research*, **8**, 362–376.

Boguski M, Lowe T and Tolstoshev C (1993) dbEST – database for expressed sequence tags. *Nature Genetics*, **4**, 332–333.

Bouck J, Yu W, Gibbs R and Worley K (1999) Comparison of gene indexing databases. *Trends in Genetics*, **15**, 159–161.

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J and Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**, 83–86.

Burke J, Wang H, Hide W and Davison DB (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Research*, **8**, 276–290.

Burset M, Seledtsov IA and Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, **28**(21), 4364–4375.

Curwen V, Eyras E, Andrews DT, Clarke L, Mongin E, Searle S and Clamp M (2004) The Ensembl automatic gene annotation system. *Genome Research*, **14**(5), 942–950.

Eyras E, Caccamo M, Curwen V and Clamp M (2004) ESTgenes: alternative splicing from ESTs in Ensembl. *Genome Research*, **14**(5), 976–987.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al*. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**(19), 5654–5666.

Hillier LD, Lemmon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A and Gish W (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807–828.

Hillier LW, Fulton RS, Fulton LA, Graves TA, Pepin KH, Wagner-McPherson C, Layman D, Maas J, Jaeger S, Walker R, *et al*. (2003) The DNA sequence of human chromosome 7. *Nature*, **424**(6945), 157–64.

Howe KL, Chothia T and Durbin R (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Research*, **12**(9), 1418–1427.

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics*, **1**(1), 76–92.

Kan Z, Castle J, Johnson JM and Tsinoremas NF (2004) Detection of novel splice forms in human and mouse using cross-species approach. *Pacific Symposium on Biocomputing*, **9**, 42–53.

Kan Z, Rouchka EC, Gish WR and States DJ (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*, **11**, 889–900.

Kan Z, States D and Gish W (2002) Selecting for functional alternative splices in ESTs. *Genome Research*, **12**(12), 1837–1845.

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel C, McCarthy M, *et al*. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Research*, **13**, 1222–1230.

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**(4), 656–664.

Kiyosawa H, Yamanaka I, Osato N, Kondo S and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*, **13**, 1324–1334.

Lehner B, Williams G, Campbell RD and Sanderson CM (2002) Antisense transcripts in the human genome. *Trends in Genetics*, **18**, 63–65.

Mironov AA, Fickett JW and Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Research*, **9**, 1288–1293.

Modrek B and Lee C (2002) A genomic view of alternative splicing. *Nature Genetics*, **30**, 13–19.

Modrek B, Resch A, Grasso C and Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, **29**, 2850–2859.

Nurtdinov RN, Artamonova II, Mironov AA and Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, **12**(11), 1313–1320.

Schuler GD (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, **75**, 694–698.

Sorek R and Safer HM (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Research*, **31**(3), 1067–1074.

Sugnet CW, Kent WJ, Ares M and Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing*, **9**, 66–77.

Thanaraj TA, Clark F and Muilu J (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, **31**(10), 2544–2552.

Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M, *et al*. (2000) Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Research*, **10**(11), 1743–1756.

Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH and Lee MP (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research*, **63**(3), 655–657.

Wheeler R (2002) A method of consolidating and combining EST and mRNA alignments to a genome to enumerate supported splice variants. *Lecture Notes in Computer Science*, **2452**, 201–209, WABI 2002.

Xie H, Zhu W, Wasserman A, Grebinskiy V, Olson A and Mintz L (2002) Computational analysis of alternative splicing using EST tissue information. *Genomics*, **80**(3), 326–330.

Xing Y, Resch A and Lee C (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Research*, **14**, 426–441.

Xu Q and Lee C (2003) Discovery of novel splice forms and functional analysis of cancer specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, **31**(19), 5635.

Xu O, Modrek B and Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, **30**(17), 3754–3766.

Specialist Review

Using ORESTES ESTs to mine gene cancer expression data

Sandro J. de Souza, Pedro A.F. Galante and Rodrigo Soares
Ludwig Institute for Cancer Research, São Paulo, Brazil

1. Introduction

The major challenge facing biologists following the completion of the Human Genome Project is the complete characterization of the human transcriptome. It is absolutely crucial that we understand how the expression of a given gene affects the onset of a given trait, either normal or pathological. For obvious reasons, one disease model that has attracted a lot of resources and efforts is cancer. The amount of data on expressed genes in cancers has grown exponentially in the last 15 years. Since the development of the EST (Expressed Sequence Tags) technology in the early 1990s (Adams *et al*., 1991), the contribution of ESTs derived from cancer samples in the public databases has been substantial. Today, more than 50% of all human ESTs in dbEST correspond to sequences derived from more than 50 types of tumors. Although efforts from individual laboratories were, and still are, important for the development of public repositories of cancer-oriented data, the emergence of large-scale sequencing projects was crucial especially due to the amount of data generated. As recently compiled by Strausberg *et al*. (2003), there are two major cancer-oriented genome projects that focus their efforts in the transcriptome: the Cancer Genome Anatomy Project (CGAP) and the Fapesp/Ludwig Human Cancer Genome Project (HCGP). The CGAP's main objective is the integration of genomics and cancer research (Strausberg *et al*., 1997; Strausberg *et al*., 2000). An index of genes expressed in cancer cells and their normal counterparts has been achieved in CGAP through the use of both the EST and the Serial Analysis of Gene Expression (SAGE) (Velculescu *et al*., 1995) technologies. HCGP centered its efforts in the generation of ESTs from a variety of tumors frequently found in the Brazilian population. The project employed a new methodology called *Open-reading frame EST sequencing* (ORESTES) (Dias-Neto *et al*., 2000), which generates ESTs from the central part of the transcripts (see details below).

The availability of the human genome sequence represented a significant boost in cancer research and increased the value of the transcriptome databases. Owing to the poor efficiency of *ab initio* gene prediction, transcriptome data proved to be essential in finding genes in the genome sequence (Lander *et al*., 2001; Venter *et al*.,

2001). Moreover, the human genome sequence provides a framework from which data from diverse sources can be viewed in an integrated manner. This integration allows the use of databases like OMIM (**http://www.ncbi.nlm.nih.gov/omin**) that record information of genetically inherited diseases, in conjunction with the cancer-oriented transcript databases. This integration has speeded up the finding of cancer-related genes.

Here, we will discuss some strategies to mine cancer-oriented transcript databases, with a special emphasis on the data produced by HCGP.

2. HCGP

The HCGP was launched in 1999 with funds provided by the Ludwig Institute for Cancer Research and Fapesp (the State of São Paulo Research Foundation). The basis of the project was the sequencing of transcripts expressed in different types of tumor cells and their normal counterparts. Here lies the most important aspect of the HCGP. The strategy used to generate the ESTs was ORESTES, which provides sequences mainly from the central part of the transcripts. In this aspect, the HCGP data complements data from other EST projects, especially CGAP, which target the extremities of the transcripts. Figure 1 illustrates this aspect of the ORESTES data. ORESTES is characterized by the generation of sequences through a low-stringency PCR reaction using arbitrary primers (Dias-Neto *et al*., 2000). Besides providing sequences from the central part of the transcript, the low-stringent PCR reaction is expected to provide a normalization effect to the set of generated sequences (Dias-Neto *et al*., 2000). The success of HCGP has stimulated the use of ORESTES in another large-scale EST sequencing project (Verjovski-Almeida *et al*., 2003).

More than 820 000 sequences were deposited in public databases (Table 1). A comparison involving ORESTES and UniGene (Build # 163) is presented in Table 2. Almost 500 000 ORESTES are present in UniGene, distributed in 52 072 clusters. UniGene clusters containing at least one ORESTES and onek nown mRNA (mostly derived from large-scale full-insert sequencing projects) comprise 18 532, while 33 540 clusters contain only ESTs. Interestingly, there are 15 508 UniGene clusters that contain only ORESTES sequences. This is unexpected because UniGene is a 3′-oriented transcript database, and ORESTES is biased toward the central part of the transcripts. However, UniGene has recently modified its strategy of building clusters with the incorporation of the genome sequence (**http://www.ncbi.nlm.nih.gov/UniGene/**). Since there is still no detailed description of this new clustering methodology, it is impossible to evaluate the



Figure 1 Schematic view of all transcripts from UniGene cluster Hs. 287.666 (Build 163). Observe that ORESTES are preferentially located at the central part of the transcript

Table 1 ORESTES in dbEST. Only main tissues are listed

	Normal	Tumor	Total
Total number deposited			823.121
Breast	59 991	73 354	133 345
Colon	8814	69 419	78 233
Head/neck	1917	149 703	151 620
Brain	46 023	20 883	66 906
Prostate	19 002	19 182	38 184
Stomach	13 756	38 916	52 672

Table 2 ORESTES in UniGene (Build # 163)

ORESTES in UniGene	497 610
Clusters with ORESTES	52 072
Clusters with ORESTES + mRNAs	18 532
Clusters with ORESTES only	15 508

Table 3 ORESTES mapped onto the genome sequence. Protocols for mapping and clustering can be found at Sakabe *et al*. (2003) and Galante *et al*. (2004)

ORESTES mapped	565 535
ORESTES defining >1 exon	185 525
Clusters with ORESTES	118 573
Clusters with ORESTES only	76 456

significance of these clusters. When integrated to the genome sequence, 565 535 mapped ORESTES sequences defined 118 573 clusters (Table 3). Details on the mapping and clustering strategy can be found in Sakabe *et al*. (2003) and Galante *et al*. (2004). A fraction of the mapped sequences (185 525) defines more than one exon at the genome level. More than 6100 new exons have been defined by these ORESTES sequences.

The unique features of the ORESTES sequences have implications for the mining process. It has been shown, for instance, that ORESTES are preferentially located at the central part of the transcript and therefore represent best the coding region (Dias-Neto *et al*., 2000; Camargo *et al*., 2001). Furthermore, the ORESTES dataset is enriched with sequences derived from rare messages (Dias-Neto *et al*., 2000; Sakabe *et al*., 2003). Those features make the ORESTES collection enriched with cDNAs corresponding to coding regions of rare transcripts. Below, we will discuss some strategies to mine the ORESTES dataset.

3. Identification of genes differentially expressed in tumors

Identification of genes that are differentially expressed in tumors is one of the most important forms of mining the cancer-oriented transcript databases. Several

reports have been published in which such genes were successfully identified by ESTs (Schmitt *et al*., 1999; Welsh *et al*., 2001; Scanlan *et al*., 2002). The use of ESTs in quantitative analysis of gene expression is a controversial issue. Owing to the heterogeneous nature of the EST data, with many sequences coming from normalized libraries, the absolute frequency of a given transcript in the EST database does not directly reflect its abundance in the cell's transcriptome. The same is true for the ORESTES collection of sequences due to the normalization effect of the methodology. It is advisable then to use the ORESTES sequences coupled to a more quantitative measurement of gene expression, such as SAGE.

We have developed an electronic protocol for the identification of genes differentially expressed in any tissue by using both ORESTES and SAGE (Leerkes *et al*., 2002). Assuming that one has a large amount of ORESTES for a given tumor and its normal counterpart, these sequences can be used to identify candidate genes to be differentially expressed in tumors. This analysis is used as a starting point to guide subsequent analyses. A critical issue when using ORESTES for the identification of differentially expressed genes is the collection of primers used to construct the respective libraries. In comparing ORESTES from two samples, it is important that all sequences came from libraries constructed using the same set of arbitrary primers. Otherwise, the pool of genes represented in each set will be different and false positive candidates will be identified. Clustering of sequences is also important to allow a comparison of the number of sequences in each cluster in the two samples being analyzed. In the final step of the protocol, SAGE is used to corroborate the analysis done using only ORESTES. We have used this protocol in a comparison of 21 437 and 37 890 ORESTES from normal and tumor breast, respectively (Leerkes *et al*., 2002). We were able to identify 154 genes as candidates for being differentially expressed in tumors. Among these, 28 have been shown to be overexpressed in tumors. We obtained 82% of success in experimental validations for 11 candidate genes.

4. Identification of new transcribed regions in the human genome

Expressed sequences have been critical to the identification of human genes (Lander *et al*., 2001; Venter *et al*., 2001). Not surprisingly, the ORESTES dataset has contributed significantly to this process, especially due to the central distribution of the sequences and the normalization effect. It is expected that the ORESTES collection is enriched with transcript sequences spanning exon–exon boundaries, especially if compared to 3′ ESTs, due to the fact that introns are less frequently found at the 3′ end of genes. ESTs that align continuously along the genome are of dubious quality since they can represent genomic contaminants. Thus, ORESTES sequences are important for the unambiguous identification of genes.

We have used 250 000 ORESTES from a variety of tumors to identify a new transcribed region in chromosome 22 (de Souza *et al*., 2000). All these sequences were mapped onto chromosome 22, and a comparison with previously annotated transcribed regions in this chromosome was made. We found 219 new transcribed

regions in chromosome 22. Since the ORESTES sequences are not indexed to a specific region of the transcript, it was impossible at the time of original mapping to define the number of new genes discovered by ORESTES in chromosome 22. For this report, we compared this 219 transcribed regions with the mapping of all known human mRNAs, and observed that for 90 of them there was already a matching known human mRNA reported after the publication of our original analysis.

The integrated approach of mapping all human ESTs onto the sequence of the human genome has been very fruitful in terms of characterizing new genes. For example, 19 additional genes were found in chromosome 21 through the use of mapped ESTs and a set of stringent criteria to identify reliable 3′ ends (Reymond *et al*., 2002). Interestingly, these new genes are small and poorly represented in the cDNA databases.

A different but powerful strategy for characterizing new human genes has been proposed by us (Camargo *et al*., 2001). In this strategy, clusters of ESTs mapped onto the genome sequence can be used for direct gap closure. RT-PCR experiments with primers designed based on the sequence of neighbor clusters form the basis of such strategy. As proof of principle, Camargo *et al*. (2001) characterize four new human genes. This transcript finishing strategy is been used in a context of a large-scale initiative that so far has characterized hundreds of new human genes (Sogayar *et al*., 2004).

5. Identification of splicing variants differentially expressed in tumors

A fascinating new perspective on the human transcriptome has emerged in the last few years. The degree of variability found at the transcriptome level greatly exceeds previous estimates. One of the major sources of variability is alternative splicing, which seems to occur in at least half of all human genes (Mironov *et al*., 1999; Modrek *et al*., 2001; Sakabe *et al*., 2003).

We recently reported an interesting case of a splicing variant of the gene *NABC1* (Correa *et al*., 2000), which uses a previously unidentified 135-bp exon (NABC1_5B), through the mapping of an ORESTES sequence to the human genome. Collins *et al*. (1998) reported that NABC1 is a strong candidate oncogene mapped to a genomic region frequently amplified in several types of tumors. These authors demonstrated that NABC1 is highly expressed in breast cancer cell lines. We confirmed the increased expression of NABC1 in breast tumor. We also showed that NABC1 and NABC1_5B are both underexpressed in colon tumors (Correa *et al*., 2000). ORESTES sequences have also allowed the characterization of splicing variants of the semaphorin 6B (Correa *et al*., 2001).

The unique features of the ORESTES sequences motivated us to perform a large-scale analysis of alternative splicing in these sequences (Sakabe *et al*., 2003). It is expected that the ORESTES dataset would be enriched with rare splicing variants, which affect the structure of the corresponding protein. We found that genes showing low expression, as evaluated by SAGE, contain more ORESTES than

other ESTs, reinforcing the normalization effect of the ORESTES methodology. Furthermore, less ORESTES are required to detect a splicing variant and the ORESTES dataset is enriched with variants that affect the coding region of human genes, a feature derived from its biased distribution along transcripts. We found that 85% of all events detected by ORESTES are within the coding region, while 77% of the events detected by conventional ESTs are within this category ($p < 0.001$) (Sakabe *et al*., 2003).

These features of ORESTES regarding alternative splicing make the methodology an efficient platform for an exhaustive coverage of the variability in the human transcriptome.

6. Final considerations

The importance of mining cancer-oriented transcript databases is clearly evident in the last 10 years. While the major efforts have been directed toward the characterization of new genes related to cancer and the identification of genes differentially expressed in tumors, new forms of exploring the data have emerged. The search for splicing variants differentially expressed in tumors is such an example. Another situation is the identification of SNPs from transcripts expressed in tumors (Brentani *et al*., 2003). Data from HCGP have been crucial in the development of the available cancer-oriented databases. More importantly, a very fruitful interaction between HCGP and CGAP teams was established (Strausberg *et al*., 2002b; Strausberg *et al*., 2003; Brentani *et al*., 2003), culminating with further contributions (Boon *et al*., 2002; Sakabe *et al*., 2003; Cerutti *et al*., 2003; Iseli *et al*., 2002; Jongeneel *et al*., 2003). This interaction was possible because of a common notion that the datasets were complementary as were the collection of tumors being approached by both projects. Fundamental was the commitment of both projects to the release of the data to the community. This common strategic view has allowed the emergence of integrated databases. SAGE Genie (Boon *et al*., 2002), for example, although structured around SAGE data, uses EST information to explore the effect of transcript variability on the gene-to-tag assignment. This and other successful interactions show that integration is the key for effective mining of cancer-oriented transcript databases.

Acknowledgments

The authors would like to thank all participants from both projects, the Ludwig/Fapesp Human Cancer Genome Project and the Cancer Genome Anatomy Project. PAFG is supported by a PhD fellowship form Fapesp.

Further reading

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420, 6915**, 520–562.

Strausberg RL, Buetow KH, Greenhut SF, Grouse LH and Schaefer CF (2002a) The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer. *Cancer Investigation*, **20**, 7–8, 1038–1050.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, de Souza SJ, *et al*. (2002) An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11287–11292.

Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA Jr, Dias Neto E, Grivet M, Gruber A, Guimaraes PEM, Hide W, *et al*. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13418–13423.

Camargo AA, Samaia HP, Dias Neto E, Simao DF, Migotto IA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, *et al*. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12103–12108.

Cerutti JM, Riggins GJ and de Souza SJ (2003) What can digital transcript profiling reveal about human cancers? *Brazilian Journal of Medical and Biological Research, USA*, **36, 8**, 975–985.

Collins C, Rommens JM, Kowbel D, Godfrey T, Tanner M, Hwang SI, Polikoff D, Nonet G, Cochran J, Myambo K, *et al*. (1998) Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8703–8708.

Correa RG, Carvalho AF, Pinheiro NA, Simpson AJ and de Souza SJ (2000) NABC1 (BCAS1): alternative splicing and downregulation in colorectal tumors. *Genomics USA*, **65**, 299–302.

Correa RG, Sasahara RM, Bengtson MH, Katayama ML, Salim AC, Brentani MM, Sogayar MC, de Souza SJ and Simpson AJ (2001) Human semaphorin 6B [(HSA)SEMA6B], a novel human class 6 semaphorin gene: alternative splicing and all-trans-retinoic acid-dependent downregulation in glioblastoma cell lines. *Genomics*, **73**, 343–348.

Dias-Neto E, Correa RG, Verjovski SA, Briones MR, Nagai MA, da Silva WA Jr, Zago MA, Bordin S, Costa FF, Goldman GH, *et al*. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **97, 7**, 3491–3496.

Galante PA, Sakabe NJ, Kirschbaum-Slager N and de Souza SJ (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.

Iseli C, Stevenson BJ, de Souza SJ, Samaia HB, Camargo AA, Buetow KH, Strausberg RL, Simpson AJ, Bucher P and Jongeneel CV (2002) Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Research, USA*, **12, 7**, 1068–1074.

Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJ, *et al*. (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4702–4705.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Leerkes MR, Caballero OL, Mackay A, Torloni H, O'Hare MJ, Simpson AJ and de Souza SJ (2002) In silico comparison of the transcriptome derived from purified normal breast cells and breast tumor cell lines reveals candidate upregulated genes in breast tumor cells. *Genomics*, **79, 2**, 257–265.

Mironov AA, Fickett JW and Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Research*, **12**, 1288–1293.

Modrek B, Resch A, Grasso C and Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, **29**, 2850–2859.

Reymond A, Camargo AA, Deutsch S, Stevenson BJ, Parmigiani RB, Ucla C, Bettoni F, Rossier C, Lyle R, Guipponi M, *et al*. (2002) Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics*, **79, 6**, 824–832.

Sakabe NJ, de Souza JE, Galante PA, Oliveira PS, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, *et al*. (2003) ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *Comptes Rendus Biologies* **326**, 10–11, 979–985.

Scanlan MJ, Gordon CM, Williamson B, Lee SY, Chen YT, Stockert E, Jungbluth A, Ritter G, Jager D, Jager E, *et al*. (2002) Identification of cancer/testis genes by database mining and mRNA expression analysis. *International Journal of Cancer*, **98**, 485–492.

Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B and Rosenthal A (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Research USA*, **1, 27, 21**, 4251–4260.

Sogayar MC, Camargo AA, Bettoni F, Carraro DM, Pires LC, Parmigiani RB, Ferreira EN, de Sa Moreira E, do Rosario D de O Latorre M, Simpson AJ, *et al*. (2004) A transcript finishing initiative for closing gaps in the human transcriptome. *Genome Research*, **14**, 1413–1423.

de Souza SJ, Camargo AA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, Andrade LE, Carrer H, El-Dorry HF, *et al*. (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12690–12693.

Strausberg RL, Buetow KH, Emmert-Buck MR and Klausner RD (2000) The cancer genome anatomy project: building an annotated gene index. *Trends in Genetics*, **16**, 103–106.

Strausberg RL, Camargo AA, Riggins GJ, Schaefer CF, de Souza SJ, Grouse LH, Lal A, Buetow KH, Boon K, Greenhut SF, *et al*. (2002b) An international database and integrated analysis tools for the study of cancer gene expression. USA. *The Pharmacogenomics Journal*, **2**, 156–164.

Strausberg RL, Dahl CA and Klausner RD (1997) New opportunities for uncovering the molecular basis of cancer. *Nature genetics*, **15** Spec No. 415–416.

Strausberg RL, Simpson AJ and Wooster R (2003) Sequence-based cancer genomics: progress, lessons and opportunities. *Nature Reviews Genetics, USA*, **4, 6**, 409–418.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995) Serial analysis of gene expression. *Science*, **20, 270, 5235**, 484–487.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **16, 291, 5507**, 1304–1351.

Verjovski-Almeida S, DeMarco R, Martins EA, Guimaraes PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY, Kitajima JP, Adamson RE, *et al*. (2003) Transcriptome analysis of the acoelomate human parasite schistosoma mansoni. *Nature Genetics, USA*, **35, 2**, 148–157.

Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF Jr and Hampton GM (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research, USA*, **15, 61, 16**, 5974–5978.

Proteome knowledge bases in the context of cancer

Djamel Medjahed and Peter A. Lemkin
National Cancer Institute at Frederick, Frederick, MD, USA

1. Introduction and motivation

The origin of most cancers can be often traced to a single transformed cell (Fearon *et al*., 1987). The evolution of the disease follows a yet to be completely understood pathway of molecular transformations occurring at both genomics and proteomics levels. Most cancers show a significant preponderance to statistically originate from well-defined parts of their respective organs. It is then only normal that investigations to identify biomarkers indicative of the early onset of the disease be focused on these organ-specific regions.

This point was elegantly demonstrated by Page *et al*. (1999) in a careful experiment, in which they used magneto-immuno-chemical purification methods to extract pure cell populations and compare the protein expression observed in experimental two-dimensional Polyacrylimide Gel Electrophoresis (2D-PAGE) (O'Farrell, 1975; O'Farrell *et al*., 1977) maps obtained from normal, milk-producing luminal epithelial cells exhibiting a tendency to develop breast carcinomas versus outer, myoepithelial cells.

This thorough characterization was achieved by using a combination of enabling technological platforms (Aebersold *et al*., 2000; Bussow, 2001; Fivaz *et al*., 2000; Kriegel *et al*., 2000; Dihazi *et al*., 2001; Angelis *et al*., 2001; Weiller *et al*., 2001; Wulfkuhle *et al*., 2000) that allowed them to flag a number of proteins exhibiting a significant differential expression between the two types of cells, and therefore warranting a closer evaluation of their potential as biomarkers of breast cancer. The time and costs involved in using these techniques can be quite prohibitive, particularly on a large scale.

This led to the initial motivation to address the need that either on a routine basis, or to establish optimal experimental conditions before hand, one might be interested in predicting the gene products likely to be detected in narrow ranges of isoelectric focusing point (pI) and molecular weight (Mw).

We believe that the initial search for cancer biomarkers can greatly benefit by formulating the hypothesis developed from knowledge-based bioinformatic tools. We will now describe in some detail two such predictive databases whose development was at least in part motivated by these pressing issues.

2. VIRTUAL2D: a Web-accessible predictive database for proteomics analysis

2.1. Introduction

The growing use of immobilized pH gradients (Gorg *et al*., 2000; Bjellqvist *et al*., 1993) and automation in producing SDS gels has allowed the emergence of reproducible, high-resolution two-dimensional separation of proteins. Furthermore, the availability of databases of primary sequences of proteins, either directly determined or inferred from genome databases allows the validation of the contents of experimental 2D-PAGE maps (WORLD-2DPAGE, 2004). VIRTUAL2D (Medjahed *et al*., 2003b) is an interactive Web-accessible collection of reference, organism-specific, synthetic (pI, Mw) maps based on the consensus, published amino acid sequences.

The approach used to determine the isoelectric focusing point and molecular mass of a peptide can then simply be summed up as follows:

1. Scan the primary sequence of the peptide.
2. Assign the pK of each contributing amino acid according to Table 1.
3. Sum up all the mass contributions.

The resulting pI/Mw for the peptide is then given by the ratio of

$$Pk_{\text{tot}} = \frac{\left\{pK_{\text{Cterm}} + \sum_{\text{int}} pK_{\text{int}} + pK_{\text{Nterm}}\right\}}{(n-2)} \tag{1}$$

and

$$M_{\text{rtot}} = \sum_i M_t i.$$

where the pI summation runs over all n contributing, internal amino acids

2.2. Database mining

The resulting plot of pI versus the molecular mass yields a theoretical 2-D PAGE map with a striking bimodal distribution centered around pH 7.4–7.5. This feature seems to be shared by all organisms analyzed (Figure 1a,d). The biochemical justification most often advanced in explanation of this observation is that the majority of proteins would tend to naturally precipitate out of solution around the cytoplasmic pH of approximately 7.2. The pI is the pH for which the protein charge is overall neutral. It, therefore, represents the point of minimum solubility due to the absence of electrostatic repulsion, resulting in maximum aggregation. While this provides an explanation for experimental 2D-PAGE maps, we must remember that no such correction was incorporated in the modeling. What then is the basis for the separation of proteins into acidic and basic domains in computed pI/Mw charts? In our efforts to answer these questions, we carried out a simulation whereby

Table 1 Values of amino acid masses and pKs (determined (Gorg *et al*., 2000) at high molar concentrations of urea) used in pI/Mw computation. The segregation is underscored by the fact that the pKs of roughly half the internal amino acids fall below pH 6.0, while for the rest they are greater than or equal to 9.0

Ionizable group	pK	Molecular mass
C-terminal	3.55	
N-terminal		
Met	7.00	132.994
Thr	6.82	102.907
Ser	6.93	88.88
Ala	7.59	72.88
Val	7.44	100.934
Glu	7.70	130.917
Pro	8.36	98.918
Internal		
Asp	4.05	116.89
Glu	4.45	130.917
His	5.98	138.943
Cys	9	104.94
Tyr	10	164.978
Lys	10	114.961
Arg	12	157.989
C-terminal side chain groups		
Asp	4.55	116.89
Glu	4.75	130.917

groups of 1545 peptides varying in length from 50 to 600 AA, in increments of 10 were randomly generated. This brings the total number of simulated sequences to 8 6520 versus 8 6518 real peptides extracted from current databases, thereby improving the prospects of any meaningful comparative statistics. As mentioned earlier, the calculation of the pI values is carried out iteratively. The pK of a peptide is calculated by tallying the contributions to the charge from the n-terminus, the c-terminus and the internal portion of the peptide. As can be observed in Figure 2, the resulting simulated pI/Mw distribution is strikingly similar to that adopted by the extracted sequences. While this may seem surprising at first, given the total absence of bias in both the lengths and content of the peptides used for the simulation, it is in fact a direct consequence of the constraints imposed by a limited proteomic alphabet of twenty amino acids with distinct pKs roughly half of which are either acidic or basic (Table 1).

In fact, as is reflected in Table 1, only seven internal amino acids make nonzero contributions to the pI of the peptide. These seven amino acids are: cysteine, aspartic acid, glutamic acid, histidine, lysine, arginine, and tyrosine. It is reasonable to suspect that a high percentage of the variation in the calculated pI values of the simulated data would be modulated by the representation of these seven amino acids,as the majority of the contribution to the charge comes from the internal portion of the peptide. To investigate the actual contribution of these seven amino acids in determining an overall pI value, a multiple regression model was developed using the adjusted numbers of these seven amino acids as predictor variables and

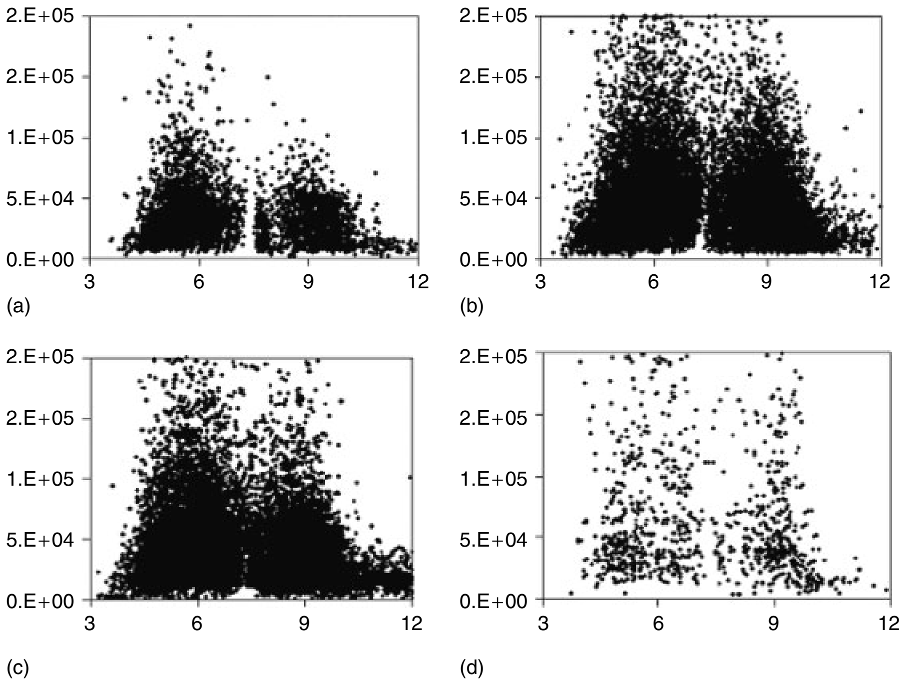


Figure 1 pI/Mw Map for several organisms. To keep in line with the experimental limits encountered in practice, the pI/Mw plot has been confined to less than 2×10^5 kD for the molecular mass and 3.0< pI < 12.0 for the isoelectric focusing point. (a) *E. coli*, (b) *Homo sapiens*, (c) mouse, (d) *Plasmodium falciparum*

the pI value as the dependent variable. The adjusted count for an amino acid is equal to the actual number of times the amino acid is found in the peptide divided by the length of the peptide. The adjusted counts will be denoted as follows:

aR = adjusted count for arginine
aC = adjusted count for cysteine
aD = adjusted count for aspartic acid
aE = adjusted count for glutamic acid
aK = adjusted count for lysine
aH = adjusted count for histidine
aY = adjusted count for tyrosine

The regression model in question uses the linear, quadratic, and cubic powers for each adjusted number of the seven amino acids that contribute to the pI calculation when they are part of the interior of the protein. A total of 21 independent variables were employed in the regression analysis. This analysis yields a multiple correlation factor R of 0.931. The coefficient of determination (the square of the multiple R) gives the proportion of the total variance in the dependent variable accounted for by the set of independent variables in a multiple regression model. For the model in question, 0.866 is the square of the multiple R. Consequently, 86.6% of the total variation in the pI values was accounted for by the aforementioned seven amino

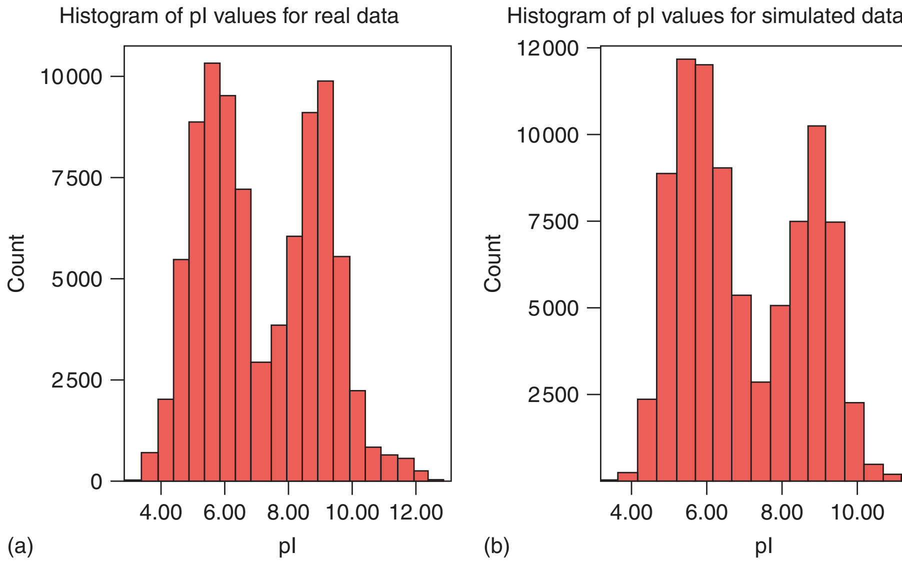


Figure 2 Side-by side comparison of ”pI/Mw” histograms for *Homo sapiens* (a) computed using amino acid sequences from TrEMBL/SWISS-PROTversus (b) randomly generated as described in the text

acids. The simulation result confirms the hypothesis that the total number of these seven amino acids is the key factor is explaining the pI value of a peptide.

The predicted pI score in the regression model is denoted as pI’, and it is the dependent (criterion) variable in the regression model. The equation for the regression model is:

$$pI' = a + \Sigma_i X_i \tag{2}$$

where a is the intercept of the model, b_i is the partial slope for the ith predictor in the model, and X_i is the ith predictor in the model. There will be 21 different predictors in the model: seven linear terms (aR, aC, aD, etc.), seven quadratic terms (aR^2, aC^2, aD^2, etc.) and seven cubic terms (aR^3, aC^3, aD^3, etc.). All parameters were estimated by ordinary least squares using the SPSS 8.0 computer package (SPSS, 2004).

The coefficient of determination or R^2 for the model is the proportion of variance of the pI values accounted for by the regression model. It is equal to the sum-of-squares regression divided by the total sum-of-squares:

$$R^2 = \frac{\Sigma pI' - <(pI)>)^2}{\Sigma(pI - pI')^2} \tag{3}$$

where

$$<(pI)> = \sum pi/N$$

A Perl script was written to process large organism-specific proteome datasets in FASTA format downloaded from the European Bioinformatics Institute (CEBI,

2004). It will output tab-delimited files of the molecular mass, pI, SWISS-PROT (SwissProt, 2004) accession number and identification for each protein entry. Inorder to increase the analytical value of Virtual2D to the scientific community, interactivity is built into these plots by implementing the following features (displayed in Figure 3).

- possibility to use the database on any JAVA-enabled computer;
- pan, zoom, and click features;
- with Internet connection, hyperlinks between each data point and popular databases (SWISS-PROT, NCBI, etc.).

2.3. Comparison with experimental data

Computed pI/Mw values were compared against those reported experimentally in two cases. In the first example, a high-resolution map for *Escherichia coli* obtained over a narrow pH range (4.5–5.5) was used. Landmarks provided by reference proteins whose characteristics were independently confirmed can be used to calibrate positions over the entire area of the image. pI, molecular masses, and relative intensities can then be determined by interpolation for all detected protein spots (Figure 4a). A minimally distorted "constellation" consisting of proteins whose predicted pI/Mw values are fairly close to their experimentally determined counterpart, displayed in Figure 4(b) can then be used in principle to "warp" (align) the experimental gel, onto the theoretical one.

2.3.1. Warping, defined

To understand warping in its simplest form, one can imagine dividing up the gel into several regions around each one of these pairs of spots so that for any given region, the local experimental landmark (brown circle) will be transformed to its predicted counterpart (blue square) by a translation specific to that neighborhood. Any experimental spot (including the landmark) within region 1, for instance, will undergo the same local translation defined by:

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_1 \qquad (4)$$
$$Y_{\text{pred}} = Y_{\text{exp}} + \Delta Y_1$$

where ΔX_1 and ΔY_1 are the components of the local translation needed to bring an experimental landmark onto its predicted counterpart. If the spot happens to be in region 3, then

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_3 \qquad (5)$$
$$Y_{\text{pred}} = Y_{\text{exp}} + \Delta Y_3$$

and so on.

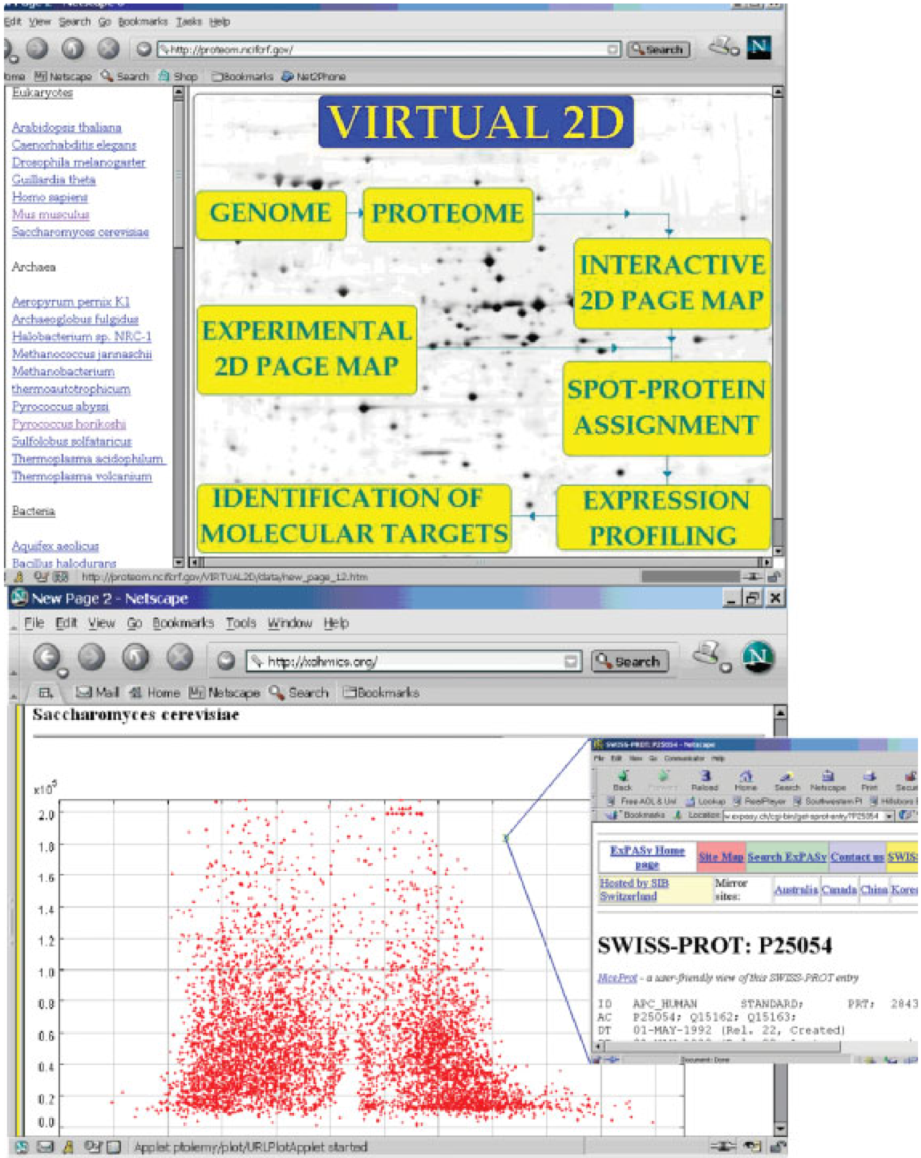
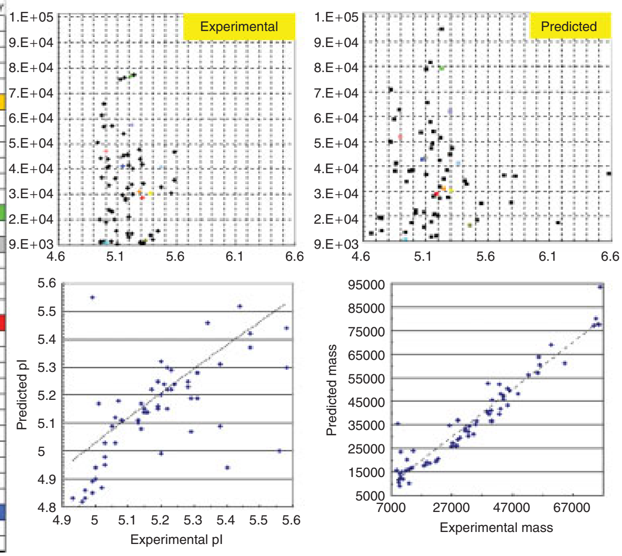


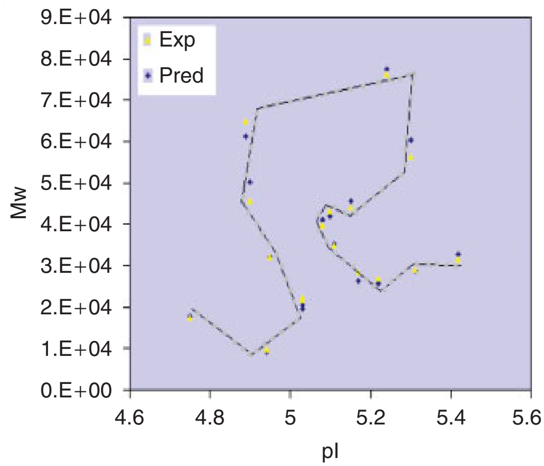
Figure 3 A snapshot of the screen display of VIRTUAL2D (**http://ncisgi.ncifcrf.gov/~medjahed**). Protein expression maps computed for 132 organisms/proteomes using data obtained from the European Bioinformatics Institute can be displayed by clicking on any of the entries on the left. On the fly interaction and identification. By using the controls, one can zoom in on a particular area. Simply moving the mouse over or clicking on any spot will either display a short description or bring up comprehensive information from the hyper-linked web server of choice (Protplot uses Java code modified from MicroArray Explorer; Lemkin *et al*., 2000)

Exp pI	Exp Mw	Ac	Pred pI	Pred Mw
4.96	9467	P00274	4.67	11675
5.44	13981	P00501	5.52	16971
5.03	39792	P00574	4.98	36512
5	45494	P00824	4.9	50194
5.34	10379	P00832	5.46	14937
5.38	28867	P00929	5.31	28724
4.99	64444	P02349	4.89	61158
5.1	8830	P02418	6.17	15769
5.31	40279	P02917	5.28	36744
5.19	31192	P02927	5.25	33368
5.23	38521	P02928	5.22	40707
5.58	45310	P02990	5.3	43182
5.21	75514	P02996	5.24	77450
5.22	32484	P02997	5.22	30292
5.13	43069	P04384	5.1	41952
4.97	59766	P04475	4.83	68984
5.56	28995	P04816	5	36911
5.13	73960	P05055	5.11	77101
4.99	9991	P05082	5.55	23586
5.31	27332	P05194	5.19	27467
5.19	44130	P05313	5.16	47522
5.22	12998	P05380	5.15	10387
4.99	55284	P06139	4.85	57138
5.32	9085	P06977	6.58	35401
5.22	56152	P08142	5.3	60441
5.17	26187	P08186	5.74	34916
5.15	43951	P08200	5.15	45757
5.2	40116	P08324	5.32	45524
4.95	18479	P08837	4.73	18120
5.17	26881	P09551	5.22	25785
5.32	32815	P11096	5.83	34358
5.14	39792	P11665	5.08	40987
5.16	74539	P13029	5.14	80024
5.06	55572	P16659	5.12	63734



(a)

Pi(Pred)	M(Pred)	Pi(exp)	M(Exp)	Ac
4.9	50194.23	4.99892	45494	P00824
5.31	28724.16	5.38	28936	P00928
4.89	61158.07	4.99	64780	P02349
5.24	77450.11	5.20869	76087	P02996
5.1	41951.63	5.13469	43079	P04384
5.3	60440.57	5.22236	56329	P08142
5.15	45756.71	5.15158	43958	P08200
5.22	25784.99	5.17088	26913	P09551
5.08	40987.03	5.1371	39816	P11665
5.42	32664.16	5.46889	31435	P16703
5.03	19572.36	5.06431	21653	P17288
4.95	32286.01	5.03085	32271	P25715
5.03	20630.25	5.02667	22398	P26427
4.94	9065.24	4.99784	9673	P26605
5.11	35088.05	5.08483	34775	P30148
4.75	17704.12	5	17510	P37901
5.17	26232.61	5.13871	28451	P39182



(b)

Figure 4 (a) Comparison of the values of isoelectric focusing points and molecular mass extracted from a high-resolution *E. coli* 2D-PAGE map downloaded from SWISS-2DPAGE and those computed in this work. In the two upper charts, a small number of corresponding data points from each set have the same color for a quicker visual inspection. (b) For a small subset of proteins, computed pI/Mw values are fairly close to the experimentalcounterparts, providing a ”constellation” of reference points that can be used for warping

For those areas without designated landmark such as region 2, one can interpolate using the translations from the surrounding neighborhoods.

$$X_{\text{pred}} = X_{\text{exp}} + \Delta X_2 \quad \text{where} \quad \Delta X_2 = \frac{\Delta X_1 + \Delta X_3 + \Delta X_6}{3} \tag{6}$$
$$X_{\text{pred}} = Y_{\text{exp}} + \Delta Y_2 \quad \text{and} \quad \Delta Y_2 = \frac{\Delta Y_1 + \Delta Y_3 + \Delta Y_6}{3}$$

The outcome of this two-dimensional alignment is not a trivial task as it is a function of several factors including the resolution of the experimental gel (the higher, the better) as well as the number and spatial distribution of landmark reference points. It involves working out the transformations that reflect the local distortions of the gel. Several software packages (Melanie, 2004; Z3, 2004;Delta 2D, 2004) currently existing on the market (as well as open source, e.g., **http://open2dprot.sourceforge.net/**) offer robust and flexible spot detection from many popular image file formats coupled with sophisticated statistical analysis and spot-pairing tools.

In the second example, we (arbitrarily) selected and downloaded from SWISS-2DPAGE, a map of human colorectal epithelia cells (Reymond *et al*., 1997). A quantitative measure of the discrepancy between the two data sets can be obtained by using the relative shift (r.s) of a protein spot between experimental and theoretical values.

$$r.s = \left[\left(\frac{\Delta pI}{pI_{\text{exp}}}\right)^2 + \left(\frac{\Delta Mw}{Mw_{\text{exp}}}\right)^2\right]^{1/2} \tag{7}$$

where

$$\Delta pI = pI_{\text{exp}} - pI_{\text{pred}} \text{ and } \Delta Mw = Mw_{\text{exp}} - Mw_{\text{pred}}$$

Despite the broad nominal intervals for pI (4–8 pH units) and Mw (0–200 kD), more than 66% of the predicted values have a relative shift less or equal to 0.12 compared to their observed counterpart. However, one must still face the reality of the numerous types of modifications occurring co- and posttranslationally, which can severely alter the electrophoretic mobility of the proteins affected. As can be seen in Figure 5, while relatively small local differences can easily be reconciled, no amount of warping will be able to totally and correctly align a collection of computed pI/Mw data points onto a set of experimentally determined protein spots, without individually identifying and incorporating the aforementioned corrections in the computation of these attributes.

3. TMAP (Tissue Molecular Anatomy Project)

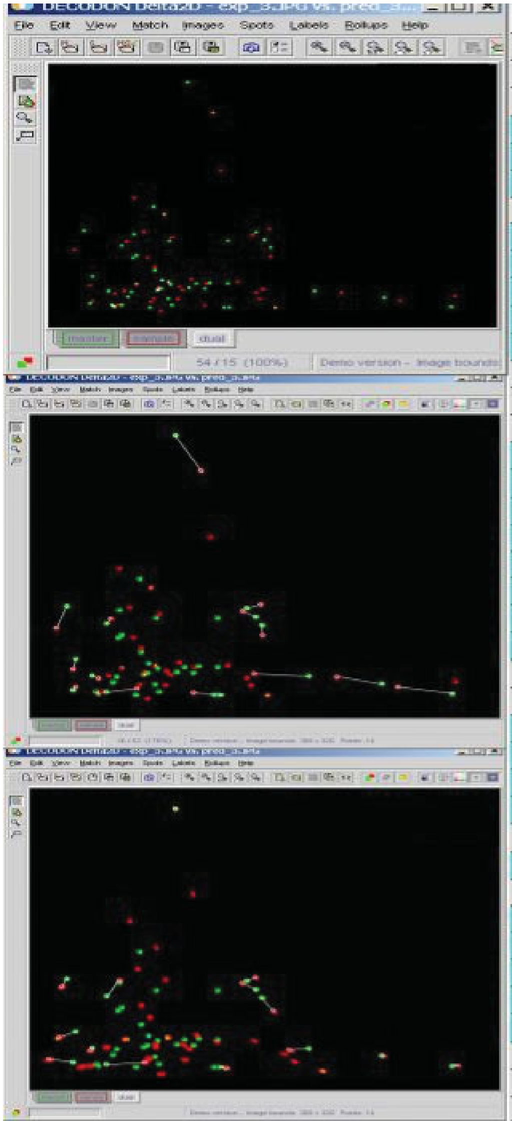
3.1. *Introduction*

By mining publicly accessible databases, we have developed a collection of tissue-specific predictive protein expression maps (PEM) as a function of cancer histological state. Data analysis is applied to the differential expression of gene

(a)

(b)

(c)



Exp pI	Exp Mw	Pred pI	Pred Mw	r.s
4.41	58700	4.29	46466	0.210
4.5	13360	4.53	15020	0.124
4.54	30228	4.5	24633	0.185
4.75	21019	4.84	19595	0.070
4.8	12166	4.88	12513	0.033
4.84	11163	4.82	11606	0.040
4.93	12076	5.34	14585	0.224
4.93	49118	5	51769	0.056
5.01	17462	5.37	16394	0.094
5.06	20750	4.97	21014	0.022
5.08	41446	5.29	41605	0.042
5.1	56317	5.24	57963	0.040
5.13	42930	5.1	79227	0.846
5.31	11038	9.48	17819	0.997
5.31	11038	5.46	8960	0.190
5.36	73173	5.51	68858	0.065
5.38	34062	5.38	35890	0.054
5.42	21346	5.44	22222	0.041
5.46	22765	5.44	23225	0.021
5.48	27623	5.56	26629	0.039
5.56	16289	5.7	15805	0.039
5.59	52522	5.61	54265	0.033
5.6	24976	5.13	24976	0.084
5.62	12121	7.03	10612	0.280
5.83	150908	6.17	132013	0.138
5.85	23060	5.71	23760	0.039
5.95	25694	5.88	28355	0.104
6.07	22824	5.77	21468	0.077
6.13	26432	6.08	25853	0.023
6.34	10683	6.81	15867	0.491
6.38	25826	6.51	26538	0.034
6.39	52160	6.28	95798	0.837
6.4	11038	6.07	11731	0.081
6.77	57695	6.95	59625	0.043
6.89	52776	6.71	56008	0.067
6.96	48310	6.97	42769	0.115
7.57	21182	6.86	22204	0.105
8.52	16844	7.93	19974	0.198
8.52	16844	6.76	20159	0.285
9.44	11331	8.73	15126	0.343

Figure 5 (a) Overlap of spots identified in 2D-PAGE map of human colorectal epithelial cell line (in green) and theoretically computed (in red). (b) Several pairs of corresponding experimentally predicted spots are connected to reflect the translations. (c) A global warping attempts to bring the computed value closer to the corresponding observed member of the pair. While in some cases, an almost exact local alignment is achieved, in many instances the differences caused by posttranslational modifications are simply too large to successfully align them. This analysis was carried out using a demonstration version of the Delta-2D package [27]

products in pooled libraries from the normal to the altered state(s). We wish to report the initial results of our survey across different tissues and explore the extent to which this comparative approach may help uncover panels of potential biomarkers of tumorigenesis that would warrant further examination in the laboratory. For the third dimension, we computed inferred gene-product translational expression levels from the transcriptional levels reported in the public databases. A number of studies have explored the feasibility of molecular characterization of the histopathological state from the mRNA abundance reported in public databases. Many potential tissue-specific cancer biomarkers were tentatively identified as a result of mining expression databases. Thus arose the motivation to explore and catalogue correlations across different tissues as a first step toward comparative cancer proteomics of normal versus diseased state. One potential clinical application is uncovering threads of biomarkers and therapeutic targets for multiple cancers.

3.2. Data mining

The Cancer Genome Anatomy Project (CGAP) (Strausberg *et al*., 2000) database of expressed sequence tags (EST), detected in different tissues, is accessible at **http://cgap.nci.nih.gov/**. It can be queried by possible histological state, source, extraction, and cloning method. In the initial construction of queries, selecting the option "ANY" from within all these fields provides an initial overview of the available libraries available. The more restrictive the search, the fewer libraries were selected. Within each library, transcripts are listed along with the number of times they were detected after a fixed number of PCR cycles. Since we were primarily interested in computing protein maps, the gene symbols associated with those ESTs that were clustered to a gene of known function were extracted from UNIGENE (UniGene, 2004). A Perl script performed the cross-reference checking between the two data sets and output a list of gene symbols and corresponding SWISS-PROT/trEMBL accession numbers (AC). The list of resulting AC was input to the pI/Mw tool server that computed the necessary pI (isoelectric focusing point) and molecular mass (Mw) for the mature, unmodified proteins. In the case of a single library, this information was married to the expression-detection counts in the following manner: The number of hits for each EST was first divided by the sum total of sequences within that library to provide a relative expression for each transcript. Finally, a renormalization was carried out by dividing relative expression levels by the maximum relative expression level. In the event a tissue search revealed several libraries fulfilling the requirements of the initial query, to improve the signal-to-noise ratio, the results are first pooled so as to generate a nonredundant list of entries and a more comprehensive expression map for that tissue and corresponding to that histological state. The flow chart is depicted in Figure 6.

3.3. ProtPlot

ProtPlot (Figure 7) is a Java-based data-mining software tool for virtual 2D gels. It was derived from Opensource MAExplorer project (**MAExplorer.sourceforge.net**)

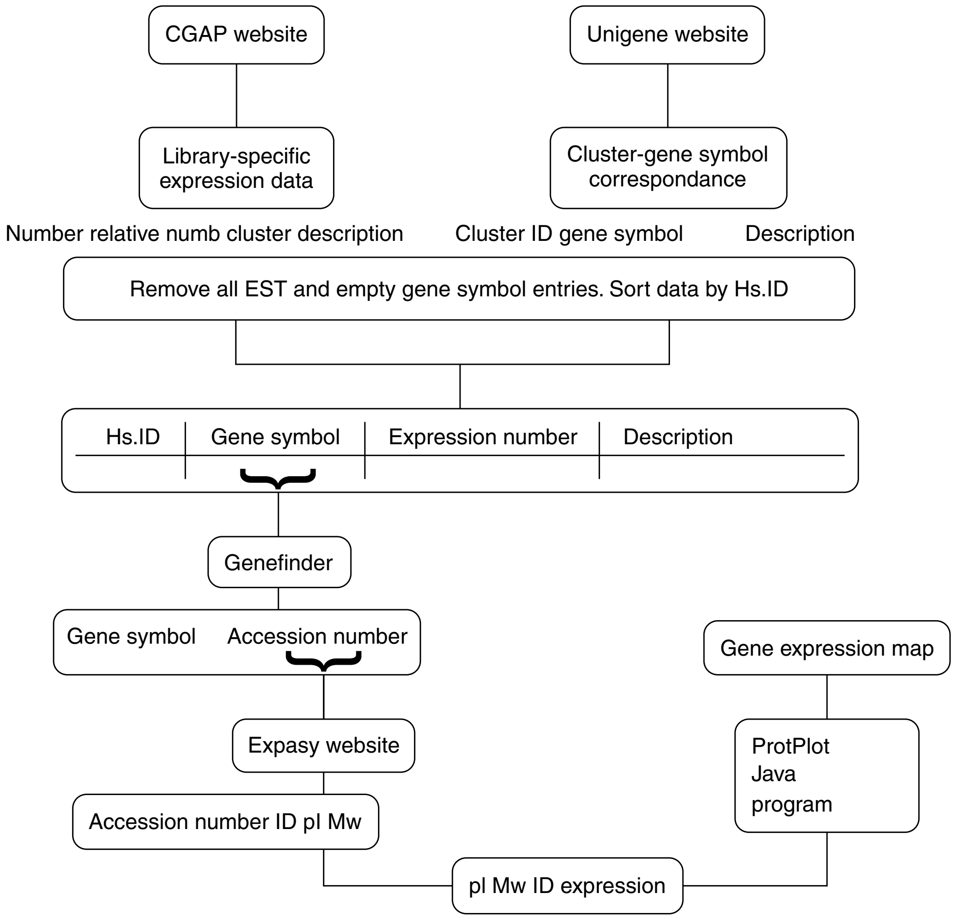


Figure 6 Flowchart describing in detail steps in the computation of expression maps

(Lemkin *et al*., 2000). It may be downloaded and run as a stand-alone application on one's computer. Its exploratory data analysis environment provides tools for the data mining of quantified virtual 2D gel (pIe, Mw, expression) data of estimated expression from the CGAP EST mRNA tissue expression database. This lets one look at the aggregated data in new ways: for example, which estimated "proteins" are in a specified range of (pI, Mw)? Or which sets of estimated "proteins" are up- or downregulated or missing between cancer samples and normal samples? Which sets or "proteins" cluster together across different types of cancers or normals? Here, one may aggregate several different normal and several different cancers as well as specify other filtering criteria.

As is well known, mRNA expression generally does not always correlate well with protein expression as seen in 2D-PAGE gels (Ideker *et al*., 2001). However, some new insights may occur by viewing the transcription data in the protein domain. If actual protein expression data is available for some of these tissues, it might be useful to compare mRNA estimated expression and actual protein

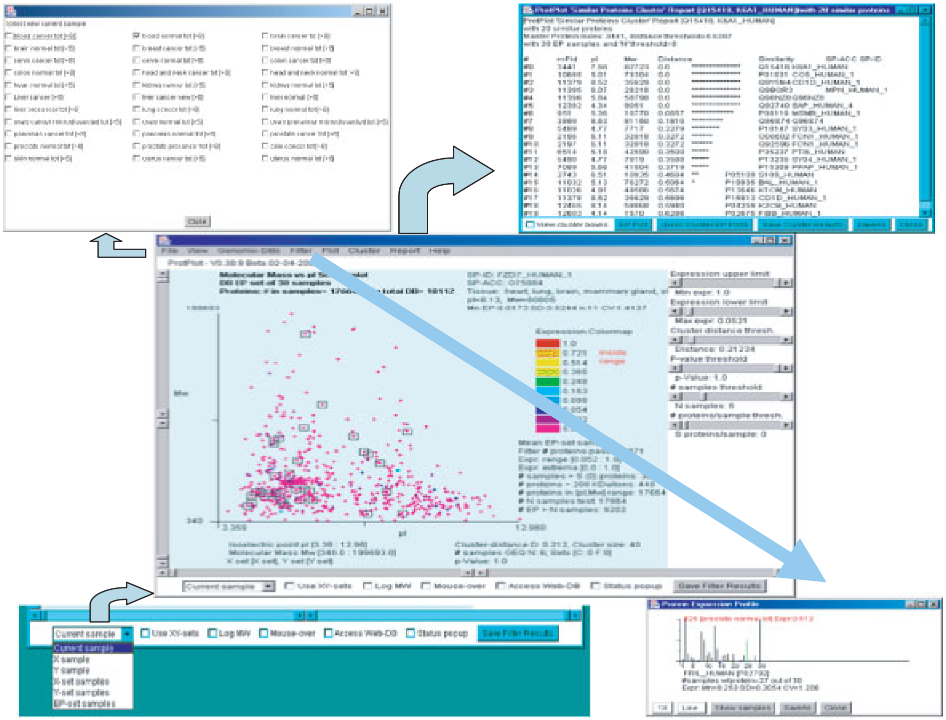


Figure 7 Snapshot of scatter plots from one sample in ProtPlot. It is also possible to create (bottom) an (X vs. Y) scatter plot or (Mean X-set vs. Mean Y-set) scatterplot when the corresponding ratio display mode is set. The following window shows the (Mean X-set vs. Mean Y-set) scatterplot; Tissue and Histology selection, panel b. This may be invoked either from the File menu or the pull-down sample selector at the lower-left corner of the main window. One can at glance obtain the expression profile of proteins or groups of proteins across tissues of choice. The small window illustrates the scrollable list of EP plots sorted by the current cluster report similarity. The spots marked by boxes belong to the same cluster

expression. This tool may help find those proteins with similar expression and those that have quite different expression. This might be useful in thinking about new hypotheses for protein postmodifications or mRNA posttranscription processing.

ProtPlot generates an interactive virtual protein 2D-gel map scatterplot based on the renormalized expression frequencies, derived from the ratios of observed counts versus maximum observed counts for each entry. This "hit" rate can be thought of as a *rough* estimate of gene expression. These ESTs were mapped to SWISS-PROT (**http://www.expasy.ch**) accession numbers and Ids, and the Mw and pI estimates were computed and used as estimates for corresponding proteins in a pseudo 2D-gel.

ProtPlot data is contained in a set of tissue- and histology-specific .prp (i.e., ProtPlot) files described in the data format documentation. These are kept in the PRP directory that comes with ProtPlot after installation. These .prp files can be updated from the ProtPlot Web server **http://tmap.sourceforge.net**.

3.3.1. Using ProtPlot for data mining virtual protein expression patterns

First, one needs to download and install ProtPlot Java program, preferably with the Java Virtual Machine(JVM) as well as the specially formatted CGAP data on a local computer. One starts ProtPlot by clicking on the “ProtPlot Startup” icon if one’s computer supports that (Windows, MacOS-X, etc.) or type ProtPlot on the command line for Unix, Linux, and other systems. Once ProtPlot is started, it loads all the library files (with a .prp extension) present within the PRP directory. The virtual protein data for each tissue is used to construct a Master Protein Index in which proteins will be present for some tissues and not for others. The data is presented in a pseudo 2D-gel image with the estimated isoelectric point (pI) on the horizontal axis and the molecular mass (Mw) on the vertical axis. Sliders on each of the axes allow one control the minimum and maximum values of pI and Mw displayed and thus the Mw versus pI scatterplot zoom region one wants to select. By clicking on a spot in the scatterplot, one will display information on that protein. One can also define that protein as the current protein. The current protein is used in some of the clustering methods, protein specific reports (Expression Profile report), and the Expression Profile plot. If one has enabled the pop-up Genomic-ID Web browser and one is connected to the Internet, it will pop up a Web page from the selected Genomic database for that protein. One then selects various options from the pull-down menus. Some of the more commonly used options are replicated as checkboxes at the bottom of the window.

3.3.2. The Scatterplot display mode

There are two primary types of pseudo 2D-gel (Mw vs. pI) scatterplot display modes (summarized in Table 2) of this derived protein expression data: expression mode or ratio mode. The expression data may be for a single sample (the current sample) or the mean expression of a list of samples (called the expression profile or EP). The ratio data is computed as the ratio of two individual samples called X and Y. Ratio data may alternatively be computed from sets of X samples and sets of Y samples. Generally, one would group a set of samples with similar characteristics together having the same condition (e.g., cancer, normal, etc.). The ratio of X and Y may be single samples in which case the ratio is computed as:

$$\text{ratio} = \frac{\text{expression X}}{\text{expression Y}} \tag{8}$$

Table 2 Summary of the four types of display modes

Display mode	Current sample	Single X/Y	X-set/Y-set	EP-set
Expression	Yes	No	No	No
Single samples ratio	No	Yes	No	No
X-set and Y-set samples ratio	No	No	Yes	No
Mean expression	No	No	No	Yes

where expression X (expression Y) is the expression of corresponding proteins. Alternatively, one may compute the ratio of the mean expression of two different sets of samples (the X set and the Y set). The X and Y sets may be thought of as experimental conditions and the members of the sets being “replicates” in some sense. For example, the X set could be cancer samples and the Y set could be normal samples. The ratio of the X/Y sets for each corresponding protein is computed as

$$\text{ratio} = \frac{\text{mean X-set expression}}{\text{mean Y-set expression}} \tag{9}$$

The following shows a screenshot of one of the (Mw vs. pI) scatterplots when the display mode was set to (X-set/Y-set) ratio mode.

3.3.3. Effect of display mode on filtering, clustering, and reporting

One selects the particular display mode using the Plot menu commands. When one selects a particular display mode, it will enable and disable Filter, View, Cluster, and Report options depending on the mode. For example, one may only use the *t*-test or missing X Y set test if one is in XY-sets ratio mode. One may only perform clustering if one is in EP-set mode. One may change the display mode using the (Plot menu | Show display mode) commands. Alternatively, since it is used so often, there is a checkbox at the bottom of the main window ”Use XY-sets” that will toggle between the XY-sets ratio mode and whatever the previous mode one had set.

3.3.4. Selecting samples

One selects samples for the current sample, X sample, Y sample, X-set samples, Y-set samples, and EP-set samples using a pop-up checkbox list chooser of all samples. For example, one may invoke this chooser for a specific tissue sample one wants to view by using the (File menu | Select samples | Select Current PRP sample) command. For X (Y) data, one invokes the choosers using (File menu | Select samples | Select X (Y) PRP sample(s)) command. One may switch between single (X/Y) and (X set/Y set) mode using the (File menu | Select samples | Use Sample X and Y sets else single X and Y samples [CB]) command.

There is an alternative display called the “Expression Profile” (EP) plot that displays a list of a subset of PRP samples for the currently selected protein. One may also display the scatterplot on the mean EP data for all proteins. The EP samples are specified using the (File menu | Select samples | Select Expression List of samples) command.

3.3.5. Listing a report on sample assignments

One may pop up a report of the current sample assignments for the: current sample single X sample, single Y sample, X sample set, Y sample set, and EP sample set using the (File menu | Select samples | List sample assignments) command.

3.3.6. Assigning the X-set and Y-set condition names

The default experimental condition names for the X and Y sample sets are “X set” and “Y set”. One may change these by the (File menu | Select samples | Assign X (Y) set name) commands.

3.3.7. Status reporting window

There is a status pop-up window that first appears when the program is started and reports the progress while the data is loading. After the data is loaded, it will disappear. One may bring it back at any time by toggling the “ Status popup” checkbox at the bottom of the window. One may also press the “Hide” button on the status pop-up window to make it disappear.

3.3.8. Data filtering

The pseudoprotein data is passed through a data filter consisting of the intersection of several tests including: pI range, Mw range, sample expression range, expression ratio(X/Y) range (either inside or outside the range), t-test comparing the X and Y sample sets, Kolmogorov–Smirnov test comparing the X and Y sample sets, missing proteins test for X and Y sample sets, tissue type filter, protein family filter (to be implemented), and clustering. The filtering options are selected in the Filter menu. If one is looking at the scatterplot in ratio mode, then one may filter by ratio of X/Y either inside or outside of the ratio range. The missing protein test defines missing as totally missing and present as having at least “N” samples present. Note that the t-test and the missing protein tests are mutually exclusive in what they are looking for, so using both results in no proteins being found. Currently, no false discovery rate correction is implemented in the data filters.

3.3.9. Saving filtered proteins in sets for use in subsequent data filtering

One may save the set of proteins created by the current data filter settings by pressing the “Save Filter Results” button in the lower-right of the main window. This set of proteins is available for use in future data filtering using the (Filter menu | Filter by AND of Saved Filter proteins [CB]) command. When one saves the state of the ProtPlot database (Filter menu | State | Save State), it will also write out the save protein sets (saved filtered proteins and saved clustered proteins) in the database “Set” folder with “.set” file name extensions. In the (Filter menu | State | Protein Sets) submenu, there are a number of commands to manipulate protein set files. One may individually save (or restore) any particular saved filtered set to (or from) a set file in the “Set” folder. There are also commands to compute the set intersection, union or difference between two protein set files and leave the resulting protein set in the saved filter set.

Table 3 Available filter options–display modes relationship

Filter name	Current sample	Single X/Y	X-set/Y-set	EP-set
>200 kDa	Yes	Yes	Yes	Yes
Tissue type	Yes	Yes	Yes	Yes
Expression (ratio) range	Expression	Ratio	Ratio	Expression
X/Y (inside/outside) range	No	Yes	Yes	No
(X-set, Y-set) t-test	No	Yes	Yes	No
(X-set, Y-set) KS-test	No	Yes	Yes	No
(X-set, Y-set) missing data	No	Yes	Yes	No
At most (least) N samples	No	No	Yes	Yes
AND of saved cluster set	Yes	Yes	Yes	Yes
AND of saved filter set	Yes	Yes	Yes	Yes

3.3.10. Filter dependence on the display mode

Note that the particular filter options available at any time depend on what the current display mode is. Table 3 shows which options are available for which display modes.

3.3.11. The data-mining "State"

The current data-mining settings of ProtPlot is called the "state". It may be saved in a named startup file called the "startup state file" in the "State" folder. The "State" folder and other folders used by ProtPlot are found in the directory in which one installed ProtPlot. Initially, there is no startup state file. If one saves the state, it creates this file. One may create as many of these saved state files as one wants to. One may change the file and thus save various combinations of settings of samples for the current, X, Y, and expression list of samples. The state also includes the various filter, view, and plot options as well as the pI, Mw, expression, ratio, cluster distance threshold, number samples threshold, p-value threshold sliders, as well as other settings. The saved Filter and Cluster sets of proteins are also written out as .set files in the "Set" folder when one saves the state. Starting ProtPlot by clicking on the ProtPlot startup icon will not read the state file when it starts up. However, if one has saved a state, clicking on the state file or a shortcut to the state file will cause it to be read when ProtPlot starts up. One may save the current state using either the (File | State | Save State) command to save it under the current name, or using either the (File | State | Save As State) command to save it under a new name one may specify. Then one may also change the current state using (File | State | Open Statefile) command.

3.3.12. The molecular mass versus pI scatterplot: expression or ratio

There are two types of scatterplots: expression for a single sample or the ratio of two samples X and Y. The Plot menu lets one switch the display mode. Ratio mode itself has two types of displays: red(X) + green(Y), or a ratio scale ranging between

<1/10 (green) and >10 (red). One may view a pop-up report of the expression or ratio values for the current protein. If "Mouse-over" is enabled, then moving the mouse over a spot will show the name of the protein and its associated data. If mouse-over is not enabled, then clicking on the spot will show its associated data. One may scroll the scatterplot in both the pI and Mw axes by adjusting the end-point scrollbars on the corresponding axes. In addition, the scatterplot can be displayed with a log transform of Mw by toggling the log Mw switch.

The pop-up plots and scatterplot may be saved as .gif image files that are put into the project's "Report" folder. Similarly, reports are saved as tab-delimited .txt text files in the "Report" folder. Because it prompts one for a file name, one may browse the entire file system and save the file in another disk location.

3.3.13. X sample(s) versus Y samples scatterplot

If one is in X/Y ratio mode (single X/Y samples or X-set/Y-set samples), one may view a scatterplot of the X versus Y expression data. The XY scatterplot can be enabled using the (Plot menu | Display (X vs Y) else (Mw vs pI) scatterplot - if ratio mode [CB]) command. One may zoom the scatterplot just as one does for the (Mw vs pI) scatterplot. The proteins displayed are those passing the data filter that have both X and Y data (i.e., expression is > 0.0).

3.3.14. Expression profile plot of a specific protein

An expression profile (EP) shows the expression for a particular protein for all samples that have that protein. The (Plot menu | Enable expression profile plot) pops up an EP plot window and displays the EP plot for any protein one selects by clicking on it. The relative expression is on the vertical axis and the sample number on the horizontal axis. Pressing on the "Show samples" button pops up a list showing the samples and their order in the plot. Pressing on the "nX" button will toggle through a range of magnifications from 1X through 50X that may be useful in visualizing low values of expression. Clicking on a new spot in the (Mw vs. pI) scatterplot will change the protein being displayed in the EP plot. Within the EP plot display, one may display the sample and expression value for a plotted bar by clicking on the bar (which changes to green with the value in red at the top). The EP plot can be saved as a GIF file. One may also click on the display to find out the value and sample. Note: since clustering uses the expression profile, the user must be in "mean EP-set display" mode.

3.3.15. Clustering of expression profiles

One may cluster proteins by the similarity of their expression profiles. First set the plot display mode to "Show mean EP-set samples expression data". The clustering method is selected from the Cluster menu. Currently, there is one cluster method. Others are planned. The cluster distance metric is the "distance" between two proteins based on their expression profile. The metric may be selected in the Cluster Menu. Currently, there is one clustering method: cluster proteins most similar to the current protein (specified by clicking on a spot in the scatterplot or using the Find

Protein by name in the Files menu). It requires one to specify (1) the current protein and (2) the threshold distance cutoff. The threshold distance is specified interactively by the ”Distance Threshold T” slider. The “Similar Proteins Cluster” Report will be updated if one changes either the current protein or the cluster distance.

The cluster distance metric must be computed in a way to take missing data into account since a simple Euclidian distance cannot be used with the type of sparse data present in the ProtPlot database. ProtPlot has several ways to compute the distance metric using various models for handling missing data. The set of proteins created by the current clustering settings can be saved by pressing the ”Save Cluster Results” button in the lower right of the cluster report window. This is available for use in future data filtering using the (Filter menu | Filter by AND of Saved Clustered proteins [CB]) command. When one saves the state of the ProtPlot database (Filter menu | State | Save State), it will also save the set of saved clustered proteins in the database ”Set” folder. One may restore any particular saved clustered set file.

One may bring up the EP plot window by clicking on the ”EP Plot” button and then click on any spot in the scatterplot to see its expression profile. Clicking on the ”Scroll Cluster EP Plots” button brings up a scrollable list of expression profiles for just the clustered proteins sorted by similarity. The proteins belonging to the cluster in the scatterplot with black boxes can be marked by selecting the “View cluster boxes” checkbox at the lower left of the cluster report window. This is illustrated in the following window.

3.3.16. Reports

Various pop-up report summaries are available depending on the display mode. All reports are tab-delimited and so may be cut and pasted into MS Excel or other analysis software. Reports also have a “Save As” button so one can save the data into a tab-delimited file. The default/Report directory is in the directory in which one installed ProtPlot. However, one may save it anywhere on one’s file system. The contents of some reports depend on the particular display mode. This is summarized in the Table 4.

3.3.17. Genomic databases

If one is connected to the Internet and has enabled ProtPlot to “Access Web-DB”, then clicking on a protein will pop up a genomic database entry for that protein. The particular genomic database to use is selected in the Genomic-DB menu.

4. Results and data analysis

Figure 8 depicts the pI/Mw maps computed by our approach for a number of these tissues. They all display the characteristic bimodal distribution that was explained previously as the statistical outcome of a limited, pK-segregated proteomic alphabet. In addition, one can quickly obtain the most significantly differentially expressed gene proteins by computing the tissue-specific charts of the ratios between normal and cancer states (Table 5).

Table 4 Characteristics of available reports as a function of display mode

Filter name	Current sample	Single X/Y	X-set/Y-set	EP-set
Statistics or proteins passing filter	SP-ACC/ID, pI, Mw, expression	SP-ACC/ID, pI, Mw, X/Y, X, Y expr, tissues	SP-ACC/ID, pI, Mw, mnX/mnY, (mn,sd,cv,n) expr for X- & Y-sets, tissues. If using t-test then (dF, t-stat, F-stat). If using KS-test then (dF, D-stat)	SP-ACC/ID, pI, Mw, (mn,sd,cv,n) expr for EP-set, tissues
Expression profiles of proteins passing filter	SP-ACC/ID, expr data EP-set	SP-ACC/ID, expr data EP-set	SP-ACC/ID, expr data EP-set	SP-ACC/ID, expr data EP-set
X & Y sets of missing proteins passing filter	No	No	SP-ACC/ID, (mn,sd,cv,n) for X- & Y-sets	No
EP set statistics of proteins passing filter	No	No	No	SP-ACC/ID, (mn,sd,cv,n) for EP-set
List of samples in current EP profile	{Nbr, sample-name, expression)	{Nbr, sample-name, expression)	{Nbr, sample-name, expression)	{Nbr, sample-name, expression)
List of all sample assignments	Current, X, Y, X-set, Y-set, EP-set	Current, X, Y, X-set, Y-set, EP-set	Current, X, Y, X-set, Y-set, EP-set	Current, X, Y, X-set, Y-set, EP-set
List of number of proteins/sample	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}	{Sample-name, # proteins in sample}
ProtPlot state	State	State	State	State

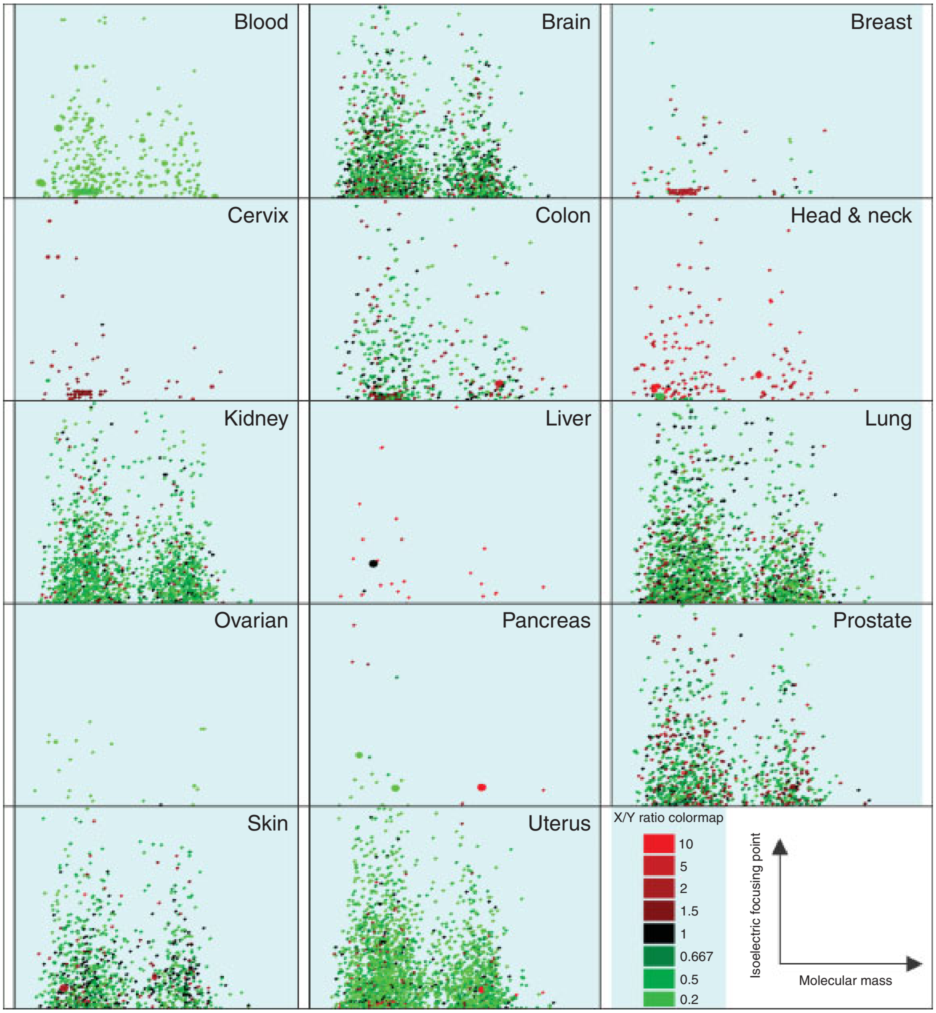


Figure 8 Tissue- and histology-specific pI/Mw interactive maps surveyed to date. The color code for scatter plots of the expression ratios (cancer /normal) is also shown in the figure

A number of proteins detected by the survey described are ribosomal or ribosomal-associated proteins such as elongation factors P04720, P26641 in colon and pancreas. Their upregulation is consistent with an accelerated cancerous cell cycle. Others may turn out to be effective tissue-specific *biomarkers* such as Phosphopyruvate hydratase (P06733 in skin). A third category will turn out to be druggable *targets.* Molecular *switches* that can be the focus of drug design for therapeutic intervention to reverse or stop the disease.

However, identification of useful potential targets requires additional knowledge of their function and cellular location. Accessibility is an obvious advantage. Such isthe case of Laminin gamma-2 (Q13753), the second highest, differentially

Table 5 For each tissue, gene products for which expression profile was the most significantly altered in the cancerous state as compared to the normal state

Blood		Brain		Breast	
Upregulated	Downregulated	Upregulated	Downregulated	Upregulated	Downregulated
	O00215	P04075	O00184	P02571	O43443
	P01907	P12277	O14498	P05388	O43444
	P01909	P41134	O15090	P12751	O60930
	P05120	P15880	O95360	P18084	O75574
	P35221	P12751	P01116	P49447	P15880
	P42704	P02570	P01118	Q05472	P17535
	P55884	P70514	P02096	Q15445	P19367
	Q29882	P99021	P20810	Q9BTP3	Q96HC8
	Q29890	Q11211	P50876	Q9HBV7	Q96PJ2
	Q99613	P46783	Q9BZZ7	Q9NZH7	Q96PJ6
	Q99848	P26373	Q9UM54	Q9UBQ5	Q9NNZ4
	Q9BD37	P26641	Q9Y6Z7	Q9UJT3	Q9NNZ5
Cervix		**Colon**		**Head & neck**	
Upregulated	Downregulated	Upregulated	Downregulated	Upregulated	Downregulated
	O75331	P00354	O14732	O75770	O60573
	O75352	P02571	P00746	P00354	O60629
	P09234	P04406	P09497	P04406	O75349
	P11216	P04687	P17066	P06702	P30499
	P13646	P04720	P18065	P09211	P35237
	P28072	P04765	P38663	P10321	P49207
	P47914	P09651	P41240	P21741	P82909
	Q02543	P11940	P53365	P30509	Q9BUZ2
	Q9NPX8	P17861	P54259	Q01469	Q9H2H4
	Q9UBR2	P26641	Q12968	Q92597	Q9H5U0
	Q9UQV5	P39019	Q9P1X1	Q9NQ38	Q9UHZ1
	Q9UQV6	P39023	Q9P2 R8	Q9UBC9	Q9Y3U8
Kidney		**Liver**		**Lung**	
Upregulated	Downregulated	Upregulated	Downregulated	Upregulated	Downregulated
O43257	O60622	P11021	P02792	O95415	O60441
O43458	Q14442	P11518		P01860	O75918
O75243	Q8WX76	P19883		P50553	O75947
O75892	Q8WXP8	P21453		P98176	O95833
O76045	Q96 T39	P35914		Q13045	P01160
Q15372	Q9H0 T6	P36578		Q15764	P04270
Q969 R3	Q9HBB5	P47914		Q92522	P05092
Q9BQZ7	Q9HBB6	Q05472		Q9BZL6	P05413
Q9BSN7	Q9HBB7	Q13609		Q9HBV7	P11016
Q9UIC2	Q9HBB8	Q969Z9		Q9NZH7	Q13563
Q9UPK7	Q9UK76	Q9BYY4		Q9UJT3	Q15816
Q9Y294	Q9UKI8	Q9NZM3		Q9UL69	Q16740

Table 5 (*continued*)

Ovarian		Pancreas		Prostate	
Upregulated	Downregulated	Upregulated	Downregulated	Upregulated	Downregulated
	P02461	P00338	P05451	O00141	O15228
	P02570	P02794	P15085	P08708	O43678
	P04792	P04720	P16233	P19013	P10909
	P07900	P05388	P17538	P48060	P11380
	P08865	P07339	P18621	Q01469	P11381
	P11142	P08865	P19835	Q01628	P98176
	P14678	P20908	P54317	Q01858	Q92522
	P16475	P26641	P55259	Q02295	Q92826
	P24572	P36578	Q92985	Q13740	Q99810
	Q15182	P39060	Q9NPH2	Q96HK8	Q9H1D6
	Q9UIS4	Q01130	Q9UIF1	Q96 J15	Q9H1E3
	Q9UIS5	Q15094	Q9UL69	Q9 C004	Q9H723
Skin		**Uterus**			
Upregulated	Downregulated	Upregulated	Downregulated		
O14947	O00622		O95432		
P01023	P12236		O95434		
P02538	P12814		O95848		
P06733	P19012		Q08371		
Q02536	P28066		Q13219		
Q02537	P30037		Q13642		
Q13677	P30923		Q9UKZ8		
Q13751	P33121		Q9UNK7		
Q13752	P36222		Q9UQK1		
Q13753	P43155		Q9Y627		
Q14733	Q01581		Q9Y628		
Q14941	Q9UID7		Q9Y630		

expressed protein in skin. It is thought to bind to cells via a high-affinity receptor and to mediate the attachment, migration, and organization of cells into tissues during embryonic development by interacting with other extracellular matrix components.

5. Conclusion

To date, the charts for 92 organisms have been assembled and are represented within VIRTUAL2D (**http://ncisgi.ncifcrf.gov/~medjahed**). TMAP (Medjahed *et al*., 2003a) (**http://tmap.sourceforge.net/**) results from the survey of most of the libraries within the CGAP public resource to produce a comprehensive list of putative gene products encompassing normal, cancerous, and when available, precancerous states for 14 tissues. These interactive, knowledge-based proteomics resources are regularly updated and available for download as open source to the research community to generate and explore in the laboratory hypothesis-driven cancer biomarkers.

Further reading

Anderson L and Seilhamer J (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 11853–11861.

Boguski MS and Schuler GD (1995) ESTablishing a human transcript map. *Nature Genetics*, **10**, 369–371.

Schuler GD (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, **75**(10), 694–698.

References

Aebersold R, Rist B and Gygi SP (2000) Quantitative proteome analysis: methods and applications. *Annals of the New York Academy Science*, **919**, 33–47.

Angelis FD, Tullio AD, Spano L and Tucci AJ (2001) Mass spectrometric study of different isoforms of the plant toxin saporin. *Mass Spectrometry*, **36**(11), 1241–1248.

Bjellqvist B, Sanchez JC, Pasquali C, Ravier F, Paquet N, Frutiger S, Hughes GJ and Hoschstrasser DF (1993) Micropreparative two-dimensional electrophoresis allowing the separation of samples containing milligram amounts of proteins. *Electrophoresis*, **14**, 1375–1378.

Bussow K (2001) Protein in gels, computers, crystals and camels. *Trends Biotechnol*, **19**(9), 328–329.

Delta 2D (Decodon) (2004) **http://www.decodon.com/**, (2004 version).

Dihazi H, Kessler R and Eschrich K (2001) In-gel digestion of proteins from long-term dried polyacrylamide gels: matrix-assisted laser desorption-ionization time of flight mass spectrometry identification of proteins and detection of their covalent modification. *Analytical Biochemistry*, **299**(2), 260–263.

European Bioinformatics Institute (EBI) (2004) **http://www.ebi.ac.uk**.

Fearon ER, Hamilton SR and Volgeinstein B (1987) Clonal analysis of human colorectal tumors. *Science*, **238**, 193–197.

Fivaz M, Vilbois F, Pasquali C and van der Goot FG (2000) Analysis of glycosyl phosphatidylinositol-anchored proteins by two-dimensional gel electrophoresis. *Electrophoresis*, **21**(16), 3351–3356.

Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildruber R and Weiss W (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **6**, 1037–1053.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L, *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Kriegel K, Seefeldt I, Hoffmann F, Schultz C, Wenk C, Regitz-Zagrosek V, Oswald H and Fleck E (2000) An alternative approach to deal with geometric uncertainties in computer analysis of two-dimensional electrophoresis gels. *Electrophoresis*, **13**, 2637–2640.

Lemkin PF, Thornwall G, Walton K and Hennighausen L (2000) The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland. *Nucleic Acids Research*, **22**, 4452–4459; **http://www-lecb.ncifcrf.gov/MAExplorer/**.

Medjahed D, Luke BT, Tontesh TS, Smythers GW, Munroe DJ and Lemkin PF (2003a) Tissue Molecular Anatomy Project (TMAP): an expression database for comparative cancer proteomics. *Proteomics*, **3**(8), 1445–1453.

Medjahed D, Smythers G, Stephens M, Powell D, Lemkin P and Munroe JD (2003b) VIRTUAL2D: A web-accessible predictive database for proteomics analysis. *Proteomics*, **2**.

Melanie (Geneva Bioinformatics) (2004) **http://www.www.expasy.ch/melanie/**, (2004 version).

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological*, **250**, 4007–4021.

O'Farrell PZ, Goodman HM and O'Farrell PH (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell*, **12**, 1133–1141.

Page MJ, Amess B, Townsend RR, Parekh R, Herath A, Brusten L, Zvelebil MJ, Stein RC, Waterfield MD, Davies SC and O'Hare MJ (1999) Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *Cell Biology*, **96**(22), 12589–12594.

Reymond MA, Sanchez J-C, Hughes GJ, Riese J, Tortola S, Peinado MA, Kirchner T, Hohenberger W, Hochstrasser DF and Kockerling F (1997) Phenotypic analysis in colorectal carcinoma: an international interdisciplinary project. *Electrophoresis*, **18**, 2842–2848.

SPSS (Statistical Package for the Social Services) (2004) **http://www.spss.com/**; (version 8.0, 2004)

Strausberg RL, Buetow KH, Emmert-Buck M and Klausner R (2000) The Cancer Genome Anatomy Project: building an annotated gene index. *Trends in Genetics*, **16**, 103–106.

SwissProt, a protein knowledgebase can be accessed at: (2004) **http://www.expasy.ch/**.

UniGene (NCBI) (2004) **http://www.ncbi.nlm.nih.gov/UniGene** (data from 2004)

Weiller GF, Djordjevic MJ, Caraux G, Chen H and Weinman JJ (2001) A specialised proteomic database for comparing matrix-assisted laser desorption/ionization-time of flight mass spectrometry data of tryptic peptides with corresponding sequence database segments. *Proteomics*, **12**, 1489–1494.

WORLD-2DPAGE (ExPosy) (2004) **http://www.expasy.ch/ch2d/2d-index.htm**.

Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS and Petricoin EF 3rd (2000) New approaches to proteomic analysis of breast cancer. *Proteomics*, **10**, 1205–1215.

Z3 (Compugen) (2004) **http://www.2dgels.com/**,(2004, version).

Short Specialist Review

Disease gene candidacy and ESTs

Mark I. McCarthy
University of Oxford, Oxford, UK

The effort to identify and characterize the genetic variants that underlie susceptibility to common multifactorial traits represents one of the main challenges in biomedical research today. Impressive success in mapping genes underlying Mendelian traits (Peltonen and McKusick, 2001) has, until recently, proved difficult to translate to complex multifactorial traits such as diabetes, heart disease, and asthma (McCarthy, 2004). For such traits, individual predisposition is influenced by variation at many genomic sites, acting in concert with diverse environmental exposures (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2). The contribution made by any one of these factors, considered alone, is generally modest, with the result that detection has only become feasible with the accumulation of large sample sizes and high-throughput technology, coupled with recent advances in genomic information and bioinformatics.

Several different methodologies are used in the search for complex trait susceptibility variants – including linkage analysis and association studies – but all of these, at some stage, require assessment of the relative biological candidacy of a list of genes (Collins, 1995). Such assessments may be based on the full complement of human genes – as in classical candidate gene studies. Alternatively, they may be implemented following a preliminary screen (for example, a genome-wide scan for linkage, or a transcriptional profiling study), which has focused attention on some subset of genes defined in terms of their chromosomal location and/or their transcriptional modulation by pathological and/or physiological perturbation. Typically, the regions identified following a genome-wide linkage scan cover as much as 1% of the genome, and can be expected to contain several hundred genes (Kruglyak and Lander, 1995). Further efforts to map the variant (or variants) responsible typically require further stages of prioritization amongst the genes on this "shortlist", based on the likelihood that each might plausibly be involved in disease pathogenesis (McCarthy *et al*., 2003).

Such prioritization is particularly difficult when, as with most multifactorial traits, the mechanisms underlying disease development are only poorly understood. Transcriptional profiling, proteomic and other novel methods are steadily improving our capacity to identify pathways apparently disturbed during disease development, but it can be difficult to attribute causality. Given these reservations, expression pattern information can play a valuable role in candidate gene prioritization. For many multifactorial traits, there is a reasonable basis for designating the tissues

and/or cell types most likely to be implicated in disease development. By matching this information to the tissue expression profiles of genes of interest, it becomes possible, in principle at least, to highlight putative susceptibility genes.

Ideally, such analyses would benefit from an explicit and exhaustive inventory of the transcriptional repertoire – including patterns of alternative splicing – of every tissue and cell type, basally and in response to physiological and pathological perturbation, and at all stages of development and aging. Despite the rapid accumulation of data from microarray studies and other sources, such a complete view remains some way off. In the meantime, EST (expressed sequence tags) data (and other transcript resequencing data derived from full-length clones or SAGE – serial analysis of gene expression – analysis) can provide some valuable indications of gene expression profiles. The ESTs catalogued in databases such as dbEST (**http://www.ncbi.nlm.nih.gov/dbEST/**) carry information not only on sequence, but on the library from which they were derived, information that typically includes an anatomical location and in some cases associated data on pathology, cell type, and (less so in humans) developmental stage (*see* Article 78, **What is an EST?**, Volume 4).

Two of the main obstacles to the use of such EST data have been overcome in recent years. First, the task of combining multiple EST reads into single transcripts has been tackled by a variety of clustering algorithms (Burke *et al*., 1999; Pertea *et al*., 2003; *see also* Article 88, **EST clustering: a short tutorial**, Volume 4), and more recently, by the availability of an increasingly complete human genome sequence against which ESTs can be aligned, as well as by a move toward sequencing of full-length clones (Imanishi *et al*., 2004). The second barrier to the use of EST data had been the absence of any systematic description of the libraries from which ESTs were derived. For example, transcripts expressed in pancreatic beta-cells might have been present in libraries whose origin was recorded under a range of valid but nonstandardized sobriquets – for example, “pancreas”, “islet”, “endocrine pancreas”, “beta-cell”, “b-cell”, and “insulinoma”. The solution here has been to map all of these descriptions onto controlled terms organized into hierarchical ontologies covering anatomical site and other descriptive dimensions (*see* Article 79, **Introduction to ontologies in biomedicine: from powertools to assistants**, Volume 8), thereby making EST (and related expression) data accessible for database queries. The eVOC system, for example, provides such a systematic description for over 7000 cDNA and SAGE libraries used as sources for expression data in terms of Anatomical Site, Cell Type, Pathology, and Developmental Stage (Kelso *et al*., 2003). eVOC terms have been incorporated within the ENSMART facility at ENSEMBL (**http://www.ensembl.org/Multi/martview**), allowing researchers to readily include expression state – along with chromosomal location and gene function – in their assessments of biological candidacy (Kasprzyk *et al*., 2004).

Nevertheless, several important limitations to the use of such data remain. Most importantly, EST (and other transcript sequencing) data are only available from a sporadic and incomplete range of tissues, with several major tissues (such as fat) poorly represented. In addition, the depth of sequencing varies from tissue to tissue, and even though much EST sequencing has been performed on normalized libraries, for all but a handful, transcript representation is far from being comprehensive. For

related reasons, EST data are semiquantitative at best: SAGE-derived transcriptional data may be more useful when a quantitative view is desired (Cras-Méneur *et al*., 2004).

Transcript sequencing data have proven most useful in the search for susceptibility genes for those diseases where the key pathophysiological events can be confidently localized to a particular tissue, especially where that tissue has a highly specialized function. Obvious examples include many monogenic causes of blindness related to retinal degeneration (Katsanis *et al*., 2002; Sohocki *et al*., 1999), and deafness arising from cochlear pathology (Skvorak *et al*., 1999); in such cases, expression information has contributed significantly to successful gene identification (Sullivan *et al*., 1999). Examples from complex traits are harder to define, not least because it is rarely possible to be so precise about the tissues involved in disease pathogenesis. In the case of type 2 diabetes for example, there are strong grounds for suspecting that fundamental molecular events might occur in a range of tissues (muscle, fat, liver, brain, pancreas): while this information can still help refine a list of positional candidates, the scope for pinpointing one or two key candidates on the basis of expression data alone is clearly modest. One of the few attempts to make systematic use of expression state information for a complex trait relates to Parkinson's disease, where, unusually, the lesional site is well defined (Hauser *et al*., 2003).

The value of EST data as a means to infer the expression distribution of transcripts will undoubtedly diminish with the advent of alternative approaches that are able to deal more effectively with questions of representation and quantification (notably, those using cDNA and/or oligonucleotide microarrays). However, it is important also to recognize the potential limitations of such hybridization-based methods: for example, the danger of cross-hybridization between closely related sequences, and the possibilities for misinterpretation that can result from alternative splicing.

Looking forward, there is no doubt that positional cloning efforts, in monogenic and complex traits alike, will make increasing use of expression state information, along with diverse other sources of data, to assist in the difficult task of assessing the biological candidacy of the long lists of genes that emerge from genome-wide approaches to susceptibility gene identification.

References

Burke J, Davison D and Hide W (1999) d2_cluster: A validated method for clustering EST and full-length cDNA sequences. *Genome Research*, **9**, 1135–1142.

Collins FS (1995) Positional cloning moves from the perditional to traditional. *Nature Genetics*, **9**, 347–350.

Cras-Méneur C, Inoue H, Zhou T, Ohsugi M, Bernal-Mizrachi E, Pape D, Clifton SW and Permutt MA (2004) An expression profile of human pancreatic islet mRNAs by serial analysis of gene expression (SAGE). *Diabetologia*, **47**, 284–299.

Hauser MA, Li Y-J, Takeuchi S, Walters R, Noureddine M, Maready M, Darden T, Hulette C, Martin E, Hauser E, *et al*. (2003) Genomic convergence: Identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Human Molecular Genetics*, **12**, 671–676.

Imanishi T, Itoh T, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Suzuki Y, *et al*. (2004) Integrative annotation of 21037 human genes validated by full-length cDNA clones. *PLOS Biology*, **2**, 1–20.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T and Birney E (2004) EnsMart: A generic system for fast and flexible access to biological data. *Genome Research*, **14**, 160–169.

Katsanis N, Worley KC, Gonzalez G, Ansley SJ and Lupski JR (2002) A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. *Proceedings of the National Academy of Sciences United States of America*, **99**, 14326–14331.

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien-Kruger S, Smedley D, Otgaar D, Greyling G, Jongeneel V, McCarthy MI, *et al*. (2003) eVOC: A controlled vocabulary for unifying gene expression data. *Genome Research*, **13**, 1222–1230.

Kruglyak L and Lander ES (1995) High-resolution genetic mapping of complex traits. *American Journal of Human Genetics*, **56**, 1212–1223.

McCarthy MI (2004) Progress in defining the molecular basis of type 2 diabetes through susceptibility gene identification. *Human Molecular Genetics*, **13**(Suppl 1), R33–R41.

McCarthy MI, Smedley D and Hide W (2003) New methods for finding disease susceptibility genes: Impact and potential. *Genome Biology*, **4**, 119-1–119-8.

Peltonen L and McKusick V (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, **291**, 1224–1229.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, *et al*. (2003) TIGR Gene Indices clustering tools (TGICL: A software system for fast clustering of large EST datasets). *Bioinformatics*, **22**, 651–652.

Skvorak AB, Weng Z, Yee AJ, Robertson NG and Morton CC (1999) Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness. *Human Molecular Genetics*, **8**, 439–452.

Sohocki MM, Malone KA, Sullivan LS and Daiger SP (1999) Localization of retina/pineal-expressed sequences: Identification of novel candidate genes for inherited retinal disorders. *Genomics*, **58**, 29–33.

Sullivan LS, Heckenlively JR, Bowne SJ, Zuo J, Hide WA, Gal A, Denton M, Inglehearn CF, Blanton SH and Daiger SP (1999) Mutations in a novel retina-specific gene cause autosomal dominant retinitis pigmentosa. *Nature Genetics*, **22**, 255–259.

Short Specialist Review

The role of nonsense-mediated decay in physiological and pathological processes

Jill A. Holbrook, Gabriele Neu-Yilik and Andreas E. Kulozik
University of Heidelberg, Heidelberg, Germany
Molecular Medicine Partnership Unit, Heidelberg, Germany

Matthias W. Hentze
Molecular Medicine Partnership Unit, Heidelberg, Germany
European Molecular Biology Laboratory, Heidelberg, Germany

1. Introduction

In eukaryotes, a conserved surveillance pathway known as nonsense-mediated decay (NMD) regulates the abundance of mRNAs containing premature termination codons (PTCs), defined as in-frame stop codons located upstream of the physiological stop codon. PTCs often arise as the result of pathological problems: to name just a few examples, insertions or deletions in DNA that change the open reading frame almost invariably lead to multiple PTCs downstream of the frameshift; deamination of methylcytosine in CG sequences causes transformation of an arginine codon (CGA) into a stop codon (TGA); or splice site mutations may result in the inclusion of an out-of-frame intronic fragment that introduces PTCs. In addition, PTCs can occur in normal transcripts as a result of physiological processes, such as use of alternative open reading frames, presence of UGA codons encoding selenocysteine, or posttranscriptional editing. NMD appears to have evolved as a means both of controlling expression of physiological transcripts containing PTCs and of limiting production of protein from abnormal PTC-containing transcripts. In this article, we describe the mechanism of NMD and briefly discuss its roles in normal cellular function and in some forms of genetic disease.

2. The molecular mechanism of NMD

In order to selectively degrade PTC-containing transcripts, the NMD machinery must distinguish between a normal stop codon and an abnormal one. Both splicing

and translation appear to be critical for this discrimination in mammals. That splicing is involved in PTC recognition is suggested by the fact that NMD-activating termination codons are generally located at least 50 nucleotides upstream of the last exon–exon junction (Figure 1). Such observational evidence has been confirmed by experiments in which insertion of an intron downstream of a physiological termination codon results in transcript degradation, and by showing that intronless PTC-containing transcripts are immune to NMD. Translation is also central to NMD, as demonstrated by a multitude of experiments in which PTC-containing transcripts are stabilized by interference with normal translational mechanisms. Taken together, these data on splicing and translation indicate that a stop codon is judged as normal or abnormal depending on its position relative to exon–exon junctions and that this positional information is shared between the splicing and translational machinery.

On a molecular level, communication appears to occur through a marker deposited at exon–exon junctions during splicing. This marker has been identified as a dynamic assembly of proteins, known as the exon junction complex (EJC), which is deposited in a sequence-nonspecific manner ∼20–24 nucleotides 5′ to every exon–exon junction (Figure 1). The EJC is involved in cellular processes in addition to NMD, and many of its constituent proteins are exchanged or removed before a transcript exits the nucleus. However, as would be expected, at least some components (proteins Y14/MAGOH, RNPS1, and eIF4AIII) persist into the cytoplasm to tag the location of the exon–exon junction. During translation, if no stop codon is located upstream of an EJC marker, the transcript is validated as normal, presumably as a result of removal of the EJC proteins. However, if a stop codon is identified upstream of an EJC complex, the persisting downstream EJC components appear to recruit or retain human NMD proteins known as UPF1, UPF2, and UPF3 (homologs of yeast NMD factors) and other protein factors (Barentsz and P29) involved in NMD, cooperating with them to trigger mRNA degradation. The protein UPF1, in particular, appears to be central in NMD. UPF1 provides a link between translation and termination, since it associates with both ribosomes and release factors. Furthermore, UPF1 is a target of regulation, undergoing an essential cycle of phosphorylation and dephosphorylation mediated by a group of factors known as SMG proteins (homologs of *Caenorhabditis elegans* NMD factors). Exactly how these interactions lead to transcript decay, however, is not yet well understood. Detailed descriptions of the mechanism of NMD, and primary literature references are contained in recent reviews (Schell *et al*., 2002; Singh and Lykke-Andersen, 2003; Wilkinson, 2003; Holbrook *et al*., 2004; Maquat, 2004).

It is important to note that the current mechanistic understanding of NMD is incomplete. Many transcripts do not behave as expected on the basis of the current model (reviewed in Holbrook *et al*., 2004). Furthermore, NMD never reduces transcript levels to zero, and residual mRNA levels usually range from 10 to 30% of wild-type levels. The fate of such residual transcripts is unclear: some evidence indicates that translation occurs, which could potentially result in biologically significant expression of truncated protein (Bamber *et al*., 1999; Donnadieu *et al*., 2003).

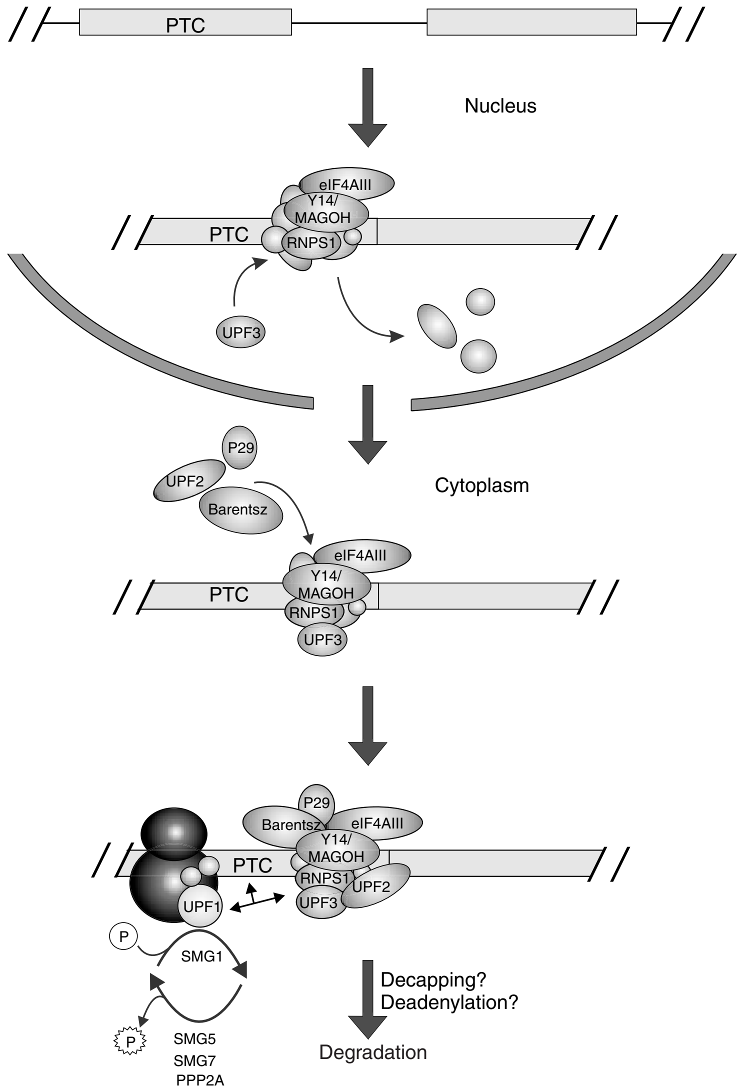


Figure 1 Model of nonsense-mediated decay. (Note: protein–protein interactions as shown in this diagram are meant to be schematic.) During splicing in the nucleus, the exon junction complex (EJC) is deposited in a nonspecific manner 20–24 nucleotides 5′ of exon–exon junctions. The EJC is initially composed of a number of proteins including NMD factors RNPS1, eIF4AIII (also known as Ddx48), and the Y14/MAGOH heterodimer (also known as Rbm8a). During maturation and export of the mRNA and transport through the nuclear pore, the EJC is remodeled and several proteins are released; others, probably including UPF3, are recruited. Both RNPS1 and Y14/MAGOH remain associated with the EJC and accompany the mRNA into the cytoplasm, where other proteins including UPF2, Barentsz (also known as Casc3), and P29 likely join the complex. During translation, the EJCs are thought to be stripped off the mRNA by the translating ribosome, and the mRNA is validated as "error-free" if no EJC is encountered downstream of a stop codon. However, if the translating ribosome (black) encounters a stop codon at least 50 nucleotides upstream of at least one EJC, as demonstrated by the PTC in the figure, NMD is triggered. The protein factor UPF1, which is of central importance in NMD, possibly interacts with the ribosome and release factors (small grey spheres). A cycle of UPF1 phosphorylation (mediated by protein SMG1) and dephosphorylation (involving proteins SMG5, SMG7, and protein phosphorylase 2 A) is essential for NMD to occur. The details of interactions between the ribosome, release factors, UPF1, the PTC, the EJC, and degradation factors are not yet well understood

3. Involvement of NMD in normal gene expression

As described in the Introduction, PTCs may arise in normal transcripts as a result of physiological processes, and one of the important roles of NMD appears to be to help balance the expression of protein or RNA isoforms of these PTC-containing transcripts. To date, the most important and widespread mechanism for introducing PTCs into normal transcripts appears to be alternative splicing (*see* Article 23, **Alternative splicing in humans**, Volume 7). Alternative splicing is potentially capable of producing thousands of PTC-containing transcripts in a normal cell. Unsurprisingly, a large proportion of these transcripts should be eligible to undergo NMD (Lewis *et al*., 2003). In line with these conjectures, a role for NMD in controlling the amounts and/or proportions of splice products has been demonstrated for a number of genes (Sureau *et al*., 2001; Gouya *et al*., 2002; Green *et al*., 2003; Lamba *et al*., 2003; Wollerton *et al*., 2004), most significantly in the case of particular splice factors that appear to autoregulate their abundance through NMD.

Further contributions from NMD to gene expression are likely in other physiologically important systems. For example, in B and T cells, the immunoglobulin and T-cell receptor genes undergo rearrangement, commonly introducing PTCs, and these transcripts are degraded by NMD. Therefore, NMD probably allows protein expression only from successfully rearranged genes, thereby ensuring a fully functioning antigen response (reviewed in Li and Wilkinson, 1998). An intriguing potential role of NMD in telomere physiology is suggested by the involvement in telomere maintenance (Reichenbach *et al*., 2003; Snow *et al*., 2003) of the human homolog of a *C. elegans* NMD protein (Chiu *et al*., 2003). Other physiological processes in which NMD or its components appear to be involved include p53 phosphorylation in response to genotoxic stress (Brumbaugh *et al*., 2004), as well as normal growth and development (Medghalchi *et al*., 2001), production of small nucleolar mRNAs (reviewed in Ruiz-Echevarria *et al*., 1996), and regulation of selenoprotein mRNAs (reviewed in Maquat, 2004).

4. Protective effects of NMD in hereditary and acquired genetic disorders

In addition to its physiological roles, NMD appears to limit production of faulty proteins by degrading pathological (i.e., potentially disease-causing) PTC-containing transcripts. Such transcripts might otherwise result in production of truncated proteins that act in a dominant negative fashion. NMD, therefore, protects against the disease arising from expression of such deleterious mRNAs. Such a protective role of NMD was first demonstrated in β-thalassemia, a condition arising from mutations in β-hemoglobin. Normal erythroid cells contain hemoglobin tetramers composed of two α- and two β-globin subunits. The common recessive form of β-thalassemia is often caused by β-globin mutations that result in production of PTC-containing NMD-sensitive transcripts. These transcripts are degraded by NMD, limiting truncated β-globin synthesis (Hall and Thein, 1994), and the resultant excess of free α-globin, which is harmful to the cell, is

degraded proteolytically. Persons homozygous for these PTC mutations produce little β-hemoglobin and are severely anemic. However, persons heterozygous for such mutations generally synthesize enough β-globin from the remaining normal allele to maintain near-normal hemoglobin levels, and are therefore healthy. Rare NMD-insensitive last-exon PTC mutations, in contrast, give rise to truncated, nonfunctional β-globin that overwhelms the cell's proteolytic system and causes toxic precipitation of insoluble globin chains (Thein *et al*., 1990). The remarkable contrast between asymptomatic heterozygotes with NMD-competent mutations and affected heterozygotes with NMD-incompetent mutations demonstrates that NMD protects most heterozygous carriers from developing dominant β-thalassemia (Kugler *et al*., 1995).

Analogous to the situation described for β-thalassemia, disease-modulating effects of NMD can explain genotype/phenotype relationships in a number of other genetic conditions (Holbrook *et al*., 2004). Furthermore, in acquired genetic disorders, NMD appears to similarly limit expression of truncated, mutant tumor suppressor genes that could give rise to harmful dominant proteins. NMD may thus inhibit cancer development in heterozygotes with a remaining intact tumor suppressor allele (reviewed in Holbrook *et al*., 2004).

5. Therapies in development for treatment of PTC-related disease

In contrast to the protective role of NMD delineated above, NMD can also contribute to disease phenotypes. This occurs when NMD destroys a PTC-containing transcript that codes for a C-terminal truncated protein that, if expressed, would be partly or fully functional. The resultant deficiency in functioning gene product could theoretically be alleviated by preventing NMD-induced transcript degradation. For genetic conditions in which PTC-containing transcripts could produce functional protein – including cystic fibrosis, Duchenne muscular dystrophy, Hurler syndrome, and X-linked nephrogenic diabetes insipidus – interventions to prevent transcript degradation are under development.

The most widely tested approach to allow production of protein from a PTC-mutated mRNA involves the use of aminoglycoside antibiotics. These drugs bind to the decoding center of the ribosome and decrease the accuracy requirements for codon–anticodon pairing, resulting in stop codon readthrough. Therefore, instead of chain termination, an amino acid is incorporated into the polypeptide chain and full-length (although missense-mutated) protein is synthesized. Aminoglycoside treatment has been shown to result in some full-length protein expression and resultant functional improvement in cell lines (Bedwell *et al*., 1997; Barton-Davis *et al*., 1999; Keeling *et al*., 2001), and in most cases, in animal models (Barton-Davis *et al*., 1999; Du *et al*., 2002; Sangkuhl *et al*., 2004). In addition, trials of aminoglycoside therapy have been carried out in humans with PTC mutations, and some promising results have been reported in a subgroup of cystic fibrosis patients in whom treatment resulted in some protein production (Wilschanski *et al*., 2000; Clancy *et al*., 2001; Wilschanski *et al*., 2003). In contrast, two very small clinical studies of patients with muscular dystrophy have been less encouraging,

showing no measurable functional improvement (Wagner *et al*., 2001; Politano *et al*., 2003).

A potentially quite different therapeutic approach is to remove PTCs themselves, rather than interfering with PTC recognition. This can be accomplished in vivo for certain splice site mutations through the use of synthetic oligonucleotide analogs. These oligos hybridize to mutant splice sites or branch point junctions of mutant pre-mRNA, thereby promoting normal splicing and eliminating PTC production that would otherwise occur because of aberrant splicing (Dominski and Kole, 1993). A similar approach has been used in a mouse model to manipulate a PTC-mutated dystrophin mRNA (Mann *et al*., 2001). In this case, antisense oligos were directed toward splice sites flanking a PTC mutation, resulting in in-frame skipping of the affected exon. This treatment removed the PTC, resulting in low-level expression of a shortened but functional dystrophin.

At this time, neither aminoglycoside nor antisense oligo treatment has produced a therapeutically significant benefit in human patients. Of the two, aminoglycoside treatment appears to be the most generally applicable possibility; however, toxicity with prolonged treatment is a concern. A possibly more significant long-term problem could result from general suppression of stop codons, since this might cause accumulation of abnormal mRNAs and abnormal translation of normal mRNAs, potentially leading to production of mutant proteins that interfere with normal cellular functions. Even more significant hurdles remain before it is feasible to treat human patients with antisense oligos, since a systemic delivery method is required, and issues of transfection efficiency, potential immune responses, and side effects must be addressed. Additionally, this treatment would not be general, but would be useful only for cases in which in-frame translation could be maintained without removing essential protein regions. In sum, however, the results from in vitro and in vivo experiments and clinical trials indicate that, at least in principle, functional protein production from PTC-mutated mRNAs is possible.

6. Conclusions

As has become increasingly evident during explorations of its molecular mechanism, NMD is one of the central conserved processes of RNA surveillance. Importantly, NMD acts to control physiological gene expression in a variety of circumstances and is capable of altering expression of both hereditary and acquired genetic diseases. It is therefore necessary to consider the potential action of NMD on transcripts that contain PTCs, regardless of whether PTCs arise in a physiological or pathological context. In specific genetic conditions characterized by protein deficiency, interest is growing in modulating NMD-mediated destruction of transcripts that could otherwise produce functional protein, although development of such treatment strategies is still at an early stage.

Acknowledgments

JAH is supported by a fellowship from the Human Frontier Science Program. The experimental work of the authors is supported by the Fritz Thyssen Stiftung and the Deutsche Forschungsgemeinschaft.

References

Bamber BA, Beg AA, Twyman RE and Jorgensen EM (1999) The Caenorhabditis elegans unc-49 locus encodes multiple subunits of a heteromultimeric GABA receptor. *The Journal of Neuroscience*, **19**, 5348–5359.

Barton-Davis ER, Cordier L, Shoturma DI, Leland SE and Sweeney HL (1999) Aminoglycoside antibiotics restore dystrophin function to skeletal muscles of mdx mice. *The Journal of Clinical Investigation*, **104**, 375–381.

Bedwell DM, Kaenjak A, Benos DJ, Bebok Z, Bubien JK, Hong J, Tousson A, Clancy JP and Sorscher EJ (1997) Suppression of a CFTR premature stop mutation in a bronchial epithelial cell line. *Nature Medicine*, **3**, 1280–1284.

Brumbaugh KM, Otterness DM, Geisen C, Oliveira V, Brognard J, Li X, Lejeune F, Tibbetts RS, Maquat LE and Abraham RT (2004). The mRNA surveillance protein hSMG-1 functions in genotoxic stress response pathways in mammalian cells. *Molecular Cell*, **14**, 585–598.

Chiu SY, Serin G, Ohara O and Maquat LE (2003) Characterization of human Smg5/7a: a protein with similarities to Caenorhabditis elegans SMG5 and SMG7 that functions in the dephosphorylation of Upf1. *RNA*, **9**, 77–87.

Clancy JP, Bebok Z, Ruiz F, King C, Jones J, Walker L, Greer H, Hong J, Wing L, Macaluso M, *et al*. (2001) Evidence that systemic gentamicin suppresses premature stop mutations in patients with cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine*, **163**, 1683–1692.

Dominski Z and Kole R (1993) Restoration of correct splicing in thalassemic pre-mRNA by antisense oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 8673–8677.

Donnadieu E, Jouvin MH, Rana S, Moffatt MF, Mockford EH, Cookson WO and Kinet JP (2003) Competing functions encoded in the allergy-associated F(c)epsilonRIbeta gene. *Immunity*, **18**, 665–674.

Du M, Jones JR, Lanier J, Keeling KM, Lindsey JR, Tousson A, Bebok Z, Whitsett JA, Dey CR, Colledge WH, *et al*. (2002) Aminoglycoside suppression of a premature stop mutation in a Cftr-/- mouse carrying a human CFTR-G542X transgene. *Journal of Molecular Medicine*, **80**, 595–604.

Gouya L, Puy H, Robreau AM, Bourgeois M, Lamoril J, Da Silva V, Grandchamp B and Deybach JC (2002) The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH. *Nature Genetics*, **30**, 27–28.

Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC and Brenner SE (2003) Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*, **19**(Suppl 1), I118–I121.

Hall GW and Thein S (1994) Nonsense codon mutations in the terminal exon of the beta-globin gene are not associated with a reduction in beta-mRNA accumulation: a mechanism for the phenotype of dominant beta-thalassemia. *Blood*, **83**, 2031–2037.

Holbrook JA, Neu-Yilik G, Hentze MW and Kulozik AE (2004) Nonsense-mediated decay approaches the clinic. *Nature Genetics*, **36**, 801–808.

Keeling KM, Brooks DA, Hopwood JJ, Li P, Thompson JN and Bedwell DM (2001) Gentamicin-mediated suppression of Hurler syndrome stop mutations restores a low level of alpha-L-iduronidase activity and reduces lysosomal glycosaminoglycan accumulation. *Human Molecular Genetics*, **10**, 291–299.

Kugler W, Enssle J, Hentze MW and Kulozik AE (1995) Nuclear degradation of nonsense mutated beta-globin mRNA: a post-transcriptional mechanism to protect heterozygotes from severe clinical manifestations of beta-thalassemia? *Nucleic Acids Research*, **23**, 413–418.

Lamba JK, Adachi M, Sun D, Tammur J, Schuetz EG, Allikmets R and Schuetz JD (2003) Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. *Human Molecular Genetics*, **12**, 99–109.

Lewis BP, Green RE and Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 189–192.

Li S and Wilkinson MF (1998) Nonsense surveillance in lymphocytes? *Immunity*, **8**, 135–141.

Mann CJ, Honeyman K, Cheng AJ, Ly T, Lloyd F, Fletcher S, Morgan JE, Partridge TA and Wilton SD (2001) Antisense-induced exon skipping and synthesis of dystrophin in the mdx mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 42–47.

Maquat LE (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Reviews Molecular Cell Biology*, **5**, 89–99.

Medghalchi SM, Frischmeyer PA, Mendell JT, Kelly AG, Lawler AM and Dietz HC (2001) Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Human Molecular Genetics*, **10**, 99–105.

Politano L, Nigro G, Nigro V, Piluso G, Papparella S, Paciello O and Comi LI (2003) Gentamicin administration in Duchenne patients with premature stop codon. Preliminary results. *Acta Myologica*, **22**, 15–21.

Reichenbach P, Hoss M, Azzalin CM, Nabholz M, Bucher P and Lingner J (2003) A human homolog of yeast Est1 associates with telomerase and uncaps chromosome ends when overexpressed. *Current Biology*, **13**, 568–574.

Ruiz-Echevarria MJ, Czaplinski K and Peltz SW (1996) Making sense of nonsense in yeast. *Trends in Biochemical Sciences*, **21**, 433–438.

Sangkuhl K, Schulz A, Rompler H, Yun J, Wess J and Schoneberg T (2004) Aminoglycoside-mediated rescue of a disease-causing nonsense mutation in the V2 vasopressin receptor gene in vitro and in vivo. *Human Molecular Genetics*, **13**(9), 893–903.

Schell T, Kulozik AE and Hentze MW (2002). Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. *Genome Biology*, **3**, REVIEWS1006.

Singh G and Lykke-Andersen J (2003) New insights into the formation of active nonsense-mediated decay complexes. *Trends in Biochemical Sciences*, **28**, 464–466.

Snow BE, Erdmann N, Cruickshank J, Goldman H, Gill RM, Robinson MO and Harrington L (2003) Functional conservation of the telomerase protein est1p in humans. *Current Biology*, **13**, 698–704.

Sureau A, Gattoni R, Dooghe Y, Stevenin J and Soret J (2001) SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *The EMBO Journal*, **20**, 1785–1796.

Thein SL, Hesketh C, Taylor P, Temperley IJ, Hutchinson RM, Old JM, Wood WG, Clegg JB and Weatherall DJ (1990). Molecular basis for dominantly inherited inclusion body beta-thalassemia. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 3924–3928.

Wagner KR, Hamed S, Hadley DW, Gropman AL, Burstein AH, Escolar DM, Hoffman EP and Fischbeck KH (2001) Gentamicin treatment of Duchenne and Becker muscular dystrophy due to nonsense mutations. *Annals of Neurology*, **49**, 706–711.

Wilkinson MF (2003) The cycle of nonsense. *Molecular Cell*, **12**, 1059–1061.

Wilschanski M, Famini C, Blau H, Rivlin J, Augarten A, Avital A, Kerem B and Kerem E (2000) A pilot study of the effect of gentamicin on nasal potential difference measurements in cystic fibrosis patients carrying stop mutations. *American Journal of Respiratory and Critical Care Medicine*, **161**, 860–865.

Wilschanski M, Yahav Y, Yaacov Y, Blau H, Bentur L, Rivlin J, Aviram M, Bdolah-Abram T, Bebok Z, Shushi L, *et al*. (2003) Gentamicin-induced correction of CFTR function in patients with cystic fibrosis and CFTR stop mutations. *The New England Journal of Medicine*, **349**, 1433–1441.

Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA and Smith CW (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Molecular Cell*, **13**, 91–100.

Short Specialist Review

Pilot gene discovery in plasmodial pathogens

Jane M. Carlton

The Institute for Genomic Research, Rockville, MD, USA

The scourge of the African subcontinent, the human malaria parasite *Plasmodium falciparum*, causes an estimated 300–500 million cases and 2–3 million deaths each year (Breman *et al*., 2001). Taken together with the most prevalent but rarely fatal species *Plasmodium vivax*, 40% of the world's population is susceptible to infection. Despite such figures, funds for malaria research have lagged behind those for more "Western" diseases such as cancer and heart disease. A malaria vaccine is yet to be produced, and the parasite has developed resistance to many of the antimalarial drugs that are the single line of defense from infection by the parasite. Although much can be blamed on the lack of research funds to counteract this terrible disease, the parasite's complex life cycle that involves multiple tissues in both a vertebrate host and mosquito vector, in conjunction with poor health care services and public infrastructure in many countries where the disease is endemic, have combined to make malaria an intractable part of day-to-day life in the poorest countries of the world.

Despite the limited amounts of funds for research of the *Plasmodium* parasite, malaria researchers were among the very first to understand the power of pilot gene discovery through EST (Expressed Sequence Tag) projects. An EST is a short (300–500 nucleotides) partial sequence of one end of a cDNA clone that provides a sequence tag for a gene (*see* Article 78, **What is an EST?**, Volume 4). In order to achieve high throughput, these sequences are usually only subject to a single pass of sequencing so the error rate can be as high as 5%. However, sequence similarity searches of the tag with public sequence databases can provide a putative identity and function of the cDNA clone, allowing for gene discovery. ESTs can also provide preliminary data concerning stage-specific gene expression, and the capacity for, and process of, alternative splicing. In organisms in which very little genome sequence data is available, EST projects provide an inexpensive, fast, and powerful means for pilot gene discovery, and can increase the number of identified genes in public databases severalfold.

A list of all *Plasmodium* ESTs generated to date is shown in Table 1. The first EST project in a species of *Plasmodium* used cDNA libraries constructed from blood stages of two laboratory clones of *P. falciparum*, from which 1115 ESTs were generated (Dame *et al*., 1996; Chakrabarti *et al*., 1994). Subsequent cDNA libraries

Table 1 Pilot gene discovery in *Plasmodium*. The number of sequenced ESTs present in GenBank's dbEST database are shown for various life-cycle stages

	P. falciparum	*P. vivax*[a]	*P. berghei*	*P. yoelii*
Life stage: number	Bs: 15 328 Gm: 5814	Bs: 806	Bs: 5582 Oc: 1485 Spz: 199 Ok: 430	Bs: 12 465 Lv: 1921 Spz: 3092 Ah: 1452
Total	21 142	806	7696	18 930

Bs: blood stage; Gm: gametocyte; Oc: oocyst; Spz: sporozoite; Ok: ookinete; Lv: liver; Ah: axenic hepatic.
[a]An additional 20 000 Bs ESTs are currently being generated.

have been constructed from gametocytes, a sexual stage of the parasite (Li *et al*., 2003), and from blood stages of other laboratory lines utilizing different cDNA library construction techniques to enhance full-length cDNA clones (Watanabe *et al*., 2002). Projects to generate additional good-quality cDNA libraries from a variety of *P. falciparum* stages are ongoing, although they are somewhat hampered by the intractability of the *P. falciparum* life cycle, which cannot be completed in a laboratory setting and which limits the amount of biological material available for certain life stages, in particular, the mosquito and liver stages. The problem of limited starting material has also hindered the construction of good-quality cDNA libraries of *P. vivax*. Unlike *P. falciparum*, no continuous culture system has been developed for the maintenance of the blood stages of *P. vivax*, and researchers have been restricted to using blood samples from patients to construct cDNA libraries for sequencing (Merino *et al*., 2003). ESTs from several other *Plasmodium* species have also been generated (Table 1). Two species of rodent malaria, in particular, *Plasmodium berghei* (Carlton *et al*., 2001a; Matuschewski *et al*., 2002; Srinivasan *et al*., 2004; Abraham *et al*., 2004) and *Plasmodium yoelii* (Suzuki *et al*., 1997; Kappe *et al*., 2001; Carlton *et al*., 2002), have significant amounts of EST data from many different life stages. Rodent species of malaria are used as in vivo models to study the human malaria parasite, since they share very many biological characteristics, and provide an analogous system with which to compare and contrast biological mechanisms.

The total number of ~50 000 ESTs generated for the *Plasmodium* genus may seem a rather insignificant number in the light of other organisms for which hundreds of thousands of ESTs exist (e.g., human, mouse, zebrafish). This is because there are several hurdles that must be overcome in order to generate good-quality cDNA libraries from different life stages of the malaria parasite that are amenable to high-throughput sequencing. First, as alluded to above, parasite material from most developmental stages is problematic to obtain in large quantities. In vitro cultivation of some stages of a few species is possible (Schuster, 2002), but in the majority of cases, in vivo material must be dissected and extracted. Second, methods to separate parasites from their host cells are necessary to prevent contamination of the parasite DNA and RNA by host nucleic acids. This is especially important since vertebrate and mosquito genomes are many fold larger

than the ~25-Mb *Plasmodium* genome, so even small amounts of contaminating host material will result in a host-biased library. Several methods, such as filtration through CFll cellulose, ultracentrifugation through Hoechst Dye 33258-CsCl, and biomagnetic separation have been tried and tested (Carlton *et al*., 2001b). Finally, the genomes of *P. falciparum*, *P. vivax*, and the four rodent malaria species show an extreme (AT) bias in their genome sequence. Coding regions are typically 70–80% (AT) and can contain tracts of poly(A) and poly(T) sequence. This can have important consequences in libraries generated by the conventional method of oligo(dT) priming at the 3′-end of mRNA, since priming may occur at regions other than the poly(A) tail. Moreover, highly (AT)-rich DNA is known to be unstable when cloned into *E. coli*, resulting in loss of insert DNA and chimeric clones that can lead to such libraries being nonrandom.

All of the *Plasmodium* ESTs mentioned here are available for downloading and searching through public EST databases such as GenBank's dbEST. Many are also available in custom *Plasmodium* databases, of which there are several. "Full-Malaria" (Watanabe *et al*., 2004) is a database of full-length-enriched cDNAs of *P. falciparum* and *P. yoelii*, which have been mapped to the full-genome sequences of both species. "ApiEST-DB" (Li *et al*., 2004) provides access to EST data from several protozoan parasites in the phylum Apicomplexa, including *Plasmodium*. This relational database can be used for gene model validation, identification of alternative splicing, and identification of phylogenetically conserved sequences. "PlasmoDB" (Bahl *et al*., 2003), the official database of the malaria parasite genome projects, also contains some of the *Plasmodium* EST datasets mentioned here. Finally, the "TIGR Protist Gene Indices" (Quackenbush *et al*., 2001) contains EST data from 15 species of protist including species of *Plasmodium*. The database provides consensus sequences of clustered ESTs and a means of identifying orthologous genes across multiple eukaryotic organisms. The methodology behind EST clustering is explained elsewhere (*see* Article 88, **EST clustering: a short tutorial**, Volume 4).

To what uses have the *Plasmodium* ESTs been put besides pilot gene discovery? A few examples are given here. A study to construct the proteomes of three *Plasmodium* species used several thousand ESTs in conjunction with all known *Plasmodium* genes in GenBank, to compare protein content between the species (Carlton *et al*., 2001a). The *P. yoelii* EST dataset has been used in conjunction with a secondary database of clusters of orthologous groups (COGs) to identify ESTs that remain uncharacterized but have matches to COG proteins, and therefore represent candidates for further protein characterization (Faria-Campos *et al*., 2003). Using the novel technique of subtractive hybridization, *P. berghei* cDNA libraries enriched for genes expressed in ookinetes (Abraham *et al*., 2004) and oocysts (Srinivasan *et al*., 2004) have been generated, and ESTs from these libraries have provided initial insight into *Plasmodium* development in the mosquito. Analysis of multiple ESTs of the same gene have revealed unique features of malaria parasite transcripts, such as the presence of multiple transcription start sites for many genes, and the long length of the 5′ untranslated region (Watanabe *et al*., 2002). And finally, the value of EST data has been emphasized recently with the publication of the genome sequences of two *Plasmodium* species (Carlton *et al*., 2002;

Gardner *et al*., 2002), which relied significantly upon the EST data for training gene finder software and for gene model verification.

The generation of further *Plasmodium* EST data continues, in particular, from other life-cycle stages of *P. falciparum* and other species such as *P. vivax*. In addition, although the *Plasmodium* EST datasets have more than proven their worth, the continued development of full-length cDNA library construction technology is promising and has the potential to produce better gene expression data for the annotation of the six additional *Plasmodium* genome sequencing projects that are currently in progress.

References

Abraham EG, Islam S, Srinivasan P, Ghosh AK, Valenzuela JG, Ribeiro JM, Kafatos FC, Dimopoulos G and Jacobs-Lorena M (2004) Analysis of the *Plasmodium* and Anopheles transcriptional repertoire during ookinete development and midgut invasion. *Journal of Biological Chemistry*, **279**, 5573–5580.

Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P, *et al*. (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Research*, **31**, 212–215.

Breman JG, Egan A and Keusch GT (2001) The intolerable burden of malaria: a new look at the numbers. *American Journal of Tropical Medicine and Hygiene*, **64**, iv–vii.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, **419**, 512–519.

Carlton JM, Muller R, Yowell CA, Fluegge MR, Sturrock KA, Pritt JR, Vargas-Serrato E, Galinski MR, Barnwell JW, Mulder N, *et al*. (2001a) Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Molecular and Biochemical Parasitology*, **118**, 201–210.

Carlton JM, Yowell CA, Sturrock KA and Dame JB (2001b) Biomagnetic separation of contaminating host leukocytes from plasmodium-infected erythrocytes. *Experimental Parasitology*, **97**, 111–114.

Chakrabarti D, Reddy GR, Dame JB, Almira EC, Laipis PJ, Ferl RJ, Yang TP, Rowe TC and Schuster SM (1994) Analysis of expressed sequence tags from *Plasmodium* falciparum. *Molecular and Biochemical Parasitology*, **66**, 97–104.

Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J, *et al*. (1996) Current status of the *Plasmodium* falciparum genome project. *Molecular and Biochemical Parasitology*, **79**, 1–12.

Faria-Campos AC, Cerqueira GC, Anacleto C, de Carvalho CM and Ortega JM (2003) Mining microorganism EST databases in the quest for new proteins. *Genetics and Molecular Research*, **2**, 169–177.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium* falciparum. *Nature*, **419**, 498–511.

Kappe SH, Gardner MJ, Brown SM, Ross J, Matuschewski K, Ribeiro JM, Adams JH, Quackenbush J, Cho J, Carucci DJ, *et al*. (2001) Exploring the transcriptome of the malaria sporozoite stage. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9895–9900.

Li L, Brunk BP, Kissinger JC, Pape D, Tang K, Cole RH, Martin J, Wylie T, Dante M, Fogarty SJ, *et al*. (2003) Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Research*, **13**, 443–454.

Li L, Crabtree J, Fischer S, Pinney D, Stoeckert CJ Jr, Sibley LD and Roos DS (2004) ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Research*, **32** Database issue, D326–D328.

Matuschewski K, Ross J, Brown SM, Kaiser K, Nussenzweig V and Kappe SH (2002) Infectivity-associated changes in the transcriptional repertoire of the malaria parasite sporozoite stage. *Journal of Biological Chemistry*, **277**, 41948–41953.

Merino EF, Fernandez-Becerra C, Madeira AM, Machado AL, Durham A, Gruber A, Hall N and del Portillo HA (2003) Pilot survey of expressed sequence tags (ESTs) from the asexual blood stages of *Plasmodium vivax* in human patients. *Malaria Journal*, **2**, 21.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

Schuster FL (2002) Cultivation of *Plasmodium* spp. *Clinical Microbiology Reviews*, **15**, 355–364.

Srinivasan P, Abraham EG, Ghosh AK, Valenzuela J, Ribeiro JM, Dimopoulos G, Kafatos FC, Adams JH, Fujioka H and Jacobs-Lorena M (2004) Analysis of the *Plasmodium* and Anopheles transcriptomes during oocyst differentiation. *Journal of Biological Chemistry*, **279**, 5581–5587.

Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A and Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene*, **200**, 149–156.

Watanabe J, Sasaki M, Suzuki Y and Sugano S (2002) Analysis of transcriptomes of human malaria parasite *Plasmodium* falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene*, **291**, 105–113.

Watanabe J, Suzuki Y, Sasaki M and Sugano S (2004) Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, *Plasmodium* species. *Nucleic Acids Research*, **32** Database issue, D334–D338.

Manufacturing EST libraries

Marcelo B. Soares and Maria F. Bonaldo
Children's Memorial Research Center, Northwestern University, Chicago, IL, USA

1. Introduction

Several articles and book chapters have been written on the topic of constructing cDNA libraries for large-scale production of expressed sequence tags (ESTs), including detailed discussions of existing methodologies and step-by-step protocols (Bonaldo *et al.*, 1996; Soares, 1994; Soares and Bonaldo, 1998; Soares and Bonaldo, 2002). Hence, rather than describing specific procedures, we identify problems that occur systematically in the manufacturing of EST libraries, discuss their cause and outcome, explain how they may be diagnosed, and indicate how they may affect analysis and interpretation of EST data.

2. The value of ESTs as tools for transcriptome analysis

ESTs are single-pass sequence reads derived from 5′ (5′ EST) and 3′ (3′ EST) ends of directionally cloned cDNAs (*see* Article 78, **What is an EST?**, Volume 4). Since EST libraries, that is, cDNA libraries utilized for production of ESTs, are almost invariably composed of oligodeoxythymidylate- (i.e., oligo(dT)-) primed, directionally cloned cDNAs, 3′ ESTs typically span the 3′ terminal noncoding exon of mRNAs, while 5′ ESTs encompass 5′ noncoding, coding and/or 3′ noncoding exons, depending on whether they are derived from full-length or truncated cDNAs (Soares and Bonaldo, 2002).

Although principally utilized for large-scale gene discovery, ESTs may reveal alternative RNA processing (splicing and polyadenylation), intronic expression, gene fusions resulting from chromosomal rearrangements, internal exon deletions, exon extensions, antisense transcription, and so on (Bonaldo *et al.*, 2004; Brentani *et al.*, 2003; Dimopoulos *et al.*, 2000; Hackett *et al.*, 2004; Hillier *et al.*, 1996; Kochiwa *et al.*, 2002; Okazaki *et al.*, 2002; Rosok and Sioud, 2004; Scheetz *et al.*, 2004a; Verjovski-Almeida *et al.*, 2003). Nevertheless, their value relies on the quality of the cDNA library from which they originate. Indeed, cDNA libraries of inadequate quality often yield artifactual ESTs that if not recognized may engender erroneous conclusions (e.g., cDNA chimeras: a cDNA containing sequences derived from more than one mRNA). Despite curation efforts, public

databases (*see* Article 80, **EST resources, clone sets, and databases**, Volume 4) do contain a significant number of artifactual ESTs (Adams *et al*., 1991).

3. Manufacturing EST libraries: a brief overview

Like some other laboratory procedures in molecular biology, the manufacturing of an EST library is indeed an art. The mere utilization of a well-established and proven protocol simply does not suffice as a guarantee of successful outcome. Certainly not, unless – and on occasion, even if ... – performed by an adept and meticulous experimentalist, who notably understands the biochemistry underlying each reaction in the process. Ultimately, the quality of an EST library is circumscribed by that of the RNA template from which it originates. Cytoplasmic mRNA is the template of choice (Carninci *et al*., 2002), but because it cannot always be obtained, total cellular poly(A)$^+$ RNA is more often utilized.

Although every reaction contributes to the quality of an EST library, first-strand cDNA synthesis is arguably the most critical step, and where problems often arise. First-strand cDNA synthesis can be initiated at the 3′ terminal poly(A) tail using an oligo(dT) primer, or at multiple sites within a transcript, simultaneously, using random primers. Although coding and 5′ noncoding regions are better represented in random primed- than in oligo(dT)-primed libraries, the lack of cloning directionality and the presence of multiple nonoverlapping truncated cDNAs make random primed libraries disadvantageous for large-scale transcript discovery.

Cloning directionality can be achieved by the inclusion of a restriction endonuclease site, such as *Not I*, in the oligo(dT) primer utilized for synthesis of first-strand cDNA (i.e., 5′ *Not I* – $[dT]_{18}$ 3′). Digestion of double-stranded cDNAs with *Not I* thus enables orientation-specific ligation to the cloning vector (Bonaldo *et al*., 1996). To maximize cloning efficiencies, cDNAs are first ligated to a synthetic adapter molecule (e.g., *Eco RI* adapter) and then digested with *Not I*. To avoid digestion at internal *Not I* sites, methyl(dCTP) may be incorporated during first-strand cDNA synthesis (Carninci *et al*., 2000); *Not I* is sensitive to CpG methylation. It is noteworthy that there are many *Not I*-truncated ESTs in public databases, predominantly derived from nonmethylated cDNAs.

The oligo(dT) primer utilized for synthesis of first-strand cDNA may be designed to contain a library-specific sequence tag, typically comprising 6–10 nucleotides, between the restriction site and the $(dT)_{18}$ sequence: 5′ *Not I* – library tag – $(dT)_{18}$ 3′. This is advantageous because it enables the identification of library/tissue of origin of ESTs derived from pooled libraries (Gavin *et al*., 2002; Laffin *et al*., 2004; Scheetz *et al*., 2004a,b). Library-specific tags have also proven invaluable to uncover library mix-ups and clone contaminations.

4. Commonly observed problems in EST libraries

A wide range of problems has been observed in EST libraries, some simple others complex (Soares, 1994; Soares and Bonaldo, 1998). Simple problems not uncommon in EST libraries include the presence of short cDNAs, cDNAs with long

poly(A/T) tails, cDNAs consisting exclusively of poly(A/T) tails, chimeric cDNAs, cDNAs derived from contaminating endogenous or exogenous DNA or RNA (e.g., bacterial DNA, nuclear DNA and RNA), and truncated cDNAs resulting from digestion at internal *Not I* sites. These problems can be greatly minimized by abiding to a few measures. First, the RNA template for synthesis of first-strand cDNA should be predigested with *RNAse*-free *DNAse I* to destroy any contaminating DNA that might otherwise be cloned. Second, a reliable size selection procedure should be utilized to exclude small cDNA fragments and excess adapter molecules. This will not only reduce the representation of clones with short inserts but will also lower the frequency of chimeric cDNAs in the library – small cDNAs often remain unaccounted in the calculations to estimate the mass of cDNA synthesized and thus determine the amount of *Eco RI* adapter and, subsequently, of cloning vector in the ligation reactions. If the amount of cDNA is underestimated, not sufficient adapter molecules are added to the reaction and chimeric cDNAs are generated. Third, methyl(dCTP) should be incorporated during first-strand cDNA synthesis if, subsequently, cDNAs are digested with a restriction endonuclease that is sensitive to CpG methylation. This will minimize the representation of truncated cDNAs resulting from digestion at such internal restriction sites. Fourth, the cloning vector, often a plasmid, should be purified from any contaminating bacterial DNA that might otherwise be cloned.

Complex problems commonly observed in EST libraries, on the other hand, cannot be as easily avoided. However, they can be minimized, and most importantly, they must be recognized in order not to cause misinterpretation of EST data. Complex problems fall into two groups: those that affect transcript representation in the library, either partially or totally, and those that have an effect on cloning directionality.

5. Problems that affect transcript representation in the library

5.1. *Problems that compromise representation of a specific region of a transcript*

The presence of an A-rich stretch at a relatively short distance (i.e., $\leq$250 nucleotides) from the 3′ terminal poly(A) tail may compromise representation of the region of the transcript between the internal A-rich sequence and the poly(A) tail. This is due both to unintended priming at the internal A-rich sequence with the 5′ *Not I* – library tag – $[dT]_{18}$ 3′ oligonucleotide, and to the fact that first-strand cDNA synthesis initiated at the 3′ terminal poly(A) tail ends at the internal priming site. The small 3′ terminal cDNA fragment generated during second-strand synthesis is eliminated during cDNA size selection. Such a problem is often observed in transcripts bearing an *Alu* repeat, in the sense orientation, within the 3′ noncoding region – full-length *Alu* transcripts contain an oligo(A) tail at the 3′ terminus. This is significant considering that Alu repeats occur in noncoding exons of approximately 10% of human mRNAs (Deininger and Batzer, 2002; Weiner, 2002). Internal priming may be minimized, but not eliminated, by an increase in the temperature of the reverse transcription reaction.

Similarly, the occurrence of a *Not I* site within a transcript may compromise library representation of the region localized between the *Not I* site and the 3′ terminal poly(A) tail – the 3′ terminal *Not I* cDNA fragment, extending from the *Not I* site in the primer (5′ *Not I* – library tag – $[dT]_{18}$ 3′) to the internal *Not I* site in the cDNA, cannot be cloned. Hence, representation of such transcripts will be limited to the region between the internal *Not I* site and the EcoRI adapter at the 5′ end of the cDNA. Such *Not I*-truncated cDNAs give rise to ESTs lacking the terminal (dA/dT)-tail. As previously discussed, this problem is mostly observed when dCTP, instead of methyl(dCTP), is used in the synthesis of first-strand cDNA.

5.2. *Problems that compromise representation of an entire transcript*

The presence of an oligo(U) stretch in the 3′ noncoding region may abrogate representation of a transcript, if it occurs within ≤250 nucleotides from the poly(A) tail. This is presumably due to the formation of a stem-loop structure in which the 3′ terminal (A) residues of the poly(A) tail of the RNA are base paired with the complementary (U) residues of the internal (U)-rich sequence. As a result, first-strand cDNA synthesis is simultaneously initiated at two sites: along the looped poly(A) tail, primed by the 5′ *Not I* – library tag – $[dT]_{18}$ 3′ oligonucleotide, and within the internal (U)-rich sequence, self-primed by the 3′ terminal (A) residues of the poly(A) tail of the RNA. The former results in a short product that is eliminated during cDNA size selection. The latter product cannot be cloned because it lacks appropriate *Not I* – *Eco RI* ends. Hence, such a transcript would not be represented in the library.

This scenario may happen with transcripts bearing an *Alu* repeat in the antisense orientation, particularly if the element is truncated. Since the full-length *Alu* repeat is approximately 300 nucleotides long, the occurrence within a transcript of a full-length copy of an *Alu* repeat in the antisense orientation is more likely to result in partial representation, rather than obliteration, of the transcript – the cDNA product corresponding to the region of the transcript extending from the poly(A) tail to the internal oligo(U)-tail of the *Alu* would be larger than 300 bp and thus escape exclusion during cDNA size selection.

6. Problems that affect cloning directionality

One should be cautious using ESTs as evidence of transcription orientation because not all cDNAs in a directional library are cloned in the intended orientation. For example, when utilizing ESTs for computational identification of putative Natural Antisense Transcripts (Kiyosawa *et al*., 2003; Lavorgna *et al*., 2004; Rosok and Sioud, 2004; Yelin *et al*., 2003).

While scrutinizing large numbers of ESTs, we have identified a problem in the manufacturing of EST libraries that causes cDNAs to be cloned in the wrong orientation. On occasion, the *Not I* site in the 5′ *Not I* – library tag – $[dT]_{18}$ 3′ oligonucleotide is destroyed during cDNA synthesis, thus generating a 5′ phosphate

terminus that can be ligated to an *Eco RI* adapter – presumably due to genuine or contaminating nucleolytic activity present in an enzyme utilized in the synthesis of double-stranded cDNA. Hence, the primer for first-strand cDNA synthesis should contain at least five nucleotides 5′ to the *Not I* site. This will not only protect the integrity of the restriction site but will also contribute to increase cleavage efficiency at an otherwise terminal *Not I* restriction site.

Two possible outcomes can be envisioned, should such an event occur. The first, and most likely, is that the cDNA cannot be cloned because it has *Eco RI* adapters at both 5′ and 3′ ends. The second, invoking the presence of an internal *Not I* site in a cDNA, is that one of the two truncated cDNAs generated upon digestion with *Not I* can only be cloned in inverted orientation: that is, the 5′ *Not I* – 3′ *EcoRI* fragment encompassing the region from the internal *Not I* site to the 3′ *Eco RI* adapter. It is noteworthy that the latter clone would give rise to a 5′ EST with an oligo(dA) tail at the 5′ end.

Acknowledgments

The authors are most grateful to the financial support provided by the US Department of Energy and the National Institutes of Health.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Bonaldo MF, Bair TB, Scheetz TE, Snir E, Akabogu I, Bair JL, Berger B, Crouch K, Davis A, Eyestone ME, *et al*. (2004) 1274 full-open reading frames of transcripts expressed in the developing mouse nervous system. *Genome Research*, **14**, 2053–2063.

Bonaldo M, Lennon G and Soares M (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, **6**, 791–806.

Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA Jr, Dias Neto E, Grivet M, Gruber A, Guimaraes PE, Hide W, *et al*. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13418–13423.

Carninci P, Nakamura M, Sato K, Hayashizaki Y and Brownstein MJ (2002) Cytoplasmic RNA extraction from fresh and frozen mammalian tissues. *Biotechniques*, **33**, 306–309.

Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M and Hayashizaki Y (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Research*, **10**, 1617–1630.

Deininger PL and Batzer MA (2002) Mammalian retroelements. *Genome Research*, **12**, 1455–1465.

Dimopoulos G, Casavant TL, Chang S, Scheetz T, Roberts C, Donohue M, Schultz J, Benes V, Bork P, Ansorge W, *et al*. (2000) Anopheles gambiae pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6619–6624.

Gavin AJ, Scheetz TE, Roberts CA, O'Leary B, Braun TA, Sheffield VC, Soares MB, Robinson JP and Casavant TL (2002) Pooled library tissue tags for EST-based gene discovery. *Bioinformatics*, **18**, 1162–1166.

Hackett JD, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Nosenko T and Bhattacharya D (2004) Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Current Biology*, **14**, 213–218.

Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, *et al*. (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807–828.

Kiyosawa H, Yamanaka I, Osato N, Kondo S and Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*, **13**, 1324–1334.

Kochiwa H, Suzuki R, Washio T, Saito R, Bono H, Carninci P, Okazaki Y, Miki R, Hayashizaki Y and Tomita M (2002) Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data. *Genome Research*, **12**, 1286–1293.

Laffin JJ, Scheetz TE, De Fatima Bonaldo M, Reiter RS, Chang S, Eyestone M, Abdulkawy H, Brown B, Roberts C, Tack D, *et al*. (2004) A comprehensive nonredundant expressed sequence tag collection for the developing *Rattus norvegicus* heart. *Physiological Genomics*, **17**, 245–252.

Lavorgna G, Sessa L, Guffanti A, Lassandro L and Casari G (2004) AntiHunter: searching BLAST output for EST antisense transcripts. *Bioinformatics*, **20**, 583–585.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.

Rosok O and Sioud M (2004) Systematic identification of sense-antisense transcripts in mammalian cells. *Nature Biotechnology*, **22**, 104–108.

Scheetz TE, Laffin JJ, Berger B, Holte S, Baumes SA, Brown R II, Chang S, Coco J, Conklin J, Crouch K, *et al*. (2004a) High-throughput gene discovery in the rat. *Genome Research*, **14**, 733–741.

Scheetz TE, Zabner J, Welsh MJ, Coco J, Eyestone Mde F, Bonaldo M, Kucaba T, Casavant TL, Soares MB and McCray PB Jr (2004b) Large-scale gene discovery in human airway epithelia reveals novel transcripts. *Physiological Genomics*, **17**, 69–77.

Soares MB (1994) Construction of directionally cloned cDNA libraries in phagemid vectors. In *Automated DNA Sequencing and Analysis*, Adams MD, Fields C and Venter JC (Eds.), Academic Press: London, pp. 110–114.

Soares MB and Bonaldo MF (1998) Construction and screening of normalized cDNA libraries. In *Genome Analysis: A Laboratory Manual*, Birren B, Green ED, Klapholz S, Myers RM and Roskams J (Eds.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, pp. 49–157.

Soares MB and Bonaldo MF (2002) cDNA libraries. In *Nature Encyclopedia of the Human Genome*, Dear P (Ed.), Nature Publishing Group Macmillan Publishers Ltd: London.

Verjovski-Almeida S, DeMarco R, Martins EA, Guimaraes PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, *et al*. (2003) Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. *Nature Genetics*, **35**, 148–157.

Weiner AM (2002) SINEs and LINEs: the art of biting the hand that feeds you. *Current Opinion in Cell Biology*, **14**, 343–350.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, *et al*. (2003) Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, **21**, 379–386.

EST clustering: a short tutorial

Winston A. Hide
South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

1. Introduction

Expressed sequence tags (ESTs) have been a cornerstone of gene discovery since their large-scale implementation in the early 1990s (Adams *et al*., 1991). By utilizing the existing sequencing technology, the concept of single pass reads "trapping" a fragment or tag of expressed genome sequence has been applied simply and with success across hundreds of species. Once generated, ESTs have potential information that can only be realized by subsequent exhaustive processing in an attempt to reconstruct the transcripts from which they have been sampled.

2. What is an EST cluster?

A cluster is fragmented, EST data and (if known) gene sequence data, consolidated, placed in correct context, and indexed by gene such that all expressed data concerning a single gene is in a single index class, and each index class contains the information for only one gene (Burke *et al*., 1999). Owing to the fragmented nature of EST reads, it is worthwhile to attempt to organize the reads into assemblies that provide a consensus view of the sampled transcripts. In order to reconstruct a true consensus reflection of the sampled parent transcripts, it is necessary to address several complex problems: Should clustering be attempted at the gene locus level or at the parent transcript level? Completeness and availability of parent genome sequence is a major factor. If clustering is at the gene level, how are the limits of the transcribed gene defined? Recent studies indicate that several forms of transcript can be generated from a single locus (Hide *et al*., 2001; Burke *et al*., 1998; Modrek *et al*., 2001; Pospisil *et al*., 2004) and that antisense RNAs and possibly other forms of RNAs revealed in recent tiling path studies can overlap their expression over the same sets of genomic nucleotides (Dahary *et al*., 2005; Hayashizaki and Kanamori, 2004, Kapranov *et al*., 2002). Understanding of transcript biology is still insufficient to allow for "true" transcript clusters to be well defined. If clustering is performed at the transcript level, overlapping transcripts that share the same genic locus must be qualified. Given a definition of overlap, transcript reconstruction can be performed via simple assembly, but the assemblers used must be able to handle unknown forms of transcript diversity that may include alternate

splice forms, identical exons from different chromosomal loci or other paralogous sequences sharing overall high sequence identity. Overclustering can also occur as a result of chimeric clones, shared vector sequence, and uncharacterized or poorly masked repeats. Underclustering can occur as a result of highly abundant transcripts, overstringent clustering parameters, or fragmentation of assemblies. Genome assemblers are designed for shotgun genome assembly, where there are several reads covering the same nucleotide. For EST assembly, the density of coverage varies directly according the level to which the nucleotide in question has been expressed and the degree to which it is sampled. It is fruitful to utilize pipelined approaches that combine algorithms to maximize the content and quality of the reconstruction of transcripts from the assembly of clusters.

3. The EST clustering process

The aim of EST clustering is simply to incorporate all ESTs that share a transcript or gene parent to the same cluster. There is usually a requirement to assemble the clustered ESTs into one or more consensus sequences (contigs) that reflect the transcript diversity, and to provide these contigs in such a manner that the information they contain most truly reflects the sampled biology. The process is confounded by a lack of absolute knowledge of the true biology.

Systems that have broad acceptance and are widely distributed or widely accessible include SANBI at University of Western Cape's StackPACK, TIGR's TGI Clustering tools (TGICL), and NCBI's Unigene (Table 1). Commonly, these systems share an overall approach, but differing in choice of algorithms used, reconstruction aims, and coverage of transcript diversity. The systems perform a preprocessing step that screens out vector, repeat, and low-complexity sequences. A database of vectors and repeats is passed across each EST, and where matches above a certain threshold occur, the matched nucleotides are substituted for a null character such as an "N" (Bedell *et al*., 2000). Masked ESTs are initially

Table 1 Clustering and mapping approaches for transcript reconstruction

System	Approach	Source
Unigene	Sequence identity	**http://www.ncbi.nlm.nih.gov/UniGene**
	Transcript-based build	**http://www.ncbi.nlm.nih.gov/UniGene/build1.html**
	Genome-based build	**http://www.ncbi.nlm.nih.gov/UniGene/g_build.html**
TIGR	Transcript-based build	Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J and Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**(5), 651–652. **http://www.tigr.org/tdb/tgi/software/**
SANBI	Transcript-based build	Miller RT, Christoffels AG, Gopalakrishnan C, Burke JA, Ptitsyn AA, Broveak TR and Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: The Sequence Tag Alignment and Consensus Knowledgebase. *Genome Research*, **9**(11), 1143–1155. **http://www.sanbi.ac.za/CODES/**

clustered by a process of initial all-against-all comparison. The resulting clusters are assigned by some form of sequence identity above a threshold, either at the level of shared word multiplicity (StackPACK's D^2 pseudometric) or sequence overlap (TIGR, NCBI). Assembly of clusters is strongly biased by the choice of assembler. Although Liang *et al*. performed a comparative analysis of the suitability of use of CAP3 (Huang and Madan, 1999; Liang *et al*., 2000), they did not take into account the flexibility of the chosen assembler for the incorporation of alternate splice forms into the generated consensus sequences. Choice of assembler is affected by the desired result. Features of the different systems are as follows: StackPACK, performs initial clustering on the basis of shared word multiplicity followed by assembly and consensus processing steps (Miller *et al*., 1999). The system is suitable for providing reliable capture of transcript diversity such as alternate splicing within gene-based clusters. StackPACK utilizes a more relaxed initial clustering, and also a more flexible assembler (PHRAP) generating more contigs per cluster, but incorporating more alternate splicing events. Less-stringent clustering, however, requires that there be a consensus management step to sort out the relationships of the contigs generated within clusters. TGICL is another broadly used procedure that combines EST clustering based on sequence similarity and subsequent transcript assembly (Quackenbush *et al*., 2001; Pertea *et al*., 2003). Initial clustering is performed by a modified version of NCBI's *megablast*, and the resulting clusters are then assembled using the CAP3 assembly program. Larger numbers of contigs in separate clusters are generated by this procedure than in StackPACK. More recently, graph-based approaches to transcript reconstruction have been developed that address the potential isoform diversity of reconstructed transcripts as a graph (Heber *et al*., 2002; Xing *et al*., 2004). The most well known clustering system is that utilized by NCBI for the frequently updated Unigene series

Table 2 Resources for EST clustering and EST description

Resource title	Link
EST links	**http://industry.ebi.ac.uk/~muilu/EST/EST_links.html**
Good description of EST clustering process at SANGER Est_db	**http://www.sanger.ac.uk/Software/analysis/est_db/**
Early explanation of clustering	**www.littlest.co.uk/software/bioinf/old_packages/jesam/jesam_paper.htm**
Simple clustering tool	Making sense of EST sequences by CLOBBing them J Parkinson, D Guiliano, M Blaxter Portable EST clustering solution freely downloaded from: **http://www.nematodes.org/CLOBB**.
Characteristics and methods to work with ESTs	Searching the expressed sequence tag (EST) databases: Panning for genes: (2000) *Briefings in Bioinformatics*, **1**(1), 76–92(17)
Public data used for transcript reconstruction	**http://www.ncbi.nlm.nih.gov/dbEST/index.html**
the TIGR TGI databases	**http://www.tigr.org/tdb/hgi/hgi.html**
Ecgene combination of clustering, AS capture and gene expression	**http://genome.ewha.ac.kr/ECgene/index2.html**
Gene Nest; Online Splicing analysis of gene indices	**http://genenest.molgen.mpg.de/**
Weizmann institute links page to gene indices	**bip.weizmann.ac.il/hg3m/databases/est.html**

of databases (**http://www.ncbi.nlm.nih.gov/UniGene**). Unigene is unique amongst these systems in that it does not attempt to reconstruct transcripts but rather attempts to define their cluster membership on the basis of NCBI data. Methods to deal with error vary widely amongst all systems. A common example of error in clustering is that of incorrect clone joining. The process of EST manufacture usually requires that a clone be sequenced from one or both ends. As a result of this experimental design, it is possible to join 5′ and 3′ ESTs that originate from the same cluster that share a parent clone. However, missannotations can result in the spurious formation of superclusters, requiring that more than one pair of ESTs sharing parent clones be required to join two clusters. Links to appropriate EST resources and information are provided in Table 2.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B and Moreno RF (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**(5013), 1651–1656.

Bedell JA, Korf I and Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**(11), 1040–1041.

Burke J, Davison D and Hide W (1999) D2_cluster: a validated METHOD for clustering EST and Full-length cDNA sequences. *Genome Research*, **9**(11), 1135–1142.

Burke J, Wang H, Hide W and Davison D (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Research*, **8**(3), 276–290.

Dahary D, Elroy-Stein O and Sorek R (2005) Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Research*, PMID: 1571075, **15**, 364–368.

Hayashizaki Y and Kanamori M (2004) Dynamic transcriptome of mice. *Trends in Biotechnology*, **22**(4), 161–167.

Heber S, Alekseyev M, Sze SH, Tang H and Pevzner PA (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18**(Suppl 1), S181–S188.

Hide WA, Babenko VN, van Heusden PA and Kelso JF (2001) The contribution of exon skipping events on Chromosome 22 to protein coding diversity. *Genome Research*, **11**(11), 1848–1853.

Huang X and Madan A (1999) Contig Assembly Program version 3 (CAP3): a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP and Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**(5569), 916–919.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL and Quackenbush J (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Research*, **28**(18), 3657–3665.

Miller RT, Christoffels AG, Gopalakrishnan C, Burke JA, Ptitsyn AA, Broveak TR and Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledgebase. *Genome Research*, **9**(11), 1143–1155.

Modrek B, Resch A, Grasso C and Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, **29**, 2850–2859.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F and Parvizi B (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**(5), 651–652.

Pospisil H, Herrmann A, Bortfeldt RH and Reich JG (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Research*, **32**, D70–D74.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**(1), 159–164.

Xing Y, Resch A and Lee C (2004) The Multiassembly problem: reconstructing multiple transcript Isoforms from EST fragment mixtures. *Genome Research*, **14**(3), 426–441.

Basic Techniques and Approaches

Using UniGene, STACK, and TIGR indices

Alan G. Christoffels

Temasek LifeSciences Laboratory, National University of Singapore, Singapore

1. Introduction

Expressed sequence tag (EST) (*see* Article 78, **What is an EST?**, Volume 4) sequencing and analysis represent a critical tool in the identification of genes and for annotation of genomic sequence despite the increase in sequenced genomes. However, the challenge in using EST data is to assign an EST to a gene without prior knowledge of the EST's origin (*see* Article 88, **EST clustering: a short tutorial**, Volume 4). This challenge has been met by three research organizations, namely, (1) National Center for Biotechnology and Information (NCBI), (2) South African National Bioinformatics Institute (SANBI), and (3) The Institute for Genome Research (TIGR) through the development of UniGene, STACK, and TIGR gene indices respectively. The gene indices represent processed EST and mRNA transcripts, where the transcripts are grouped or clustered into nonredundant transcripts associated with distinct gene loci. Each of these gene indices shares a common framework of data cleaning, clustering, and assembly, with additional modifications to meet a specific goal. A comparison among the gene indices' protocols is presented as a flowchart (Figure 1) followed by information that can be gleaned from the user interface designed for each of the gene indices (Figures 2–4).

2. UniGene

The UniGene builds protocol implements: (1) a transcript-based approach utilizing ESTs and mRNAs exclusively and (2) a genome-based approach where transcripts are mapped to genomic sequences to identify gene loci (see Figure 1 for the transcript-based approach; Pontius *et al*., 2003). UniGene data can be queried using a UniGene identifier, GenBank EST accession number, cDNA library, or a chromosome location (Wheeler *et al*., 2003). The retrieved UniGene record will have cross-links to databases, including ProtEST for protein similarities, Digital Differential Display for expression profiles, and Homologene to identify putative orthologous relationships (Figure 2). Notice the absence of assembly information due to the inclusion of nonoverlapping ESTs that share a cloneID (Figure 1). Queries against the UniGene database can be more specific by restricting the query

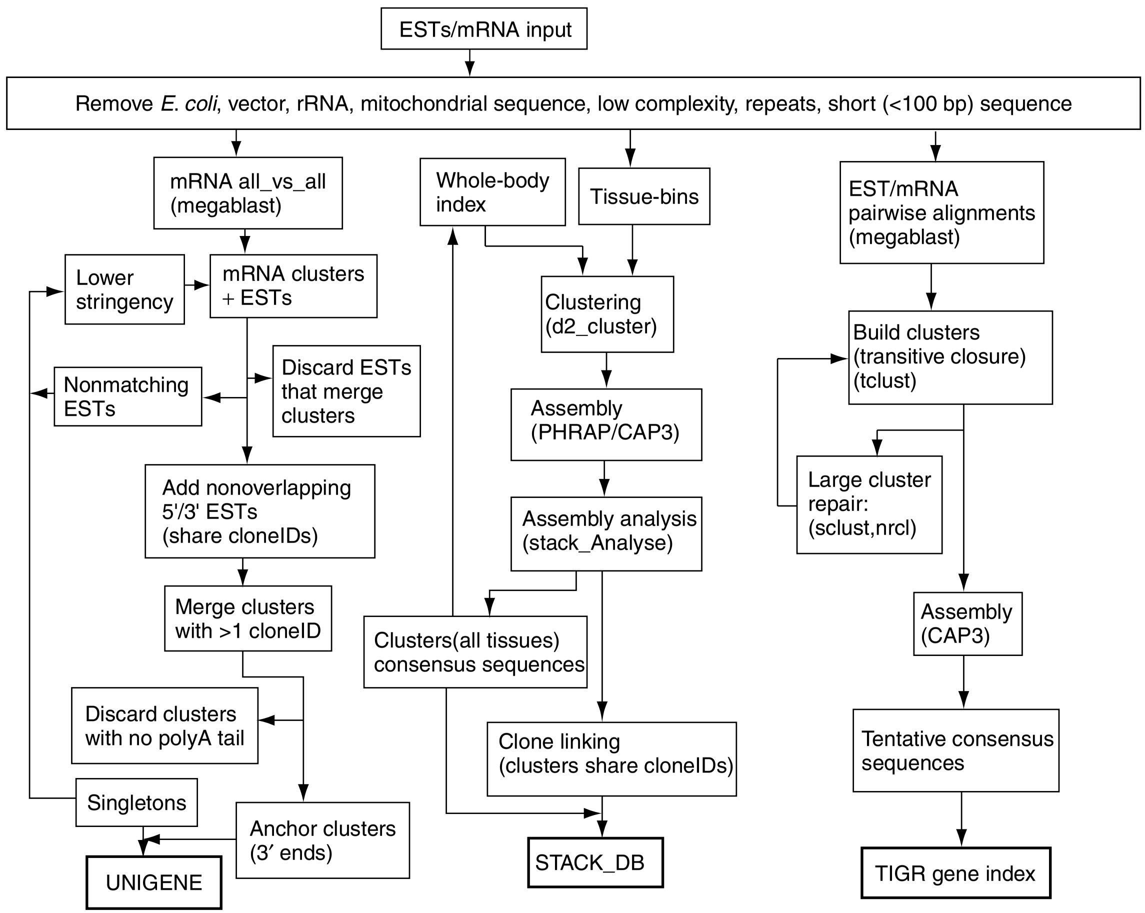


Figure 1 Comparison of the clustering approaches implemented in UniGene, STACK, and TIGR gene indices. The clustering protocol begins with a data-cleaning step followed by pairwise alignment clustering of ESTs (UniGene and TIGR) or word-based clustering (STACK). EST clusters are assembled using PHRAP or CAP3 (STACK and TIGR respectively) and further assembly assessment carried out in STACK processing

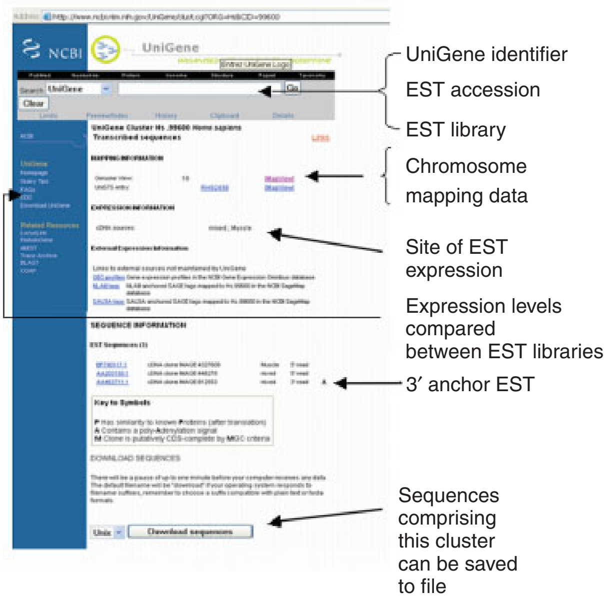


Figure 2 UniGene cluster information as displayed on the NCBI website. UniGene-id Hs.99600 was used to retrieve the cluster report

to specific fields, for example, searching for clusters containing one to ten ESTs using "1:10[ESTC]". Additional field restriction terms are available from the NCBI website at **http://www.ncbi.nlm.nih.gov/UniGene/query_tips.html**.

3. Sequence tag alignment and consensus knowledgebase (STACK)

STACK implements a loose clustering system, d2_cluster (Burke *et al*., 1999), which does not rely on pairwise alignment but instead assesses the composition and multiplicity of words within a sequence (Figure 1). This approach allows for the capture of otherwise discarded low-quality sequence with the result that the integrity of each cluster is assessed by error-checking tools (Figure 1; Miller *et al*., 1999).

STACK allows the user to explore 14 tissue categories, a disease category, and a whole-body index. The STACK database can be viewed via the Web at **http://www.sanbi.ac.za/stacksearch.html** using a sequence as input into a BLAST engine. The BLAST results are hyperlinked to the STACK viewer, which allows for extraction of detailed information pertaining to a matching STACK sequence. Alternatively, all clustered data for a specific STACK tissue can be accessed via WebProbe (**http://www.sanbi.ac.za/stackpack/webprobe.html**). A query from this page returns a summary report with links to detailed information for all clusters and linked clusters contained within a specific tissue category. For example, Figure 3(a) illustrates a query using an EST accession number (e.g., H64402). The

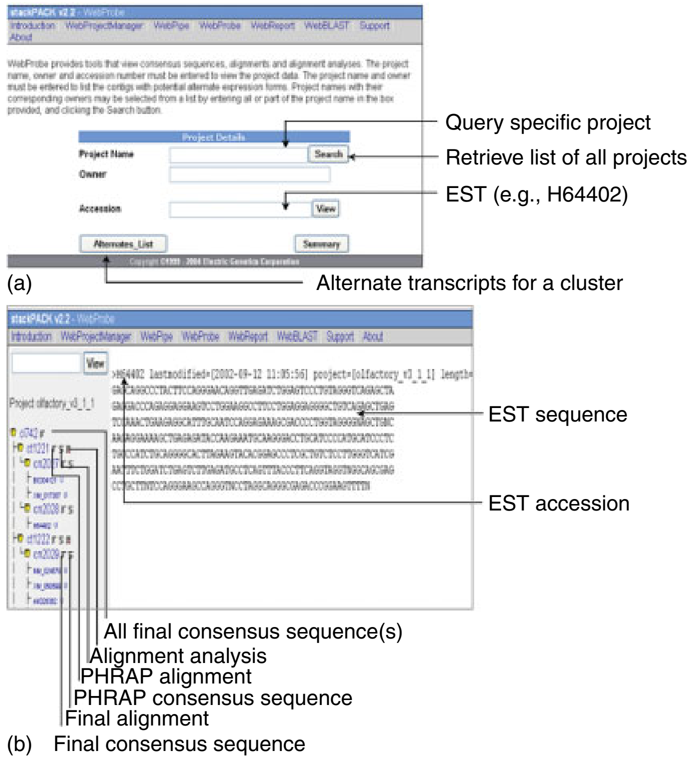


Figure 3 WebProbe, the STACK database extraction and viewing tool. (a) An EST is used as input to retrieve all clustered information relating to the specified EST identifier (H64402). (b) The corresponding cluster (cl742) is presented in a tree format with its contigs (ct1221 and ct1222). Consensus sequences cn2027 and 2028 corresponding to ct1221 and cn2029 corresponding to ct 1221 are displayed with constituent ESTs/mRNAs branching off from it

retrieved data are presented in two panels, namely, the right panel displays the EST sequence in FASTA format and the panel on the left displays the corresponding cluster in a tree format (Figure 3b) with its constituent EST accession numbers and hyperlinks to additional information such as PHRAP alignments, PHRAP consensus sequences, assembly analysis data, and final consensus sequences. Each EST identifier in the left panel has a UniGene hyperlink to search for corresponding UniGene records. Detailed documentation on each of the stackPACK^{TM} tools is available at **http://www.sanbi.ac.za/stackpack/webprobe.html.**

4. TIGR gene indices

The TIGR gene indices represent a collection of species-specific transcript databases. EST data and gene sequences are clustered and assembled into representative gene sequences or tentative consensus (TC) sequences using an alignment-based approach (Figure 1). The TC sequences provide putative genes with functional

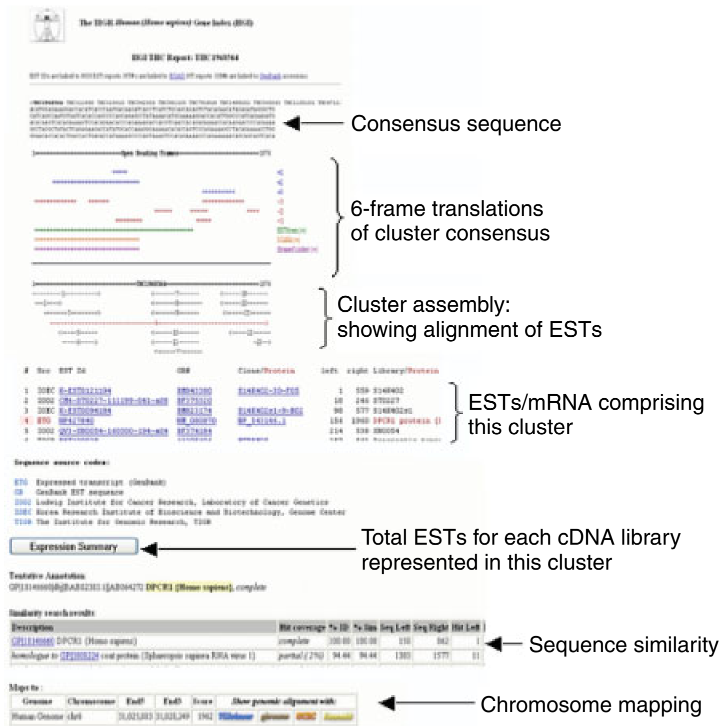


Figure 4 TIGR gene index report for THC1968564. The retrieved record contains a FASTA formatted contig sequence, followed by a 6-frame translation, assembly, constituent EST/mRNAs, relative level of expression, and sequence annotation. The assembly information has been reduced

annotation, a link to mapping data, and provide links between orthologous and paralogous genes through the TIGR Orthologous Groups Alignment (TOGA) database (see Figure 4; Quakenbush *et al*., 2001).

Sequence searches against TIGR gene indices are carried out with WUBLAST (**http://tigrblast.tigr.org/tgi/**). The BLAST results are hyperlinked to matching TC reports containing additional information in the context of the specific assembled transcripts (Figure 4). Alternatively, TIGR gene indices can be searched using a TC number that represents the most current version of a particular assembly, or an EST accession number. Integration of radiation hybrid mapping data allows the user to search for TC sequences that map to a specific genomic location.

5. Availability of UniGene, STACK, and TIGR gene index

UniGene cluster sequences, updated weekly, may be downloaded from the NCBI FTP site. The STACK clustering and visualization tools, stackPACK^{TM}, can be downloaded for different platforms from the SANBI FTP site. TIGR's clustering tools (TIGRCL), available for multiple platforms, can be downloaded from the

TIGR website. The TIGR gene indices are available for download as species-specific datasets.

Further reading

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T and Wide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Research*, **29**, 234–238.

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, *et al*. (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research*, **12**, 493–502.

References

Burke J, Davison D and Hide W (1999) d2_cluster: A validated method for clustering EST and full-length cDNA. *Genome Research*, **9**, 1135–1142.

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR and Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Research*, **9**, 1143–1155.

Pontius JU, Wagner L and Schuler GD (2003) UniGene: A unified view of the transcriptome. In *The NCBI Handbook*, National Center for Biotechnology Information: Bethesda, Also available on the World Wide Web: **http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene**.

Quakenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, **31**, 28–33.

Introductory Review

Microarrays: an overview

Jörg D. Hoheisel
Deutsches Krebsforschungszentrum, Heidelberg, Germany

1. Introduction

In recent years, DNA microarrays have become a synonym for genome-wide functional analyses. The power of the technology as an experimental tool derives from the specificity and relatively strong affinity of duplex formation of complementary sequences paired with the accessibility to analysis of very many fragments in parallel. In terms of molecular biology, DNA microarrays are an old technology already. This is documented by the widespread use of at least some microarray-based applications in laboratories across the world. At the same time, however, it is still an exciting and quickly developing field, since the full potential of assay formats that can be performed is currently not even conclusively defined let alone met. Also, accuracy, sensitivity, and speed of analysis are frequently not yet adequate and have large margins for improvement. As a result, the availability of arrays for routine testing is just starting.

2. The past

To some, it might look as if DNA chips appeared rather suddenly. However, as ever so often in research, it was a technology in the waiting. Gillespie and Spiegelman (1965) observed that single-stranded DNA binds strongly to nitrocellulose membranes in a way that prevents the strands from reassociating with each other but permits hybridization to complementary RNA. Development of a blotting technique by Southern (1975) proved to be a milestone in assaying nucleic acids. The technique, in all its modifications, was essential for many kinds of analyses, such as the identification of disease genes, and crucial for many achievements in the area of genomic research. Clone-filter arrays were introduced by Hans Lehrach (Poustka *et al*., 1986), permitting a reverse analysis format. The resulting parallelism in data production was a crucial step toward genome-wide analyses. The technique was initially designed for mapping purposes (e.g., Hoheisel *et al*., 1993) but quickly stretched out toward other applications (Lennon and Lehrach, 1991).

DNA microarrays on glass surfaces (Figure 1) made their debut then in the late 80s of the last century. Their parallel and independent conception by several groups (Bains and Smith, 1988; Khrapko *et al*., 1989; Drmanac *et al*., 1989; Southern

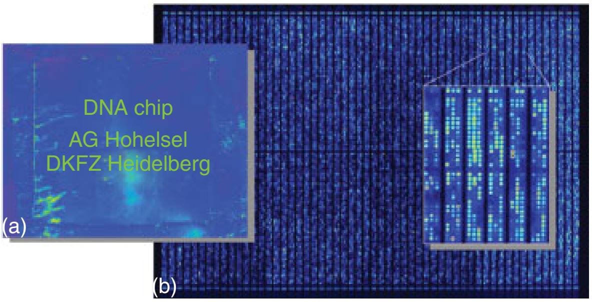


Figure 1 Six years difference: a small step in scale but a large step in complexity. Oligonucleotides were synthesized in situ, controlled by a lithographic mask that resembled some letters (a) or a micromirror system (b), respectively, and hybridized with a labeled DNA sample. While one oligonucleotide was synthesized in the former system, about 40 000 different ones are present on the latter chip, and more complex patterns are possible

et al., 1992) documents the fact that the time was right. Initially, the aim was the development of a tool that would permit the sequencing of the human genome, which at the time was thought to be achievable only with entirely new techniques. While proven wrong for the first (few) human genome(s), the methodology still is a promising contender for routine comparative sequencing. The development of means for the in situ synthesis of oligonucleotide arrays (Fodor *et al*., 1991; Maskos and Southern, 1992) made a relatively cheap and eventually reproducible production possible and made complex microarrays of many features more readily available. A first meeting on the technology was organized at the Engelhardt Institute in Moscow in 1991 (Cantor *et al*., 1992), assembling basically all scientists who were active in this field at the time, an accomplishment never to be repeated.

In terms of impact, the publication of Patrik Brown and colleagues on a preliminary transcriptional analysis of 27 *Arabidopsis* genes (Schena *et al*., 1995) was instrumental in attracting the attention of the wider research community. They did their analysis by a clever combination of techniques – basically all of which had been utilized previously but not in such combination – attaching cDNA PCR products to microscopic slides and using dual fluorescence coloring for detection. Apart from reporting on how such microarrays could be produced and analyzed in a format that biologists were used to, their main achievement really was the provision of a new perspective on what could be achieved by this technology.

3. The present

On this basis, DNA microarrays have become a tool that is commonly used for research purposes in both academia and commerce. Transcriptional profiling is currently the most frequently performed assay on DNA microarrays and starts to yield data, which not only improve the basic knowledge on biological processes but also support diagnosis and prognosis (e.g., Golub *et al*., 1999). Concomitantly, the data analysis tools – an integral and essential part of the analysis – are developing fast. While naturally starting out from many different approaches and protocols,

standardization begins to occur (MIAME; Brazma *et al*., 2001), although still at an insufficient level. Especially, the preparative steps that are prior to the actual chip analysis require much more rigorous standardization and more accurate recording. But even at the level of basic, relatively simple processes, such as normalization, different algorithms often result in a distinct data interpretation. A recent review on transcriptional studies on pancreatic tumors (Brandt *et al*., 2004), for example, demonstrated the inconsistency of such analyses. Of 978 genes that were found to be differentially transcribed in the 10 studies published to date, only 148 were identified in at least two studies, and not a single gene was identified in all studies. Worse even, frequently, there is a huge discrepancy between the genes prioritized by microarray studies and those identified by other means (Miklos and Maleszka, 2004). Much more care is needed in the experimental design, record taking, the actual analysis, and data evaluation to end up with reproducible results.

At the same time, new assay formats emerge, such as studies on epigenetic variations, RNA splicing, protein–DNA interactions, the selection of antisense or RNAi molecules, and such alike. It is apparent that not one microarray format alone will be optimal for all possible applications. Initially, microarrays were mostly made of oligomers, but PCR products or clone fragments took over relatively quickly, before oligonucleotides, short and long ones, became more dominant again. However, for analyses of protein–DNA interactions, for example, the use of DNA fragments is favorable, since most if not all of a genome could be covered that way. Therefore, it seems likely that several chip formats will coexist for a foreseeable future, although in situ synthesis and thus oligomer microarrays are superior with respect to achieving high and verifiable quality.

Currently, microarray technology is beginning to divide into two distinct areas, routine assays on the one hand and research studies on the other. The former type of analysis will eventually require microarrays of low complexity but produced in really large numbers. Research also moves away from the very complex arrays of global coverage, which will mainly be used for initial screens only. At the same time, the need for flexible design and production processes is increasing substantially, which is not surprising given the increase in the number of users, assays, targets, and applications as well as the need of combining all these for a more comprehensive picture. Coincidentally – or maybe not – several large companies that are related to electronic chips rather than biology are investing in the field of biological microarrays. Apart from the financial and commercial aspects, they bring in the much needed expertise and experience in matters such as stringent quality control and marketing, which are essential ingredients of making microarrays a success in the diagnostics market.

4. The future

Apart from the diversification and partition of the field of DNA microarrays, the technology is advancing still with respect to technology. Miniaturization not only of the chips themselves but even more importantly of their environment is actively pursued. Simultaneously, since dependent on the above, signal detection down to individual molecule interactions is another focus of development. And concurrent to

this, the quantification of the absolute signal intensities is important, too. Toward routine application, efforts are ongoing to simplify the usually rather complex process of sample preparation. Various methods exist for avoiding both initial amplification and labeling of the sample material prior to analysis. In conjunction with this, alternative modes of measurement are developed. This includes the identification of mass changes on cantilevers (Arntz *et al*., 2003), for example, the detection by mass spectrometry of phosphates of nucleic acids that hybridize to catching molecules made of phosphate-free DNA-mimics (Brandt *et al*., 2003) or a direct electronic detection by means of conductance or impedance measurements (e.g., Hahm and Lieber, 2003).

On the basis of the experience gathered at the level of nucleic acids, also other molecule classes become more and more the focus of developments for analyses that are similar in type, scale, and complexity to what is achieved on DNA chips. As a matter of course, much is still to be learnt before such experiments are performed at a level achieved with nucleic acids today. Even more difficult will be the interpretation of the resulting data or their translation into information that is useful for diagnostic purposes. However, developments are rapid and the demand for such assays is increasing. This kind of information is needed for the ability to describe a (simple) biological model in silico. Termed "systems biology", such an approach is hampered by the lack of data, since only then the appropriate algorithms can be developed and evaluated and eventually the complex and dynamic interactions in biological systems be described.

References

Arntz Y, Seelig JD, Lang HP, Zhang J, Hunziker P, Ramseyer JP, Meyer E, Hegner M and Gerber C (2003) Label-free protein assay based on a nanomechanical cantilever array. *Nanotechnology*, **14**, 86–90.

Bains W and Smith G (1988) A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology*, **135**, 303–307.

Brandt O, Feldner J, Stephan A, Schröder M, Schnölzer M, Arlinghaus HF, Hoheisel JD and Jacob A (2003) PNA-microarrays for hybridisation of unlabelled DNA-samples. *Nucleic Acids Research*, **31**, e119.

Brandt R, Grützmann R, Bauer A, Jesnowski R, Ringel J, Löhr M, Pilarsky C and Hoheisel JD (2004) DNA-microarray analysis of pancreatic malignancies. *Pancreatology*, **4**, 587–597.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al*. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

Cantor CR, Mirzabekov A and Southern E (1992) Report on the sequencing by hybridisation workshop. *Genomics*, **13**, 1378–1383.

Drmanac R, Labat I, Brukner I and Crkvenjakov R (1989) Sequencing of megabase plus DNA by hybridisation: theory of the method. *Genomics*, **4**, 114–128.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT and Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.

Gillespie D and Spiegelman S (1965) A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *Journal of Molecular Biology*, **12**, 829–842.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hahm J and Lieber CM (2003) Direct ultrasensitive electrical detection of DNA and DNA sequence variations using nanowire nanosensors. *Nano Letters*, **4**, 51–54.

Hoheisel JD, Maier E, Mott R, McCarthy L, Grigoriev AV, Schalkwyk LC, Nizetic D, Francis F and Lehrach H (1993) High-resolution cosmid and P1 maps spanning the 14-Mbp genome of the fission yeast *Schizosaccharomyces pombe*. *Cell*, **73**, 109–120.

Khrapko K, Lysov Y, Khorlyn A, Shick V, Florentiev V and Mirzabekov A (1989) An oligonucleotide hybridization approach to DNA sequencing. *FEBS Letters*, **256**, 118–122.

Lennon GG and Lehrach H (1991) Hybridisation analyses of arrayed cDNA libraries. *Trends in Genetics*, **7**, 314–317.

Maskos U and Southern EM (1992) Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridisation properties of oligonucleotides synthesised in situ. *Nucleic Acids Research*, **20**, 1679–1684.

Miklos GLG and Maleszka R (2004) Microarray reality checks in the context of a complex disease. *Nature Biotechnology*, **22**, 615–621.

Poustka A, Pohl T, Barlow DP, Zehetner G, Craig A, Michiels F, Ehrich E, Frischauf A-M and Lehrach H (1986) Molecular approaches to mammalian genetics. *Cold Spring Harbor Symposia on Quantitative Biology*, **51**, 131–139.

Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, **98**, 503–517.

Southern EM, Maskos U and Elder JK (1992) Analysing and comparing nucleic acid sequences by hybridisation to arrays of oligonucleotides: evaluation using experimental models. *Genomics*, **13**, 1008–1017.

Creating and hybridizing spotted DNA arrays

Ivana V. Yang
Duke University Medical Center, Durham, NC, USA

1. Introduction

The ability to study transcription of all genes in an organism in parallel using gene-expression profiling has revolutionized both basic and clinical science fields (Ramaswamy and Golub, 2002; Staudt and Brown, 2000). Over the past decade, microarrays have evolved from a specialized and expensive tool used in small numbers in a few laboratories to an almost standard technique available and affordable to many researchers. Although not fully mature, both microarray technology and data analysis protocols have been refined to the point that high-quality data can be obtained relatively easily and quickly. This review focuses on recent advancements in printing and hybridizing spotted DNA arrays, and discusses some general aspects of experimental design and data analysis for two-color hybridization assays.

2. What to print?

The first step in creating spotted DNA microarrays is the choice of the DNA material to print. DNA fragments generated by the polymerase chain reaction (PCR) amplification from bacterial cDNA clone sets were the original source of the printing material (Schena *et al*., 1996) and are still used by a number of established microarray laboratories. Sequence-verified cDNA clone sets for many organisms, including mouse, rat, and human, have been made available to researchers through public efforts (Table 1a) (Lennon *et al*., 1996). Analysis and clustering of expressed sequence tags (ESTs) to identify underlying genes (Boguski and Schuler, 1995; Quackenbush *et al*., 2000) provides annotation for cDNA clones and typically includes information such as identifiers from multiple databases, gene ontology (GO) functional categories, chromosome locations, and the presence of orthologs (*see* Article 80, **EST resources, clone sets, and databases**, Volume 4).

An alternative PCR strategy is employed for microarray studies of organisms with smaller and simpler genomes (bacteria, parasites, yeast) (DeRisi *et al*., 1997) and for organisms that are not well represented in publicly available clone

Table 1 Resources for spotted microarray production and hybridization

(a) Clone sets and oligonucleotide libraries	
Compugen	**http://www.labonweb.com/**
Illumina	**http://www.illumina.com**
I.M.A.G.E. Consortium[a]	**http://image.llnl.gov/**
MWG	**http://www.mwg-biotech.com**
Operon Biotechnologies	**http://www.operon.com**
(b) Coated slide substrates	
Asper Biotech	**http://www.asperbio.com**
Corning	**http://www.corning.com**
Erie Microarray	**http://www.eriemicroarray.com**
Full Moon Biosystems	**http://www.fullmoonbio.com**
Genetix	**http://www.genetix.com**
Quantifoil	**http://www.quantifoil.com**
Telechem International	**http://www.arrayit.com**
(c) Microarray spotting robots	
Amersham Biosciences	**http://www.amershambiosciences.com**
Brooks Automation (formerly Intelligent Automation Systems)	**http://www.brooks.com**
Genetix	**http://www.genetix.com**
Genomic Solutions	**http://www.genomicsolutions.com**
Perkin Elmer (formerly Packard Biosciences)	**http://las.perkinelmer.com/**
Telechem International	**http://www.arrayit.com**
(d) RNA amplification and labeling kits and protocols	
Ambion	**http://www.ambion.com**
Agilent Technologies	**http://www.agilent.com**
Arcturus	**http://www.arctur.com**
National Human Genome Research Institute	**http://research.nhgri.nih.gov/nhgri_cores /microarray/array_protocols.html**
Stanford University	**http://cmgm.stanford.edu/pbrown/protocols/index.html**
The Institute for Genomic Research (TIGR)	**http://pga.tigr.org/protocols.shtml**
(e) Microarray scanners and image processing software	
Amersham Biosciences	**http://www.amershambiosciences.com**
Agilent Technologies	**http://www.agilent.com**
CSIRO Mathematical and Information Sciences	**http://experimental.cmis.csiro.au/Spot/index.php**
Genomic Solutions	**http://www.genomicsolutions.com**
Molecular Devices (formerly Axon Instruments)	**http://www.moleculardevices.com**
Perkin Elmer (formerly Packard Biosciences)	**http://las.perkinelmer.com**
The Institute for Genomic Research (TIGR)	**http://www.tigr.org/software/tm4**
(f) Common RNA reference samples	
BD Biosciences (Clontech)	**http://www.bdbiosciences.com/clontech**
Stratagene	**http://www.stratagene.com**

[a] Provides links for clone distributors in Europe and the United States.

collections (Kim *et al*., 2003). In these cases, DNA fragments are prepared by PCR amplification from genomic DNA using gene-specific primer sets; the challenge in this approach lies in effective primer design to ensure efficient high-throughput PCR amplification.

A major shift toward printing long, unmodified oligonucleotide probes (generally 70-mers) instead of PCR products has emerged in the past couple of years owing to both an increase in the availability of genomic sequence and the decrease in the cost of oligonucleotide synthesis. Oligonucleotide libraries are an attractive alternative not only because clone sets and PCR reactions are eliminated but also because they provide more versatility in the types of projects that are possible to conduct using microarrays. For example, alternative splicing can be examined by designing probes for multiple exons within a given gene (Clark *et al*., 2002; Johnson *et al*., 2003); it is estimated that as many as 50% of human genes have splice variants and many of these may be relevant to human disease.

Melting-temperature matched libraries of 70-mers with high specificities for target genes are commercially available (Table 1a) or can be designed by individual investigators using software packages designed for this specific purpose (Rouillard *et al*., 2003; Wang and Seed, 2003). Direct comparison studies of spotted oligonucleotide and PCR product microarrays have shown that the two types of probes provide comparable sensitivity and specificity, and that there is an excellent correlation between expression ratios obtained using the two approaches (Wang *et al*., 2003; Woo *et al*., 2004).

3. Array printing

The most popular printing substrates are aminosilane-coated slides because they require no modifications in the DNA molecules for the efficient immobilization on the slide surface. Positively charged primary amine groups form ionic bonds with negative charges on the DNA phosphate backbone, and this electrostatic interaction is enhanced by DNA cross-linking to the slide surface using ultraviolet light and/or heat. Alternative attachment strategies include in-house coated poly-L-lysine slides and commercially available aldehyde and epoxide substrates (Table 1b). In addition to the amount of DNA immobilized on the slide surface, important criteria in evaluating printing substrates are spot morphology and background signal (see Section 5).

Spotted arrays are created by printing very small volumes of DNA material onto precise locations on glass substrates using robots (Table 1c) equipped with spotting pins. The choice of printing buffer, PCR purification protocol, temperature, and humidity are crucial to obtaining high-quality arrays (Hegde *et al*., 2000). Improvements in spotting pin technology, robot precision, and substrate chemistries have allowed investigators to print large quantities of high-density arrays (up to 60 000 elements) in as little as a few days.

4. Target preparation and hybridization assays

Labeled targets for two-color hybridization assays have typically been generated by incorporation of fluorescently labeled deoxynucleotide triphosphate (dNTP) analogs into reverse transcribed cDNA, but this direct incorporation strategy suffers from low incorporation efficiencies and dye bias due to the bulky nature of cyanine 3

(Cy3) and cyanine 5 (Cy5) dyes. In indirect or aminoallyl labeling, 5-aminoallyl-dUTP is incorporated into cDNA during reverse transcription (RT), and this is followed by a chemical coupling reaction of Cy dye esters to the aminoallyl linker. The third approach, dendrimer labeling, differs from both direct and indirect incorporation methods in that it is not dependent on modified dNTP incorporation but rather relies entirely on hybridization kinetics. Both aminoallyl and dendrimer labeling protocols offer the advantages of improved incorporation efficiencies and reduced dye bias as compared to direct incorporation (Manduchi *et al*., 2002).

To facilitate studies of small numbers or even single cells in applications such as fine needle aspirate biopsies, laser-capture microdissected, or flow sorted cells, RNA amplification protocols have been developed and optimized for microarray purposes by a number of laboratories and commercial entities (Table 1d). These protocols are based on the T7-linear amplification method first described by Eberwine (Van Gelder *et al*., 1990). The main concern in hybridizing amplified products is that the original transcript representation may be distorted during the amplification process. Several studies have shown that transcript representation is preserved during amplification, thus validating the use of RNA amplification schemes in microarray analysis (Iscove *et al*., 2002; Zhao *et al*., 2002). The amplification strategy has facilitated the first microarray profiling of human blood samples (Whitney *et al*., 2003), an important step toward bringing gene-expression profiling into the clinic.

Cy3- and Cy-5-labeled samples are cohybridized to the immobilized PCR products or oligonucleotides in a competitive hybridization assay. Following a prehybridization step to block nonspecific binding to the positively charged substrate surface surrounding the DNA spots (Hegde *et al*., 2000), labeled targets are incubated on the slide overnight. After a series of washing steps to remove unbound targets, arrays are scanned using instruments equipped with scanning confocal lasers, and image analysis and feature extraction are performed using either software provided with the scanner or other software packages (Table 1e).

5. Quality control

Quality control measures have been implemented in most laboratories for both array production and hybridization assays. Several slides from each printing batch are usually stained with an intercalating fluorescent dye that is detected either in the green or in the red channel (Wang *et al*., 2003). Screening the slides in this way allows for the detection of problems that may have occurred during the printing run, such as a hung up pin or a change in humidity or temperature. Spot morphology is also indicative of substrate quality and any defects in the slide surface are also diagnosed using this DNA staining technique.

Most laboratories have also adopted a practice of printing positive and negative controls in multiple positions on each slide. Negative controls are important for assessing the level of nonspecific hybridization and include spotting buffer, polyA DNA, salmon sperm DNA, and random oligonucleotide sequences. Positive or spike-in controls are genes from a different species that do not have sequence homology to the species of interest (e.g., *Arabidopsis thaliana* for mammalian

arrays). cRNA for these genes is spiked into the labeling reactions at different relative copy numbers, and any deviations from expected ratios are indicative of problems in hybridization conditions (Wang *et al*., 2003). Printing controls at multiple positions on each slide allows for detection of any spatial gradients and artifacts. Finally, a number of quality control metrics for the image processing step have been proposed (Brown *et al*., 2001; Wang *et al*., 2001), and almost all image analysis software packages have incorporated some type of quality control measure.

6. Experimental design

One of the biggest challenges in applying microarray technology is the question of experimental design. Investigators need answers to questions such as which samples should be cohybridized, how much replication is necessary, robustness of design, dye bias, sample size, and statistical power needed to ensure that the experimental objectives are achieved. Although much has been learned over the years through trial and error and statisticians have provided recommendations based on mathematical models (Churchill, 2002; Dobbin *et al*., 2003; Yang and Speed, 2002), there is still no simple answer for how a given experiment should be designed. Experimental design will depend heavily on the type of study and the amount of funding available for microarray analysis.

In the simplest design, a query sample (treated in some way) is compared to the reference (untreated) sample on a single array. However, most projects require comparisons of many samples across many arrays, and this type of analysis is facilitated by the use of a common reference. Each sample of interest is cohybridized with the reference sample, which can be either a pool of all samples included in the project or an unrelated "universal" reference sample. Universal RNA reference pools are constructed from a large number (10 or more) of cell lines or tissues and are commercially available (Table 1f). The idea underlying such a design is that by combining RNAs from diverse cell lines, one might obtain a more complete representation of the genes spotted on the array. An alternative to the common reference design is "loop" design, in which each experimental sample is hybridized with two other experimental samples to form a loop (Churchill, 2002). The advantage of the loop design is that more data are collected on samples of interest instead of an irrelevant reference sample. The disadvantage is that information is lost for many genes when samples that are further apart in the loop are compared. Furthermore, data coming out of loop designs are more complex, thus requiring specialized analysis tools.

In addition to careful experimental design, incorporating replicate experiments allows researchers to deal with the unwanted sources of variation in their microarray experiments. Spotting the same DNA on each slide multiple times is one type of replication, but within-arrays replicas have generally shown good correlations and thus have often been eliminated to make room on the array for additional sequences that provide better genome coverage. Hybridization assay replicas include both technical replicates, in which the same RNA sample is labeled and hybridized multiple times, and biological replicas, in which RNA samples from multiple cell

line growths or individuals are used. A special type of technical replication is dye flip assays; in this case Cy3 and Cy5 labels are reversed to eliminate any dye bias. The most commonly asked question in experimental design is how much replication is necessary. Several recommendations have been made recently based on statistical models derived from the body of data available to date (Dobbin *et al*., 2003; Pavlidis *et al*., 2003; *see also* Article 52, **Experimental design**, Volume 7).

7. Data analysis

After data have been collected and images processed, the first step in the analysis of two-color array data is to normalize the relative fluorescence intensities in Cy3 and Cy5 channels. Normalization and data transformation are dealt with elsewhere in this encyclopedia and are only discussed briefly in this chapter. Normalization adjusts for experimental variation such as differences in labeling and detection efficiencies for the fluorescent dyes and for differences in the quantity of input RNA from the two samples examined in the assay. A number of global normalization approaches have been described (Quackenbush, 2002); however, none of these methods take into account systematic, intensity-specific biases that are often seen in the data. Locally weighted linear regression (lowess) (Cleveland and Devlin, 1988) analysis has been shown to successfully remove intensity-specific effects in Cy5/Cy3 ratios (Yang *et al*., 2002b), and lowess normalization has since become the most widely used approach for balancing the two channels. This method can be applied to either the entire array or subsets of elements based on the position to eliminate any spatial effects that may appear in the data. Once data have been normalized, we can begin to make meaningful comparisons of expression levels across many experimental conditions.

Microarray analysis is used to ultimately answer the question of which genes are differentially regulated between control and test groups. Deciding what the cutoff value used to distinguish differential expression from natural variability in the data is a challenging task. A cutoff of twofold up- or down-regulation was chosen to define differential expression in early microarray studies (Schena *et al*., 1996). However, using a single cutoff value may not be appropriate since background noise is much greater, relative to signal, at lower fluorescence intensities. Several groups have therefore proposed and tested methods for intensity-dependent estimation of differential expression (Baggerly *et al*., 2001; Ideker *et al*., 2000; Yang *et al*., 2002a). Statistical testing has also become a widely used approach to defining differential expression across many experiments. Both existing statistical methods, such as analysis of variance (ANOVA) (Cui and Churchill, 2003), and algorithms developed specifically for microarray data (significance analysis of microarrays or SAM) (Tusher *et al*., 2001) have been successfully applied in microarray data analysis. Detailed description of application of statistical methods to gene-expression analysis can be found in Article 53, **Statistical methods for gene expression analysis**, Volume 7 and Article 59, **A comparison of existing tools for ontological analysis of gene expression data**, Volume 7. Subset(s) of genes that pass the test, that is, are significantly differentially expressed between experimental

groups, can then be used for further analysis and data mining (Quackenbush, 2001; Slonim, 2002).

8. Validation

It has become a common practice to validate relative expression levels of a small number of genes identified as differentially regulated in a microarray experiment by either more traditional techniques such as semiquantitative RT-PCR, ribonucleotide protection assay (RPA), and Northern blot analysis, or by recently developed real-time quantitative RT-PCR assays (Chuaqui *et al.*, 2002). The conclusion drawn from a large number of studies to date is that expression ratios obtained from microarrays and single-gene techniques are in agreement in terms of the direction of change (up- or downregulated) but frequently differ in detected fold changes. Investigators often choose the genes with the highest differential expression ratios for further study, but this strategy may overlook potentially interesting genes that have small fold changes in a microarray experiment but could be more differentially expressed when studied using other approaches. A better alternative for selecting genes to be validated is based on their function and relevance to the biological process that the investigator is addressing in the study.

9. Conclusions

Since their conception 10 years ago, spotted microarrays have evolved into a tool that produces reliable gene-expression data at the whole-genome level. Availability of genomic sequence combined with improvements in substrate chemistries and spotting technologies have led to the ability to print probes for entire genomes at least once on a single slide. Advances in protocols for RNA amplification and target labeling allow researchers to obtain high-quality data from samples with limited amounts of RNA. Finally, sophisticated algorithms for normalization, integration of replicate experiments, and identification of differentially expressed genes have been developed to deal effectively with undesirable technical variability in microarray data. The challenge in the field of spotted microarrays now lies primarily in data mining and integration of gene-expression profiles, with findings from other "-omic" sciences into a system-level understanding of biological systems.

References

Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV and Zhang W (2001) Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology*, **8**, 639–659.

Boguski MS and Schuler GD (1995) ESTablishing a human transcript map. *Nature Genetics*, **10**, 369–371.

Brown CS, Goodwin PC and Sorger PK (2001) Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8944–8949.

Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, Hewitt SM, Phillips JL, Krizman DB, Tangrea MA, Ahram M, *et al*. (2002) Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*, **32**(Suppl), 509–514.

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, **32**(Suppl), 490–495.

Clark TA, Sugnet CW and Ares M Jr (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.

Cleveland W and Devlin S (1988) Locally weighted linear regression: an approach to regression analysis by local fitting. *Journal American Statistical Association*, **83**, 596–609.

Cui X and Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, **4**, 210.

DeRisi JL, Iyer VR and Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dobbin K, Shih JH and Simon R (2003) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute*, **95**, 1362–1369.

Hegde P, Qi R, Abernaty K, Gay C, Dharap C, Gaspard R, Hughes JE, Snesrud E, Lee N and Quackenbush J (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 552–562.

Ideker T, Thorsson V, Siegel AF and Hood LE (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805–817.

Iscove NN, Barbara M, Gu M, Gibson M, Modi C and Winegarden N (2002) Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnology*, **20**, 940–943.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R and Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.

Kim H, Snesrud EC, Haas B, Cheung F, Town CD and Quackenbush J (2003) Gene expression analyses of Arabidopsis chromosome 2 using a genomic DNA amplicon microarray. *Genome Research*, **13**, 327–340.

Lennon GG, Auffray C, Polymeropoulos M and Soares MB (1996) The I.M.A.G.E. consortium: an Integrated Molecular Analysis of Genomes and their Expression. *Genomics*, **33**, 151–152.

Manduchi E, Scearce LM, Brestelli JE, Grant GR, Kaestner KH and Stoeckert CJ Jr (2002) Comparison of different labeling methods for two-channel high-density microarray experiments. *Physiological Genomics*, **10**, 169–179.

Pavlidis P, Li Q and Noble WS (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Quackenbush J (2001) Computational analysis of microarray data. *Nature Reviews Genetics*, **2**, 418–427.

Quackenbush J (2002) Microarray data normalization and transformation. *Nature Genetics*, **32**(Suppl), 496–501.

Quackenbush J, Liang F, Holt I, Pertea G and Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, **28**, 141–145.

Ramaswamy S and Golub TR (2002) DNA microarrays in clinical oncology. *Journal of Clinical Oncology*, **20**, 1932–1941.

Rouillard JM, Zuker M and Gulari E (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Research*, **31**, 3057–3062.

Schena M, Shalon D, Heller R, Chai A, Brown PO and Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10614–10619.

Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, **32**(Suppl), 502–508.

Staudt LM and Brown PO (2000) Genomic views of the immune system. *Annual Review of Immunology*, **18**, 829–859.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD and Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 1663–1667.

Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J and Lee NH (2003) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology*, **4**, R5.

Wang X, Ghosh S and Guo SW (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, **29**, E75.

Wang X and Seed B (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.

Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA and Brown PO (2003) Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1896–1901.

Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J and Churchill G (2004) A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *Journal of Biomolecular Techniques*, **15**, 276–284.

Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, *et al*. (2002a) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology*, **3**, research0062.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002b) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, e15.

Yang YH and Speed T (2002) Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**, 579–588.

Zhao H, Hastie T, Whitfield ML, Borresen-Dale AL and Jeffrey SS (2002) Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. *BMC Genomics*, **3**, 31.

Using oligonucleotide arrays

Lindsay W. Mitchell and Eric P. Hoffman
Children's National Medical Center, Washington DC, USA

1. Introduction

Manufacture of nucleic acid microarrays typically involves the placement of purified DNA fragments onto specific "addresses", or locations, of an underlying solid support. The DNA fragments can be double-stranded PCR products, cloned plasmids, or shorter single-stranded oligonucleotides (25–70-mers). In each case, every specific purified DNA on the solid support comprises a "feature" of the array and an array may contain thousands or millions of different features, each having an independent specific address on the microarray. Complex solutions of mRNAs from cells or tissues are labeled (typically directly or indirectly labeled with fluorescent dyes) to enable detection by laser scanning. Labeled solutions are then placed onto the microarray, and hybridization between the solution-based probe and immobilized target-sequence on the microarray occurs. Laser scanning is then used to quantify the amount of labeled mRNA on each feature, and normalization of resulting signals allows derivation of a relative concentration of each mRNA in the original solution. Sensitivity and specificity of microarray assays depend on many factors including the length of the target sequence, GC content, and consistency and efficiency of mRNA labeling. This review focuses on Affymetrix arrays, a particular, but commonly used microarray platform for mRNA profiling (see **http://www.affymetrix.com**).

Other important applications of microarrays include analysis of genomic DNA. Linkage analysis of disease traits within families can be accomplished using microarrays that detect polymorphisms throughout the human genome. In this application, 25-mer oligonucleotides are designed to detect single nucleotide polymorphisms (SNPs) by differential hybridization of the genomic DNA probe to the allele-specific target sequences. The ratio of hybridization signals to features querying each of the two alleles of an SNP (e.g., either G or T polymorphism) determines the genotype of that subject's DNA. Current Affymetrix arrays for detecting polymorphisms query 100 000 distinct positions in the genome (loci) using a two microarray set, providing coverage of the genome with a polymorphism roughly every 30 000 bp (*see* Article 77, **Genotyping technology: the present and the future**, Volume 4). Comparative genomic hybridization (CGH) queries the relative amount of each of the 100 000 loci (independent of the genotype at the polymorphic site), so these same "SNP chips" can be used to scan for deletion or amplification of regions of the genome in patient tumors as well (*see* Article 93, **Microarray CGH**, Volume 4).

All microarrays work on the same principle of hybridization of a complex solution of labeled nucleic acids (mRNA or genomic DNA), to specific sequences placed in specific addresses on a solid support (the array). For this reason, the data obtained by the 25-mer in situ synthesized Affymetrix arrays is, in principle, similar to the data obtained by mechanically spotted cDNA microarrays (PCR products), or mechanically spotted arrays of longer, machine-synthesized oligonucleotides (e.g., 70-mers; see **www.appliedbiosystems.com**). In practice, however, there are some clear distinctions between these experimental platforms that serve to make the data obtained quite distinct. First, mechanically spotted arrays of oligonucleotides and PCR products (cDNA arrays) typically suffer from greater variation in the amount of nucleic acid spotted and retained on the surface of the solid support than do in situ synthesized arrays. Thus, it is important to have methods of normalizing the resulting hybridization signals to the amount of target in the feature, and this is typically executed by cohybridizing a "control" RNA sample labeled with a different color fluorescent molecule followed by derivation of a ratio of hybridization signals (experimental/control) for each feature. In contrast, Affymetrix arrays are chemically synthesized in situ using photolithography, resulting in very consistent array-to-array amounts of target nucleic acid at each feature. Thus, RNA samples can be hybridized individually (no cohybridization) and signals are normalized across the entire array or across all arrays within a project. Generally, there are fewer technical sources of variation with Affymetrix arrays than with mechanically spotted arrays. It should be noted that there are other in situ synthesized microarrays, most notably from Agilent, whose arrays contain a single 60-mer probe for each gene; these arrays are generally used in two-color assays.

A second major distinction between platform types concerns the number of measurements taken per transcript, sometimes referred to as "redundancy" or repeated measurements. Spotted oligonucleotide or cDNA arrays typically use one feature per transcript measured. Thus, the resulting signal is generated from a single fluorescent measurement. Affymetrix arrays use "probe sets", which contain multiple perfect match (PM) and mismatch (MM) oligonucleotide features (targets) querying each transcript. Current arrays use 11 perfect match and 11 paired mismatch (a single substitution relative to the reference in the center – 13th position – of the oligonucleotide) for each probe set (Figure 1). As a result, there are 11 relatively independent measurements that are used to derive a single "signal" for a corresponding mRNA as well as 11 estimates of the corresponding background. This repeated measurement provides better control over technical variation (feature-to-feature variability), and variations in performance (hybridization efficiency, etc.) of different features detecting a specific mRNA. The interpretation of probe set–generated data (probe set algorithms) has become an active area of research, as covered in more detail below. The third distinction between platform types is the relative ease of providing quality control, standard operating procedures, and database resources for standardized, factory-manufactured platforms such as Affymetrix microarrays (see Tumor Analysis Best Practices Working Group, 2004). While spotted oligonucleotide and cDNA microarrays are much more flexible in their content, this same asset becomes a liability when attempting to standardize data reporting and interpretation between

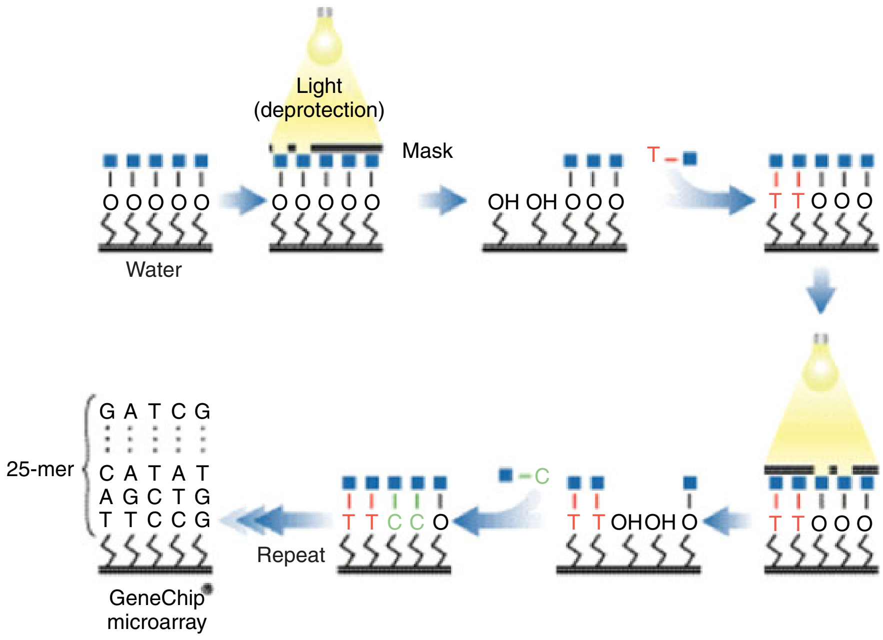


Figure 1 Photolithographic masks are strategically placed so that only features requiring the nucleotide being added are deprotected and made ready for nucleotide binding. (Reproduced by permission of Affymetrix, Inc.)

different laboratories (*see* Article 91, **Creating and hybridizing spotted DNA arrays**, Volume 4).

In what follows, we focus on Affymetrix arrays, describing microarray manufacture, use, and data interpretation.

2. Production of Affymetrix arrays

Manufacture of Affymetrix microarrays begins with a piece of quartz. A washing process serves to render natural surface hydroxylation uniform so the quartz can then be bathed in silane that reacts with the hydroxyl groups to create a network of covalently linked molecules. This network serves as the platform for the synthesis of anchored oligonucleotides. The distance between covalently linked silane molecules is the limiting factor in determining what synthesis density is possible on the array. (Current Affymetrix microarrays contain over 1 million features in a 1.2 cm^2 solid support, and this density is increasing rapidly.) Photoresponsive linker molecules are then attached to the silane and photolithographic masks are strategically placed over the surface of the quartz with gaps in all locations where growing oligonucleotides require addition of whichever nucleotide base is to be included in the next wash. For each step of nucleotide addition, masks are placed in different locations to allow different oligos to be synthesized in parallel. UV light above the mask passes through holes in the mask to activate discrete regions on the surface (Figure 1) by deprotecting linker molecules, making them receptive to nucleotide

coupling. Once the placement of masks, UV excitation, and nucleotide wash is repeated a sufficient number of times to generate the desired ~25-mer probes at each feature, the quartz is placed in flow cell cartridges for ease of use in the hybridization and staining wash steps. Each batch of arrays synthesized in this manner is subject to hybridization with control cocktails to assess the quality of the manufacture process.

The selection of the 25-mer sequence to be synthesized at each array address is a critical part of microarray design. Each 25-mer is only a small fraction of the complete mRNA being queried, thus, the selection of the best set of well-matched sensitive and specific probes is important for the accurate measure of transcript abundance. To make this decision, Affymetrix compiles sequence and annotation details from databases such as GenBank, RefSeq, and dbEST (see **http://www.ncbi.nih.gov**) and uses information therein to group transcripts into similar clusters. These clusters are aligned to the human genome to identify which transcripts are splicing and polyadenylation variants of one another. On the basis of this information, Affymetrix either chooses one representative "exemplar" transcript to use in probe set design for a given gene or selects a series of consensus sequences in cases when a superior representative transcript choice is not available. The 3′ and 5′ orientation of the exemplar and consensus sequences is then determined, and in cases in which this cannot be unequivocally established, probe sets are designed for both strand orientations. Of all the possible 25-mer probes that could query an exemplar or consensus sequence, a few are selected that best fit several criteria. First, the probes must have minimal possible chance of cross-hybridization with other, unrelated transcripts. Second, probe sets are selected that are located at the 3′ end of the transcript for compatibility with the poly-A-based target preparation (amplification) process described in the next section. Finally, probe sets are evaluated for anticipated hybridization efficiency under the given pH, salt, and temperature conditions used in assays.

Similar hybridization microarrays are used for purposes other than mRNA profiling. Affymetrix microarrays are available that detect DNA polymorphisms (single-nucleotide polymorphism genotyping, or "SNP typing"), or provide DNA sequence from a patient (resequencing arrays). The microarrays that query DNA sequence are also produced in the manner outlined above and work by detecting the amount of hybridization from probes in solution to target sequences on the quartz surface. The major distinction is that relative hybridization to different sequences at a single base position are detected. For example, to genotype or sequence DNA segment X, four 25-mer features are synthesized with sequence variation only in the center of the oligo: . . . ATA . . . ; . . . AGA . . . ; . . . ACA . . . ; . . . AAA . . . If the probe hybridizes best to the . . . AGA . . . feature, this indicates that the correct DNA sequence in that patient contains Guanine in the position being queried.

3. Target preparation and hybridization for mRNA profiling

Preparation of target samples for use with a gene expression array first involves the isolation of total RNA from the cells or tissue of interest. This entails homogenizing

the tissue or lysing the cells in combination with chaotropic agents to disperse the tissue and prevent RNA degradation. RNA is then extracted from this lysate using phase separation with chloroform followed by cleanup to remove DNA and protein contamination. Following isolation, RNA is quantified by spectrophotometer UV absorption at 260 wavelength. Quality is assessed by determining the integrity of the 18s and 28s ribosomal RNA bands by gel electrophoresis or other similar approaches.

The total RNA is then used as starting material for amplification and labeling prior to hybridization to the microarrays. Labeling is executed to enable detection of the bound probe on the microarray feature after hybridization. Amplification is necessary to increase the amount of mRNA present in isolated total RNA to a quantity sufficient for effective hybridization and detection (at least 400 times the starting amount). Depending on the amount of RNA isolated, one (~1–15 μg starting material) or two (~10–100 ng of starting material) rounds of amplification are required. While the principles in play are identical for both one- and two-round amplification protocols, maintenance of transcript integrity is generally poorer in samples processed through two rounds, making more lenient QC necessary and complicating the comparison of data generated by one- and two-round protocols.

In both cases, the first step of amplification is the generation of complementary DNA from polyadenylated transcripts, which correspond to protein-coding sequences (mRNA), as distinguished from transfer RNA (tRNA), ribosomal RNA (rRNA), and other RNA species (hnRNA). This is accomplished through the use of an oligo(dT) primer that binds to the mRNA poly-A tail and primes reverse transcription into a DNA/RNA hybrid molecule using reverse transcriptase (Figure 2). DNA polymerase I and random primers in combination with RNase H are then used to replace the RNA strand with DNA, leading to the synthesis of double-stranded cDNA. The amplification step follows, with the cDNA serving as a template for T7 RNA polymerase, leading to production of 400 or more RNA molecules from each cDNA molecule (Figure 2). Biotinylated nucleotide precursors are included in the reaction, so that the resulting copy RNA (cRNA) is internally labeled with biotin (Figure 2). If the first round of cRNA production does not lead to adequate amounts of probe, then the process can be repeated (cRNA to cDNA to second round cRNA). All amplification procedures introduce some bias, however, as some transcripts amplify better than others. Thus, it is essential to follow the same protocol for samples that will be compared as evidence suggests this bias is sequence specific.

Quantity and quality of amplified, labeled cRNA is determined before fragmentation (UV absorbtion at 260 and cRNA smear between 100–300 bp on an agarose gel). 15–20 μg of quality cRNA is fragmented by heating to produce ~50 bp segments. This size is optimal for interaction of target with probe and successful consistent hybridization. Because the targets themselves are only 25-mer in length and query a unique part of the entire transcript, maintenance of the entire transcript as a complete unit is unnecessary for measuring transcript abundance. Fragmented cRNA is checked again by gel electrophoresis to confirm size (~50 bp) and is then combined with reagents for array hybridization. Also included in the hybridization reagent mix is Oligo B2, a synthetic oligonucleotide that binds to probes located around the perimeter of the array and arranged in a checkerboard

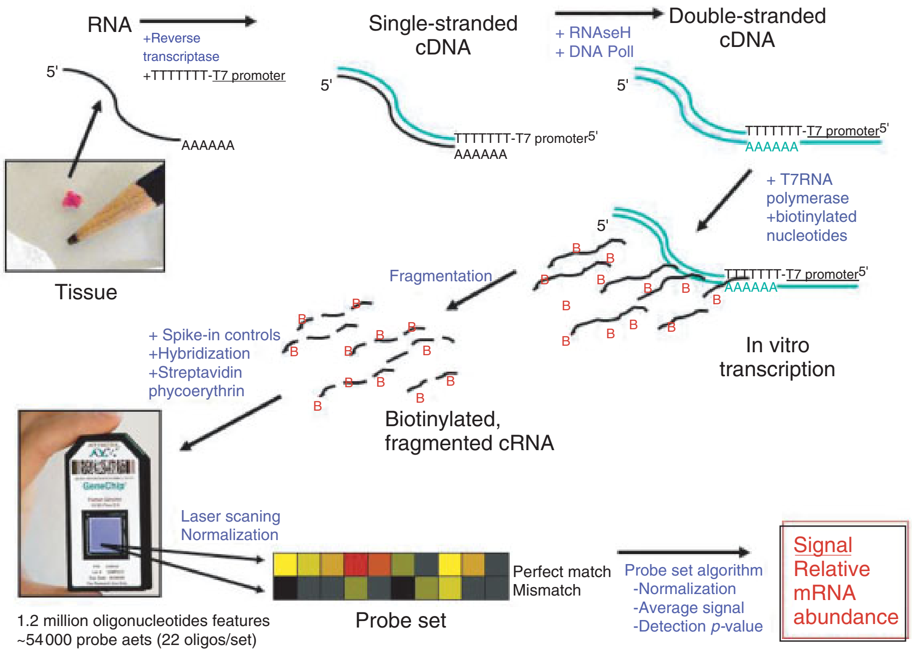


Figure 2 The steps used to process tissue or cells through RNA isolation, cDNA synthesis, cRNA amplification, labeling, hybridization, and image generation of affymetrix microarrays

pattern at each corner. These oligos assist in alignment of an interpretive software grid with fluorescent signal on the array.

Hybridization cocktails containing fragmented, biotin-labeled cRNA, spike-in controls, and buffers are introduced into the flowcell casing that encloses the array. The arrays are then incubated for 16 h at controlled temperature with mixing to allow the hybridization of the labeled probes in solution to the target feature on the microarray surface. After incubation, the hybridization cocktail is removed, the array washed, and the bound biotinylated probes are stained with streptavidin (biotin binding) complexed with the fluorescent molecule phycoerythrin. This streptavidin-phycoerythrin antibody complex (SAPE) provides a fluorescence-based "report" of how much transcript is bound to any given feature on the array. This fluorescent report can be amplified by subsequent staining with biotinylated goat anti-streptavidin antibody and then again with phycoerythryn-labeled streptavidin antibody. The amount of fluorescence on the array is assessed before and/or after amplification with a laser scanner also manufactured by Affymetrix (Figure 2).

4. Data analysis; brief overview: feature – probe set – microarray – project – interproject

The scanned image of a hybridized array is stored as a pixel-based image file (.DAT) and this image is processed to determine the amount of fluoresence signal

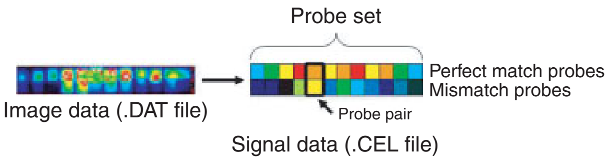


Figure 3 The anatomy of a probe set. Each mRNA transcript is queried by a series of oligonucleotides, with either perfect match or mismatch design against specific 25-bp regions of the transcript. Typically, 11 probe pairs (perfect match and mismatch partners) are used for each probe set. The resulting scanned image of a probe set is shown in the lower left panel, with the resulting normalization and averaging shown in the right panel pair. Signal data for each feature is stored in a .cel file, and probe set algorithms typically use this file to derive a single expression "signal" for each probe set

at each of the 1 million features (.CEL file) (Figure 3). This data is then processed to derive a single, numerical signal for each probe set. As discussed above, one of the advantages of the Affymetrix platform is the multiple measurements that are taken for each transcript; 11 probe pairs (composed of perfect match and mismatch partners) are designed against distinct parts of the mRNA to be queried (Figure 3). The mismatch probe contains a single substitution in the center position, which destabilizes specific hybridization, providing an estimate of nonspecific hybridization to the perfect match probe. Thus, the simplest measure of transcript abundance is to subtract the mismatch fluorescence signal from the perfect match for each of 11 probe *pairs* within a probe *set* and then average and normalize the signal for each probe set within the *microarray* (Figure 3). The normalized "signal" for each transcript (probe set) is then a single number that can be compared between microarrays within the same *project*, or between different projects (though consistent laboratory protocol is crucial for *interproject* comparison). In this flow of data interpretation, one progressively works from the feature – to probe pair – to probe set – to microarray – to project – to interproject analysis.

There are many approaches to both derivation of a single "signal" from a probe set, and also many methods of normalization of signals so that different microarrays can be accurately compared to each other. Methods for deriving signal are typically called "probe set algorithms", and these aim to extract the greatest amount of accurate information content from a probe set (sensitivity) while also reducing the amount of noise (specificity) (*see* Article 54, **Algorithms for gene expression analysis**, Volume 7). We have recently compared a number of probe set algorithms, and have found that different algorithms are optimal depending on the specific application. For example, a more "stringent" probe set algorithm may provide better data analysis in a "noisy" experiment where there are many uncontrolled confounding variables. On the other hand, an algorithm that ignores the mismatch signal can be much more sensitive in detecting low-level hybridization signals (Seo *et al*., 2004).

To briefly describe the logic of a particular probe set algorithm (MAS5.0 provided by Affymetrix), the first step in calculating expression value (probe set signal) is an estimation of background signal (fluoresence) present on the array. The array is divided into 16 regions and the lowest 2% of signals in that region are taken to represent background signal in that zone. The finalized background value takes into

account the background in all other zones (each weighted by its distance from the zone in question), in order to smoothen background estimations across the entire array. Probe signal intensities in each zone are then adjusted by subtraction of that zone's background from the intensity of each feature contained therein. A threshold is set to ensure that this subtraction does not result in any negative intensity values. These background-adjusted intensities can then be used to calculate gene expression values. As described above, the basic principle involves finding the difference in intensity between each PM and MM probe *pair* and averaging the differences of every probe pair in a given probe *set*. This results in a single "average difference" value for each probe set and represents a relative level of expression of the gene queried by that particular probe set.

In addition to providing this numerical representation of transcript behavior, the MAS5.0 software also provides an "absolute" call for each probe set. This metric results in every probe set being labeled as "Present", "Absent", or "Marginal" and serves as an indicator of the reliability of the numerical expression value assigned to the probe set. To assign an absolute call to a probe set, an absolute call is first assigned to each probe pair within that probe set. A probe pair is considered Present if its difference (PM-MM) is greater than or equal to a set threshold and if its ratio (PM/MM) is greater than or equal to a second set threshold. A probe pair is considered Absent when neither of these conditions is met (MM-PM and MM/PM are both greater than or equal to the set thresholds). The difference and ratio thresholds are established on the basis of calculated noise (defined as the variation in digitized signal of the laser scanner as it reads intensity from the array surface) and whether the array was stained once with phycoerythryn streptavidin or amplified by staining twice. On the basis of the absolute call of each probe *pair*, an overall absolute call for the probe *set* can be calculated. Three values are used in making this determination. The first value is the ratio of probe pairs called Present to those called Absent. The second value is the positive fraction, or the number of probe pairs called Present divided by the total number of probe pairs in the probe set. The third value, Log Average Ratio, which takes into consideration the numerical expression value, provides an estimation of the influence of cross hybridization. (The Log Average Ratio is calculated by averaging the log of PM/MM values of all the probe pairs in the probe set and multiplying by 10). These three parameters, Positive/Negative Ratio, Positive Fraction, and Log Average Ratio, are considered in the context of a scale with user defined thresholds below which a probe set is called Absent, above which a probe set is called Present, and between which a probe set is called Marginal. It has been demonstrated recently that cross-hybridization on arrays is largely dependent on the GC content of each particular probe. For this reason, a proposal for newer versions of Affymetrix arrays includes implementation of more general cross-hybridization assessment in place of the sequence-specific MisMatch probes that are currently employed. Each probe would then be compared to one of several general cross-hybridization controls that matched its particular GC content profile. This approach would free space on the array for additional target querying probes while still maintaining effective assessment of noise and probe specificity.

MAS5.0 software also includes array-wide normalization of expression values to a user-defined target intensity value. This normalization maintains all *relative*

relationships between probe sets/transcripts on an array but brings them within the user specified scale. This allows arrays normalized to the same target intensity to be more effectively compared to one another. The software reports the scaling factor that was used to adjust all intensity values on the array closer to the set target intensity, which allows the user to identify arrays within a project for which more extreme adjustment was necessary (indicating possible problems with target transcript integrity, hybridization etc.).

The standardized Affymetrix microarrays permit the development of robust probe set algorithms, and it should be noted that there are a number of other widely used algorithms for estimating expression on the basis of probe set intensities, including the method of Li and Wong in the dCHIP software package (Li and Wong, 2001) and the Robust Multichip Average algorithm (Irizarry *et al*., 2003) in the BioConductor suite of tools (**http://www.bioconductor.org**). Standardization has also enabled the development of quality control and standard operating procedures (QC/SOP). A recent publication explains standard QC/SOP suggestions for Affymetrix arrays, and the reader is referred to this publication for further explanations and metrics (Tumor Analysis Best Practices Working Group, 2004).

5. Project analysis software for microarray projects

Determination of numerical expression values (normalized "signal") for probe sets is the initial step to extracting biological information from Affymetrix array data. After definition of signals using one or more of the probe set algorithms, data analysis then turns to interpretation of the project (microarray comparisons) (*see* Article 57, **Low-level analysis of oligonucleotide expression arrays**, Volume 7). Typically, there are a number of biological replicates (two or more microarrays representing the same "group"; for example, a control or experimental group), and any number of groups within a project. A relatively simple project would have just two groups; for example, "normal muscle at rest" and "exercised muscle", with two or more replicates per group. This basic experimental design can be expanded to a time series (time points after a bout of exercise), or to multiple groups (endurance training, power training, weightless in space).

For an individual with considerable Microsoft Excel or other spreadsheet savvy, it is feasible to execute some manipulation of expression value data by writing and implementing one's own equations for normalization and statistical comparisons. However, because this is beyond the scope of many researcher's time and other constraints, and because the Affymetrix manufacturing consistency makes it possible, most array data analysis is executed with the assistance of software programs such as MAS5.0 (**http://www.affymetrix.com**), GeneSpring (**http://www.silicongenetics.com**), dCHIP (**http://www.biostat.harvard.edu/complab/dchip**), SpotFire (**http://www.spotfire.com**), and HCE (**http://www.cs.umd.org/hcil/hce**), among others. While there is a large degree of overlap in the features offered by each of these programs (as all endeavor to be stand-alone analysis packages), each tend to have their own strengths and can usually be used in combination to make use of these unique advantages. MAS5.0, GeneSpring, and SpofFire are commercially

available software packages, and the reader is referred to the corresponding website for more information. Here, we briefly describe two freely available software packages, dCHIP (Li and Wong, 2001) and HCE (Seo *et al*., 2004).

dCHIP software (**http://www.biostat.harvard.edu/complab/dchip**) focuses on calculating expression values and absolute calls (present/absent), but takes a "project-based" view, rather than a "single microarray" view. As such, it analyzes all microarrays within a project simultaneously, thereby detecting "chip outliers" (a microarray that is too dissimilar from others in the project), and providing project-based normalization. Whole project filtering is made possible allowing the user to create lists of genes that meet certain criteria. Filtering criteria offered in the current version of dCHIP include: selection of genes whose variation across all arrays falls within a certain range, selection of genes that are called Present in at least some minimum percentage of arrays, and selection of genes whose raw expression value exceeds a certain threshold in a certain percentage of arrays. This sort of filtering allows the user to pare away what is uninteresting from the 20 000+ probe sets for which data has been collected. In addition, dCHIP also allows the user more versatile tools for executing statistical comparisons of arrays. The user is able to specify any group of arrays to be compared to any other group of arrays and the program offers several different criteria by which to execute the comparison. The first selects genes which have a "fold change" greater than a specified value when the groups are compared (expression values within groups are averaged to execute the comparison). The second criteria selects genes with an absolute difference greater than or less than a specified threshold when the groups are compared. A third identifies genes that are called present in a at least a certain percentage of the groups being compared. The software also allows selection of genes with p-value less than a specified threshold as determined by either standard or paired t-test analysis. One major advantage of this statistical comparison package is that it allows several comparisons to be executed simultaneously through an option to combine comparisons. For example, the user could select all genes passing a certain fold change threshold between two groups that also pass a fold change threshold between two entirely different groups.

The outputs of all such filtering and statistical comparison executed in dCHIP are in the format of lists, which can be linked with ontological data for each probe set provided that the user downloads such information into the program. There are not, however, direct links to annotation data bases. Another limitation of the program is that there is no corresponding visual representation of these filtering and comparison results. dCHIP does contain a PM/MM data view interface that allows the user to obtain very detailed information about the behavior of each probe set, but the user must toggle between views to observe the behavior of any given probe set from one array to another. This sort of data visualization is more useful as a retrospective analysis tool to confirm the behavior of probe sets deemed to be interesting by the user via preliminary analysis within dCHIP or other analysis programs. dCHIP also contains an option for unsupervised sample and gene clustering.

HCE, or Hierarchical Clustering Explorer (**http://www.cs.umd.org/hcil/hce**) is focused on data visualization and interpretation, with the goal of optimizing human/computer interaction. It is a downstream analysis program in that it requires

prior calculation of expression values (signals) by another program. HCE also lacks a provision for filtering or comparison-based selection/exclusion of genes so this, too, if desired, must be executed prior to loading data into the program. Once data is loaded into HCE, it is normalized for visualization, with allowance for user adjustment. The main HCE display consists of sample or gene clustering. Before executing the clustering, the user is allowed to specify which linkage and similarity measures are to be used, how the dendrogram should be organized as it is being created, and which samples to include data from when executing the clustering. The main display is interactive, allowing the user to zoom in on heatmap/dendrogram areas of interest, and to hide features that are uninteresting. Such data display allows visual identification of gene expression patterns of interest such as a region of the heat map indicating upregulation of a cluster of genes in treatment but not in control samples and lists of the genes in regions of interest can be exported for further analysis.

The unique aspects of HCE that set it apart from other analysis programs are those referred to as “rank-by-feature” functions. These functions were implemented with the aim of simplifying multidimensional data analysis. Any data that has more than three dimensions (expression array data, economic trend data, etc.) becomes difficult to visualize and often interesting correlations and relationships among the various dimensions go unnoticed. HCE attempts to overcome this difficulty somewhat by assisting the user in executing a brief but thorough survey of the data to identify the “shape” of the data. To this end, the data can first be examined in one dimension with the program’s Histogram Ordering function. Within this function, the user examines histograms for each array, where each data point is representative of a single probe set expression value. These histograms can then be ordered according to criteria such as normality, uniformity, number of potential outliers, number of unique values, and size of biggest data gap. Color coding of lists allow the user to readily identify which arrays are at either extreme with respect to the currently selected criteria (i.e., which array has the most and which has the least normal histogram distribution). If a region of interest is noted on the histogram, it can be selected to determine the identity of the data points contained therein. Lists of data point selected in this manner can be exported, and the program interface also provides limited database links for selected data points.

From there, the user moves on to examine the data in two dimensions with the program’s Scatterplot Ordering function. In this function, dimensions, or arrays, are compared to one another, and the degree of their correlation, based on user selected criteria, is color-coded. The comparison criteria available to the user are the correlation coefficient, least square error for linear regression, quadracity, number of potential outliers, and number of potential items in a region of interest. As with the histogram function, scatterplots at either extreme of a given comparison criteria can be readily identified by the color-coding scheme and regions of interest can be selected for identity and database link information.

These rank-by-feature functions present both challenges and unique opportunities in the analysis of microarray data. Challenges include biological interpretation of a noted pattern or relationship. For example, if one finds that disease versus control comparisons have a very clear quadratic (as opposed to linear) correlation, what does this reveal about physiological processes of the disease? A unique opportunity

is found in the fact that this program has the potential to accept both array and nonarray data simultaneously and to look for correlations based on these criteria as well. This would be of particular advantage for an array project in which considerable clinical data (weight, age, height, etc.) was also available for each of the samples.

6. Practical hints regarding project analyses: false discovery rate correction, stats, fold change

Each microarray mRNA profiling project contains both "desired" experimental variables under study and potential "confounding" variables. The goal of any good experimental project is to control as many of the confounding variables as possible. For example, in experiments using mice, the sex, age, and mouse strain used are all potential confounding variables. This variation can be easily controlled (e.g., only mice of the same sex, age, and inbred mouse strain are utilized in the project). However, in human projects, these same variables can be much more difficult, if not impossible to control, and they become confounding variables.

Potential confounding variables in microarray experiments that require careful consideration include tissue heterogeneity (e.g., variability and sampling errors in use of small tissue samples) and variability in tissue culture conditions that may lead to differences between "identical" cultures. Though tissue culture is generally considered a very "clean" experimental system, variations in cell density, temperature, and media content can all lead to relatively dramatic changes in expression profiles. Another potentially significant source of confounding noise is technical variability in the processing of the RNA and arrays. However, use of quality control and standard operating procedures can reduce the technical variation to a level of relatively small concern (Tumor Analysis Best Practices Working Group, 2004).

In practice, it is often best to run *unsupervised clustering by chip* as the initial step in data analysis (see Seo *et al*., 2004). In this analysis, hierarchical clustering is used to determine which microarrays are most closely related in the pattern of signals over all probe sets (Figure 4). Ideally, one should find that the microarrays in a project are grouped in the dendrogram by the desired biological variable. An example of this is shown in Figure 4, where inbred mice injected with lipopolysaccharide (LPS) were sacrificed and exsanguinated at defined time intervals (0, 24, 48 h). Specific cell types (T_H1, T_H2, and platelets) were then isolated from the peripheral blood using a negative cell isolation technique (StemCell Technologies, Inc). Two rounds of RNA amplification were performed, and the resulting microarrays clustered by sample (Figure 4). It can be seen that the T_H1 samples at 24 and 48 h form the left-most branch, while T_H2 samples (0, 24, 48 h) form a cluster in the center of the dendrogram. The platelet profiles (PLT) seem more variable, with two samples clustering among T_H1 samples and two more distantly related samples forming their own branches to the right of the dendrogram. One could surmise that the microarrays from the T_H1 and T_H2 cell populations are relatively consistent within each group, while the platelets are

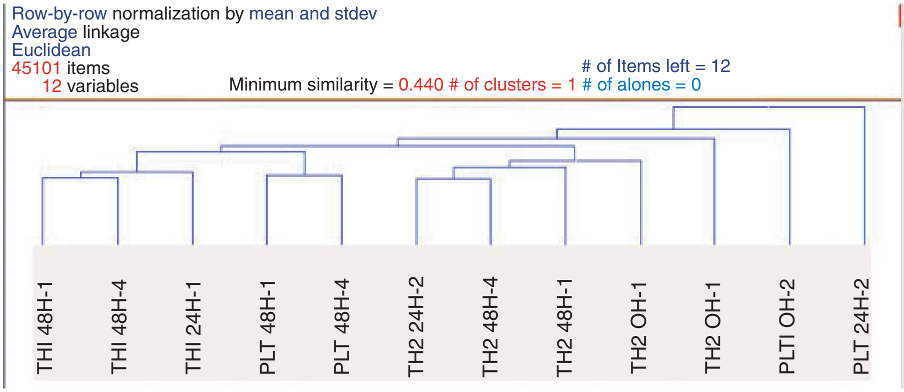


Figure 4 Unsupervised hierarchical clustering to determine extent of confounding variables. In this experiment, mice were exsanguinated at three time points (0, 24, 48 h) following LPS injection. Three different cell types (T_H1, T_H2, and platelets [PLT]) were then purified from the blood. Unsupervised hierarchical clustering of arrays was performed using HCE, resulting in the dendrogram (tree) shown below. This figure is further interpreted in the chapter text above. Note that the T_H1 and T_H2 cell arrays generally cluster together, while the platelet profiles are less consistent in locus within the dendrogram. This analysis suggests that there are more uncontrolled confounding variables with platelets, and that resulting data should be considered "noisy" in nature. An alternative interpretation would be that platelets at 48 h take on many features of T_H1 cells. More replicates would be needed to distinguish between these possibilities

more variable. This suggests that there is less uncontrolled confounding noise with T_H1 and T_H2 cells, compared to platelets and that the findings of the T_H1 and T_H2 data are likely more reliable than the platelet data. As follow-up, one could reassess platelet preparation purity and otherwise attempt to identify variables in the procedure that are contributing to confounding noise in platelet samples. In general, the better the unsupervised clustering appears (e.g., unsupervised segregation into the known biological groups), the cleaner and more robust the resulting data interpretation will be (e.g., upregulated and downregulated transcripts).

Another generalization is that it is best to calculate expression values using two different probe set algorithms (e.g., MAS5.0 and dCHIP difference model), to execute all statistical group comparisons using both sets of expression values in parallel, and to rank all resulting data by significance (p-value) between groups. It is best to rank by p-value rather than by fold change because large fold changes can have no significance (high variability), while small changes can show very strong p-values. Those transcripts that show consistent comparison results (fold change and p-value) with both MAS5.0 and dCHIP can be considered the most robust and prioritized for follow-up. Other transcripts may show a very strong p-value with one algorithm, and not the other; these can be assigned to the next tier of confidence. It is not unusual to find only 10–30% concordance between data obtained from different probe set algorithms. Concordance, as one might expect, decreases as experimental noise (uncontrolled confounding variables) increases.

A very active area of debate within the research community is how to best correct for multiple testing in microarray data analyses. Clearly, when one is testing millions of oligonucleotides, and tens of thousands of probe sets, a p-value of 0.05 is likely to identify many "differences" by chance, as 5% of the 40 000 probe sets

on an array is 2000 probe sets. There is no easy answer to the "multiple testing" problem. However, there are three pragmatic ways around this issue. First, use of a different method/platform (such as quantitative RT-PCR) to validate or confirm the transcripts of interest (ideally using a distinct sample set). A second option is to examine only a small subset of the data. For example, if one is interested in cytokines and only considers the probe sets for cytokines, the 40 000 probe sets at hand could be reduced to a few hundred, lessening the concern about multiple testing. Finally, statistical corrections can be applied to account for multiple testing, but these increase specificity by reducing sensitivity and in reality select only those p-values that are most significant without changing their overall ranking.

7. Practical hints on experimental design

A major challenge in microarray projects is to obtain the maximum sensitivity and specificity, which can be accomplished by removing all or most potential confounding variables, and/or by increasing replication. Data analysis always entails balancing of signal/noise issues, and few clear definitions exist for minimum number of replicates that should be used in microarray investigation. Some generalizations can be made; inbred mouse projects using whole organs require relatively few replicates (often as few as $n = 3$ per group), whereas human tissue experiments, with numerous confounding variables, require much larger groups. This paradigm serves to make a project using mice reasonable in scope and cost, and the equivalent project using human samples much more daunting (and noise-ridden).

One way to avoid the many uncontrolled confounding variables in human studies is longitudinal sampling. If multiple blood samples are taken from the same person, then each person serves as his own control, and the effects of uncontrolled variables such as ethnicity, age, and sex, can be minimized. For example, we have studied muscle biopsies from volunteers in an aerobic training program, with four muscle biopsies taken per subject (Hittel *et al*., 2003; Hittel *et al*., 2005). The longitudinal design allowed robust data to be obtained from just three males and three females.

One last practical note is that rats and knockout transgenic mice are typically "out-bred". Thus, one cannot assume that littermates are genetically identical, and interindividual variation becomes a major confounding variable, as it is in humans.

8. Database resources

A large amount of Affymetrix microarray data is publicly available at three databases; NCBI GEO (**http://www.ncbi.nih.gov/geo**), ArrayExpress (**http://www.ebi.ac.uk/arrayexpress**), and PEPR (Public Expression Profiling Resource; **http://pepr.cnmcresearch.org**). GEO and ArrayExpress contain many types of experimental platforms (spotted microarrays, SAGE, Affymetrix arrays, other), while PEPR is an Affymetrix-only resource. To build cross-platform databases, an international consortium was formed to develop a consensus regarding key information

required for each microarray experiment; namely, the MIAME guidelines (Minimal Information About a Microarray Experiment) (**http://www.mged.org**; Brazma *et al*., 2001). Consistent with the MIAME standards, many scientific journals require that microarray data discussed in publications be made publicly available with sufficient information to allow interpretation.

Acknowledgments

The authors acknowledge the generous support of the National Heart, Lung, and Blood Institute (NHLBI) and National Institute for Neurological Disorders and Stroke (NINDS) of the National Institutes of Health, and the Department of Defense (W81XWH-04-01-0081).

References

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **4**, 365–371.

Hittel DS, Kraus WE and Hoffman EP (2003) Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome. *Journal of Physiology*, **548**, 401–410.

Hittel DS, Kraus WE, Tanner CJ, Houmard JA and Hoffman EP (2005) Exercise training increases electron and substrate shuttling proteins in muscle of overweight men and women with the metabolic syndrome. *Journal of Applied Physiology*, **98**, 168–179.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**(4), e15.

Li C and Wong W (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31–36.

Seo J, Bakay M, Chen YW, Hilmer S, Shneiderman B and Hoffman EP (2004) Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. *Bioinformatics*, **20**, 2534–2544.

Tumor Analysis Best Practices Working Group (2004) Expression profiling–best practices for data generation and interpretation in clinical trials. *Nature Reviews Genetics*, **5**, 229–237.

Specialist Review

Microarray CGH

Denis A. Smirnov
Immunicon Corporation, Huntington Valley, PA, USA

Vivian G. Cheung
University of Pennsylvania, Philadelphia, PA, USA

1. Introduction

Comparative genomic hybridization (CGH) is a molecular cytogenetic method for detection of chromosomal abnormalities. In a CGH experiment, differentially labeled "test" and "reference" DNA samples are simultaneously hybridized to "target" DNA sequences bound to glass slides (Figure 1). The basic assumption of a CGH experiment is that the ratio of the binding of test and control DNA is proportional to the ratio of the concentrations of sequences in the two samples. So the intensity ratio of the two fluorescence signals is indicative of the relative DNA copy number in the "test" relative to the "reference" DNA sample. Traditionally, CGH is performed using metaphase chromosome targets (it is referred to as *chromosome* CGH). This technique has contributed significantly to the current understanding of gross chromosomal imbalances and variations associated with certain diseases, particularly cancer, since regions of amplifications and deletions are likely to contain novel oncogenes and tumor suppressor genes (reviewed in Lichter *et al*., 2000). Although chromosome CGH allows genome-wide analysis of gross DNA copy number imbalances, there are two major limitations that restrict its usefulness as a comprehensive screening tool. First, CGH to metaphase chromosomes provides only limited resolution in the identification of deletions and gains, which is at best is in the order of 2–10 Mb (Lichter *et al*., 2000). In addition, chromosome CGH cannot be easily automated because of the need to identify individual chromosomes. Recent advances in microarray technology allow one to circumvent these limitations by replacing metaphase chromosomes with a collection of mapped cDNA clones, oligonucleotides, or genomic clones placed on glass slides (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998; Beheshti *et al*., 2002). This new platform for genome-wide analysis of chromosomal imbalances is generally referred to as *microarray* CGH. Yano and colleagues have recently determined that although gains and losses of DNA sequence copy numbers detected by array CGH and conventional CGH coincided highly, with correspondence rates of 94 and 95%, respectively, microarray CGH had higher sensitivity for the detection of copy number losses compared with that for conventional CGH (Yano *et al*.,

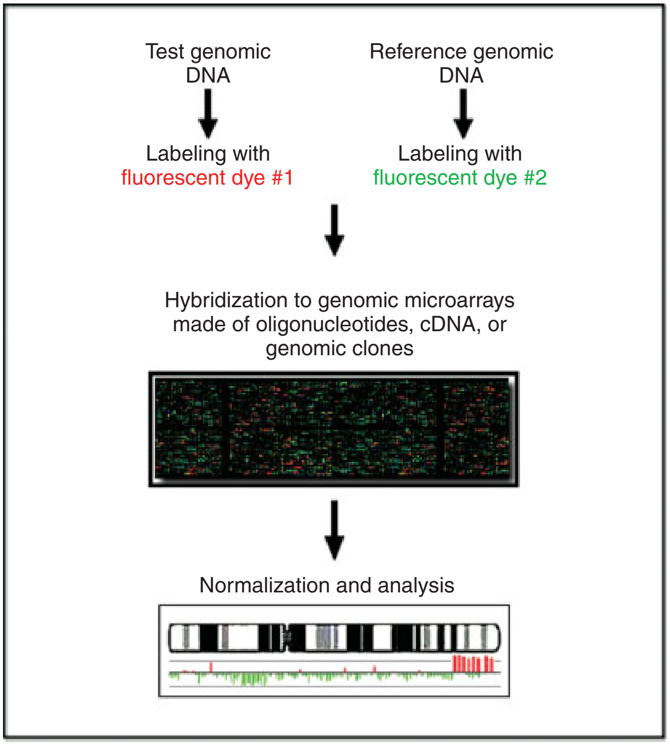


Figure 1 Schematic diagram of a typical microarray CGH experiment. Differentially labeled "test" and "reference" DNA samples are simultaneously hybridized to DNA sequences bound to glass slides. The intensity ratio of the two fluorescence signals is indicative of the relative DNA copy number in the "test" relative to the "reference" DNA sample

2004). In an earlier study, it was also observed that despite overall concordance between results obtained with microarray CGH and conventional CGH, not all amplified genes detected by microarrays CGH were detected by chromosome CGH (Arai *et al*., 2003). These results can be potentially explained by higher resolution achieved by microarray CGH platforms (currently down to 30∼70 kb), which is significantly higher than that of conventional CGH (Lucito *et al*., 2003; Ishkanian *et al*., 2004). This article will focus on current techniques utilized to manufacture microarrays for CGH analysis. We will also discuss approaches that are used to prepare and hybridize labeled genomic DNA probes and methods employed in analysis of microarray CGH hybridization experiments.

2. Microarrays for comparative genomic hybridization

Currently, there are three major approaches for manufacturing microarrays for genomic hybridizations: cDNA array CGH, oligonucleotide array CGH, and BAC array CGH.

2.1. cDNA array CGH

In cDNA array CGH, one utilizes conventional cDNA arrays, such as those used in gene expression studies. The utility of this approach was first demonstrated by Pollack *et al*. (1999). They used a 3360-feature cDNA microarray platform to perform parallel CGH and gene-expression studies on a collection of breast cancer cell lines and tissues. This approach allowed them to study relationships between gene amplification and overexpression. Later on, Pollack and colleagues expanded their initial gene expression/copy number alteration profiling study to 44 advanced, primary breast tumors and 10 breast cancer cell lines by using cDNA microarrays containing 6691 different mapped human cDNAs (Pollack *et al*., 2002). Similarly, Guo *et al*. (2002) utilized a 21 632-feature cDNA microarray to profile gene expression and copy number alterations in nasopharyngeal carcinomas. Beheshti *et al*. (2003) performed CGH analysis of 10 neuroblastomas using a 19 200-feature cDNA microarray. cDNA CGH approach offers several advantages. The most important one is the availability of a significant number of commercial cDNA array platforms and cDNA clone sets. In addition, with ever-increasing characterization of a complete human gene set, cDNA array CGH can serve as a useful tool for direct identification of genes in the amplified and/or deleted regions. Finally, as mentioned above, cDNA arrays have an advantage in allowing analysis of DNA copy number alterations and mRNA expression of the same genes in the same sample through easy parallel hybridizations experiments. However, there are several major limitations of this approach. First, cDNA sequences do not represent any intergenic sequences such as introns and promoter sequences, where potential alterations can occur. Even a complete set of full-length cDNA clones will represent at best only 5% of human genomic sequences. Also, because of the small size and reduced complexity of the cDNA target sequences, cDNA array CGH approach may not have sufficient sensitivity to reproducibly detect low copy number imbalances (Beheshti *et al*., 2002; Beheshti *et al*., 2003). Beheshti *et al*. (2003) estimated that deletions or copy number gains below the level of 10 to 20 copies cannot be determined with certainty. Finally, there is an extensive sequence similarity between paralogous genes (gene families), which can interfere with the interpretation of cDNA array CGH data.

2.2. Oligonucleotide array CGH

A second, and currently the least utilized, approach for manufacturing genomic microarrays involves oligonucleotides. Recently, three reports were published that utilized microarrays made of oligonucleotides for high-resolution detection of copy number changes (Lucito *et al*., 2003; Bignell *et al*., 2004; Zhao *et al*., 2004). Lucito *et al*. (2003) arrayed 70-mer oligonucleotide probes designed from the human genome sequence with an average resolution of a probe every 30 kb and determined copy number alterations in genomic DNA samples from cancer cells. In their experiments, differentially labeled "test" and "reference" DNA samples were simultaneously hybridized to the 85 000 feature oligonucleotide arrays. In a different approach, two groups utilized commercially available SNP detection microarrays (like Affymetrix GeneChip 10 000 SNP assay; **http://www.affymetrix.com**)

for copy number changes detection (Bignell *et al*., 2004; Zhao *et al*., 2004). Current resolution of this microarray is approximately 210 kb on the genome. Affymetrix has recently released a 100 000 SNP GeneChip that increases resolution to ~20 kb. This approach also produces genotyping data in conjunction with the copy number alterations analysis.

It should be noted that all groups utilized a “genomic complexity reduction” approach when preparing genomic DNA samples for hybridization (Lucito *et al*., 1998). Genomic DNA samples were digested to completion using BglII or Xba1 prior to ligation of adaptors and amplification of the ligation products using an adaptor-specific primer (Lucito *et al*., 2003; Bignell *et al*., 2004; Zhao *et al*., 2004). Oligonucleotide probes on microarrays utilized by all groups were specifically selected to represent genomic sequences in the vicinity of BglII or Xba1 restriction endonuclease sites that will be amplified using a “genomic complexity reduction” method. This approach was shown to reduce the complexity of the genome by 98% and thereby improve hybridization kinetics since it produces more copies of particular sequences within a given mass of DNA that are complimentary to the arrayed oligonucleotides (Lucito *et al*., 1998).

Oligonucleotide genomic arrays offer several advantages. First, they provide very high resolution coverage of human genome. Because of the small size of oligonucleotide probes, they can observe very small deletions and amplifications that would be missed with high-density BAC since size of the BAC (150–200 kb) obscures high resolution. Since sequence composition of oligonucleotide arrays is precisely specified, these arrays can be easily converted into highly reproducible fabricated industrial standard product that can be made available for wide usage. One of the disadvantages of oligonucleotide arrays is their high production cost and limited availability. Another disadvantage of oligonucleotide array is their current reliance on “genomic complexity reduction” methodology. Some genomic sequences will be poorly represented by this approach because of lack of particular restriction endonuclease sites in certain regions. Polymorphism of restriction endonuclease sites also presents a challenge in interpretation of CGH data (Lucito *et al*., 2003).

2.3. CGH with arrays made of genomic clones

The third and most frequently used microarray CGH platform uses genomic DNA derived from genomic clones such as yeast artificial chromosomes (YAC), bacterial artificial chromosomes (BAC), P1, PAC, and cosmids as targets on the microarray. Very frequently, when people mention microarray CGH they refer to CGH analysis that utilizes microarrays made of genomic clones. This array CGH platform was first described by Solinas-Toldo *et al*. (1997). Early experiments with genomic microarrays were limited by the absence of good collections of mapped clones and therefore utilized small collections of mapped cosmids, P1, YAC and BAC (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998). Since then, our laboratory and several other groups had assembled several collections of mapped BAC clones spanning the human genome at an average interval of ~1 Mb (Morley *et al*., 2001; Cheung *et al*., 2001; Snijders *et al*., 2001; Fiegler *et al*., 2003). For example, we

have assembled in our laboratory a collection of about 5000 mapped RPCI-11 BAC clones that covers the human genome at approximately 1-Mb resolution (Morley *et al*., 2001; Cheung *et al*., 2001; Smirnov *et al*., 2004a). The clones were anchored to STS markers from the GeneBridge 4 radiation hybrid map. Each clone was mapped by filter hybridization and verified by PCR to contain the STS marker. Chromosomal locations of some of the clones were verified by fluorescent in situ hybridization (Kirsch and Reid, 2000). The mapped clones were further characterized by HindIII fingerprinting and end-sequencing. The clones were anchored to the current human genome sequence assemblies by aligning the end sequences of the BAC clones to the DNA sequences at the UCSC Genome Browser. Information about the mapped BAC clones is available through our Web-based database GenMapDB, **http://genomics.med.upenn.edu/genmapdb** and through the UCSC Genome Browser (**http://genome.ucsc.edu** on the GenMapDB clone track). Glycerol stocks of all the clones are available through various repositories (**http://genomics.med.upenn.edu/genmapdb**). Recently, this collection was expanded by groups at Genome Sciences Centre, Vancouver, BC, Canada, and BACPAC Resource Center at Children's Hospital Oakland Research Institute in Oakland, California. They have assembled a collection of 32 433 BAC clones that covers the entire human genome at a final resolution of ∼75 kb (Ishkanian *et al*., 2004). The identity of all the clones has been validated by fingerprinting. Information about this collection is available at **http://bacpac.chori.org/pHumanMinSet.htm**. This collection of BAC clones was recently arrayed on glass slides, and successful array CGH experiments were performed (Ishkanian *et al*., 2004).

2.3.1. Strategies for construction of microarrays made of genomic clones

Despite the availability of mapped clones, the manufacturing of genomic microarrays is still challenging because of the difficulty of preparing adequate amounts of DNA from genomic clones (Beheshti *et al*., 2002). Initially, for microarray CGH studies, DNA from genomic clones was extracted using a large-scale alkaline lysis method followed by the anion-exchange chromatography (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998). When expanded to thousands of clones, this approach becomes too labor intensive and expensive. Unfortunately, small-scale DNA preparations do not produce sufficient amounts of genomic DNA to manufacture genomic microarrays. In addition, *E. coli* DNA is a very common contaminant of DNA preparations from large insert clones using small-scale silica- or filtration-based DNA preparation miniprep kits (10–25% of total purified DNA) (Foreman and Davis, 2000). There are currently three strategies that are used to manufacture large genomic microarrays. All these procedures involve amplification step that aims at producing large quantities of DNA from a library of clones, and generating spotting solutions with high DNA concentration sufficient for microarray manufacturing. The first strategy utilizes degenerate oligonucleotide primer (DOP) PCR that is designed to amplify representative fragments of the BAC DNA with degenerate primers in a single step (Telenius *et al*., 1992). Recently, Fiegler *et al*. (2003) used a bioinformatic approach to select three novel degenerate DOP-PCR primers primers that were efficient in the amplification of human DNA but were poor at amplifying

E. coli DNA, a common contaminant of DNA preparations from large insert clones. Using this approach, a BAC array was constructed that covers entire human genome at about 1-Mb intervals. Similar mouse-specific DOP primers were also designed and a 1-Mb mouse genome microarray was constructed (Chung *et al*., 2004).

Another strategy that is utilized for manufacturing large genomic arrays involves ligation-mediated (LM) PCR. LM-PCR starts with restriction enzyme digestion and linker ligation prior to two rounds of PCR amplification. This approach requires initial affinity chromatography-based purification of BAC DNA and does not prevent potential amplification of contaminating *E. coli* DNA. DNA preparations are monitored for contamination with *E. coli* DNA by gel electrophoresis prior to linker ligation (Snijders *et al*., 2001). LM-PCR approach was successfully used to manufacture very large genomic arrays (Snijders *et al*., 2001; Ishkanian *et al*., 2004).

A third strategy that is used for manufacturing genomic arrays employs strand-displacement rolling circle amplification (RCA) methodology. This technique utilizes random primers and Phi29 DNA polymerase to amplify circular DNA templates selectively and exponentially even in the presence of contaminating bacterial genomic DNA (Dean *et al*., 2001). This approach does not require thermocyclers and initial purification of BAC DNA. Recently, large genomic microarrays made of 4500 BAC clones were prepared using this technique (Smirnov *et al*., 2004a).

There are several other ways that can be potentially used for manufacturing large genomic arrays. GenomePlex Whole Genome Amplification technology developed by Rubicon Genomics can be potentially used to amplify BAC DNA to make large genomic arrays (Gribble *et al*., 2004). In about 3 h, microgram quantities of DNA can be generated from nanograms of starting material. Potentially, T7-based linear amplification system recently adapted for genomic DNA can also be used for preparation of BAC DNA for genomic array manufacturing (Liu *et al*., 2003). Also, inter-Alu PCR was utilized to manufacture large genomic arrays (Cheung *et al*., 1998; Smirnov *et al*., 2004b). This technique involves the amplification of a specific DNA region present between two nearby Alu sequences on the genome. Its principle is based on the widespread distribution of so-called Alu sequences (about 300 bp in size) among the human genome. There are about one million copies of it spaced along the genome about every 4 kb. Human DNA sequences can be selectively amplified using suitable Alu-primer pairs even in the presence of contaminating genomic DNA such as *E. coli* genomic DNA.

Genomic microarrays made of genomic DNA derived from genomic clones are currently most widely used platform for array CGH experiments. It offers excellent, potentially even complete representation of all human genomic sequences, making it very powerful for high-resolution CGH as well as other applications, such as chromatin precipitations (ChIP) and methylation status studies. However, this platform also has several disadvantages. As mentioned above, genomic microarrays made of large number of genomic clones are still difficult to manufacture. Some genomic clones contain repetitive sequences, which can lead to erroneous hybridization patterns. This problem can be particularly severe for certain repeat-rich regions of the genome. Large quantities of Cot-1 DNA have to be added to hybridization reactions to suppress hybridization of repetitive sequences. Certain

genomic clones may also represent duplicated regions in the genome that may complicate interpretation of hybridization results. Finally, large size of genomic clone also can potentially mask detection of small copy number abnormalities.

3. Probe preparation and hybridization to microarrays

Prior to microarray CGH experiment, genomic DNA isolated from "test" and "reference" samples are differentially labeled by incorporation of fluorescently labeled nucleotides, for example, Cyanine-3 and Cyanine-5 dCTPs. Sufficient amounts of starting genomic DNA (typically in a range of 50–500 ng per sample per hybridization) is required to perform hybridization to microarray slides. There are numerous instances where only a very small number of cells are available for cytogenetic evaluation. Considerable efforts are being directed toward utilization of whole genome amplification (WGA) methodologies to allow CGH analysis when only very few or even single cell is available for analysis. Many of the methods that are being utilized for manufacturing genomic arrays are also employed for unbiased amplification of very small amounts of "test" genomic DNA for CGH experiments. These methods include degenerate oligonucleotide primed PCR (DOP-PCR), primer extension PCR (PEP), phi29 strand-displacement rolling circle amplification (RCA), and OmniPlex technology (Zhang *et al*., 1992; Telenius *et al*., 1992; Dean *et al*., 2002; Barker *et al*., 2004). Recent amplification fidelity studies indicate that two methods, RCA and Omniflex, amplify genomic DNA in a highly representative manner (Dean *et al*., 2002; Barker *et al*., 2004; Paez *et al*., 2004). Unfortunately, these studies were not extended to samples with very small amounts genomic DNA (in pg range) to demonstrate the fidelity of amplification of genomic DNA sampled extracted from single cell. Following the preparation of differentially labeled "test" and "reference" samples, hybridization is performed to microarray slides in the presence of high concentrations (10 μg or more) of human Cot1 DNA to suppress nonspecific hybridization to repetitive sequences.

4. Analysis of microarray CGH data

In a typical microarray CGH experiment, the test genomic samples are labeled with one fluorescent tag (Cy3 dye, for example) and co-hybridized with a control sample labeled with a second fluorescent tag (Cy5 dye, for example). The fluorescent intensity ratios of each clone produced by hybridization of differentially labeled experimental and control genomic DNAs to the array are used as an indication of copy number changes. The ratio of experimental to control fluorescence is proportional to the relative dosage between the two samples. Since the samples hybridized to the microarrays were labeled with two different dyes, Cy3 and Cy5, that have different incorporation efficiency, it is often necessary to normalize the signal intensities. A method that is used frequently in microarray data analysis is to normalize all the Cy3 and Cy5 intensities to the value of one, assuming that the overall differences between the two channels are negligible. However, this global normalization method is not appropriate for many CGH experiments, especially

those with tumor samples that have high percentages of chromosomal aberrations (Wessendorf *et al*., 2002). Normalization based on regions that are known to have normal chromosome profiles is more precise but requires prior knowledge of the copy number changes in that sample. Alternatively, the rank invariant normalization method can be used, which allows one to perform normalization without prior knowledge of the degree of chromosomal aberrations in the samples (Tseng *et al*., 2001). In this method, normalization is based on clones whose rankings by intensity signal in the two channels are the most similar. These clones most likely represent regions with normal chromosomal profiles.

After normalization, the fluorescence ratio between experimental (test) and control genomic DNA (reference) samples should be 1.0 in regions where the relative DNA sequence copy number is the same in those samples, greater than 1.0 if the relative DNA sequence copy number in the test sample is greater, and less than 1.0 in regions where the relative DNA copy number is lower in the test sample. CGH profiles are interpreted by comparing the fluorescence intensity ratios of each clone to the "normal" ratio of 1.0 (reviewed in Molinaro *et al*., 2002). One method of interpretation of array CGH data is by deviation of the ratio from the thresholds of 1.25:0.75 (above or below threshold) (Wessendorf *et al*., 2002). A different method uses a standard deviation of replicate measurements obtained from reference to reference hybridizations to establish threshold levels. Deviations of the fluorescence ratio above or below the mean +/- 2SD are considered as significant changes in copy number (Schwaenen *et al*., 2004). Alternatively, a t-statistic (or nonparametric alternatives such as Wilcoxon Rank Sum Test) can be used to compare replicate fluorescence ratio measurements obtained from a test:reference CGH analysis with the results of reference:reference CGH hybridizations (Moore *et al*., 1997; Smirnov *et al*., 2004a). Several software packages (e.g. See GH; **http://www.bccrc.ca/cg/ArrayCGH_Group.html**) have been developed to help researchers to normalize, analyze, and visualize fluorescence data from array CGH experiments.

Analysis of CGH results obtained with oligonucleotide arrays offers a unique set of challenges. For example, CGH method based on Affymetrix SNP GeneChip differs from spotted-array-based CGH in that the normal and tumor DNAs are hybridized to different arrays in a similar fashion to Affymetrix expression array experiments. However, this approach still requires a set of separate experiments with genomic DNA from "normal" donors for the subsequent analysis of tumor samples (Bignell *et al*., 2004). In addition, when estimating copy number, the output from the each feature is often considered in the context of signal intensities from oligonucleotides in neighboring or flanking positions.

5. Applications of array CGH and genomic array technology

Microarray CGH is a very powerful method that allows detection and localization of DNA copy number aberrations, such as deletions and amplifications. From numerous traditional chromosome CGH experiments, it is apparent how widespread copy

number alterations are in cancer and other diseases (reviewed in Beheshti *et al*., 2002; see SKY/M-FISH and CGH database **http://www.ncbi.nlm.nih.gov/sky/**; Tumor CGH database **http://amba.charite.de/~ksch/cghdatabase/start.htm**; Database of Recurrent Chromosomal Aberrations in Cancer **http://cgap.nci.nih.gov/Chromosomes/RecurrentAberrations**). Microarray CGH allows improved high-resolution quantitative assessment of copy number changes. Array CGH genome analysis techniques have already revealed numerous regions that are frequently abnormal in multiple types of tumors. Several well-established oncogenes, tumor suppressor genes, and genes associated with cancer susceptibility have been mapped to these regions of recurrent abnormality. These include MYC (amplification at 8q24), AKT2 (amplification at 19q13), ERBB2 (amplification at 17q21.2), CCND1 (amplification at 11q13), p53 (deletion at 17p13), BRCA1 (deletion at 17q21), and BRCA2 (deletion at 13q12); BCL2 (amplification at 18q21) (Forozan *et al*., 2000; Pollack *et al*., 2002; Ishkanian *et al*., 2004; Smirnov *et al*., 2004a). These data strongly support the notion that regions of recurrent abnormality encode genes that contribute to cancer progression when differentially expressed because of mutation, loss, or amplification. Increased resolution of array CGH will facilitate identification of new genes in specific chromosomal regions and functional characterization of already-known genes. Similarly to cDNA array gene expression profiling, CGH profiles of cancer samples also have a potential to be used to predict survival and choose treatment (van't Veer *et al*., 2002). Diseases and conditions other than cancer will also benefit from recent advances in microarray technology. For instance, Shaw-Smith and colleagues employed DNA microarray to identify copy number changes in patients with learning disability and dysmorphism (Shaw-Smith *et al*., 2004). Hu *et al*. (2004) used arrays to screen for evidence of aneuploidy in single cells. Thus, microarray CGH technology could also potentially be used for analysis of preimplantation embryos, uncultured amniocytes, and single fetal cells isolated noninvasively from peripheral blood of pregnant women.

In addition to array CGH, genomic array technology can have a utility in other areas of genomics. Genomic microarrays can be used to identify promoter sequences isolated by chromatin immunoprecipitations (reviewed in Buck and Lieb, 2004). Genomic microarrays can also be used to identify genomic sequences and genes regulated through epigenetic modification (Ballestar *et al*., 2003). Genomic arrays were successfully used for mapping regions shared identical-by-descent (IBD) between two individuals without locus-by-locus genotyping or sequencing (Cheung *et al*., 1998; Smirnov *et al*., 2004b). Recently, a technique named *genomic microarray painting* was described that allows high-resolution analysis of composition and the breakpoints of aberrant chromosomes (Gribble *et al*., 2004). Several new technologies, such as multiplex amplifiable probe hybridization (MAPH) and multiplex ligation dependent probe amplification (MLPA), when combined with genomic microarrays have the potential to provide sensitive and precise targeting of the unique regions of the genome where rearrangement occurs (Hollox *et al*., 2002; Schouten *et al*., 2002). In the next several years, these and other applications of genomic microarrays will be developed for basic research and clinical applications.

6. Remaining challenges

Despite enormous progress and promise, many obstacles remain in the way of genomic microarrays to become a clinical diagnostics tool. When compared to expression arrays, it is clear that in the field of genomic microarrays there are still no widely accepted commercial genomic array platforms. In fact, only few companies offer genomic microarray products. For clinical applications, it is essential that all of the targets placed onto genomic arrays were extensively characterized. For instance, all BAC clones will have to be FISH-mapped to reveal mismapped clones and those that hybridize to multiple sites within human genome. Moreover, precise standards for signal normalization and genomic microarray data analysis are yet to be established. When working with very small amounts of DNA from clinical samples (picogram range), methods utilized for whole genome amplification are yet to be extensively characterized. Finally, extensive clinical evaluations will have to be performed in order for the microarray CGH technology to be widely accepted as a routine diagnostic tool.

References

Arai H, Ueno T, Tangoku A, Yoshino S, Abe T, Kawauchi S, Oga A, Furuya T, Oka M and Sasaki K (2003) Detection of amplified oncogenes by genome DNA microarrays in human primary esophageal squamous cell carcinoma: comparison with conventional comparative genomic hybridization analysis. *Cancer Genetics and Cytogenetics*, **146**, 16–21.

Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH and Esteller M (2003) Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *The EMBO Journal*, **22**, 6335–6345.

Barker DL, Hansen MS, Faruqi AF, Giannola D, Irsula OR, Lasken RS, Latterich M, Makarov V, Oliphant A, Pinter JH, *et al*. (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Research*, **14**, 901–907.

Beheshti B, Park PC, Braude I and Squire JA (2002) Microarray CGH. *Methods in Molecular Biology*, **204**, 191–207.

Beheshti B, Braude I, Marrano P, Thorner P, Zielenska M and Squire JA (2003) Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia*, **5**, 53–62.

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, **14**, 287–295.

Buck MJ and Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Cheung VG, Gregg JP, Gogolin-Ewens KJ, Bandong J, Stanley CA, Baker L, Higgins MJ, Nowak NJ, Shows TB, Ewens WJ, *et al*. (1998) Linkage-disequilibrium mapping without genotyping. *Nature Genetics*, **18**, 225–230.

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, *et al*. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, **409**, 953–958.

Chung YJ, Jonkers J, Kitson H, Fiegler H, Humphray S, Scott C, Hunt S, Yu Y, Nishijima I, Velds A, *et al*. (2004) A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. *Genome Research*, **14**, 188–196.

Dean FB, Nelson JR, Giesler TL and Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*, **11**, 1095–1099.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, *et al*. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5261–5266.

Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IP, *et al*. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes, Chromosomes & Cancer*, **36**, 361–374.

Foreman PK and Davis RW (2000) Real-time PCR-based method for assaying the purity of bacterial artificial chromosome preparations. *BioTechniques*, **29**, 410–412.

Forozan F, Mahlamaki EH, Monni O, Chen Y, Veldman R, Jiang Y, Gooden GC, Ethier SP, Kallioniemi A and Kallioniemi OP (2000) Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. *Cancer Research*, **60**, 4519–4525.

Gribble SM, Fiegler H, Burford DC, Prigmore E, Yang F, Carr P, Ng BL, Sun T, Kamberov ES, Makarov VL, *et al*. (2004) Applications of combined DNA microarray and chromosome sorting technologies. *Chromosome Research*, **12**, 35–43.

Guo X, Lui WO, Qian CN, Chen JD, Gray SG, Rhodes D, Haab B, Stanbridge E, Wang H, Hong MH, *et al*. (2002) Identifying cancer-related genes in nasopharyngeal carcinoma cell lines using DNA and mRNA expression profiling analyses. *International Journal of Oncology*, **21**, 1197–1204.

Hollox EJ, Akrami SM and Armour JA (2002) DNA copy number analysis by MAPH: molecular diagnostic applications. *Expert Review of Molecular Diagnostics*, **2**, 370–378.

Hu DG, Webb G and Hussey N (2004) Aneuploidy detection in single cells using DNA array-based comparative genomic hybridization. *Molecular Human Reproduction*, **10**, 283–289.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

Kirsch IR and Ried T (2000) Integration of cytogenetic data with genome maps and available probes: present status and future promise. *Seminars in Hematology*, **37**, 420–428.

Lichter P, Joos S, Bentz M and Lampel S (2000) Comparative genomic hybridization: uses and limitations. *Seminars in Hematology*, **37**, 348–357.

Liu CL, Schreiber SL and Bernstein BE (2003) Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics*, **4**, 19.

Lucito R, Nakimura M, West JA, Han Y, Chin K, Jensen K, McCombie R, Gray JW and Wigler M (1998) Genetic analysis using genomic representations. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 4487–4492.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research*, **13**, 2291–2305.

Molinaro AM, van der Laan MJ and Moore DH (2002) *Comparative Genomic Hybridization Array Analysis*. U.C. Berkeley Division of Biostatistics Working Paper Series. Working paper 106. **http://www.bepress.com/ucbbiostat/paper106**

Moore DH II, Pallavicini M, Cher ML and Gray JW (1997) A *t*-statistic for objective interpretation of comparative genomic hybridization (CGH) profiles. *Cytometry*, **28**, 183–190.

Morley M, Arcaro M, Burdick J, Yonescu R, Reid T, Kirsch IR and Cheung VG (2001) GenMapDB: a database of mapped human BAC clones. *Nucleic Acids Research*, **29**, 144–147.

Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, *et al*. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Research*, **32**, e71.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, **23**, 41–46.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL and Brown PO (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12963–12968.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F and Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, **30**, e57.

Schwaenen C, Nessling M, Wessendorf S, Salvi T, Wrobel G, Radlwimmer B, Kestler HA, Haslinger C, Stilgenbauer S, Dohner H, *et al*. (2004) Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 1039–1044.

Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, *et al*. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *Journal of Medical Genetics*, **41**, 241–248.

Smirnov DA, Burdick JT, Morley M and Cheung VG (2004a) Method for manufacturing whole-genome microarrays by rolling circle amplification. *Genes, Chromosomes & Cancer*, **40**, 72–77.

Smirnov D, Bruzel A, Morley M and Cheung VG (2004b) Direct IBD mapping: identical-by-descent mapping without genotyping. *Genomics*, **83**, 335–345.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263–364.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T and Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, Chromosomes & Cancer*, **20**, 399–407.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA and Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, **13**, 718–725.

Tseng GC, Oh MK, Rohlin L, Liao JC and Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549–2557.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wessendorf S, Fritz B, Wrobel G, Nessling M, Lampel S, Goettel D, Kuepper M, Joos S, Hopman T, Kokocinski F, *et al*. (2002) Automated screening for genomic imbalances using matrix-based comparative genomic hybridization. *Laboratory Investigation*, **82**, 47–60.

Yano S, Matsuyama H, Matsuda K, Matsumoto H, Yoshihiro S and Naito K (2004) Accuracy of an array comparative genomic hybridization (CGH) technique in detecting DNA copy number aberrations: comparison with conventional CGH and loss of heterozygosity analysis in prostate cancer. *Cancer Genetics and Cytogenetics*, **150**, 122–127.

Zhang L, Cui X, Schmitt K, Hubert R, Navidi W and Arnheim N (1992) Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 5847–5851.

Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*, **64**, 3060–3071.

Specialist Review

Expression and localization of proteins in mammalian cells

Jennifer L. Stow and Rohan D. Teasdale
The University of Queensland, Brisbane, Australia

1. Introduction

The ability to transfer foreign DNA into eukaryotic cells for the expression of recombinant genes and proteins has revolutionized experimentation in biology and has created new approaches for the treatment of genetic disease in humans. Basic molecular biology techniques allow engineering of DNA to dictate the conditions and locations of gene expression, to alter the form and function of gene products, and to tag gene products in any number of ways. The techniques of cell biology then come to play in analyzing the subcellular location of a specific gene product, its function within cells and whole organisms, and its role in disease. Genome enterprises present us with large cohorts of known and novel genes. At the cutting edge of dissecting the biology of whole genomes is our ability to harness the forces of molecular and cellular biology for the large-scale, high-throughput expression of genes, and the analysis of protein location and function in fixed cells or in living cells and organisms.

This review will focus on the methodologies for gene transfer and protein expression in mammalian cells, and on the emerging technologies for high-throughput analysis of in vitro protein expression. One important end point for the development of gene transfer and protein expression techniques is their application to gene therapy with the goal of curing, treating, or circumventing defects in genetic disease in humans. This is a field of intense investigation and one that is constantly and thoroughly reviewed in the literature (Chernajovsky *et al*., 2004; Scollay, 2001). The production of biopharmaceuticals represents another emerging application for recombinant gene technology (van Berkel *et al*., 2002). However, this current review will focus on experimental applications of gene and protein expression and the move toward high-throughput application of these techniques.

2. Methods for gene expression – many ways in, on, and off

Molecular and cellular biology have provided us with the means to introduce foreign genes into eukaryotic cells in order to express recombinant proteins for many applications in research, agriculture and medicine. While typically applied to

the analysis or production of single recombinant gene products, there is increasing need for large-scale gene or protein expression to analyze whole genomes and proteomes. The vectors and methods for introducing foreign DNA into mammalian cells have been continuously refined over the past two decades to improve the efficiency of transfection, the levels of expressed protein, and to manipulate the timing, location, and stimuli for gene expression. Much of this development has been driven by the need to obtain ultimately safe and effective expression systems for gene therapy, with research applications reaping benefit along the way.

A variety of physical, chemical, and biological approaches have been exploited for introducing recombinant genetic material and proteins into cells (see Table 1). There are pros and cons to each of these methods, and the method of choice in any situation is based largely on the type of cell being transfected and the goal of the expression (Colosimo *et al*., 2000). For experimental purposes, perhaps the most commonly used techniques utilize a wide range of commercially available cationic

Table 1 Summary of methods used for transfection and gene expression in mammalian cells

Method	References
Calcium phosphate	
One of the earliest transfection methods. Coprecipitates of calcium phosphate and DNA are added to and taken up by cells. This plasmid method allows simultaneous transfer of multiple DNA constructs for coexpression.	Kato *et al*. (1986) Batard *et al*. (2001) Jordan and Wurm (2004)
Physical methods	
Purified plasmid DNA can be directly micro-injected into the nuclei or cytoplasm of individual mammalian cells; up to hundreds of cells can be injected. GFP-tagged gene products can be visualized immediately upon synthesis and imaged in live individual cells.	Boyd (2002) Kreitzer *et al*. (2003) Trezise (2002) Li (2004)
Electroporation techniques use electric current to transiently perforate cells or in vivo tissues for entry of plasmid DNA. Electroporation is used for cell types that are not amenable to transfection by other methods and for multiple transgene transfer to large numbers of cells.	
Lipofection	
Synthetic, cationic lipids mixed with plasmid DNA to form liposomes that fuse with or are taken up by cells. Many liposome formulations are now commercially available and offer wide-ranging functionality and efficient transfer in experimental systems and in some clinical situations.	Felgner *et al*. (1987) Safinya (2001) Tranchant *et al*. (2004)
Viral vectors	
Designer, recombinant viral vectors with packaged transgenes but excluding virally encoded genes give high efficiency transfer and high levels of protein expression. Vectors based on adenoviral and adeno-associated DNA viruses are the most commonly used to date for gene therapy, infecting a wide range of cell types. Vectors based on RNA retroviruses integrate stably into the host genome for sustained transgene expression. Lentiviral vectors such as HIV-1 infect both dividing and nondividing cells.	Cepko *et al*. (1984) St George (2003) Delenda (2004) Blesch (2004) Lu (2004)

liposome formulations as vehicles for inserting DNA plasmids into mammalian cells (Safinya, 2001; Tranchant *et al*., 2004). Viral vectors, which make use of the inherent propensity of viruses to invade mammalian cells, are undergoing the most intense refinement for potential therapeutic use in humans. Of note are the lentiviral vectors, particularly those based on HIV-1, which hold promise for highly efficient, stable, and safe expression of recombinant proteins in many cell types and tissues for research purposes, and in the clinic (Blesch, 2004; Delenda, 2004).

The plasmids for introducing recombinant DNA into mammalian cells are also under constant development. Viral promoters such as the cytomegalovirus early promoter (CMV) were first introduced to drive the constitutive, medium-level expression of genes in mammalian cells, and other viral promoters offer either higher or lower levels of constitutive expression (Fitzsimons *et al*., 2004). An important advance was then in the development of plasmids for regulated gene expression. Regulatory elements from the bacterial *lac* operator and repressor provide tight, reversible expression of genes normally involved in lactose uptake and metabolism, but can be utilized for inducible expression of transgenes in mammalian cells. As a readily visible demonstration of how the *lac* operator-repressor system works, Cronin *et al*. (2001) used it to drive IPTG-inducible expression of an exogenous tyrosinase gene in mice, which in turn resulted in inducible coat-color variation (Cronin *et al*., 2001). The tetracyclin resistance operon from *Escherichia coli* is the basis for *tet*-inducible expression in mammalian cells (Knott *et al*., 2002) by which gene expression can be turned on or off. Hormone-based inducible expression systems, namely, progesterone and estrogen make use of ligand-binding nuclear hormone receptors to switch transgene expression on and off (Evans, 1988). Viral plasmids can be engineered in a variety of sophisticated ways to give targeted insertion into the genome and to regulate gene expression (Galla *et al*., 2004; Grez and von Melchner, 1998; Tan *et al*., 2004). Expression of specific proteins in whole animals, especially for gene therapy purposes, relies on the use of customized plasmids, often with promoters to drive cell- or tissue-specific expression and tags, such as green fluorescent protein (GFP), to signify expression (Sasmono *et al*., 2003).

Regulated gene expression can also be used to knock out or knock down the expression of endogenous genes in cells and animals. Earlier approaches included antisense or missense DNA, and the most popular method now relies on the introduction of short interfering RNAs (siRNA) and short hairpin RNAs (shRNA) to block transcription of specific genes (Crooke, 2004; Hannon and Rossi, 2004; Wadhwa *et al*., 2004). Together, different transfection modes, coupled with regulated expression plasmids and interfering RNAs, generate a powerful "toolbox" for the investigation and manipulation of gene expression and protein function in mammalian cells. Elements of the toolbox are increasingly adapted for high-throughput analysis of recombinant protein localization and function in whole proteomes.

3. Bypassing genes with protein expression

In addition to recombinant genetic approaches, the introduction of biologically active macromolecules (polypetides, genes, drugs, or other materials) directly into

mammalian cells offers another approach for experimentation and therapeutics. Small cationic, often arginine-rich peptides known as protein transduction domains (PTDs; usually of 9–16 amino acids) can be used to gain entry into cells with macromolecules fused or cross-linked to them. Some of the PTDs in use include those derived from nucleic acid binding proteins such as the human immunodeficiency virus Type 1 TAT protein, the *Drosophila* homeotic transcription protein antennapedia (Antp), and synthetic peptides such as nonaarginine (R(9)) (Becker-Hapak *et al*., 2001; Beerens *et al*., 2003). The exact mechanism by which PTDs actually enter cells is still not completely clear but appears to involve binding of the charged peptides to heparan sulfate residues on cell surfaces, and then either direct membrane penetration, endocytosis, or macropinocytosis in lipid rafts (Beerens *et al*., 2003; Wadia *et al*., 2004). Protein transduction technology is potentially amenable to high-throughput assays and provides exciting possibilities for the future. On the basis of the high efficiency of delivery promised by this method, one might envisage the use of protein transduction technology in high-throughput assays for investigating signaling or enzymatic pathways and for screening the effects of drugs or small molecules on a transduced protein.

4. High-throughput approaches: transfected mammalian cell arrays

Analysis of gene function in mammalian cells traditionally involves the transfection of DNA constructs into cells to overproduce individual protein products and then monitoring its influence on the functions in various cellular processes. Typically, the adaptation of these approaches to screen large numbers of proteins for various cellular phenotypes is not practical. However, recent studies have started to develop the technologies, both transfection and image analysis, required to perform such cell transfection–based screens that go beyond the 96- or 364-well-based format.

Recently, a microarray-based transfection strategy termed reverse transfection or cell microarrays, has been described (Ziauddin and Sabatini, 2001). Using standard microarray robotics, nanoliter quantities of DNA products in an aqueous gelatin solution are printed onto a glass slide that can be stored. After exposure to the transfection reagent, the slides are placed in a culture dish and covered with cell suspensions of adherent mammalian cells. Cells growing on the printed areas take up the DNA, creating regions of localized transfection throughout a lawn of nontransfected cells. This high-throughput adaptation allows thousands of different expression constructs to be expressed in adherent mammalian cells in parallel on an individual glass slide. The effect of each expression construct can be gauged by studying the cluster of living cells present at each spot location. For example, a single spot of 200–500 μm diameter can contain 30–80 transfected cells (Figure 1). Fluorescent products can be detected at low resolution using microarray scanners or at high resolution using microscopy. In addition, methods have been developed to detect radiolabeled products using autoradiography film and the transfer of the cell arrays to nitrocellulose for analysis by western blotting. Further optimization of cell microarray techniques, image capture, and analysis has now been reported (Baghdoyan *et al*., 2004).

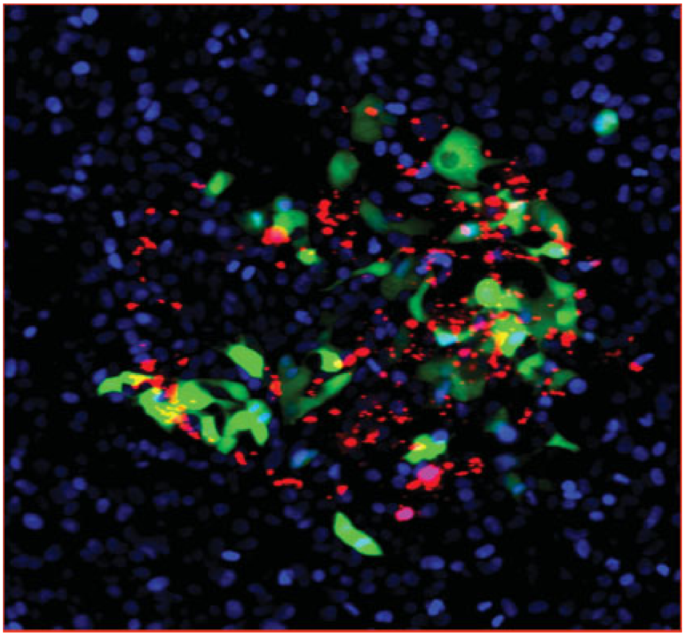


Figure 1 Reverse transfection microarray. This staining demonstrates the transfection technology first developed by the laboratory of David Sabitini (Ziauddin and Sabatini, 2001), and details of their methods are available on the following website (**http://jura.wi.mit.edu/sabatini_public/reverse_transfection/paper_home_page.htm**). This image, produced by A. Forrest, Institute for Molecular Bioscience, University of Queensland, shows a single printed array spot (500 µm in diameter), highlighted using a red Cy3-labeled protein carrier, containing an expression plasmid encoding GFP. Only the cells located over the array spot have been transfected with the plasmid as indicated by the green fluorescence from the expressed GFP. All cell nuclei have been staining with blue fluorescent DNA probe DAPI

To illustrate the utility of cell microarrays, a number of proof-of-principle assays have been established to detect gene products that can induce a range of cellular phenotypes. These assays, in principle, could be scaled to application on a genome-wide scale. Ziauddin and Sabitini (2001) identified gene products that increased the activity of kinase signaling pathways based on a general increase in tyrosine phosphorylation as detected by an anti-phosphotyrosine antibody. A more specific signaling pathway reporter has also been developed on the basis of the transcription reporter assay involving the serum response enhancer (SRE) element coupled to GFP to monitor transcriptional activity modulated by the transfection of individual gene products (Webb *et al*., 2003). A further level of analysis was superimposed in this case by using selective chemical inhibitors to block specific signaling pathways capable of activating SRE.

Protein microarrays that are generated by printing recombinant proteins onto slides at high spatial density have been screened for their ability to interact with proteins and chemical agents (Zhu *et al*., 2001; *see also* Article 24, **Protein arrays**, Volume 5). Similar applications have been developed for cell microarrays, which have the significant benefit of generating the protein and performing the assay

in the context of intact mammalian cells, rather than using purified recombinant individual proteins (Bailey *et al*., 2002). Ziauddin and Sabitini (2001) were able to detect radiolabeled small molecules FK506 and dopamine D1, specifically binding transfected cell arrays expressing their respective receptors FKBP12 and dopamine D1. Cell arrays have been used to analyze the ligand-induced activation of GPCR (Mishina *et al*., 2004). In this work, cell arrays that were prepared expressing 36 GPCRs with a promiscuous Gα subunit were screened for agonist stimulation using a fluorescence-based calcium flux assay. This approach allowed the parallel analysis of GPCRs to rapidly determine receptor specificity for individual agonists and was adaptable to assay the agonist stimulation or antagonist inhibition.

The cell array technology is also suitable for transfection of RNAi probes, either siRNA (Erfle *et al*., 2004; Kumar *et al*., 2003; Mousses *et al*., 2003; Silva *et al*., 2004; Vanhecke and Janitz, 2004) or DNA constructs that direct the expression of shRNA (Silva *et al*., 2004) to assay for knockdown phenotypes in mammalian cells. Using the cell arrays, Silva *et al*. (2004) validated putative positive shRNAs from a previous screen performed in a 96-well format (Paddison *et al*., 2004) for the ability to antagonize the proteosome function using a fluorescent reporter.

A fully automated high-content screening platform capable of automatically acquiring and analyzing high-resolution multicolor images of fluorescent cells has recently been developed (Liebel *et al*., 2003). Both the cell transfection and immunofluorescence labeling in this system were fully automated using liquid handling robotics in a 96-well plate format. While the transfection efficiency was lower (around 30%) than manual transfection, it was sufficient to obtain high-resolution images. Image acquisition required the development of software able to automatically identify the focal plane for each field of view to maximize the capture of "in-focus" images. These images were obtained using a motorized xyz-stage to automatically change samples. Recently, this image acquisition platform has been adapted to live cell arrays (Conrad *et al*., 2004) and further developed to include machine-learning-based image classification methods that allow for the automatic phenotype classification of the acquired images. While these methods were initially used to differentiate image localization (Conrad *et al*., 2004), this machine-learning approach can be readily trained to classify other cellular phenotypes. The further development of these automated high-content screening platforms will greatly expand the repertoire of screens for which the cell arrays can be applied.

5. Location – key information for analyzing expressed proteins

Eukaryotic cells contain a number of distinct, membrane-bound subcompartments or organelles. These organelles function to compartmentalize distinct biochemical pathways and physiological processes. In this context, organelles within the cell include distinct structures such as mitochondria and nuclei, more dynamic and changeable organelles such as recycling endosomes or the trans-Golgi network and the plasma membrane (cell surface) itself, all of which are further subdivided into discrete membrane domains (Johnson *et al*., 2004; Miaczynska and Zerial,

2002; Simons and Vaz, 2004). The cytoplasm forms another cellular compartment that is now recognized to mediate frequent exchange of proteins with the nucleus, endoplasmic reticulum, and other organelles. Determining organelle association or subcellular localization is an essential step in characterizing the possible function of a novel protein or in assessing the expression and function of known recombinant proteins. Methods and approaches for investigating subcellular localization are therefore important to all applications of recombinant technology, from analyzing the expression of individual proteins in cells and organisms through to the analysis of whole proteomes (Simpson *et al*., 2001). The subcellular localization of proteins is often complex and subject to change, in keeping with the growing awareness that many proteins have more than one function in the cell. A protein can be a stable, stationary resident of one organelle, exemplified by glycosyl transferases that are targeted to and function in a specific cisterna of the Golgi complex (Opat *et al*., 2001). Alternatively, proteins can have more than one functional location, an example of which is seen in cell surface receptors that reside initially on the plasma membrane where they function to bind ligand and are then internalized into endosomes from where they generate cell signaling (Gonzalez-Gaitan, 2003). Yet, further complexity arises during protein synthesis and processing during which time newly synthesized proteins pass transiently through the organelles of the secretory pathway, including the endoplasmic reticulum and Golgi complex and a series of carrier vesicles, on their way to their final destination (Lippincott-Schwartz *et al*., 2000). The membrane domain or organelle to which a protein is targeted and then resides is determined by information encoded in its amino acid sequence or by interactions with other proteins or lipids (van Vliet *et al*., 2003).

6. Tagging expressed proteins – the fluorescence revolution

For several decades, antibodies have been used as the main tool for detecting, localizing, and isolating proteins of interest (Myers, 1989). With the advent of recombinant protein expression have come new methods for tagging and visualizing proteins (Fritze and Anderson, 2000). Epitope tags are amino acid sequences that are added to DNA constructs to encode short peptides derived from mammalian proteins, such as the myc epitope, or from synthetic sequences, such as FLAG (Einhauer, 2001; Ellison and Hochstrasser, 1991). In the expressed protein, the epitope tags are then detected using antibodies for localization or isolation of the recombinant protein (Fritze and Anderson, 2000). Tagging recombinant proteins has become another essential component in the "toolbox" for gene and protein expression in individual experiments and in high-throughput systems. The introduction of fluorescent protein tags is surely the most significant advance in this field, revolutionizing our ability to track expressed proteins in fixed and in live eukaryotic cells. Just like antibodies, fluorescent protein tags have been developed for scientific use by exploiting nature itself. The prototypical fluorescent tag is GFP, a bioluminescent protein cloned from *Aequorea victoria* jellyfish with the genetically encoded ability to stably and brightly fluoresce (Chalfie *et al*., 1994). GFP can be expressed in the context of many expression vectors to form chimeras with

a protein of interest, which can then be introduced and expressed in cells or whole organisms. GFP is remarkably inert in eukaryotic cells. Perhaps the biggest bonus in most experimental systems is how little GFP overexpression appears to affect the biology and survival of mammalian cells, whole animals, or plants. Equally valuable is the general finding (with notable exceptions) that chimeric proteins incorporating GFP as a tag retain their ability to correctly fold, be trafficked within the cell, and to function. One of its potential drawbacks of using GFP as a tag is its large size (27 kDa), and of course there are (usually unreported) instances where the addition of GFP does interfere with protein synthesis and or function, and this must be assessed empirically in each case.

The discovery of other bioluminescent proteins such as DsRed (Baird *et al*., 2000) and further diversity in GFP probes arising from natural spectral variants of GFP or those created by specific mutagenesis of GFP (e.g., yellow or cyan fluorescent proteins, YFP and CFP respectively) have driven the development of techniques for sophisticated multiprobe fluorescence imaging and analysis (Labas *et al*., 2002). The simultaneous expression of multiple proteins tagged with distinct variants of GFP is now a common method for examining the colocalization of the expressed proteins or of organelle markers (Figure 2). There are now also many fluorescent probes available as constitutive or regulatable tags for

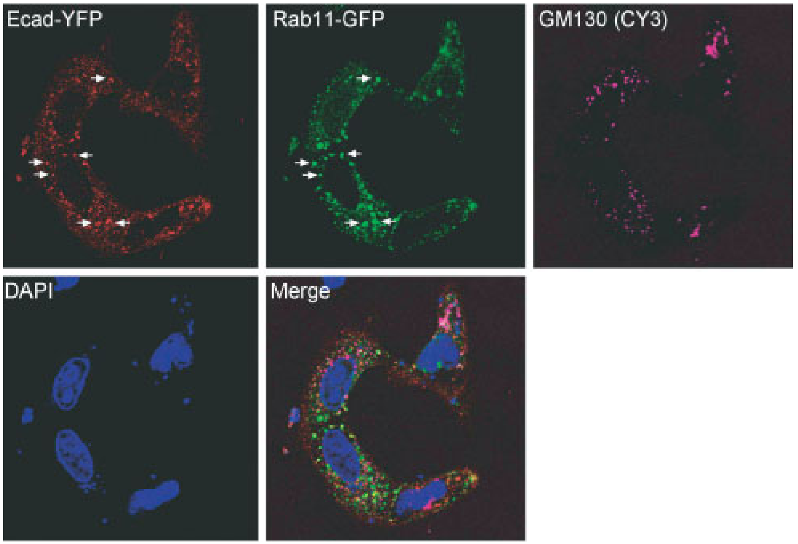


Figure 2 Fluorescence imaging and simultaneous subcellular localization with multiple probes. DNAs encoding two recombinant proteins tagged with different fluorescent tags (green fluorescent protein (GFP) or yellow fluorescent protein (YFP)) have been cotransfected and coexpressed in HeLa cells. After fixation, the cells were immunolabeled to detect a third protein using a CY-3 fluorescent antibody conjugate, and the cells were stained with DAPI to label DNA in the nuclei. Cells were imaged on a Zeiss confocal microscope, and different fluorescent signals were resolved using spectral unmixing. The images show the partial colocalization of a newly synthesized cell surface protein (E-cadherin) that has left the Golgi complex (marked with GM-130) and now transiently resides in recycling endosomes (marked with Rab11) on its way to the cell surface. Images supplied by J. Lock, Institute for Molecular Bioscience, University of Queensland

expressed proteins (Patterson and Lippincott-Schwartz, 2004; Zhang *et al*., 2002). Photoactivatable GFP is one example where, when exposed to bright light at specific wavelengths, the 100-fold increase in the fluorescence emission of this probe allows proteins to be located in a precise and temporal fashion as they move about the cell (Patterson and Lippincott-Schwartz, 2002). The chameleon probes combine multiple, domain-specific GFP variant probes into one gene product to detect calcium based on conformational induced fluorescence resonance energy transfer (FRET) (Miyawaki *et al*., 1997). Taken to an extreme, the multitag, multiexpression strategy will eventually be applied to visualizing complicated protein pathways or complexes under specific physiological conditions.

Techniques such as FRET, which in one form measures fluorescence quenching of a donor fluorophore emission by a closely adjacent acceptor, can be used to study protein–protein interactions in cells expressing CFP- and YFP-tagged proteins, or can be used to study conformational states of a single protein multiply labeled with tags on different domains (Jares-Erijman and Jovin, 2003). Individual events in cell signaling and whole signaling networks are being studied and assembled in cells by FRET analysis (Miyawaki, 2003). The activation state of heterotrimeric G-proteins was resolved in *Dictyostelium* by using FRET labeled with YFP and CFP on opposing G-protein subunits (Janetopoulos *et al*., 2001). The expression of fluorescently labeled proteins and FRET analysis revealed novel locations on the membranes of the endoplasmic reticulum and Golgi complex for critical protein interactions in the Ras signaling pathway (Chiu *et al*., 2002).

In fixed cells, GFP can be detected as a protein tag by virtue of its own fluorescence under UV light, or using antibodies to GFP, which can in turn be detected by standard secondary antibodies in immunofluorescence or by immunogold labeling at an ultrastructural level (Wylie *et al*., 2003). These secondary probes often enhance the detection of small amounts of GFP-expressed protein or enable double labeling with other probes (Gleeson *et al*., 2004).

In live cells, the use of GFP and its variants to tag recombinant proteins has almost single handedly opened up the world of real-time cell imaging (Lippincott-Schwartz *et al*., 2003; Shav-Tal *et al*., 2004). The ability to track individual proteins, carrier vesicles, and compartments in mammalian cells has had a significant impact on the field of protein trafficking. Live imaging of GFP-tagged cargo proteins has provided new insights into the nature of tubulovesicular carriers that perform trafficking, showing information not previously revealed by either light microscopy or electron microscopy (Wylie *et al*., 2003). Intravital imaging of GFP-tagged expressed proteins or other fluorescence markers in live whole animals (Bouvet *et al*., 2002) has now begun to show the real-time behavior of cells during disease processes.

7. Prediction and annotation of subcellular localization of expressed proteins

Analyzing the subcellular localization of individually tagged and expressed proteins is time consuming, and alternative methods for high-throughput or genome-wide

localization studies are only now being developed. To this end, protein localization will increasingly rely on computational methods to predict organellar locations as a prelude to experimental analysis (*see* Article 37, **Signal peptides and protein localization prediction**, Volume 7). Computational methods currently being developed are based on comparison of either the overall amino acid sequence or the sequences of specific domains of proteins with different subcellular localizations. In some cases, the localization of a protein can be attributed to the presence of linear amino acid motifs or sorting signals that actively target proteins to different regions of the cell or to specific organelles. The computational techniques aim to differentiate sets of proteins with known organelle distributions and with this as a base to develop predictions by which to categorize the localization of novel sequences. To date, the most successful methods can differentiate soluble proteins that localize to either the nucleus or cytoplasm, or those that enter the secretory pathway. Other methods have focused on predicting proteins associated with individual organelles including the nucleus (Cokol *et al*., 2000), peroxisome (Emanuelsson *et al*., 2003; Neuberger *et al*., 2003), mitochondrion (Emanuelsson *et al*., 2000; Emanuelsson and von Heijne, 2001; Guda *et al*., 2004), Golgi complex (Wrzeszczynski and Rost, 2004; Yuan and Teasdale, 2002), endoplasmic reticulum (Wrzeszczynski and Rost, 2004), or extracellular space (Chen *et al*., 2003; Grimmond *et al*., 2003). These methods report varying accuracy rates based on their ability to predict the localization of sets of known proteins that frequently overlap with the training sets used in their development, and yet further validation is needed to establish these computational methods for predicting the subcellular localization of novel proteins.

The "components" section of Gene Ontology (GO) terms contains the nomenclature used to describe subcellular localization (*see* Article 82, **The Gene Ontology project**, Volume 8), and then databases such as RefSeq and MGI that incorporate the GO terms can be used as sources to retrieve information regarding protein localization. In addition, Swiss-Prot contains a subcellular localization field (*see* Article 97, **Ensembl and UniProt (Swiss-Prot)**, Volume 8). However, one inherent source of inaccuracy in these databases is that their information on protein localization may be context specific. Many experimental and physiological factors can influence the subcellular localization of an individual protein in temporal and spatial fashions. Some of the compounding factors include:

1. C*ell type*: Different cell types have specialized or unique organelles, for instance, melanosomes in melanocytes or the distinct cell surface domains of polarized cells. Targeting proteins to such organelles or locations often involves unique sorting signals and protein machinery.
2. A*rtifacts of overexpression*: During synthesis and processing, many endogenous proteins form hetero-oligomers that are targeted to a specific organelle by a single constituent subunit. Gross overexpression of a recombinant protein can saturate the biosynthetic and sorting machinery of the cell, disrupting hetero-oligomerization and resulting in aberrant or unrepresentative localization of the foreign gene product.
3. *Overexpressed and tagged proteins*: The location and type of an epitope tag engineered into an open reading frame (ORF) can disrupt the subcellular localization

of the target protein. The addition of protein residues to the amino or carboxyl terminus of a protein can disrupt terminal sorting signals such as mitochondrial transit peptides, endoplasmic reticulum signal peptides, and endoplasmic reticulum sorting signals (Emanuelsson and von Heijne, 2001; Teasdale and Jackson, 1996). Also, although few cases have been documented, the addition of epitope tags can disrupt polypeptide folding resulting in the formation of aggresomes (Johnston *et al*., 1998). The addition of GFP with its relatively large mass (27 kDa) can sterically dominate some small proteins or render them too large for size-dependent sorting events like passage through the nuclear pore. Finally, the epitope tag can dictate subcellular localization in its own right, for example, GFP displays a nuclear and cytoplasmic localization and appears to contain a nuclear localization signal (NLS).

4. *Physiological status of the cell*: The subcellular localization of proteins can change as a direct result of, or incidental to, the stage of cell the cycle or in response to cell signaling events. At present, critical evaluation of these features in experimental approaches is still needed to fully and accurately determine subcellular localization.

8. High-throughput subcellular localization

Experimental approaches and computational approaches for subcellular localization rely on defining a staining pattern for an expressed protein and determining whether this matches one of the characteristic patterns known to represent a particular organelle (Figure 3). Staining patterns for organelles such as the endoplasmic reticulum, the Golgi complex, nuclei, cell surface, and mitochondria are readily recognizable in most mammalian cells, and obtaining one of these patterns for a novel protein gives a first indication as to its likely residence in the cell. Further experimentation or bioinformatics can begin to address some of the caveats to this apparent localization mentioned above; nonetheless, this shot-gun subcellular localization of expressed proteins remains an important information-gathering exercise. Recently, a genome-wide analysis of the subcellular localization of proteins in *Saccharomyces cerevisiae* proteins has been reported (Huh *et al*., 2003). Each of the 6234 annotated ORFs was tagged at the C-terminus with GFP using homologous recombination, and expressed proteins were assessed by fluorescence microscopy combined with colocalization experiments. Seventy-five percent of the yeast proteome was assigned a subcellular localization on the basis of analysis of the micrographs and in selected cases on colocalization with organelle markers. No such extensive study has been performed in mammals, with the majority of protein subcellular localization data being generated by individual groups focused on small sets of proteins. However, recently a number of projects have begun to systematically analyze the subcellular localization of mammalian proteins. First, the European Molecular Biology Labs (EMBL) has established a GFP-cDNA human protein localization project (Simpson *et al*., 2000; **http://gfp-cdna.embl.de/**), which aims to determine the subcellular localization of novel proteins encoded by transcripts generated by the German cDNA consortium (Wiemann *et al*., 2003). It has used the Gateway cloning system (Invitrogen) to

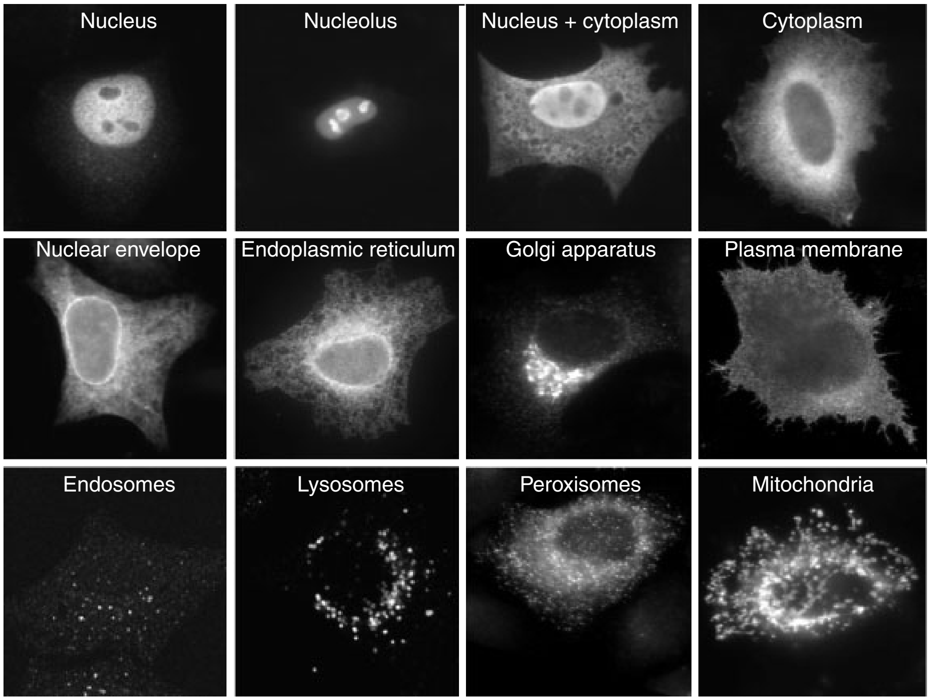


Figure 3 Organelle-staining patterns in cells. Localization of various known and novel expressed proteins in HeLa cells shows a range of diverse staining patterns, many of which are characteristic of subcellular organelles. Images supplied by K. Hanson, M. Kerr and R. Aturaliya, Institute for Molecular Bioscience, University of Queensland

enable rapid generation of N – and C – terminally GFP-tagged human ORFs. To date, over 800 novel human proteins have been analyzed and assigned to subcellular locations. Second, we have initiated a mouse protein localization database (**http://membrane.imb.uq.edu.au/**), which is focused on determining the subcellular localization of sets of membrane proteins predicted using a computational pipeline (Kanapin *et al*., 2003). A megaprimer approach (Suzuki *et al*., 2001) was used to generated expression constructs encoding novel or known membrane proteins with an 11 amino acid myc epitope fused to the N-terminal cytoplasmic domain of each membrane protein, excluding those predicted to contain a N-terminal endoplasmic reticulum signal peptide or sorting signal. To date, over 450 membrane proteins have been experimentally analyzed and assigned a subcellular localization. This experimental data is integrated with the primary literature describing the subcellular localization of membrane proteins into the one resource that currently contains primary subcellular localization data for over 25% of the 4169 predicted mouse membrane proteins.

In the future, the expression and localization of large numbers of proteins in mammalian cells will involve many of the techniques and approaches described herein. A typical approach might entail initial computational screening and sequence-based prediction of protein location, followed by experimental validation of localization and functional analysis of the expressed protein. Such experiments

will require methods incorporating sophisticated fluorescent tagging and efficient gene delivery. Proteins will need to be expressed in multiple cell lines and, perhaps, in animals. Finally, data from these validation experiments will need to be collected using automated microscopy and imaging, and collated and interpreted using intelligent pattern recognition software. The researcher will then have the task of piecing together the many overlapping protein locations, the functional information about expressed proteins and working these into models of mammalian cells in health and disease.

Acknowledgments

We would like to acknowledge colleagues and members of our laboratories, as noted, for providing the data in the figures. Our thanks to Darren Brown for help with preparation of the manuscript. JLS and RDT are each supported by fellowships and grants from the National Health and Medical Research Council of Australia and the Australian Research Council of Australia.

References

Baghdoyan S, Roupioz Y, Pitaval A, Castel D, Khomyakova E, Papine A, Soussaline F and Gidrol X (2004) Quantitative analysis of highly parallel transfection in cell microarrays. *Nucleic Acids Research*, **32**, e77.

Bailey SN, Wu RZ and Sabatini DM (2002) Applications of transfected cell microarrays in high-throughput drug discovery. *Drug Discovery Today*, **7**, S113–S118.

Baird GS, Zacharias DA and Tsien RY (2000) Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 11984–11989.

Batard P, Jordan M and Wurm F (2001) Transfer of high copy number plasmid into mammalian cells by calcium phosphate transfection. *Gene*, **270**, 61–68.

Becker-Hapak M, McAllister SS and Dowdy SF (2001) TAT-mediated protein transduction into mammalian cells. *Methods*, **24**, 247–256.

Beerens AM, Al Hadithy AF, Rots MG and Haisma HJ (2003) Protein transduction domains and their utility in gene therapy. *Current Gene Therapy*, **3**, 486–494.

Blesch A (2004) Lentiviral and MLV based retroviral vectors for ex vivo and in vivo gene transfer. *Methods*, **33**, 164–172.

Bouvet M, Wang J, Nardin SR, Nassirpour R, Yang M, Baranov E, Jiang P, Moossa AR and Hoffman RM (2002) Real-time optical imaging of primary tumor growth and multiple metastatic events in a pancreatic cancer orthotopic model. *Cancer Research*, **62**, 1534–1540.

Boyd A (2002) Exogenous DNA expression in eukaryotic cells following microinjection. *Methods in Cell Science*, **24**, 115–122.

Cepko CL, Roberts BE and Mulligan RC (1984) Construction and applications of a highly transmissible murine retrovirus shuttle vector. *Cell*, **37**, 1053–1062.

Chalfie M, Tu Y, Euskirchen G, Ward WW and Prasher DC (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802–805.

Chen Y, Yu P, Luo J and Jiang Y (2003) Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mammalian Genome*, **14**, 859–865.

Chernajovsky Y, Gould DJ and Podhajcer OL (2004) Gene therapy for autoimmune diseases: quo vadis? *Nature Reviews. Immunology*, **4**, 800–811.

Chiu VK, Bivona T, Hach A, Sajous JB, Silletti J, Wiener H, Johnson RL II, Cox AD and Philips MR (2002) Ras signalling on the endoplasmic reticulum and the Golgi. *Nature Cell Biology*, **4**, 343–350.

Cokol M, Nair R and Rost B (2000) Finding nuclear localization signals. *EMBO Reports*, **1**, 411–415.

Colosimo A, Goncz KK, Holmes AR, Kunzelmann K, Novelli G, Malone RW, Bennett MJ, and Gruenert DC (2000) Transfer and expression of foreign genes in mammalian cells. *BioTechniques*, **29**, 314–318, 320–322, 324 passim.

Conrad C, Erfle H, Warnat P, Daigle N, Lorch T, Ellenberg J, Pepperkok R and Eils R (2004) Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research*, **14**, 1130–1136.

Cronin CA, Gluba W and Scrable H (2001) The lac operator-repressor system is functional in the mouse. *Genes & Development*, **15**, 1506–1517.

Crooke ST (2004) Antisense strategies. *Current Molecular Medicine*, **4**, 465–487.

Delenda C (2004) Lentiviral vectors: optimization of packaging, transduction and gene expression. *The Journal of Gene Medicine*, **6**(Suppl 1), S125–S138.

Einhauer AJA (2001) The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. *Journal of Biochemical and Biophysical Methods*, **49**, 455–465.

Ellison MJ and Hochstrasser M (1991) Epitope-tagged ubiquitin. A new probe for analysing ubiquitin function. *The Journal of Biological Chemistry*, **266**, 21150–21157.

Emanuelsson O, Elofsson A, von Heijne G and Cristobal S (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. *Journal of Molecular Biology*, **330**, 443–456.

Emanuelsson O, Nielsen H, Brunak S and von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**, 1005–1016.

Emanuelsson O and von Heijne G (2001) Prediction of organellar targeting signals. *Biochimica et Biophysica Acta*, **1541**, 114–119.

Erfle H, Simpson JC, Bastiaens PI and Pepperkok R (2004) siRNA cell arrays for high-content screening microscopy. *BioTechniques*, **37**, 454–458, 460, 462.

Evans RM (1988) The steroid and thyroid hormone receptor superfamily. *Science*, **240**, 889–895.

Felgner PL, Gadek TR, Holm M, Roman R, Chan HW, Wenz M, Northrop JP, Ringold GM and Danielsen M (1987) Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 7413–7417.

Fitzsimons HL, Bland RJ and During MJ (2004) Promoters and regulatory elements that improve adeno-associated virus transgene expression in the brain. *Methods*, **28**, 227–236.

Fritze CE and Anderson TR (2000) Epitope tagging: general method for tracking recombinant proteins. *Methods in Enzymology*, **327**, 3–16.

Galla M, Will E, Kraunus J, Chen L and Baum C (2004) Retroviral pseudotransduction for targeted cell manipulation. *Molecular & Cellular Proteomics*, **16**, 309–315.

Gleeson PA, Lock JG, Luke MR and Stow JL (2004) Domains of the TGN: coats, tethers and G proteins. *Traffic*, **5**, 315–326.

Gonzalez-Gaitan M (2003) Signal dispersal and transduction through the endocytic pathway. *Nature Reviews. Molecular Cell Biology*, **4**, 213–224.

Grez M and von Melchner H (1998) New vectors for gene therapy. *Stem Cells*, **16**(Suppl 1), 235–243.

Grimmond SM, Miranda KC, Yuan Z, Davis MJ, Hume DA, Yagi K, Tominaga N, Bono H, Hayashizaki Y, Okazaki Y, *et al* (2003) The mouse secretome: functional classification of the proteins secreted into the extracellular environment. *Genome Research*, **13**, 1350–1359.

Guda C, Fahy E and Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.

Hannon GJ and Rossi JJ (2004) Unlocking the potential of the human genome with RNA interference. *Nature*, **431**, 371–378.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS and O'Shea EK (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Janetopoulos C, Jin T and Devreotes P (2001) Receptor-mediated activation of heterotrimeric G-proteins in living cells. *Science*, **291**, 2408–2411.

Jares-Erijman EA and Jovin TM (2003) FRET imaging. *Nature Biotechnology*, **21**, 1387–1395.

Johnson HM, Subramaniam PS, Olsnes S and Jans DA (2004) Trafficking and signalling pathways of nuclear localizing protein ligands and their receptors. *BioEssays*, **26**, 993–1004.

Johnston JA, Ward CL and Kopito RR (1998) Aggresomes: a cellular response to misfolded proteins. *The Journal of Cell Biology*, **143**, 1883–1898.

Jordan M and Wurm F (2004) Transfection of adherent and suspended cells by calcium phosphate. *Methods*, **33**, 136–143.

Kanapin A, Batalov S, Davis MJ, Gough J, Grimmond S, Kawaji H, Magrane M, Matsuda H, Schonbach C, Teasdale RD, *et al* (2003) Mouse proteome analysis. *Genome Research*, **13**, 1335–1344.

Kato S, Anderson RA and Camerini-Otero RD (1986) Foreign DNA introduced by calcium phosphate is integrated into repetitive DNA elements of the mouse L cell genome. *Molecular and Cellular Biology*, **6**, 1787–1795.

Knott A, Garke K, Urlinger S, Guthmann J, Muller Y, Thellmann M and Hillen W (2002) Tetracycline-dependent gene regulation: combinations of transregulators yield a variety of expression windows. *BioTechniques*, **32**, 796, 798, 800 passim.

Kreitzer G, Schmoranzer J, Low SH, Li X, Gan Y, Weimbs T, Simon SM and Rodriguez-Boulan E (2003) Three-dimensional analysis of post-Golgi carrier exocytosis in epithelial cells. *Nature Cell Biology*, **5**, 126–136.

Kumar R, Conklin DS and Mittal V (2003) High-throughput selection of effective RNAi probes for gene silencing. *Genome Research*, **13**, 2333–2340.

Labas YA, Gurskaya NG, Yanushevich YG, Fradkov AF, Lukyanov KA, Lukyanov SA and Matz MV (2002) Diversity and evolution of the green fluorescent protein family. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4256–4261.

Li S (2004) Electroporation gene therapy: new developments in vivo and in vitro. *Current Gene Therapy*, **4**, 309–316.

Liebel U, Starkuviene V, Erfle H, Simpson JC, Poustka A, Wiemann S and Pepperkok R (2003) A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Letters*, **554**, 394–398.

Lippincott-Schwartz J, Altan-Bonnet N and Patterson GH (2003) Photobleaching and photoactivation: following protein dynamics in living cells. *Nature Cell Biology* **5**, S7–S14.

Lippincott-Schwartz J, Roberts TH and Hirschberg K (2000) Secretory protein trafficking and organelle dynamics in living cells. *Annual Review of Cell and Developmental Biology*, **16**, 557–589.

Lu Y (2004) Recombinant adeno-associated virus as delivery vector for gene therapy–a review. *Stem Cells and Development*, **13**, 133–145.

Miaczynska M and Zerial M (2002) Mosaic organization of the endocytic pathway. *Experimental Cell Research*, **272**, 8–14.

Mishina YM, Wilson CJ, Bruett L, Smith JJ, Stoop-Myer C, Jong S, Amaral LP, Pedersen R, Lyman SK, Myer VE, *et al* (2004) Multiplex GPCR assay in reverse transfection cell microarrays. *Journal of Biomolecular Screening*, **9**, 196–207.

Miyawaki A (2003) Visualization of the spatial and temporal dynamics of intracellular signalling. *Developmental Cell*, **4**, 295–305.

Miyawaki A, Llopis J, Heim R, McCaffery JM, Adams JA, Ikura M and Tsien RY (1997) Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature*, **388**, 882–887.

Mousses S, Caplen NJ, Cornelison R, Weaver D, Basik M, Hautaniemi S, Elkahloun AG, Lotufo RA, Choudary A, Dougherty ER, *et al* (2003) RNAi microarray analysis in cultured mammalian cells. *Genome Research*, **13**, 2341–2347.

Myers JD (1989) Development and application of immunocytochemical staining techniques: a review. *Diagnostic Cytopathology*, **5**, 318–330.

Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A and Eisenhaber F (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of Molecular Biology*, **328**, 581–592.

Opat AS, van Vliet C and Gleeson PA (2001) Trafficking and localization of resident Golgi glycosylation enzymes. *Biochimie*, **83**, 763–773.

Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, *et al* (2004) A resource for large-scale RNA-interference-based screens in mammals. *Nature*, **428**, 427–431.

Patterson GH and Lippincott-Schwartz J (2002) A photoactivatable GFP for selective photolabeling of proteins and cells. *Science*, **297**, 1873–1877.

Patterson GH and Lippincott-Schwartz J (2004) Selective photolabeling of proteins using photoactivatable GFP. *Methods*, **32**, 445–450.

Safinya CR (2001) Structures of lipid-DNA complexes: supramolecular assembly and gene delivery. *Current Opinion in Structural Biology*, **11**, 440–448.

Sasmono RT, Oceandy D, Pollard JW, Tong W, Pavli P, Wainwright BJ, Ostrowski MC, Himes SR and Hume DA (2003) A macrophage colony-stimulating factor receptor-green fluorescent protein transgene is expressed throughout the mononuclear phagocyte system of the mouse. *Blood*, **101**, 1155–1163.

Scollay R (2001) Gene therapy: a brief overview of the past, present, and future. *Annals of the New York Academy of Sciences*, **953**, 26–30.

Shav-Tal Y, Singer RH and Darzacq X (2004) Imaging gene expression in single living cells. *Nature Reviews. Molecular Cell Biology*, **5**, 855–861.

Silva JM, Mizuno H, Brady A, Lucito R and Hannon GJ (2004) RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 6548–6552.

Simons K and Vaz WL (2004) Model systems, lipid rafts, and cell membranes. *Annual Review of Biophysics and Biomolecular Structure*, **33**, 269–295.

Simpson JC, Neubrand VE, Wiemann S and Pepperkok R (2001) Illuminating the human genome. *Histochemistry and Cell Biology*, **115**, 23–29.

Simpson JC, Wellenreuther R, Poustka A, Pepperkok R and Wiemann S (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Reports*, **1**, 287–292.

St George JA (2003) Gene therapy progress and prospects: adenoviral vectors. *Gene Therapy*, **10**, 1135–1141.

Suzuki H, Fukunishi Y, Kagawa I, Saito R, Oda H, Endo T, Kondo S, Bono H, Okazaki Y and Hayashizaki Y (2001) Protein-protein interaction panel using mouse full-length cDNAs. *Genome Research*, **11**, 1758–1765.

Tan W, Zhu K, Segal DJ, Barbas CF III and Chow SA (2004) Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. *Journal of Virology*, **78**, 1301–1313.

Teasdale RD and Jackson MR (1996) Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the golgi apparatus. *Annual Review of Cell and Developmental Biology*, **12**, 27–54.

Tranchant I, Thompson B, Nicolazzi C, Mignet N and Scherman D (2004) Physicochemical optimisation of plasmid delivery by cationic lipids. *The Journal of Gene Medicine*, **6**(Suppl 1), S24–S35.

Trezise AE (2002) In vivo DNA electrotransfer. *DNA and Cell Biology*, **21**, 869–877.

van Berkel PH, Welling MM, Geerts M, van Veen HA, Ravensbergen B, Salaheddine M, Pauwels EK, Pieper F, Nuijens JH and Nibbering PH (2002) Large scale production of recombinant human lactoferrin in the milk of transgenic cows. *Nature Biotechnology*, **20**, 484–487.

Vanhecke D and Janitz M (2004) High-throughput gene silencing using cell arrays. *Oncogene*, **23**, 8353–8358.

van Vliet C, Thomas EC, Merino-Trigo A, Teasdale RD and Gleeson PA (2003) Intracellular sorting and transport of proteins. *Progress in Biophysics and Molecular Biology*, **83**, 1–45.

Wadhwa R, Kaul SC, Miyagishi M and Taira K (2004) Vectors for RNA interference. *Current Opinion in Molecular Therapeutics*, **6**, 367–372.

Wadia JS, Stan RV and Dowdy SF (2004) Transducible TAT-HA fusogenic peptide enhances escape of TAT-fusion proteins after lipid raft macropinocytosis. *Nature Medicine*, **10**, 310–315.

Webb BL, Diaz B, Martin GS and Lai F (2003) A reporter system for reverse transfection cell arrays. *Journal of Biomolecular Screening*, **8**, 620–623.

Wiemann S, Bechtel S, Bannasch D, Pepperkok R and Poustka A (2003) The German cDNA network: cDNAs, functional genomics and proteomics. *Journal of Structural and Functional Genomics*, **4**, 87–96.

Wrzeszczynski KO and Rost B (2004) Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. *Cellular and Molecular Life Sciences*, **61**, 1341–1353.

Wylie FG, Lock JG, Jamriska L, Khromykh T, Brown DL and Stow JL (2003) GAIP participates in budding of membrane carriers at the trans-Golgi network. *Traffic*, **4**, 175–189.

Yuan Z and Teasdale RD (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains. *Bioinformatics*, **18**, 1109–1115.

Zhang J, Campbell RE, Ting AY and Tsien RY (2002) Creating new fluorescent probes for cell biology. *Nature Reviews. Molecular Cell Biology*, **3**, 906–918.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al* (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Ziauddin J and Sabatini DM (2001) Microarrays of cells expressing defined cDNAs. *Nature*, **411**, 107–110.

Specialist Review

Integrating genotypic, molecular profiling, and clinical data to elucidate common human diseases

Eric E. Schadt and Alan Sachs
Rosetta Inpharmatics, LLC, Seattle, WA, USA

1. Introduction

One classic approach in widespread use to identify key drivers of complex traits such as common human diseases is the genetic approach proposed by Botstein *et al*. (1980). In this genetic approach, variations in DNA are treated as markers and tested in human and experimental populations for cosegregation with diseases of interest. Markers that cosegregate with a given disease highlight regions in the genome that are considered linked to the disease and that at least partially explain the susceptibility to the disease. Positional cloning strategies and direct testing of positional candidate genes in such regions are then pursued to identify the gene supporting the linkage signal.

This forward genetics approach has proven successful in human and experimental systems for identifying broad regions in the genome associated with disease, and in the case of Mendelian disorders in actually identifying genes underlying such disorders. However, identifying genes harboring common variants that are associated with common human diseases has been less successful, given a number of issues that make such identifications difficult: (1) inadequate sample size, given the frequency of disease or size of the genetic component of disease that can be explained in the population under study; (2) disease heterogeneity (affects ability to initially detect and to replicate); (3) size of regions supported by a given linkage can contain hundreds or even thousands of genes, and narrowing down the region to the gene or genes underlying the linkage is a long and laborious process that is not guaranteed to lead to success; (4) phenotyping inconsistencies resulting in misclassification of individuals into disease groups, which contaminates the signal genetic studies aim to detect; and (5) failure to account for unknown environmental influences or interactions among genes or between genes and environment.

Genetic association studies, in contrast to linkage studies, have increased power to detect genes leading to disease, but these types of studies come with their own set of issues that have prevented broad success. Identifying candidate genes for

consideration in a genetic association test is often highly subjective, driven by biological intuition of what a gene's functional role may be in a disease, rather than more objective data directly implicating the gene's role in disease. Population substructure and admixture may also confound the interpretation of results from these types of studies (Ziv and Burchard, 2003). Further, linkage disequilibrium operating over regions supporting multiple genes containing many variants makes it difficult to identify the gene supporting the association in some cases and the causal variant in the great majority of the cases. Replication is an issue as well given many of the associations detected in one population fail to replicate in other independent populations, and in cases in which replication is achieved the replicated association is often to a different haplotype, complicating the interpretation of the results (Helgadottir, 2004). Even with advances in genotyping technologies and the nearing completion of a human haplotype map that make genome-wide association scans a reality, issues related to the application of appropriate statistical models, multiple testing, and the need to replicate discoveries in such a setting are still problematic. In cases in which associations are detected, it is often the case that the haplotype or causal variants identified as associated with disease suggest no obvious functional role for the associated gene (Helgadottir, 2004; Gretarsdottir, 2003; Pajukanta, 2004; Stefansson, 2002). Finally, genetic studies fail altogether to identify genes central to a complex trait when common variants in such genes simply do not exist in the general population, making detection by genetic linkage or association impossible. Genes such as HMG-CoA reductase regulating important biological processes such as cholesterol synthesis have not yet had common variants identified that associate with traits related to those key biological processes.

Molecular profiling strategies, on the other hand, have the potential to provide much of the needed functional insights into genes and networks more generally found to be associated with disease. The ability to score tens of thousands of phenotypes (e.g., gene expression traits) simultaneously and link those phenotypes to disease traits of interest offers the ability to take an unparalleled look at the molecular processes underlying disease. There has been an explosion of the use of gene expression microarrays and proteomics technologies over the past 5 years toward this end, and a number of novel discoveries have been obtained through the use of this technology, including identification of genes and patterns of expression causally associated with complex disease phentoypes (Karp, 2000; Schadt *et al*., 2003a; Schadt *et al*., 2005), uncovering novel gene structure and alternative splicing events in genes known to be associated with complex disease traits (Johnson, 2003; Schadt, 2004; Shoemaker, 2001), identification of biomarkers (DePrimo, 2003), identification of disease subtypes (Schadt *et al*., 2003a; Mootha, 2003; van 't Veer, 2002), and identification of mechanisms of toxicity (Waring *et al*., 2002).

However, molecular profiling technologies on their own do not provide the information necessary to directly identify key drivers of disease. Technologies such as gene expression microarrays allow for the detection of gene–gene interactions, but do not provide the causal information necessary to infer direction among the many interactions. Others have explored establishing causal relationships among genes by integrating data from multiple sources, including gene expression and proteomic data, single-gene perturbation experiments, and databases of experimentally determined transcription regulatory domains (Lee, 2002; Luscombe,

2004; Sabo, 2004). More recently, the integration of gene expression and genotypic data has been explored as another approach to inferring causal relationships among genes and disease traits. This approach makes use of the naturally occurring DNA variations in human or experimental populations and the perturbations these variations give rise to at the molecular level, and that in turn lead to the complex traits associated with diseases of major public health concern. Jansen and Nap (2001) were among the first to suggest the use of expression profiles in segregating populations toward this end. Studies in *Drosophila* (Jin, 2001; Wayne and McIntyre, 2002), yeast (Brem *et al*., 2002; Yvert, 2003), and mouse (Schadt *et al*., 2003a; Schadt *et al*., 2005; Zhu *et al*., 2004) soon followed, and more recently general surveys on the genetics of gene expression in humans have appeared (Monks, 2004; Morley, 2004).

These recent studies combining genotypic and molecular profiling data to uncover the complex gene networks underlying complex physiological traits related to common human diseases have begun to demonstrate the power that exists in segregating populations to establish causal associations to a wide array of complex traits of interest. In what follows, we review some of these approaches to elucidating common human diseases using integrative genomics approaches where genotypic, molecular profiling, and phenotypic data are combined to refine the definition of a disease trait, identify disease subtypes, make use of the causal information provided by the genotypic data in segregating populations to identify key drivers of disease, and to more generally reconstruct gene networks associated with disease.

2. Experimental setup

While the approach discussed here can be applied to almost any organism, we focus primarily on mouse models for common human diseases. The use of inbred strains of mice to dissect the genetic complexity of common human diseases offers a viable alternative to human studies given that experimental parameters such as environment, breeding scheme, and detailed phenotyping, can be controlled. In addition, mouse models are arguably more relevant to common human diseases than other frequently used experimental systems such as cell lines, yeast, and *Drosophila*. Finally, the mouse models discussed here illustrate the potential power this integrative genomics approach may bring to elucidating the complexity of common human diseases.

The aim of genetic studies in the type of mouse population discussed here is to identify quantitative trait loci (QTL) for molecular profiling and disease traits of interest, and ultimately, to identify the gene or genes underlying these QTL. A standard experimental design employed in these studies consists of crossing two inbred strains of mice that often differ with respect to a phenotype of interest. The two inbred lines of interest are bred and the F1 progeny from this mating are either backcrossed to one of the original inbred lines, giving rise to a backcross (BC) population of mice, or the F1 animals are intercrossed to give rise to an F2 population of mice. While many variations to this type of breeding scheme exist, these types of crosses have been one of the primary workhorses for mouse genetics over the past 30 years. The inbred lines, F1 animals, and BC or F2 animals are then taken

through an experimental protocol that may involve exposure to different environmental assaults, phenotyping with respect to traits of interest, and the BC or F2 animals are genotyped at markers uniformly spaced throughout the genome. In addition, tissues may be extracted and processed for molecular profiling purposes. All of these data taken together form the starting point of the approaches reviewed here.

The integrative genomics approach to dissecting complex traits are highlighted below on a previously described F2 intercross population constructed from the C57BL/6 J and DBA/2 J strains of mice (Schadt *et al*., 2003a; Drake, 2001). In this population, referred to here as the BXD data set, 111 female F2 mice were maintained on a rodent chow diet for up to 12 months of age, and then switched to an atherogenic high-fat, high-cholesterol diet for another 4 months. At 16 months of age, the mice were phenotyped and their livers extracted for gene expression profiling. As described by Drake *et al*. (2001), these mice model the spectrum of disease in a natural population, with many mice developing atherosclerotic lesions, and others having significantly higher fat-pad masses, higher cholesterol levels, and higher bone densities than others in the same population.

The expression studies highlighted for the BXD data set were carried out using oligonucleotide microarrays containing more than 23 000 genes (Schadt *et al*., 2003a; Hughes, 2001). RNA from liver samples of the BXD animals were hybridized against a pool of RNA composed of equal aliquots of RNA from each F2 animal. The gene expression traits treated as quantitative traits in genetic analyses were formed as the log ratio of the intensity measures for each gene represented on the microarray for each sample profiled. While the analyses highlighted in the examples below are focused on these gene expression traits, the general methods described could be applied to any molecular profiling data, including gene expression, proteomics, and metabolomics.

3. Refining the definition of complex diseases and identifying disease subtypes

3.1. A general approach to the identification of disease subtypes

One of the complexities encountered in dissecting common human diseases is the degree of disease heterogeneity represented in almost every population studied. Different pathways associated with different molecular processes may underlie a particular disease, so that any given population represents a mixture of these different subtypes, as depicted in Figure 1(a). Therefore, while individuals may appear phenotypically similar with respect to a physiological trait, the molecular processes that underlie the disease may vary widely from individual to individual. Several studies have highlighted such heterogeneity using microarrays, including a recent study associating patterns of gene expression from breast cancer tumor samples with survival (van ’t Veer, 2002; van de Vijver, 2002). Other studies have demonstrated patterns of expression in different tumor types that allow for disease classification beyond what could be achieved via clinical (e.g., histological) methods (Nutt, 2003; Pomeroy, 2002; Singh, 2002).

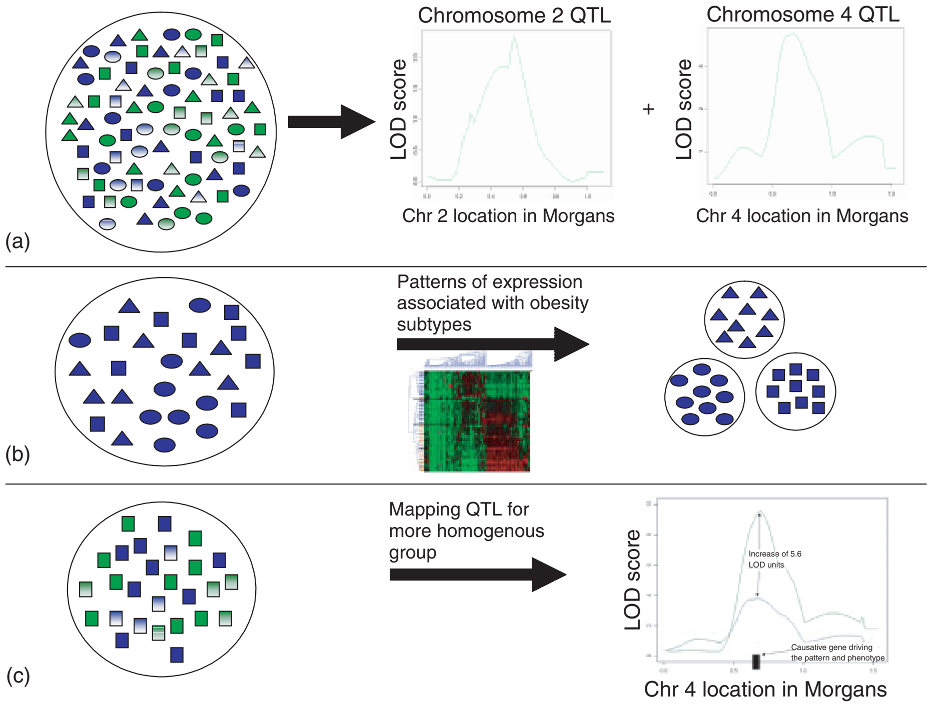


Figure 1 Integrating molecular profiling and genotypic data to refine the definition of disease and identify subtypes of disease. (a) A common scenario in genetic studies is represented here, where the population under study (represented by the large circle) is composed of many families segregating the disease, but with different families potentially segregating different subtypes of the disease. The shapes represent different subtypes of a given disease (here we assume obesity), and the colors represent the extent of disease. In this example, green represents individuals that are lean and blue represents individuals who are obese, with shaded individuals falling somewhere in between. A genetic analysis of this hypothetical population reveals two suggestive linkages to obesity, but neither achieves significance given contamination of the genetic signals by the different subtypes. (b) Expression data from the most obese (or most lean) individuals in the population are examined to identify patterns of expression that are specific to subgroups within the extreme group and that are associated with disease. Therefore, the definition of disease is redefined using the expression data, and the different patterns of expression can result in the identification of disease subtypes. (c) Patterns of expression defining disease subtypes are used to reclassify all individuals in a given population, and then genetic analyses are carried out on these more homogenous subgroups, resulting in a boost to the genetic signal associated with a particular subtype. In this instance, the chromosome 4 linkage here is seen to increase 5.6 LOD units over what was detected when the whole population was considered as a single group (a)

Schadt *et al*. (2003a) demonstrated a novel approach to refining the definition of complex disease states and identifying subtypes by integrating gene expression and genotypic data, providing direct evidence that different subtypes of disease represented in a given population can arise owing to different genetic causes. Figure 1 demonstrates the process used in this study to refine the definition of an obesity phenotype in the BXD data set and then to identify subtypes of the obesity phenotype under the control of distinct genetic loci. Examining a disease trait in a population without knowledge of the different subtypes of the disease represented can lead to significantly reduced power to identify associations to the disease, as depicted in Figure 1(a) and (c). However, if molecular profiling data can

be leveraged to identify more homogenous subgroups, as depicted in Figure 1(b), then performing genetic analyses on these more homogenous subgroups may lead to significantly increased power to identify associations, as shown in Figure 1(c).

3.2. *Application of this approach highlights subtypes of obesity under control of distinct genetic loci*

In the study by Schadt *et al*. (2003a), expression data were used to identify a set of gene expression traits that were able to distinguish the most obese and lean animals. As shown in Figure 2, the set of genes able to discriminate the obese and lean animals gave rise to two distinct patterns of expression in the obese animals, indicating that multiple molecular processes specific to each of the obese groups may explain fat-mass variation in each group. The patterns of gene expression associated with the different high-fat animals refine the definition of the obesity trait beyond what would be possible with the gross clinical characterization of fat mass alone.

To establish whether the different patterns of expression associated with the obesity phenotype were associated with different causes of obesity, Schadt *et al*.

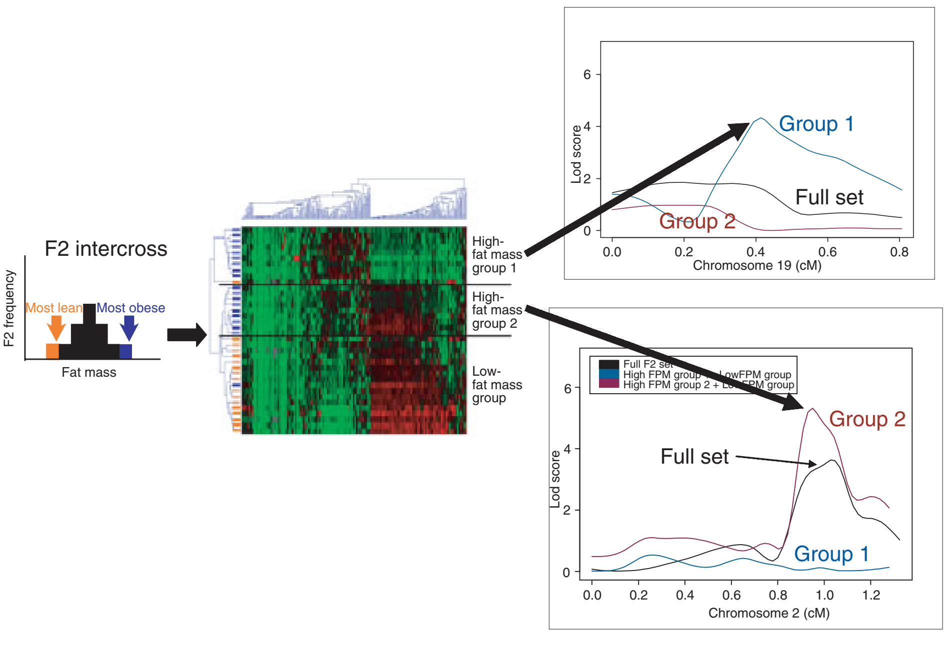


Figure 2 The type of procedure depicted in Figure 1 was applied to identify different subtypes of obesity in a segregating mouse population (Schadt *et al*., 2003a). Extreme lean and obese animals from an F2 population were identified, and expression traits from the most transcriptionally active genes in the liver of these animals were clustered in two dimension using an agglomerative hierarchical clustering procedure. The obese and lean groups clustered together and two distinct patterns of expression were detectable in the obese animals. The two obese groups were seen to represent two different subtypes of obesity, as verified by genetic analysis. The LOD score curves indicate that high-fat group 1 is controlled by a chromosome 19 locus, but not by the chromosome 2 locus, whereas high-fat group 2 is controlled by the chromosome 2 locus but not the chromosome 19 locus

classified all of the F2 animals into one of the three groups defined by the expression data: high-fat group 1, high-fat group 2, or low-fat group. Genetic analysis of each of the two high-fat groups revealed they were driven by distinct genetic loci. As shown in Figure 2, the chromosome 2 QTL completely vanishes when considering mice with expression patterns most similar to that of the high-fat group 1 mice, but increases by almost 2 LOD units over the original LOD score when considering mice most similar to the high-fat group 2 mice. A similar such locus was identified on chromosome 19, where in this case the QTL was specific to animals most similar to high-fat group 1 mice, not high-fat group 2 mice.

This approach provides for subtyping methods that may be less biased than more classic subtyping approaches that seek to identify individuals most informative for a given genetic association. This is because the expression data are an independent source of information used to refine the definition of disease, identify putative subtypes of disease, and ultimately classify the population into the subtypes. In addition, the accuracy of the classifications based on gene expression data can be supported by the simultaneous increases and decreases in linkage strength with respect to the different subtypes. While microarray platforms are subject to a variety of preprocessing issues that can impact interpretation of the data (Miklos and Maleszka, 2004), especially in the classification setting, others have shown that variation in these types of experiments can be controlled (He, 2003), and successes in refining the definition of disease using microarray data have been realized (van 't Veer, 2002; van de Vijver, 2002; Singh, 2002; He, 2003). Still, the application of molecular profiling data in this setting is in its infancy, and further investigations will be required before the utility of the approach can be more generally achieved.

In addition to providing for increased power to detect linkages to disease, refining disease definition using molecular profiling strategies has potentially broad implications for drug discovery as well (Schadt *et al*., 2003b). For instance, because the different subtypes of a given disease may be driven by different, unrelated pathways, different drugs may be needed to target each subtype effectively. In the example described above, developing a drug that targets the gene underlying the chromosome 2 QTL in Figure 2 may only prove effective for the fat group 2 obesity subtype. Identifying that part of the population that stands to benefit the most from such a drug would result in greater efficacy, which not only has the potential to better impact public health, but may also enable greater productivity in the drug development and diagnostic components of the pharmaceutical industry.

4. Tests for causality to distinguish between causal and reactive RNAs

As discussed above, standard molecular profiling experiments are not able to distinguish "causal" events (e.g., RNAs that serve as causative drivers of disease phenotypes) from "reactive" events (e.g., those RNAs whose levels change as a consequence of the disease state) without additional experimentation. This has limited the utility of molecular profiling data (gene expression data in particular) in identifying the key drivers of disease. Several groups have previously spoken

of a strategy of identifying key drivers of complex traits by examining genes located in regions of the genome genetically linked to a complex trait, and then looking for colocalization of cis-acting expression QTL for those genes residing in the region linked to the disease (Schadt *et al*., 2003a,b; Brem *et al*., 2002; Monks, 2004; Morley, 2004). Those genes with (1) expression values that are significantly correlated with the disease trait, (2) transcript abundances controlled by QTL that colocalize with the disease QTL, and (3) physical locations supported by the disease and expression QTL are natural causal candidates for the disease trait. This approach highlights how in segregating populations it is possible to identify variations in DNA sequence that cosegregate with variations in a disease trait, or, more generally, with variations in any complex trait. Whether occurring naturally in a segregating population or induced artificially by methods such as construction of gene knockout in transgenic animals, or gene knockdown using shRNA expression, perturbations such as these allow for the direct study of causal associations between molecular profiling and disease traits, as depicted in Figure 3(a).

To demonstrate how variations in DNA can be associated with variations in molecular profiling traits and disease traits, and how the degree of association can be used to order the molecular profiling traits relative to one another and relative to the disease traits, consider the following simple example. Given that two traits (e.g., an RNA trait and disease phenotype) are driven by the same DNA locus, only three relationships between the two traits are possible, with respect to the DNA locus. Figure 3(c–e) highlight the three possible relationships between a hypothetical expression trait and obesity trait associated with a given locus in a hypothetical mouse population. Figure 3(c) represents the simplest causal RNA model in which the DNA locus (L) controls the variation in the RNA levels of expression (R), which in turn modulates the obesity trait (C). Figure 3(d) depicts the reactive model, in which the DNA locus controls the variation in the obesity trait, which in turn modulates the RNA levels of expression. The independent model is given in Figure 3(e), where the DNA locus independently controls the variation in the obesity trait and RNA levels.

To illustrate how the relationship best supported by the data can be established given gene expression data, or molecular profiling data more generally, and genotypic data in a given population, consider the hypothetical mouse population in Figure 3(b) in which half of the mice have the AA genotype and the other half have the BB genotype at a given locus. As depicted in Figure 3(b), all mice with the BB genotype are obese, while 87.5% of the mice with the AA genotype are lean and the other 12.5% are obese. Further, 87.5% of the BB mice have higher transcript levels with respect to a specific gene, while the other 12.5% have unchanged levels, assuming here that the expression levels for a given mouse are taken relative to the average level over all mice. Similarly, 87.5% of the AA mice have lower transcript levels with respect to the same gene, while the other 12.5% have unchanged transcript levels. To avoid issues of statistical significance surrounding these frequencies, we assume that the hypothetical population is arbitrarily large. In this case, if the obesity and expression trait were uncorrelated with the genotype at locus L (i.e., not significantly linked to this locus), we would expect an equal percentage for each of the expression/obesity trait combinations for each genotype

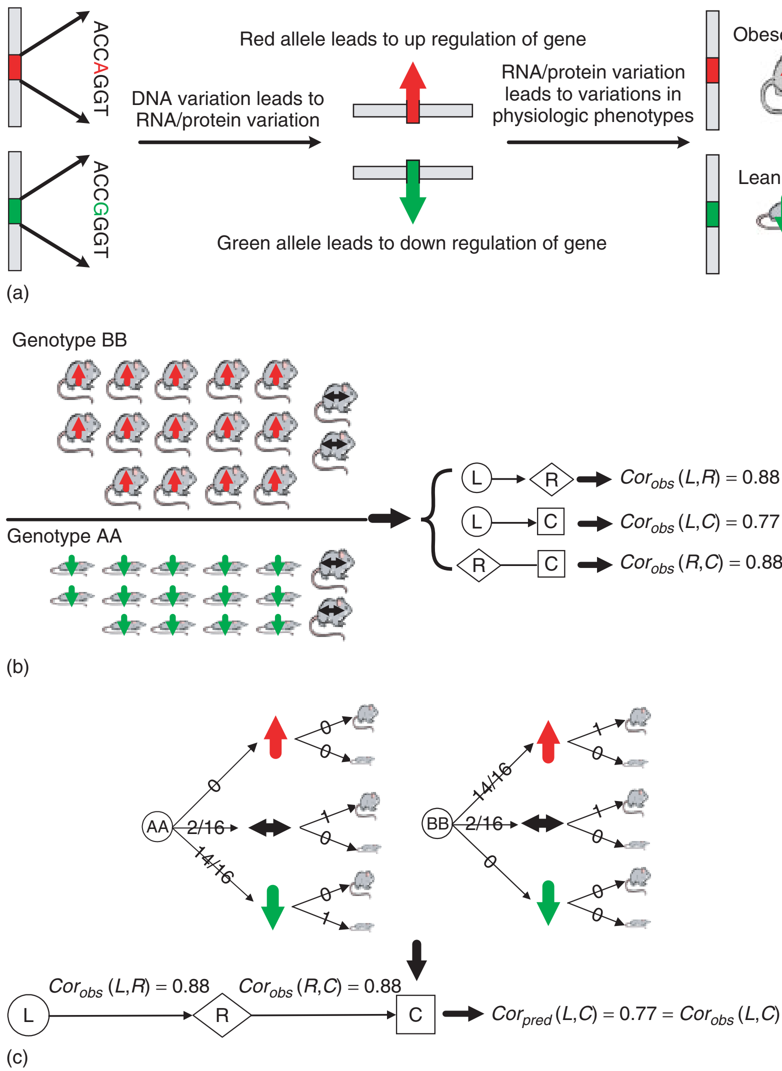
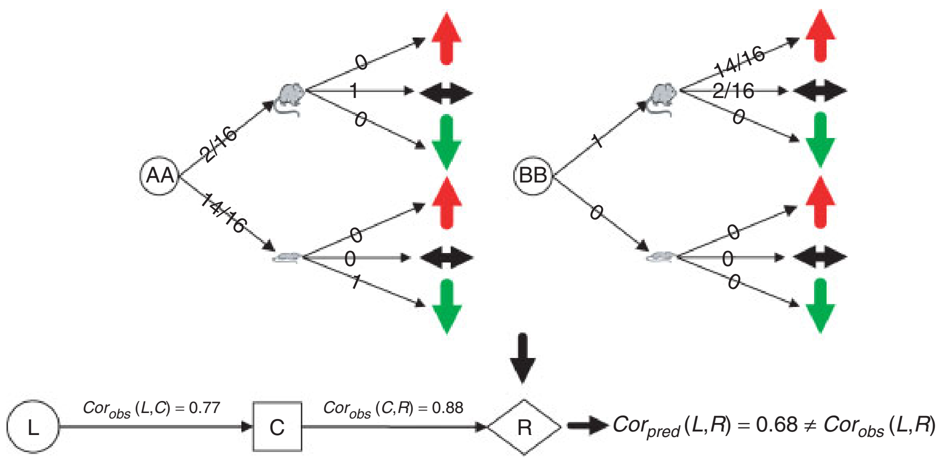
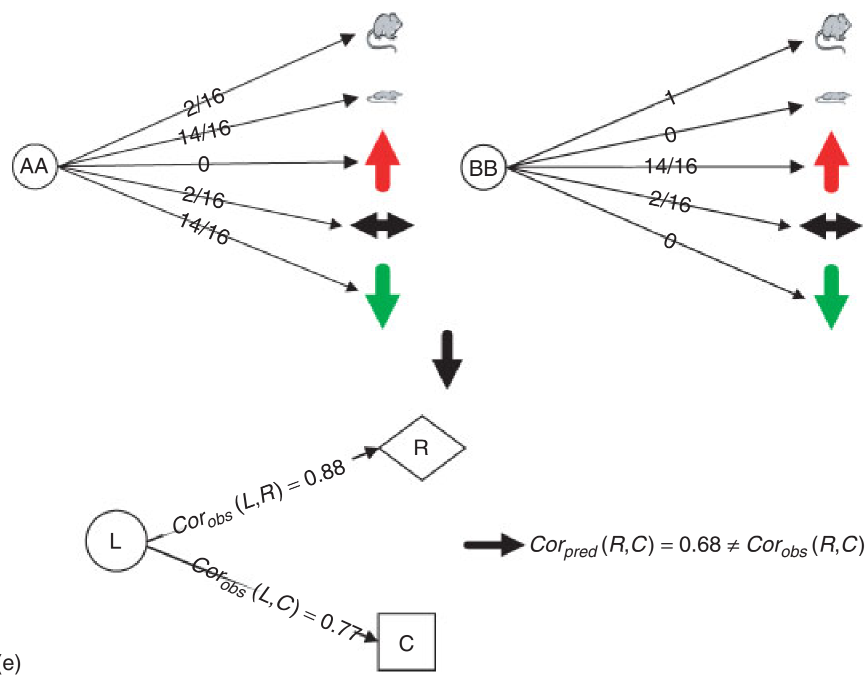


Figure 3 Inferring causality from a hypothetical mouse population. (a) Inferring causality from molecular profiling data follows naturally from the Central Dogma of Biology. Here the red allele ("A") and green allele ("G") are shown to lead to up- and downregulation, respectively, of a gene expression trait associated with the DNA polymorphism. In turn, these changes lead to perturbations in the transcriptional network that in turn lead to obese and lean mice. (b) Obesity is shown to cosegregate with genotypes at a hypothetical locus in a hypothetical mouse population. The expression of a particular gene is also seen to be associated with this same locus, where red, black, and green arrows indicate up-, neutral, and downregulation of the gene, respectively. The structure of the population depicted here allows us to compute the observed correlation coefficients provided to the right of the diagram (Cox and Hinkley, 1974), where L represents the locus associated with the AA and BB genotypes, R represents the transcript abundances of the gene, and C represents the obesity phenotype (either obese or lean). (b, c, d, and e) From the observed correlation structure between the locus, gene expression trait, and obesity trait given in (b), we can assess whether the relationship between the expression trait and the obesity trait is best supported by the model depicted in (c), (d), or (e). For each diagram, the numbers given along the directed edges represent the fraction of animals with the trait value given at the end of the edge, conditional on the trait value at the beginning of the edge. These fractions can be derived from the information in (b) and form the basis of the correlation coefficient calculations. As described in the text, the expression trait as causal for the obesity trait (c) is the relationship best supported by the data



(d)



(e)

Figure 3 (*continued*)

at locus L. Since this is clearly not the case in Figure 3(b), we can conclude that the expression and obesity traits are significantly linked to locus L.

To determine in this example if the RNA is a cause or consequence of the clinical state, the correlation structures between the variables of interest are examined in the context of each of the possible models shown in Figure 3(c–e). Figure 3(c) highlights the causative model, with respect to the RNA trait, where the correlation

between the genotypes and obesity trait predicted from the model is seen to be consistent with the observed correlation. This scenario is equivalent to the situation in which the correlation between the obesity trait and genotype, given the gene expression state, is 0 (i.e., the obesity trait and genotypes are conditionally independent, given the expression trait). Because the obesity trait and genotype are uncorrelated, once we condition on transcript abundances, we can conclude that the RNA is causal for the clinical trait (Pearl, 1983). Figure 3(d) highlights the reactive model, in which the observed correlation between the gene expression trait and genotype is 0.88, but the predicted value is 0.68. Therefore, the correlation between the expression trait and genotype predicted from the model does not equal the observed correlation. In this case, the expression trait and genotypes will still be significantly correlated after we condition on the obesity trait, confirming that the RNA levels are not responding to the obesity trait. Finally, Figure 3(e) highlights the independent model, where again the correlation between the gene expression and obesity traits predicted from the model is not consistent with the observed correlation. Therefore, given the results of the fits to these three models, the data for this hypothetical example indicate that the causative model is the most parsimonious and thus is the best explanation of the underlying biology. We conclude that the AA/BB locus controls variation in the RNA levels and that this RNA, in turn, controls variation in the obesity trait, rather than the RNA levels changing as a consequence of the obesity.

This type of reasoning can be formalized in statistical algorithms and applied systematically to genotypic, molecular profiling, and clinical data to identify causal relationships among molecular profiling traits and between molecular profiling and clinical traits. Schadt *et al*. (2005) develop this approach and apply it to the B × D data set, identifying and validating three novel susceptibility genes for obesity. However, it should be noted that the example described above was simplified in order to describe how causal information can be derived from the genotypic and molecular profiling data. In practice, there are a number of issues to consider before such an approach can be carried out. One such issue is the dependency of the identification of the correct relationship among traits on measurement and modeling errors. Suppose RNA trait R is actually causal for trait C, but the measurement errors related to the expression of R far exceed that of C. This may lead to a failure to detect R as causal for C, or worse, would incorrectly identify C as causal for R. Therefore, it is important that measurement errors be appropriately controlled for, either through repeated measures, employing a more accurate means of measuring, or more precisely modeling the error in the likelihood equations. There are issues in the model selection procedure as well in determining the relationship among traits driven by a common locus. For example, two traits driven independently by a given locus may not be conditionally independent given the locus genotypes. Therefore, it is important to model hidden variables that allow for correlation between two traits that is independent of the given locus. Finally, the high-dimensional nature of this problem in which tens of thousands of molecular profiling traits are considered simultaneously gives rise to serious multiple testing issues. These types of statistical issues are complex, and have only recently begun to be systematically addressed in the published literature.

5. Elucidating genetic networks underlying disease

The ability to infer causal associations among traits of interest can be more generally applied to one of the most pressing problems in complex disease research: the reconstruction of gene networks associated with disease. In order to make connections between the large sets of genes screened in molecular profiling experiments and disease traits, it is now commonplace to map genes screened in such experiments to known pathways or other annotation sets that can then be tested for enrichment compared to a more complete reference gene set (Mootha, 2003; Lenburg, 2003; Wu, 2002). Forming relationships between different traits as discussed above can be more generally modeled using graphical structures constructed from the experimental data. Such models allow for the efficient identification and representation of relationships among genes and between genes and disease traits. As discussed above, causal inferences can be derived from molecular profiling and disease trait data via QTL data, thereby providing a novel source of information that complement gene expression data and that can be incorporated into methods that seek to identify graphical models (gene networks) of gene interactions.

In order to systematically study complex gene networks in living systems, genes have to be systematically perturbed and the responses to such perturbations recorded. Common systematic perturbation methods include knocking in/out genes, inhibiting gene activity using chemical compounds (e.g., drugs), inhibiting/activating a gene's expression using RNAi technologies (e.g., siRNA, shRNA), and constructing transgenics. Many of these techniques are time consuming and may lack the multifactorial context needed to achieve complex phenotypes of interest, given that many complex phenotypes are the result of multiple genes interacting in complex ways with different genetic backgrounds and environmental conditions, as depicted in Figure 4. Other techniques such as chemical and siRNA inhibition can be accomplished efficiently, but these techniques frequently give rise to off-target effects that cannot be resolved in the absence of additional experimentation (Jackson, 2003), and they cannot always be carried out in the appropriate biological context. For example, typically such experiments need to be carried out using cell-based assays instead of living animals being exposed to the environmental conditions giving rise to the clinical phenotypes of interest.

Naturally occurring genetic variations have the potential to address many of these shortcomings. This type of perturbation has yet to be fully exploited on a genome-wide scale using gene expression or other molecular profiling data, despite it being one of the more comprehensive forms of multifactorial perturbations available in mammalian systems. Zhu *et al*. (2004) recently extended the classic Bayesian network reconstruction methods to incorporate QTL information for the transcript abundances of each gene expression trait considered in a set of experiments of interest. Networks in this setting are composed of nodes, representing genes, and directed edges, representing causal interactions between two genes, as depicted in Figure 5. The causal information used to direct edges in such a network obtain from the genetic information described above. That is, if RNA levels for two genes are genetically controlled by a similar set of genetic loci, then their expression QTL should overlap. In a previously described work (Zhu *et al*., 2004), the extent of QTL overlap was measured by computing the correlation coefficient between vectors of

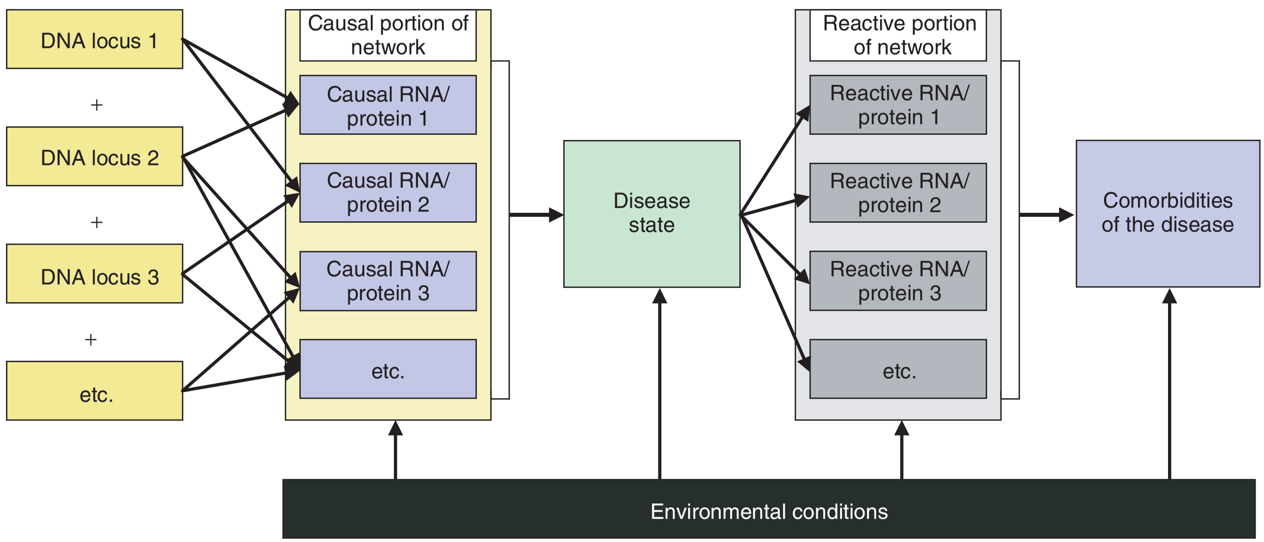


Figure 4 A simple diagram of the molecular causal and reactive events associated with common human diseases such as obesity. The DNA loci represent variations in DNA that in turn lead to changes in RNA or protein. These changes give rise to a larger perturbation in the genetic network that ultimately leads to a given disease. The disease state can in turn cause other changes in the genetic network that lead to comorbidities of the disease (e.g., obesity can increase risk for other diseases such as diabetes and heart disease). In addition to the genetic causes of disease, environmental factors also play a critical role, as common human diseases are the result of complex interactions among multiple genetic loci and between the genetic loci and environmental factors

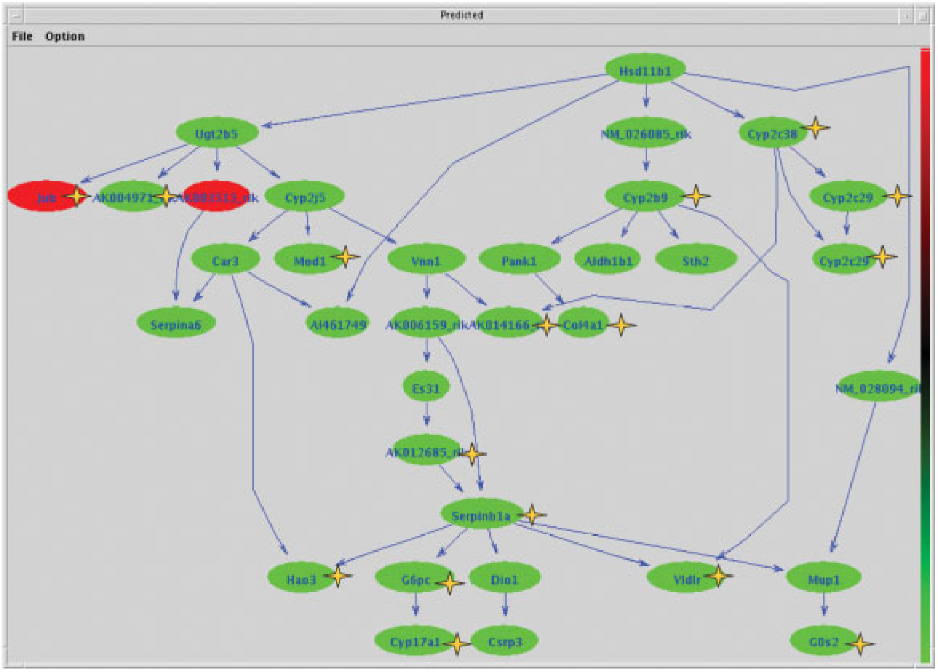


Figure 5 A previously reported (Zhu *et al*., 2004) predicted subnetwork from the liver expression data of the BXD cross described in the text. The subnetwork is centered at the gene HSD11B1 node. The nodes (genes) making up this subnetwork were identified from the larger liver gene network constructed from the BXD data set, by constraining to those nodes having a path to HSD11B1 no longer than three links. Pictured here are the 33 gene nodes meeting this criteria. The stars indicate the 20 genes represented in the Ppara signature previously described (Zhu *et al*., 2004). The gene expression state of each node is colored according to the predicted state when HSD11B1 is in the downregulated state. Red indicates a gene is upregulated relative to the reference pool, green indicates a gene is downregulated relative to the reference pool, and no fill indicates a gene is not differentially expressed relative to the reference pool

LOD scores associated with eQTL identified over entire chromosomes for each gene.

Employing this type of technique to the BXD data set described above, Zhu *et al*. demonstrated that the genetic information allowed for a more predictive network than could be realized using standard, state-of-the-art Bayesian network reconstruction methods, when validated against an independent, single-gene perturbation experiment. The network depicted in Figure 5 from this study gave rise to significant predictive capabilities in that 20 of the 33 genes predicted to change in response to perturbing HSD11B1 indirectly were identified as significantly changed in response to specific drug target experiments in which HSD11B1 activity was perturbed. In this experiment, perturbing HSD11B1 was achieved by activating the peroxisome proliferators activated receptor alpha (Ppara) gene. Only 10 genes would have been expected by chance (the *p*-value for enrichment reported by Zhu *et al*. was as 1.7×10^{-4}). When Zhu *et al*. applied the standard state-of-the-art Bayesian methods to these same data, but without making use of the QTL data, the resulting network predicted only five genes to change in response to downregulating HSD11B1. Only 1 of the 5 genes was detected as significantly

changed in the independent drug target experiment, and given 1 to 2 would have been expected by chance, these results were completely insignificant. That is, the QTL information was necessary to achieve significant predictive capabilities.

The causality arguments given above and in Figure 3 could also be incorporated into the Bayesian network reconstruction methods by altering the conditional mutual information measure that is standard in Bayesian reconstruction methods. The mutual information measure central to Bayesian reconstruction methods could be applied to conditional probabilities that take into account genotypic, molecular profiling, and clinical trait data. While the mutual information measure is useful in more general network reconstruction problems, the advance described here is potentially more powerful given the causal anchoring the QTL data provide, and therefore may potentially lead to a more robust and more powerful test for establishing the relationship between any two traits.

There are numerous advantages gained from the reliable reconstruction of gene networks. Novel genes can be placed in the context of known genes, novel pathways associated with disease traits of interest can be identified, interactions among known pathways and between known and unknown pathways can be characterized, and hypotheses generated to further elucidate disease traits and living organisms more generally. However, beyond these basic life sciences applications there is broad utility in the pharmaceutical setting as well. The networks and causal associations between genes and disease traits will result in the identification of targets for therapeutic intervention. For example, genes such as HMG-CoA reductase that do not have common DNA variations associated with, say, cholesterol levels, could be identified in the broader network as they respond to other genes that do harbor such variations and that in turn cause differential regulation in them. Further, in the network context, the potential exists to prioritize targets on the basis of their association with multiple disease traits, so that targets showing not only causal association with diseases such as obesity, but with diabetes and heart disease as well, can be considered in this more comprehensive light.

6. Conclusion

Classic genetic approaches to dissecting common human diseases have sought to push the limits in applying statistical methods to extract the most information from the association of DNA variations and complex trait variation in human populations and model organism systems. However, in the absence of a broader range of phenotypes, the classic genetic approaches have not fully lived up to their potential in identifying the key drivers of common human diseases. Here we have reviewed a novel approach to the dissection of complex traits that involves the integration of genotypic, molecular profiling, and clinical data. The large-scale molecular phenotyping data is used in this process to identify sets of genes associating with and even defining the molecular processes underlying complex disease traits, and ultimately identifying the key drivers of those processes by associating variation in the molecular profiling and disease traits with variations in DNA.

The types of procedures detailed here begin to form the basis for expanding our understanding of pathways associated with common human diseases and complex systems more generally. The networks that can be reconstructed from the types of procedures reviewed here can themselves serve as a source of annotations for future experiments, given they begin to elucidate the actual pathways underlying complex traits. This work represents only the beginning in an exciting new area of biology where integration of data and the reconstruction of pathways and mapping these pathways to meaningful annotations can, in turn, elucidate the functional roles a given pathway may play in living systems. This type of approach offers a promising avenue of biological research to elucidate the complexity of living systems, given the ability to identify the key drivers of complex traits, order genes into pathways associated with these complex traits, and more completely annotate the functional roles genes underlying the complex traits may play.

Of course, these types of methods are not without problems, and they have limitations that should be recognized and taken into consideration. For example, the network reconstruction procedure discussed above does not model feedback control, a control mechanism that is well known to occur in living systems. Many other methods could be explored for these purposes, including ordinary differential equations, dynamic differential equations, and structural equation models, to name just a few. Further, there are many statistical issues to consider in assessing the significance of the networks resulting from methods that integrate genetic and expression data, including how accurately the topologies of such networks reflect the connectivity of the true underlying pathway and confidence measures for the causal associations identified between any two genes. There are other issues to discuss as well, such as the relevance of gene networks over a diversity of tissues to a particular disease state, and how network connections between tissues will vary as networks spanning multiple tissues are constructed. Recent papers have suggested that significant topological changes in gene networks may obtain in response to changing environmental conditions (Luscombe, 2004), which, if true in mammalian systems, would significantly increase the complexity of the network reconstruction problem. However, even in cases in which the actual recovery of the whole network is impossible, it may be possible using the types of techniques described in Figure 3 to partition the network with respect to a given disease trait into causal, reactive, and independent pieces, as depicted in Figure 6. While this would leave one some distance from actual goal of reconstructing whole networks, it may allow molecular profiling data to be partitioned into causal and reactive sets, thereby facilitating more rapid identification of the key drivers of disease.

However, these problems notwithstanding and the additional investigations needed to enhance confidence in causal inferences made from the integration of genotypic and molecular profiling data, the progressive methods discussed here offer new ways to tackle some of the most pressing problems in the biomedical and life sciences research. The integration of large-scale phenotypic data with genome-wide genotypic data provide the ingredients necessary to inform more fully the molecular processes underlying complex diseases. These techniques have the potential to significantly impact the drug discovery process by increasing the probability of success for identifying targets, minimizing costs of clinical trials by targeting the appropriate disease subtypes, and providing patients with drugs

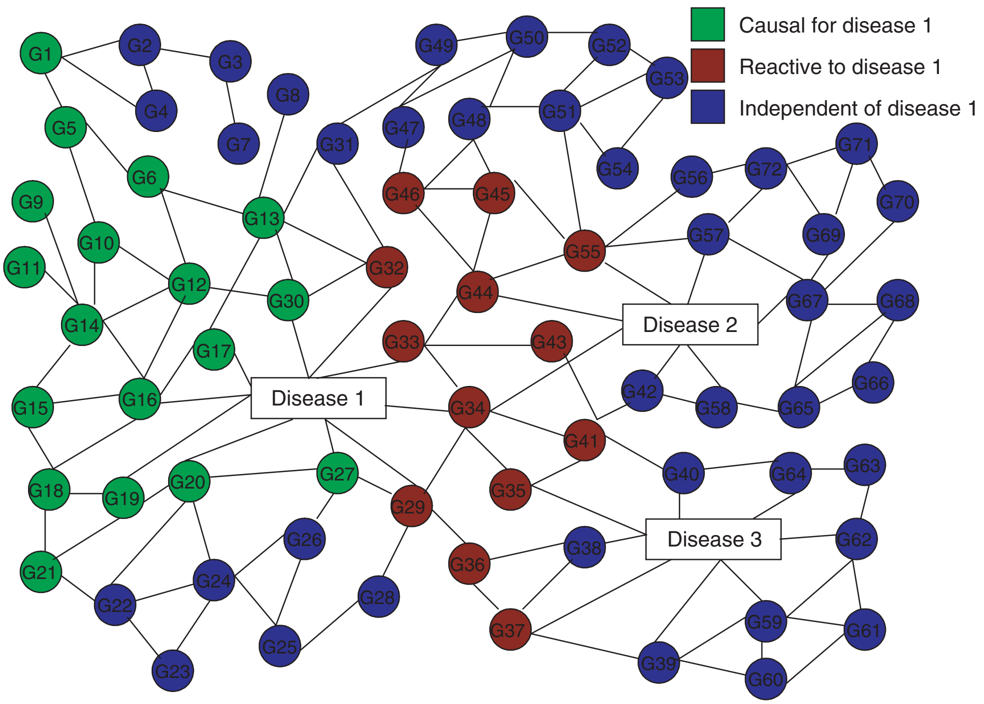


Figure 6 A hypothetical disease-specific genetic network. This simplified view of a genetic network highlights the complex interactions among genes that are associated with common human diseases. While reconstruction of the actual network underlying complex disease traits is still a difficult, if not impossible, problem, partitioning the network with respect to a given disease is possible. Here we illustrate how the nodes in a genetic network can be identified as causal, reactive, or independent of a given disease trait using the type of approach highlighted in Figure 3

that are highly effective for specific diseases that today are only loosely or even incorrectly grouped according to gross clinical symptoms such as obesity.

References

Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**(3), 314–331.

Brem RB, Yvert G, Clinton R and Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**(5568), 752–755.

Cox DR and Hinkley DV (1974) *Theoretical Statistics*, Chapman & Hall: London.

DePrimo SE, Wong LM, Khatry DB, Nicholas SL, Manning WC, Smolich BD, O'Farrell AM and Cherrington JM (2003) Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC Cancer*, **3**(1), 3.

Drake TA, Schadt E, Hannani K, Kabo JM, Krass K, Colinayo V, Greaser LE 3rd, Goldin J and Lusis AJ (2001) Genetic loci determining bone density in mice with diet-induced atherosclerosis. *Physiological Genomics*, **5**(4), 205–215.

Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, Jonsdottir T, Gudmundsdottir T, Bjarnadottir SM, Einarsson OB, Gudjonsdottir HM, *et al*. (2003) The

gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genetics*, **35**(2), 131–138.

He YD, Dai H, Schadt EE, Cavet G, Edwards SW, Stepaniants SB, Duenwald S, Kleinhanz R, Jones AR, Shoemaker DD, *et al*. (2003) Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics*, **19**(8), 956–965.

Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, Gudmundsson G, Grant SF, Thorgeirsson G, *et al*. (2004) The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nature Genetics*, **36**(3), 233–239.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, **19**(4), 342–347.

Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G and Linsley PS (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, **21**(6), 635–637.

Jansen RC and Nap JP (2001) Genetical genomics: the added value from segregation. *Trends in Genetics TIG*, **17**(7), 388–391.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G and Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics*, **29**(4), 389–395.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R and Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**(5653), 2141–2144.

Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S, *et al*. (2000) Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nature Immunology*, **1**(3), 221–226.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**(5594), 799–804.

Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT and Christman MF (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**(1), 31.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA and Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**(7006), 308–312.

Miklos GL and Maleszka R (2004) Microarray reality checks in the context of a complex disease. *Nature Biotechnology*, **22**(5), 615–621.

Monks, SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs AB and Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics*, **75**(6), 1094–1105.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**(3), 267–273.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS and Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**(7001), 743–747.

Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, *et al*. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, **63**(7), 1602–1607.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, Duan XJ, Soro-Paavonen A, Naukkarinen J, Saarela J, *et al*. (2004) Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nature Genetics*, **36**(4), 371–376.

Pearl J (1983) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers: San Francisco.

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, *et al*. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**(6870), 436–442.

Sabo PJ, Humbert R, Hawrylycz M, Wallace JC, Dorschner MO, McArthur M and Stamatoyannopoulos JA (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(13), 4537–4542.

Schadt EE (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology*, **5**, R73.

Schadt EE, Lamb J, Zhu J, Edwards S, Reitman M, GuhaThakurta D, Monks SA, Lum PY, Drake TA, Lusis AJ, *et al*. (2005) An Integrative Genomics Approach to Infer Causal Associations between Gene Expression and Disease. *Nature Genetics*, in press.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, *et al*. (2003a) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**(6929), 297–302.

Schadt EE, Monks SA and Friend SH (2003b) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochemical Society Transactions*, **31**(2), 437–443.

Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G, *et al*. (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**(6822), 922–997.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cells*, **1**(2), 203–209.

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, *et al*. (2002) Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics*, **71**(4), 877–892.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, **347**(25), 1999–2009.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–556.

Waring JF, Gum R, Morfitt D, Jolly RA, Ciurlionis R, Heindel M, Gallenberg L, Buratto B and Ulrich RG (2002) Identifying toxic mechanisms using DNA microarrays: evidence that an experimental inhibitor of cell adhesion molecule expression signals through the aryl hydrocarbon nuclear receptor. *Toxicology*, **181–182**, 537–550.

Wayne ML and McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(23), 14903–14906.

Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R and Altschuler SJ, *et al*. (2002) Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, **31**(3), 255–265.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R and Kruglyak L (2003) Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, **35**(1), 57–64.

Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, *et al*. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and Genome Research*, **105**(2–4), 363–374.

Ziv E and Burchard EG (2003) Human population structure and genetic association studies. *Pharmacogenomics*, **4**(4), 431–441.

Specialist Review

The promise of gene signatures in cancer diagnosis and prognosis

Jimmy C. Sung, Alice Y. Lee and Timothy J. Yeatman
The H. Lee Moffitt Cancer Center and Research Institute at the University of South Florida, Tampa, FL, USA

1. Introduction

Precise tumor diagnosis is the first step in cancer management since therapy generally stems from the initial tumor classification (*see* Article 65, **Complexity of cancer as a genetic disease**, Volume 2). While many tumor biopsies are diagnostic and form the cornerstone of cancer therapy, classification of tumor type and site of origin is a significant clinical challenge that is often underestimated. In fact, up to 10% of all metastatic tumors have no definable primary site of origin (Blaszyk *et al*., 2003). The current standard of pathologic practice, using morphologic criteria and *semiquantitative* immunohistochemical (IHC) analyses, is often limited in its capacity to define tumor type or site of origin. Tumors with seemingly identical pathology under the microscope may follow different disease trajectories because of underlying molecular heterogeneity (Golub *et al*., 1999). This diagnostic imprecision, on the whole, may adversely affect the quality of patient care while unnecessarily increasing health care expenditure. Moreover, our inability to predict the future biological behavior of tumors is problematic, leading to the overtreatment of many patients to help an unknown few.

With the mapping of the human genome (*see* Article 24, **The Human Genome Project**, Volume 3), DNA microarray technology, coupled with large-scale methods of gene-expression analysis, has the power to revolutionize cancer diagnosis and treatment. By studying the activity of many genes simultaneously – an approach known as *molecular profiling* (Liotta and Petricoin, 2000) – a genetic portrait of an individual cancer can be generated, providing investigators with greater insight into the cancer's etiology, its responsiveness to therapy, its potential to metastasize, and the probability of patient survival. Molecular profiling holds promise to overcome the limitations of traditional cancer diagnosis with the potential for more accurate cancer classification, improved treatment stratification, and the discovery of new therapeutic targets (Alizadeh *et al*., 2000).

1.1. Molecular signatures for cancer diagnosis

A major challenge of cancer research has been to direct specific treatments to distinct tumor types so that efficacy is maximized while toxicity is minimized (Golub

et al., 1999). Thus, advances in cancer classification are critical to advancing the quality of care available to cancer patients. Historically, classification has relied on subjective interpretations of the morphological appearance of the tumor. Molecular profiling using microarray technology (*see* Article 90, **Microarrays: an overview**, Volume 4) has great potential as a systematic and unbiased approach for assigning particular tumor samples to previously defined classes, as well as discovering new tumor subtypes. Golub *et al*. (1999) were among the first to demonstrate that patterns in gene-expression data could be used to distinguish among different tumor types. On the basis of gene-expression profiling alone, it was possible to distinguish tumor samples of acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL). A gene predictor derived from an analysis of microarray data using 38 known acute leukemia samples was capable of accurately classifying 29 of 34 new independent leukemia samples. The gene-expression signature used to distinguish AML from ALL included genes encoding for cell surface proteins (CD11c, CD33, MB-1), those involved in cell cycle progression (Cyclin D3, Op18, MCM3), and known oncogenes (c-MYB, E2A, HOXA9). Next, turning to the question of class discovery, a technique based on self-organizing maps that automatically discovered the distinction between AML and ALL, as well as the distinction between B-cell and T-cell ALL was used. Gene-expression monitoring was shown to provide new and important insights into tumor pathogenesis, as well as future clinical outcome.

Using 218 samples spanning 14 common cancer types, Ramaswamy *et al*. (2001) constructed a classifier based on a support vector machine algorithm. Overall classification accuracy was 78%, exceeding the accuracy of random classification (9%), suggesting the feasibility of cancer diagnosis across the common malignancies based on a comprehensive catalog of gene-expression profiles. This initial work was then extended by Bloom *et al*. (2004) who demonstrated that large public datasets could be combined, normalized, and scaled to produce a multitumor classifier capable of discriminating 21 different tumor types with high degrees of diagnostic accuracy. To accomplish this, sizable gene sets ($n > 500$) were used in combination with a neural network approach to weight the genes in the classifier (Bloom *et al*., 2004). Figure 1 demonstrates the large number of discriminating genes that can be identified in each defined tumor type and subsequently exploited to develop a multitumor classifier.

1.2. Gene signatures for prognosis of lymphoma

The correlation between gene-expression patterns and clinical outcome was first demonstrated by Alizadeh *et al*. (2000) in a cDNA microarray-based study of diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma. Forty percent of patients responded favorably to standard therapy achieving a durable remission, while the remainder succumbed to the disease. This variability in clinical outcome was shown to reflect molecular heterogeneity within the tumors. Hierarchical clustering identified two molecularly distinct types of DLBCL with gene-expression patterns indicative of different stages of B-cell differentiation: one type had a gene-expression pattern characteristic of germinal center B cells, while the second type expressed genes normally induced

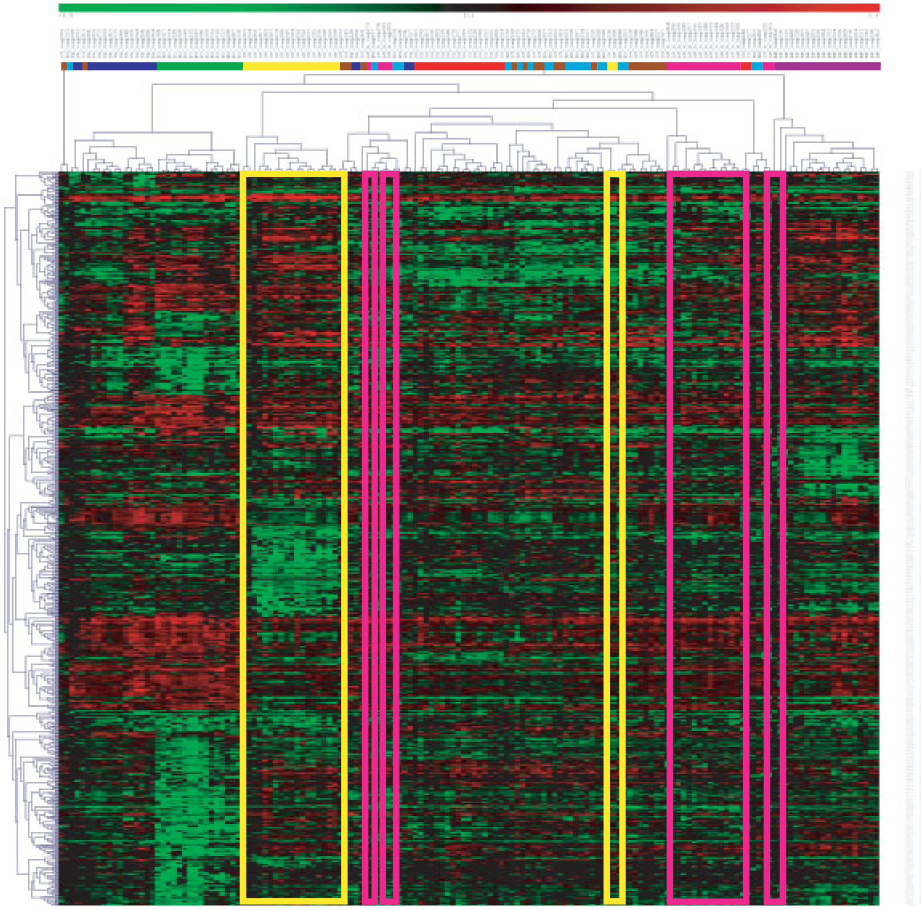


Figure 1 Hierarchical clustering of eight different types of adenocarcinomas. The Kruskal–Walls H-test was used to identify those genes most correlated with each tumor type, selecting ∼700 genes from 30 849 distinct transcripts on the cDNA chip. Average linkage hierarchical clustering of spotted cDNA array expression data using a Pearson correlation coefficient distance matrix illustrates the problems with this approach to classification, which typically weights each gene equally. Even for ovartan cancer samples (yellow boxes), which are generally well classified, there are two cutlying samples that are grouped within a set of diverse tumors. For other tissues of origin such as lung (pink boxes), the situation is worse. Similar results are obtained for samples assayed using Affymetrix GeneChips. Although hierarchical clustering can be used with weights for each gene, we have no a priori means of determining the appropriate weights. This is the rationale that underlies the use of the ANN in tumor classification. (Reproduced with permission from Bloom *et al*. (2004) Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology*, **164**(1), 9–16. © The American Society for Investigative Pathology)

during in vitro activation of peripheral blood B cells. Patients with the germinal center B-like DLBCL showed a significantly higher posttreatment survival rate (76% versus 16%) than patients with the peripheral blood B type, demonstrating that gene expression-based tumor classification can identify clinically significant subtypes of cancer. In fact, the clinical divergence between the two types was so remarkable that it was suggested that the two subtypes be regarded as distinct diseases. Alizadeh *et al*. (2000) anticipated that such large-scale microarray data

analyses would identify a small group of marker genes that can be used to stratify patients into molecularly distinct categories, thus improving the power and precision of clinical therapies. For example, by determining which patients are unlikely to respond to standard chemotherapy, more aggressive therapies, such as bone marrow transplantation, could be recommended early on. Additionally, as the two classes of DLBCL differentially expressed entire transcriptional modules consisting of hundreds of genes, new drugs might be developed to target the upstream signal-transducing molecules that are responsible for the expression of pathological transcriptional programs.

More recently, Dave *et al*. (2004) used gene-expression profiling to construct a molecular predictor of survival length for the second most common form of non-Hodgkin's lymphoma – follicular lymphoma. The gene-expression signatures used to construct the survival predictor, which included genes that encode T-cell markers (e.g., CD7 and STAT4) and genes highly expressed in macrophages (e.g., ACTN1 and TNFSF13B), allowed patients to be divided into four quartiles with widely disparate median lengths of survival (ranging from 13.6 to 3.9 years), independently of any clinical prognostic variables. Surprisingly, these signatures were derived from nonmalignant cells in the tumors, revealing the important role of the host immune system in the development of this type of malignancy and pointing the way to possible new targets for therapy.

1.3. Gene signatures for prognosis of solid malignancies

Not only has molecular profiling revealed subtypes and expression signatures that could improve prognostic classification in blood cancers (leukemia and lymphoma), but high-throughput gene profiling has shown great promise in the study and treatment of solid tumors as well. Using cluster analysis on cDNA microarray data representing 8201 genes, Perou *et al*. (2000) characterized variation in gene-expression patterns in set of 65 specimens of breast tumors from 42 different patients. These gene-expression patterns provided a distinctive molecular portrait of each tumor, with patterns in two samples from the same individual almost always being more similar to each other than either was to any other sample. Moreover, variations in growth rate, in the activity of certain signaling pathways, and in the cellular composition of the tumors, were all reflected in the corresponding variation in the expression of specific subsets of genes. For example, the expression of a large cluster of genes regulated by the interferon pathway (including STAT1) varied widely among the tumor samples. Dendogram diagramming also divided the samples into two major subdivisions – an estrogen-receptor positive group, characterized by the relatively high expression of many genes expressed by breast luminal cells, and an estrogen-receptor negative group. It was discovered that the estrogen-positive group could be further subdivided into two groups, each with a distinctive expression profile and significantly different survival outcomes (Sorlie *et al*., 2001). Interestingly, Perou *et al*. (2000) discovered that a metastasis and primary tumor were as similar in their overall pattern of gene expression as were repeated samplings of the same primary tumor, suggesting that the molecular program of a primary tumor may generally be retained in its metastases.

Further, gene expression–based research into the clinical behavior of breast cancer was performed by van't Veer *et al*. (2002). Using supervised clustering on microarray data taken from 117 breast cancer patients, van't Veer *et al*. identified a gene-expression signature predictive of a short interval to distant metastases ("poor prognosis" signature) in patients without tumor cells in local lymph nodes at diagnosis. A prognosis classifier consisting of 70 genes predicted to 83% accuracy the actual outcome of disease. Genes significantly overexpressed in the poor prognosis signature included those involved in cell cycle, invasion and metastasis, angiogenesis, and signal transduction (e.g., cyclin E2, MCM6, MMp9, PK428, and the VEGF receptor FLT1) and represented potential targets for the rational development of new cancer drugs. This gene-expression profile outperformed all currently used clinical parameters in predicting disease outcome, thus providing an improved strategy for selecting patients who would benefit from adjuvant therapy.

Bhattacharjee *et al*. (2001) have proposed a gene expression-based taxonomy of lung cancer, the leading cause of cancer death worldwide. Using oligonucleotide microarrays (*see* Article 92, **Using oligonucleotide arrays**, Volume 4), mRNA expression levels in 186 lung tumor samples were analyzed, including 139 adenocarcinomas resected from the lung. Hierarchical and probabilistic clustering of expression data defined distinct subclasses of lung adenocarcinoma, including one cluster most likely representing metastatic adenocarcinomas from the colon. One subclass, characterized by several neuroendocrine markers such as dopa decarboxylase and achaete-scute homolog1, showed a significantly less favorable survival outcome than all the other groups. Another subclass, characterized by the highest expression of type II alveolar pneumocyte markers, was associated with a more favorable clinical outcome. Interestingly, two adenocarcinoma subclasses were associated with lower tobacco-smoking histories. More recently, Tomida *et al*. (2004) created a prediction classifier based on the expression profiling of 50 patients with non-small-cell lung cancer, which comprise 80–85% of all lung cancers. The resultant classifier yielded 82% accuracy for forecasting survival or death 5 years after surgery. Furthermore, unsupervised hierarchical clustering analysis revealed for the first time the existence of clinicopathologically relevant subclasses of squamous cell carcinomas with marked differences in invasive growth and prognosis. Overall, the discovery of such distinct subtypes of lung carcinomas suggests that the integration of gene-expression profile data with clinical parameters could significantly aid in the effective diagnosis and individualized treatment of lung cancer patients.

A previously unrecognized taxonomy based on molecular profiling has also been proposed by Bittner *et al*. (2000) for the most common human cancer – melanoma. Currently, no well-established histopathological, molecular, or immunohistochemical marker defines subclasses of this neoplasm. Through cluster analysis, a distinct subset of melanomas was discovered whose characteristic genes were differentially regulated in invasive melanomas that form primitive tubular networks in vitro, a feature of some of the most highly aggressive melanomas. Notably, no statistically significant association was found between the cluster groups and any clinical or histological variable, such as age, sex, biopsy site, or in vitro pigmentation. In their analysis, Bittner *et al*. generated a ranked list of individual genes (e.g., WNT5A,

MART-1, pirin, HDHB) with the most power to define the clusters of the 19 samples studied, providing a sound molecular basis for the dissection of other clinically relevant melanoma subtypes and for greater understanding of the more aggressive forms of this cancer.

Another study (Ramaswamy *et al*., 2003) using expression profiling has challenged the prevailing model of metastasis, which holds that most primary tumor cells have low metastatic potential, while rare cells (less than 1 in 10 million) within primary tumors acquire metastatic capacity through mutation. A gene-expression signature was discovered that distinguished primary from metastatic adenocarcinomas. Solid tumors carrying this signature were most likely to be associated with metastasis and poor clinical outcome, suggesting that the metastatic potential of human tumors is encoded in the bulk of a primary tumor, rather than in just a small minority of the cells. The findings support the existence of a molecular program of metastasis that is shared by multiple types of solid tumors, suggesting the possibility of common therapeutic targets shared by different cancers.

Other solid tumors for which gene-expression profiling has proven a valuable tool for pattern-based class discovery and prognostication include gliomas (the most common malignant primary brain tumor) (Freije *et al*., 2004), kidney cancer (Vasselli *et al*., 2003), parathyroid tumors (Haven *et al*., 2004), leiomyosarcoma (sarcoma of the smooth muscle) (Lee *et al*., 2004), and prostate cancer (Lapointe *et al*., 2004), and more recently colorectal cancer (Eschrich *et al*., 2005). A gene expression–based classifier for gliomas was found to be a more powerful survival predictor than histologic grade or age (Freije *et al*., 2004). This classifier was validated with an additional external and independent data set from another institution, indicating that gene expression–based predictors are robust and can be applied across multiple institutions and platforms.

Vascular cell adhesion molecule-1 (VCAM-1) was found to be the gene most predictive in patients with kidney cancer, whose clinical course can be highly variable, with many patients dying within 1 year of diagnosis and others living for years with slowly progressive disease (Vasselli *et al*., 2003). Survival for patients with metastatic renal cancer was correlated with the expression of various genes based solely on the molecular profile of the primary kidney tumor.

Although parathyroid tumors are heterogeneous, histological differences are subtle, thus confounding classification. A class discovery approach identified three broad cluster groupings with possibly different molecular pathways of tumorigenesis (Haven *et al*., 2004). Similarly, a gene-expression signature, containing many genes previously reported to be associated with metastasis in a variety of cancers, was found to be predictive of metastatic outcome in leiomyosarcomas (Lee *et al*., 2004). And, for prostate cancer, three previously unknown subclasses were discovered with high-grade and advanced stage tumors disproportionately represented among two of the subclasses. Expression levels of MUC1, highly expressed in the more aggressive tumor subgroups, and AZGP1, highly expressed in the nonadvanced subgroup, were found to be strong predictors of tumor recurrence independent of tumor grade, stage, and preoperative prostate-specific antigen levels. Finally, for colorectal cancer, Eschrich *et al*. (2005) have identified a profile of genes with the potential to outperform Dukes' staging in predicting outcome for this disease. Of interest was that their classifier gene set was tested in an

independent Danish subpopulation with favorable results. Of the markers identified, osteopontin played a dominant role, and had been previously identified by the same laboratory using a pooled RNA approach with a different microarray platform and a different clinical endpoint (stage versus overall survival) (Agrawal *et al.*, 2002). The discovery of such differentially expressed genes as well as distinct molecular subtypes has great potential not only in uncovering new leads to understanding the mechanisms of cancer progression but also toward tailoring powerful and precise treatments to the individual patient.

2. Conclusions

The new insights into the complexity of cancer that are currently being generated by gene-profiling technology are certainly groundbreaking and exciting. Nonetheless, the true promise of molecular profiling lies in the translation of the technology into cost-effective clinical use. Many diagnostics and prediction assays are being developed at this time; however, as of today, none has undergone the scrutiny of prospective randomized trials. Once the scientific merits of gene expression–based therapy are firmly established, rational therapy may be designed on the basis of sound diagnosis, accurate prediction of response, and the likelihood of beneficial outcome. The clinical advantage and economic benefits of such advancement in cancer treatment will be enormous for the most important elements in this endeavor – our patients.

Related articles

Article 65, **Complexity of cancer as a genetic disease**, Volume 2; Article 24, **The Human Genome Project**, Volume 3; Article 90, **Microarrays: an overview**, Volume 4; Article 92, **Using oligonucleotide arrays**, Volume 4

Acknowledgments

We thank Lindsay Rodzwicz for her editorial assistance in preparing this manuscript. Supported by NCI Grants: R21-CA101355, R01-CA098522, K24-CA85429 and U01-CA85052.

References

Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D and Yeatman TJ (2002) Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *Journal of the National Cancer Institute*, **94**(7), 513–521.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–511.

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(24), 13790–13795.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**(6795), 536–540.

Blaszyk H, Hartmann A and Bjornsson J (2003) Cancer of unknown primary: clinicopathologic correlations. *APMIS: Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, **111**(12), 1089–1094.

Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J and Yeatman TJ (2004) Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology*, **164**(1), 9–16.

Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, Fisher RI, Braziel RM, Rimsza LM, Grogan TM, *et al*. (2004) Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, **351**(21), 2159–2169.

Eschrich S, Yang I, Bloom G, Kwong KY, Boulware D, Cantor A, Coppola D, Kruhoffer M, Aaltonen L, Orntoft TF, *et al*. (2005) Molecular staging for survival prediction of colorectal cancer patients. *Journal of Clinical Oncology*, in press.

Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, Mischel PS and Nelson SF (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, **64**(18), 6503–6510.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Haven CJ, Howell VM, Eilers PH, Dunne R, Takahashi M, van Puijenbroek M, Furge K, Kievit J, Tan MH, Fleuren GJ, *et al*. (2004) Gene expression of parathyroid tumors: molecular subclassification and identification of the potential malignant phenotype. *Cancer Research*, **64**(20), 7405–7411.

Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(3), 811–816.

Lee YF, John M, Falconer A, Edwards S, Clark J, Flohr P, Roe T, Wang R, Shipley J, Grimer RJ, *et al*. (2004) A gene expression signature associated with metastatic outcome in human leiomyosarcomas. *Cancer Research*, **64**(20), 7201–7204.

Liotta L and Petricoin E (2000) Molecular profiling of human cancer. *Nature Reviews Genetics*, **1**(1), 48–56.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–752.

Ramaswamy S, Ross KN, Lander ES and Golub TR (2003) A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, **33**(1), 49–54.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(26), 15149–15154.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(19), 10869–10874.

Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, Some M, Yanagisawa K, Takahashi T and Osada H (2004) Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene*, **23**(31), 5360–5370.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, Torres-Cabala C, Tabios R, Mariotti A, Stearman R, *et al*. (2003) Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(12), 6958–6963.

Short Specialist Review

Seven years of yeast microarray analysis

Gavin Sherlock
Stanford University, Stanford, CA, USA

Paul T. Spellman
Lawrence Berkeley Laboratory, Berkeley, CA, USA

1. Introduction

One of the specific goals of the Human Genome Project (*see* Article 24, **The Human Genome Project**, Volume 3), defined by the National Research Council study in 1988, was to sequence the genomes of a number of model organisms, with the expectation that an understanding of the genomes of model organisms would guide us in our investigations of the human genome. *Saccharomyces cerevisiae* was the first eukaryote to be fully sequenced, which revealed approximately 6000 simply constructed genes. At the same time, DNA microarrays were invented at both Stanford University and Affymetrix (*see* Article 91, **Creating and hybridizing spotted DNA arrays**, Volume 4 and Article 92, **Using oligonucleotide arrays**, Volume 4), yielding the ability to simultaneously quantify the abundance of thousands of nucleic acid species in complex mixtures. Easy genetic manipulation, compact gene structures and a complete genome sequence made yeast the center of genomics, and microarrays provided an unprecedented opportunity to assay an entire biological system. This short review examines the development of microarray technology, and the variety of innovative uses to which it has been put in furthering our understanding of yeast biology, illustrating how yeast was, and continues to be, at the forefront of genomic research (*see* Article 43, **Functional genomics in *Saccharomyces cerevisiae***, Volume 3).

2. Large-scale studies of gene expression

In 1997, Joe DeRisi and colleagues, from Pat Brown's laboratory at Stanford, published a landmark microarray study (DeRisi *et al.*, 1997). On the basis of the recently published genome sequence, the Brown and Botstein labs PCR-amplified 6400 predicted open reading frames from yeast, and spotted each PCR product to create the first whole-genome microarray in an 18 mm × 18 mm square.

DeRisi, Iyer, and Brown investigated the diauxic shift, which is the point at which yeast switch to ethanol as a carbon source, after having exhausted the available glucose. Seven samples taken every 2 h resulted in more than 43 000 expression measurements, their analysis of which showed that changes in gene expression during the diauxic shift result in metabolic reprogramming consistent with the environmental changes that were occurring.

The following year, two groups analyzed the transcriptional program of the yeast mitotic cell cycle, using high-density oligonucleotide arrays from Affymetrix (Cho *et al*., 1998), or homemade microarrays with spotted PCR products (Spellman *et al*., 1998, which includes the authors of this chapter). Together, they used four different methods of whole cell culture synchronization to determine the periodically transcribed genes through the cell cycle. While the two studies were not in exact agreement as to the number and identity of those transcripts, there was a large overlap in their findings. At that time, our cell cycle dataset was (and continued to be for some time) the largest single dataset published with microarray experiments from 62 unique arrays. It should be noted that both the Davis and Brown/Botstein groups made their data available on the Internet and that in our cell cycle study we actually reinterpreted the Davis lab data, which was the first time a dataset was reanalyzed. This marked the beginning of the age of microarray data reanalysis, with our study (and a number of the following studies) showing that complex reanalysis of yeast gene expression data was possible.

3. Yeast gene deletions and DNA microarrays

Concurrent to the efforts to characterize global gene expression profiles was an effort to generate systematic deletion mutants, a set of 24 000 yeast strains available as homo- and heterozygous diploids and as haploids of each mating type. This project was both driven by and drove microarray technology because Ron Davis' lab, in conjunction with Affymetrix, had developed a system of using a molecular barcode for each deletion, which could be detected by hybridization. In the proof of principle experiment, Shoemaker *et al*. (1996) demonstrated that these bar-coded yeast strains could be grown competitively under a given condition and the relative abundance of each strain could be assayed using microarrays to assess if deletion of a given gene resulted in a measurable quantitative trait.

Winzeler *et al*. (1999b) extended this approach by assaying complex mixtures of more than 500 homozygous bar-coded deletion mutants, under growth in either rich or minimal medium, to determine whether the deletion of these nonessential genes affected the fitness of the strains. While generation of the mutants initially was an enormous effort, the ability to rapidly assay their phenotypes in competitive mixtures was relatively simple to implement. Thus, after generation of the initial reagents, it is relatively easy to assay a large number of growth conditions to uncover the functional roles of previously uncharacterized genes.

A recently developed method, Synthetic Lethality Analyzed by Microarray (SLAM; Ooi *et al*., 2003), has taken this concept one step further. Instead of assaying the fitness of individual deletion mutants, a pool of mutants is transformed with a deletion construct for a gene of interest. Synthetic interactions are detected

using microarrays to determine if a deletion mutant in the original pool has been depleted. The advantage of this method over traditional screens of this nature, in addition to microarray identification of candidates, is that it is quantitative, and candidates can be ranked. Theoretically, it could also be used to find deletion mutants that enhance the ability of another deletant to grow.

4. Assaying drug modes of action

At the same time as the first whole genome expression studies were being carried out to characterize biological pathways in *S. cerevisiae*, Marton *et al*. (1998) demonstrated that expression signatures from deletion mutants treated with different drugs could be used to identify candidate drug targets. Their logic was that the typical drug-induced signature would be largely absent if the cell were deleted for the protein that was the target of the drug. Additionally, they characterized "off-target" effects, which remained even in the absence of the target, thus elucidating drug specificity. This result established that drug action could be systematically studied in yeast, which contains many genes for which there exist homologs in human cells.

Hughes *et al*. (2000) built on this idea and created a compendium of yeast gene expression data, generated both from deletion mutants and from cell treated with compounds with known molecular targets. Their goal was to construct a reference database of gene expression profiles, such that interpreting expression profiles of new mutants or cell treated with drugs of unknown targets could be done in the context of the preexisting data. If treatment with a drug resulted in a similar expression *signature* as resulted from deletion or mutation of a particular gene, then that drug may target that gene, either directly or indirectly. Hughes *et al*., as had Eisen *et al*. (1998) before them, were among the first to realize the cumulative value of microarray data and that different microarray experiments are not islands but instead each add to the value of the others, such that the sum of biology that can be derived from a global set is greater than that which can be derived from any single set alone.

5. Mapping traits using microarrays

In an illustration of the diverse applications of microarrays, Winzeler *et al*. (1998) used high-density oligonucleotide arrays to scan for allelic variation in the genome, on the basis of the premise that polymorphisms in the genomic DNA of an yeast strain would cause decreased hybridization to oligonucleotides that were designed on the basis of the sequenced strain of *S. cerevisiae*, S288 C. They were able to identify markers spaced roughly every 3.5 kilobases throughout the genome of a strain unrelated to S288 C. Using these markers, they were able to map the phenotype of cyclohexamide sensitivity in one of the parental strains unambiguously to the Pdr5 multidrug resistance pump. Steinmetz *et al*. (2002) extended this technique by identifying the causative alleles of quantitative trait loci (QTLs) responsible for a clinically important phenotypic trait in yeast, the ability to

grow at high temperature. In an even more far-reaching study, Brem *et al*. (2002) used the same technique as Winzeler *et al*. (1999a) to track genetic markers, but instead of scoring simple measurable traits in plate or growth assays, they measured considered genome-wide transcription as the phenotype. They used whole-genome microarrays to measure the expression of the two parental strains and 40 haploid segregants and were able to link the expression levels of 570 genes to one or more loci. Furthermore, they identified eight loci that appeared to encode trans-acting factors responsible for some of the expression variation that they observed. Thus, microarrays can be used to track both genome-wide genotypes and genome-wide phenotype, and identify causative linkages between them.

6. Systems biology

While there are many applications of microarrays that can be used to probe the state or contents of a cell or biological system, microarray data can be greatly enhanced by the addition of other systematic data. In an attempt to characterize the network of interactions that control galactose metabolism, Ideker *et al*. (2001) systematically perturbed the galactose metabolic network, both genetically and environmentally. They used microarrays to assay the gene expression changes, and tandem mass spectrometry of tagged proteins to assay changes in protein abundance for several hundred proteins. These data were then combined with preexisting 2-hybrid interaction data and protein–DNA interaction data from transcription factor databases to build a model of galactose metabolism within the cell. The approach of combining microarray data (potentially of different types) with other functional genomics data is likely the way by which we will elucidate how a cell works as a system.

7. Assessing the state and content of the genome

Raghuraman *et al*. (2001) used microarrays to define Origins of Replication for the whole genome using microarrays. Using the classic, Heavy/Light isotope trick, they purified EcoRI fragmented DNA whose ratio of Heavy to Light isotopes depended on proximity to an origin. Comparing Heavy–Heavy and Heavy–Light fractions by hybridization to a microarray identified regions of the genome that acted as ARS and determined when they fired in the cell cycle as well as fork migration rates.

8. Identification of protein–DNA binding sites

One of the most important advances in moving toward a full systems biology approach is the use of DNA microarrays to determine the in vivo binding sites for transcription factors using Chromatin immunoprecipitation (ChIP). Rick Young's group at MIT was the first group to microarray-based ChIP data (Ren *et al*., 2000), characterizing genes whose regulatory regions are bound by Ste12 and Gal4.

They followed up this result with a truly impressive body of work, systematically studying 106 yeast transcription factors to determine their in vivo specificities (Lee *et al*., 2002). In this one paper, they were able to characterize nearly all of the sequence-specific transcription factors and were able to identify network motifs, such as autoregulation, multicomponent loops, and feedforward loops. These data were used, in conjunction with other existing expression data, to produce network models for various cellular machinery, such as the cell cycle.

9. Protein chips

While the majority of microarray experiments have used DNA fragments on the microarray, either oligonucleotides, PCR products, or cDNA clones, protein microarrays have also been developed. Zhu *et al*. (2001) purified proteins encoded by 5800 different yeast ORFs, and then spotted the purified proteins onto a glass substrate to create a protein microarray. The microarray was then used to screen for proteins that interacted with calmodulin or phospholipids. This approach allowed them to identify a putative calmodulin-binding site that was present in many of the proteins that interacted with calmodulin. Again, while the initial reagents are time-consuming to prepare, once they are in hand, they can rapidly be used to assay various properties of proteins.

10. Summary and future

Since their development in 1996, microarrays have become a powerful tool in the biologists' arsenal. While they are most often thought of in connection with microarray expression studies, several landmark studies have demonstrated a wide variety of applications for microarrays, the majority of which were carried out in yeast. Using microarrays, researchers are able to assay almost every aspect of a living cell to characterize it, from the parental origin of genes, to their copy number, to the replicative state of the genome, to many aspects of the transcriptome and its properties, to the proteome itself (Figure 1). Owing to space constraints, we have limited the scope of our review but we would like to point out that microarrays have been used to study RNA stability and decay, RNA splicing, transcript length, RNA localization, and even the kinetics of protein translation (Holstege *et al*., 1998; Wang *et al*., 2002; Grigull *et al*., 2004; Clark *et al*., 2002; Hurowitz and Brown, 2003; Marc *et al*., 2002; Arava *et al*., 2003). Furthermore, microarrays can be used as a screening technology to identify the effect upon the fitness of a cell in the presence of various deletions or mutations.

Great strides have been made in understanding several fundamental aspects of yeast cell biology using microarrays since 1996, yet by no means can we say that yeast is "solved" as an organism. First, and most surprisingly, we do not yet know with certainty the number and location of all yeast genes nor do we know the set of transcripts produced from these genes. Second, we do not know the functions of almost one-third of the genes that are identified. Tiling arrays, in which nearly the entire yeast genome sequence can be interrogated, will enable us to elucidate all the transcripts encoded by the genome and accurately

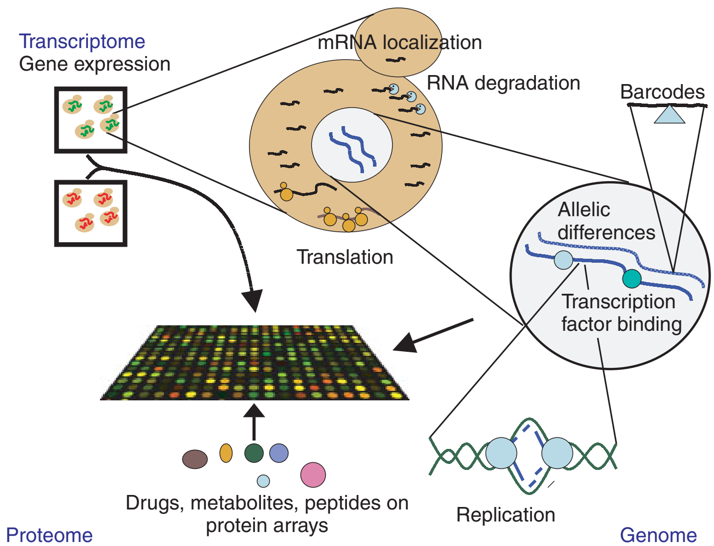


Figure 1 Illustration of an yeast cell to show how microarrays can be applied to interrogate all aspects of the central dogma of molecular biology. In the central dogma, DNA (the genome) is used as a template to make RNA (the transcriptome), which in turn is translated into protein (the proteome). Although microarrays are often thought of in connection with assaying transcript abundance, they can be used to assay various properties of each of the aspects of the central dogma, as shown

determine their 5′ and 3′ ends. It is unclear at this point how many additional, as yet unrecognized genes exist in the genome, and how many of those currently marked as dubious by the community database, SGD (Christie *et al*., 2004) really have no function, or do not produce a transcript. Comprehensive understanding of yeast biology on an organismal scale requires additional microarray experiments, using more comprehensive microarrays, coupled with the integration of additional nonmicroarray data types, such as 2-hybrid and other protein–protein interaction data. In addition, it will be necessary to integrate these results with a comprehensive analysis of the yeast metabolome. Once we understand, through experimentation, how all the components in the central dogma fit together (the genes in the genome, when and what transcripts are produced, and the functions of the encoded products) and how the metabolites and their abundance and fluxes change during various physiological programs, we will understand, to a first approximation, how yeast works on an organismal level. Application and extension of the fundamental principles that have been learned from such studies will be instrumental in understanding how more complex, multicellular organisms work.

Further reading

Cho RJ, Fromont-Racine M, Wodicka L, Feierbach B, Stearns T, Legrain P, Lockhart DJ and Davis RW (1998) Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3752–3757.

Diehn M, Eisen MB, Botstein D and Brown PO (2000) Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genetics*, **25**, 58–62.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F and Botstein D (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16144–16149.

Ferea TL, Botstein D, Brown PO and Rosenzweig RF (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9721–9726.

Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO and Petes TD (2000) Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 11383–11390.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A and Davis RW (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics*, **21**, 278–283.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, **19**, 342–347.

Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, *et al*. (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genetics*, **25**, 333–337.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M and Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO and Davis RW (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 13057–13062.

Nagy PL, Cleary ML, Brown PO and Lieb JD (2003) Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6364–6369.

Ooi SL, Shoemaker DD and Boeke JD (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science*, **294**, 2552–2556.

Shalon D, Smith SJ and Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, **6**, 639–645.

Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG, Young RA, Bell SP and Aparicio OM (2001) Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science*, **294**, 2357–2360.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R and Kruglyak L (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genetics*, **35**, 57–64.

Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA and Snyder M (2000) Analysis of yeast protein kinases using protein chips. *Nature Genetics*, **26**, 283–289.

References

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO and Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3889–3894.

Brem RB, Yvert G, Clinton R and Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, **2**, 65–73.

Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Research*, **32**, D311–D314.

Clark TA, Sugnet CW and Ares M Jr (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.

DeRisi JL, Iyer VR and Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Grigull J, Mnaimneh S, Pootoolal J, Robinson MD and Hughes TR (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molecular And Cellular Biology*, **24**, 5534–5547.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES and Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Hurowitz EH and Brown PO (2003) Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. *Genome Biology*, **5**, R2.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Marc P, Margeot A, Devaux F, Blugeon C, Corral-Debrinski M and Jacq C (2002) Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO Reports*, **3**, 159–164.

Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, *et al*. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, **4**, 1293–1301.

Ooi SL, Shoemaker DD and Boeke JD (2003) DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nature Genetics*, **35**, 277–286.

Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ and Fangman WL (2001) Replication dynamics of the yeast genome. *Science*, **294**, 115–121.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Shoemaker DD, Lashkari DA, Morris D, Mittmann M and Davis RW (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genetics*, **14**, 450–456.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of The Cell*, **9**, 3273–3297.

Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH and Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326–330.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D and Brown PO (2002) Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5860–5865.

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Winzeler EA, Lee B, McCusker JH and Davis RW (1999a) Whole genome genetic-typing in yeast using high-density oligonucleotide arrays. *Parasitology*, **118**(Suppl), S73–S80.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al*. (1999b) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Short Specialist Review

Bacterial genome organization: comparative expression profiling, operons, regulons, and beyond

Scott N. Peterson
The Institute for Genomic Research, Rockville, MD, USA
The George Washington University, Washington DC, USA

1. Bacterial ecosystem diversity is mirrored at the level of the genome

The diversity of bacterial species is not known, however, it has been estimated that less than 1% of the bacterial species have been characterized to date (DeLong and Pace, 2001). Bacteria occupy virtually every ecosystem present on our planet, including "extreme" environments wherein certain bacteria survive in temperatures greater than 110°C or less than −10°C, ranges of pH from <1.0 to >9.0, salt concentrations as high as 5M NaCl, and pressures in excess of 130 Mpa (Rothschild and Mancinelli, 2001). Since the environment provides the selective pressure responsible for bacterial genome evolution, the wide diversity of environments occupied by microbes offer a partial explanation for the genomic diversity observed thus far. A major insight resulting from widespread sequencing of bacterial genomes is the apparent prevalence of gene acquisition through horizontal gene transfer events in bacterial genome evolution (Lawrence and Hendrickson, 2003). DNA sequence data (from more than 130 microbial genomes) indicate that bacterial chromosomes come in a wide variety of shapes, sizes, and gene complements, and emphasize that exceptions are the rule in bacterial genomics. One genomic feature conserved in all bacterial genomes is the organization of genes in cotranscribed clusters referred to as *operons*. Conserved genomic features shared among all characterized bacteria are rare and thus justify the importance that we attribute to them.

The universal occurrence of operons in bacterial genomes strongly suggests that the selective forces responsible for their formation are powerful and pervasive. The "selfish operon" model is based on the assumption that horizontal gene acquisition is the force that drives operon formation (Lawrence and Roth, 1996). Alternative models involving tandem gene duplication have also been proposed (Reams and Neidle, 2004). Genes located within operons frequently encode proteins that act in concert to carry out a specific function, such as steps in a metabolic

pathway. According to the selfish operon model, the functional congruency of genes within operons is a natural consequence of the fact that the inheritance of novel, advantageous, multigenic traits can only be coselected and fixed in the genome if all required genes are acquired in a single DNA acquisition event. Most genes within horizontally acquired genomic fragments provide no selective advantage to the host and are lost over time, thereby increasing the physical linkage of coselected genes. The increased physical linkage between coselected genes further increases the probability that they will be coacquired in a single event. This self-perpetuating process may reach a practical end point when coselected genes achieve maximal linkage. The fact that coregulated genes tend to be "like-minded" with respect to function is a useful means of forming hypotheses pertaining to the function of genes of unknown function (at least 30% of all genes in microbial genomes). A refinement of this idea considers the conservation of gene linkage across bacterial species as a means of increasing the power of functional predictions (Overbeek *et al*., 1999). Operon structure (gene identity and order within the operon) while sometimes conserved is more frequently variable, especially over large phylogenetic distances (Lathe *et al*., 2000; Wolf *et al*., 2001). The heterogeneity of operon structure may result from numerous opposing evolutionary forces that generate heterogeneity in bacterial populations through genome rearrangements, mobile elements, and recombination. It is important to remember that a genome sequence represents a "snapshot" of evolution and as such is a "work in progress".

2. Global analyses provide a new way to visualize microbial genomes

There are an increasing number of algorithms being applied to identify functional elements within genomes, such as transcriptional terminators (Ermolaeva *et al*., 2000), DNA regulatory motifs (Roth *et al*., 1998; McGuire *et al*., 2000), and operons (Salgado *et al*., 2000; Ermolaeva *et al*., 2001). The sensitivity, specificity, and precision of operon, promoter, and transcription terminator identification are ever improving and are remarkably efficient at identifying these genomic elements (Bockhorst *et al*., 2003). Experiments to better define the *Escherichia coli* transcriptome, using high-density Affymetrix chips, confirmed the existence of many previously predicted transcription units (operons) and identified many novel transcripts, particularly those expressed from small open reading frames (Tjaden *et al*., 2002; Bockhorst *et al*., 2003). The colocalization of genes exhibiting similar transcriptional behavior represent a powerful means of defining operon structure and simultaneously suggest where 5′ promoter elements are likely to reside. Transcription unit (operon) definition, whether achieved experimentally or computationally, are complementary and promise an enhanced ability to experimentally test computationally based motif predictions. Likewise, experimental datasets will aid in the validation and improvement of future motif-finding algorithms. Tjaden and colleagues assessed gene expression of *E. coli* grown in minimal and rich medium, at four distinct growth phases, both aerobically and anaerobically. The semi-independent nature of each growth condition examined

provided a dataset wherein the complexity of gene expression inherent to each growth condition can be dissected, and in some cases, allows one to define a regulon (a group of coregulated operons). For those attempting to understand the relationship between cellular signals and the associated gene expression response, it is useful to view bacterial genomes not as a set of genes but as a set of operons and regulons. It is likely that an understanding of transcription will require a detailed understanding not only of operons and regulons but also of their interaction and cross-talk under specific environmental conditions.

3. Global gene expression analysis of *S. pneumoniae* competence development

Streptococcus pneumoniae cells growing at low cell densities produce and respond to a signal transduction, competence stimulating peptide (CSP). Cells exposed to CSP undergo a rapid and synchronous response resulting in competence (DNA uptake and genomic integration) (Morrison and Baker, 1979). At the molecular level, it had been established that CSP induces a cell signaling cascade, resulting in the upregulation of various RNA transcripts and proteins required to mediate the uptake of extracellular DNA and its subsequent homology-dependent introduction into the genome, mediated by recA (Havarstein *et al*., 1995; Peterson *et al*., 2000; Rimini *et al*., 2000). The global gene expression analysis using DNA microarrays identified all 40 previously known and 84 new CSP-induced genes (Peterson *et al*., 2004a,b). Induced genes exhibit three discrete temporal phases of activation. The first wave (max = 7.5 min post-CSP) of transcripts includes those encoding CSP and the membrane apparatus for its export (signal amplification), transcriptional regulators, known to regulate early and late gene expression and a set of genes related to bacteriocin production. The second wave (max = 12.5 min post-CSP) of transcriptional activation (81 genes) encodes proteins required for DNA uptake, processing, and recombination. The final wave (max = 20 min post-CSP) includes 20 genes encoding a complete set of heat shock effectors. Concurrent with our gene expression analysis, members of Don Morrison's laboratory (The University of Illinois, Chicago) constructed and analyzed numerous mutant strains bearing deletions in CSP-induced genes. Among the 91 genes analyzed, 23 were essential for growth, whereas 67 were individually dispensable for competence development. Only one new essential gene (unknown function) was identified. Despite the high efficiency with which global gene expression analysis identified CSP-induced genes, it was surprising that such a great fraction of those genes appeared to be dispensable for competence development.

4. Molecular dissection of DNA double-strand break repair in *D. radiodurans*

Deinococcus radiodurans is best known for its extraordinary resistance to ionizing radiation (X rays). The molecular events associated with this resistance remain largely uncharacterized. *D. radiodurans*, like many other bacteria, maintains its

genome (two chromosomes and two extrachromosomal plasmids) in multiple copies and is able to repair as many as 150 DNA double-strand breaks in a recA-dependent process. Remarkably, the recovery of cells following high exposures of ionizing radiation (5 kGy) occurs without lethality or apparent mutagenesis (Battista *et al*., 1999). The cell density of cultures begins to increase within 60 min following irradiation, suggesting that the repair of DNA double-strand breaks and the oxidative damage caused by radiation-dependent free-radical formation is complete in some cells within the population (Tanaka *et al*., 2004). *D. radiodurans* is also unusually resistant to a number of other stresses, such as desiccation, UV radiation, heat, oxidative damage, and osmotic stress. It has been speculated that the wide array of resistance observed in *D. radiodurans* arose primarily as an adaptation to survival in desiccating environments (Battista, 1997). Analysis of the *D. radiodurans* genome sequence revealed a rather ordinary repertoire of DNA metabolism and repair genes and, therefore, few clues to help explain the observed resistance of this bacterium (White *et al*., 1999). Mattimore and Battista (1996) established a connection between the resistance to ionizing radiation and desiccation genetically by demonstrating that a collection of *D. radiodurans* radiation sensitive strains were also sensitive to desiccation, suggesting a shared molecular basis for the two resistance profiles.

Since desiccation also generates DNA double-strand breaks, we exploited the established association between resistance to ionizing radiation and cellular desiccation to identify functionally important genes involved in DNA double-strand break repair. Since cellular desiccation and the delivery of high doses of ionizing radiation are not instantaneous, our expression analyses examines a transcriptional response already in progress and therefore does not capture the onset and induction kinetics of the induced genes. The gene expression changes in response to ionizing radiation and cellular desiccation are of limited complexity and, somewhat unexpectedly, small in magnitude (Tanaka *et al*., 2004). A total of 72 genes (2.2% of genome) were induced following exposure to ionizing radiation (3 kGy), of which the majority encoded functions belonging to three functional classes. The largest class (32 genes, 44%) encodes proteins of unknown function. The second group (*recA, ruvB, uvrA, uvrB, gyrA,* and *gyrB*) encodes functions involved in DNA metabolism. Five genes with an apparent role in oxidative damage protection make up the third class. The gene expression pattern associated with recovery of cells from desiccation (5% humidity) involved 73 genes, of which 33 were in common with those induced in response to irradiation. Among the overlapping transcripts were all six DNA metabolism and repair genes and 21 genes of unknown function; only two of the five genes associated with adaptation to oxidative stress were induced in cells recovering from desiccation.

The five most highly induced genes were identical in both responses and encode proteins of unknown function. Mutational analysis of strains bearing deletions of these five genes individually and in all pairwise combinations revealed that these genes encode proteins that participate in both recA-dependent and independent processes. Only three of the five single gene knockout strains displayed increased sensitivity to ionizing radiation.

This poor correlation between gene activation and functional essentiality appears to mirror that observed for *S. pneumoniae* competence development. However,

when these single gene deletions were examined in pairwise combination and in a *recA*-deficient background, a different picture emerged. Two of the proteins were shown to possess partially redundant or complementary activities. With the exception of paralogous gene families, functional redundancy in bacterial genomes is, to some extent, unexpected.

The overrepresentation of genes of unknown function in *D. radiodurans* related to DNA damage repair strongly imply that this process occurs either by a largely novel mechanism or by well-characterized mechanisms involving genes bearing little sequence identity to functionally equivalent proteins found in other bacteria. The lack of sequence identity between DNA repair proteins in *D. radiodurans* to other bacterial proteins may not be a simple matter of divergent evolution but rather the consequence of nonorthologous gene relationships. Nonorthologous genes encode proteins that perform identical or similar functions within a cell but owing to their independent evolution have no ancestral relationship to their functional counterparts in other species. In other words, nonorthologous genes are those that have evolved more than once during evolution. Examples of nonorthologous genes have been identified in nature (Koonin *et al*., 1996). The fact that each successive microbial genome sequencing project reveals a set of genes of unknown function at nondeclining frequency suggests that nonorthologous genes may be prevalent and as such pose a daunting problem within the field of microbial genomics. The assignment of gene function based on sequence identity (orthology) has had enormous impact on our ability to "understand" genomic sequence information. In the absence of orthology, the only means of obtaining information on the function of a protein is through focused laboratory experimentation. It is conceivable that our reliance on the functional data generated as the result of the efforts made in a few select model systems, while powerful, cannot account for all of the coding potential in any genome. It is possible that the next phase of genomics effort should address this question in more detail and that ultimately functional characterization of lesser known bacterial models will be required if we hope to reinvigorate the orthology-based functional assignments that we have come to rely so heavily upon.

5. Transcriptional analysis of *Bacillus anthracis* sporulation and determination of the endospore proteome

The dormant spore of *Bacillus anthracis*, the bacterium thought to cause anthrax, is the infectious agent. Upon entry into a host through inhalation, the spore efficiently infects macrophages where it rapidly germinates within the harsh environment of the phagolysosome (Guidi-Rontani and Mock, 2002). The determination of the gene expression in *B. anthracis* cells undergoing sporulation (induced by nutrient starvation) and of the spore as it germinates are of fundamental interest. We recently reported on the gene expression during sporulation and the protein content of the spore as determined by multidimensional chromatography and tandem mass spectrometry (Lui *et al*., 2004). The combined use of these two global approaches provided a means of relating the observed gene expression to the proteome of the spore itself. We compared the RNA expression patterns of 20 samples (15-min intervals) over a 5-h period corresponding to exponential, nonexponential, and

stationary phase growth, as well as, the entirety of sporulation. It is known that the onset and progression of sporulation in the closely related *B. subtilis* involves the sequential activation of a number of sigma factors that mediate and coordinate the temporal expression of the so-called sporulation genes (Kroos and Yu, 2000). In *B. anthracis*, a significant and reproducible change in gene expression was observed for over 3500 genes (63% of the genome). Just over 2000 genes displayed growth-phase regulated expression in two or more consecutive time points (>30 min). Hierarchical clustering of the expression data revealed a complex set of expression patterns. Through a slightly oversimplified view of the data, the gene expression patterns can be represented by five distinct patterns, based on the temporal onset of gene expression changes (activation or repression). These five patterns appear to correspond to biologically meaningful phases, as judged by the number of transcripts and transcriptional regulators involved at each phase. Approximately 1100 genes displayed altered RNA expression within the first two phases, which correspond to cells exiting exponential growth phase and entering stationary phase growth and stationary phase growth just prior to the onset of sporulation. The heterogeneity of transcriptional behavior among genes within these two phases is substantial especially when compared to the transcript profiles corresponding to ~900 genes observed in the subsequent three phases occurring during sporulation.

Once inside the macrophage, the *B. anthracis* spore enters into a race with the host immune response. Consequently, the spore has become uniquely adapted to germinate and begin vegetative growth rapidly. For this reason and others, it had been assumed that the spore contained proteins to ensure efficient reentry into a growth phase. Among the 750 proteins identified in the endospore, half are encoded by genes expressed constitutively throughout the entire experimental time course. Consistent with prior assumptions, these transcripts encode proteins biased in functional representation, the most overrepresented classes being those involved in protein synthesis, nucleoside/tide biosynthesis, and energy metabolism. The remaining half of the endospore proteome is composed of proteins encoded by genes displaying altered RNA expression in each of the five expression phases in approximately equivalent proportions. It was not anticipated that gene expression changes occurring during the first two growth phases would contribute equally to the endospore as genes differentially expressed during sporulation. It is somewhat surprising that approximately 15% of the spore proteome is derived from strongly repressed genes, an interesting subset of which are upregulated (~75 min) as cells enter saturation phase. These genes are subsequently repressed as cells enter into sporulation. This gene expression pattern may suggest that preparations for sporulation occur prior to the cell's actual commitment to this pathway. It is noteworthy that less than 20% of the genes displaying sporulation-limited RNA expression encode proteins present in the endospore. Taken together, the two global analyses support the view that *B. anthracis* sporulation is a highly controlled and complex process. The proteins (raw materials) and enzymes (tools) required to construct and load the endospore are expressed in a growth-phase dependent and independent manner. The proteins encoded by genes displaying sporulation-specific expression contribute several functions related to the structure and architecture of the endospore. The function and significance of many proteins differentially expressed during

sporulation remain to be determined. However, our data suggest that the majority of differential gene expression associated with sporulation is intended to provide proteins required for the processing of the raw materials and the physical construction of the spore. In this regard, the spore, like most building projects, necessitates the acquisition of a specialized set of tools and "some assembly is required".

6. *S. pneumoniae* growth-phase-dependent gene expression: comparative expression profiling (CEP)

Growth-phase-dependent gene expression in *S. pneumoniae* is of comparable complexity (∼1000 genes) to that observed in *B. anthracis* but represents a significantly greater proportion of its genome's coding capacity (∼45% compared to ∼22% for *B. anthracis*). In *S. pneumoniae*, the majority (90%) of growth-phase-regulated genes are downregulated as cells entered stationary phase growth. The potential commonality of cellular signals (e.g., glucose limitation) generated by cells as they enter saturation phase growth provides an opportunity to determine whether the comparisons of gene expression profiles, as illustrated by the *D. radiodurans* study, can be fruitfully applied to a wide range of species. We refer to this approach as comparative expression profiling (CEP). *B. anthracis* and *S. pneumoniae* are both members of the low G+C gram-positive lineage. *S. pneumoniae* does not sporulate in response to nutrient limitations. It is possible that aspects of the transcriptional and metabolic response to nutrient limitation are common to these distantly related cousins. Nutrient limitation may represent a primordial selective pressure faced by all bacterial species throughout time. Given the potential commonality of the selective pressure, we were interested in determining whether mechanisms to cope with nutrient limitations are highly conserved across species. Early stage analysis of the differentially regulated genes in *S. pneumoniae* indicates that a large number of similarities and differences exist to growth-phase-regulated genes in *B. anthracis*. This comparative analysis may also allow us to refine our focus toward those genes specifically related to and functionally required for sporulation and endospore formation (Figure 1).

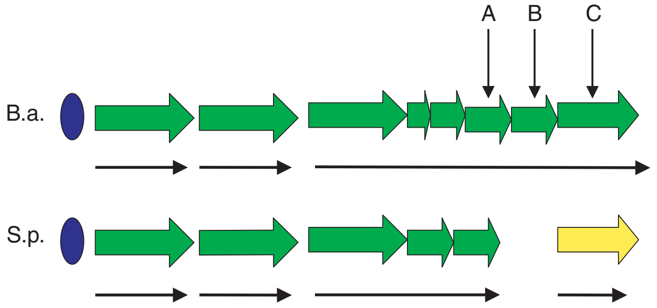


Figure 1 Comparative expression profiling (CEP)

7. The relationship between gene expression and functional significance: *S. pneumoniae* competence gene expression revisited

The universal occurrence of operons in bacterial genomes and the organization of genes encoding proteins of functional relationship in cotranscribed operons and coexpressed regulons are consistent with the frugality and efficiency that we expect from bacteria. The fact that only 18% of the genes differentially expressed in response to CSP in *S. pneumoniae* appear to be essential for wild-type competence development suggests a lack of frugality inherent to this response. While performing CEP using the *S. pneumoniae* competence and growth-phase gene expression data, a fortuitous observation was made. There was a total of 34 genes differentially expressed in both studies. Upon closer inspection, the transcription units predicted by several CSP-induced genes differed from those predicted by the growth-phase expression study. The most compelling interpretation of this observation is that many of the genes induced during competence development are the result of failures to terminate transcription efficiently. Potential transcriptional read-through affecting 37 genes (4 early, 33 late) was noted. Overall, this modification increases the percentage of essential genes induced during competence development to 27%, a value that still falls short of our expectations of a highly evolved bacterial system. The 22 gene products whose functions are required for competence reside in 13 operons (5 early and 8 late). All of the genes (8) within the five early operons are essential, whereas 14 of the 19 genes present in the eight late operons are essential. All instances of "read-through" transcription occurred downstream of the previously characterized promoter motifs associated with early and late gene expression and correlated with the transcripts undergoing the largest increases following CSP-induction (~100- to 1000-fold, Snesrud *et al*., 2004). The relationship between promoter strength and the occurrence of read-through transcription is not yet clear and may be trivially related to the ease of detection for highly expressed genes. Alternatively, it may be that some operons contain weak transcriptional terminators "by design" (see below).

A useful concept in microbial genomics invokes a view of genomes as consisting essentially of two types of genes, a core set of conserved functions required for cell growth and existence and a larger set of conditionally dispensable contingency genes that provide the broadened functional repertoire demanded for success in a particular microenvironment (Mushegian and Koonin, 1996; Peterson and Fraser, 2001). It may be that the competence regulons can be viewed in a similar manner. In this regard, 22 of the 27 core competence genes (81%), residing in 13 operons, expressed are essential for wild-type competence. A larger set of nonessential genes such as those mediating chaperonin and proteins folding and degradation and those involved in bacteriocin production that are not required for competence may provide subtle fitness advantages to the cell. Another possibility suggested by a parallel study to our own is that CSP-induced genes play a role in a phenomenon referred to as *saturation phase cell autolysis* (Dagkessamanskaia *et al*., 2004). Competence, a low-cell-density phenomenon, is not detectable in saturation phase cultures where autolysis occurs. The possibility that CSP-induced regulons play a role in biological

processes other than competence suggests that it is our expectation that differentially expressed genes be functionally required that is misguided. Genes within operons deemed dispensable for wild-type competence may encode proteins with activities required for cell autolysis and vice versa. Stationary phase autolysis is known to be mediated by *lytA* by an unknown mechanism (Tomasz and Zanati, 1971). It is interesting that *lytA* is induced in response to CSP as the result of transcriptional read-through and serves as fair warning not to be too hasty to assume that the role of CSP-induced operons in stationary phase autolysis is a biological anomaly. After all, it is through "accidents" that genomic novelty and expansion of functional capabilities are achieved in nature. It is conceivable that only a small number of genes in any genome are dedicated to functin in a single scenario. The induction of operons in multiple cellular contexts wherein the definition of essentiality is dynamic support the viewpoint that despite appearances, gene organization within bacterial genomes is remarkably efficient and may account for bacteria's ability to respond to a wide variety of environmental signals with a relatively limited number of genes.

8. Summary and conclusions

The conservation of gene pair (coexpression) relationships across species has been used extensively as a predictor of protein interaction pairs and the coparticipation of proteins in the same cellular process. In this manner, CEP represents an extension of the same idea. The application of CEP to closely related species may enhance the identification of the core gene sets that are functionally required to accomplish the cellular response under investigation. When applied to species of distant phylogenetic relationships, CEP may provide a systematic means of identifying candidate nonorthologous gene relationships and, more importantly, valuable clues to those scientists attempting to functionally annotate those genes. Ultimately, our ability or inability to interpret global gene expression is primarily limited by our knowledge of the number and identity of pertinent cellular signals and/or the details of the molecular events associated with the response to those signals. The expansion of tools and methods used for gene expression data "mining", together with the efforts of computational biologists generating high-quality predictions of regulatory sequences and transcription units in bacterial genomes, will allow our understanding of the complexities of gene regulation at the level of the operon and regulon to reach new heights. As we begin to document the interplay and cross-talk occurring between regulons in various physiological contexts, we are certain to gain a greater understanding of bacterial genome evolution and define new inroads for the development of antimicrobial therapies.

References

Battista JR (1997) Against all odds: the survival strategies of *Deinococcus radioudurans*. *Annual Review of Microbiology*, **51**, 203–224.

Battista JR, Earl AM and Park MJ (1999) Why is *Deinococcus radioudurans* so resistant to ionizing radiation? *Trends in Microbiology*, **7**, 362–365.

Bockhorst J, Qui Y, Glasner J, Liu M, Blattner F and Craven M (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34–i43.

Dagkessamanskaia A, Moscoso M, Henard V, Guiral S, Overweg K, Rueter M, Martin B, Wells J and Claverys JP (2004) Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae:* competence triggers stationary phase autolysis of *ciaR* mutant cells. *Molecular Microbiology*, **51**, 1071–1086.

DeLong EF and Pace NR (2001) Environmental diversity of bacteria and archaea. *Systematic Biology*, **50**, 470–478.

Ermolaeva MD, Khalak HG, White O, Smith HO and Salzberg SL (2000) Prediction of transcription terminators in bacterial genomes. *Journal of Molecular Biology*, **301**, 27–33.

Ermolaeva MD, White O and Salzberg SL (2001) Prediction of operons in microbial genomes. *Nucleic Acids Research*, **29**, 1216–1221.

Guidi-Rontani C and Mock M (2002) Macrophage interactions. *Current Topics in Microbiology and Immunology*, **271**, 115–141.

Havarstein LS, Coomaraswamy G and Morrison DA (1995) An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences of United States of America*, **92**, 11140–11144.

Koonin EV, Mushegian AR and Bork P (1996) Non-orthologus gene displacement. *Trends in Genetics*, **12**, 334–336.

Kroos L and Yu Y (2000) Regulation of sigma factor activity during *Bacillus subtilis* development. *Current Opinion in Microbiology*, **3**, 553–560.

Lathe WC III, Snel B and Bork P (2000) Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences*, **25**, 474–479.

Lawrence JG and Hendrickson H (2003) Lateral gene transfer: When will adolescence end? *Molecular Microbiology*, **50**, 739–749.

Lawrence JG and Roth JR (1996) Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.

Lui H, Bergman NH, Thomason B, Shallom S, Hazen A, Crossno J, Rasko DA, Ravel J, Read TD, Peterson SN, *et al*. (2004) Formation and composition of the *Bacillus anthracis* endospore. *Journal of Bacteriology*, **186**, 164–178.

Mattimore V and Battista JR (1996) Radioresistance of *Deinococcus radioudurans:* functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *Journal of Bacteriology*, **178**, 633–637.

McGuire AM, Hughes JD and Church GM (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research*, **10**, 744–757.

Morrison DA and Baker MF (1979) Competence for genetic transformation in pneumococcus depends on synthesis of a small set of proteins. *Nature*, **282**, 215–217.

Mushegian AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of United States of America*, **93**, 10268–10273.

Overbeek R, Fonstein M, D’Souza MD, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of United States of America*, **96**, 2896–2901.

Peterson CT, Ahn S, Tettelin H and Peterson SN (2004b) *The S. pneumoniae transcriptome as defined by DNA microarrays, DNA sequence motif identification and comparative expression profiling*. (manuscript in preparation).

Peterson SN, Cline RT, Tettelin H, Sharov V and Morrison DA (2000) Gene expression analysis of the *Streptococcus pneumoniae* competence regulons by use of DNA microarrays. *Journal of Bacteriology*, **182**, 6192–6202.

Peterson SN and Fraser CM (2001) The complexity of simplicity. *Genome Biology*, **2**, 2002.1–2002.8.

Peterson SN, Sung CK, Cline RT, Desai BV, Snesrud EC, Luo P, Walling J, Li H, Mintz M, Tsegaye G, *et al*. (2004a) Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays. *Molecular Microbiology*, **51**, 1051–1070.

Reams AB and Neidle EL (2004) Selection for gene clustering by tandem duplication. *Annual Review of Microbiology*, **58**, 119–142.

Rimini R, Jannson B, Feger G, Roberts TC, deFrancesco M and Gozzi A (2000) Global Analysis of transcription kinetics during competence development in *Streptococcus pneumoniae* using high density DNA arrays. *Molecular Microbiology*, **36**, 1279–1292.

Roth FP, Hughes JD, Estep PW and Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, **16**, 939–945.

Rothschild LJ and Mancinelli RL (2001) Life in extreme environments. *Nature*, **409**, 1092–1101.

Salgado H, Moreno-Hagelsieb G, Smith TF and Collado-Vides J (2000) Operons in *Escherichia coli:* genomic analyses and predictions. *Proceedings of the National Academy of Sciences of United States of America*, **97**, 6652–6657.

Snesrud EC, Fleischmann RD and Peterson SN (2004) *Quantitative limits of DNA microarray gene expression data* (manuscript in preparation).

Tanaka M, Earl AM, Howell HA, Park MJ, Eisen JA, Peterson SN and Battista JR (2004) Analysis of *Deinoccous radiodurans'* transcriptional response to ionizing radiation and desiccation reveals novel proteins that contribute to extreme radioresistance. *Genetics*, **168**, 21–33.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E and Rosenow C (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Research*, **30**, 3732–3738.

Tomasz A and Zanati E (1971) Appearance of a protein 'agglutinin' on the spheroplast membrane of pneumococci during induction of competence. *Journal of Bacteriology*, **105**, 1213–1215.

White O, Eisen JA, Heidleberg JF, Hickey EK and Peterson JD (1999) Genome sequence of the radioresistant bacterium *Deinococcus radioudurans* R1. *Science*, **286**, 1571–1577.

Wolf YI, Rogozin IB, Kondrashov AS and Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and the prediction of gene function using genomic context. *Genome Research*, **11**, 356–372.

Short Specialist Review

Genomic analysis of host pathogen interactions

Paul Kellam and Catherine V. Gale
University College London, London, UK

In the era of postgenomic biology, an area of much interest is how a multicellular organism interacts with the microbial world. This is perhaps not surprising when one considers the importance the human body places on protection from microbes. At a very rudimentary level of classification, of the approximately 200 cell types of the human body, 20 are concerned entirely with protection of the host from microbial attack. This review is broadly applicable to all types of pathogen, however, we will focus primarily on viral infections because of the close association between virus life cycles with the host cell environment.

Of the available postgenomic tools for the investigation of host pathogen interactions, gene expression profiling currently seems the most useful. This is aided by the fact that microbes often interact with host cells that are homoeostatically stable and therefore are in relatively steady state with respect to gene expression dynamics. Microbial encounter results in one of three outcomes although transition between the states can occur, namely, acute pathogenic microbial replication resulting in either host death or resolution of infection; microbial persistence without overt pathology and without microbial clearance; or commensalism where the microbe and host exist in a mutually beneficial way. In all cases, the major effect on host cells and tissues is a remodeling of their gene expression program to accommodate or combat the particular microbial exposure (Kellam, 2001). Dynamic changes in host gene expression programs are seen in the large number of studies where microarrays are used in an infection context (Figure 1) (Kellam, 2001; Bryant *et al*., 2004; Fruh *et al*., 2001). However, only recently has the data become of sufficient size to allow computational integration (Jenner and Young, 2005). Overall, common gene expression programs can be identified in many cell types, revealing a coordinated view of expression for many known infection-responsive genes. These include cytokines and chemokines, genes involved in the innate antiviral responses, especially interferon-inducible gene sets, cell cycle control, and apoptotic control gene sets. Gene expression profiling studies have shown that these common cellular responses to infection can be induced by related viruses, but by different mechanisms. Exposing human embryonic lung (HEL) cells to either IFN-α or a herpes simplex virus KM110 that contains inactivating mutations preventing lytic viral replication produce a very similar remodeling of the cellular transcriptome (20 of 27 modulated genes were induced by HSV-1

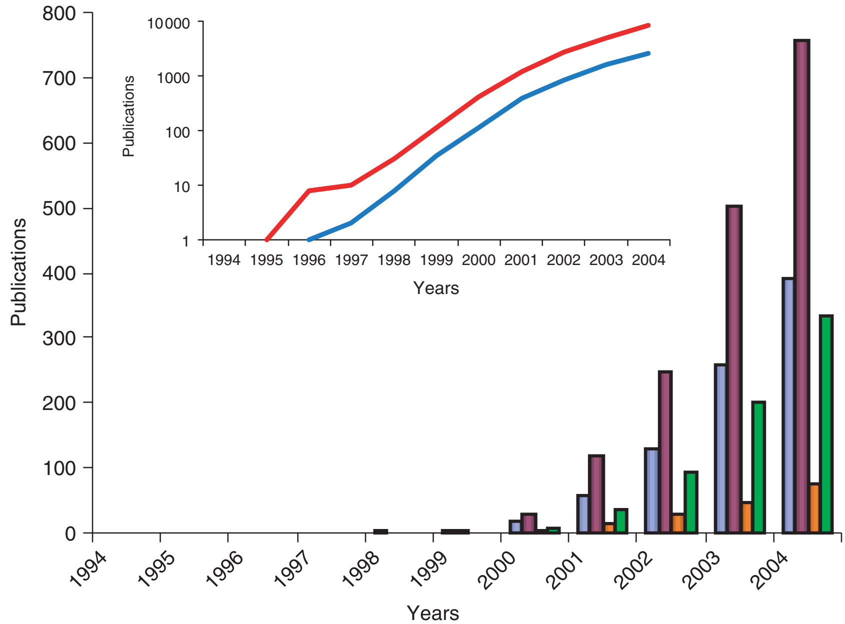


Figure 1 The continuing rise in the use of microarrays in infectious disease research. Pubmed (**http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed**) was searched with the terms "*microarray*" (red line, inset graph), "*cancer and microarray*" (blue line, inset graph), "*bacteria and microarray*" (purple bar), "*virus and microarray*" (blue bar), "*immune and microarray*" (green bar), and "*(host and pathogen) AND microarray*" (orange bar) for each year, with the number of publications recorded. While this is a crude relative measure, it serves to illustrate the trends of cancer biology driving the rapid rise in microarray usage, but with other areas of biology such as infectious disease research adopting microarray methods from the late 1990s onward

KM110 and IFN-α) (Mossman *et al*., 2001). The effects of the HSV-1 mutant occurred following the internalization of the viral particles but in the absence of viral gene expression, suggesting that HSV-1 activates an intracellular sensor that induces IFN-responsive genes at an early step during virus infection. In contrast, the herpesvirus, human cytomegalovirus (HCMV) induces IFN-responsive genes in primary human fibroblasts following attachment of virions to an unknown cellular receptor. In this context, the HCMV envelope glycoprotein B is necessary and sufficient for the IFN response. As with HSV-1, this happens in the absence of viral gene expression, but in contrast to HSV-1 occurs before virion internalization (Boyle *et al*., 1999), demonstrating that different cellular interactions, with distinct viruses, can converge on common antiviral transcriptional programs.

The infected cell transcriptome must, however, be put into context. Virus-infected cells in a host organism do not exist in isolation. The integrated induction of the innate and adaptive immune response to combat an infection requires the activation and terminal differentiation of diverse hematopoietic cell types. This multicell-type immune response, coordinated through transcriptional reprogramming of diverse hematopoietic cells, is therefore intimately coupled to the infected cell response. At the level of the innate arm, pathogen encounter by professional

antigen-presenting cells such as macrophages and (monocyte derived) dendritic cells, which share 96% of expressed genes prior to antigen encounter (Chaussabel *et al*., 2003), results in conserved macrophage or dendritic cell transcriptional responses that can be interpreted as "core" effector responses for the presence of a pathogen (Chaussabel *et al*., 2003; Nau *et al*., 2002; Huang *et al*., 2001; Ehrt *et al*., 2001; Granucci *et al*., 2001). The degree of transcriptional programs conservation between macrophages and (monocyte derived) dendritic cells to the same pathogen is less, with only 40% of the regulated genes being shared by both cell types (Chaussabel *et al*., 2003). These transcriptional changes are most likely due to the pathogen-specific engagement of a repertoire of innate immune receptor molecules such as the toll-like receptors (TLRs). The combinatorial integration of such receptor stimulation results in the induction of both the "core" activation and maturation pathways together with pathogen-specific transcriptional responses. This has been demonstrated in detail for macrophages and (monocyte derived) dendritic cells for distinct bacterial pathogens (or pathogen components) (Chaussabel *et al*., 2003) and for (monocyte derived) dendritic cells for diverse pathogens such as the yeast *Candida albicans*, the bacteria *Esherichia coli*, and Influenza Virus (Huang *et al*., 2001). In this context, diverse pathogen input into signaling pathways results in pathogen-specific outputs. Whereas this may be expected for professional antigen-presenting cells, allowing the formation of pathogen-encoded bridges between the innate and adaptive immune response, it raises the question as to what level other cell types respond differently to related or diverse pathogens.

Different herpesviruses infecting the same cell type show how viruses influence discrete transcriptional networks. An extensive study of the alphaherpesviruses Pseudorabies virus (PRV), and HSV-1, infecting rat embryonic fibroblasts was analyzed by grouping coregulated genes into functional pathways. This showed that both viruses upregulated 32% of cellular genes consistently, such as the insulin growth factor pathways and apoptosis pathway components (Ray and Enquist, 2004). Interestingly, the two viruses also differentially regulated important functional pathways, with PRV downregulating Notch pathway components and HSV-1 upregulating interferon and interleukin 1 pathway components. Differences in cellular responses at much lower levels of viral variation are also seen in studies addressing the impact of viral strain variation on gene expression and the potential functional relevance. Most of these studies address the issue of intraspecies variation by looking at examples where clear differences in infection phenotype lie. For example, a comparison of adenovirus type-5 and -12 transformation of human embryonic retina, kidney, and lung cells revealed differential gene expression of TGFβ factors that might contribute to the differences in oncogenic behavior of these two viruses (Vertegaal *et al*., 2000). Similarly, differential expression of proinflammatory cytokines in response to pathogenic and nonpathogenic pneumovirus infection, in a mouse model, have been described (Domachowske *et al*., 2002). Microarrays have also been used to propose a model of cell cycle control by reoviruses that differs between two viral serotypes (Poggioli *et al*., 2002) and define the molecular mechanism of pathogenesis of the hantaviruses, Sin Nombre and Prospect Hill (Khaiboullina *et al*., 2004). In addition, a study of different measles virus strains has also showed differences in host cell gene expression patterns that could help explain their distinct growth phenotypes in vitro, although the

problem of low multiplicity, asynchronous infection on a variable cell substrate, make analysis difficult in this case. It is becoming increasingly evident that this problem is common when low titer viruses are studied (Bolt *et al*., 2002), and great attention to the virology of such experiments will be required to extract biological meaning and functional relevance from these intraspecies comparative studies.

At a whole-organ level, gene expression profiling of Hepatitis B (HBV) and C (HCV) viruses has also revealed that viruses can influence common cellular environments in different ways, in these cases causing hepatocellular carcinoma (HCC) by distinct mechanisms (Iizuka *et al*., 2002; Okabe *et al*., 2001). HBV and HCV chronic hepatitis could be distinguished on the basis of differentially expressed genes. A broad analysis of the types of genes involved showed that HBV upregulated genes involved in apoptosis, cell cycle arrest, and extracellular matrix degradation, whereas HCV upregulated genes involved in cell cycle acceleration and extracellular matrix storage. The major difference between HBV- and HCV-derived tumors was the upregulation in HCV and downregulation in HBV of genes involved in activating chemotherapeutic drugs or detoxifying xenobiotic carcinogens. This suggests that a greater contribution of carcinogenic metabolites may be involved in the development of the HCV-associated tumors and that decreased expression in HBV HCC is specific for this virally derived tumor type. The gene expression differences in the two HBV/HCV studies reflect the fact that HBV and HCV are unrelated viruses that infect the same organ. They may also reflect, however, the fact that the viruses infect common cell types at different stages of the cellular differentiation and development and alter the cellular transcriptome at the stage infected.

Gene expression profiling in viral oncogenesis illustrated above for HBV and HCV raises the possibility that viruses are able to manipulate intracellular transcriptional programs to drive the infected cells to a different state that is beneficial to the virus life cycle. Ultimately, this is likely to be a rationale for all viral remodeling of host cell transcriptional environments. When the virus life cycle results in persistent infection, however, this trade-off between virally induced processes and the cellular defense system must be well balanced. Perhaps the greatest evidence for this comes from the studies of human herpesvirus 8, also known as Kaposi's sarcoma associated herpesvirus (KSHV). This virus can infect and persist in a latent form in cells of the B-cell and endothelial cell lineage and, mainly in immune deficient settings, induce tumors of both cell types. Transcriptional profiling of viral and nonviral B-cell lymphomas and leukemias showed that latently KSHV-infected B-cells are arrested at the penultimate stage of B-cell differentiation into plasma cells (Jenner *et al*., 2003). These tumors are clearly transcriptionally distinct from the CD19 positive B-cells the virus is known to infect (Dittmer *et al*., 1999; Blackbourn *et al*., 2000). This suggests that KSHV, like the related human herpesvirus Epstein Barr Virus (EBV) is able to direct B-cells along differentiation pathways to a B-cell type more suitable for virus latency and possibly lytic replication (Jenner *et al*., 2003). This does not only apply to KSHV and B-cells, as recently KSHV reprogramming of the endothelial transcriptome was demonstrated (Wang *et al*., 2004; Hong *et al*., 2004; Carroll *et al*., 2004) showing that KSHV can infect either blood or lymphatic endothelial cells and drive their gene expression programs to a common intermediary (Wang *et al*., 2004). In both B-cells and endothelial cells, KSHV is thus able

to remodel the host cell transcriptional environment without leading to cell crisis and death. The coordinated molecular mechanisms of such transcriptional remodeling are unknown but suggest that just as viruses helped reveal detailed molecular mechanisms of the cell at the beginning of the molecular biology era, they may also be instrumental in determining authentic transcriptional control networks at a "systems" level of understanding.

The progress in gene expression profiling, however, largely ignores the multicell type, coordinated, and time-dependent dynamic of host and pathogen interaction (Figure 2). The real challenge for the future is to be able to integrate this genome scale complexity into models of how various pathogens survive in the hostile environment of the host and how this maybe exploited to design new strategies and therapies for the treatment of infectious diseases (Fruh *et al*., 2001). Within this setting, it is important to remember that the pathogen is not inert but is also regulating its gene expression to replicate, survive an immune response, and transfer to a new host. Understanding the coordinated, multihost cell choreography of gene

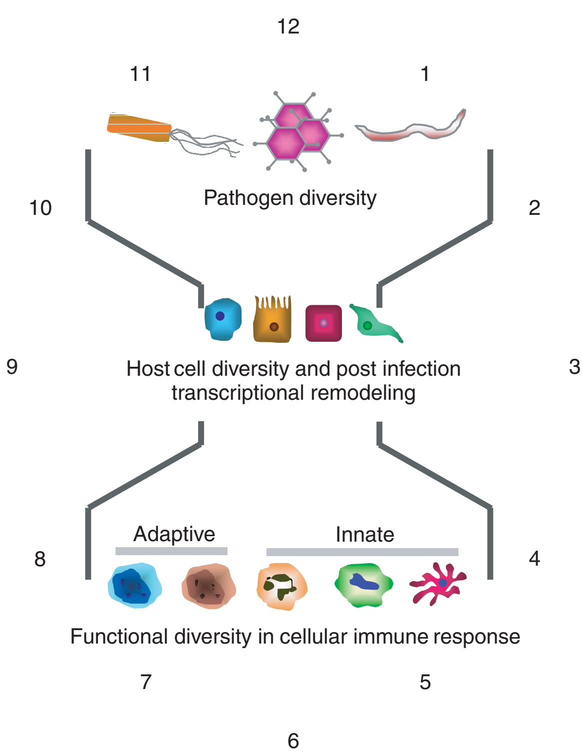


Figure 2 The multicell type, multipathogen context of infectious disease genomics. Different prokaryotic, eukaryotic, and viral microorganisms can interact with or infect probably all cell types of complex multicellular organisms. At the beginning of interaction, the "infection clock" starts, providing the time-dependent context for the remodeling of infected cells gene expression networks. The temporal regulation of both the innate and adaptive arms of the immune system also involves transcriptional remodeling with both conserved and pathogen-specific gene expression programs contributing to a coordinated immune response

expression changes to pathogens represents a challenge to biology at least as great as that posed in understanding developmental biology. With infectious disease accounting for 20–45% of a country's disease burden, and being intimately linked with social, environmental, and economic factors (Weiss and McMichael, 2004), this is a challenge of immense complexity and importance.

Acknowledgments

We are grateful to R. A. Weiss for critically reading the manuscript. C.V. Gale is supported by the Special Hospital Trustees at the Middlesex Hospital, London.

References

Blackbourn DJ, Lennette E, Klencke B, Moses A, Chandran B, Weinstein M, Glogau RG, Witte MH, Way DL, Kutzkey T, *et al*. (2000) The restricted cellular host range of human herpesvirus 8. *Aids*, **14**(9), 1123–1133.

Bolt G, Berg K and Blixenkrone-Moller M (2002) Measles virus-induced modulation of host-cell gene expression. *The Journal of General Virology*, **83**(Pt 5), 1157–1165.

Boyle KA, Pietropaolo RL and Compton T (1999) Engagement of the cellular receptor for glycoprotein B of human cytomegalovirus activates the interferon-responsive pathway. *Molecular and Cellular Biology*, **19**(5), 3607–3613.

Bryant PA, Venter D, Robins-Browne R and Curtis N (2004) Chips with everything: DNA microarrays in infectious diseases. *The Lancet Infectious Diseases*, **4**(2), 100–111.

Carroll PA, Brazeau E and Lagunoff M (2004) Kaposi's sarcoma-associated herpesvirus infection of blood endothelial cells induces lymphatic differentiation. *Virology*, **328**(1), 7–18.

Chaussabel D, Tolouei Semnani R, McDowell MA, Sacks D, Sher A and Nutman TB (2003) Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, **102**(2), 672–681.

Dittmer D, Stoddart C, Renne R, Linquist-Stepps V, Moreno ME, Bare C, McCune JM and Ganem D (1999) Experimental transmission of Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) to SCID-hu Thy/Liv mice. *The Journal of Experimental Medicine*, **190**(12), 1857–1868.

Domachowske JB, Bonville CA, Easton AJ and Rosenberg HF (2002) Differential expression of proinflammatory cytokine genes in vivo in response to pathogenic and nonpathogenic pneumovirus infections. *The Journal of Infectious Diseases*, **186**(1), 8–14.

Ehrt S, Schnappinger D, Bekiranov S, Drenkow J, Shi S, Gingeras TR, Gaasterland T, Schoolnik G and Nathan C (2001) Reprogramming of the macrophage transcriptome in response to interferon-gamma and Mycobacterium tuberculosis: signaling roles of nitric oxide synthase-2 and phagocyte oxidase. *The Journal of Experimental Medicine*, **194**(8), 1123–1140.

Fruh K, Simmen K, Luukkonen BG, Bell YC and Ghazal P (2001) Virogenomics: a novel approach to antiviral drug discovery. *Drug Discovery Today*, **6**(12), 621–627.

Granucci F, Vizzardelli C, Pavelka N, Feau S, Persico M, Virzi E, Rescigno M, Moro G and Ricciardi-Castagnoli P (2001) Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunology*, **2**(9), 882–888.

Hong YK, Foreman K, Shin JW, Hirakawa S, Curry CL, Sage DR, Libermann T, Dezube BJ, Fingeroth JD and Detmar M (2004) Lymphatic reprogramming of blood vascular endothelium by Kaposi sarcoma-associated herpesvirus. *Nature Genetics*, **36**(7), 683–685.

Huang Q, Liu D, Majewski P, Schulte LC, Korn JM, Young RA, Lander ES and Hacohen N (2001) The plasticity of dendritic cell responses to pathogens and their components. *Science*, **294**(5543), 870–875.

Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, Takemoto N, Tangoku A, Hamada K, Nakayama H, *et al*. (2002) Comparison of gene expression profiles between hepatitis B virus- and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data on the basis of a supervised learning method. *Cancer Research*, **62**(14), 3939–3944.

Jenner RG and Young RA (2005) Insights into host responses against pathogens from transcriptional profiling. *Nature Reviews Microbiology*, **3**(4), 281–294.

Jenner RG, Maillard K, Cattini N, Weiss RA, Boshoff C, Wooster R and Kellam P (2003) Kaposi's sarcoma-associated herpesvirus-infected primary effusion lymphoma has a plasma cell gene expression profile. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(18), 10399–10404.

Kellam P (2001) Post-genomic virology: the impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Reviews in Medical Virology*, **11**(5), 313–329.

Khaiboullina SF, Rizvanov AA, Otteson E, Miyazato A, Maciejewski J and St Jeor S (2004) Regulation of cellular gene expression in endothelial cells by sin nombre and prospect hill viruses. *Viral Immunology*, **17**(2), 234–251.

Mossman KL, Macgregor PF, Rozmus JJ, Goryachev AB, Edwards AM and Smiley JR (2001) Herpes simplex virus triggers and then disarms a host antiviral response. *Journal of Virology*, **75**(2), 750–758.

Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES and Young RA (2002) Human macrophage activation programs induced by bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(3), 1503–1508.

Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, Tsunoda T, Furukawa Y and Nakamura Y (2001) Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. *Cancer Research*, **61**(5), 2129–2137.

Poggioli GJ, DeBiasi RL, Bickel R, Jotte R, Spalding A, Johnson GL and Tyler KL (2002) Reovirus-induced alterations in gene expression related to cell cycle regulation. *Journal of Virology*, **76**(6), 2585–2594.

Ray N and Enquist LW (2004) Transcriptional response of a common permissive cell type to infection by two diverse alphaherpesviruses. *Journal of Virology*, **78**(7), 3489–3501.

Vertegaal AC, Kuiperij HB, van Laar T, Scharnhorst V, van der Eb AJ and Zantema A (2000) cDNA micro array identification of a gene differentially expressed in adenovirus type 5- versus type 12-transformed cells. *FEBS Letters*, **487**(2), 151–155.

Wang HW, Trotter MW, Lagos D, Bourboulia D, Henderson S, Makinen T, Elliman S, Flanagan AM, Alitalo K and Boshoff C (2004) Kaposi sarcoma herpesvirus-induced cellular reprogramming contributes to the lymphatic endothelial gene expression in Kaposi sarcoma. *Nature Genetics*, **36**(7), 687–693.

Weiss RA and McMichael AJ (2004) Social and environmental risk factors in the emergence of infectious diseases. *Nature Medicine*, **10**(12 Suppl), S70–S76.

Short Specialist Review

Protein microarrays as an emerging tool for proteomics

Ji Qiu and Sam Hanash
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

1. Introduction

Although the concept of a "multianalyte microspot immunoassay" system was first proposed back in late 1980s (Ekins, 1989), it was not until the development and success of DNA microarrays that protein microarrays gained attention, with substantial progress in recent years. Protein microarrays have emerged as a promising approach to meet the pressing need for systematic analysis of thousands of proteins in parallel. Protein microarrays provide a high-throughput, sensitive, and low-volume sample consumption platform for various assays, including the determination of the interactions between proteins and other biomolecules of interest, such as other proteins, antibodies, drugs, and various small ligands. Such assays are essential for basic biological research, for a better understanding of disease processes, and for the identification of novel diagnostic biomarkers and therapeutic targets.

This review addresses protein microarray applications, with a focus on clinical applications, particularly cancer. Additional information about protein microarrays may be found in several review articles that have been published in various journals (Cutler, 2003; Espina *et al*., 2004; Huels *et al*., 2002; Liotta *et al*., 2003; MacBeath, 2002; Templin *et al*., 2002). Furthermore, a recent issue of the journal *Proteomics* (November, 2003) was devoted to protein microarrays and contains both review articles and original research articles.

2. Technical overview – classes and detection methods

From a conceptual point of view, a protein microarray is an array of protein spots on solid supports, typically modified glass slides. Untreated glass slides bind protein poorly, and early on, protein microarrays relied on surface treatments established for DNA microarrays and histological techniques. Heterogeneous or homogeneous protein samples are immobilized on slides through noncovalent ionic interactions with a positively charged surface (such as with polylysine) (Haab *et al*., 2001) or covalent interaction between the primary amines (such as lysine on protein) with the aldehyde group on the slide surface (MacBeath and Schreiber,

2000). However, a need for optimizing the surface chemistries specifically for protein microarrays has been recognized to improve binding capacity, increase signal-to-noise ratio, prevent immobilized proteins from denaturing, and orient proteins properly. Various surfaces using specialized immobilization chemistries have been reported, including the use of nitrocellulose-coated slides to immobilize proteins, with reliance on hydrophobic interactions (Stillman and Tonkinson, 2000), immobilization of biotinylated proteins onto streptavidin-coated surface (Lesaicherre *et al*., 2002), and immobilization of His-tagged proteins onto Ni^{2+}-chelating surfaces (Zhu *et al*., 2001). The experience in the authors' laboratory has shown that glass slides coated with nitrocellulose membrane provided sufficient capacity and a good signal-to-noise ratio (Madoz-Gurpide *et al*., 2001).

There are generally two major classes of protein microarrays (Figure 1). One is intended for protein profiling in which multiple protein capture agents (usually antibodies) are spotted to assay the abundance of corresponding antigens or epitopes in a biological sample such as cell or tissue lysate or biological fluid (Figure 1a–c). Alternatively, multiple biological samples are spotted to assay for proteins that interact with a specific analyte in the biological samples applied to the array (Figure 1d). Another class is intended for functional protein analysis in which large numbers of purified proteins are spotted to study their biochemical properties (Figure 1e–f). Using such arrays, one could efficiently determine potential binding targets of a drug under development or substrates for a kinase of interest in a whole proteome.

The detection of binding in protein microarrays can be based on a sandwich immunoassay in which an immobilized antibody or any other type of capture agent is utilized to capture proteins of interest in a biological sample and a second labeled antibody or capture agent is used to detect and determine the abundance of the protein (Figure 1a). Such an approach, therefore, requires two capture agents that recognize different epitopes on proteins of interest. This dual-capture agent requirement creates constraints for interrogating the abundance of large numbers of proteins simultaneously in a biological sample because it is difficult to have access to high-quality matched antibody pairs.

An alternative to sandwich assays is to label the biological samples before they are hybridized to the arrays. The biological samples can be labeled with either a fluorophore for the direct detection method (Figure 1b) or a hapten tag such as biotin for the indirect detection method (Figure 1c). In the latter case, the arrays are further incubated with fluorophore-conjugated streptavidin. The levels of the proteins in the biological sample, for which corresponding capture agents are arrayed, is determined on the basis of quantifying the amount of fluorophore bound to the immobilized capture agents. Although it is possible to get an accurate measurement of the absolute concentration of proteins of interest in a biological sample using protein microarrays, often the relative concentration between two samples is determined as established for DNA arrays (Schena *et al*., 1996). In one of the earliest studies, 115 antibodies were printed on polylysine-coated slides (Haab *et al*., 2001). Different dilutions of the corresponding 115 antigens mixture were labeled with Cy5 and the reference mixture of the same 115 antigens was labeled with Cy3. Each Cy5-labeled antigen mixture and Cy3-labeled reference were applied to the antibody arrays simultaneously. The ratios of Cy5-to-Cy3 intensity

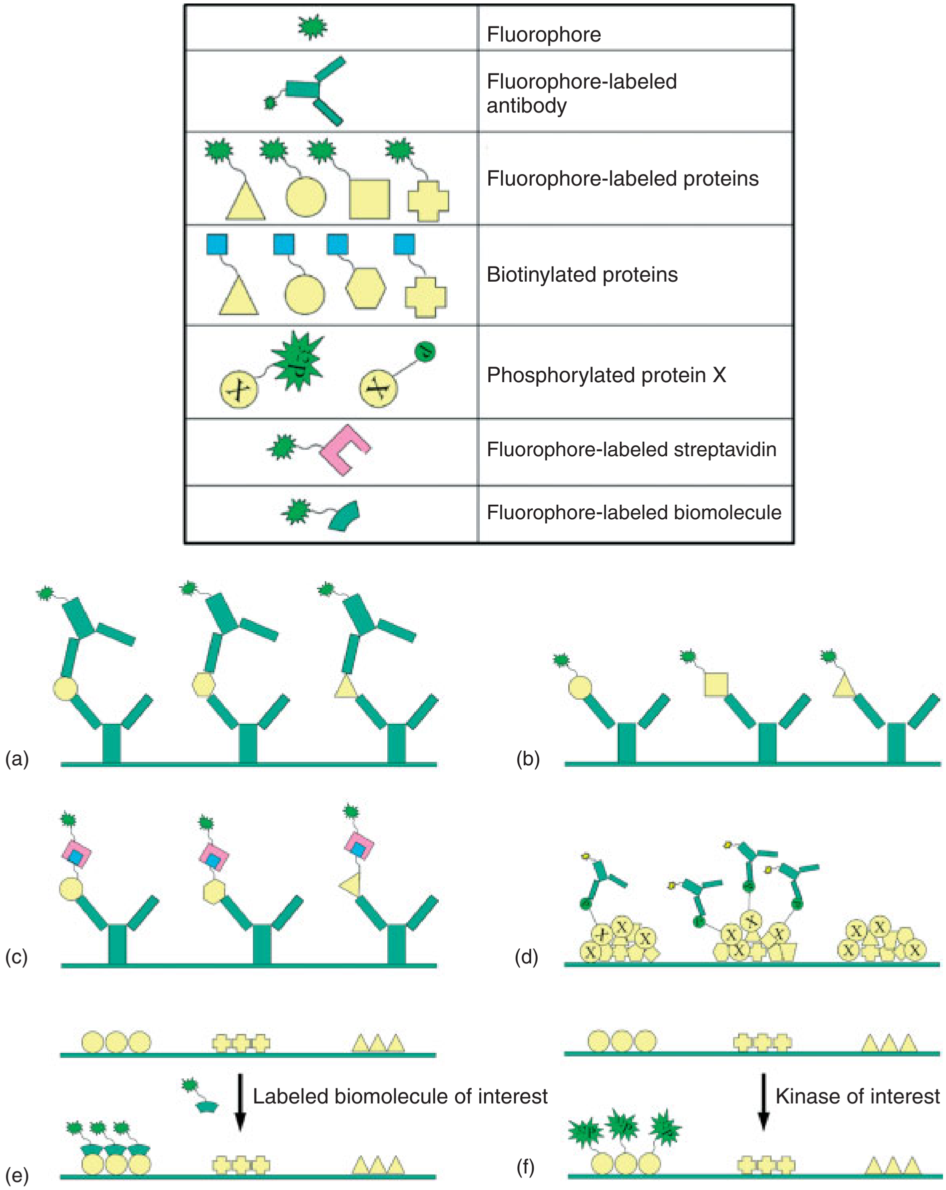


Figure 1 Different classes of protein microarrays. (a–d) Protein profiling microarrays. (a–c) Antibodies are spotted to assay the abundance of respective proteins in biological samples. (a) Sandwich assay detection; (b) direct detection; (c) indirect detection; (d) biological samples are spotted and probed with antiphosphorylated protein X to assay the abundance of phosphorylated protein X in each of the spotted biological samples. (e, f) Two examples of functional protein microarrays. Pure proteins are spotted on protein microarrays. (e) Labeled biomolecules, such as drugs in development or other proteins, are applied to the array to detect potential binding targets of the labeled biomolecules; (f) a kinase of interest is applied to the array to determine potential kinase substrates

at each spot for different antigen dilutions were plotted for each antibody–antigen pair to provide quantitative assessments of antibody–antigen interactions.

The success of protein microarrays relies on sensitive detection methods, as only nanoliter volumes of protein samples are printed for each spot. Tremendous efforts have been made to improve detection sensitivity without loss of resolution or increase in background. Schweitzer *et al*. adopted isothermal rolling-circle amplification (RCA) to protein microarrays to increase sensitivity (Schweitzer *et al*., 2002) and Haab *et al*. successfully implemented this method with the two-color approach (Zhou *et al*., 2004). An alternative is the tyramide signal amplification system that amplifies both fluorescent and chromogenic signals (Paweletz *et al*., 2001).

One concern with the labeling of biological samples is that the labeled biotin or fluorophore might interfere with protein interactions. Methodologies have been developed to detect binding of proteins to the arrays without any pretreatment of the biological samples using either mass spectrometry, surface plasmon resonance (SPR) (Rich *et al*., 2002), or atomic force microscope (AFM) (Lynch *et al*., 2004). These label-free approaches provide more flexibility to experimental design and preserve protein conformation in the biological samples that may be disrupted by labeling. However, these approaches are still at the proof-of-concept stage.

3. Applications

Clinical applications of protein microarrays cover a wide range from their use to detect disease or assess response to therapy based on profiling of biological fluids, to their use for profiling of disease tissue to determine disease subtypes and decide on the most appropriate therapy.

Belov *et al*. utilized a microarray of 60 antibodies directed against a cluster of differentiation (CD) antigens to immunophenotype various types of leukemia (Belov *et al*., 2001; Belov *et al*., 2003). Whole cells were applied to the array to determine expression of CD antigens on the cell surface. Distinctive expression patterns of CD antigens were observed for normal leukocytes and for different types of leukemia. The microarray results compared well with data from established flow cytometry techniques.

As a model to better understand how proteins shape the tissue microenvironment, Knezevic *et al*. analyzed protein expression in tissue derived from squamous cell carcinomas of the oral cavity with an antibody microarray approach (Knezevic *et al*., 2001). Utilizing laser capture microdissection to procure total protein from specific cell populations, they demonstrated that quantitative, and potentially qualitative, differences in expression patterns of multiple proteins within epithelial cells reproducibly correlated with oral cavity tumor progression. Differential expression of multiple proteins was found in stromal cells surrounding and adjacent to regions of diseased epithelium that directly correlated with tumor progression of the epithelium. Most of the proteins identified in both cell types were involved in signal transduction pathways. They hypothesized, therefore, that extensive molecular communications involving complex cellular signaling between epithelium and stroma play a key role in driving oral cavity cancer progression.

Cytokines have become a popular target for analysis using microarrays. In one study, Schweitzer *et al*. printed 51 cytokine antibodies on chemically derivatized glass slides and investigated cytokine secretion from human dendritic cells induced with lipopolysaccharide or tumor-necrosis factor-α (Schweitzer *et al*., 2002). Signal amplification by RCA achieved high specificity, femtomolar sensitivity, 3 log quantitative range, and economy of sample consumption.

There is substantial interest in assembling panels of markers that are simultaneously assayed to allow disease diagnosis or monitoring. Haab *et al*. assembled protein microarrays with antibodies that target 184 distinct proteins to profile prostate cancer serum samples for potential biomarkers (Miller *et al*., 2003). From the profiles of 33 prostate cancer and 20 control serum samples using a two-color fluorescence assay, five proteins (von Willebrand Factor, immunoglobulinM, Alpha1-antichymotrypsin, Villin, and immunoglobulinG) had significantly different levels between the prostate cancer samples and the controls. However, the specificity of any of the reactive proteins for prostate cancer was not assessed. The mining of gene expression data for genes with a distinctive expression pattern in a particular cancer may be a useful source to identify potential serum markers for which antibodies are produced and tested with protein microarrays.

A complementary approach for cancer marker identification consists of the analysis of serum for autoantibodies against tumor proteins. There is increasing evidence for an immune response to cancer in humans, demonstrated in part by the identification of autoantibodies against a number of intracellular and surface antigens detectable in sera from patients with different cancer types (Hanash, 2003). The identification of panels of tumor antigens that elicit an antibody response may have utility in cancer screening, diagnosis, or in establishing prognosis. Such antigens may also have utility in immunotherapy against the disease. There are several approaches for the detection of tumor antigens that induce an immune response. For most antigenic proteins that induce an antibody response in cancer identified using proteomics, posttranslational modifications contributed to the immune response. For example, in a study of lung cancer, sera from 60% of patients with lung adenocarcinoma and 33% of patients with squamous cell lung carcinoma, but none of the noncancer controls, exhibited IgG-based reactivity against proteins identified as glycosylated annexins I and II (Brichory *et al*., 2001). However, these modifications are generally not captured using either recombinant proteins or antibodies that do not distinctly recognize specific forms of a protein. One approach for comprehensive analysis of proteins in their modified forms is to array proteins directly isolated from cells and tissues following protein fractionation schemes. Proteins in reactive fractions are identified by mass spectrometry.

Microarrays that contain natural proteins derived from tumor cells have the potential to substantially accelerate the pace of discovery of tumor antigens to yield a molecular signature for immune responses directed against protein targets in different types of cancer. In a study of lung cancer, protein lysates from the A549 human lung adenocarcinoma cell line were separated into 1840 fractions that were spotted on nitrocellulose-coated slides (Qiu *et al*., 2004). Sera from lung cancer patients and healthy controls were each hybridized to an individual microarray (Figure 2). A total of 63 of the 1840 arrayed fractions demonstrated

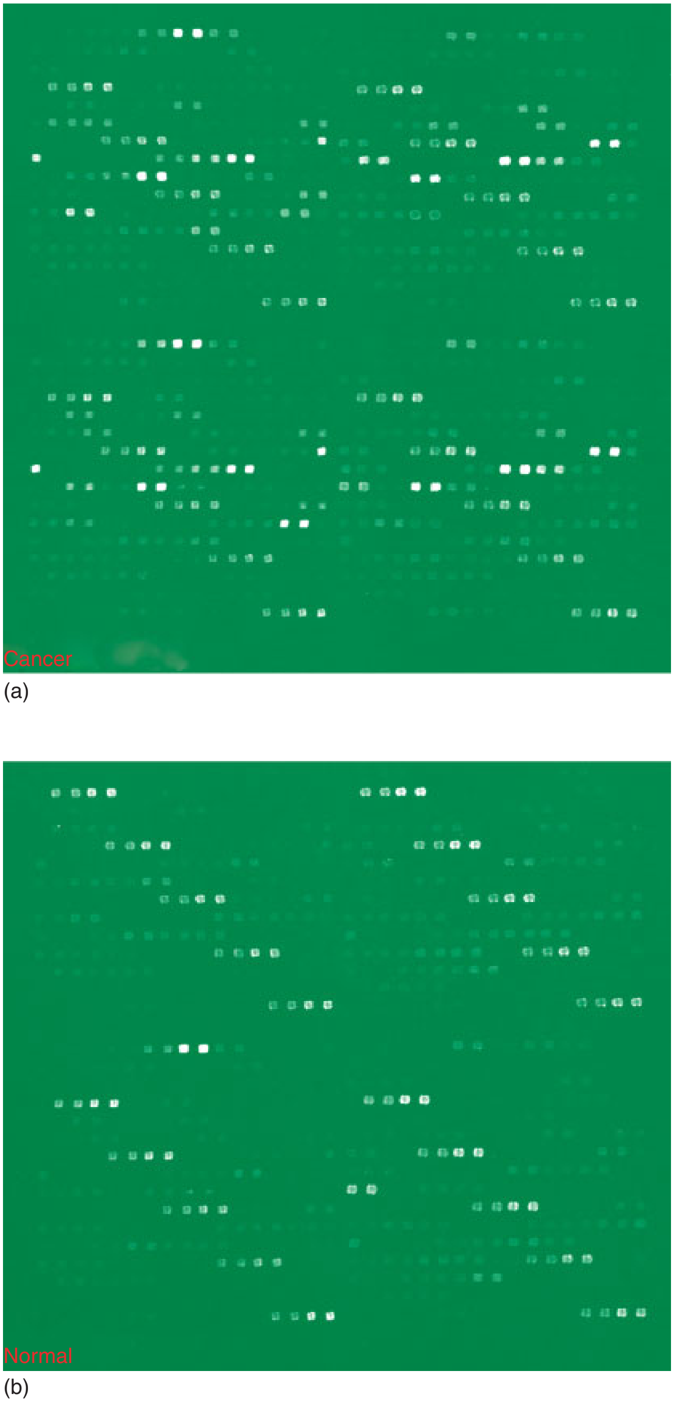


Figure 2 Natural protein microarray images. Whole-cell lysates from A549 lung cancer cell line were separated into 1840 fractions and spotted onto nitrocellulose-coated slides. (a) was hybridized with a lung cancer patient serum sample and (b) a normal serum sample

increased reactivity in cancer patients relative to controls as measured by a rank-based statistic ($p < 0.008$). In a study of colon cancer, microarrays printed with 1760 distinct protein fractions, prepared from the LoVo colon adenocarcinoma cell line, were hybridized with individual sera (Nam *et al*., 2003). A fraction that exhibited IgG-based reactivity with 9/15 colon cancer sera was found to contain Ubiquitin C-terminal hydrolase L3 (UCH-L3) by tandem mass spectrometry (ESI-Q-TOF). The highest levels of UCH-L3 mRNA among the 329 tumors of different types analyzed by DNA microarrays were found in colon tumors. Independent validation by Western blotting demonstrated that UCH-L3 antibodies existed in 19/43 sera from patients with colon cancer, and in 0/54 sera from subjects with lung cancer, colon adenoma or otherwise healthy. These data point to the utility of microarrays printed with natural proteins for the identification of tumor antigens that have induced an antibody response in patients with specific cancers.

Another clinically relevant application of protein microarrays is the identification of proteins that induce an antibody response in autoimmune disorders (Hueber *et al*., 2002; Robinson *et al*., 2002). Microarrays were produced by attaching several hundred proteins and peptides to the surface of derivatized glass slides. Arrays were incubated with patient sera, and fluorescent labels were used to detect autoantibody binding to specific proteins in autoimmune diseases, including systemic lupus erythematosus and rheumatoid arthritis. In a more recent study (Robinson *et al*., 2003), a "myelin proteome" microarray was developed to profile the evolution of autoantibody responses in experimental autoimmune encephalomyelitis, a model for multiple sclerosis. Increased diversity of autoantibody responses predicted a more severe clinical course. Chronic experimental autoimmune encephalomyelitis was associated with previously undescribed extensive intra- and intermolecular epitope spreading of autoreactive B-cell responses. Proteomic monitoring of autoantibody responses provided a useful approach to monitor autoimmune disease and to develop and tailor disease- and patient-specific tolerizing DNA vaccines. Protein microarrays have also been used to predict resistance or susceptibility to future autoimmune disease from the present antoantibody repertoire (Quintana *et al*., 2004). Reaction patterns of mouse sera obtained before cyclophosphamide treatment, with 27 different antigens (out of 266 antigens printed on the microarrays), could predict whether a particular mouse would develop cyclophosphamide-accelerated diabetes.

Unlike DNA microarrays, which provide only one measure of gene expression, namely RNA levels, many different features of proteins may be addressed by different types of protein microarrays, including, on the one hand, determination of protein abundance in biological samples, and on the other, determination of their functional states, as may be deduced, for example, from their extent of phosphorylation. Many cellular processes are regulated by reversible protein phosphorylation (Cohen, 2002; Eckhart *et al*., 1979). Genetic changes related to oncogenes or tumor suppressor genes that lead to cancer development also result in alterations of protein phosphorylation patterns (Vogelstein and Kinzler, 2004). A systematic profiling of functional states of proteins involved in signal transduction pathway provides insight into disease development and treatment. The development of reagents that allow assessment of protein modification such as phosphorylation, glycosylation, or other functionally relevant protein changes has substantial utility

for clinical investigations. Therefore, protein microarrays using antibodies specific to phosphorylated protein have been the focus of a number of recent studies.

Paweletz *et al*. arrayed whole-protein lysates prepared from multiple microdissected cancer samples (Paweletz *et al*., 2001). Each array was interrogated with a single antibody to determine the level or posttranslational modification of a particular protein in the arrayed lysates. A high degree of sensitivity, precision, and linearity was demonstrated, making it possible to quantify the phosphorylation state of signaling proteins in human tissue lysates. Using this approach, Paweletz *et al*. analyzed pro-survival checkpoint proteins at the transition stage from patient-matched histologically normal prostate epithelium to prostate intraepithelial neoplasia and to invasive prostate cancer. Cancer progression was associated with increased phosphorylation of Akt, suppression of apoptosis pathways, as well as decreased phosphorylation of ERK. At the transition from histologically normal epithelium to intraepithelial neoplasia, a statistically significant increase in phosphorylated Akt and a concomitant suppression of downstream apoptosis pathways preceding the transition into invasive carcinoma were observed.

Nielsen *et al*. profiled the ErbB receptor typrosine kinase in human tumor cell lines (Nielsen *et al*., 2003). They demonstrated that antibody arrays accurately profiled the functional state of the ErbB signaling pathway, not just ErbB abundance. Data obtained from microarrays correlated very well with data obtained using traditional approaches. Only a handful of antibodies were used for this prototype study. Scaled up studies based on the availability of capture agents would permit the systematic analysis of cellular signal transduction networks in cancer.

Antibody microarrays with a relatively limited content, compared to DNA microarrays, have become commercially available. For example, the BD Clontech Ab Microarray 500 consists of 500 distinct antibodies that detect a variety of human proteins either in circulation or in cell and tissue extracts. The antibody set targets a broad range of functional classes involved in signal transduction, cell growth and proliferation, apoptosis, inflammation, and the immune response. A complete list of the antibodies on this array is available at **http://bioinfo2.clontech.com/abinfo/abinfo/array-list-action.do**. Other commercially available antibody microarrays that interrogate a more limited or a particular class of proteins are also available. The content and variety of commercially available microarrays are likely to continue to expand. A potential limiting factor for the use of commercial microarrays is their cost.

Traditionally, protein biochemical properties are assayed one at a time. However, protein microarrays provide a platform that protein functions can be assayed at proteome scale with the advantages of low reagent consumption, rapid interpretation of results, and the ability to easily control experimental conditions. For this application, microarrays are constructed with purified recombinant proteins encoded by the whole proteome or subproteome of interest and are used to analyze protein biochemical activities, protein–protein interactions, or protein–small molecule interactions, as has been successfully demonstrated with the yeast whole-proteome microarray (Michaud *et al*., 2003; Zhu *et al*., 2001; Zhu *et al*., 2000). For reviews on functional protein microarrays, see Schweitzer *et al*. (2003) and Wilson and Nock (2002).

4. Challenges and perspectives

There is increased appreciation of the merits of protein microarrays given their multiplexing nature and high sensitivity. The number of applications using protein microarrays is expanding rapidly, especially in clinical research in view of the limited amount of samples available from human subjects. However, substantial challenges still remain, both for strategies that rely on arrayed capture agents and for strategies that rely on arrayed proteins. Currently, antibodies are widely used as the capture agents for protein microarrays because they can provide excellent specificity. However, the limited availability of such antibodies with the requisite specificity limits the coverage that can be achieved at the present time with protein microarrays. The development of antibodies is a laborious and costly process. The challenge is to devise strategies to produce highly specific antibodies at a proteome scale. The compelling need for capture agents with the prerequisite specificity has led numerous biotechnology companies and research labs to devise novel strategies to replace antibodies that include aptamers (SomaLogic, **http://www.somalogic.com/**), phage display (Dyax, **http://www.dyax.com**), ribozymes (Archemix, **http://www.archemix.com/**), and partial-molecule imprints (Aspira Biosystems, **http://www.aspirabio.com**). Additionally, in principle, when used in combination, some combinatorial strategies are being considered in which a limited repertoire of capture agents that target epitopes shared by sets of proteins such as protein families may be devised that could provide proteome-scale coverage. Such strategies may yield distinctive patterns of reactivity that may be clinically or biologically informative but would have difficulties in identifying the specific reactive protein(s).

For strategies that rely on arrayed proteins, there is a substantial challenge in producing proteins at a proteome scale. *Escherichia coli* has been the dominant expression system to produce recombinant proteins for functional protein microarrays. Some use has been made of synthetic peptides, for limited applications in which protein conformation is not of concern, such as linear epitope mapping (Pellois *et al*., 2002; Reimer *et al*., 2002). Eukaryotic proteins expressed in *E. coli* usually fail to fold into their native conformations and be posttranslationally modified, as occurs in the tissues in which they are originally expressed. Therefore, data obtained from microarrays arrayed with bacterially expressed recombinant proteins are not fully informative. Consequently, there is a substantial need to develop alternative expression systems to produce pure proteins with proper folding and posttranslational modification. High-throughput protein production in some eukaryotic expression systems that rely on yeast, insect, or mammalian cells is currently being explored (Braun and LaBaer, 2003; Gilbert *et al*., 2004).

It is expected that the reach of protein microarrays will continue to expand gradually as resources are developed. Eventually, the use of such arrays will become widespread, thus contributing substantially to proteome research and diagnostic use.

References

Belov L, de la Vega O, dos Remedios CG, Mulligan SP and Christopherson RI (2001) Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Research*, **61**, 4483–4489.

Belov L, Huang P, Barber N, Mulligan SP and Christopherson RI (2003) Identification of repertoires of surface antigens on leukemias using an antibody microarray. *Proteomics*, **3**, 2147–2154.

Braun P and LaBaer J (2003) High throughput protein production for functional proteomics. *Trends in Biotechnology*, **21**, 383–388.

Brichory FM, Misek DE, Yim AM, Krause MC, Giordano TJ, Beer DG and Hanash SM (2001) An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9824–9829.

Cohen P (2002) The origins of protein phosphorylation. *Nature Cell Biology*, **4**, E127–E130.

Cutler P (2003) Protein arrays: the current state-of-the-art. *Proteomics*, **3**, 3–18.

Eckhart W, Hutchinson MA and Hunter T (1979) An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. *Cell*, **18**, 925–933.

Ekins RP (1989) Multi-analyte immunoassay. *Journal of Pharmaceutical and Biomedical Analysis*, **7**, 155–168.

Espina V, Woodhouse EC, Wulfkuhle J, Asmussen HD, Petricoin EF III and Liotta LA (2004) Protein microarray detection strategies: focus on direct detection technologies. *Journal of Immunological Methods*, **290**, 121–133.

Gilbert M, Edwards TC and Albala JS (2004) Protein expression arrays for proteomics. *Methods in Molecular Biology*, **264**, 15–23.

Haab BB, Dunham MJ and Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**, RESEARCH0004.

Hanash S (2003) Harnessing immunity for cancer marker discovery. *Nature Biotechnology*, **21**, 37–38.

Hueber W, Utz PJ, Steinman L and Robinson WH (2002) Autoantibody profiling for the study and treatment of autoimmune disease. *Arthritis Research*, **4**, 290–295.

Huels C, Muellner S, Meyer HE and Cahill DJ (2002) The impact of protein biochips and microarrays on the drug development process. *Drug Discovery Today*, **7**, S119–S124.

Knezevic V, Leethanakul C, Bichsel VE, Worth JM, Prabhu VV, Gutkind JS, Liotta LA, Munson PJ, Petricoin EF III and Krizman DB (2001) Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics*, **1**, 1271–1278.

Lesaicherre ML, Lue RY, Chen GY, Zhu Q and Yao SQ (2002) Intein-mediated biotinylation of proteins and its application in a protein microarray. *Journal of the American Chemical Society*, **124**, 8768–8769.

Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J and Petricoin EF III (2003) Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell*, **3**, 317–325.

Lynch M, Mosher C, Huff J, Nettikadan S, Johnson J and Henderson E (2004) Functional protein nanoarrays for biomarker profiling. *Proteomics*, **4**, 1695–1702.

MacBeath G (2002) Protein microarrays and proteomics. *Nature Genetics*, **32**(Suppl 2), 526–532.

MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Madoz-Gurpide J, Wang H, Misek DE, Brichory F and Hanash SM (2001) Protein based microarrays: a tool for probing the proteome of cancer cells and tissues. *Proteomics*, **1**, 1279–1287.

Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, Snyder M, Predki PF and Schweitzer BI (2003) Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnology*, **21**, 1509–1512.

Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Butler EB, Teh BS and Haab BB (2003) Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. *Proteomics*, **3**, 56–63.

Nam MJ, Madoz-Gurpide J, Wang H, Lescure P, Schmalbach CE, Zhao R, Misek DE, Kuick R, Brenner DE and Hanash SM (2003) Molecular profiling of the immune response in colon cancer using protein microarrays: occurrence of autoantibodies to ubiquitin C-terminal hydrolase L3. *Proteomics*, **3**, 2108–2115.

Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G and Sorger PK (2003) Profiling receptor tyrosine kinase activation by using Ab microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9330–9335.

Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, Petricoin IE and Liotta LA (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Pellois JP, Zhou X, Srivannavit O, Zhou T, Gulari E and Gao X (2002) Individually addressable parallel peptide synthesis on microchips. *Nature Biotechnology*, **20**, 922–926.

Qiu J, Madoz-Gurpide J, Misek DE, Kuick R, Brenner DE, Michailidis G, Haab BB, Omenn GS and Hanash S (2004) Development of natural protein microarrays for diagnosing cancer based on an antibody response to tumor antigens. *Journal of Proteome Research*, **3**, 261–267.

Quintana FJ, Hagedorn PH, Elizur G, Merbl Y, Domany E and Cohen IR (2004) Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(Suppl 2), 14615–14621.

Reimer U, Reineke U and Schneider-Mergener J (2002) Peptide arrays: from macro to micro. *Current Opinion in Biotechnology*, **13**, 315–320.

Rich RL, Hoth LR, Geoghegan KF, Brown TA, LeMotte PK, Simons SP, Hensley P and Myszka DG (2002) Kinetic analysis of estrogen receptor/ligand interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 8562–8567.

Robinson WH, DiGennaro C, Hueber W, Haab BB, Kamachi M, Dean EJ, Fournel S, Fong D, Genovese MC, de Vegvar HE, *et al*. (2002) Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Medicine*, **8**, 295–301.

Robinson WH, Steinman L and Utz PJ (2003) Protein arrays for autoantibody profiling and fine-specificity mapping. *Proteomics*, **3**, 2077–2084.

Schena M, Shalon D, Heller R, Chai A, Brown PO and Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10614–10619.

Schweitzer B, Predki P and Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. *Proteomics*, **3**, 2190–2199.

Schweitzer B, Roberts S, Grimwade B, Shao W, Wang M, Fu Q, Shu Q, Laroche I, Zhou Z, Tchernev VT, *et al*. (2002) Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nature Biotechnology*, **20**, 359–365.

Stillman BA and Tonkinson JL (2000) FAST slides: a novel surface for microarrays. *Biotechniques*, **29**, 630–635.

Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF and Joos TO (2002) Protein microarray technology. *Drug Discovery Today*, **7**, 815–822.

Vogelstein B and Kinzler KW (2004) Cancer genes and the pathways they control. *Nature Medicine*, **10**, 789–799.

Wilson DS and Nock S (2002) Functional protein microarrays. *Current Opinion in Chemical Biology*, **6**, 81–85.

Zhou H, Bouwman K, Schotanus M, Verweij C, Marrero JA, Dillon D, Costa J, Lizardi P and Haab BB (2004) Two-color, rolling-circle amplification on antibody microarrays for sensitive, multiplexed serum-protein measurements. *Genome Biology*, **5**, R28.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA and Snyder M (2000) Analysis of yeast protein kinases using protein chips. *Nature Genetics*, **26**, 283–289.

Short Specialist Review

Tissue microarrays

Ronald Simon, Martina Mirlacher and Guido Sauter
University Medical Center Hamburg-Eppendorf, Hamburg, Germany

1. Introduction

A variety of powerful high-throughput technologies are available for the identification of potentially relevant genes in individual tissue samples. These tools generally identify more interesting genes than the resources that are available for their further analysis. In situ analyses methods are optimally suited for studying gene alterations. Technologies such as immunohistochemistry (IHC), fluorescence in situ hybridization (FISH), and to some extent also RNA in situ hybridization (RNA ISH) allow a cellular and subcellular localization of the gene product of interest and a clear distinction between diseased and nondiseased cells. However, traditional slide-by-slide analysis cannot match the speed of modern high-throughput target-identification technologies. Tissue microarrays (TMAs) permit in situ analysis of hundreds of tissue samples on one microscope glass slide. Since studies have shown that minute tissue samples as small as 0.6 mm in diameter are sufficiently representative of their donor tissues (reviewed in Sauter *et al*., 2003), TMAs have become a widely used research tool.

2. TMA technology

Hundreds of cylindrical samples measuring up to 4 mm (usually 0.6 mm) in diameter are removed from formalin-fixed or frozen tissues and placed into one recipient block (Kononen *et al*., 1998). Simple commercial or homemade arrayers can be used for this purpose. Sections from such TMA blocks containing hundreds of different tissues are then placed on a glass slide. All methods for in situ tissue analysis can be performed on TMAs using similar protocols as for conventional large sections (Figure 1).

3. TMA applications

After its initial description in 1998 (Kononen *et al*., 1998), TMAs have become widely used in recent years. More than 150 publications published in 2003 have utilized TMAs. Some of the most powerful applications of TMAs are described below.

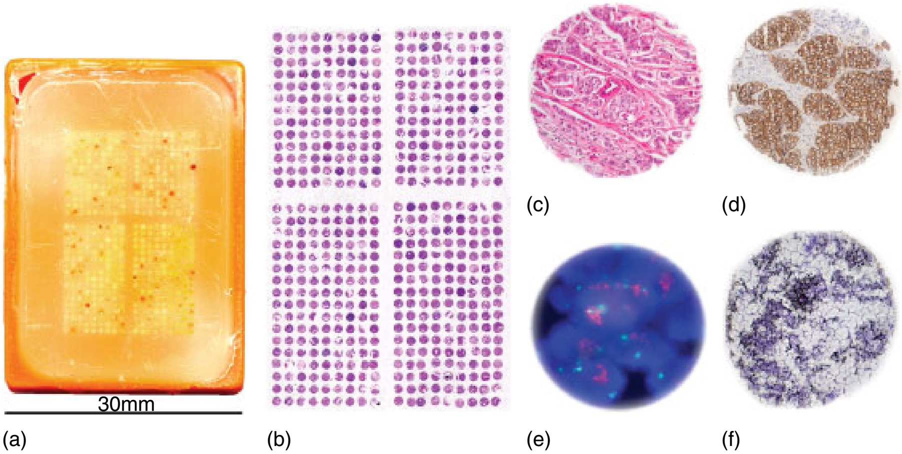


Figure 1 Tissue microarrays and TMA applications. (a) TMA block; (b) hematoxylin and Eosin (H&E)-stained section (5 µm) of the TMA; (c) magnification of a H&E-stained tissue spot (diameter 0.6 mm); (d) immunohistochemistry of a breast cancer tissue spot. Brownish membranous staining indicates strong expression of the Her2 receptor protein. (e) FISH analysis. Magnification (1000x) of cell nuclei (blue staining) showing two green signals corresponding to the centromere of chromosome 17 and multiple red signals indicating *Her2* gene amplification. (f) RNA in situ hybridization on a TMA spot from frozen tissue using an radioactively labeled oligonucleotide probe

3.1. TMAs for molecular epidemiology

Data on the expression of a given gene are often available from the literature. However, examples of extensively analyzed genes suggest that published data often lack practical value. For example, *HER2*, the target gene for Trastuzumab (Herceptin) has been analyzed in thousands of publications. Yet, the data are highly controversial. Published expression frequencies range from <10 to >90% for many important tumor types (reviewed in Sauter *et al*., 2003). The paramount impact of experimental parameters such as antibody selection, IHC protocol, tissue characteristics, and scoring criteria on the results of IHC analyses greatly reduce the comparability of IHC data derived from different groups. TMAs containing samples from all different tumor types make it possible to analyze a gene of interest under fully standardized conditions in one experiment. For example, we used a set of TMAs containing more than 3500 different samples from more than 120 different tumor categories to investigate target genes for established drug targets such as KIT/CD117 (Imatinib; Glivec) (Went *et al*., 2004) or epidermal growth factor (EGFR) (Tarceva, Iressa, Cetuximab, Erbitux) (Sauter *et al*., 2003). Such studies often show expression of a target gene in a wide variety of tumor entities. The analysis of a representative number of cases from various tumor types typically results in a representative ranking list of marker expression in different tumor entities (Sauter *et al*., 2003).

3.2. TMAs for prognosis evaluation

TMAs containing a large number of samples from one tumor type with attached clinico-pathological information allow an estimation of the biological importance

of a gene alteration. Associations with advanced stage, high grade, presence of metastases, or poor clinical outcome argue for a role in tumor progression. TMAs are highly efficient in identifying associations between molecular features and prognosis. For example, significant associations were found between estrogen or progesteron expression (Torhorst *et al*., 2001) or HER2 alterations (Barlund *et al*., 2000) and survival in breast cancer patients, between vimentin expression and prognosis in kidney cancer (Moch *et al*., 1999), and between Ki67 labeling index and prognosis in urinary bladder cancer (Nocito *et al*., 2001), soft tissue sarcoma (Hoos *et al*., 2001a), or Hurthle cell carcinoma (Hoos *et al*., 2002).

3.3. TMAs for normal tissue evaluation

The expression pattern of genes of interest in normal tissues is often important. Often it allows conclusions on the potential biological role of a gene product. Normal tissue analysis is critically important for potential drug targets. Expression of a target gene in important normal cell types may predict the site of potential side effects. Cell type–specific expression analysis would not be possible if other methods than in situ techniques were used. For example, EpCam, a target for several anticancer therapies, is expressed in bile ducts of the liver. Since bile ducts constitutes are a very small (<1%) component of the liver, EpCam analyses by non–in situ techniques would suggest either no or low expression in liver (Figure 2). IHC analysis, however, reveals high EpCam levels in a small but vital liver compartment.

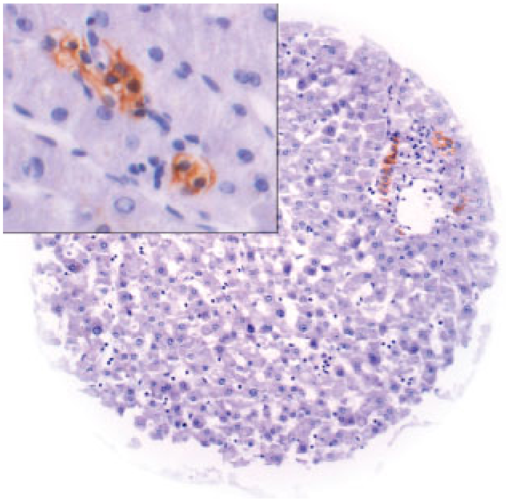


Figure 2 EpCam expression in a tissue spot from normal liver. Inset: 1000x magnification. EpCam expression if confined to small areas of bile ducts. A similar staining intensity, however, can be seen as well in cancer cells. This emphasizes the importance of in situ analyses for normal tissue evaluation. The use of disaggregated tissue for such an expression comparison (like Northern or Western blots) would suggest a negligible low expression level in normal liver

3.4. *Important technical aspects*

Although the technology is simple and easily applicable, some aspects of TMAs are controversially discussed and others are underestimated. Understanding the impact of tissue heterogeneity, the potential for automation, and the importance of pathology expertise is particularly important for the successful use of TMAs.

4. Tissue heterogeneity

The utility of TMAs for reliably finding associations between molecular and clinicopathological features is the breakthrough discovery of the TMA field. Placing multiple tissues into one paraffin block has been described before (Wan *et al*., 1987). At this time, it was found that molecular analysis of small tissue samples would be insufficient for the identification of clinically relevant information. Technical improvements, therefore, aimed at the deposition of as large as possible tissues on multitissue blocks (Battifora, 1986). Remarkably, all studies comparing molecular features obtained on TMAs (with a spot diameter of 0.6 mm) with clinical or pathological data found the expected associations if these were previously well established (reviewed in Sauter *et al*., 2003). Results of these studies considerably weaken the practical importance of studies showing an increasing concordance between TMA data and corresponding large section results if 3–4 tissue samples are analyzed per tumor (Rubin *et al*., 2002; Engellau *et al*., 2001; Camp *et al*., 2000; Fernebro *et al*., 2002; Hoos *et al*., 2001b). It is obvious that TMA results will match large section results more closely if the highest possible number of cores per tumor block is analyzed. However, as shown in a p53 analysis in breast cancer, this has also disadvantages. Notably, the risk of false positivity is also increasing with the size of analyzed tissue. Torhorst *et al*. (2001) found p53 positivity to be prognostically relevant in four independent TMAs composed of one core each of >500 breast cancers but not in corresponding large sections. The use of TMAs containing only one core per tumor is therefore recommended for the vast majority of applications.

5. Pathology expertise and automation

Although TMA technology is often discussed in the context of other array techniques such as DNA or protein arrays, for which array construction, analysis, and data recording can easily be automated, the TMA method is fundamentally different. TMAs represent the ultimate miniaturization of molecular pathology and share the inherent strengths and limitations of pathology analyses. The level of pathology expertise is most decisive for the success of TMA studies, while classical array tools like automation are less relevant. Automated TMA manufacturing lacks importance, because sufficiently large ready-to-use "tissue libraries" are lacking.

Classical pathology skills are critical for the classification of arrayed tissues, experimental design, and staining interpretation. More than 100 studies can be

based on one TMA block. As the molecular data will be compared with pathological information, the quality of "basic pathology data" is of highest importance. The difficulties related to the development of an adequate IHC protocol are greatly underestimated. Nonspecific positivity is a frequent problem and requires the use of multiple controls. A significant fraction of antibodies cannot be optimized for use on tissue sections. In general, IHC staining results are greatly dependent on antibody selection, antigen retrieval strategy, staining protocol, and on minor variables such as the section age (Mirlacher *et al*., 2004). For example, the use of three different antibodies for EGFR resulted – after state-of-the-art protocol development – in an up to fivefold difference in the rate of positivity (Sauter *et al*., 2003).

These inherent difficulties reduce the importance of automated TMA analyses. Although studies have demonstrated the feasibility of automated quantitation of brightfield or fluorescence signals on TMAs, automated systems have limited capability for recognition of staining artifacts, necrotic, crushed or damaged tissue elements, or for distinction of neoplastic versus nonneoplastic cells. Nevertheless, modern imaging tools have high value for image documentation and preliminary screening of staining results in a high-throughput scale. At the same time, manual (visual) analysis of the TMA staining by an experienced pathologist provides the greatest level of reading precision and will remain the gold standard for TMA staining interpretation. Attempts to improve automated TMA analysis are paralleled by an increased use of lysate arrays composed of protein extracts from tumor tissues (Nath and Chilkoti, 2003; Nishizuka *et al*., 2003). These are spotted on glass slides in a DNA array-like manner and can be analyzed with the same software used for DNA arrays. If morphologic information is not needed, lysate arrays may represent a better option than complicated and expensive systems for automated TMA analysis.

References

Barlund M, Forozan F, Kononen J, Bubendorf L, Chen Y, Bittner ML, Torhorst J, Haas P, Bucher C, Sauter G, *et al*. (2000) Detecting activation of ribosomal protein S6 kinase by complementary DNA and tissue microarray analysis. *Journal of the National Cancer Institute*, **92**, 1252–1259.

Battifora H (1986) The multitumor (sausage) tissue block: novel method for immunohistochemical antibody testing. *Laboratory Investigation*, **55**, 244–248.

Camp RL, Charette LA and Rimm DL (2000) Validation of tissue microarray technology in breast carcinoma. *Laboratory Investigation*, **80**, 1943–1949.

Engellau J, Akerman M, Anderson H, Domanski HA, Rambech E, Alvegard TA and Nilbert M (2001) Tissue microarray technique in soft tissue sarcoma: immunohistochemical Ki-67 expression in malignant fibrous histiocytoma. *Applied Immunohistochemistry & Molecular Morphology: AIMM*, **9**, 358–363.

Fernebro E, Dictor M, Bendahl PO, Ferno M and Nilbert M (2002) Evaluation of the tissue microarray technique for immunohistochemical analysis in rectal cancer. *Archives of Pathology & Laboratory Medicine*, **126**, 702–705.

Hoos A, Stojadinovic A, Mastorides S, Urist MJ, Polsky D, Di Como CJ, Brennan MF and Cordon-Cardo C (2001a) High Ki-67 proliferative index predicts disease specific survival in patients with high-risk soft tissue sarcomas. *Cancer*, **92**, 869–874.

Hoos A, Urist MJ, Stojadinovic A, Mastorides S, Dudas ME, Leung DH, Kuo D, Brennan MF, Lewis JJ and Cordon-Cardo C (2001b) Validation of tissue microarrays for

immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. *American Journal of Pathology*, **158**, 1245–1251.

Hoos A, Stojadinovic A, Singh B, Dudas ME, Leung DH, Shaha AR, Shah JP, Brennan MF, Cordon-Cardo C and Ghossein R (2002) Clinical significance of molecular expression profiles of Hurthle cell tumors of the thyroid gland analyzed via tissue microarrays. *American Journal of Pathology*, **160**, 175–183.

Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G and Kallioniemi OP (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, **4**, 844–847.

Mirlacher M, Kasper M, Storz M, Knecht Y, Durmuller U, Simon R, Mihatsch MJ and Sauter G (2004) Influence of slide aging on results of translational research studies using immunohistochemistry. *Modern Pathology* **17**, 1414–1420.

Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP and Sauter G (1999) High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *American Journal of Pathology*, **154**, 981–986.

Nath N and Chilkoti A (2003) Fabrication of a reversible protein array directly from cell lysate using a stimuli-responsive polypeptide. *Analytical Chemistry*, **75**, 709–715.

Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Espina V, *et al*. (2003) Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14229–14234.

Nocito A, Bubendorf L, Tinner EM, Suess K, Wagner U, Forster T, Kononen J, Fijan A, Bruderer J, Schmid U, *et al*. (2001) Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. *The Journal of Pathology*, **194**, 349–357.

Rubin MA, Dunn R, Strawderman M and Pienta KJ (2002) Tissue microarray sampling strategy for prostate cancer biomarker analysis. *The American Journal of Surgical Pathology*, **26**, 312–319.

Sauter G, Simon R and Hillan K (2003) Tissue microarrays in drug discovery. *Nature Reviews Drug discovery*, **2**, 962–972.

Torhorst J, Bucher C, Kononen J, Haas P, Zuber M, Kochli OR, Mross F, Dieterich H, Moch H, Mihatsch M, *et al*. (2001) Tissue microarrays for rapid linking of molecular changes to clinical endpoints. *American Journal of Pathology*, **159**, 2249–2256.

Wan WH, Fortuna MB and Furmanski P (1987) A rapid and efficient method for testing immunohistochemical reactivity of monoclonal antibodies against multiple tissue samples simultaneously. *Journal of Immunological Methods*, **103**, 121–129.

Went P, Dirnhofer S, Bundi M, Mirlacher M, Schraml P, Mangialaio S, Dimitrijevic S, Kononen J, Simon R and Sauter G (2004) Prevalence of KIT expression in human tumors. *Journal of Clinical Oncology*, **22**, 4514–4522.

The use of external controls in microarray experiments

Joop M.L.M. van Helvoort, Dik van Leenen and Frank C.P. Holstege

University Medical Center Utrecht, Utrecht, The Netherlands

1. Introduction

Expression profiling using DNA microarrays results in either ratios of mRNA expression levels or in an absolute measurement that is correlated to the expression level. This chapter deals with how the accuracy of such experiments can best be evaluated using external experimental controls where the accuracy refers to how close the observed value is to the real biological value. Generally speaking, there are at least two different scenarios that require a good assessment of microarray accuracy. The first scenario is determining the accuracy in order to decide which technology works best. For example, do the microarrays from one particular supplier work as well as those from other sources? Or, how will various sample handling and labeling strategies affect the accuracy of a microarray experiment? The question of accuracy can of course be posed for the refinement of every step of the experimental procedure or production process, whether it is the use of different oligomer lengths for manufacture, assessment of different hybridization protocols, or even assessment of different data-processing strategies. A second scenario that requires assessment of accuracy is at the level of every individual assay, in order to check that each hybridization gives results of similar accuracy, that is, after the optimization process.

Despite the obvious importance in determining accuracy and the great interest in microarray technology, limited information is available concerning systematic analysis of microarray accuracy during technique optimization or for collections of individual experiments. One method that may be chosen to optimize protocols is to measure and increase signal intensity. The assumption made here is that increased signal to noise will yield more accurate experiments. For instance, the Tumor Analysis Best Practices Working Group states that the percentage of present calls on the Affymetrix platform is an important QC criterion (Hoffman *et al*., 2004). Anecdotally, we know that increase in fluorescent signals have often been used as the main criteria for technique optimization. It should be noted, however, that in particular, on the other array platforms, an increase in signal may in fact be the result of an increase in aspecific target binding, for example, due to increased

cross-hybridization. It should be further noted that signal saturation can lead to loss of dynamic range and the inability of microarrays to measure real changes in expression levels.

A second method to evaluate if a technique has been fully optimized is to compare a new method with the one that has already been widely accepted, using a correlation coefficient to determine how similar the results of the two methods are. However, a correlation coefficient only gives information on how closely the two protocols behave and does not give any information on the accuracy of either one. A high correlation may mean that the compared technologies simply suffer from the same systematic errors. Whether ratios or absolute intensities are being compared, a low correlation still begs the question of which technique is the most accurate. Using correlation coefficients of intensities in order to monitor and improve reproducibility, for example between two channels, can also suffer from the drawback that the technology is being optimized for yielding identical results rather than accurately reporting differences in expression.

A third method for evaluation is to use an established cell line or tissue culture experiment that reproducibly shows differences that can be verified with other techniques such as Northern blotting or quantitative RT-PCR. This assumes that these alternate techniques accurately report differences in expression. Nevertheless, differential cell culture experiments are a good method because this does involve optimization of reporting established differences in expression, which is the goal of most microarray experiments. One disadvantage is that verification and optimization is driven by the differences reported in the microarray experiment. There is no test for false-negatives, unless RT-PCR, for example, is carried out on many thousands of genes not reported as differentially expressed by the microarray experiment. A second drawback is that this does not lend itself to the routine assessment of every single microarray experiment carried out after optimization. This is especially important for projects that require reproducible accuracy across many hybridizations or if collections of microarray experiments are to be compared with each other.

In this chapter, we discuss a fourth approach for optimizing nearly any aspect of microarray technology that does not suffer from the disadvantages outlined above and that also lends itself for reporting accuracy within every subsequent microarray experiment after optimization. This approach is based on the use of external controls, also known as spikes, spike-in controls, exogenous controls, or external standards. We describe what external controls are, how they may contribute toward systematic analyses of accuracy, and how such controls should be implemented. Other uses of external controls such as normalization, determining sensitivity, and reporting absolute mRNA copies per cell are also discussed.

2. External controls

External controls are in vitro synthesized RNA molecules that can be exogenously added to the RNA sample that is to be assayed. The distinct advantage of control RNAs is that there is complete control over their use, in particular, over the amount that is added to samples, which makes such controls extremely

versatile. Although external controls have been introduced and used sporadically at quite early stages in microarray technology (Holstege *et al*., 1998; Lockhart *et al*., 1996), it is apparent that their virtue is only starting to be fully recognized through the increased demand for the standardization of microarray experiments. A good example is the External RNA Controls Consortium initiative (ERCC, **http://www.ctsl.nist.gov/biotech/workshops/ERCC2003/**), which is presently being launched to supply the microarray community with standard external controls that are "platform-independent control materials needed for the performance evaluation of reproducibility, sensitivity, and robustness in gene expression analysis."

When external controls are added in fixed amounts to experimental RNA samples, they can be used to test how well their representation within each sample, or their ratios between two samples, are determined by the microarray experiment. For example, they may be added at fixed ratios in any quantity between two otherwise identical RNA samples in order to mimic differences in mRNA expression for a small number of genes. In this case, the microarray features representing the external controls are the only features on the microarray that should be reporting differences in RNA levels. Different control RNAs can also be added to RNA samples at defined picomolar amounts in order to test the sensitivity of an experiment by covering the entire range of possible mRNA expression values (Lockhart *et al*., 1996; Ramakrishnan *et al*., 2002). A combination of different ratios and different quantities is especially useful, because experiments can be evaluated on the basis of sensitivity as well as accuracy in reproducing the actual spiked ratios (Badiee *et al*., 2003; Hughes *et al*., 2001; Kane *et al*., 2000; Richter *et al*., 2002; Wang *et al*., 2003).

The accuracy of a microarray experiment is affected by the individual steps that make up such an experiment: RNA isolation, target synthesis and labeling, hybridization, and image and data analysis. The external controls can be used to determine how the accuracy is influenced during a microarray experiment from the moment they have been added. To give an example, when the controls are labeled separately and added just before hybridization, the accuracy of hybridization and subsequent steps can be determined (Dudley *et al*., 2002; Hughes *et al*., 2001; Lockhart *et al*., 1996). To test the effect of the complete experimental procedure on accuracy, the external controls could also be added to the lysis buffer of the RNA isolation method or as encapsulated RNA prior to RNA isolation, as suggested by the ERCC.

When data from individual hybridization assays are to be compared, any differences introduced by variation in the experimental procedure have to be countered by normalization. In the vast majority of publications involving microarrays, the data of the gene-specific features of a microarray are used for normalization because these data cover the whole array layout and the complete intensity range. However, using gene-specific features has the disadvantage that normalized differences in gene expression will always be balanced. That is, approximately as many genes will be downregulated as upregulated (Figure 1a and c) (Van De Peppel *et al*., 2003) because the underlying assumption is that gene expression, on average, does not change between samples. This does not necessarily represent what is really taking place in the biological samples, especially

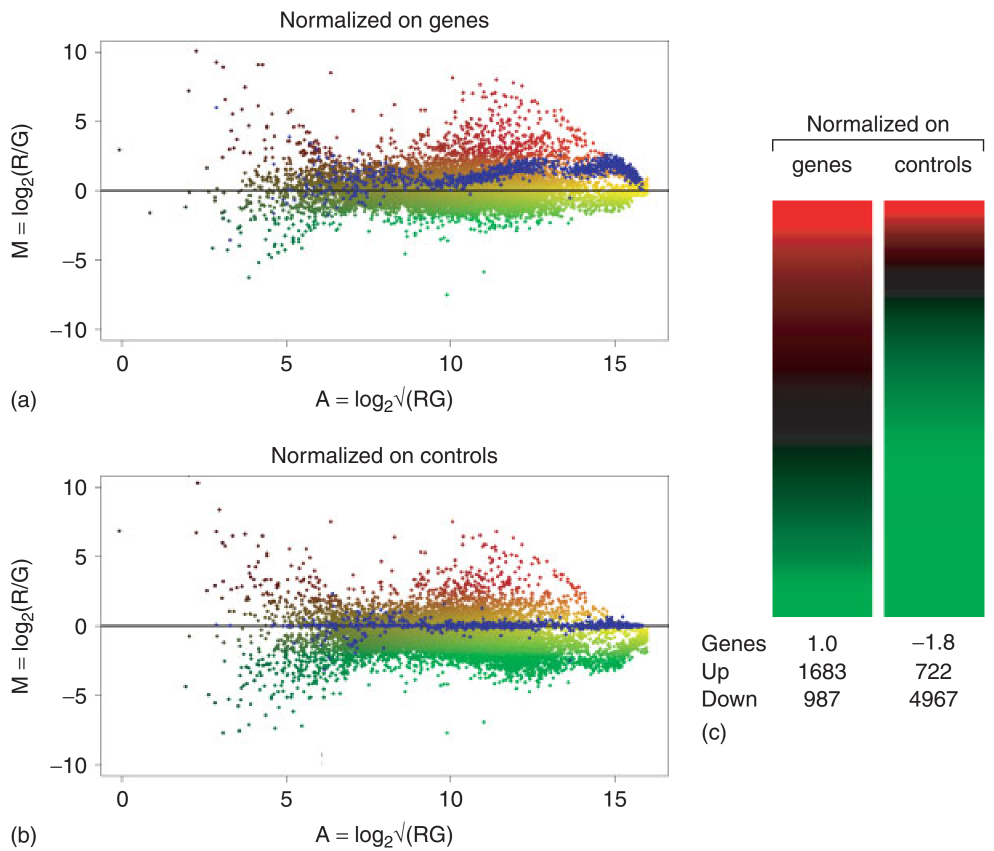


Figure 1 Use of external controls in normalization. Yeast cells were either grown exponentially in log phase or cultured during stationary phase. Nine different external controls were added to the total RNA of mid-log phase and stationary phase samples in a 1:1 ratio. cDNA of the stationary phase sample was Cy5-labeled and cDNA of the log-phase sample was Cy3-labeled. Hybridization was performed on in-house spotted yeast slides. (a) and (b) are MA scatterplots after lowess normalization on genes (a) or controls (b). (c) illustrates the difference between the two normalization strategies in the amount of genes that are up- or downregulated. After normalization on genes, a similar number of genes is up- as is downregulated (a). On the other hand, after normalization on controls, most genes show decreased expression (b). (Reproduced with permission from Van de Peppel *et al*. (2003). © EMBO)

when global changes in gene expression are expected. In other words, the accuracy of a measurement can be severely affected, when genes are used for normalization. However, external controls avoid these limitations and can be used to determine global or unbalanced changes. An example of global changes in gene expression is the comparison of exponentially grown cells with cells in stationary phase, where mRNA synthesis is expected to be decreased. In spite of this reduction in mRNA levels, fixed amounts of labeled material that corresponds to equal amounts of mRNA are applied on a microarray for both exponentially grown as well as stationary phase cells. By adding an identical amount of external control RNA to the same amount of total RNA, the external controls can be used to normalize for equal amounts of total RNA (Figure 1b and c) (Van De Peppel *et al*., 2003).

Another experimental situation that may require external control normalization is the use of microarrays carrying a relatively small number of features, such as those dedicated to a specific biological process or type of genes. Here, changes in expression may be unbalanced because of the preselection of genes as these arrays are typically designed to use genes that are consistently differentially expressed under the experimental condition under study.

It should be noted that the use of external controls does not limit the approaches one can use to normalize data; external controls have been used for total intensity normalization (Holstege *et al*., 1998; Kane *et al*., 2000) but can also be used in local intensity–dependent normalization (e.g., lowess, Yang *et al*., 2002), provided that a number of control RNAs are used with a range of spike-in concentrations.

The amount of external controls that is added is normally expressed as a molar quantity or as picograms per micrograms of total RNA or mRNA. Although the representation of external controls as copy number per cell is informative for the sensitivity in a microarray experiment, the quantitative conversion of the amount of external controls to a copy number per cell is not trivial. To express the sensitivity and accuracy of a method as mRNA copy numbers per cell (e.g., Kane *et al*., 2000), several assumptions have to made. The important variables used in various publications to calculate the copy number per cell are average transcript length (for humans, this has varied between 1000 and 2700 bp, Hughes *et al*., 2001, **http://www.ncbi.nlm.nih.gov/Web/Newsltr/spring03.human.html**), total RNA per cell (e.g., 20 pg, Kane *et al*., 2000), total number of polyA transcripts per cell (e.g., 100 000 or 360 000, Hughes *et al*., 2001; Kane *et al*., 2000), and percentage of mRNA within the total RNA used. The simple formula below can be used to estimate the spiked-in equivalent, expressed in transcripts per cell, when a known amount of external control RNA has been added to total RNA.

$$transcript\ \ per\ \ cell = \frac{Y(g)}{X(g)} \times \frac{total\ \ transcripts\ \ per\ \ cell}{perc.of\ \ mRNA\ \ in\ \ total\ \ RNA} \times 100 \qquad (1)$$

where Y is the amount of external control in grams and X is the amount of total RNA in grams.

3. Requirements for external controls

In view of the fact that external controls are used to monitor changes in sample RNA, it is imperative that external control RNA behaves similarly to the sample mRNA and, consequently, both should have the same physical properties, including similar lengths and GC content. When used for eukaryotic samples, external controls should also have a polyA tail, preferably of the same length as for eukaryotic mRNA. This is important as the length of the polyA tail can affect the behavior of sample mRNA and external control RNA in at least four ways. First, because a polyA tail is a region of low complexity, the risk of cross-hybridization is increased if the polyA tail is also included in the probe. Second, when modified (d)UTPs are used for labeling, they will be incorporated at the position of the polyA tail. A difference in the length of the polyA tail will result in a different

number of labeled (d)UTPs at this position between control RNA and sample mRNA. This can cause a discrepancy in the way the signal intensity corresponds to the number of transcripts for similar amounts of control mRNA and sample mRNA. Third, when labeled nucleotides are in close proximity to one another in the polyA tail, in particular at this position, the risk for quenching of fluorescence will increase with higher labeling percentages ('t Hoen *et al*., 2003). The use of a two nucleoside anchored primer at the 5′ end of the polyA tail or random primers in target synthesis can circumvent these problems. Fourth, many RNA amplification and labeling protocols rely on priming from the polyA tail and it is important that the controls and the experimental RNA have the same priming efficiencies.

4. Concluding remarks

As more and more laboratories will become involved in microarray technology, the demand for a format that can be used as an aid in data comparison will become greater. The use of external controls, especially when standardized, is extremely helpful in evaluating the diversity of data that is being generated. Eventually, these external controls will enable scientists to perform more accurate analyses of microarray experiments from different sources.

References

Badiee A, Eiken HG, Steen VM and Lovlie R (2003) Evaluation of five different cDNA labeling methods for microarrays using spike controls. *BMC Biotechnology*, **3**, 23.

Dudley AM, Aach J, Steffen MA and Church GM (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7554–7559.

Hoffman EP, Awad T, Palma J, Webster T, Hubbell E, Warrington JA, Spira A, Wright G, Buckley J, Triche T, *et al*. (2004) Guidelines: Expression profiling - best practices for data generation and interpretation in clinical trials. *Nature Reviews. Genetics*, **5**, 229–237.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES and Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, **19**, 342–347.

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD and Madore SJ (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, **28**, 4552–4557.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.

Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M, *et al*. (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Research*, **30**, e30.

Richter A, Schwager C, Hentze S, Ansorge W, Hentze MW and Muckenthaler M (2002) Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays. *Biotechniques*, **33**, 620–628, 630.

’t Hoen PA, de Kort F, van Ommen GJ and den Dunnen JT (2003) Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Research*, **31**, e20.

Van De Peppel J, Kemmeren P, Van Bakel H, Radonjic M, Van Leenen D and Holstege FC (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports*, **4**, 387–393.

Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J and Lee NH (2003) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology*, **4**, R5.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, e15.

SAGE

Seth Blackshaw
Johns Hopkins University School of Medicine, Baltimore, MD, USA

1. Introduction

Serial analysis of gene expression (SAGE) shares much in common with other digital gene expression technologies such as expressed sequence tag (EST) (*see* Article 78, **What is an EST?**, Volume 4) sequencing and massively parallel signature sequencing (MPSS). These techniques all aim to comprehensively profile gene expression by obtaining short stretches of DNA sequence from randomly selected cDNAs in a sample of interest. As when conducting an opinion poll, the more cDNAs that are sampled, the more accurate and comprehensive is the resulting expression profile. SAGE distinguishes itself from other technologies by the cloning and sequencing of multiple concatenated sequence tags – short DNA sequences obtained from a defined point within each cDNA that are long enough to uniquely identify each transcript (Velculescu *et al*., 1995; Saha *et al*., 2002).

2. How is SAGE performed?

mRNA molecules from a sample of interest are bound at their 3′ ends using magnetic beads conjugated to oligo (dT), and first- and second-strand cDNA is synthesized. The sample is then digested to completion with a four-base cutter restriction enzyme (usually NlaIII by convention, although any enzyme can be used) (see Figure 1, Step 1). All sequences that lie 5′ of the 3′-most NlaIII site is thus lost, and this NlaIII site will define the position of the sequence tag.

A linker containing both a PCR primer and a type IIS restriction site is then ligated to the 5′ end of the cDNA (Figure 1, Step 2). The sample is typically divided into two at this point, with linkers that differ only in the PCR primer sequence being used. The sample is then digested with the type IIS enzyme whose site is encoded in the linker. Depending on the enzyme used, this enzyme cleaves within the cDNA anywhere from 10 to 22 bp 3′ of the NlaIII site, generating the SAGE tag for that particular transcript. Cleaved tags are typically blunt-ended, after which cleaved tags ligated to the two sets of linkers are combined, ligated together to form ditags, and amplified by PCR (Figure 1, Step 3).

Ditags are then recleaved with NlaIII, gel purified, ligated to form concatemers, and subcloned into plasmid vectors to form a SAGE library. Individual clones are

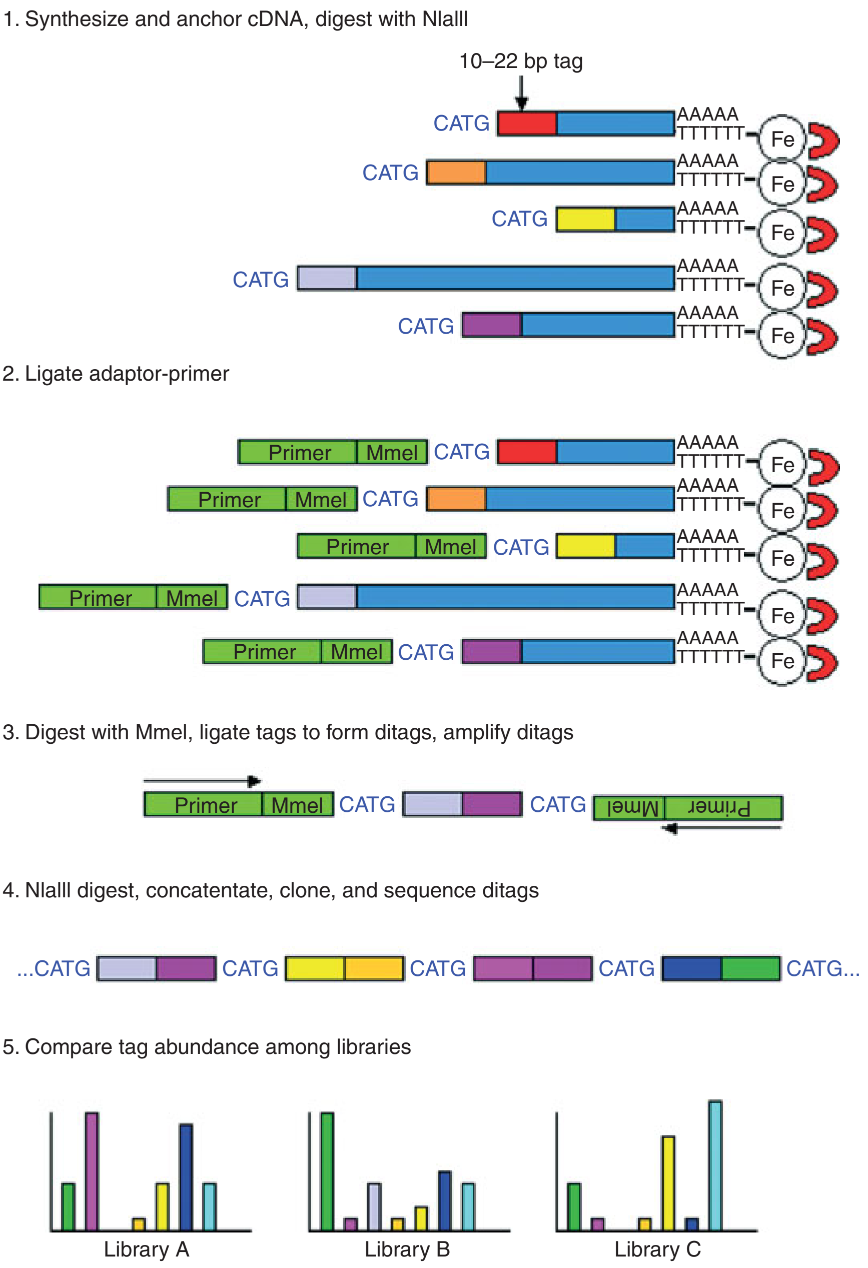


Figure 1 SAGE library construction and analysis

then isolated and sequenced (Figure 1, Step 4). Finally, SAGE tags are matched to genes using appropriate software, and tag abundance levels from the library of interest are compared to those from other libraries to identify differentially expressed genes (Figure 1, Step 5) where gene expression levels are determined simply by tag count. Duplicate ditags are discarded from the analysis, thus controlling for any tag-specific biases in PCR amplification.

3. Digital- versus hybridization-based expression profiling

Digital-based expression profiling technologies such as SAGE have a number of advantages over hybridization-based approaches such as microarray analysis (*see* Article 90, **Microarrays: an overview**, Volume 4). Digital-based approaches are nearly comprehensive – sampling all expressed genes, rather than being limited by one's choice of probes. Absolute levels of gene expression (i.e., individual tags as a percentage of total tags) are obtained – which making it straightforward to compare data sets obtained by different labs and large amounts of public digital expression data are already available from sites like the NCBI SAGE map project (**http://www.ncbi.nlm.nih.gov/SAGE**). It should be noted, however, that tag abundance measurements are more accurate for highly expressed transcripts than for those expressed at low levels. Only the widely used molecular approaches of cDNA library construction and DNA sequencing are used to generate the data, making it easy to initiate these studies. Very little experimental error is introduced in the construction of the libraries (Blackshaw *et al*., 2003), with PCR amplification biases being controlled by the elimination of duplicate ditags from analysis. The sensitivity of digital-based expression profiling is only limited by one's ability to sequence – one could, in principle, sequence enough tags to cover every mRNA in the sample of interest. Finally, since there is no hybridization background or variations in probe quality to consider, interpretation of digital gene expression data is quite straightforward and easily performed with a simple set of statistical tools (*see* Article 53, **Statistical methods for gene expression analysis**, Volume 7).

Hybridization-based systems have a number of advantages of their own, however, cost and speed being the main ones. Commercially available oligonucleotide arrays (*see* Article 92, **Using oligonucleotide arrays**, Volume 4) giving nearly complete coverage of the human genome can be hybridized in triplicate for ∼$3000, while constructing and sequencing 50 000 tags from a SAGE library can cost ∼$15 000. Furthermore, construction of SAGE libraries can require two weeks of work, and performing the required number of sequencing reactions can take months, while microarray hybridization can be completed in two days and many arrays can be run in parallel. The considerable variation in gene expression seen among biological samples (Blackshaw *et al*., 2003; Pritchard *et al*., 2001) is best controlled for by obtaining multiple replicates of expression profiles – a proposition that is difficult to accomplish with digital-based methods. Since all digital-based methods rely on sampling, the "real" tag count in a sample actually represents a Poisson distribution around the observed tag count in a library (Audic and Claverie, 1997), introducing a level of uncertainty that is unacceptable for certain studies. Though expression profiling of very small amounts of starting material (i.e., <1 µg total RNA) typically

requires amplification of the starting material and leads to skewed representation of certain transcripts, the comparative hybridization of amplified samples can control for this, making it the method of choice in such situations. Finally, variations in efficiency of cDNA-target hybridization among probes can result in expression of some rare genes being detected quite efficiently, whereas rare genes are only efficiently detected at high tag numbers by digital-based methods.

4. Hybridization or digital-based expression profiling – which to use?

The choice of a digital or hybridization-based approach to profiling gene expression depends to a large extent on one's experimental aims. If the primary aim of a study is gene discovery, or more generally, to analyze gene expression in a limited number of samples in great depth and at high resolution, digital-based methods such as SAGE are probably the method of choice. Digital-based methods are particularly well suited to generating publicly accessible databases of gene expression. Hybridization-based methods, on the other hand, are the method of choice when a large number of samples need to be screened or when very small amounts of starting material are used. More generally, hybridization-based approaches are probably the best choice for exploratory studies where a large initial investment of resources is not desired.

5. Comparison of SAGE with other methods of digital-based expression profiling

Since 20–40 SAGE tags are obtained with each sequencing reaction as opposed to 1 tag per reaction with conventional EST sequencing, SAGE is much more efficient at profiling gene expression than EST sequencing. Likewise, the great majority of public EST data is obtained from libraries that have been both normalized and subtracted, and thus do not accurately reflect mRNA levels in the sample in question. MPSS has many of the same advantages of SAGE and is, in principle, more rapid, although costs per sample are high at present and the technology is not widely available. Neither SAGE nor MPSS, however, gives extensive information about alternative splicing of mRNAs, and both approaches require a fully sequenced genome to be maximally useful. Conventional EST sequencing may be the method of choice in cases in which either of these factors is a concern.

6. Variations and choice of methodology in SAGE

A number of variations of SAGE have been developed in recent years, including the generation of 5′-anchored libraries using affinity to the 5′ cap structure or, in *Caenorhabditis elegans*, to the trans-spliced 5′ leader exon found in virtually all mRNAs (Hwang *et al*., 2004; Wei *et al*., 2004). The choice of SAGE tag

length varies with the complexity of the genome one is analyzing and the specific topic addressed. All things being equal, a shorter tag is preferable, since more tags can then be obtained in each sequencing reaction. Since each tag specifies 4^n possible combinations, where $n =$ tag length, a 10-bp tag can specify over 10^6 combinations – more than enough to unambiguously define the majority of transcripts in most organisms, provided cDNA sequences for the transcripts are available. However, to identify a unique match in genomic DNA and allow de novo transcript annotation, anywhere from 10^{10} to 10^{11} combinations must be specified, tags longer than 13 bp (13 unique bp + 4 bp from the NlaIII site = 4^{17} combinations) are required.

The number of tags sequenced will depend on the number of transcripts per cell in the sample tested, the sensitivity desired, and the cellular complexity of the sample examined. Abundant transcripts will be accurately profiled with relatively low tag counts, but rare mRNAs in abundant cell types or abundant mRNAs selectively expressed in rare cell types will require much higher tag counts to profile accurately.

References

Audic S and Claverie JM (1997) The significance of digital gene expression profiles. *Genome Research*, **7**, 986–995.

Blackshaw S, Kuo WP, Park PJ, Tsujikawa M, Gunnersen JM, Scott HS, Boon WM, Tan SS and Cepko CL (2003) MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues. *Genome Biology*, **4**, R17.

Hwang BJ, Muller HM and Sternberg PW (2004) Genome annotation by high-throughput 5′ RNA end determination. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 1650–1655.

Pritchard CC, Hsu L, Delrow J and Nelson PS (2001) Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13266–13271.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW and Velculescu VE (2002) Using the transcriptome to annotate the genome. *Nature Biotechnology*, **20**, 508–512.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET and Ruan Y (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(32), 11701–11706.

Introductory Review

Core methodologies

Gerard Cagney
University College Dublin, Dublin, Ireland

There is no science without fancy, and no art without facts – Vladimir Nabokov

These are exciting times for proteomics. While technological progress at the frontiers of the science continues unabated, the basic ability to identify proteins in a mass spectrometer justifies much of the fuss. This section describes techniques that form the "bread and butter" of proteomics – techniques that are within the reach of many laboratories and are increasingly provided as "core facilities" within research institutes. Indeed, one of the major challenges is not merely technical but to communicate to the general research community exactly what proteomics can and cannot do. In this respect, Section 1, "Core Methodologies", goes some way to expounding the state of the art. Most of the methods described can be considered routine in a proteomics lab, yet they greatly extend the power of virtually any type of experiment involving proteins. The significance of protein and peptide ionization methods coupled to mass spectrometry (MS) was recognized by the award of the Nobel Prize for Chemistry in 2002 to John Fenn and Koichi Tanaka. But before the late 1990s, proteins were almost exclusively identified using antibody-based methods such as ELISA and western blot, or using Edman degradation for protein microsequencing.

Nowadays, researchers who can accurately and rapidly identify proteins have three very powerful advantages in pursuing their research goals. First, because protein identification can be achieved in a semiautomatic manner, the type of project that can be undertaken is greatly extended. Examples in this and subsequent sections range from the analysis of chemical modification of single proteins, to the characterization of protein complexes isolated by affinity precipitation, to examining the proteome dynamics of whole cells under conditions of health and disease. These studies are possible partly because the feedback cycle between experiment and result has been dramatically shortened with proteomics. It is therefore feasible to develop a new model in a reasonable amount of time, or to analyze more samples under more conditions in order to give greater statistical power. Second, researchers are not limited to identifying only those proteins for which tests (or antibodies) are available. Even where tests are available, they may not be uniformly sensitive or robust, making large-scale or parallel studies problematic or impossible. Meanwhile, although the goal of comprehensive genome-wide qualitative and quantitative analysis (already realized for mRNA)

is not yet a reality for proteins, recent studies in organisms such as yeast and the malaria parasite can clearly be described as "systemwide", with thousands of proteins being identified. A third advantage of proteomics is that all or most of the proteins in a sample can be identified, not just those specifically chosen for testing according to some criteria. This opens the door to chance discovery. Proteomics studies, in common with other so-called omics approaches, have sometimes been criticized for lacking central hypotheses, and these criticisms deserve to be taken on board. In the most satisfying work however, experiment design that directly addresses a hypothesis has been combined with the possibility of making new discoveries. In other words, hypothesis testing and discovery science need not be incompatible. If "finding out new things" is what makes for successful science, then we can expect many more successes from proteomics.

These successes are being driven by machines, mass spectrometers, whose origins, like much of molecular biology itself, can be traced to pioneering physicists like Joseph John Thompson and Francis William Aston at the beginning of the twentieth century. It would be a disservice to the many outstanding intellectual achievements to attempt to summarize the history of MS (mass spectrometry) here, but the key methods contributing to the analysis of peptides and proteins followed the development of soft desorption and ionization techniques, particularly, matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI). These methods allow large biomolecules such as peptides to enter a mass spectrometer. Getting biomolecules into the instrument as intact charged species was the major barrier to the development of the field. Both methods are discussed in these pages (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5), and it is interesting that both continue to be widely used with no single ionization method dominating to date. One may generalize and say that MALDI is more amenable to automation and high-throughput experimentation, while ESI has been adapted very successfully to separation methods such as liquid chromatography. Almost by definition, a scientist engaged in a proteomics experiment will be dealing with mixtures of proteins of some complexity, and in general, the complexity needs to be reduced before or during the ionization process. This is because of limits resulting from design of the various instruments, for instance, the duty cycle, the dynamic range, and signal saturation.

The complexity of proteomics samples is addressed through a mixture of sample preparation and component separation. Here, an overview of sample preparation is provided (*see* Article 2, **Sample preparation for proteomics**, Volume 5), including proteolytic digestion with trypsin, a very important method in most proteomics laboratories (*see* Article 16, **Improvement of sequence coverage in peptide mass fingerprinting**, Volume 5). Strategies to separate proteins may be applied to intact proteins or to their proteolytic digests. Often, separation by one feature of the chemistry of proteins or peptides is sufficient, but the complexity of many proteomic samples is such that at least one additional feature must be used. Two-dimensional gels (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) separate proteins according to their isoelectric points and mass for instance, while the MUDPIT technique (Multidimensional Protein Identification Technology) separates peptides according to their charge and hydrophobicity (*see* Article 13, **Multidimensional liquid chromatography**

tandem mass spectrometry for biological discovery, Volume 5). Many variations are possible, constrained only by the need to keep the peptides solvated and ultimately to ionize them. Capture technologies that trap and separate proteins or peptides on the basis of the presence of moieties such as carbohydrate or phosphate groups have been described, and in principle, almost any type of chemistry that can separate proteins or peptides on the basis of their distinct chemistries can be exploited. However, unlike say oligonucleotides, proteins are chemically heterogeneous, and certain classes of proteins, while very important biologically, are difficult to adapt to normal sample preparation protocols. Here strategies that may be used to prepare membrane proteins for MS are described (*see* Article 15, **Handling membrane proteins**, Volume 5).

One trend that has been consistent across all types of proteomics technologies is miniaturization. This has been driven by factors such as the need to analyze thousands of proteins in parallel, limited amounts of sample, economics, and especially in the case of mass spectrometry, by the need for increased sensitivity. The development of micro- and nanoscale liquid chromatography was especially important, and both Nano-MALDI and Nano-ESI MS are described here (*see* Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5). A very useful description of making nanocolumns and tips is also included (*see* Article 19, **Making nanocolumns and tips**, Volume 5). This method can lead to very significant gains in the sensitivity of electrospray experiments with complex peptide mixtures. An important new technique for analysis at the cell and tissues level is laser-based microdissection (*see* Article 6, **Laser-based microdissection approaches and applications**, Volume 5), and this approach is likely to have important clinical research applications.

The range of different mass spectrometers can be overwhelming to the newcomer (and latecomer!). Both traditional and emerging aspects of the workhorse instruments are treated here: time-of-flight MS (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5), quadrupole MS (*see* Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5), and quadrupole ion trap MS (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5). The most useful features of different architectures have been intelligently combined to create hybrid instruments (*see* Article 10, **Hybrid MS**, Volume 5). Similarly, the use of the extremely powerful class of Fourier transform ion cyclotron instrument is becoming more widespread (*see* Article 5, **FT-ICR**, Volume 5). An important aspect of these multistage MS techniques is the use of instrument software that automates the process. This is particularly the case for tandem mass spectrometry experiments where the ions chosen for fragmentation following a primary "scan" can be chosen using predetermined rules, eliminating the need for constant supervision of the experiment. As discussed (*see* Article 18, **Techniques for ion dissociation (fragmentation) and scanning in MS/MS**, Volume 5), tandem mass spectrometry is a major means of acquiring structural information about peptide molecules, particularly, the amino acid sequence. Normally associated with quadrupole or ion trap instruments, the somewhat related phenomenon of post-source decay means that this form of structural analysis can now be carried out by time-of-flight instruments, considerably extending its application.

However, it is often forgotten that the power of proteomics is not fully realized solely by a powerful instrument, or even by an optimal combination of separation method, ionization method, and instrument. Without the automated methods for interpreting mass spectra, these protein "signatures" would accumulate many times faster than the ability of skilled workers to determine what they represent. The development of algorithms and programs that assist the process of identifying proteins from their cognate mass spectra was therefore another key technological breakthrough, allowing the science of proteomics to evolve. Several workers in the early 1990s independently devised methods for using publicly available DNA sequence databases to reduce the task of interpreting protein and peptide mass spectra to a tractable problem. Although similar in principle, the approaches of peptide mass fingerprinting (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5) and tandem mass spectrum interpretation (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5) require different programs. Tutorials describing both approaches included here will be very helpful (*see* Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5).

The observation at the top of this text was made by Nabakov during an interview in 1966. Many of us are fortunate to arrive now onto a scene where approaches that required highly original thought and effort to develop are sufficiently embedded that we may cast our proteomic nets across the water in expectation of a plentiful catch. The work of these pioneering scientists should be applauded, as should the authors of the chapters following, many of whom are one and the same. To them we owe the ability of proteomics to combine fancy and facts.

Introductory Review

Advances of LC-MS and CZE-MS in proteome analysis

Hookeun Lee
Institute for Molecular Systems Biology, ETH Zürich, Switzerland

Ruedi Aebersold
Institute for Molecular Systems Biology, ETH Zürich, Switzerland
University of Zürich, Zürich, Switzerland

1. Introduction

Mass spectrometry (MS) is an essential core technology in proteomics, the comprehensive analysis of the proteins expressed in a cell or tissue (Aebersold and Goodlett, 2001). In the late 1980s, two new ionization methods were developed that have revolutionized the analysis of large biomolecules by MS.

Both methods, electrospray ionization (ESI, Fenn *et al*., 1989) and matrix-assisted laser desorption/ionization (MALDI, Karas and Hillenkamp, 1995), convert nonvolatile analytes like proteins and peptides to ions in gas phase without derivatization and minimal fragmentation. These ions are then selected to mass analysis. Because of the lack or minimal extent of analyte fragmentation during ionization, ESI and MALDI are also referred to as *soft* ionization methods (Aebersold, 2003). ESI generates ions by spraying a solution at atmospheric pressure through a small-diameter needle and by applying electric potential between the needle and the inlet of a mass spectrometer. Through evaporation and splitting, the size of the droplets is reduced. This desolvation process ultimately produces bare analyte ions in the gas phase, which are then transmitted to the mass spectrometer. Because of its solution-based nature, ESI is easily coupled on line with liquid phase separation techniques, particularly with high-performance liquid chromatography (HPLC) and capillary electrophoresis (CE). Furthermore, due to the propensity of ESI to produce multiple charged ions, simple quadrupole instruments and other types of mass analyzers with limited mass/charge (m/z) range could be used to detect ions with masses exceeding the nominal m/z range of the instrument (typically 0–4000 Da). For MALDI, analytes are cocrystallized with a UV-absorbing compound (matrix) on a sample probe. Pulses of UV laser light are used to vaporize small amounts of the matrix and analyte ions to produce protonated gas phase ions. The method is relatively resistant to contaminating species and, due to the propensity to generate predominantly singly charged ions, MALDI spectra are simple to explain (Aebersold and Goodlett, 2001).

Over the last decade increasingly sophisticated and powerful mass spectrometers for peptide analysis have been developed and interfaced with high-performance separation systems. Separations prior to MS analysis remove salts and other impurities and concentrate analytes in the narrow elution peaks, resulting in increased sensitivity. This review will focus on the recent advances of on-line high-performance separation techniques such as HPLC and capillary zone electrophoresis (CZE) for MS analysis in proteomics.

2. Reversed-phase microcapillary chromatography and ESI-MS

Nanospray ionization, a low flowrate ESI, gained wide popularity due to its high sensitivity. Wilm and Mann proposed that the desorption efficiency of analyte peptide ions from the electrosprayed droplet increases as the size of the droplets decreases. This is thought to be due to the larger surface area of the droplet in relation to its total volume (Wilm and Mann, 1996). As a result, a greater proportion of the available analyte molecules is ionized and transmitted to mass spectrometers. This enhanced sensitivity of the nanospray method reduced the amount of peptide required for MS analysis to a few femtomoles and below in the mid-1990s (Shevchenko *et al*., 1996). Efforts to miniaturize HPLC to utilize nanospray ionization in combination with on-line peptide separation lead to the development of packed microcolumns using fused silica capillaries with a 20–150-μm inner diameter (i.d.) and a flowrate of less than 1 μl min^{-1} (Ishihama, 2005; Moseley *et al*., 1991; Holland and Jorgenson, 1995; Emmett and Caprioli, 1994). Even though the sensitivity of MS analysis increases with decreasing column i.d., most microcapillary liquid chromatography (μLC) work is done with 75 or 100 μm i.d. capillary columns because these clog less frequently than capillary columns of ≤50 μm i.d. Using this microscale format, reversed-phase liquid chromatography has most often been combined with mass spectrometric analysis of a protein digest because the on-line separation also effectively removes contaminants such as salts and detergents that are detrimental to ESI performance. Figure 1 shows an integrated μLC/nanospray ionization device (Gatlin *et al*., 1998). The microcross works as a flow splitter providing nanoliter per minute scaled flow to the column and as a junction point of liquid and the metal wire supplying the electric potential required for nanospray ionization. One advantage of this type of device is the use of the capillary tip, tapered to ∼5 μm, to hold the C_{18}-derivatized particles in place rather than use of a sintered frit connected to a separate ESI emitter via a union. This design significantly reduces the dead volume and band broadening. This type of integrated device has gained popularity from work by Yates and coworkers (Gatlin *et al*., 1998), most notably as multidimensional protein identification technology (MudPIT) using a biphasic capillary column (Link *et al*., 1999).

The integrated device has been further developed by several groups (Yi *et al*., 2003; Meiring *et al*., 2002; Licklider *et al*., 2002). The main advantage is the introduction of a sample trapping and enriching precolumn to enable automated consecutive runs for high-throughput analysis. The precolumn allows decreased sample loading time, rapid sample desalting, and facilitates automation. Despite

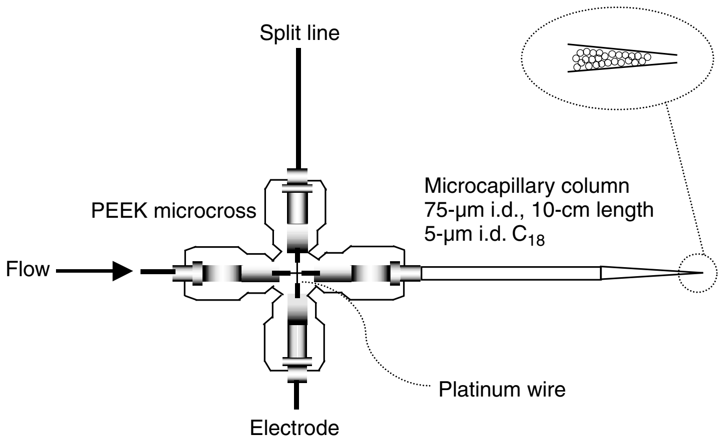


Figure 1 Schematic representation of μLC/ESI device

these advantages, the dead volume between the precolumn and the microcapillary analytical column can be a limiting factor. The consequent band broadening will decrease ESI sensitivity for those analytes present at low relative stoichiometry. To minimize the dead volume, a precolumn and a microcapillary column were integrated as one piece (Meiring *et al*., 2002; Licklider *et al*., 2002), or the two columns were placed in close proximity (Yi *et al*., 2003). The ruggedness and sensitivity of such devices was demonstrated by performing more than 60 consecutive μLC separations on complex peptide mixtures without degradation of chromatographic resolution and by detecting 1 amol of a standard peptide using a conventional ion-trap mass spectrometer (Yi *et al*., 2003).

Further increases of sensitivity into the zeptomole (10^{-21} moles) range was achieved by combining ultra high-pressure μLC separating at a pressure of 10 000 psi with Fourier transform ion-cyclotron resonance (FTICR) mass spectrometer (Shen *et al*., 2003; Belov *et al*., 2004).

Recently, Yin *et al*. (2005) have developed a novel integrated nano-LC/ESI microfluidic chip in which an enrichment column, a reversed-phase column (75-μm width × 50-μm depth × 40-mm length), and an ESI tip have been laser-ablated into a polyimide film. The compact format of the chip reduces the number of capillary tubing and fittings required and greatly simplify operation skills. It also facilitates troubleshooting for blockages and leakages in capillaries and connectors. Initial efforts using this chip coupled to ion-trap mass spectrometer have resulted in the identification of 47 protein clusters from albumin and IgG-depleted rat plasma sample by the chip separation alone, and 111 clusters by two-dimensional separation of strong cation exchange (SCX) chromatography and the chip (Fortier *et al*., 2005).

3. Capillary zone electrophoresis and ESI-MS

For CZE or simply CE, separations occur inside a capillary filled with a moderately conductive liquid. The degree of separation of analytes inside the capillary depends

on their relative electrophoretic mobilities under the applied electrical field. CE has been widely utilized for separations of proteins and peptides due to the high separation efficiency of the method that reaches $\sim 10^6$ theoretical plates, and high speed (Shen and Smith, 2002; Simpson and Smith, 2005). A critical disadvantage of CE, however, is the limited sample injection volume, typically below 20 nl ($\sim$1% of the total capillary column volume) (Shen *et al*., 2000), which limits the sample concentration and limits of detection. To overcome this limitation, various preconcentration methods have been implemented including sample stacking, transient isotachophoresis, and solid-phase extraction (Stroink *et al*., 2001). Among the various preconcentration methods, solid-phase extraction has been effectively coupled with CE for on-line interfacing with mass spectrometers. Figeys and coworkers used a microfluidic device integrating a C_{18}-based preconcentrator and CE for the analysis of protein digests (Figeys *et al*., 1999; Figeys *et al*., 1998).

ESI is the predominant method for interfacing CE to MS, while MALDI has also been used extensively. ESI-based interfaces divide into two categories: sheath-flow interfaces and sheathless interfaces (Schmitt-Kopplin and Frommberger, 2003; Moini, 2002). For the sheath-flow interface, a second tube of a larger diameter surrounds the separation capillary in a coaxial arrangement. The sheath liquid is driven through the outer tube by an external pump. The sheath-flow interfaces provide system stability and reduce the requirement for the background electrolyte to be compatible with ESI-MS. These advantages make those most widely used interface designs (Simpson and Smith, 2005). The main disadvantage of the sheath-flow configuration, however, is low detection sensitivity due to the dilution of the analyte by the sheath liquid (Moini, 2002). Sheathless interfaces offer higher sensitivity because there is no sheath liquid present to dilute the CE effluent. A variety of the sheathless interfaces have been reviewed by Issaq *et al*. (2004).

4. Conclusions

Major analytical challenges in proteomics are the enormous complexity of the proteome and the large variance in protein abundances. Thus, with advances in MS technology, the initial efforts of proteomics have focused on improving interfaces between the separation techniques and MS to increase detection sensitivity. Concurrently, another effort to combine other separation techniques, such as SCX, affinity chromatography, free-flow electrophoresis, capillary isoelectric focusing, and one-dimensional electrophoresis, with liquid chromatography - mass spectrometry (LC-MS) and CE-MS have been carried out to enhance the resolution and dynamic range of proteome-wide analysis (Han *et al*., 2001; Moritz *et al*., 2004; Chen *et al*., 2003; Aebersold and Goodlett, 2001). It can be expected that with these developments complete proteome analysis will become a reality.

References

Aebersold R (2003) A mass spectrometric journey into protein and proteome research. *Journal of the American Society for Mass Spectrometry*, **14**, 685–695.

Aebersold R and Goodlett DR (2001) Mass spectrometry in proteomics. *Chemical Reviews*, **101**, 269–295.

Belov ME, Anderson GA, Wingerd MA, Udseth HR, Tang K, Prior DC, Swanson KR, Buschbach MA, Strittmatter EF, Moore RJ *et al*. (2004) An automated high performance capillary liquid chromatography-Fourier transform ion cyclotron resonance mass spectrometer for high-throughput proteomics. *Journal of the American Society for Mass Spectrometry*, **15**, 212–232.

Chen J, Balgley BM, DeVoe DL and Lee CS (2003) Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. *Analytical Chemistry*, **75**, 3145–3152.

Emmett MR and Caprioli RM (1994) Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. *Journal of the American Society for Mass Spectrometry*, **5**, 605–613.

Fenn JB, Mann M, Meng CK, Wong SF and Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**, 64–71.

Figeys D, Corthals GL, Gallis B, Goodlett DR, Ducret A, Corson MA and Aebersold R (1999) Data-dependent modulation of solid-phase extraction capillary electrophoresis for the analysis of complex peptide and phosphopeptide mixtures by tandem mass spectrometry: application to endothelial nitric oxide synthase. *Analytical Chemistry*, **71**, 2279–2287.

Figeys D, Zhang Y and Aebersold R (1998) Optimization of solid phase microextraction – capillary zone electrophoresis – mass spectrometry for high sensitivity protein identification. *Electrophoresis*, **19**, 2338–2347.

Fortier MH, Bonneil E, Goodley P and Thibault P (2005) Integrated microfluidic device for mass spectrometry-based proteomics and its application to biomarker discovery programs. *Analytical Chemistry*, **77**, 1631–1640.

Gatlin CL, Kleemann GR, Hays LG, Link AJ and Yates III JR (1998) Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid-chromatography-microspray and nanospray mass spectrometry. *Analytical Biochemistry*, **263**, 93–101.

Han DK, Eng J, Zhou H and Aebersold R (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*, **19**, 946–951.

Holland LA and Jorgenson JW (1995) Separation of nanoliter samples of biological amines by a comprehensive two-dimensional microcolumn liquid chromatography system. *Analytical Chemistry*, **67**, 3275–3283.

Ishihama Y (2005) Proteomic LC-MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *Journal of Chromatography. A*, **1067**, 73–83.

Issaq HJ, Janini GM, Chan KC and Veenstra TD (2004) Sheathless electrospray ionization interfaces for capillary electrophoresis-mass spectrometric detection. *Journal of Chromatography. A*, **1053**, 37–42.

Karas M and Hillenkamp F (1995) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, **60**, 2299–2301.

Licklider LJ, Thoreen CC, Peng J and Gygi SP (2002) Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Analytical Chemistry*, **74**, 3076–3083.

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM and Yates III JR (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, **17**, 676–682.

Meiring HD, Heeft E van der, Hove GJ ten and Jong APJM de (2002) Nanoscale LC-MS$^{(n)}$: technical design and applications to peptide and protein analysis. *Journal of Separation Science*, **25**, 557–568.

Moini M (2002) Capillary electrophoresis mass spectrometry and its application to the analysis of biological mixtures. *Analytical and Bioanalytical Chemistry*, **373**, 466–480.

Moritz RL, Ji H, Schutz F, Connolly LM, Kapp EA, Speed TP and Simpson RJ (2004) A proteome strategy for fractionating proteins and peptides using continuous free-flow

electrophoresis coupled off-line to reversed-phase high-performance liquid chromatography. *Analytical Chemistry*, **76**, 4811–4824.
Moseley MA, Deterding LJ, Tomer KB and Jorgenson JW (1991) Nanoscale packed-capillary liquid chromatography coupled with mass spectrometry using a coaxial continuous-flow fast atom bombardment interface. *Analytical Chemistry*, **63**, 1467–1473.
Shen Y, Berger SJ, Anderson GA and Smith RD (2000) High-efficiency capillary isoelectric focusing of peptides. *Analytical Chemistry*, **72**, 2154–2159.
Shen Y, Moore RJ, Zhao R, Blonder J, Auberry DL, Masselon C, Pasa-Tolic L, Hixson KK, Auberry KJ and Smith RD (2003) High-efficiency on-line solid-phase extraction coupling to 15-150-microm-i.d. column liquid chromatography for proteome analysis. *Analytical Chemistry*, **75**, 3596–3605.
Shen Y and Smith RD (2002) Proteomics based on high-efficiency capillary separations. *Electrophoresis*, **23**, 3106–3124.
Shevchenko A, Wilm M, Vorm O and Mann M (1996) Mass spectrometric sequencing of proteins characterization in silver-stained polyacrylamide gels. *Analytical Chemistry*, **68**, 850–858.
Simpson DC and Smith RD (2005) Combining capillary electrophoresis with mass spectrometry for applications in proteomics. *Electrophoresis*, **26**, 1291–1305.
Stroink T, Paarlberg E, Waterval JCM, Bult A and Underberg WJM (2001) On-line sample preconcentration in capillary electrophoresis, focused on the determination of proteins and peptides. *Electrophoresis*, **22**, 2374–2383.
Schmitt-Kopplin P and Frommberger M (2003) Capillary electrophoresis-mass spectrometry: 15 years of developments and applications. *Electrophoresis*, **24**, 3837–3867.
Wilm M and Mann M (1996) Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry*, **68**, 1–8.
Yi EC, Lee H, Aebersold R and Goodlett DR (2003) A microcapillary trap cartridge-microcapillary high-performance liquid chromatography electrospray ionization emitter device capable of peptide tandem mass spectrometry at the attomole level on an ion trap mass spectrometer with automated routine operation. *Rapid Communications in Mass Spectrometry*, **17**, 2093–2098.
Yin H, Killeen K, Brennen R, Sobek D, Werlich M and Goor Tvande (2005) Microfluidic chip for peptide analysis with an integrated HPLC column, sample enrichment column, and nanoelectrospray tip. *Analytical Chemistry*, **77**, 527–533.

Sample preparation for proteomics

Thierry Rabilloud
CEA-Laboratoire d'Immunochimie, Grenoble, France

1. Introduction

Sample preparation can be defined as the process that goes from the biological sample of interest to the final protein extract that can be handled by the desired proteomics method. This process has to achieve parallel goals:

- disruption of the initial structure of the biological material;
- prevention of any spurious degradation of the analyte (i.e., proteins);
- removal of compounds that can interfere with the further processing of the sample, for example, protein separation or chemical labelling or controlled digestion;
- keeping proteins soluble in conditions and media that are compatible with this downstream processing.

It can be easily seen that sample preparation strongly depends on the type of processing that will be used for the proteomics analysis *per se*. Two major subsets of sample preparation can be defined. In the first one, the protein–protein interactions must be kept as close as possible to the *in vivo* situation. This is of major importance when protein complexes are to be analyzed. In the second one, individual polypeptide chains are analyzed. In this case, separation of proteins into undegraded polypeptide chains must be achieved.

2. Native sample preparation

This means that the sample preparation must respect as much as possible the three-dimensional structure of the proteins, including the links that can exist between proteins. Whatever the subsequent separation method is, two main problems are encountered with these techniques. The first problem is to prevent degradation of the sample by hydrolases, especially proteases, which are very difficult to inhibit under the conditions that allow the three-dimensional structure of proteins to be kept intact. The second and most important problem is to keep the protein assemblies

as close as possible from their *in vivo* situation while keeping conditions that are compatible with downstream processing.

This poses in turn the problem of choosing extraction and purification conditions that look as similar as possible to those prevailing in a cell, but still being practicable. It must be recalled that a cytosol is a 10% by weight solution of proteins, that is, a milieu of very high viscosity and of rather unknown chemical parameters (dielectric constant, ionic strength, etc.). It is therefore obvious that weakly buffered water solutions do not represent a good model of the intracellular medium and that problems with protein assemblies in such media are likely to occur. However, increasing the ionic strength is probably not adequate, as this is known to extract many subunits from complexes. The work or Schägger and coworkers (Schägger and Von Jagow, 1991) can provide interesting tracks to follow. In this paper, it is shown that high concentrations of a dielectric compound (e.g., aminocaproic acid) dramatically enhances the extraction of protein complexes. Thus, dielectric compounds could show some of the beneficial effects of salts (salting in effects), without showing their dissociating effects. However, aminocaproic acid has a relatively low dipolar moment, and might not be the ideal dielectric compound. Chemicals with stronger dipolar moments such as sulfobetaines (Vuillard *et al*., 1995a) have been shown to increase protein solubility and may prove useful for protein complexes isolation. Their efficiency as salt mimics has been shown for the purification of halophilic proteins (Vuillard *et al*., 1995b). Thus, dielectric compounds with varying dipolar moments or hydrophobic parts may be worth testing to enhance the stability of protein complexes during their fractionation either by electrophoresis or chromatographic techniques.

Apart from these general parameters, sample preparation is also driven by the separation method used. When chromatographic separations are used, for example, by affinity chromatography (Rigaut *et al*., 1999), the conditions are relatively flexible for many parameters such as ionic strength or detergent choice. This is not the case when native electrophoresis is used (Schägger and Von Jagow, 1991). In this case, the ionic strength must be kept low and an ionic charge must be given to the extracted proteins. This is achieved either by the use of a specially designed detergent (Hisabori *et al*., 1991) or via a charged protein-binding molecule such as Coomassie Blue (Schägger and Von Jagow, 1991). Apart from its use for electrophoresis, this charge-shift process has the important benefit to increase protein solubility.

3. Denaturing sample preparation: general considerations

In this case, the constraint of keeping proteins intact is replaced by separating them as much as possible in the form of individual polypeptide chains. This means in turn breaking all bonds, keeping proteins together, and allowing binding to other compounds. Apart from disulfide bridges, most of these bonds are noncovalent interactions, for example, ionic bonds, hydrogen bonds, and hydrophobic interactions. These interactions can be broken by chaotropes, detergents (especially ionic ones), or a combination thereof, while disulfide bridges must be broken by special reagents. Here again, one of the major difficulties is

to prevent spurious degradation of the proteins by hydrolases. Under denaturing conditions, only very resistant hydrolases are active, that is, mainly proteases, but this can be a very important problem, especially when complete proteins are to be analyzed downstream. The problem is further enhanced by the fact that denatured, unfolded proteins are most sensitive to proteases, even if the latter are only partly active.

Last but not least, in many samples, protein denaturation induces removal of DNA-binding proteins from DNA, and thus massive swelling of DNA in the sample, correlating with increased viscosity and additional problems, depending on the downstream methods. Other nonproteinaceous compounds, such as salts or lipids, can also be deleterious for downstream methods and must be taken into account in the sample preparation protocols.

Biological samples are also generally very complex, and no single separation technique is able to resolve the sample into individual components. This means in turn that complex separation schemes, usually multidimensional (i.e., relying on different parameters), must be used. In most of these schemes, the interface process between the various separations is robust enough, so that the constraints applying to sample preparation are only those induced by the first stage of separation.

3.1. The case of disulfide bridges

Breaking of disulfide bridges is usually achieved by adding to the solubilization medium an excess of a thiol compound, most of the time dithiothreitol (DTT). However, DTT is still not a perfect reducing agent. Some proteins are not fully reduced by DTT. In addition, DTT must be used in an important excess over the protein disulfide bridges. This is a problem when thiol derivatization must be performed, as the derivatization agent will have to be present in a further excess over DTT. In these cases, phosphines are very often an effective answer. The reaction is stoichiometric, which allows us in turn to use very low concentrations of the reducing agent (a few millimolar). The most powerful compound is tributylphosphine, which was the first phosphine used for disulfide reduction in biochemistry (Ruegg and Rüdinger, 1977). However, the reagent is volatile, toxic, has a rather unpleasant odor, and needs an organic solvent to make it water-miscible. Dimethylsulfoxide (DMSO) or dimethylformamide (DMF) are suitable carrier solvents, which enable the reduction of proteins by 2-mM tributylphosphine (Kirley, 1989). These drawbacks have disappeared with the introduction of water-soluble phosphines, for example tris (carboxyethyl) phosphine, which seem however to be less potent reducers.

4. Denaturing sample preparation rationales for selected proteomics methods

The last part of this chapter will deal with sample preparation available according to the downstream proteomics approach. For space limitation reasons, only the

rationales can be detailed in this chapter, and detailed protocols can be found in the references cited herein.

4.1. *Sample preparation for proteomics based on zone electrophoresis*

In these proteomics methods, the separation process is split in two phases (e.g., Bell *et al*., 2001). The first phase is a protein separation by denaturing zone electrophoresis, that is, in the presence of denaturing detergents, most often sodium dodecyl sulfate (SDS). The second phase is carried out by chromatography on the peptides produced by digestion of the separated proteins. As mentioned above, this has no impact on the sample preparation itself, which just needs to be compatible with the initial zone electrophoresis.

This is by far the simplest case. Sample preparation is achieved by mixing the initial sample with a buffered, concentrated solution of an ionic detergent, usually containing a reducer to break disulfide bridges and sometimes an additional nonionic chaotrope such as urea. Ionic detergents are among the most powerful protein denaturing solubilizing agents. Their strong binding to proteins makes all proteins to bear an electric charge of the same type, whatever their initial charge may be. This induces in turn a strong electrostatic repulsion between protein molecules, and thus maximal solubility. The system of choice is based on SDS, as this detergent binds rather uniformly to proteins. However, SDS alone at room temperature, even at high concentrations, may not be powerful enough to denature all proteins. This is why heating of the sample in the presence of SDS is usually recommended. The additional denaturation brought by heat synergizes with SDS to produce maximal solubilization and denaturation, even of the most resistant proteases.

The use of SDS is not always without drawbacks. One of the most important is encountered when the sample is rich in DNA. A terrible viscosity results, which can hamper the electrophoresis process. Moreover, some protein classes (e.g., glycoproteins) bind SDS poorly and are thus poorly separated in the subsequent electrophoresis. In such cases, it is advisable to use cationic detergents. They are usually less potent than SDS, so that a urea-detergent mixture must be used for optimal solubilization (MacFarlane, 1989). Moreover, electrophoresis in the presence of cationic detergents must be carried out at a very acidic pH, which is not technically simple but still feasible (MacFarlane, 1989). This technique has however gained recent popularity as a double zone electrophoresis method able to separate even membrane proteins (Hartinger *et al*., 1996).

4.2. *Sample preparation for proteomics based on two-dimensional gel electrophoresis*

In this scheme, the proteins are first separated by isoelectric focusing (IEF) followed by SDS electrophoresis. The constraints made on sample preparation are thus those induced by the IEF step. One of these constraints is the impossibility to use ionic

detergents at high concentrations, as they would mask the protein charge and thus dramatically alter its isoelectric point (pI). Ionic detergents can however be used at low doses to enhance initial solubilization (Wilson *et al*., 1977), but their amount is limited by the capacity of the IEF system (in terms of ions tolerated) and by the efficiency of the detergent exchange process that takes place during the IEF step. Another major constraint induced by IEF is the requirement for low ionic strength, induced by the high electric fields required for pushing the proteins to their isoelectric points. This means in turn that only uncharged compounds can be used to solubilize proteins, that is, neutral chaotropes and detergents. The basic solubilization solutions for IEF thus contain high concentrations of a nonionic chaotrope, historically urea but now more and more a mixture of urea and thiourea (Rabilloud *et al*., 1997), together with a reducing agent and a nonionic detergent. While CHAPS and Triton X-100 are the most popular detergents, it has been recently shown that other detergents can enhance the solubility of proteins and give better performances (Chevallet *et al*., 1998; Luche *et al*., 2003). In this solubilization process, detergents play a multiple role. They bind to proteins and help keep them in solution, but they also break protein–lipid interactions and promote lipid solubilization. This is a problem in lipid-rich samples, in which the amount of detergent present in the sample preparation solution can be limiting. In this case, lipid removal should be included in the sample preparation process. However, lipid removal is based on protein precipitation in solvents in which lipids are soluble. The major problem is that many proteins cannot be solubilized from the precipitate under IEF-compatible conditions (e.g., Tastet *et al*., 2003).

Apart from lipids, IEF is also very sensitive to many other compounds, such as salts or nucleic acids, which must be removed from many samples. Salt removal is carried out either by dialysis or by precipitation of proteins (e.g., by trichloroacetic acid (TCA) or organic solvents). The classical drawback of these approaches is loss of proteins due to their sticking to the dialysis membrane in the former method and by being soluble in the precipitation medium or irreversibly precipitated in the latter method.

Nucleic acids are present at problematic concentrations in most cell extracts, not to speak of nuclear extracts. One removal method is digestion by nucleases, initially by a mixture of RNAses and DNAses (O'Farrell, 1975). As with most of enzyme-based removal methods, the main drawbacks are linked to the parallel action of proteases (Castellanos-Serra and Paz-Lago, 2002), thereby degrading the proteins, and to the addition of extraneous proteins (the nucleases). A more versatile strategy is to use a high pH during extraction, so that most proteins are anions and are repelled from the anionic nucleic acids. To avoid overswelling of the nucleic acids, which decreases the subsequent removal by ultracentrifugation, this increase of pH can be mediated by the addition of a basic polyamine (e.g., spermine, Rabilloud *et al*., 1994), which will precipitate the nucleic acids. However, the most basic proteins are still cations at the pH 10 obtained in the spermine extraction method, and thus stick to nucleic acids. Extraction of these basic proteins can be obtained either with competing cations such as protamine (Sanders *et al*., 1980) or lecithins at acid pH (Willard *et al*., 1979). These methods are efficient but introduce high amounts of charged compounds, so that only low sample amounts can be loaded.

Another method is based on TCA precipitation (Damerval *et al*., 1986), which denatures nucleic acids and gives them a positive charge, thereby repelling the proteins. While this method is able to extract basic proteins such as ribosomal proteins (Görg *et al*., 1998), some proteins are lost at the resolubilization step after precipitation.

Another major problem encountered in sample preparation for two-dimensional electrophoresis resides in proteolysis. While ionic detergents are known to efficiently inactivate proteases, urea and nonionic detergents are unable to do so, as recently shown for example on yeast extracts (Harder *et al*., 1999). Thorough protease inactivation is clearly not an easy task. Most efficient methods are based either on TCA precipitation (with the risk of protein losses) or on the use of ionic detergents, which is always limited. It has been recently described, however, that thiourea helps in preventing spurious proteolysis (Castellanos-Serra and Paz-Lago, 2002).

4.3. Sample preparation for peptide separation-based approaches

In these methods, the sample preparation process is designed to offer optimal digestion of the proteins into peptides. This means that proteins must be extracted from the sample and denatured to maximize exposure of the protease cleavage sites. This also means that the protease used for peptide production must be active in the separation method. In this case, the robustness of many proteases is a clear advantage. Classical extraction media usually contain either multimolar concentrations of chaotropes or detergents. In the latter case, the sample is usually solubilized and denatured in high concentrations of ionic detergents, and simple dilution is used to bring the detergent concentration down to a point compatible with other steps such as chemical labeling or proteolysis.

The choice between chaotropes and detergent is driven mainly by the constraints imposed by the peptide separation method. In the wide-scope approach based on on-line two-dimensional chromatography of complex peptide mixtures (Washburn *et al*., 2001), both the ion exchange and reverse-phase steps are very sensitive to detergents. It must be mentioned that these on-line two-dimensional chromatographic methods are one of the rare cases in which the interface between the two separation methods does not bring extra robustness, so that the sample preparation must be compatible with both chromatographic methods. The modification approach (Gevaert *et al*., 2003), which uses extensive reverse-phase chromatography, is also very sensitive to detergent interference. This rules out the use of detergent and favors the use of chaotropes, generally nonionic ones because of the ion exchange step. Urea is used for these methods, with possible artifacts induced by urea-driven carbamylation of the sample during the lengthy digestion process. Inclusion of thiourea and lowering of the urea concentration could decrease the incidence of carbamylation in these methods. In methods in which a detergent-resistant method is used, for example, avidin selection of biotinylated peptides (Gygi *et al*., 1999), extraction by SDS is clearly the method of choice, as the above-mentioned drawbacks are absent.

5. Concluding remarks

It should be obvious from the above that the sample preparation process is dependent on many parameters that are sample-dependent, such as protein concentration, protein–lipid and protein–nucleic acid ratios, and so on. This explains why sample preparation, although critical for the quality of the final proteomics results, is so ill mastered and is likely to stay so.

References

Bell AW, Ward MA, Blackstock WP, Freeman HN, Choudhary JS, Lewis AP, Chotai D, Fazel A, Gushue JN, Paiement J, *et al*. (2001) Proteomics characterization of abundant Golgi membrane proteins. *The Journal of Biological Chemistry*, **276**, 5152–5165.

Castellanos-Serra L and Paz-Lago D (2002) Inhibition of unwanted proteolysis during sample preparation: evaluation of its efficiency in challenge experiments. *Electrophoresis*, **23**, 1745–1753.

Chevallet M, Santoni V, Poinas A, Rouquié D, Fuchs A, Kieffer S, Rossignol M, Lunardi J, Garin J and Rabilloud T (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis*, **19**, 1901–1909.

Damerval C, De Vienne D, Zivy M and Thiellement H (1986) Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. *Electrophoresis*, **7**, 52–54.

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR and Vandekerckhove J (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature Biotechnology*, **21**, 566–569.

Görg A, Boguth G, Obermaier C and Weiss W (1998) Two-dimensional electrophoresis of proteins in an immobilized pH 4-12 gradient. *Electrophoresis*, **19**, 1516–1519.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Harder A, Wildgruber R, Nawrocki A, Fey SJ, Larsen PM and Görg A (1999) Comparison of yeast cell protein solubilization procedures for two-dimensional electrophoresis. *Electrophoresis*, **20**, 826–829.

Hartinger J, Stenius K, Hogemann D and Jahn R (1996) 16-BAC/SDS-PAGE: a two-dimensional gel electrophoresis system suitable for the separation of integral membrane proteins. *Analytical Biochemistry*, **124**, 126–133.

Hisabori T, Inoue K, Akabane Y, Iwakami S and Manabe K (1991) Two-dimensional gel electrophoresis of the membrane-bound protein complexes, including photosystem I, of thylakoid membranes in the presence of sodium oligooxyethylene alkyl ether sulfate/dimethyl dodecylamine oxide and sodium dodecyl sulfate. *Journal of Biochemical and Biophysical Methods*, **22**, 253–260.

Kirley TL (1989) Reduction and fluorescent labelling of cyst(e)ine-containing proteins for subsequent structural analyses. *Analytical Biochemistry*, **180**, 231–236.

Luche S, Santoni V and Rabilloud T (2003) Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics*, **3**, 249–253.

MacFarlane D (1989) Two dimensional benzyldimethyl-n-hexadecylammonium chloride-sodium dodecyl sulfate preparative polyacrylamide gel electrophoresis: a high capacity high resolution technique for the purification of proteins from complex mixtures. *Analytical Biochemistry*, **176**, 457–463.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological Chemistry*, **250**, 4007–4021.

Rabilloud T, Adessi C, Giraudel A and Lunardi J (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **18**, 307–316.

Rabilloud T, Valette C and Lawrence JJ (1994) Sample application by in-gel rehydration improves the resolution of two-dimensional electrophoresis with immobilized pH gradients in the first dimension. *Electrophoresis*, **15**, 1552–1558.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Ruegg UT and Rüdinger J (1977) Reductive cleavage of cystine disulfides with tributylphosphine. *Methods Enzymology*, **47**, 111–116.

Sanders MM, Groppi VE and Browning ET (1980) Resolution of basic cellular proteins including histone variants by two-dimensional gel electrophoresis: evaluation of lysine to arginine ratios and phosphorylation. *Analytical Biochemistry*, **103**, 157–165.

Schägger H and von Jagow G (1991) Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Analytical Biochemistry*, **199**, 223–231.

Tastet C, Charmont S, Chevallet M, Luche S and Rabilloud T (2003) Structure-efficiency relationships of zwitterionic detergents as protein solubilizers in two-dimensional electrophoresis. *Proteomics*, **3**, 111–121.

Vuillard L, Braun-Breton C and Rabilloud T (1995a) Non-detergent sulphobetaines: a new class of mild solubilization agents for protein purification. *The Biochemical Journal*, **305**, 337–343.

Vuillard L, Madern D, Franzetti B and Rabilloud T (1995b) Halophilic protein stabilization by the mild solubilizing agents nondetergent sulfobetaines. *Analytical Biochemistry*, **230**, 290–294.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Willard KE, Giometti C, Anderson NL, O'Connor TE and Anderson NG (1979) Analytical techniques for cell fractions. XXVI. A two-dimensional electrophoretic analysis of basic proteins using phosphatidyl choline/urea solubilization. *Analytical Biochemistry*, **100**, 289–298.

Wilson D, Hall ME, Stone GC and Rubin RW (1977) Some improvements in two-dimensional gel electrophoresis of proteins. Protein mapping of eukaryotic tissue extracts. *Analytical Biochemistry*, **83**, 33–44.

Specialist Review

Tandem mass spectrometry database searching

Jimmy K. Eng, Daniel B. Martin and Ruedi Aebersold
Institute for Systems Biology, Seattle, WA, USA

1. Introduction

Mass spectrometry (MS) has become a principal technology in proteomics analysis largely due to improvements in sequence database availability and development of software tools to interpret mass spectra. Software tools of the modern proteomic era were first published around 1990 (Johnson and Biemann, 1989; Bartels, 1990; Yates *et al*., 1991). These programs facilitate the de novo sequencing process that became possible after collision-induced dissociation (CID) was shown to induce peptide fragmentation primarily along the amide backbone. In 1993, the era of sequence database searching began using protein fingerprinting or peptide mass mapping, with the development of software to query sequence databases with protein digestion products. This method remains in use today as it is fast and relatively accurate, but typically requires purified proteins as input substrate such as those excised from a gel spot or band. Despite these advances, the rapid adoption of MS as a proteomics platform is directly correlated with the automation of computational tools for the analysis of spectra from peptides analyzed by tandem mass spectrometry. As a peptide-centric identification strategy, MS/MS database searching or peptide fingerprinting allows for the analysis of mixtures of unseparated proteins such as those from protein complexes or whole-cell lysates. This review will touch on computational methods associated with peptide identification using MS/MS sequence database searching.

One of the earliest attempts to combine MS/MS analysis with database search strategies was published by Mann and Wilm (1994). Their method queried sequence databases with "peptide sequence tags", which are short amino acid sequences obtained by partial de novo sequence interpretation of tandem mass spectra combined with the corresponding mass constraints of the start and end mass of the sequence interpretation within the tandem mass spectrum. Sequence tag searches can be performed quickly because string matches can be coded efficiently and the database search space is typically limited to the analysis of peptides generated from an in silico enzymatic digestion of the database entries. Additionally, by relaxing search constraints such as the prescribed peptide mass, the sequence tag approach can identify interesting posttranslational modifications and homologous sequences

(Shevchenko *et al*., 1997). The major drawback to the sequence tag approach is the tag generation process as it has historically been performed manually. This is a rate-limiting step particularly when one considers the extraordinary rate at which modern mass spectrometers can acquire tandem mass spectra. Sequence tags can be generated in an automated fashion; however, automated tag-generation success rates are typically limited to good-quality spectra in the tag interpretation region and, in the case of ions generated by electrospray ionization, to the tags that are generated from a specific precursor ion charge state (Tabb *et al*., 2003).

The year 1994 also saw the first published method of querying uninterpreted tandem mass spectra of peptides against sequence databases (Eng *et al*., 1994). The embodiment of the described program, SEQUEST, uses a two-stage score function. The first stage generates a preliminary score for all peptides in a sequence database on the basis of summed ion intensities of a processed spectrum, with preliminary score adjustment factors for ion continuity, immonium ion contributions, and the percentage of fragment ions matched. The second stage of the score function, applied to the 500 best preliminary scoring peptides, generates a cross-correlation score between each theoretical and acquired tandem mass spectra. The cross-correlation score used in SEQUEST is effectively the correlation coefficient or vector dot product computed between the two spectra with a correction factor. Putative peptide identifications are ranked and reported on the basis of the cross-correlation score. Information from the preliminary score, such as the preliminary score rank and percentage of matched fragment ions, are also useful indicators for assessing positive identifications and are reported as well. Subsequent publications describe applications of the algorithm to identify posttranslational modifications (Yates *et al*., 1995a), query of nucleic acid sequence databases (Yates *et al*., 1995b), application to spectra generated by different instruments (Griffin *et al*., 1995; Yates *et al*., 1996), implementation to run on parallel computer clusters (Sadygov *et al*., 2002), and a modification to normalize the cross-correlation score for length dependence (MacCoss, 2002). SEQUEST is among the most widely used search algorithms.

In 1999, Perkins *et al*. published a search tool named *Mascot* that queries uninterpreted tandem mass spectra among one of its search options (Perkins *et al*., 1999). The Mascot score function is based on the MOWSE scoring algorithm (Pappin *et al*., 1993) previously used in peptide mass fingerprinting but modified to be probability-based. Although the direct details of the implementation of the score function are unpublished, it is likely that statistics on fragment ion frequencies, the direct extension of the MOWSE peptide mass statistics, are the major component of the score function. It has been noted that for each candidate peptide, Mascot iteratively searches a spectrum's peak intensities to find the set of most intense peaks that yields the maximum score. Using this method, the optimal number of peaks in a spectrum are independently determined and scored for each peptide in a database. The database search scores are interpreted as p-values and significant peptide identifications are based on the p-value scores. Owing to its independent development, fast searches, and probability-based score function, Mascot has garnered popularity as it facilitates the analysis of data acquired from most of the major instrument manufacturers.

Bafna and Edwards (2001) implement a probabilistic model for MS/MS database searching in a program named SCOPE. The score function in SCOPE incorporates fragment ion probabilities, spectral noise, and instrument measurement error. The authors employ a two-stage stochastic model for the observed MS/MS spectrum, given a peptide. The first step generates fragment ions from peptides, given a probability distribution estimated from training data. For this step, sufficient training data were not available to learn the values empirically so the authors defined the probabilities on the basis of consultation with "experienced mass spectrometry operators". The second step generates a spectrum using these fragment ions according to a distribution based on known instrument measurement error. Peptides are scored by the probability they would give rise to the observed spectrum, as opposed to all alternative spectra. Dynamic programming is used to efficiently sum up contributions from each possible pattern of peptide ion fragmentation. Since the scores for each peptide with respect to an observed spectrum are computed independently of all other peptides under consideration, it is not impossible for several peptides to receive equally high scores. The authors thus compute an additional p-value aimed at estimating the likelihood that each peptide score arose by chance. Although only 28 MALDI-TOF-TOF test spectra are used to evaluate the scoring scheme, the prototype implementation demonstrates the effectiveness of the model as 26 of the 28 spectra are identified correctly with scores significantly separated from the incorrect peptides. Additionally, the two incorrect identifications exhibited scores not well separated from the other top-scoring peptides in each particular search. This indicates that the significance test can differentiate correct from incorrect identifications, at least in this small set.

Havilio *et al*. (2003) describe a computational method using a statistical scoring approach based on the probability of detecting fragment ions in the experimental spectrum that incorporates a measurement of ion intensity. A set of score functions is developed using parameters incorporating both experimental observations and theoretical knowledge of peptide fragmentation. Examples of parameters used in the score functions include terms for random or noise peaks, the relative intensity of a given fragment ion peak, and the relative frequency of unmatched fragment ion peaks. The authors generate their own test set of highly probable peptides to evaluate fragment ion statistics. To do so, a large dataset collected from 46 LC-MS/MS runs of an ion trap mass spectrometer is analyzed. The score function used in the de novo sequencing program described in Dancik *et al*. (1999) is implemented and the criteria for highly probable peptides, those that pass a p-value of .01 cutoff, exhibit over 50% matched fragment ions, and are greater than 10 amino acid residues in length, are used to generate statistics for Havilio's model. These include, for each fragment type, the probability of detecting an ion of a given intensity across a range of fragment mass to peptide mass ratios, the probability of failing to detect a fragment ion across the fragment mass to peptide mass range, and the probability of a random ion of a given mass to have a particular intensity. The authors implement two score functions, intensity-based and non-intensity-based, using these statistics, with specific details described in the manuscript. Presumably tested on the same data used to generate the ion statistics, the manuscript shows that the intensity-based statistical scorer exhibits significant improvements over

the non-intensity-based scorer, the initial score function of Dancik *et al*., and an implementation of the cross-correlation score function as presented in Eng *et al*.

An approach to identify peptides through sequence database searching using Bayesian statistics is implemented by ProbID (Zhang *et al*., 2002). This Bayesian approach attempts to quantify the likelihood that the query spectrum is generated from a given candidate peptide sequence in the database. Background information used in computing the probability includes the protease used to digest the input sample (which is used in the score function and not as a peptide filter in the database search itself), immonium ion presence, matched and unmatched ions, and consecutive or complementary matched fragment ions referred to as a pattern match. Relatively simple models are used to compute contributions for each of the background information types. For example, a normal distribution around each theoretical ion is used to weigh matched ions on the basis of their distance from the theoretical peak. The standard deviation of the normal distribution is defined by the user-specified mass accuracy of the input data. Similarly, the total number of unmatched ions is counted and used as an exponent for a term involving just the mass difference between the highest and lowest peak in the experimental spectrum. The authors compare ProbID against SEQUEST on two reference datasets and show that both software tools similarly identify a majority of the peptides correctly. However, a still significant ∼10% of the identifications are detected by only one of the search engines. In summary, this manuscript shows that a relatively simple Bayesian probabilistic score function can perform as well as an industry standard tool including identifying spectra not identified by SEQUEST.

Hernandez *et al*. introduce a new method for protein identification using MS/MS data, named *Popitam*, that uses a nondeterministic heuristic approach to confront the combinatorial problems associated with analyzing modified peptides (Hernandez *et al*., 2003). This problem causes extraordinary increase in search times as sequences that have multiple potential modification sites require an independent search for each combination of modifications. Popitam's algorithm selects a given number of peaks in a spectrum, transforms the peaks into a spectrum graph, and compares them against a sequence database to generate a list of scored candidate peptides. (A spectrum graph is a mathematical representation in which peaks in an MS/MS spectrum are represented as nodes in the graph. Linking the nodes with edges that differ by amino acid residue masses generates a peptide sequence.) The initial implementation of Popitam, referred to as the "Full Path algorithm", works by looking for complete paths through the MS/MS spectrum graph where the sequence database is used to direct the search for matched sequences. If the spectrum graph contains gaps of two or more fragmentation positions, the initial algorithm struggles to identify peptides of these spectra because a full path through the graph is required. In Popitam's "Tag algorithm", short tags are generated where the sequence database is used to direct and emphasize relevant sections in the graph from which tags can be generated. As the tag discovery process can be combinatorially complex, a heuristic search using Ant Colony Optimization (ACO) is performed. In this implementation, ACO (inspired by the pheromone trail laying and following behavior of real ants) is a method of exploring different solutions to find good scoring paths in the spectrum graphs. The resulting ant score is a combination of terms such as the coverage measure between database

peptide sequence and the generated sequence based on parsing the spectrum graph, and a regression score representing the quality of the correlation between the experimental masses included in the path through the spectrum graph and the corresponding theoretical masses computed from the database peptide sequence. With the tag algorithm, combining unconnected relevant sections of the graph in order to maximally cover the theoretical peptides suggests that Popitam will be able to identify mutated or modified peptides without any prior knowledge of the modification.

A fragment ion frequency–based model by Sadygov *et al*. uses a hypergeometric distribution and is named PEP_PROBE (Sadygov and Yates, 2003). The theory involves formulating a model based on the frequency of fragment ion matches. The hypergeometric distribution is related to the binomial distribution but handles cases of sampling without replacement, in this case, the sampling of matched fragment ions. Four variables are used to calculate the hypergeometric probability. The first term is the total number of theoretical fragment ions in a sequence database for peptides of a given mass. The second term is the number of these theoretical fragment ions that match a peak in the input spectrum. The third term is the number of possible fragment ions for the peptide sequence being scored. The last term is the number of matched fragment ions for the peptide sequence being scored. To assess the significance of peptide matches, the authors implement a p-value that is calculated from the cumulative hypergeometric distribution. An implementation of the hypergeometric algorithm considering only b- and y-ions is tested against a set of approximately 59 000 ion-trap MS/MS spectra that contains 5000 true identifications. An analysis against this characterized dataset shows that false-positive error rate of 50% to less than 5% can be achieved depending on how high the score threshold cutoff is set. However, there is no mention of the corresponding sensitivity rates for each displayed error rate.

The MS/MS database search methods presented above are only a selection of the published tools in this growing field of computational proteomics. MS-Tag (Clauser *et al*., 1999), which implements the MOWSE score function, mutant tolerant identifications with PENDATA (Pevzner *et al*., 2001), vector dot product analysis with Sonar (Field *et al*., 2002), and signal detection theory and extended matches in OLAV (Colinge *et al*., 2003) are other noteworthy database search tools not covered in detail here. However, research and development in MS/MS database searching has gone beyond simply finding the optimal score functions used in MS/MS database searching. A new class of software tools have recently been developed that do not currently perform MS/MS database searches but rather assist in distinguishing between correct and incorrect database search assignments, as produced by the search tools described above. Three models have been described: one statistically based model, one model centered on machine learning using full datasets, and one intensity-based machine-learning model. A fourth introduces the concept of ion mobility and how peptide composition and charge affect MS/MS fragmentation and thus database search analysis.

Moore *et al*. have developed an analysis tool called *Qscore* to evaluate the quality of SEQUEST search results at the protein level that the authors suggest may decrease false-positive assignments and manual curation time (Moore *et al*., 2002). This algorithm is pseudo-probability based as it incorporates a quality

score of an individual peptide match into a calculation of the probability that a protein identification has occurred by chance. The probabilistic portion of the calculation is based on the number of peptides identified from a particular protein, the number of tryptic peptides from the particular protein in the database, and the total number of identified peptides. The quality score portion of the score is calculated in a fashion similar to SEQUEST where the spectrum is divided into a number of small m/z bins. The total ion current of each bin in a normalized measured spectrum is multiplied by the total ion current in the same range in a normalized predicted spectrum and all products are summed to a single value. A perfect match in this algorithm gives a quality score of 1. The quality score is converted to a probability by taking the inverse of 1 minus the quality score. The composite Qscore incorporates this quality score into the probability calculation. Testing by the authors produced good correlation with hand curation with reduced false-positive results.

Keller *et al*. introduced a method based on the use of the expectation maximization (EM) algorithm to derive a mixture model of correct and incorrect peptide identifications from the data (Keller *et al*., 2002). The software tool, PeptideProphet, learns to distinguish between correct and incorrect peptide identifications based on observed information, primarily database search scores and other peptide properties, from each peptide in a dataset. Peptide properties that are used to discriminate between correct and incorrect identifications include the difference between the observed versus calculated peptide masses, the number of missed enzymatic cleavage sites, and the number of peptide termini consistent with the expected enzymatic cleavage. A peptide discriminant score optimally combines the individual search scores, such as those produced by SEQUEST, into a single discriminant score. The coefficients of the discriminant function are developed for each type of MS/MS database search program and can be optimized for each type of mass spectrometer. On the basis of the distribution of scores mapped through the discriminant function, PeptideProphet applies EM to find mixture model distributions, Gaussian for the positive population and gamma for the negative population, that best fit the observed data. Other parameters of the data, such as the distribution of correct enzymatic termini exhibited by the positive and negative populations, are also automatically learned. Using the learned distributions of the positive and negative populations, PeptideProphet assigns computed probabilities (as opposed to p-values) that, in an implementation using ion trap data with the SEQUEST database search program, are shown to accurately reflect the confidence of peptide identifications being correct. Such accurate computed probabilities furthermore enable the prediction of false-positive error rates. The algorithm learns the correct and incorrect peptide assignments from the information in each individual dataset, rather than relying on a single set of universal parameters, which makes it tolerant of variations in factors such as data quality and protease digestion efficiency. Additionally, because PeptideProphet can be extended to analyze data generated from various mass spectrometers and run through different search engines, it can serve to standardize publication criteria and facilitate comparisons of large-scale datasets, particularly if data of a constant error rate are reported. It should be noted that the authors extend this peptide analysis to identifying proteins with the tool ProteinProphet (Nesvizhskii *et al*., 2003).

Elias *et al*. describe an intensity-based model developed using a machine-learning algorithm on 27 000 high-quality 2+ peptide-spectrum matches (Elias *et al*., 2004). This strategy uses a decision tree to model the probability of seeing a fragment ion based on 63 peptide fragment attributes. The decision tree was also applied to a set of mismatched peptides (the second best peptide from the high-quality matches). The authors calculate the likelihood of observing each fragment ion from a potential peptide match using both trees and calculate an odds ratio for that fragment ion. The individual odds ratios for each peptide are summed to generate a cumulative odds ratio (LOD score) for that peptide match. The authors tested their LOD score against other commonly used single scoring tools, such as SEQUEST's XCorr and dCn, and were able to demonstrate better sensitivity (proportion of correct identifications noted as correct) and specificity (proportion of incorrect identifications noted as being incorrect) than any of the commonly used strategies. In the authors' analysis, the LOD score performed approximately as well as the discriminant score used in PeptideProphet and when combined with the Xcorr and dCn had a slight advantage in specificity for a given sensitivity; comparisons against the full PeptideProphet analysis were not performed. The authors tested a publicly available dataset previously evaluated using a standardized scoring system and report finding 24% more correct identifications at the same false-positive rate using a LOD-based analysis.

Kapp *et al*. mined a database of 5500 tandem mass spectra of unique peptides obtained by tryptic digestion in order to determine which factors affect tandem MS fragmentation under low-energy CID conditions (Kapp *et al*., 2003). A residue's preference for cleavage (N- or C-terminus) was determined on the basis of two independent methods. First, a cleavage intensity ratio (CIR) value is calculated to quantify the extent of fragmentation occurring both N- and C-terminal to each amino acid residue within a peptide. The CIR is calculated by dividing the summed intensity for each amide bond cleavage by the average intensity of all cleavage sites within a peptide. CIR values greater than 1 indicate enhanced cleavage and below 1 indicate reduced cleavage. Second, a linear model that takes into account positional factors as well as peptide length was used to determine whether certain residues showed a propensity to cleave N- and/or C-terminal. Linear regression was performed to estimate effects of variables, such as the relative position of the cleavage along the peptide backbone and the specific effect of the adjacent amino acid on either side of the cleavage site, in order to select and retain only those variables that have a real effect on the fragmentation process. On the basis of using both CIR values and the linear model, the authors have classified peptides as nonmobile if the number of available protons is less than or equal to the number of arginine residues. For the remaining peptides, if the number of available protons was greater than the number of arginine residues but less than or equal to the total number of basic residues (arginine, lysine, and histidine), these were classified as partially mobile, otherwise, they were classified as mobile (more protons than basic residues). This has been termed the relative proton mobility scale. Analysis of SEQUEST and Mascot against the 5500 unique manually validated peptide MS/MS spectra shows that the search score varies considerably depending on the charge state and proton mobility classification. In relation to MS/MS database search, two conclusions can be drawn from this analysis. The first is that search

algorithm cutoff filters or score thresholds perform poorly for certain classes of spectra (e.g., nonmobile peptide), so score cutoffs should be reevaluated by taking into account the relative proton mobility scale described. Second, score functions can be improved by taking into account fragment ion abundances that are based on sequence, charge state, and proton mobility classification.

2. Conclusion

Database search software has made extraordinary strides in the last 10 years. Numerous programs now exist to allow rapid assignment of peptides of uninterpreted spectra. Many such programs are quite good and easily implemented for both large and small laboratories. Comparisons show that a number of the best search programs are fairly similar in their performance, each finding roughly similar numbers of peptides and assigning identifications to a few spectra "invisible" to other algorithms. Novel algorithms show promise of higher sensitivity at a given error rate compared to existing tools but need to be validated for real-world applicability. The future will almost certainly bring advances in allowing assignment to modified peptides as well as a refinement of search criteria as large datasets become available for testing. Further, the era of manual curation is coming to a conclusion, with the development of software that can assign a quality or probability score to a peptide assignment on the basis of a variety of spectral features or search scores. These developments will likely culminate with tools that allow rapid optimization for a set of experimental conditions such as MS platform and enzymatic treatment, ultimately streamlining sampling processing in both high- and low-throughput environments.

References

Bafna W and Edwards N (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17**(Suppl 1), S13–S21.

Bartels C (1990) Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical and Environmental Mass Spectrometry*, **19**, 363–368.

Clauser KR, Baker P and Burlingame AL (1999) Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anaytical Chemistry*, **71**, 2871–2882.

Colinge J, Masselot A, Giron M, Dessingy T and Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, **3**, 1454–1463.

Dancik V, Addona TA, Clauser KR, Vath JE and Pevzner PA (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **6**, 327–342.

Eng JK, McCormack AL and Yates JR III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.

Elias JE, Gibbons FD, King OD, Roth FP and Gygi SP (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, **22**, 214–219.

Field HI, Fenyo D and Beavis RC (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, **2**, 36–47.

Griffin PR, MacCoss MJ, Eng JK, Blevins RA, Aaronson JS and Yates JR III (1995) Direct database searching with MALDI-PSD spectra of peptides. *Rapid Communications in Mass Spectrometry*, **9**, 1546–1551.

Havilio M, Haddad Y and Smilansky Z (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Analytical Chemistry*, **75**, 435–444.

Hernandez P, Gras R, Frey J and Appel RD (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, **3**, 870–878.

Johnson RJ and Biemann K (1989) Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical and Environmental Mass Spectrometry*, **18**, 945–957.

Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP and Simpson RJ (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry*, **75**, 6251–6264.

Keller A, Nesvizhskii AI, Kolker E and Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, **74**, 5383–5392.

MacCoss MJ, Wu CC and Yates JR III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical Chemistry*, **74**, 5593–5599.

Mann M and Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399.

Moore RE, Young MK and Lee TD (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, **13**, 378–386.

Nesvizhskii AI, Keller A, Kolker E and Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, **75**, 4646–4658.

Pappin DJC, Hojrup P and Bleasby AJ (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, **3**, 327–332.

Perkins DN, Pappin DJ, Creasy DM and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Pevzner PA, Mulyukov Z, Dancik V and Tang CL (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Research*, **11**, 290–299.

Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ and Yates JR III (2002) Code developments to improve the efficiency of automated MS/MS spectra interpretation. *Journal of Proteome Research*, **1**, 211–215.

Sadygov RG and Yates JR III (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical Chemistry*, **75**, 3792–3798.

Shevchenko A, Wilm M and Mann M (1997) Peptide sequencing by mass spectrometry for homology searches and cloning of genes. *Journal of Protein Chemistry*, **16**, 481–490.

Tabb DL, Saraf A and Yates JR III (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*, **75**, 6415–6421.

Yates JR III, Eng JK and McCormack AL (1995a) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to nucleotide sequences. *Analytical Chemistry*, **67**, 3202–3210.

Yates JR III, Eng JK, McCormack AL and Schieltz D (1995b) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry*, **67**, 1426–1436.

Yates JR III, Eng JK, Clauser KR and Burlingame AL (1996) Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *Journal of the American Society for Mass Spectrometry*, **7**, 1089–1098.

Yates JR III, Zhou J, Griffin PR and Hood LE (1991) Computer aided interpretation of low energy MS/MS mass spectra of peptides. *Techniques in Protein Chemistry II*, **46**, 477–485.

Zhang N, Aebersold R and Schwikowski B (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.

Specialist Review

Interpreting tandem mass spectra of peptides

Richard S. Johnson
Amgen Corporation, Seattle, WA, USA

1. Introduction

Given the success of database search programs at identifying proteins (Clauser *et al*., 1999; Eng *et al*., 1994; Fenyo *et al*., 1998; Mann and Wilm, 1994; Perkins *et al*., 1999), one could reasonably question the importance of knowing how to manually interpret tandem mass spectra (MS/MS) of peptides. This is especially true when working in a high-throughput environment where the data load precludes any sort of manual evaluation. Nevertheless, it can be intuitively grasped that occasionally looking at data might have some benefit. In addition to developing a sense of mass spectral aesthetics, and therefore an eye for data quality, knowing how to sequence peptides is essential when working with organisms for which no extensive sequence database is available. However, for organisms with sequenced genomes, this skill may still be important for validating database matches. Such validation is essential when a database search reveals only one or two peptides matching to a particular protein, which is frequently the case for experiments involving amino acid–specific fractionation (e.g., ICAT; Gygi *et al*., 1999). In addition, anyone who has ever performed a database search using hundreds of spectra obtained from an LC/MS/MS experiment has found that well over half of the spectra do not match with anything. The more curious among us might wonder what gives rise to all the unmatched data, and whether there is anything to be learned from it. For example, the author once sequenced a number of carbamylated peptides (carbamylation was not considered in the database search). This information was useful for subsequent experiments. The scope of the following tutorial is limited to determining sequences from MS/MS spectra resulting from low energy collision induced dissociation of positively charged (protonated) peptides in which the majority of the fragment ions are singly charged.

2. Peptide fragmentation

For low-energy CID MS/MS spectra, the only sequence-defining fragmentations are the so-called *b*-type and *y*-type ions (Figure 1a). These ion types arise via

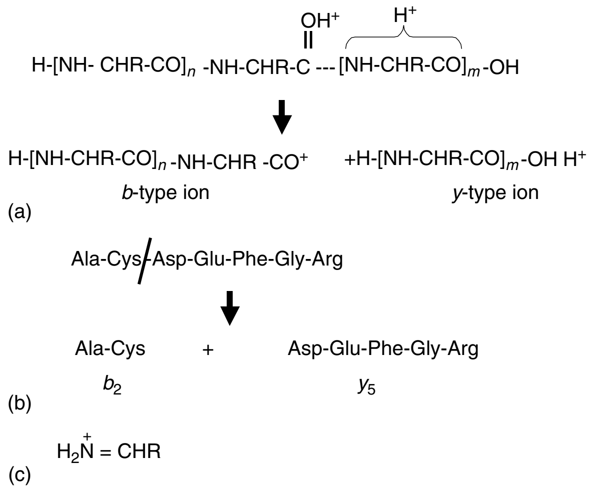


Figure 1 Peptide fragment ions. (a) A doubly protonated peptide has one proton sequestered within the C-terminal portion of the molecule and the other is protonating the carbonyl oxygen. This latter protonation weakens the amide linkage (dotted line) to the point that the peptide ion fragments into a C-terminal y-type ion and an N-terminal b-type ion. (b) This example indicates that b- and y-type ions are numbered according to the number of residues each ion contains. (c) Structure of immonium ions

a protonation of the peptide backbone, which subsequently weakens the amide linkage to the point that cleavage between the nitrogen and carbonyl carbon is favored. The y-type and b-type ions can both lose a molecule of water (18 Da) or ammonia (17 Da) in secondary fragmentations; in addition, the b-type ions can lose a molecule of carbon monoxide (28 Da) to give rise to a-type ions. These sequence-specific ions are labeled according to the number of residues they contain (e.g., Figure 1b). To calculate the mass of a specific b-type ion for a known peptide sequence, one sums the relevant amino acid residue masses (Table 1), and then adds the mass of the N-terminal proton. For y-type ions, the residue masses are summed, and the mass of the C-terminal −OH group is added, plus two additional protons (one for the N-terminus and one to provide the charge).

For the sake of simplicity, Figure 1a depicts b-type ions as acylium cations. However, it is currently thought that a more likely reaction mechanism involves nucleophilic attack of the protonated carbonyl carbon by the adjacent carbonyl oxygen located N-terminal to the cleavage site (Arnott *et al*., 1994; Schlosser and Lehmann, 2000). This would result in a five-member oxazolone ring structure on the C-terminal end of b-type ions. Such a fragmentation mechanism is appealing in that it accounts for certain well-known observations. First, one does not observe b_1 ions (b-type ions comprised only of the N-terminal amino acid in the peptide) for peptides with free N-termini, since there is no carbonyl group to induce cleavage. In contrast, if a carbonyl is added to the N-terminal amino group, either by acylation or carbamylation, such modified peptides can fragment to produce b_1 ions. Second, fragmentation on the C-terminal side of proline residues is much reduced, since the side chain ring structure constrains the attack of the carbonyl.

Table 1 Masses for monoisotopic amino acid residues (defined as −NH-CHR-CO-) calculated using all light isotopes. To calculate the mass of a *y*-type ion, sum the appropriate residue masses and add 19.0184 (three hydrogen plus one oxygen); to calculate the mass of a *b*-type ion, sum the appropriate residue masses and add 1.0078 (hydrogen atom)

Amino acid	Single-letter code	Mass
Glycine	G	57.0215
Alanine	A	71.0371
Serine	S	87.0320
Proline	P	97.0528
Valine	V	99.0684
Threonine	T	101.0477
Cysteine	C	103.0092
Isoleucine	I	113.0841
Leucine	L	113.0841
Asparagine	N	114.0429
Aspartic acid	D	115.0270
Glutamine	Q	128.0586
Lysine	K	128.0950
Glutamic acid	E	129.0426
Methionine	M	131.0405
Histidine	H	137.0589
Oxidized methionine		147.0354
Phenylalanine	F	147.0684
Arginine	R	156.1011
Carbamidomethylcysteine		160.0307
Tyrosine	Y	163.0633
Tryptophan	W	186.0793

A few ion types are observed that are not sequence specific. When a *b*-/*y*-type fragmentation (Figure 1a) occurs twice in the same molecule, one generates the so-called internal fragment ions. These possess neither the C- nor N-termini, and have a structure analogous to a *b*-type ion that is missing a piece of the N-terminus. These are particularly prominent when proline is on the N-terminal side of the internal fragment, and they rarely contain more than two to four amino acids. Immonium ions are composed of protonated single amino acid residues that have eliminated a carbon monoxide molecule (Figure 1c).

Shown in Figure 1a is the presence of a fragmentation-promoting proton located on the carbonyl oxygen. Although in acidic aqueous solutions amines and histidine imidazole rings are preferentially protonated over amide carbonyls, there is less of a difference in basicity in the gas phase. Once in the vacuum of a mass spectrometer, the protons on a peptide ion are relatively free to move about, and this "mobile proton theory" provides an important framework for understanding and predicting how peptides will fragment. An important exception seems to occur when the number of arginine residues exceeds the number of charging protons (Kapp *et al.*, 2003). The guanido group of arginine has such high gas-phase basicity that it tends to sequester protons and makes them unavailable for inducing amide bond fragmentation. In the event that there is no amide protonation, higher collision energy is needed to mobilize the proton on the arginine. The higher collision energy can result in alternative and sometimes unexpected and incomplete fragmentation

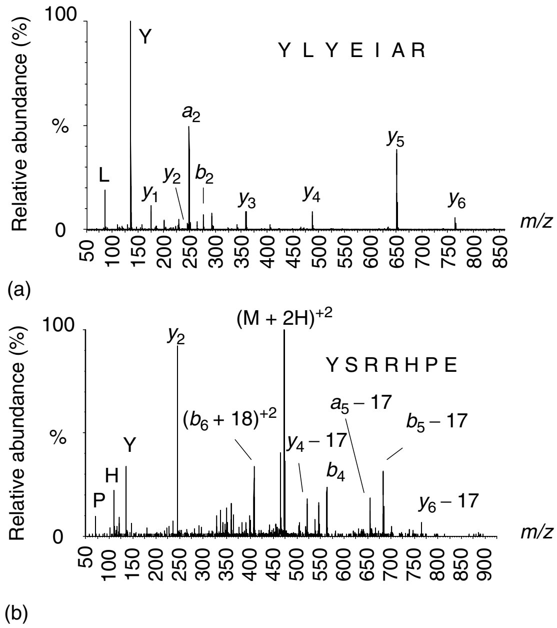


Figure 2 Effect of a mobile proton on peptide fragmentation. (a) A doubly protonated peptide ion containing only one arginine leaves a single mobile proton available for promoting cleavages at amide bonds, as indicated in Figure 1(a). This spectrum exhibits a contiguous series of *y*-type ions that can be easily sequenced. (b) A doubly protonated peptide ion containing two arginine residues, leaving no mobile proton. In this spectrum, there are no contiguous series of ions from which the sequence can be deduced, but there are several unusual fragment ions – doubly charged *b*+18 ions, plus a preponderance of neutral losses from a few *y*-, *b*-, and *a*-type ions

(e.g., Figure 2). The MS/MS spectrum of a doubly protonated peptide ion that contains only one arginine residue (i.e., there is one mobile proton) contains a contiguous series of *y*-type ions and is easily sequenced (Figure 2a). The spectrum of a peptide containing equal numbers of arginines and protons (no mobile proton) exhibits unusual fragment ions that do not delineate the entire sequence (Figure 2b). Electrospray-ionized tryptic peptides tend to produce easily sequenced MS/MS spectra (e.g., Figure 2a), since they typically have no more than one arginine and are usually multiply protonated (i.e., they have a mobile proton). Furthermore, by placing the arginine at the C-terminus, this helps ensure that a complete series of *y*-type ions are formed.

The presence or absence of a mobile proton accounts for some well-known observations. First, cleavage on the C-terminal side of aspartic acid is enhanced for peptide ions lacking a mobile proton (Yu *et al*., 1993; Wysocki *et al*., 2000), which is thought to be due to self-protonation of the amide backbone by the acidic side chain. Second, cleavage on the N-terminal side of proline is enhanced for peptide ions containing a mobile proton (Wysocki *et al*., 2000). This is presumably due to skewing the population of amide protonation to favor N-alkylated amides, which have higher gas-phase basicity (Nair *et al*., 1996). Thus, the N-terminal side of proline is more likely to be protonated, and therefore suffer a cleavage.

3. General approach to sequencing peptides

If a single complete and contiguous series of a single ion type is present in a spectrum, then deducing the sequence is a simple matter of subtracting one fragment ion from the next (e.g., Figure 3). Each mass difference would correspond to one of the amino acid residue masses (Table 1). However, since it is not known if the ion series contains the C- or N-terminus (i.e., whether the series is due to y-type or b-type ions), the directionality of the derived sequence is not known. To determine this, one needs to be able to plausibly connect the series with one of the termini. For tryptic peptides, which have a C-terminal lysine or arginine, this can be done by determining if the lowest mass ion in the series corresponds to a y_1 ion containing either lysine ($y_1 = m/z$ 147) or arginine ($y_1 = m/z$ 175) (Figure 3a). If the peptide is not derived from tryptic proteolysis, then one would look for y_1

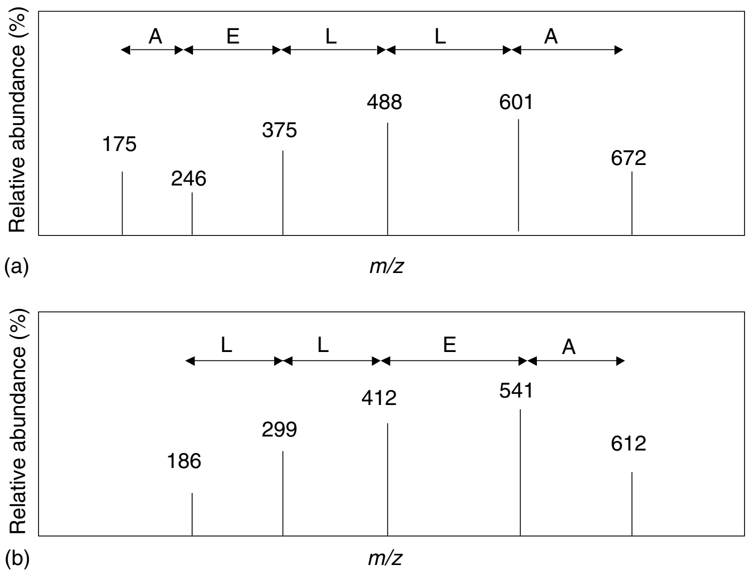


Figure 3 General approach to peptide sequencing. Hypothetical spectra are shown of a peptide with a molecular weight of 785. (a) Fragment ion mass differences delineate a partial sequence; however, the directionality is not known (i.e., the partial sequence could be either AELLA or ALLEA). The lowest mass ion in the series at m/z 175 corresponds to a y_1 ion containing an arginine residue; therefore, a likely explanation is that the series contains the C-terminus – ALLEAR. This partial sequence is 114 Da short of the peptide mass (corresponding to N or two Gs), so two candidate sequences would be NALLEAR or GGALLEAR. (b) The fragment ion mass differences delineate two possible sequences (LLEA or AELL, directionality is unknown). The lowest mass ion in the series does not correspond to any possible y_1 ion ($186 - 19 = 167$, which does not correspond to an amino acid residue mass listed in Table 1). If the ion series was due to b-type ions, then the highest mass ion might be due to the loss of a single C-terminal residue. Subtract 17 Da from the peptide mass and then subtract the highest mass ion in the series ($785 - 17 - 612 = 156$). The residue mass of arginine is 156, so this could be assumed to be a b-type ion series. Hence, a partial sequence can be deduced as LLEAR, which is 185 Da less than the peptide mass. Therefore, there are two or more unsequenced amino acids with a combined mass of 185 Da that is appended to the N-terminus – [185] LLEAR

ions corresponding to any other possible amino acid (y_1 = residue mass plus 19). Alternatively, one can check to see if the highest mass ion in the series is one residue short of accounting for the entire mass of the peptide (e.g., Figure 3b).

Interpreting MS/MS spectra is never as simple as what is depicted in Figure 3. Usually, there is more than one ion series, and, in addition, secondary fragmentations frequently occur (losses of water or ammonia). Internal fragment ions will cause confusion, and sometimes certain y- or b-type ions are missing. For example, it is very common to not observe fragmentation between the two N-terminal amino acids. In these cases, the exact sequence cannot be determined; one can only say that there are two or more amino acids with a certain combined mass that are located at the N-terminus. Additional ambiguity results from the fact that leucine and isoleucine have identical masses and cannot be distinguished. Glutamine and lysine differ by only 0.036 Da, and phenylalanine and oxidized methionine by 0.033 Da. These mass differences can be detected only with higher-mass-accuracy instruments that are well calibrated. Also, one needs to remember that certain pairs of amino acids have similar or identical masses as some single amino acids (G + G = N, A + G = Q, G + V ∼ R, A + D ∼ W, S + V ∼ W). Some of these difficulties are demonstrated in the examples described below.

Different types of instruments will provide data with different characteristics. Two tryptic peptide examples will be shown below – one from an ion trap (Cooks and Kaiser, 1990) and the other from a quadrupole/time-of-flight (Qtof) hybrid (Morris *et al*., 1996). The ion trap that was used in this example exhibited lower resolution (unit resolved) and lower mass accuracy (+/− 0.4 Da); another characteristic of ion trap spectra is that part of the lower mass end of the MS/MS spectrum is absent. In contrast, the Qtof mass spectrometer provided full scans with sufficient resolution and accuracy to distinguish lysine from glutamine. Another difference is that ion trap spectra contain abundant high-mass b-type ions, whereas Qtof MS/MS spectra rarely exhibit b-type ions containing more than a few amino acids.

3.1. *Example #1: ion trap spectrum*

The approach taken in this example (Figure 4) is to first see if there are any partial sequences that can be easily determined from the least noisy region of the spectrum. For MS/MS spectra of multiply charged peptide ions, the least noisy region is located above the precursor m/z value (Figure 4b). In this region, there are two ion series (ions marked * or ˆ); the ion series marked by an asterisk also contains a series due to neutral loss of water. These ion series define two partial sequences – LSLV and LVES – however, the directionality is not known.

Directionality is surmised by trying to connect an ion series to one of the termini by assuming the series is either b-type or y-type (as described in Figure 3). Ion trap MS/MS spectra are missing the lower m/z end, so one cannot expect to find a y_1 ion to help identify the y-type series. On the other hand, high molecular weight b-type ions are frequently observed in ion trap spectra, so the next step is to see if the higher m/z ion from each series will correspond to the loss of arginine or lysine

to form the highest mass *b*-type ion. In this example, the calculation is done by subtracting 17 (the mass of the C-terminal OH group) from the peptide molecular weight (1228, in this example), and then subtracting the highest *m/z* value from each of the two series (thus, $1228 - 17 - 987 = 224$ and $1228 - 17 - 1083 = 128$). The mass difference of 128 Da corresponds to the residue mass of lysine, and so the series marked by an asterisk can now be assumed to be a *b*-type ion series.

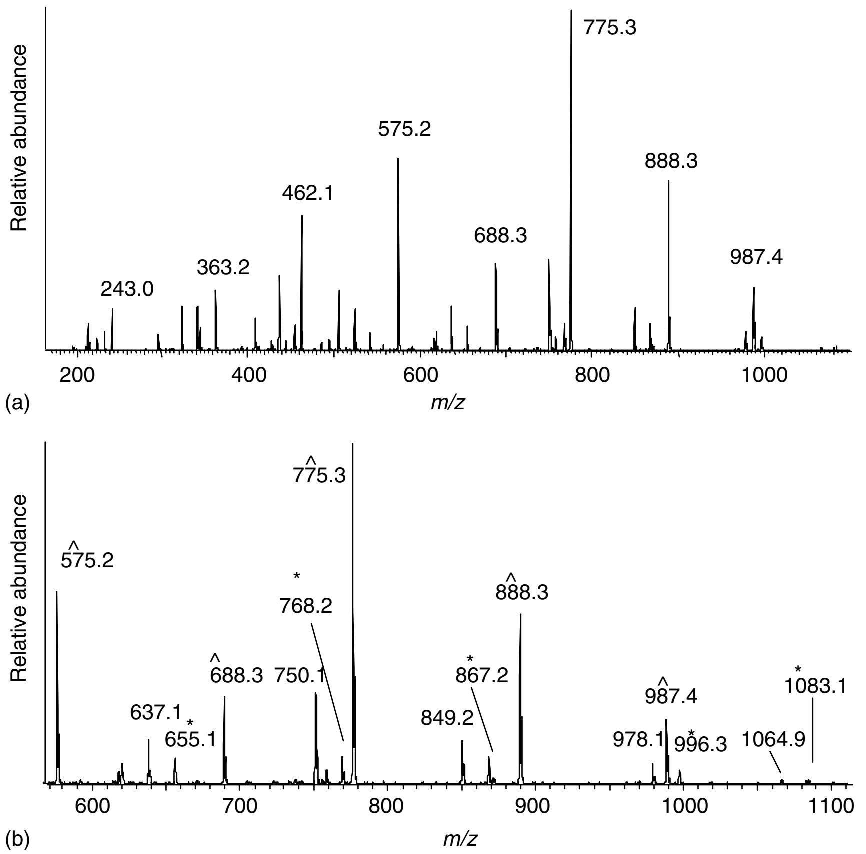


Figure 4 Sequencing a peptide of molecular weight 1228.7 from an ion trap MS/MS spectrum of a doubly charged precursor at *m/z* 615.4. (a) Entire spectrum is shown with a few of the major peaks labeled. (b) The region above the precursor ion *m/z* usually contains the fewest ions and exhibits the simplest fragmentation pattern. One series of ions (labeled ˆ) have mass differences corresponding to the partial sequence LSLV. The other series of ions (labeled *) all have satellite peaks 18 Da lower and delineate a different partial sequence LVES. The ion at *m/z* 1083 has the correct mass for a *b*-type ion resulting from the loss of a single lysine residue; therefore, the ion series marked with an asterisk is hypothesized to be a *b*-type ion series that delineates the partial sequence LVESK. The other series is presumed to be a *y*-type series. (c) From the partial sequence LVESK, the low-mass *y*-type ions (y_1 to y_5) can be calculated (see Table 1) –*m/z* 147, 234, 363, 462, and 575. The ions matching these calculated *y*-type ions are labeled (ˆ), and are contiguous with the similarly labeled (presumed *y*-type ions) in panel (b). The low-mass *b*-type ions can be calculated from the high-mass *y*-type ions (see text) –*m/z* 243, 342, 455, 542, and 655 – which are contiguous with the similarly labeled ions in panel (b)

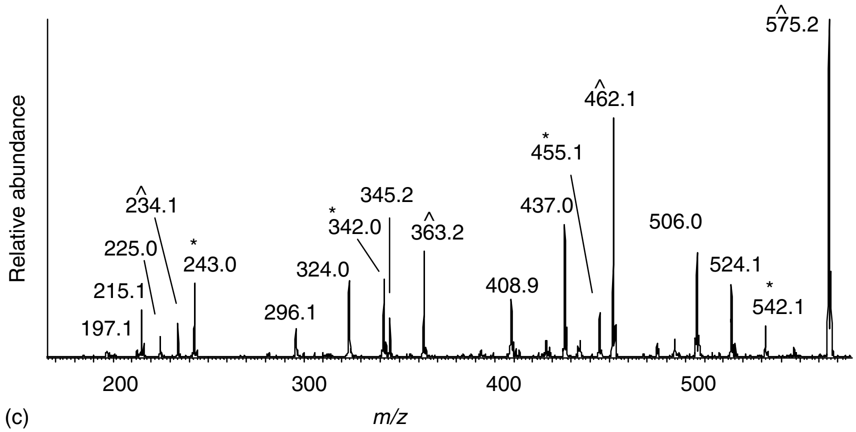


Figure 4 (*continued*)

If no series can be identified as a b-type series, then it is sometimes possible to see if the mass difference obtained could correspond to lysine or arginine plus another amino acid. For example, if m/z 1083 was absent and only m/z 996 was visible (the next ion in the series below 1083), then the mass difference would be 215. Since no combination of lysine or arginine plus another amino acid can equal 224, and since serine plus lysine equals 215, one could still tentatively hypothesize that the asterisk-labeled series is due to b-type ions. In any case, two partial sequences are known at this point – the sequence LVESK (derived from the b-type ion series), plus VLSL located toward the N-terminal part of the sequence (which by the process of elimination is now assumed to be a y-type ion series).

The next step is to connect these two sequences and possibly extend them. From the partial sequence LVESK, the low-m/z y-type ions can be calculated – add 19 to the residue mass of lysine to obtain y_1, add the residue mass of serine to the y_1 mass to get y_2, and so on (Figure 4c). The mass of y_1 is usually below the mass range of an ion trap MS/MS spectrum, and one would not expect to see it. The other y-type ions (y_2 through y_5) are all present. The y_5 ion happens to be the lowest m/z ion in the series that defined the partial sequence VLSL (Figure 4b), which indicates that the two partial sequences are linked (VLSLLVESK). Similarly, low-mass b-type ions can be calculated from the high-m/z y-type ions by adding the mass of two protons to the peptide molecular weight and then subtracting the y-type ion mass (i.e., b-type ions = peptide MW + 2 hydrogen atoms − y-type ion). For example, the lowest mass b-type ion is calculated from the highest mass y-type ion (b-type ion $= 1228 + 2 - 987 = 243$). This b-type ion is composed of more than one amino acid, most likely two; hence, one could tentatively assign the ion at m/z 243 to be b_2. The b_3 through b_6 can be calculated in a similar fashion, all of which are found in the spectrum. The b_6 ion matches the lowest mass b-type ion in the series that defined the partial sequence of LVESK, which helps confirm that the two partial sequences are linked (VLSLLVESK). The only missing piece is at the N-terminal region where there is what is presumed to be a dipeptide of mass 242 Da. This region cannot be sequenced, since there is no b_1 ion or y_{10} ion present.

The sequence, as best can be defined by low-energy CID, is [242]VLSLLVESK, where the number in brackets represents the combined mass of two or more amino acids. Also, the notation "L" implies either leucine or isoleucine, since the two cannot be distinguished.

At this point, one should verify that the majority of fragment ions can be accounted for, given this hypothetical sequence. This is accomplished by checking to see if there are any doubly charged b- or y-type ions, which is done by adding the mass of a proton to the singly charged b- or y-type ion and dividing the sum by two. It is unusual to see multiply charged fragment ions unless the precursor ion had more than two charges, although higher-mass doubly charged y-type ions are sometimes observed from doubly charged precursor ion. Internal fragments (fragments not containing either the C- or N-terminal amino acids) with 2 to 5 residues are sometimes seen, and these m/z values are calculated by summing the amino acid residue masses and then adding a proton. Finally, and most likely, are ions resulting from neutral losses of water and ammonia from the b- and y-type ions. Given these additional fragment ions, plus the y- and b-type fragment ions, all of the major ions can be accounted for, and in fact, the correct sequence is ELVISLIVESK. Losses of water from b-type ions are particularly prominent in this example (Figure 4), which is due to the presence of an N-terminal glutamic acid that readily formed an N-terminal pyroglutamyl N-terminus during the fragmentation process. Similarly, prominent losses of ammonia are seen for peptides with glutamine at the N-terminus.

3.2. *Example #2: Qtof spectrum*

Most tryptic peptides have lysine or arginine at their C-terminus, and the approach taken in this example (Figure 5) is to first see if there are any y_1 ions containing lysine or arginine (m/z 147.113 and 175.119, respectively). The aim is to build the sequence toward the N-terminus in a stepwise fashion. In this approach, the directionality of the sequence is not an issue, since one starts out with what is presumed to be a C-terminal y_1 ion. In this spectrum (Figure 5a), an ion at m/z 147.111 is observed, and therefore a C-terminal lysine is presumed to be present. The mass of the y_1 ion is subtracted from higher-m/z fragment ions in a search for mass differences corresponding to an amino acid residue mass, and two such ions are found at m/z 262.141 (y_1 plus aspartic acid) and 333.177 (y_1 plus tryptophan). Since tryptophan has about the same mass as aspartic acid plus alanine, one could potentially account for both ions by assuming that the penultimate residue is Asp and that the y_2 ion is m/z 262.141. Using the same approach to find y_3 ions, m/z 262.141 is subtracted from higher-m/z fragment ions, and two ions are found at m/z 333.177 and 409.283, corresponding to alanine and phenylalanine. Whereas the mass difference between m/z 333.177 and 262.141 is almost exactly the correct mass for alanine (Table 1), the mass difference between m/z 409.283 and 262.141 is in error by 0.074 Da compared to the correct mass for phenylalanine. For Qtof data, this error is too much, and the y_3 ion is probably m/z 333.177. At this point, likely $y_1 - y_3$ ions have been identified that define a partial sequence at the C-terminal end of the peptide (ADK). Continuing this process, two potential

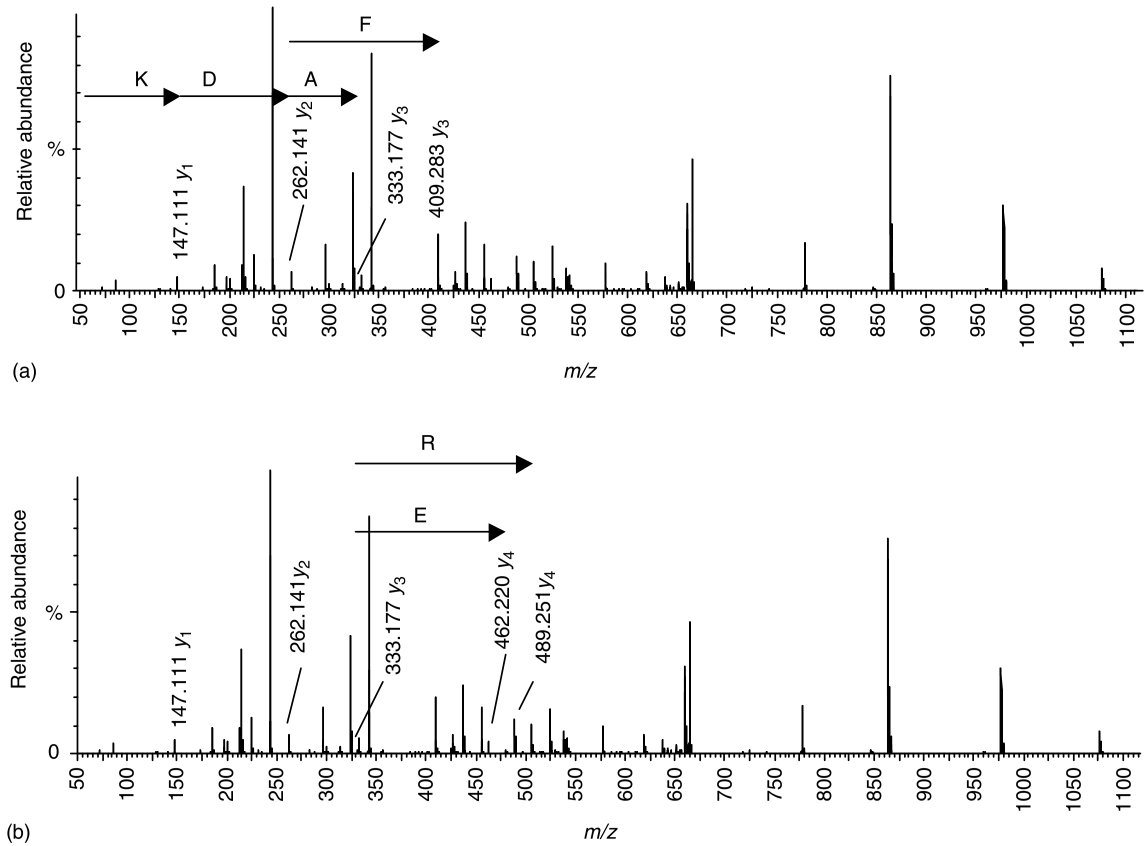


Figure 5 Sequencing a peptide of molecular weight 1317.67 from a Qtof MS/MS spectrum of a doubly charged precursor at *m/z* 659.8. (a) The y_1 ion representing a C-terminal lysine is present at *m/z* 147.111. There are two ions at *m/z* 262.141 and 333.177 that differ in mass from the y_1 ion by an amino acid residue mass (aspartic acid and tryptophan, respectively). Since the residue mass of tryptophan is approximately the same as the combined mass of aspartic acid and alanine, an additional ion is accounted for by assuming that the C-terminal sequence is ADK (y_2 = *m/z* 262 and y_3 = *m/z* 333.177) rather than WK (y_2 = 333.177). The mass accuracy of the Qtof instrument excludes an alternative y_3 ion at *m/z* 409.283. (b) There are two potential y_4 ions; however, for *m/z* 489.251, there are no other ions at higher *m/z* that differ by an amino acid residue mass (i.e., it is a dead end). (c) The putative y_4 ion at *m/z* 462.220 is the low-mass end of a series of ions that unambiguously define a partial sequence containing six amino acids. (d) Most of the remaining ions not hypothesized to be *y*-type ions (as described in panels a–c) are identified as *a*- or *b*-type ions, plus some neutral losses

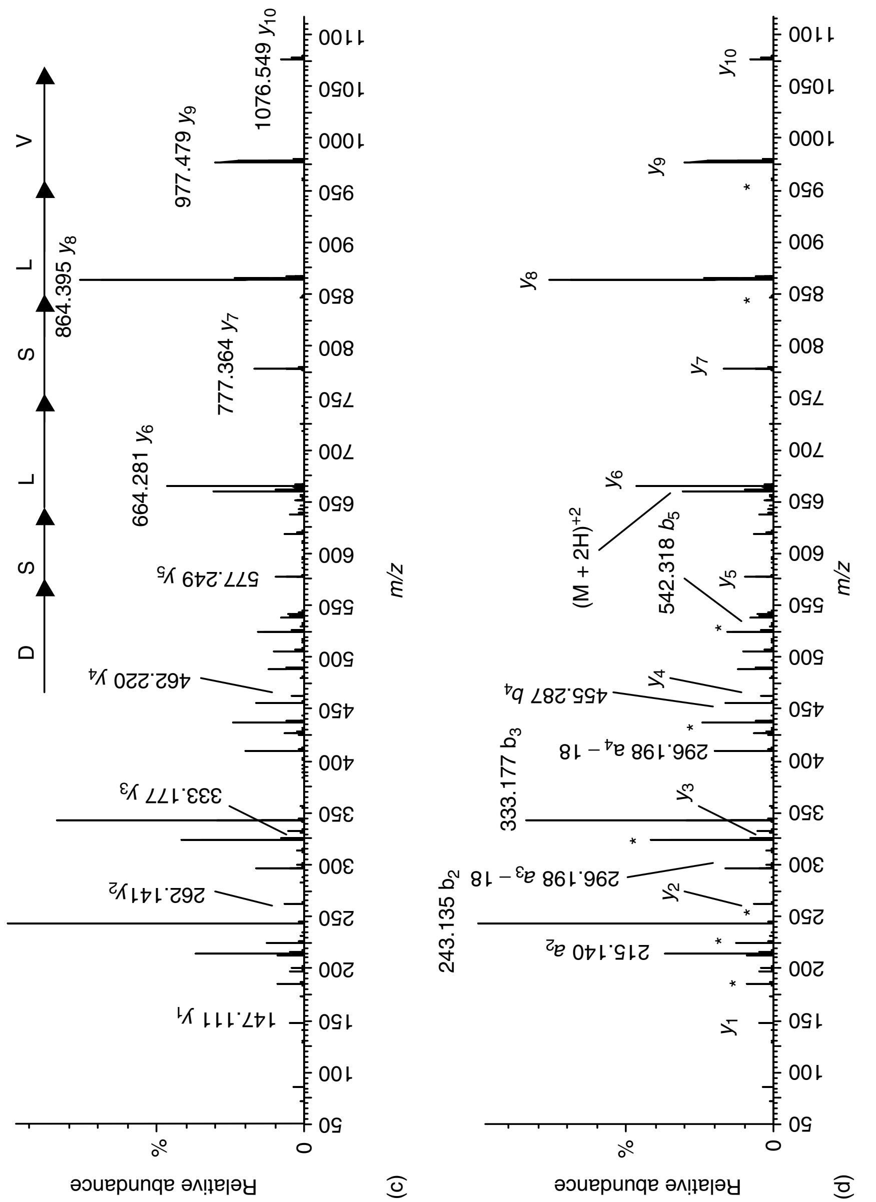


Figure 5 *(continued)*

y_4 ions are identified at m/z 462.220 and 489.251, corresponding to additions of glutamic acid and arginine respectively. The errors in mass differences between these two ions and the y_3 ion at m/z 333.177 are both sufficiently small that it is not possible to choose between the two. Hence, further sequence extensions have to proceed, keeping both possibilities in mind. However, there are no ions above m/z 489.251 whose mass difference corresponds to a single residue. In contrast, the ion at m/z 577.249 is almost exactly the mass of aspartic acid above m/z 462.220. The remainder of the spectrum contains a series of ions from m/z 664.281 to 1076.549 whose mass differences delineate a partial sequence of SLSLV.

From the highest-mass y-type ion at m/z 1076.549, one can calculate a low-mass b-type ion at m/z 243.14 (b-type ions = peptide MW + 2 hydrogen atoms − y-type ion). This b-type ion is composed of more than one residue, most likely two. As in Example #1, the next step is to verify that the remaining abundant ions can be accounted for as either b- or y-type ions, multiply charged ions, internal fragment ions, neutral losses, or immonium ions. Given the limitations of low-energy CID, one can rationally determine the sequence to be [242]VLSLSDEADK, where 242 implies the presence of more than one residue whose residue mass sum is 242. Likewise, L signifies either isoleucine or leucine; however, using the answer to Example #1 as a clue, the reader may be able to correctly guess the entire sequence.

4. Conclusion

Deriving sequences from tandem mass spectra is slow and tedious, but not particularly difficult. Most people have acquired the necessary mathematics skills (addition and subtraction) by the age of eight, although the possibility of multiple charging resulting from electrospray ionization may require more advanced math skills (multiplication and division). The main impediment for most people is that there is not enough time to interpret all the spectra generated. For this reason, there have been a number of attempts to automate the process using various computer programs and algorithms (e.g., Bartels, 1990; Biemann *et al*., 1966; Chen *et al*., 2001; Dancik *et al*., 1999; Fernandez-de-Cossio *et al*., 2000; Ma, 2003); although this is an important topic, further discussion of these programs is beyond the scope of this tutorial.

References

Arnott D, Kottmeier D, Yates N, Shabanovitz J and Hunt D (1994) Fragmentation of multiply protonated peptides under low energy conditions presented at the 42nd ASMS. *Conference on Mass Spectrometry and Allied Topics, Chicago, 29 May – 3 June 1994*.

Bartels C (1990) Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical and Environmental Mass Spectrometry*, **19**, 363–368.

Biemann K, Cone C, Webster BR and Arsenault GP (1966) Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *Journal of the American Chemical Society*, **88**, 5598–5606.

Chen T, Ming-Yang K, Tepel M, Rush J and Church GM (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **8**, 325–337.

Clauser K, Baker P and Burlingame A (1999) Role of accurate mass measurement in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **71**, 2871–2882.

Cooks RG and Kaiser R Jr (1990) Quadrupole ion trap mass spectrometry. *Accounts of Chemical Research*, **23**, 213–219.

Dancik D, Addona T, Clauser K, Vath J and Pevzner P (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **6**, 327–342.

Eng JK, McCormack AL and Yates IIIJR (1994) An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.

Fenyo D, Qin J and Chait BT (1998) Protein identification using mass spectrometric information. *Electrophoresis*, **19**, 998–1005.

Fernandez-de-Cossio J, Gonzalez J, Satomi Y, Shima T, Okumura N, Besada V, Betancourt L, Padron G, Shimonishi Y and Takao T (2000) Automated interpretation of low-energy CID spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. *Electrophoresis*, **21**, 1694–1699.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Kapp E, Schutz F, Reid G, Eddes J, Moritz R, O'Hair R, Speed T and Simpson R (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry*, **75**, 6251–6264.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A and Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communication in Mass Spectrometry*, **17**, 2337–2342.

Mann M and Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399.

Morris H, Paxton T, Dell A, Langhorne J, Berg M, Bordoli R, Hoyes J and Bateman R (1996) High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole / orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communication in Mass Spectrometry*, **10**, 889–996.

Nair H, Somogyi A and Wysocki V (1996) Effect of alkyl substitution at the amide nitrogen on amide bond cleavage: electrospray ionization/surface-induced dissociation fragmentation of substance P and two alkylated analogs. *Journal of Mass Spectrometry*, **31**, 1141–1148.

Perkins D, Pappin D, Creasy D and Cottrell J (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Schlosser A and Lehmann W (2000) Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *Journal of Mass Spectrometry*, **35**, 1382–1390.

Wysocki V, Tsaprailis G, Smith L and Breci L (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, **35**, 1399–1406.

Yu W, Vath J, Huberty M and Martin S (1993) Identification of the facile gas-phase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix-assisted laser desorption time-of-flight mass spectrometry. *Analytical Chemistry*, **65**, 3015–3023.

Specialist Review

FT-ICR

Mark P. Barrow, William I. Burkitt and Peter J. Derrick
University of Warwick, Coventry, UK

1. Introduction

It is Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry (Amster, 1996; Comisarow and Marshall, 1974a,b,c; Jacoby *et al*., 1992; Marshall and Schweikhard, 1992; Marshall *et al*., 1998) which among mass spectrometric methods (*see* Article 10, **Hybrid MS**, Volume 5, Article 7, **Time-of-flight mass spectrometry**, Volume 5, Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5, and Article 75, **Mass spectrometry**, Volume 6) holds the greatest potential at the forefront of research because of the inherently ultrahigh resolution and mass accuracy which can be achieved. FT-ICR mass spectrometry allows unequivocal mass assignment and resolution of species that could not otherwise be distinguished if using some other types of mass spectrometer. This latter capability is particularly relevant for the analysis of multi-component mixtures (Barrow *et al*., 2003; Wu *et al*., 2003b). All varieties of mass spectrometer must involve the analysis of gas-phase ions to obtain a mass-to-charge ratio (m/z) and ion sources must be used to generate the gas-phase ions from a neutral sample. FT-ICR is no different in these respects.

2. Basic principles of FT-ICR mass spectrometry

In FT-ICR mass spectrometry, ions are usually generated externally in a separate ion source and then injected into a container known as the cell (see Figure 1), which is typically cubical or cylindrical in geometry. The cell is located within a strong magnetic field, typically generated using a superconducting magnet. Charged particles moving in the presence of a magnetic field, with a component of their velocity perpendicular to the magnetic field axis, will experience a force known as the Lorentz Force. The axes of this force, motion of the (positively) charged particle, and magnetic field are mutually perpendicular, as dictated by Fleming's Left Hand Rule. A charged particle, such as an ion, will begin to precess about the center of the magnetic field axis, resulting in an orbit. This motion is known as the cyclotron motion. The cyclotron frequency is related to the mass-to-charge

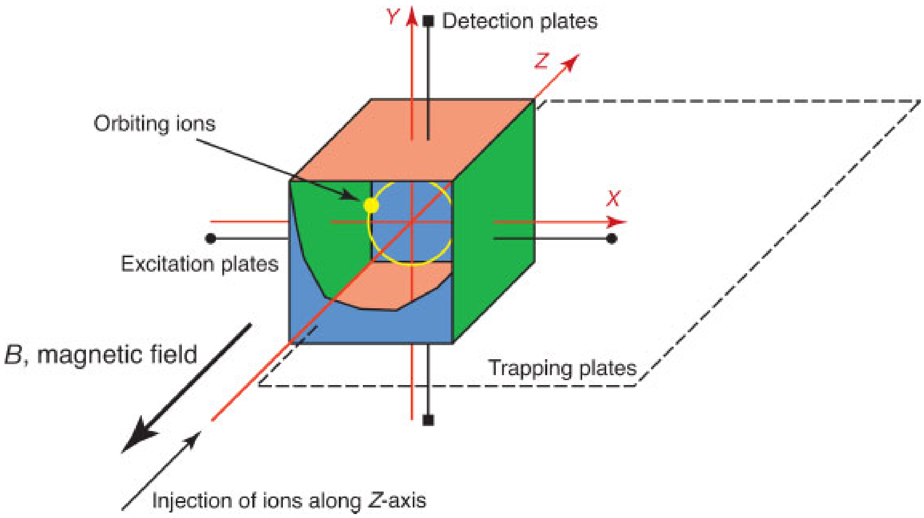


Figure 1 Schematic representation of a cubic FT-ICR cell. The cell is aligned with the bore of the magnet so that the magnetic field axis is coaxial with the trapping axis (z-axis). The excitation and detection plates can be seen, with the trapping electrodes at each end of the cell, and the orbiting ions are shown in yellow

ratio of the ion, as shown in equations (1a) and (1b):

$$\omega = \frac{qB}{m} \tag{1a}$$

$$f = \frac{qB}{2\pi m} \tag{1b}$$

where q is the charge on the ion, B is the magnetic field strength, m is the mass of the ion, ω is the cyclotron frequency in terms of radians per second, and f is the cyclotron frequency in terms of Hertz. Ions of lower m/z have higher cyclotron frequencies than ions of higher m/z.

The FT-ICR cell consists of the two excitation and two detection plates, but, in order to restrain the ions' motion along the axis of the magnetic field, it is necessary to include two trapping plates. A low potential (typically of an order of 1 V) is applied to the trapping plates. When the ions enter the FT-ICR cell, the radii of the cyclotron orbits are too small to be detectable. The ions must therefore be excited to detectable radii and this occurs through usage of a radio frequency (RF) potential applied to two excitation plates operating at the resonant frequency (i.e., resonant with the cyclotron frequency) of the ions. All ions are excited to orbits of the same radii, though their cyclotron frequencies differ. Figure 2 depicts the excitation of ions to a detectable orbit radius. Detection of the ions occurs as the ion packets pass two detector plates. As the ion packets move past these plates, charge moves within the detection circuit to counteract the proximity of the ions. This current can be measured as a function of time and it is from here that the raw data (known as a transient, time-domain data, or sometimes as a free induction decay or FID) is

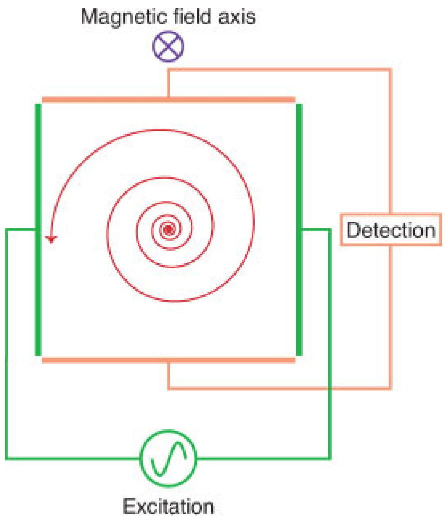


Figure 2 A cross section of a cubic FT-ICR cell, depicting excitation of ions (in red) through the application of an RF potential to the excitation electrodes. Once the ions have been excited to a cyclotron motion of suitable radius, the image current of the orbiting ions can be detected on the detection plates

obtained. It should be noted that the ions are remeasured because of the fact that they repeatedly pass the detector plates, as nondestructive detection is employed.

The raw data will represent the detection at the same time of all the ions, with their different cyclotron frequencies. Information about the ion packets' frequencies is obtained through usage of the mathematical procedure known as a Fourier transform (FT) (Marshall and Comisarow, 1975; Marshall and Verdun, 1990) by which frequency information is obtained from time-domain data. A spectrum is produced where signal intensity is plotted as a function of the cyclotron frequencies of the ions present. From equations (1a) and (1b), it can be seen that the cyclotron frequency is related to m/z. It is therefore possible to convert the plot and to perform a calibration (Shi *et al*., 2000), thereby creating a mass spectrum in which signal intensity is plotted as a function of m/z. The conversion of raw data to a mass spectrum using a Fourier transform is depicted in Figure 3.

3. Advantages of FT-ICR mass spectrometry: mass accuracy and resolution

The excellent mass accuracy and resolution of FT-ICR mass spectrometers are amongst the technique's strongest assets. Mass accuracy is a measurement of how well the observed m/z correlates with the theoretical value. Many other types of modern commercial mass spectrometers cite specifications for mass accuracy of the order of tens of ppm over a specific mass range, whereas specifications for FT-ICR mass spectrometers typically cite mass accuracies of 1 ppm or better (lower). Equation (2) is the relationship for determining the mass accuracy of a result when

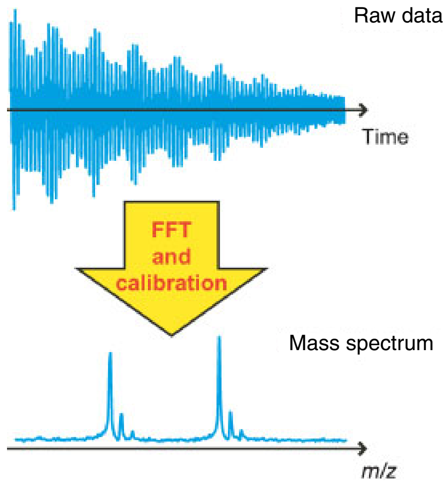


Figure 3 Schematic diagram showing the measured data of a FT-ICR experiment and the subsequent fast fourier transform (FFT) and calibration of the measured data into the final mass spectrum

comparing the observed m/z with the value believed to be "true," called here the theoretical m/z:

$$\text{Mass accuracy} = \frac{m_{\text{observed}} - m_{\text{theory}}}{m_{\text{theory}}} \times 1000000 \qquad (2)$$

m_{observed} is the m/z of the peak of interest in the mass spectrum obtained and m_{theory} is the theoretical m/z that would have been expected for the species. The mass accuracy is defined in terms of ppm (equation 2).

Resolution is extremely important for resolving closely spaced signals, such as in the cases of multiply charged ions (as generated by electrospray ionization, for instance). The ability to use instruments with high resolution can sometimes make the difference between identifying and not identifying a species of interest. Equation (3) defines the resolution:

$$\text{Resolution} = \frac{m}{\Delta m} \qquad (3)$$

where m is the m/z of the peak of interest and Δm is the width, in terms of m/z, determined using any one of the many methods of mass spectrometry. Resolution can be measured using the "10% valley" definition, as most commonly applied when using magnetic-sector instruments, or by using the full width at half maximum (FWHM) definition, which is typically used during the analysis of data acquired from time-of-flight (TOF) (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) or FT-ICR mass spectrometers, illustrated in Figure 4. FT-ICR mass spectrometers can routinely reach resolutions of hundreds of thousands in "broadband mode" ("normal" experimental conditions) or even reach a resolution of a few million in "heterodyne mode" (where a very narrow

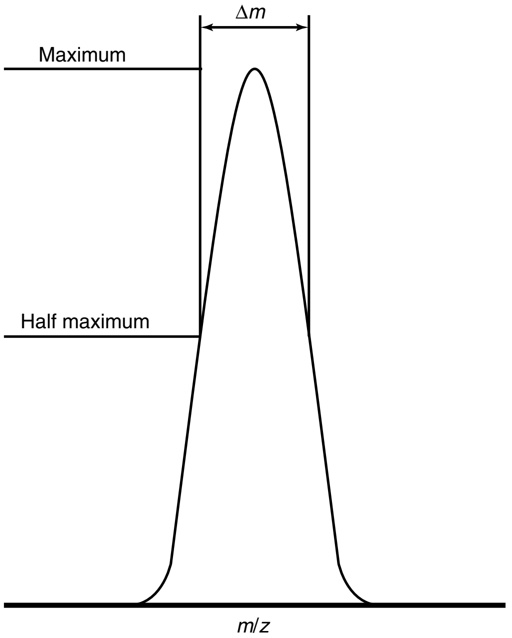


Figure 4 Definitions of resolution, showing the "full width at half maximum" definition, which is most typically used in conjunction with time-of-flight and FT-ICR mass spectrometers

m/*z* range is studied but under very high-resolution conditions), compared with commercial time-of-flight mass spectrometers approaching approximately 10 000 and commercial magnetic-sector mass spectrometers approaching a resolution of 100 000.

4. Electrospray ionization

Electrospray ionization (ESI) (Dole *et al*., 1968; Fenn *et al*., 1989; Mann, 1992; Yamashita and Fenn, 1984a,b) is the ionization technique nowadays most frequently coupled with FT-ICR mass spectrometry. ESI has the advantage of minimizing fragmentation, which is particularly important for labile biological molecules, where fragments must not interfere with the mass spectrum when trying to determine the original constituents. ESI frequently leads to the formation of multiple charged ions, particularly in the case of large macromolecules. As mass spectrometry is based upon the determination of *m*/*z*, not mass alone, it is frequently the case that a mass spectrum will contain the same molecules but in a variety of charge states, and therefore at several different *m*/*z* values. As ions become more highly charged, the *m*/*z* becomes lower and the spacing between the "isotopomers" (peaks due to the presence of other isotopes) becomes narrower. As a result, it becomes more difficult to resolve the signals and the resolution of the mass analyzer becomes more important. Note that the charge state of an ion can be determined by examining the spacing between the isotopomer signals. The growth of electrospray

ionization can, therefore, be seen to be promoting the growth in FT-ICR mass spectrometry, for FT-ICR mass spectrometers offer the ultrahigh resolution that is particularly well suited to a coupling with ESI. For these reasons, ESI FT-ICR mass spectrometry has become increasingly important for the study of biological molecules in recent years.

5. Tandem mass spectrometry techniques

For structural elucidation, "tandem mass spectrometry" (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5) experiments are performed, in which particular ions can be selected and dissociated. The resulting dissociation pattern provides structural information. One of the most common forms of tandem mass spectrometry experiment performed on FT-ICR mass spectrometers is sustained off-resonance irradiation collision-induced dissociation (SORI-CID) (Amster, 1996; Laskin *et al*., 2000; Palmblad *et al*., 2000). SORI-CID is based upon multiple collisions of the selected ions with gas particles, resulting in a "slow heating effect" and the ions' internal energies are increased, leading to dissociation. Other varieties of tandem mass spectrometry experiment include infra-red multiphoton dissociation (IRMPD) and electron capture dissociation (ECD). IRMPD (Stace, 1998; Tsybin *et al*., 2003) entails the usage of a laser, most commonly a CO_2 laser, to irradiate ions whilst trapped in the cell, and can be compared with the "slow heating effect" of SORI-CID. IRMPD has the advantage of avoiding the requirement to put a collision gas into the cell, as a higher pressure leads to a more rapid damping of the transient (time-domain data) and therefore decreases the resolution. SORI-CID and IRMPD are usually considered to be ergodic processes (Derrick *et al*., 1995), where the weakest bonds are cleaved. When studying peptides and proteins, backbone amide bonds and posttranslational modifications, tend to be cleaved, leaving to the formation of b-fragments and y-fragments from peptides and proteins (Sheil *et al*., 1990). ECD (Zubarev *et al*., 1998, 1999; Zubarev, 2003) involves a beam of low-energy electrons (0.2 eV or less, for instance) being emitted into the cell. Ions capture these low-energy electrons and dissociation occurs. The dissociations are thought to be nonergodic processes. With peptides and proteins, N–C and S–S bonds are preferentially cleaved. Utilization of ECD leads primarily to the formation of c and $z^\bullet$ fragment ions, and one advantage is that noncovalent bonds and phosphorylation sites (*see* Article 61, **Posttranslational modification of proteins**, Volume 6, Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6, and Article 73, **Protein phosphorylation analysis – a primer**, Volume 6) are left intact. ECD mass spectra can therefore be relatively less complex due to avoidance of unwanted fragmentation.

Fourier transform ion cyclotron resonance mass spectrometry represents an increasingly powerful tool for the study of biological molecules, especially within the context of "proteomics" (*see* Article 2, **Sample preparation for proteomics**, Volume 5) (Aebersold and Mann, 2003; Sali *et al*., 2003; Tyers and Mann, 2003). An example of a mass spectrum of a biological sample, in this case the protein known as ubiquitin, is shown in Figure 5. FT-ICR mass spectrometry, in particular, offers the greatest potential for the analysis of complex mixtures such as enzyme

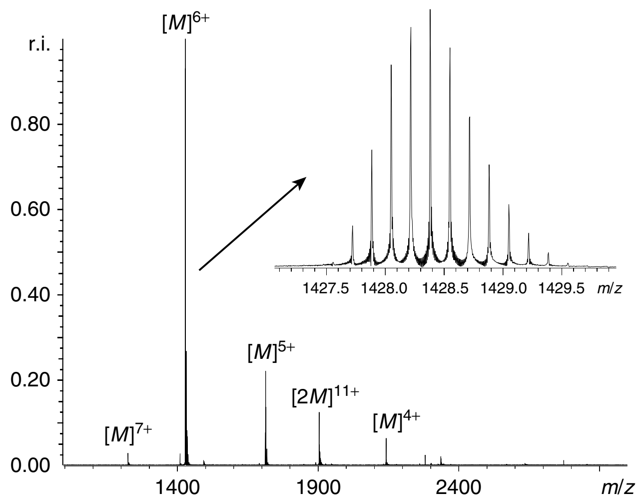


Figure 5 ESI FT-ICR mass spectrum of 20-µM ubiquitin in an aqueous 10-mM ammonium acetate solution, measured in terms of relative intensity versus m/z. The ubiquitin ions are measured in the mass spectrum in a number of charge states, there were also some noncovalently bound ubiquitin ions present. The inset shows a closeup of the ubiquitin ions in the 6+ charge state, the isotopomers were clearly resolved

digest owing to its inherently high mass accuracy, resolving power, and dynamic range (Wu, 2003a). These attributes allow for the positive identification of a greater number of species within a mixture than is possible with other mass spectrometric techniques, including species of low abundance such as posttranslationally modified proteins. However, with the advent of the low-excitation ionization afforded by ESI, mass spectrometry has increasingly been used for the characterization of the solution-phase properties of biomolecules. The ability to disperse into the gas-phase and ionize noncovalent complexes has allowed for the study protein–protein and protein–ligand interactions (Loo, 1997). The high resolution and mass accuracy of FT-ICR mass spectrometry offers decisive advantages when studying such interactions. For example, the ability of FT-ICR to resolve the isotopes of proteins up to 100 kDa in mass (Kelleher *et al*., 1997) allows the definitive determination of the oligomeric state of a protein, or the ligands binding to a particular complex to be decisively determined. The structural characterization of such complexes is currently being attempted using tandem mass spectrometry techniques such as SORI-CID and ECD. Increases in sensitivity of detection and magnetic field strength will only serve to increase the potential of FT-ICR mass spectrometry to allow the understanding of life at the molecular level.

References

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
Amster IJ (1996) Fourier transform mass spectrometry. *Journal of Mass Spectrometry*, **31**, 1325–1337.

Barrow MP, McDonnell LA, Feng X, Walker J and Derrick PJ (2003) Determination of the nature of naphthenic acids present in crude oils using nanospray Fourier transform ion cyclotron resonance mass spectrometry: The continued battle against corrosion. *Analytical Chemistry*, **75**, 860–866.

Comisarow MB and Marshall AG (1974a) Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, **25**, 282–283.

Comisarow MB and Marshall AG (1974b) Selective-phase ion cyclotron resonance spectroscopy. *Canadian Journal of Chemistry*, **52**, 1997–1999.

Comisarow MB and Marshall AG (1974c) Frequency-sweep Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, **26**, 489–490.

Derrick PJ, Lloyd PM and Christie JR (1995) In *13th International Mass Spectrometry Conference: Physical Chemistry of Ion Reactions*, Cornides I, Horváth G and Vékey K (Eds.), John Wiley & Sons: Budapest.

Dole M, Mack LL, Hines RL, Mobley RC, Ferguson LD and Alice MB (1968) Molecular beams of macroions. *Journal of Chemical Physics*, **49**, 2240–2249.

Fenn JB, Mann M, Meng CK, Wong SF and Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**, 64–71.

Jacoby CB, Holliman CL and Gross ML (1992) Fourier transform mass spectrometry: features, principles, capabilities, and limitations. In *Mass Spectrometry in the Biological Sciences: A Tutorial*, Jacoby CB, Holliman CL and Gross ML (Eds.), Kluwer Academic Publishers: Dordrecht, pp. 93–116.

Kelleher NL, Senko MW, Siegel MM and McLafferty FW (1997) Unit resolution mass spectra of 112 kDa molecules with 3 Da accuracy. *Journal of the American Society for Mass Spectrometry*, **8**, 380–383.

Laskin J, Byrd M and Futrell J (2000) Internal energy distributions resulting from sustained off-resonance excitation in FTMS. I. Fragmentation of the bromobenzene radical cation. *International Journal of Mass Spectrometry*, **195/196**, 285–302.

Loo JA (1997) Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrometry Reviews*, **16**, 1–23.

Mann M (1992) Electrospray mass spectrometry. In *Mass Spectrometry in the Biological Sciences: A Tutorial*, Mann M (Ed.), Kluwer Academic Publishers: Dordrecht, pp. 145–163.

Marshall AG and Comisarow MB (1975) Fourier and Hadamard transform methods in spectroscopy. *Analytical Chemistry*, **47**, 491A–504A.

Marshall AG and Schweikhard L (1992) Fourier transform ion cyclotron resonance mass spectrometry: technique developments. *International Journal of Mass Spectrometry and Ion Processes*, **118/119**, 37–70.

Marshall AG and Verdun FR (1990) *Fourier Transforms in NMR, Optical, and Mass Spectrometry: A User's Handbook*, Elsevier: Amsterdam; Oxford.

Marshall AG, Hendrickson CL and Jackson GS (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews*, **17**, 1–35.

Palmblad M, Hakansson K, Hakansson P, Feng X, Cooper HJ, Giannakopulos AE, Green PS and Derrick PJ (2000) A 9.4 T Fourier transform ion cyclotron resonance mass spectrometer: description and performance. *European Journal of Mass Spectrometry*, **6**, 267–275.

Sali A, Glaeser R, Earnest T and Baumeisters W (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.

Sheil MM, Guilhaus M and Derrick PJ (1990) Collision-activated decomposition of peptides by Fourier transform ion cyclotron resonance spectrometry. *Organic Mass Spectrometry*, **25**, 671–680.

Shi SDH, Drader JJ, Freitas MA, Hendrickson CL and Marshall AG (2000) Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry. *International Journal of Mass Spectrometry*, **196**, 591–598.

Stace AJ (1998) Infrared photophysics in an ion trap. *Journal of Chemical Physics*, **109**, 7214–7223.

Tsybin YO, Witt M, Baykut G, Kjeldsen F and Hakansson P (2003) Combined infrared multiphoton dissociation and electron capture dissociation with a hollow electron beam in

Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Communications in Mass Spectrometry*, **17**, 1759–1768.

Tyers M and Mann M (2003) From genomics to proteomics. *Nature*, **422**, 193–197.

Wu S-L, Choudhary G, Ramström M, Bergquist J and Hancock WS (2003a) Evaluation of shotgun sequencing for proteomic analysis of human plasma using HPLC coupled with either ion trap or Fourier transform mass spectrometry. *Journal of Proteome Research*, **2**, 383–393.

Wu ZG, Jernstrom S, Hughey CA, Rodgers RP and Marshall AG (2003b) Resolution of 10,000 compositionally distinct components in polar coal extracts by negative-ion electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Energy & Fuels*, **17**, 946–953.

Yamashita M and Fenn JB (1984a) Electrospray ion source. Another variation on the free-jet theme. *Journal of Physical Chemistry*, **88**, 4451–4459.

Yamashita M and Fenn JB (1984b) Negative ion production with the electrospray source. *Journal of Physical Chemistry*, **88**, 4671–4675.

Zubarev RA, Kelleher NL and McLafferty FW (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society*, **120**, 3265–3266.

Zubarev RA, Kruger NA, Fridriksson EK, Lewis MA, Horn DM, Carpenter BK and McLafferty FW (1999) Electron capture dissociation of gaseous multiply-charged proteins is favored at disulfide bonds and other sites of high hydrogen atom affinity. *Journal of the American Chemical Society*, **121**, 2857–2862.

Zubarev RA (2003) Reactions of polypeptide ions with electrons in the gas phase. *Mass Spectrometry Reviews*, **22**, 57–77.

Specialist Review

Laser-based microdissection approaches and applications

Rosamonde E. Banks
University of Leeds, Leeds, UK

1. Introduction

There is considerable heterogeneity in terms of both the numbers and types of cells present in different organs or tissues, reflecting their different biological functions. As an example, the cellular composition of the kidney comprises many diverse cell types including proximal tubule epithelial cells, distal tubule epithelial cells, collecting duct epithelial cells, endothelial cells, podocytes, mesangial cells, fibroblasts, and muscle cells. The proteome of any tissue, therefore, is actually a composite, reflecting the proteome of all its constituent cell types. The cellular makeup of tissues or organs can vary, however, not only during pathological states but also as a result of normal physiological changes. A prime example of this is the profound cellular changes that occur in the ovary and endometrium during the menstrual cycle. With diseases such as cancer, much of the normal tissue architecture may be lost with tumor areas consisting largely of malignant cells, but containing increased numbers of endothelial cells and infiltrating lymphocytes as part of the pathology. Additionally, within such a tumor and the surrounding tissue, cells that are at different stages of genetic evolution may be present, representing the transition from normal to frankly malignant and invasive pathology, and different microenvironments may be present, for example, normoxic and hypoxic areas.

Such tissue heterogeneity undoubtedly impacts on comparative studies, introducing more background "noise" to the analysis than would be present if the diseased cells alone were compared with their normal counterpart. The ability to isolate specific cell types or areas for experimental analysis would overcome this and generate more easily interpretable and readily applicable data. Conversely, however, while analysis of whole-tissue lysates may be more difficult to interpret, it does involve minimal manipulation of the in vivo state and the interplay between different elements of tissue, such as stromal–epithelial and endothelial–epithelial interactions, are undoubtedly important. The relative merits of either the whole-tissue approach or the selection of specific cells should be taken into account and the most appropriate method determined by the specific questions being addressed.

In terms of enriching for defined cell populations, various manual microdissection, immunoisolation, or primary cell culture generation strategies have been

explored. However, all these have inherent problems ranging from the laborious nature of the work to the possibility of introducing in vitro artifacts. One recent area of technological development that has facilitated the isolation of specific cell types or cellular areas without enzymatic digestion or cell culture is that of laser-based microdissection systems. Used extensively in experiments involving subsequent analysis of DNA or RNA from the isolated cells, the numbers of reports of successful use in proteomics-based studies are far fewer but nevertheless show the potential of the approach. The principles underlying the various laser-based microdissection instruments are outlined below, together with a consideration of the technical issues, and their use in various proteomics-based research illustrated.

2. Principles of laser-based microdissection systems

There are three main commercially available systems in current use. The first developed was the PixCell laser capture microdissection (LCM) system, which was developed by scientists at the National Institutes of Health in the United States (Emmert-Buck *et al*., 1996) and is marketed by the bioengineering company Arcturus. The system is based on a high-precision inverted microscope with a low energy, near-infrared laser mounted above it, a manipulator arm, and a computerized laser control and image capture platform. Using normal cleaned glass slides, the tissue section to be dissected is fixed and stained using appropriate protocols and the slide is placed on the microscope platform. A 6-mm-diameter Perspex cap, whose lower surface is covered with a transparent thermolabile ethylene vinyl acetate film, is then lowered into place onto the section with the film in direct contact with the tissue. By moving the stage so that the area of interest is in the field of view and by focusing the laser beam of adjustable diameter 7.5 to 30 μm through the cap onto an area containing cells of interest, the laser is then fired and the heat generated causes localized melting of the film and fusion with the selected cells directly below. The combination of relatively low energy and short-duration pulses (up to 5 ms) minimizes the heat damage to the tissue. Using the micromanipulator arm, the cap is then subsequently lifted from the section, complete with the fused dissected cells (Figure 1). Captured cells can then be solubilized using an appropriate extraction buffer either immediately or following further dissection and collection of cells. The most recently developed version of this system allows much of the process to be automated once appropriate areas of tissue are selected on the section image.

The PALM MicroBeam system couples “laser microbeam microdissection” (LMM) to ablate areas of tissue surrounding the cells of interest with “laser pressure catapulting” (LPC) to subsequently retrieve the selected area of cells (Schutze and Lahr, 1998). Again, the system is based around an inverted microscope to visualize the tissue section with the microdissection, that is, the movement of the stage and laser firing being automated. The operator delineates essentially the area containing the cells of interest on a computer screen and the process of ablation of the surrounding area is then computer-controlled. This strategy is therefore noncontact and nonheating, with cells of interest being cut round using cold photolysis generated by an ultraviolet N_2 laser microbeam rather than positively

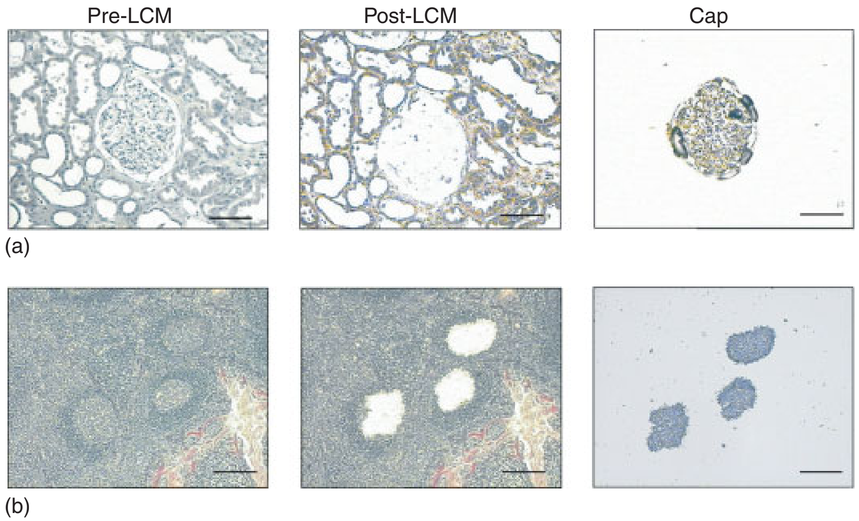


Figure 1 Examples of laser capture microdissection of (a) a renal glomerulus from kidney and (b) germinal centers from lymphoid tissue. Shown are the appearance of the tissue prior to and following capture and the captured cells present on the cap. The bar represents 100 µm (Reproduced from Banks, Dunn, Forbes, Stanley, Pappin, Navan, Gough, Harnden, and Selby, 1999, by permission of Wiley VCH)

targeted. The minimum width of the laser track separating adjacent areas is quoted as $<1\,\mu m$, thus allowing quite selective ablation and delineation of infiltrating cells within areas of interest. To support the section and to allow subsequent visual examination of the retrieved material, a polyethylene naphthalate membrane backing is often used between the section and the glass slide. After microdissection, the laser is used to catapult the dissected tissue, complete with support membrane, into a collecting tube containing appropriate extraction buffer, a process termed *laser pressure catapulting*.

The most recently developed commercially available system is the Leica AS LMD. Similar in principle to the PALM system in that it employs a similar UV laser ablative microdissection strategy to remove surrounding tissue, it has a stated minimum diameter cut of $<2.5\,\mu m$. However, the microscope is a conventional upright version and the slide is inverted such that the dissected material, once freed from surrounding tissue, is collected in a tube below the section, largely under the force of gravity (Kolble, 2000). Again the system is automated in that the operator delineates the dissection area, and in this case the stage remains still with the laser being moved and fired, a process controlled by the computer, to achieve the required ablation of surrounding areas.

3. Technical issues

The relative technical merits of the three systems with particular regard to resolution, ease of use and environment requirements, and sample visualization

have been recently reviewed (Cornea and Mungenast, 2002). Particular issues pertaining to sample processing are briefly discussed below.

The predominant use of laser-based microdissection has been for the study of nucleic acids in selected cell populations, combined with a variety of downstream analytical techniques. The analysis of nucleic acids has the major advantage of requiring very little starting material due to the use of PCR amplification so that single-cell analysis is achievable. Additionally, paraffin-embedded, formalin-fixed material can often be used, allowing better visualization of the microdissection and providing access to a wealth of archival material. The uses of laser-based microdissection with nucleic acids have been reviewed and will not be covered further here (see Further reading).

As there is no amplification equivalent to PCR for proteins, proteomic analysis techniques that can be employed with microdissected material are dependent on a balance between their sensitivity and the time taken to accumulate sufficient material. For example, the dissection required to procure sufficient cells for large format 2D-PAGE requires tens of thousands of shots, which may represent several days of dissection, and there is a question, even with tissues that are relatively easy to dissect, as to whether, generally, only the most abundant protein species will therefore be examined. Those tissues requiring extensive microdissection will also inevitably result in a degree of compromise in terms of the purity achieved, and the degree of enrichment afforded in the protein profile relative to the starting material will also vary between tissue types (Craven *et al*., 2002).

Material fixed in cross-linking agents such as formalin is not suitable for proteomic investigation and hence frozen sections are used, although unfortunately of poorer morphology, particularly when viewed with inverted optics. Tissue sections need to be stained using protocols designed to allow good visualization but minimizing artifacts such as protein degradation and allowing subsequent analysis of extracted proteins. The initial "proof of principle" demonstration used a modified rapid hematoxylin and eosin staining method with proteinase inhibitors on frozen kidney and cervical tissue and showed that material collected by LCM could generate 2D-PAGE (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5) protein profiles, with minimal effects of processing on either the profile or the subsequent mass spectrometric (MS; *see* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5) sequencing results (Banks *et al*., 1999). Since this initial report, the examination of different tissue-processing regimens indicate that the type of fixative for the frozen sections is critical, with ethanol being the preferred option; various stains including hematoxylin and eosin, and methyl green and toluidine blue can be adapted and used successfully, as can the modified immunolabeling protocols using silver-enhanced gold labeling or fluorescent labeling (Craven *et al*., 2002; Mouledous *et al*., 2003a; Ahram *et al*., 2003). More recently, the process of "navigated LCM" has been used, which incorporates the use of fixed unstained sections for the microdissection, thus avoiding staining artifacts, but relies on images generated from adjacent stained coverslipped sections with consequent improved visualization of tissue architecture to delineate and guide the dissection process (Mouledous *et al*., 2003b).

4. Selected examples of applications in proteomics

The majority of published proteomics-based studies using laser-based microdissection have utilized the PixCell LCM system unless indicated otherwise. The different studies are described below, subdivided on the basis of the subsequent proteomic techniques used to analyze the dissected material.

4.1. *Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)*

One of the main separation techniques is 2D-PAGE. The use of 2D-PAGE (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5) with protein lysates generated from the use of LCM provides a very encouraging endorsement of the approach, with the caveat that there are limitations with respect to sample and tissue type in terms of the degree of enrichment achieved (Craven *et al*., 2002). The initial demonstration of this combination of approaches showed the feasibility of using cells dissected from frozen tissue sections subjected to brief fixation and hematoxylin and eosin staining, with only modest effects of processing on the protein profile and compatibility with subsequent MS sequencing (*see* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5) (Banks *et al*., 1999). As described previously, several other studies have also since explored other technical aspects of tissue processing. The main limiting factor with 2D-PAGE is the amount of material required, which may necessitate tens of thousands of laser shots and microdissection times of several hours or even days depending on the tissue type and still result in only a few hundred spots, that is, far less than normally visualized on large format gels. Inevitably, therefore, even with tissues that are relatively easy to dissect, only the most abundant protein species are likely to be examined.

In the first comparative analysis of microdissected normal versus malignant cells, analysis of lysates generated from approximately 50 000 cells from each matched sample from two patients with esophageal cancer allowed visualization of almost 700 proteins, with 98% similarity between normal and malignant protein profiles (Emmert-Buck *et al*., 2000). Ten proteins were tumor-specific with seven other proteins only found in normal epithelium, one of which was identified as annexin I. Using a combination of Western blotting of microdissected cell lysates and immunohistochemistry, an extension of this study confirmed complete loss or dramatic reduction of annexin I expression in both esophageal and prostate cancers, occurring even at the premalignant stage (Paweletz *et al*., 2000b). This study illustrates the use of a small number of gels requiring large amounts of microdissected protein as the initial step to determine differentially expressed proteins and then validating on a larger sample set using Western blotting with the necessity for much less material. A similar later study on prostate cancer found six tumor-specific changes in two cases examined, one of which was the tumor marker PSA, and which were only present in the microdissected epithelial

compartments but not in stroma (Ornstein *et al*., 2000b). An extension of this study to an examination of 12 matched normal and high-grade tumor samples from 8 patients with cells obtained by LCM or manual microdissection and loading 100 000 to 140 000 cells per gel revealed 40 tumor-specific differences. However, disappointingly, only six of these were present in more than one case. Several of the proteins had been implicated before in tumor biology, for example, the 67-kDa laminin receptor and lactate dehydrogenase. (Ahram *et al*., 2002).

Similar success has also been seen with other cancers such as breast, ovarian, and pancreatic cancers. Examination of six matched pairs of normal breast ductal cells and ductal carcinoma in situ at either the whole tissue level or following LCM with up to 100 000 cells per gel identified 57 proteins to be differentially expressed and which met various criteria including being consistent between samples (Wulfkuhle *et al*., 2002). Only four proteins were identified by both whole tissue – and LCM-based approaches. Using a similar strategy, the 52-kDa FK506 binding protein, RhoGDI, and glyoxalase I were demonstrated to be uniquely overexpressed in invasive ovarian cancer compared with the low malignant potential ovarian tumors (Brown Jones *et al*., 2002), with downstream validation by a combination of Western blotting and reverse-phase protein arrays. In pancreatic cancer, microdissection and the subsequent 2D-PAGE analysis revealed about 800 spots per gel, with nine spots consistently differentially expressed including S100A6, trypsin, and annexin III (Shekouh *et al*., 2003).

The possibility of using fluorescence 2D-difference gel electrophoresis (DIGE; *see* Article 25, **2D DIGE**, Volume 5) has been explored, which avoids the difficulties of gel–gel comparisons and allows more ready identification of spots from parallel gels for sequencing. Using limiting labeling, in a comparison of 250 000 cells from a matched tumour-normal esophageal sample, 1000 to 1100 protein spots were visualized (Zhou *et al*., 2002), with 58 proteins upregulated by >3-fold in tumors and 107 proteins downregulated by >3-fold. Using a matched SYPRO Ruby stained gel (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5), annexin I was identified again as being downregulated and the tumor rejection antigen gp96 was upregulated, both of which were confirmed by Western blotting. Using a similar approach but with saturation labeling, normal murine intestinal epithelium and adenoma tissues from Min mice were compared (Kondo *et al*., 2003). Using only 2.7 to 6.6 µg of protein, which represented a 1-mm^2 area of tissue dissected using the Leica LMD system from a 10-µm-thick frozen section with a collection time of 1 min, a 24-cm 2D-PAGE profile with 1500 spots was generated, 37 of which differed. Using a murine cell line run in parallel to generate sufficient material for MS analysis and subsequent Western blotting for validation, a number of proteins were identified including prohibitin, which is a putative tumor suppressor, HSP84, and a number of 14-3-3 zeta forms.

4.2. *Western blotting*

Western blotting of microdissected material provides not only information about the cellular localization of a protein but also confirmation of its molecular size,

and, depending on antigen and antibody, requires relatively small amounts of material. Examples include the detection of actin, major vault protein, GAPDH, aldolase isoforms, pyruvate kinase, RhoGDI, annexin I, and gp96, with numbers of microdissected cells varying from approximately 250 to 40 000 (Craven and Banks, 2002; Unwin *et al*., 2003; Zhou *et al*., 2002; Paweletz *et al*., 2000b; Brown Jones *et al*., 2002; Wulfkuhle *et al*., 2002). Illustrating the potential of the approach not only in validating identified potential new markers but also in addressing questions regarding known proteins, microdissected material from benign and malignant prostatic epithelium was used to demonstrate that the vast majority of intracellular PSA exists in the free unbound form, but has the capability of binding to α_1-antichymotrypsin form once outside the cell (Ornstein *et al*., 2000a).

4.3. *Immunoassay*

With the increasingly sensitive immunoassays employing chemiluminescent detection, laser-based microdissection represents a source of material for assay. Immunoassay of the amounts of the tumor marker prostate-specific antigen (PSA) in normal and malignant prostate epithelium using only several tens to hundreds of microdissected cells acquired in <15 min, showed that malignant cells were found to contain up to sevenfold more PSA than normal epithelial cells, although there was overlap between the sets (Simone *et al*., 2000). Calculation of the actual number of PSA molecules per cell resulted in estimates of average values ranging from 10 000 to 1 000 000. More recently, hepatitis C virus core protein has been quantified by immunoassay in lysates prepared from microdissected liver cells prepared using the Leica LMD system and found to vary from 7 to 56 pg/cell in infected samples, being more sensitive than immunohistochemistry (Sansonno *et al*., 2004).

4.4. *Multiplex protein and antibody microarrays*

The development of protein microarrays (*see* Article 24, **Protein arrays**, Volume 5, Article 32, **Arraying proteins and antibodies**, Volume 5, and Article 33, **Basic techniques for the use of reverse phase protein microarrays for signal pathway profiling**, Volume 5) enabling the simultaneous high-throughput profiling of many proteins in large numbers of samples, with only relatively small amounts of sample being required, is an area of considerable activity. Various formats exist including ones essentially analogous to cDNA arrays but with either antibodies or ligands arrayed onto which labeled tissue lysates can be applied. Alternatively, tissue lysates are arrayed, which are then probed with antibodies. It is increasingly evident that microdissected cell lysates are suitable for such applications.

Using a reverse-phase array approach, solubilized microdissected prostate tissue has been spotted onto nitrocellulose slides to generate microarrays containing up to 1000 tissue lysate spots/slide. Typically, these were produced using lysates resulting from only 500 to 3000 laser shots (equivalent to 2500–15000 cells). By using lysates from prostate cells covering the transition from normal epithelium

through prostate intraepithelial neoplasia (PIN) to invasive cancer and by probing with a range of antibodies, changes during tumor progression were shown. These included decreased levels of cleaved poly(ADP-ribose) polymerase (PARP) and cleaved caspase-7, increased phosphorylation of Akt, but decreased phosphorylation of Erk and were interpreted as representing activation of prosurvival events (Paweletz *et al*., 2001). Results using antibodies against additional phosphorylated cell signaling molecules provided further evidence of activation of the Akt pathway (Grubb *et al*., 2003). A similarly designed study examining the activation status of several key signaling molecules involved in cell survival and proliferation in microdissected ovarian cancer also found changes in many of the variables examined but with generally no type or stage-specific patterns (Wulfkuhle *et al*., 2003). This may support the need to individualize patient samples in terms of the biological changes and the structuring of therapeutic strategies accordingly.

Antibody arrays have also been used to analyze microdissected samples. Using biotinylated extracts from 2500 to 3500 cells (approximately 0.5 µg) of both stromal and epithelial elements of squamous cell carcinoma of the oral cavity, together with an array containing 368 different antibodies against a variety of proteins, multiple protein differences were found with 11 proteins consistently changing in either relative levels or state of phosphorylation. Several of these were proteins involved in signal transduction pathways, with many exhibiting selective changes in either the epithelial cells or the stroma (Knezevic *et al*., 2001). This illustrates the importance of stromal–epithelial cell interactions in cancer.

4.5. Mass spectrometry

Although more conventionally used for protein identification on the basis of peptide mass fingerprints (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5), matrix-assisted laser desorption ionization mass spectrometers equipped with time-of-flight detectors (MALDI-TOF; *see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) are increasingly being used to generate profiles of intact proteins or fragments in complex biological samples in the context of disease-specific patterns, requiring only small amounts of protein. Using MALDI-TOF, direct analysis of 1250 cells from normal stroma, normal epithelium, carcinoma in situ, and invasive carcinoma prepared from a mastectomy sample and following addition of sinapinic acid to the LCM film immobilized on a standard stainless steel MALDI target generated distinct reproducible spectra containing 20–50 peaks from four cell types (Palmer-Toy *et al*., 2000). A subsequent investigation examined a range of factors and their effects on the quality of the profile generated, including section staining, ethanol treatment, and matrix application and showed that under optimal conditions, good quality profiles of at least the more abundant peaks could be generated from as few as 10 cells and again demonstrated that this approach could define distinct differences in profiles between normal and invasive breast epithelium (Xu *et al*., 2002).

Several studies have used the SELDI (surface-enhanced laser desorption and ionization) system developed by Ciphergen Biosystems (Fremont, CA, USA) to profile laser microdissected material. Based around a MALDI-TOF MS,

proteins from complex mixtures are immobilized onto ProteinChip arrays containing 8–24 sample spots, by selective capture using the conventional chromatographic chemistries of the chip surfaces. The first study to use SELDI-MS (surface-enhanced laser desorption and ionization-mass spectrometry) to analyze microdissected material was part of a larger study examining the specific forms of several prostate cancer markers in tissue samples and biological fluids. Using only 2000–3000 cells, several proteins were found to be differentially expressed between normal and malignant samples with a peak detected only in the prostate carcinoma samples being found to have the same molecular weight as purified prostate-specific membrane antigen (PSMA) (Wright *et al*., 1999). Several studies have now shown that microdissected material can be examined in this way to discriminate between normal and diseased cells in a variety of cell types and diseases, although mainly cancers (Paweletz *et al*., 2000a; von Eggeling *et al*., 2000; Batorfi *et al*., 2003; Zheng *et al*., 2003; Melle *et al*., 2003). The numbers of peaks generated in a sample vary with ProteinChip type, extraction buffer, and tissue and although reproducible protein "fingerprints" can be generated using as few as 25–100 cells, far more proteins can be seen with increasing loads and in most cases between 500 and 10 000 cells have been used, the amount of protein present in these dissections being very dependent on factors such as cell type and section thickness however. Although in a number of cases, particular discriminatory peaks have been found, the challenge now lies in being able to identify the particular proteins or protein fragments that will provide further information. One approach that was used successfully to identify a peak of approximately 36 kDa as annexin V, which discriminated between normal pharyngeal epithelium and head and neck squamous cell carcinoma was that of subsequent 2D-PAGE and MS sequencing (Melle *et al*., 2003). This was possible due to the size of this particular peak, making it amenable to such separation. However, for smaller proteins or fragments, this is still a challenge.

5. Summary

The number of studies using laser-based microdissection methods as a prelude to subsequent proteomic analyses is still small, but sufficient to illustrate the promise of such technology in overcoming the problems of tissue heterogeneity. The use of highly sensitive downstream analysis techniques that require relatively small amounts of sample, such as protein and antibody microarrays and MS profiling are likely to be the main areas in which such microdissection techniques will contribute more extensively in the future. The combination with multidimensional liquid chromatography tandem MS (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5) is also a possibility, although not yet explored. The analysis of different cell populations in abnormal tissues will clarify the molecular basis of changes occurring in disease and during progression, and in defining the important role of interactions between the different cellular compartments. Ultimately, in addition to increasing our understanding of the underlying disease pathogenesis, this has the potential to contribute in various areas of clinical utility ranging from marker

discovery for diagnosis and prognosis, to identification of specific disease-related alterations in individuals and appropriate tailoring of targeted therapy.

Related articles

Article 7, **Time-of-flight mass spectrometry**, Volume 5; Article 22, **Two-dimensional gel electrophoresis**, Volume 5; Article 24, **Protein arrays**, Volume 5; Article 25, **2D DIGE**, Volume 5; Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5; Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5; Article 32, **Arraying proteins and antibodies**, Volume 5; Article 33, **Basic techniques for the use of reverse phase protein microarrays for signal pathway profiling**, Volume 5

Further reading

Conn PM (Ed) (2002) *Methods in Enzymology*, "Laser Capture Microscopy" – Volume 356, Academic Press: San Diego, CA.

References

Ahram M, Best CJM, Flaig MJ, Gillespie JW, Leiva IM, Chuaqui RF, Zhou G, Shu H, Duray PH, Linehan WM, *et al*. (2002) Proteomic analysis of human prostate cancer. *Molecular Carcinogenesis*, **33**, 9–15.

Ahram M, Flaig MJ, Gillespie JW, Duray PH, Linehan WM, Ornstein DK, Niu S, Zhao Y, Petricoin EF III and Emmert-Buck MR (2003) Evaluation of ethanol-fixed, paraffin-embedded tissues for proteomic applications. *Proteomics*, **3**, 413–421.

Banks RE, Dunn MJ, Forbes MA, Stanley A, Pappin D, Naven T, Gough M, Harnden P and Selby PJ (1999) The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis – preliminary findings. *Electrophoresis*, **3**, 689–700.

Batorfi J, Ye B, Mok SC, Cseh I, Berkowitz RS and Fulop V (2003) Protein profiling of complete mole and normal placenta using ProteinChip analysis on laser capture microdissected cells. *Gynecologic Oncology*, **88**, 424–428.

Brown Jones M, Krutzsch H, Shu H, Zhao Y, Liotta LA, Kohn EC and Petricoin EF (2002) Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics*, **2**, 76–84.

Cornea A and Mungenast A (2002) Comparison of current equipment. *Methods in Enzymology*, **356**, 3–12.

Craven RA and Banks RE (2002) Use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis. *Methods in Enzymology*, **356**, 33–49.

Craven RA, Totty N, Harnden P, Selby PJ and Banks RE (2002) Laser capture microdissection and two-dimensional polyacrylamide gel electrophoresis: evaluation of tissue preparation and sample limitations. *American Journal of Pathology*, **160**, 815–822.

Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA and Liotta LA (1996) Laser capture microdissection. *Science*, **274**, 998–1001.

Emmert-Buck MR, Gillespie JW, Paweletz CP, Ornstein DK, Basrur V, Appella E, Wang QH, Huang J, Hu N, Taylor P, *et al*. (2000) An approach to proteomic analysis of human tumors. *Molecular Carcinogenesis*, **27**, 158–165.

Grubb RL, Calvert VS, Wulfkuhle JD, Paweletz CP, Linehan WM, Phillips JL, Chuaqui R, Valasco A, Gillespie J, Emmert-Buck M, *et al*. (2003) Signal pathway profiling of prostate cancer using reverse phase protein arrays. *Proteomics*, **3**, 2142–2146.

Knezevic V, Leethanakul C, Bichsel VE, Worth JM, Prabhu VV, Gutkind JS, Liotta LA, Munson PJ, Petricoin EF and Krizman DB (2001) Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics*, **1**, 1271–1278.

Kolble K (2000) The LEICA microdissection system: design and applications. *Journal of Molecular Medicine*, **78**, B24–B25.

Kondo T, Seike M, Mori Y, Fujii K, Yamada T and Hirohashi S (2003) Application of sensitive fluorescent dyes in linkage of laser microdissection and two-dimensional electrophoresis as a cancer proteomic study tool. *Proteomics*, **3**, 1758–1766.

Melle C, Ernst G, Schimmel B, Bleul A, Koscielny S, Wiesner A, Bogumil R, Moller U, Osterloh D, Halbhuber K-J, *et al*. (2003) Biomarker discovery and identification in laser microdissected head and neck squamous cell carcinoma with ProteinChip technology, two-dimensional electrophoresis, tandem mass spectrometry, and immunohistochemistry. *Molecular and Cellular Proteomics*, **2**, 443–452.

Mouledous L, Hunt S, Harcourt R, Harry J, Williams KL and Gutstein HB (2003a) Proteomic analysis of immunostained, laser-capture microdissected brain samples. *Electrophoresis*, **24**, 296–302.

Mouledous L, Hunt S, Harcourt R, Harry J, Williams KL and Gutstein HB (2003b) Navigated laser capture microdissection as an alternative to direct histological staining for proteomic analysis of brain samples. *Proteomics*, **3**, 610–615.

Ornstein DK, Englert C, Gillespie JW, Paweletz CP, Linehan WM, Emmert-Buck MR and Petricoin EF (2000a) Characterization of intracellular prostate-specific antigen from laser capture microdissected benign and malignant prostatic epithelium. *Clinical Cancer Research*, **6**, 353–356.

Ornstein DK, Gillespie JW, Paweletz CP, Duray PH, Herring J, Vocke CD, Topalian SL, Bostwick DG, Linehan WM, Petricoin EF, *et al*. (2000b) Proteomic analysis of laser capture microdissected human prostate cancer and *in vitro* prostate cell lines. *Electrophoresis*, **21**, 2235–2242.

Palmer-Toy DE, Sarracino DA, Sgroi D, LeVangie R and Leopold PE (2000) Direct acquisition of matrix-assisted laser desorption/ionization time-of-flight mass spectra from laser capture microdissected tissues. *Clinical Chemistry*, **46**, 1513–1516.

Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, Petricoin III EF and Liotta LA (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Paweletz CP, Gillespie JW, Ornstein DK, Simone NL, Brown MR, Cole KA, Wang Q-H, Huang J, Hu N, Yip T-T, *et al*. (2000a) Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Research*, **49**, 34–42.

Paweletz CP, Ornstein DK, Roth MJ, Bichsel VE, Gillespie JW, Calvert VS, Vocke CD, Hewitt SM, Duray PH, Herring J, *et al*. (2000b) Loss of annexin 1 correlates with early onset of tumorigenesis in esophageal and prostate carcinoma. *Cancer Research*, **60**, 6293–6297.

Sansonno D, Lauletta G and Dammacco F (2004) Detection and quantitation of HCV core protein in single hepatocytes by means of laser capture microdissection and enzyme-linked immunosorbent assay. *Journal of Viral Hepatitis*, **11**, 27–32.

Schutze K and Lahr G (1998) Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnology*, **16**, 737–742.

Shekouh AR, Thompson CC, Prime W, Campbell F, Hamlett J, Herrington CS, Lemoine NR, Crnogorac-Jurcevic T, Buechler MW, Friess H, *et al*. (2003) Application of laser capture microdissection combined with two-dimensional electrophoresis for the discovery of differentially regulated proteins in pancreatic ductal adenocarcinoma. *Proteomics*, **3**, 1988–2001.

Simone NL, Remaley AT, Charboneau L, Petricoin EF, Glickman JW, Emmert-Buck MR, Fleisher TA and Liotta LA (2000) Sensitive immunoassay of tissue cell proteins procured by laser capture microdissection. *American Journal of Pathology*, **156**, 445–452.

Unwin RD, Craven RA, Harnden P, Hanrahan S, Totty N, Knowles M, Eardley I, Selby PJ and Banks RE (2003) Proteomic changes in renal cancer and co-ordinate demonstration of both the glycolytic and mitochondrial aspects of the Warburg effect. *Proteomics*, **3**, 1620–1632.

von Eggeling F, Davies H, Lomas L, Fiedler W, Junker K, Claussen U and Ernst G (2000) Tissue-specific microdissection coupled with ProteinChip array technologies: applications in cancer research. *Biotechniques*, **29**, 1066–1070.

Wright Jr GL, Cazares LH, Leung S-M, Nasim S, Adam BL, Yip T-T, Schellhammer PF, Gong L and Vlahou A (1999) Proteinchip surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer and Prostatic Diseases*, **2**, 264–276.

Wulfkuhle JD, Aquin JA, Calvert VS, Fishman DA, Coukos G, Liotta LA and Petricoin EF (2003) Signal pathway profiling of human tissue specimens using reverse-phase protein microarrays. *Proteomics*, **3**, 2085–2090.

Wulfkuhle JD, Sgroi DC, Krutzsch H, McLean K, McGarvey K, Knowlton M, Chen S, Shu H, Sahin A, Kurek R, *et al*. (2002) Proteomics of human breast ductal carcinoma *in situ*. *Cancer Research*, **62**, 6740–6749.

Xu BJ, Caprioli RM, Sanders ME and Jensen RA (2002) Direct analysis of laser capture microdissected cells by MALDI mass spectrometry. *Journal of the American Society for Mass Spectrometry*, **13**, 1292–1297.

Zheng Y, Xu Y, Ye B, Lei J, Weinstein MH, O'Leary MP, Richie JP, Mok SC and Liu BC-S (2003) Prostate carcinoma tissue proteomics for biomarker discovery. *Cancer*, **98**, 2576–2582.

Zhou G, Li H, DeCamp D, Chen S, Shu H, Gong Y, Flaig M, Gillespie MW, Hu N, Taylor PR, *et al*. (2002) 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Molecular and Cellular Proteomics*, **1**, 117–124.

Specialist Review

Orbitrap mass analyzer

Alexander Makarov
Thermo Electron, Bremen, Germany

Michaela Scigelova
Thermo Electron, Hemel Hempstead, UK

1. Introduction

The orbitrap mass analyzer is rightfully considered to be one of the newest analyzers. Its roots, however, can be traced back to 1923, when the principle of orbital trapping was proposed by Kingdon (1923). Experiments over the next half a century have shown that charged particles could be indeed trapped in electrostatic fields but they offered no hint of how this could be used for mass analysis. Simultaneously, advances in charged particle optics led to the development of a multitude of electrostatic fields (Korsunskii and Basakutsa, 1958). It was a quadro-logarithmic field used for orbital trapping of laser-produced ions that enabled Knight to perform crude mass analysis by applying axial resonant excitation to trapped ions (Knight, 1981). At the same time, this attempt demonstrated that the quest for a new mass analyzer would require a great deal of improvement in all key areas, most notably a more accurate definition of the quadro-logarithmic field, an ability to inject ions from an external ion source, and an improvement in ion detection.

These issues have been successfully addressed in the seminal work of Makarov (1999, 2000). A number of very significant technological advances were implemented, thus making the orbitrap analyzer usable in practical settings, namely, the development of pulsed injection from an external ion storage device (Hardman and Makarov, 2003; Hu *et al*., 2005; Makarov *et al*., 2006). Unique to the history of mass spectrometry, both proof of principle and product development have been carried out entirely within industry: from the first public announcement in 1999 till the entry into mainstream mass spectrometry in 2005.

The term *orbitrap* was first coined to define a mass analyzer wherein ions combine rotation around an electrode system with harmonic oscillations along the axis of rotation at a frequency characteristic of their *m/q* value (Makarov, 2000). The image current from coherently oscillating ions is detected on receiver plates as a time-domain signal. Fourier transform of this signal then gives us a mass spectrum. Therefore, the orbitrap analyzer extends the family of Fourier transform mass analyzers, which, until recently, contained only one widely used analyzer: Fourier transform ion cyclotron resonance (FTICR).

From the point of view of an ion optics design, the orbitrap analyzer belongs in the class of closed electrostatic traps. In theory, the ions could remain confined within such traps for an indefinite period of time. Other examples of closed electrostatic traps include ion storage rings, linear electrostatic rings, multiple-reflection ion mirrors, and so on. The harmonic nature of oscillations enables the orbitrap mass analyzer to provide the highest frequency and the highest quality of focusing in this group. This, in turn, gives the orbitrap analyzer its superior performance with respect to the dynamic range, mass accuracy, and resolving power that it is capable of achieving.

2. Fundamental principles of operation

2.1. Geometry of the analyzer

The geometry of the orbitrap mass analyzer is shown in Figure 1. The trap consists of an outer barrel-like electrode and a central spindlelike electrode along the axis. These electrodes are shaped in such a manner so as to produce the quadro-logarithmic potential distribution:

$$U(r, z) = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2} \cdot (R_m)^2 \cdot \ln\left[\frac{r}{R_m}\right] + C \quad (1)$$

where r, z are cylindrical coordinates ($z = 0$ being the plane of the symmetry of the field), C is a constant, k is field curvature, R_m is the characteristic radius. In this trap, stable trajectories combine rotation around the central electrode with oscillations along the axis, resulting in an intricate spiral (Makarov, 2000).

2.2. Motion of trapped ions

There are three characteristic frequencies of ion motion:

- frequency of rotation around the central electrode
- frequency of radial oscillations (between maximum and minimum radii)
- frequency of axial oscillations (along z-axis).

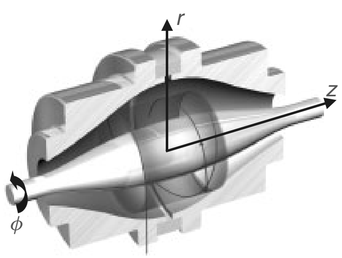


Figure 1 Diagram of the orbitrap mass analyzer showing a spiral trajectory of an ion

Only the last one is completely independent of initial velocities and coordinates of the ions. Therefore, only this frequency could be used for determinations of mass-to-charge ratios *m/q*:

$$\omega = \sqrt{\frac{q}{m} \cdot k} \tag{2}$$

wherein constant k comes from equation (1) and changes in proportion to the voltage between the central and the outer electrodes.

2.3. Ion detection

Axial oscillation frequencies can be directly detected by measuring the image current on the outer orbitrap electrodes as shown in Figure 1. A broadband detection is followed by a fast Fourier transformation (FFT) to convert the recorded time-domain signal into a mass-to-charge spectrum (Marshall and Verdun, 1990). The image current is amplified and processed exactly in the same way as for FTICR, resulting in a similar sensitivity and signal-to-noise ratios. However, there is a minor but important distinction: the square-root dependence originating from the electrostatic nature of the field causes a much slower drop in resolving power observed for ions of increased *m/q*. As a result, the orbitrap analyzer may outperform FTICR in this respect for masses above a particular *m/q* (typically, above 1–2000 *m/z*).

There could be an alternative way of detecting ions that follows the original proposal of Knight (1981): to excite ions axially using a voltage at a resonant frequency and to scan the mass range by sweeping this frequency. However, this approach would offer no advantage over conventional traps while at the same time lacking the ability to carry out MS^n experiments. Therefore, detection of the image current remains the major mode of operation for the orbitrap mass spectrometers.

2.4. Formation of coherent ion packets

The most important prerequisite for detection of the image current is the ability to concentrate all ions of the same *m/q*. The relevant dimensions of the ion cloud must be smaller than the amplitude of oscillations we wish to detect. This could be achieved in one of two ways:

- Broadband excitation of ions from the equatorial plane. Traditionally, this approach is used in FTICR and is compatible with well-known types of external ratio frequency (RF) storage devices. But such an approach demands substantial complexity of an ion introduction apparatus (Makarov, 1999).
- Excitation by off-axis injection of pulsed ion packets ("excitation by injection"). This approach minimizes perturbations of the quadro-logarithmic field but requires a very fast ejection of large ion population from an ion source or an external RF storage device.

Ultimately, the second approach has proved to be more practical and robust. It is performed according to the following sequence:

1. Ions are trapped in a gas-filled RF-only set of rods, preferably a linear ion trap. In principle, this also allows various manipulations of the ions, including isolation, fragmentation, MS^n, and so on.
2. Pulsed voltages are applied to the end electrodes (Hardman and Makarov, 2003; Hu *et al*., 2005) or across the RF electrodes (Makarov *et al*., 2006) so that the ions find themselves in a strong extraction field. The probability of collisions and collision-induced dissociation during the ion extraction is minimized by storing ions near the exit orifice. Additional lenses are used for the final spatial focusing of the ion beam into the entrance of the orbitrap analyzer, as well as to facilitate differential pumping to achieve the very high vacuum necessary for effective mass measurement.
3. Ions of individual mass-to-charge ratios arrive at the entrance of the orbitrap analyzer as a tight packet with dimensions considerably smaller than the amplitude of their axial oscillations. When ion packets are injected into the orbitrap analyzer off-axis (Figure 1), they start coherent axial oscillations without the need for any additional excitation.
4. Upon entering the orbitrap analyzer, the ion packets are "squeezed" by increasing the electric field to move the ions toward the equator and the central electrode. This increase in electric field is created by ramping up the voltage on the central electrode.
5. Because of the strong dependence of rotational frequencies on ion energies, angles, and initial positions, each ion packet soon spreads over the angular coordinate, forming a thin rotating ring. This has important ramifications: more ions can be present in the orbitrap mass analyzer before the space charge effects start impacting the mass resolution and accuracy of the measurement.

Following the injection, the voltages on both the central electrode and deflector are stabilized so that no mass drift can take place during the next stage (ion detection).

2.5. *Decay of coherent ion packets*

As mentioned earlier, under ideal conditions, the ions could remain in an electrostatic analyzer indefinitely. Unfortunately, the collisions with residual gas in the orbitrap cause ions to scatter and limit the time a transient can be detected for a period of a few seconds. A decay of the transient in the orbitrap mass analyzer and consequent limitation of resolving power is further caused by a loss of coherence due to miniscule imperfections of the orbitrap electrode manufacturing, and due to space charge repulsion. The transient could be further extended by improving the accuracy of electrodes and the vacuum.

The space charge repulsion is greatly reduced because of the shielding action of the central electrode that screens ions on one side of the ion ring from influencing ions on the other side. However, a due diligence in the design of the orbitrap

electrodes is needed to avoid more complex nonlinear effects caused by the interaction between nonlinear field perturbations and space charge.

2.6. Main analytical parameters of the orbitrap mass analyzer

Similar to other mass analyzers (e.g., a quadrupole), analytical parameters are determined to a large extent by the present status of manufacturing technology and electronics. The current level of machining precision enables the resolving power of the orbitrap mass analyzer to reach a value of several hundred thousands. Internal and thermal noise of electronic components impacts the sensitivity of detection of the image current in the orbitrap, making a couple of dozen ions its limit of detection. The mass error using external mass calibration is a few parts per million and remains stable over a 24-hour period. The accuracy is limited principally by the noise of electronic components. The repetition rate of a few hertz reflects the present speed at which we are able to apply a high voltage to the central electrode. While this combination of analytical parameters remains inferior to that of FTICR mass analyzers, it appears to be attractive enough given the absence of a superconducting magnet and of lower m/q ion losses suffered during transfer from an external ion storage device.

In comparison with other accurate-mass analyzers (e.g., time-of-flight analyzers), the orbitrap system offers much higher dynamic range over which accurate masses can be determined (extent of mass accuracy).

2.7. Fragmentation in the orbitrap mass analyzer

The ions trapped in the orbitrap analyzer have energies in the range of kiloelectron volts. A high-energy fragmentation caused by collision with residual gas happens automatically, and its extent could be regulated by gas pressure. Pulsed lasers could be another way to induce fragmentation of ions in the orbitrap. Unfortunately, when an ion decays under the dynamic trapping, its fragments will have the same velocity. As their energy is proportional to their individual mass-to-charge ratios, the trajectories become highly elliptical. Therefore, low-mass fragments (with m/q typically below 30–50% of that of the precursor ion) will fall onto the central electrode, while lower-charge state fragments (with m/q typically above 50% of the precursor ion) will hit the outer electrodes. This property seriously limits the utility of the analyzer for MS^n, especially taking into consideration the absence of collisional cooling, increased cycle time, inferior resolving power and sensitivity, cost, and complexity of such an apparatus.

2.8. The orbitrap analyzer used as an accurate-mass detector in hybrid mass spectrometers

The challenging nature of the technical issues related to performing MS/MS in the orbitrap mass analyzer was the main reason behind the concept of using it as an

accurate-mass detector for another mass analyzer: linking two mass spectrometers into one hybrid instrument. If we also insert an ion storage device between the first mass analyzer and the orbitrap, the analyzers are effectively decoupled from each other and any mass analyzer could be used as the first stage.

In the first commercial orbitrap-based instrument, a linear ion trap with radial ejection (Schwartz 2005) was chosen as a "partner" for the orbitrap mass analyzer because of its very high sensitivity, superb control of the ion population, short cycle time, and MS^n capability. Depending on the requirements for the analysis, the two analyzers can be used independently or in concert. It is worth pointing out that the MS/MS spectra generated in the linear ion trap and the orbitrap mass analyzer are very similar, with the only major difference being the resolution and mass accuracy of the observed peaks.

The ion storage device linking the linear ion trap to the orbitrap analyzer is called the *C-trap*. The C-trap has a high space charge capacity (Makarov *et al*., 2006), and enables several intriguing modes of operation:

- Ions can be fragmented by injecting them into the C-trap at higher energies to yield fragmentation patterns similar to those in triple-quadrupole mass spectrometers.
- The C-trap supports multiple fills. An injection of a fixed number of ions of a known compound can be followed by injection of analyte ions. Both sets of ions are then injected simultaneously into the orbitrap. This allows for a robust internal calibration of each spectrum.
- Multiple injections of ions fragmented or selected at different conditions could be stored together and acquired in a single orbitrap spectrum.
- Additional fragmentation methods could be used within the ion storage device, for example, IRMPD (infrared multiphoton dissociation), ion-molecule reactions, and ETD (electron transfer dissociation).

The next section describes selected analytical applications of the hybrid LTQ orbitrap mass spectrometer.

3. Analytical applications of the orbitrap mass analyzer

The key attributes of the orbitrap analyzer, namely, ruggedness, high mass accuracy, and excellent resolving power, make it suitable for both proteomics and small-molecule analyses. Details regarding recently published proteomics applications can be found in Scigelova and Makarov (2006), and a short overview is presented here.

The main benefit to the user is that a very accurately measured mass of a peptide allows for a considerable reduction of false positive identifications, thus resulting in a *highly confident identification*. The high resolving power helps deal effectively with extremely complex peptide mixtures. Figure 2 illustrates this point; the mass of the two coeluting isobaric peptides can be accurately determined only if peaks are sufficiently resolved.

On the other hand, the fragmentation (MS/MS) spectra of peptides for database searches might not be necessarily required with such high precision. When MS/MS

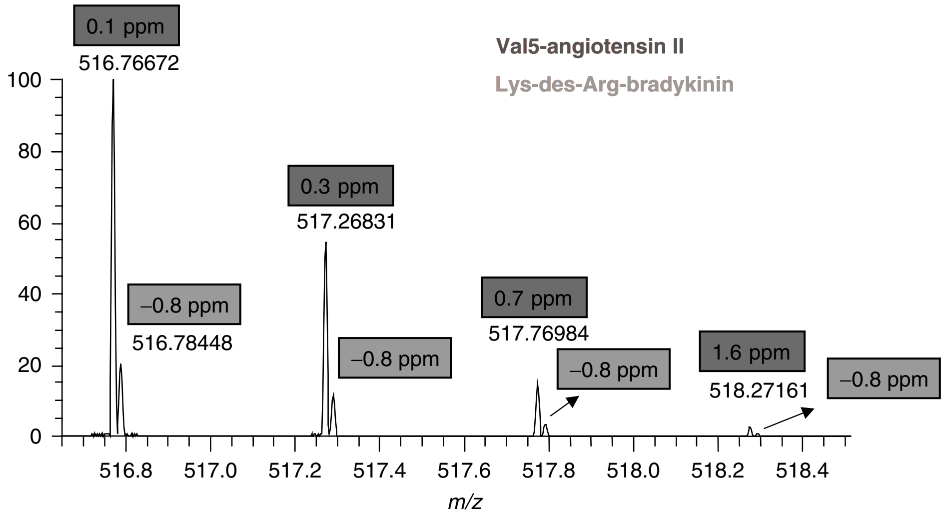


Figure 2 Analyzing isobaric peptides. The two peptides need to be sufficiently resolved (resolving power 60 000 was used in the experiment) to allow for an accurate-mass assignment

spectra is recorded with the linear ion trap detector, three spectra per second can be comfortably obtained. Both mass analyzers can indeed work in parallel: while a high-resolution/mass accuracy spectrum of the precursor is being acquired in the orbitrap, the fast linear ion trap carries out fragmentation and detection of MS/MS (or higher-order MS^n) spectra of selected peptides. A true parallel operation is achieved by allowing for a short preview of ions that are being measured in the orbitrap analyzer. This preview defines the parent ions that the linear ion trap will fragment. Within one second, one orbitrap spectrum acquired at resolving power 60 000 together with three linear ion trap fragmentation spectra can be obtained (Makarov *et al*., 2006).

Analysis of *posttranslational modifications* is an area of proteomics that greatly benefits from accurate-mass measurement as well as multiple levels of fragmentation (Olsen *et al*., 2006). An accurate measurement of the neutral loss of phosphate from the precursor ion in the MS/MS spectrum allows for a selective targeting of phosphopeptides in complex mixtures, while the site of phosphorylation is confidently identified in the following MS^3 spectrum.

De novo sequencing of peptides is arguably the biggest challenge in proteomics. The problem is compounded by a “combinatorial explosion” caused when considering multiple modifications. Mass accuracy and high resolution considerably improve the results of de novo interpretation of peptide MS/MS spectra using computer algorithms (Figure 3).

The orbitrap mass analyzer offers sufficiently high resolving power to encourage attempts at analyzing medium-sized proteins (10–25 kDa, Figure 4). This approach, called a *top-down analysis*, enables a detailed characterization of the protein, including the determination of posttranslational modifications.

In the areas related to small molecule analysis, the ability to conduct *biotransformation profiling* via tandem mass spectrometry coupled with accurate-mass

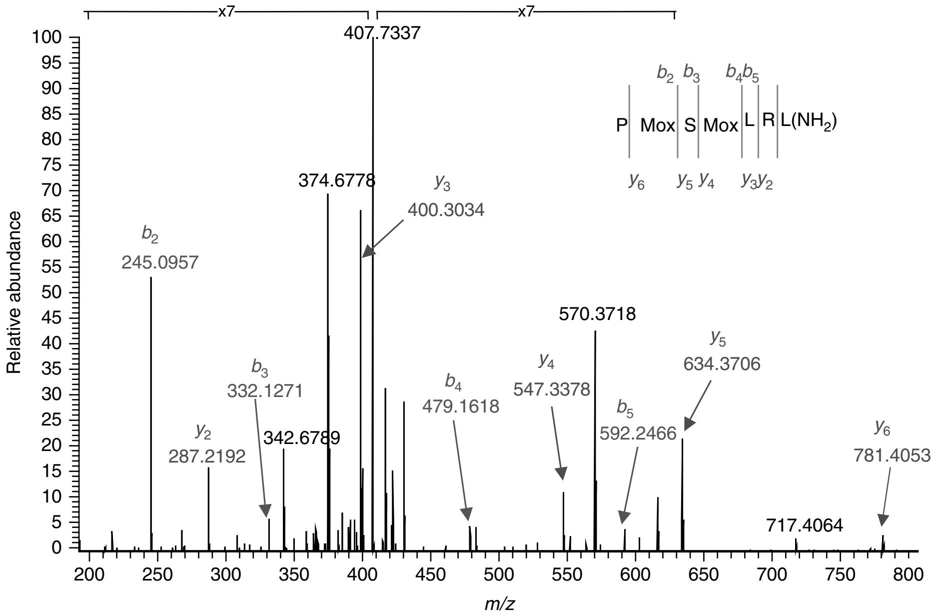


Figure 3 De novo sequence assignment of a snail peptide. Note the presence of three modified amino acid residues. It is worth pointing out that the orbitrap mass analyzer confidently differentiates between oxidized methionine and phenylalanine candidates (mass difference 0.033 u). (The sample was kindly provided by Dr Ka Wan Li of the Free University, Amsterdam. Reprinted from Scigelova M and Makarov A (2006) Orbitrap mass analyzer – overview and applications in proteomics. *Practical Proteomics*, 6, 16–21, with the permission of the copy-right owner)

measurement, all in a single experiment, is clearly one of the most attractive features of the orbitrap mass analyzer. The tight mass tolerance reduces or eliminates background chemical noise. Thus, suspected metabolites can be confirmed or refuted using (1) the predictive chemical formula and corresponding mass error of the analysis, (2) ring-plus double bond equivalent rule, and (3) accurate-mass measurement of product ion spectra of suspected metabolites (Peterman *et al*., 2006).

In the context of whole biological systems, the study of *metabolic networks* has so far been hindered by the lack of techniques that identify metabolites and their biochemical relationship. The orbitrap mass analyzer is proving to be very useful in this field (Breitling *et al*., 2006).

In some areas of research, the use of a high-resolution/accurate-mass analyzer with MS^n capabilities is poised to cause a major shift in experimental approaches and strategies. For instance, the study of lipid mixtures traditionally using parent and neutral ion scanning on triple-quadrupole mass spectrometers is now able to adopt a “profiling” approach relying on high resolution/accurate mass (Ejsing *et al*., 2006). This approach has the potential to revolutionize the *study of lipids*, making it a high-throughput global approach akin to biomarker discovery in the areas of proteomics and metabolomics.

Yet another interesting application of the orbitrap mass spectrometer to small molecule analysis is in the area of *drugs of abuse*. The orbitrap analyzer allows for a

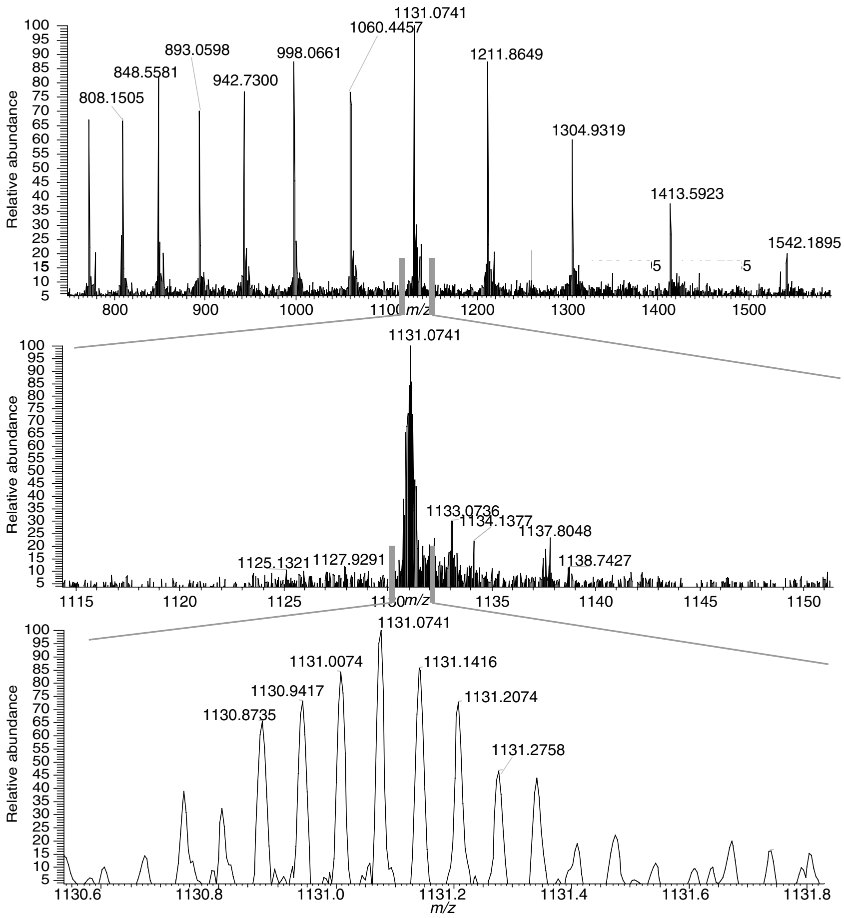


Figure 4 Intact myoglobin is measured in the orbitrap mass analyzer at resolving power 100 000 in an infusion experiment. (Reprinted from Scigelova M and Makarov A (2006) Orbitrap mass analyzer – overview and applications in proteomics. *Practical Proteomics*, **1**, In press, with the permission of the copy-right owner)

thorough characterization of novel compounds that have a high potential for misuse in sports (Thevis *et al*., 2006a, 2006b). The MS^n capability of the hybrid linear ion trap–orbitrap mass spectrometer is used to determine fragmentation pathways while a routine analysis is then conducted on a triple-quadrupole mass spectrometer.

In conclusion, it is clear that the orbitrap mass analyzer is fast becoming a unique and powerful addition to the scientific toolbox for probing biological systems and increasing selectivity and confidence of routine analyses. The above shortlist of applications is bound to expand considerably in the near future as the orbitrap systems are becoming more widespread and penetrate into other areas of research. Undoubtedly, we will see major breakthroughs in the design and development of the orbitrap analyzer technology.

References

Breitling R, Pitt AR and Barrett MP (2006) Precision mapping of the metabolome. *Trends in Biotechnology*, doi: 10.1016/j.tibtech.2006.10.006.

Ejsing C, Moehring T, Bahr U and Duchoslav E (2006) Collision-induced dissociation pathways of yeast sphingolipids and their molecular profiling in total lipid extracts: a study by quadrupole TOF and linear ion trap–orbitrap mass spectrometry. *Journal of Mass Spectrometry*, **41**, 372–389.

Hardman M and Makarov A (2003) Interfacing the orbitrap mass analyzer to an electrospray ion source. *Analytical Chemistry*, **75**, 1699–1705.

Hu Q, Noll RJ, Li H, Makarov A, Hardman M and Cooks RG (2005) The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, **40**, 430–443.

Kingdon KH (1923) A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review*, **21**, 408–418.

Knight RD (1981) Storage of ions from laser-produced plasmas. *Applied Physics Letters*, **38**, 221–222.

Korsunskii MI and Basakutsa VA (1958) A study of the ion-optical properties for a sector-shaped electrostatic field of the difference type. *Soviet Physics-Technical Physics*, **3**, 1396.

Makarov AA (1999) *US Pat.* 5,886,346.

Makarov A (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, **72**, 1156–1162.

Makarov A, Denisov E, Kholomeev A, Balschun W Lange O, Strupat K and Horning S (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry*, **78**, 2113–2120.

Marshall AG and Verdun FR (1990) *Fourier Transforms in NMR, Optical, and Mass Spectrometry: A User's Handbook*, Elsevier, Amsterdam, Netherlands.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, in vivo, and site-specific phosphorylation dynamics in signalling networks. *Cell*, **127** 635–648.

Peterman SM, Duczak N, Kalgutkar AS, Lame ME and Soglia JR (2006) Application of a linear ion trap/orbitrap mass spectrometer in metabolite characterisation studies: examination of the non-tricyclic anti-depressant nefazodone using data-dependent accurate mass measurement. *Journal of the American Society for Mass Spectrometry*, **17**, 363–375.

Schwartz J, (2005) *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley & Sons, Ltd, Chichester, UK.

Scigelova M and Makarov A (2006) Orbitrap mass analyzer – overview and applications in proteomics. *Practical Proteomics*, **6**, 16–21.

Thevis M, Kamber M and Schaenzer W (2006a) Screening for metabolically stable aryl-propionamide derived selective androgen receptor modulators for doping control purposes. *Rapid Communications in Mass Spectrometry*, **20**, 870–876.

Thevis M, Krug O and Schaenzer W (2006b) Mass spectrometric characterization of efaproxiral (RSR13) and its implementation into doping controls using liquid chromatography–atmospheric pressure ionization-tandem mass spectrometry. *Journal of Mass Spectrometry*, **41**, 332–338.

Short Specialist Review

Time-of-flight mass spectrometry

Robert J. Cotter
Johns Hopkins University School of Medicine, Baltimore, MD, USA

1. Introduction

Time-of-flight (TOF) mass spectrometers were introduced in the 1940s (Stephens, 1946; Cameron and Eggers, 1948) and commercialized in 1955 (Wiley and McLaren, 1955). Regarded for many years as instruments with low mass range and low resolving power, TOF mass spectrometers were "rediscovered" following the introduction of ionization techniques capable of ionizing biological macromolecules including proteins and peptides, oligosaccharides, glycolipids and other glycoconjugates, and polynucleotides. Thus, they have become important for proteomics not only in their ability to obtain molecular weights of intact proteins but also for obtaining amino acid sequences of peptides and for locating and characterizing posttranslational modifications.

2. Mass analysis and mass range

The basic TOF mass spectrometer consists of a small source region s, a longer drift length D and a detector (Figure 1a). Ions are formed in the source by any number of ionization techniques that have included *electron ionization* (EI) (Wiley and McLaren, 1955), *chemical ionization* (CI) (Futrell *et al*., 1968), *plasma desorption mass spectrometer* (PDMS) (Torgerson *et al*., 1974), *laser desorption* (LD) (Van Breemen *et al*., 1983), *matrix-assisted laser desorption/ionization* (MALDI) (Karas and Hillenkamp, 1988; Tanaka *et al*., 1988), and *electrospray ionization* (ESI) (Verentchikov *et al*., 1994). Ions of all masses m are then accelerated as they exit the source to final kinetic energies $\frac{1}{2}mv^2 = eV$, where e is the charge on an electron and V is the voltage across the source. The ions then travel through a drift length at velocities v, which are constant but different for each mass, before striking the detector. The TOF for an ion is then given by

$$t = \left(\frac{m}{2eV}\right)^{1/2} (2s + D) \approx \left(\frac{m}{2eV}\right)^{1/2} D \qquad (1)$$

In theory, the mass range of such an instrument is unlimited. For current instruments using MALDI, detectors limit the mass range, as the efficiency for conversion

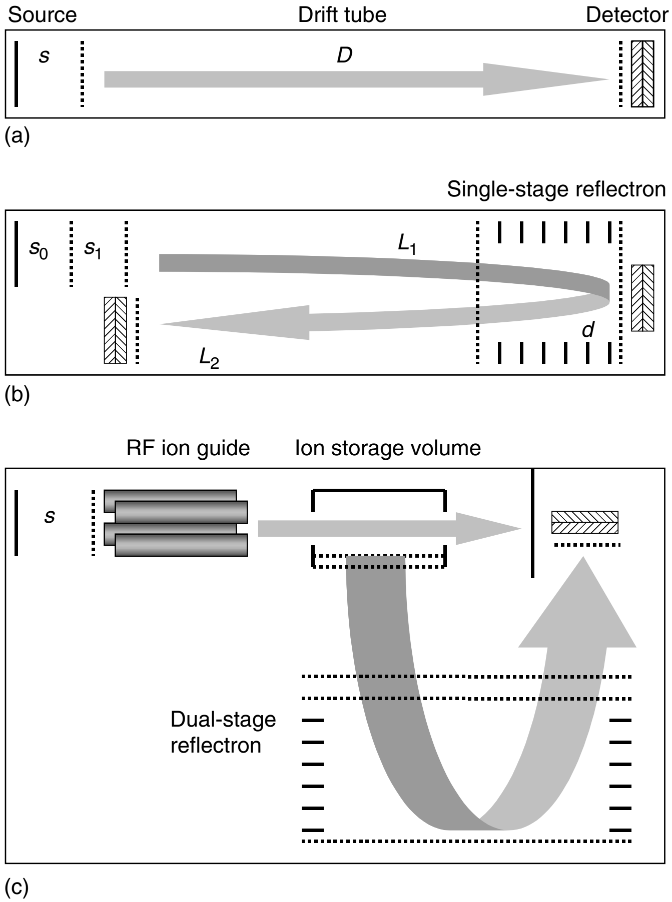


Figure 1 Basic configurations of time-of-flight mass spectrometers: (a) a simple linear TOF mass analyzer with a single-stage ionization source, (b) a *reflectron* TOF mass analyzer with a dual-stage ion extraction source, and (c) an *orthogonal acceleration* mass analyzer with a *quadrupole ion guide* and a dual-stage reflectron

of incoming ions to an electron cascade by electron multipliers and channel plate detectors depends upon ion velocity. That velocity is slower for higher mass ions and is the reason that current instruments operate at high (generally 20 kV) accelerating voltages. At the same time, cryo-cooled superconducting tunneling junction (SJT) (Hilton *et al*., 1998) detectors are now providing megadalton detection, but at lower time resolution.

3. Spatial, energy distributions and mass resolution

The expression for mass resolving power: $m/\Delta m = t/2\Delta t$ indicates that this will be only as good as our ability to parse the time measurement into very small time intervals. The *boxcar* method used by early commercial instruments (which

enabled recording of only one time interval in each TOF cycle) (Holland *et al.*, 1983) was capable of resolving time intervals ranging from 10 to 40 ns. Today, fast analog-to-digital converters (transient recorders, waveform recorders, and digital oscilloscopes) have digitization rates up to 8 Gsample s^{-1}, corresponding to time intervals of 125 ps. When single ion counting is used, time-to-digital converters (TDCs) provide an accuracy of 200–500 ps.

Mass resolution is also affected by the initial conditions of ions. For example, ions formed at different distances between the back of the source and the extraction grid will have different kinetic energies as they leave the source, and therefore different velocities and different TOFs. This is because an initial *spatial distribution* is effectively a different distance s, and a different energy due to acceleration $eV = eEs$, where E is the electric field in the source. Ions may, of course, have different kinetic energies before or as a result of ionization. This initial *kinetic energy distribution* U_0 means that ions will have different final energies as well: $eV + U_0$. The effects of both of these are given in a more complex TOF equation:

$$t = \frac{(2m)^{1/2}}{eE}\left[(U_0 + eEs)^{1/2} \mp U_0^{1/2}\right] + \frac{(2m)^{1/2} D}{2\,(U_0 + eEs)^{1/2}} \tag{2}$$

The first term is the time in the source and includes the so-called turn-around time (Wiley and McLaren, 1955), since the initial velocity component along the flight path may be in the forward or reverse direction. The second term is the time in the flight tube and reflects the final energy. While the turn-around time is important in gas phase ionization methods, such as EI, it is generally not in methods such as MALDI in which ions are formed on a surface.

The effects of the initial spatial distribution are the easier of the two to correct. Ions formed about a small distance $s + \Delta s$ will be focused to first order at a distance $d = 2s$ from the source, the so-called *space-focus plane*. Because this distance is not far enough from the source to enable sufficient mass/time dispersion and achieve good resolving power, it is more common to use a *dual-stage extraction* source (Figure 1b), in which the field strength E_1 in the second region s_1 is much greater than that E_0 in the first s_0. In this case, the space focus is given by

$$d = 2\sigma^{1/2}\left[\frac{1}{s_0^{1/2}} - \frac{2s_1}{s_0^{1/2}\left(\sigma^{1/2} + s_0^{1/2}\right)^2}\right] \tag{3}$$

where $\sigma = s_0 + (E_1 E_0)s_1$. In addition, it is possible to achieve better (second order) focusing for an optimal set of parameters.

4. Time-lag focusing and delayed extraction

The greatest challenge for achieving high resolving power is the simultaneous focusing of the initial spatial and energy distributions. The *time-lag focusing* method of Wiley and McLaren introduced a delay time between the ionizing

electron beam pulse forming the ions and the extraction pulse accelerating them from the source. During that time, ions with larger forward velocity would move closer to the extraction grid, so that they would receive less energy from acceleration than slower moving ions. Prior to the development of MALDI this approach was used for infrared laser desorption (IRLD) (Van Breemen *et al.*, 1983; Tabet and Cotter, 1984).

Laser microprobes (Hillenkamp *et al.*, 1975) and PDMSs (Torgerson *et al.*, 1974) relied on the fact that ions formed directly from a surface would have a very minimal spatial distribution, so that correction of the kinetic energy could be accomplished, without pulsed extraction, using a reflectron (see below). Following the introduction of MALDI[7] (Karas and Hillenkamp, 1988; Tanaka *et al.*, 1988) there was a return of interest in pulsed extraction methods (Whittal and Li, 1995; Brown and Lennon, 1995; Vestal *et al.*, 1995), which have produced very high mass resolutions. While known commonly as *delayed extraction*, King *et al.* (1995) have termed this approach *velocity-space correlated focusing* because the spatial distribution at the time of the extraction pulse (which is absent initially) is derived entirely from the initial velocities (energies).

Time-lag focusing and the more recent pulsed extraction methods are generally accomplished with dual-stage extraction and are all mass dependent. Because the velocity distributions will be different for different masses, motion during the delay period will be different. Thus, either the delay time or the ratio E_1/E_0 is different for each mass. Over many years a number of dynamic methods have been developed, including *impulse field focusing* (Marable and Sanzone, 1974), *postsource pulsed focusing* (Kinsel and Johnston, 1989), *velocity compaction* (Muga, 1988), and *dynamic-field focusing* (Yefchak *et al.*, 1989). *Mass-correlated acceleration*, developed more recently (Kovtoun *et al.*, 2002), takes advantage of the fact that heavier ions enter the second region later than lighter ions and dynamically changes the field E_1 by raising the flight tube voltage. In many cases, it is found that focusing in the middle mass range (5000–6000 Da) produces reasonable focus throughout the useful mass range for MALDI.

5. Reflectrons and postsource decay

Mamyrin and coworkers first reported their *reflectron* or *ion mirror* in 1973 (Mamyrin *et al.*, 1973). Shown in Figure 1b, this mass analyzer incorporates a retarding electric field that reverses the direction of the ions. The device works to correct for an initial kinetic energy spread because the higher velocity ions penetrate more deeply into the reflectron and take longer to turn around. For a single-stage reflectron:

$$t = \left(\frac{m}{2eV}\right)^{1/2} [L_1 + L_2 + 4d] \tag{4}$$

where L_1 and L_2 are the drift paths toward and away from the reflectron, and d is the penetration depth. The single-stage reflectron (Tang *et al.*, 1988) provides first order energy focusing, while the dual-stage reflectron (an example is shown

in Figure 1c) can provide second order focusing. A *quadratic reflectron* (Mamyrin, 1994) provides (in theory) infinite order focusing, that is, the flight time is independent of energy, but the nonlinear field along the center axis also includes off-axis field lines that cause divergence of the ion beam and loss of transmission.

When an ion of mass m_1 originating from the source dissociates to form a fragment m_2 in the flight tube, its flight time through a single-stage reflectron is

$$t = \left(\frac{m_1}{2eV}\right)^{1/2}\left[L_1 + L_2 + 4\frac{m_2}{m_1}d\right] \quad (5)$$

This offers the opportunity to select a given precursor ion m_1 using an electronic gate in the flight tube L_1 and to record the flight times of all the product ions m_2, a method generally referred to as *postsource decay* (PSD) (Kaufmann *et al*., 1994). Focusing is poorer for lower mass product ions that do not approach the optimal penetration depth, a problem that is generally addressed by acquiring multiple spectral segments at reduced reflectron voltage. The *curved-field reflectron* (CFR) (Cornish and Cotter, 1993; Cornish and Cotter, 1994) eliminates the need for stepping the reflectron voltage, focusing the entire product ion mass range simultaneously. Though this reflectron uses a nonlinear field, it is closer to the constant field reflectron than to the quadratic, so that high ion transmission is maintained.

6. Orthogonal acceleration

The development of *orthogonal acceleration* (oaTOF) mass spectrometers has not only provided a means for achieving high mass resolution but has also provided the opportunity to utilize continuous ionization sources, most notably electrospray ionization. The first such instrument by Dawson and Guilhaus (1989) used an EI source and extraction into a linear TOF mass analyzer, while a later instrument by Mirgorodskaya *et al*. (1994) used an ESI source with a reflectron. Essential to the operation of the instrument (and to its ultimate mass resolution) is the collimation of the ion beam into a storage volume (Figure 1c) so that the remaining component of the velocity spread is directed along a single axis. Ions are then extracted orthogonal to that axis, using space focusing to correct for any width in the collimated beam. While a collimated beam can be achieved by passing the beam through a small orifice, much better transmission can be achieved using a quadrupole, hexapole, or octapole ion guide (Krutchinsky *et al*., 1998). With such a configuration, the mass analyzer is effectively isolated from the source and has no memory of the initial conditions of the ions. Thus, multiple sources can be used, as can atmospheric pressure sources such as electrospray and atmospheric pressure MALDI (AP MALDI) (Laiko *et al*., 2000).

7. Hybrid instruments

The development of oaTOF mass spectrometers has provided the opportunity for a very successful tandem (hybrid) configuration that combines a quadrupole mass

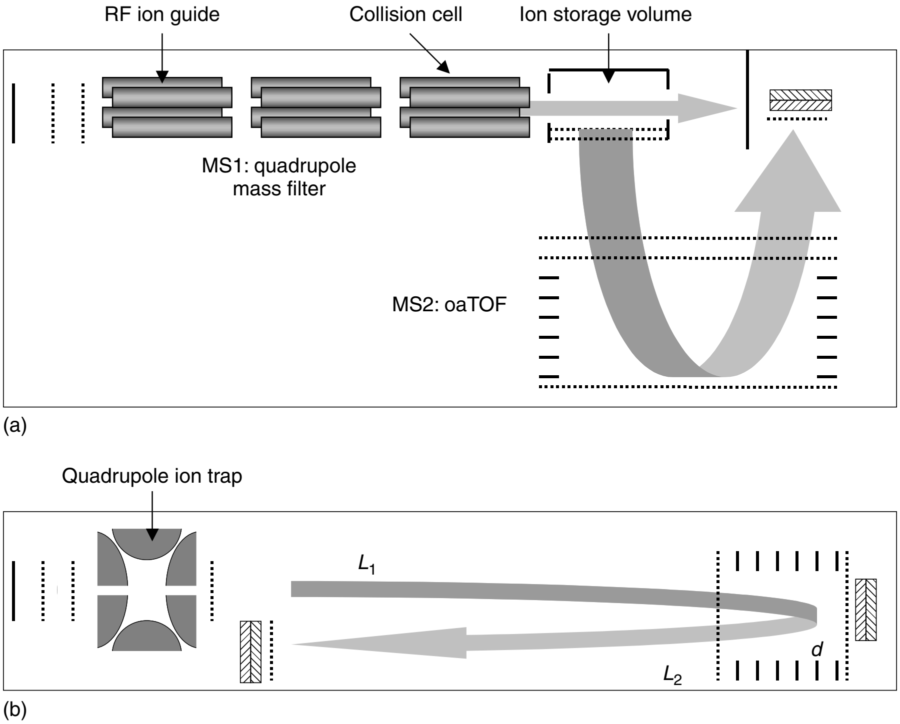


Figure 2 Hybrid instruments: (a) a tandem quadrupole/time-of-flight mass spectrometer with an RF ion guide and collision chamber, and (b) a tandem ion trap/time-of-flight mass spectrometer

analyzer in the first stage and an orthogonal TOF as the second mass analyzer (Figure 2a) (Shevchenko *et al*., 1997; *see also* Article 10, **Hybrid MS**, Volume 5). In a tandem experiment, the first mass analyzer is generally used to pass a precursor mass while the second mass analyzer scans its product ion spectrum; thus, the quadrupole is an appropriate mass filter as the first stage while the TOF analyzer provides the multichannel advantage for recording the product ion spectrum. The mass/charge range of quadrupoles is more limited than the TOF, but is nicely compatible with the multiply charged ions from an electrospray source, which it accommodates using an RF ion guide as described above. An additional quadrupole provides a low energy *collision-induced dissociation* (CID) chamber for inducing fragmentation. Because fragmentation of multiple-charged ions is generally accompanied by charge reduction, product ions may have considerably higher *m/z* than their precursors, but this is easily accommodated by the TOF analyzer. An additional advantage is that selection of the precursor ion mass is much better than it is on the TOF/TOF instruments described below.

A configuration utilizing an electrostatic ion trap (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) with the TOF mass spectrometer was introduced by Chien and Lubman (1994), and was also intended to provide compatibility between a continuous beam ESI source and the pulsed TOF mass analyzer. A *trapTOF* instrument with a quadratic reflectron was reported by

Doroshenko and Cotter (1998). An advantage of the combined ion trap/TOF (Figure 2b) is that one can carry out MS^n analyses. While this is possible on an ion trap alone, the recorded spectrum at any stage of MS^n on the ion trap/TOF is always a TOF spectrum with higher mass resolution than can be obtained on the trap.

8. Tandem time-of-flight mass spectrometers

The difficulties inherent in stepping the reflectron voltage for PSD, as well as an interest in providing collision-induced dissociation of larger singly charged ions, has motivated the development of tandem TOF (or TOF/TOF) mass spectrometers. Such instruments are not entirely new; a very simple linear system designed by Jardine *et al*. (1992) used a voltage step after the collision cell to differentiate the flight times of the product ions. Beussman *et al*. (1995) and Cotter and Cornish (1993) both reported instruments using two reflectrons in a *Z* geometry in 1993. In the first instrument, photodissociation of the precursor ion was followed by additional acceleration of the product ions to accommodate the bandwidth of the reflectron; in the latter, pulsed CID was followed by a CFR to enable focusing of the product ions.

The most recent approaches to TOF/TOF design use a linear, pulsed extraction first analyzer, a mass selection gate at (or near) the space-velocity focus point, a collision chamber, and a second reflectron mass analyzer (Figure 3). These instruments differ in how they address the focusing of product ions in the reflectron. Specifically, precursor ions of mass m_1 produce product ions m_2 whose kinetic energies are proportional to their masses: (m_2/m_1) eV. Thus, product ion energies can range from the precursor ion energy (typically 20 keV) to a few eV, while single and dual-stage reflectrons focus only a fraction of that energy range. In one configuration (Figure 3a), precursor ions are decelerated to 2 keV prior to collision. After the collision, the product ions are pulse extracted from (effectively) a second source, which reaccelerates the ions by an additional 18 keV, so that ions with an energy range from 18 to 20 keV enter the reflectron (Medzihradszky *et al*., 2000). In a second approach (Figure 3b), precursor ions leave the source and enter the collision chamber at 8 keV and are not decelerated. The product ions (still traveling at the same velocity) enter a field-free *lift cell* whose potential is raised (by 19 kV) while the ions are in residence (Suckau *et al*., 2003). A third configuration (Figure 3c) uses a CFR to focus all the product ions resulting from 20-keV collisions, obviating the need for decelerating, reextracting, or lifting the product ions (Iltchenko *et al*., 2004). These configurations provide much-improved approaches compared with stepping of the reflectron voltage, plus the opportunity to locate mass selection at the space-velocity focus point and to carry out CID between the mass analyzers. Interestingly, MS/MS spectra on these instruments are not very different with and without a collision gas, which suggests that the processes are fundamentally the same: PSD or *laser induced dissociation* producing high internal energies, with fragmentation then aided by the collisions. At the same time, one observes the appearance of (or increases in) internal and single amino acid ions, which may result from a second, collision-induced fragmentation (Medzihradszky *et al*., 2000; Iltchenko *et al*., 2004; *see also* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5).

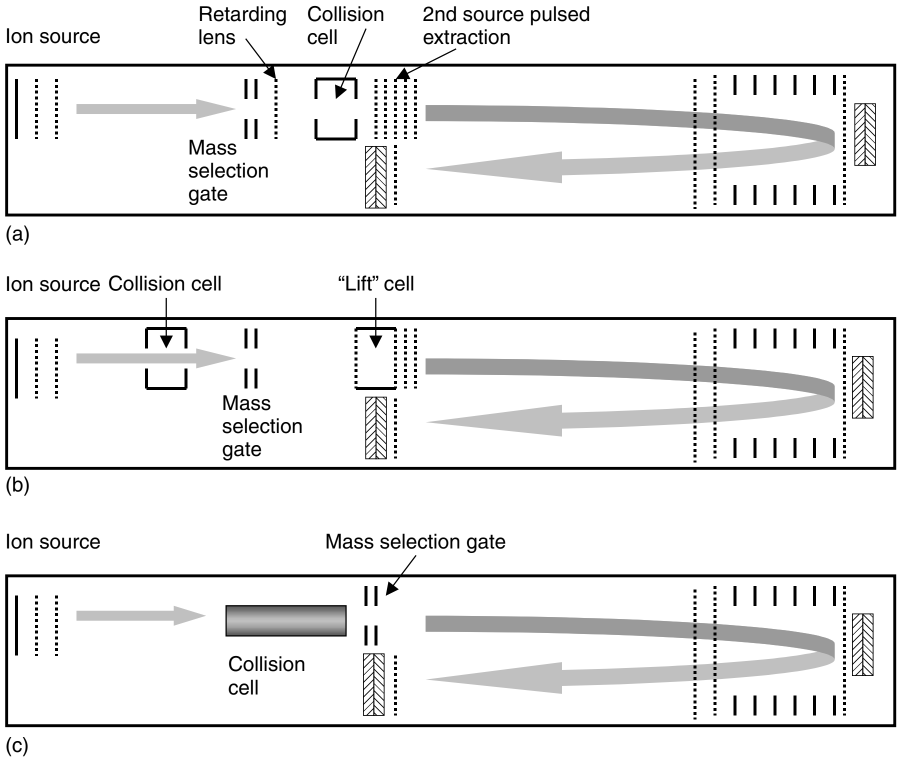


Figure 3 Tandem time-of-flight mass spectrometers using different approaches for focusing product ions: (a) deceleration prior to collision followed by reacceleration of product ions, (b) potential lifting of product ions, and (c) focusing by a curved-field reflectron

9. Future prospects

There is considerable interest in the development of compact, field-portable and/or miniaturized mass spectrometers; in this area, the TOF offers the possibility for retaining the same high mass range as its larger, high-performance counterparts (English *et al*., 2003). While such instruments are expected to play a significant future role in proteomics-based, point-of-care diagnostics, it is also clear that there is currently an increasing demand for high-performance instrument for the identification and validation of biomarkers. This includes both high mass resolution and MS/MS capabilities.

Further reading

Cotter RJ (1997) *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research*, American Chemical Society: Washington DC, ISBN 0-8412347-4-4.

Kinter M and Sherman NE (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, ISBN 0-4713224-9-0.

Siuzdak G (2003) *The Expanding Role of Mass Spectrometry in Biotechnology*, ISBN 0-9742-4510-0.

Warscheid B, Jackson K, Sutton C and Fenselau C (2003) MALDI analysis of *Bacilli* in spore mixtures by applying a quadrupole ion trap time-of-flight tandem mass spectrometer. *Analytical Chemistry*, **75**, 5608–5617.

References

Beussman DJ, Vlasak PR, McLane RD, Seeterlin MA and Enke CG (1995) Tandem reflectron time-of-flight mass spectrometer utilizing photodissociation. *Analytical Chemistry*, **67**, 3952–3957.

Brown RS and Lennon JJ (1995) Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer. *Analytical Chemistry*, **67**, 1998–2003.

Cameron AE and Eggers DF Jr (1948) An Ion Velocitron. *The Review of Scientific Instruments*, **19**, 605.

Chien BM and Lubman DM (1994) Analysis of the fragments from collision-induced dissociation of electrospray-produced peptide ions using a quadrupole ion trap storage/reflectron time-of-flight mass spectrometer. *Analytical Chemistry*, **66**, 1630–1536.

Cornish TJ and Cotter RJ (1993) A curved field reflectron for improved energy focusing of product ions in time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry: RCM*, **7**, 1037–1040.

Cornish TJ and Cotter RJ (1994) A curved field reflectron time-of-flight mass spectrometer for the simultaneous focusing of metastable product ions. *Rapid Communications in Mass Spectrometry: RCM*, **8**, 781–785.

Cotter RJ and Cornish TJ (1993) A tandem time-of-flight (TOF/TOF) mass spectrometer. *Analytical Chemistry*, **65**, 1043–1047.

Dawson JHJ and Guilhaus M (1989) Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry: RCM*, **3**, 155.

Doroshenko VM and Cotter RJ (1998) A quadrupole ion trap/time-of-flight mass spectrometer with a parabolic reflectron. *Journal of Mass Spectrometry: JMS*, **33**, 305–318.

English RD, Warscheid B, Fenselau C and Cotter RJ (2003) *Bacillus* spore identification via proteolytic peptide mapping with a miniaturized MALDI TOF mass spectrometer. *Analytical Chemistry*, **75**, 6886–6893.

Futrell JH, Tiernan TO, Abramson FP and Miller CD (1968) Modification of a Time-of-Flight Mass Spectrometer for Investigation of Ion-Molecule Reactions at Elevated Pressures. *The Review of Scientific Instruments*, **39**, 340–345.

Hillenkamp F, Unsold E, Kaufmann R and Nitsche R (1975) Laser microprobe mass analysis of organic materials. *Nature*, **256**, 119.

Hilton GC, Martinis JM, Wollmann DA, Irwin KD, Dulcie LL, Gerber D, Gillevet PM and Twerenbold D (1998) Impact energy measurement in time-of-flight mass spectrometry with cryogenic microcalorimeters. *Nature*, **391**, 672–675.

Holland JF, Enke CG, Allison J, Stults JT and Pinkston JD (1983) Mass spectrometry in the chromatographic time frame. *Analytical Chemistry*, **65**, 997A–1112A.

Iltchenko S, Gardener B, English RD and Cotter RJ (2004) Tandem time-of-flight mass spectrometry with a curved field reflectron. *Analytical Chemistry*, **76**, 1976–1981.

Jardine DR, Morgan J, Alderdice DS and Derrick PJ (1992) A tandem time-of-flight mass spectrometer. *OMS, Organic Mass Spectrometry*, **27**, 1077–1083.

Karas M and Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Analytical Chemistry*, **60**, 2299–2301.

Kaufmann R, Kirsch D and Spengler B (1994) Sequencing of peptides in a time-of-flight mass spectrometer: evaluation of postsource decay following matrix-assisted laser desorption ionisation (MALDI). *International Journal of Mass Spectrometry and Ion Processes*, **131**, 355–385.

King TB, Colby SM and Reilly JP (1995) High resolution MALDI-TOF mass spectra of three proteins obtained using space-velocity correlation focusing. *International Journal of Mass Spectrometry and Ion Processes*, **145**, L1–L7.

Kinsel GR and Johnston MV (1989) Post source pulse focusing: a simple method to achieve improved resolution in a time-of-flight mass spectrometer. *International Journal of Mass Spectrometry and Ion Processes*, **91**, 157–176.

Kovtoun SV, English RD and Cotter RJ (2002) Mass correlated acceleration in a reflectron MALDI TOF mass spectrometer: an approach for enhanced resolution over a broad mass range. *Journal of the American Society for Mass Spectrometry*, **13**, 135–143.

Krutchinsky AN, Loboda AV, Spicer VL, Dworschak R, Ens W and Standing KG (1998) Orthogonal injection of matrix-assisted laser desorption/ionization ions into a time-of-flight spectrometer through a collisional damping interface. *Rapid Communications in Mass Spectrometry: RCM*, **12**, 508–518.

Laiko VV, Baldwin MA and Burlingame AL (2000) Atmospheric pressure matrix-assisted laser desorption/ionization mass spectrometry. *Analytical Chemistry*, **72**, 652–657.

Mamyrin BA (1994) Laser assisted reflectron time-of-flight mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes*, **131**, 1–19.

Mamyrin BA, Karatajev VJ, Shmikk DV and Zagulin VA (1973) Mass reflectron. New nonmagnetic time-of-flight high-resolution mass spectrometer. *Soviet Physics JETP*, **37**, 45–48.

Marable NL and Sanzone G (1974) High-resolution time-of-flight mass spectrometry. Theory of the impulse-focused time-of-flight mass spectrometer. *International Journal of Mass Spectrometry and Ion Physics*, **13**, 185–194.

Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML and Burlingame AL (2000) The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Analytical Chemistry*, **72**, 552–558.

Mirgorodskaya OA, Shevchenko AA, Chernushevich IV, Dodenov AF and Miroshnikov AI (1994) Electrospray-ionization time-of-flight mass spectrometry in protein chemistry. *Analytical Chemistry*, **66**, 99–107.

Muga ML (1988) Velocity compaction time-of-flight mass spectrometer for mass range 1000–10,000 mu. *Analytical Instrumentation*, **16**, 31.

Shevchenko A, Chernushevich I, Ens W, Standing KG, Thomson B, Wilm M and Mann M (1997) Rapid 'de Novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of flight mass spectrometer. *Rapid Communications in Mass Spectrometry: RCM*, **11**, 1015–1024.

Stephens WE (1946) A Pulsed Mass Spectrometer with Time Dispersion. *Physical Review*, **69**, 691.

Suckau D, Resemann A, Schuerenberg M, Hufnagel P, Franzen J and Holle A (2003) A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Analytical and Bioanalytical Chemistry*, **376**, 952–965.

Tabet JC and Cotter RJ (1984) Laser desorption time-of-flight mass spectrometry of high mass molecules. *Analytical Chemistry*, **56**, 1662.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida y and Yoshida T (1988) Protein and polymer analyses up to m/z 100 000 by laser ionization time-of flight mass spectrometry. *Rapid Communications in Mass Spectrometry: RCM*, **2**, 151–153.

Tang X, Beavis R, Ens W, LaFortune F, Schueler B and Standing KG (1988) A secondary ion time-of-flight mass spectrometer with an ion mirror. *International Journal of Mass Spectrometry and Ion Processes*, **85**, 43–67.

Torgerson DF, Skowronski RP and Macfarlane RD (1974) New approach to the mass spectroscopy of non-volatile compounds. *Biochemical and Biophysical Research Communications*, **60**, 616–621.

Van Breemen RB, Snow M and Cotter RJ (1983) Time resolved laser desorption mass spectrometry: I. The desorption of preformed ions. *International Journal of Mass Spectrometry and Ion Physics*, **49**, 35–50.

Verentchikov AN, Ens W and Standing KG (1994) Reflecting time-of-flight mass spectrometer with an electrospray ion source and orthogonal extraction. *Analytical Chemistry*, **66**, 126–133.

Vestal ML, Juhasz P and Martin SA (1995) Delayed extraction matrix-assisted laser desorption time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry: RCM*, **9**, 1044–1050.

Whittal RM and Li L (1995) High-resolution matrix-assisted laser desorption/ionization in a linear time-of-flight mass spectrometer. *Analytical Chemistry*, **67**, 1950–1954.

Wiley WC and McLaren IH (1955) Time-of-flight mass spectrometer with improved resolution. *The Review of Scientific Instruments*, **26**, 1150–1157.

Yefchak GE, Enke CG and Holland JF (1989) Models for mass-independent space and energy focusing in time-of-flight mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, **87**, 313–330.

Short Specialist Review

Quadrupole mass analyzers: theoretical and practical considerations

Frank A. Kero and **Richard A. Yost**
University of Florida, Gainesville, FL, USA

Randall E. Pedder
Ardara Technologies LP, Monroeville, PA, USA

1. Introduction

Quadrupole mass analyzers are dynamic mass filters that play a central role in mass spectrometric proteomic investigations. Quadrupoles are used both alone and in tandem for the isolation (by mass-to-charge) of ions; they may also be used (in RF-only mode) to contain ions during collision-induced dissociation (CID) (Yost and Enke, 1979). Gas-phase ions are generated by a source external to the quadrupole mass analyzer, often at atmospheric pressure. It should be noted that ion production at one atmosphere is an important practical advantage, in that it allows for the easy exchange of ion sources without disturbing the vacuum of the quadrupole region (Cole, 1997). Peptides and proteins are typically introduced into the ion source in the liquid phase either from an HPLC column (for LC/MS), by flow injection analysis (with no HPLC column), or by continuous direct infusion through a fused silica capillary. For heterogeneous samples, a sample cleanup or separation technique such as HPLC or solid-phase extraction is typically incorporated into the method development scheme to allow for the sequential introduction of multiple analytes into the ion source. Simultaneous ionization of multiple analytes, particularly at widely varying concentrations, may result in suppression and prove deleterious in quantitative investigations (Cole, 1997; King *et al*., 2000; Constantopoulos *et al*., 1999).

The most common ionization source coupled with quadrupole mass analyzers for proteomics studies is electrospray ionization (ESI), since it is ideal for compounds such as peptides and proteins that can be ionized in solution. The fundamentals of ESI mechanisms have been described elsewhere (Cole, 1997; Kebarle, 2000; Cech and Enke, 2001). A key advantage of ESI is the ability to impose multiple

charges on an analyte upon ionization. Recall that ions are separated in mass analyzers as a function of their respective mass-to-charge ratios (*m/z*) (Throck Watson, 1997). It follows that multiple charging reduces the mass-to-charge ratio of ions, and thereby extends the mass range of the instrument. Atmospheric pressure chemical ionization (APCI) is a highly efficient alternative to ESI for less polar compounds. ESI and APCI are complementary techniques, as reflected in the phase in which charge transfer processes occur (King *et al*., 2000). Ionization for ESI occurs in the liquid phase, whereas ionization for APCI occurs in the gas phase. Both APCI and ESI are suitable ionization sources for a quadrupole mass analyzer because the ions are generated and transmitted as a continuous beam of ions. The other primary ionization method for proteomics studies is matrix-assisted laser desorption/ionization (MALDI). Since MALDI produces narrow pulses of ions, it is a far better marriage with pulsed mass analyzers (such as time-of-flight mass analyzers; *see* Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5 and Article 14, **Sample preparation for MALDI and electrospray**, Volume 5) than with the continuous mass analyzers such as quadrupoles. Furthermore, MALDI yields almost exclusively singly charged ions, and thus benefits from the essentially unlimited mass range of the time-of-flight mass analyzer. Quadrupole mass analyzers are also commonly used in mass spectrometers employing electron ionization (EI) and chemical ionization (CI), such as those coupled with gas chromatographs (GC/MS). These ionization techniques do not play a significant role in modern proteomic investigations, however.

The purpose of this communication is to provide the reader with a balanced treatment of practical and theoretical considerations for quadrupole mass analyzers, and to present the advantages of interfacing quadrupoles in tandem.

2. Theory and practice of quadrupole mass filters

A quadrupole mass analyzer consists of four electrically isolated hyperbolic or cylindrical rods linked to RF (radio frequency) and DC (direct current) voltages, as described by the schematic in Figure 1 (Pedder, 2001). The combination of RF and DC voltages creates a region of strong focusing and selectivity known as a hyperbolic field. In simplest terms, the ratio of RF/DC allows for the selective transmission of ions of a narrow range of mass-to-charge from the total population of ions introduced from the ionization source.

The idealized hyperbolic field can be described in terms of Cartesian coordinates (x and y directions toward the rods, and z direction along the rods' axis). Ions of a selected *m/z* follow a stable trajectory around the center of the field, and are transmitted in the z direction through the device (Dawson and Douglas, 1999). Motion of these "stable ions" in the x and y directions are small in amplitude, as the lowest energy pathway lies toward the center of the hyperbolic field. Ions of other *m/z* will have unstable trajectories, with increasing displacement in the x and y directions away from the center of the hyperbolic field, and thus strike the quadrupole rods, where they neutralize upon contact. These ions will not transmit efficiently through the quadrupole to the detector. The application of RF and DC voltages creates a region of stability for transmission of ions of a limited *m/z*

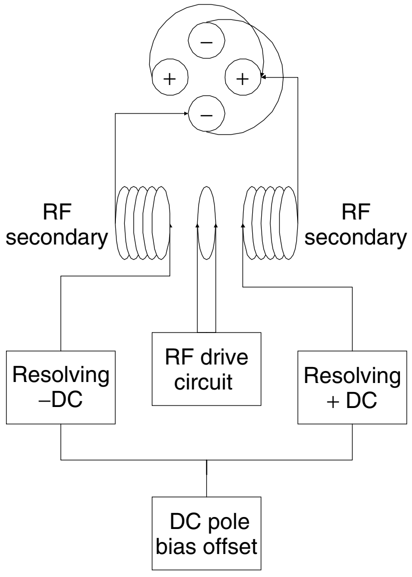


Figure 1 The quadrupole mass filter and its power supply

range. In this regard, the quadrupole analyzer is operating as a mass filter. A mass spectrum is produced by ramping the RF and DC voltages at a nearly constant ratio. Ions of increasing *m/z* will sequentially achieve stable trajectories and reach the detector in order of increasing *m/z*.

To understand the behavior of ions in the quadrupole, a brief introduction to the mathematics associated with ion motion is worthwhile. For a complete discussion, the reader is referred to the work of March and Hughes (1988).

The motion of ions through a quadrupole is described by a second-order linear differential equation called the Mathieu equation, which can be derived starting from the familiar equation relating force to mass and acceleration, $F = ma$, yielding the final parameterized form, with the following substitutions for the parameters a and q (Dawson, 1976):

$$\frac{d^2u}{d\xi^2} + (a_u - 2q_u \cos 2\xi)u = 0 \quad a_u = \frac{8eU}{m{r_0}^2\Omega^2} \quad q_u = \frac{4eV}{m{r_0}^2\Omega^2} \qquad (1)$$

The u in the above equations represents position along the coordinate axes (x or y), ξ is a parameter representing Ωt/2, t is time, e is the charge on an electron, U is the applied DC voltage, V is the applied zero-to-peak RF voltage, m is the mass of the ion, r_0 is the effective radius between electrodes, and Ω is the applied RF frequency in radians s^{-1}. The parameters a and q are proportional to the DC voltage U and the RF voltage V, respectively.

The analytical solution to this second-order linear differential equation is:

$$u(\xi) = \Gamma \sum_{n=-\infty}^{\infty} C_{2n} \exp(2n + \beta)i\xi + \Gamma' \sum_{n=-\infty}^{\infty} c_{2n} \exp -(2n + \beta)i\xi \qquad (2)$$

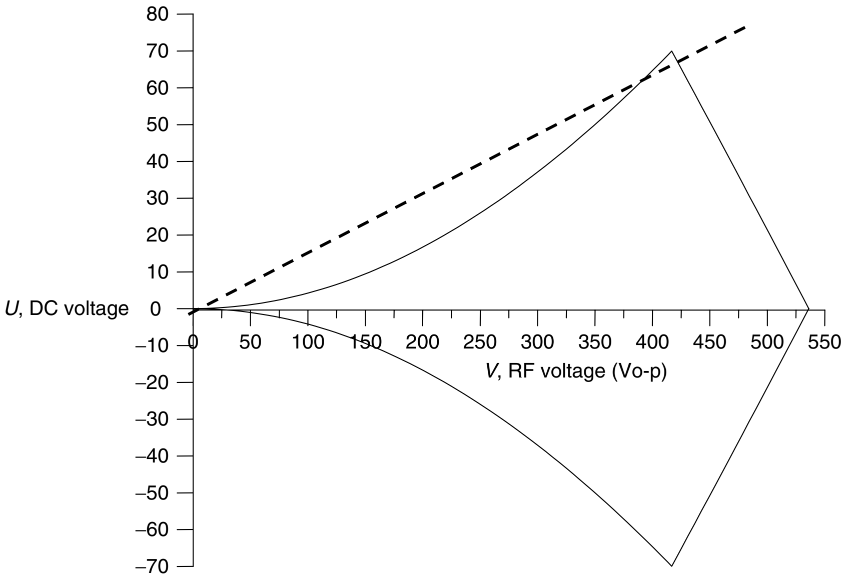


Figure 2 Mathieu stability diagram for an ion of *m/z* 219 in a quadrupole mass filter with 9.5-mm diameter round rods and an RF frequency of 1.2 MHz (Pedder, 2001)

which reduces to a similar infinite sum of sine and cosine functions:

$$u(\xi) = \sum_{n=-\infty}^{\infty} A_n \sin \omega_n + \sum_{n=-\infty}^{\infty} A_n \cos \omega_n \tag{3}$$

The solutions to the Mathieu equation can be presented graphically, as shown in Figure 2 in a so-called stability diagram. Points in (U, V) space (DC, RF voltage space) within the lines lead to stable trajectories for the *m/z* 219 ion; points outside the lines will lead to an unstable trajectory. The dotted line is called the scan line, and shows the ramp of RF and DC voltage at a nearly constant ratio (the slope of the line). If the line passes just under the apex of the stability region, as shown here, the *m/z* 219 ions will have a stable trajectory at those voltages, and will be transmitted to the detector.

Figure 3 shows the stability diagram for three different ions, of *m/z* 28, 69, and 219. Note that ions of lower *m/z* have stable trajectories at lower RF and DC voltages (i.e., they are stable at points in (U, V) space closer to the origin). The dotted scan line again shows the ramp of RF and DC voltage at a constant ratio (Pedder, 2001). The scan line passes through the apices of the stability regions for all three ions. As the RF and DC voltages ramp up in amplitude, ions of increasing *m/z* have stable trajectories and are transmitted, as shown in the dotted mass spectrum below the stability diagram. If the scan line has a lower slope (the solid line), it passes through a larger portion of each stability region, resulting in a mass spectrum with lower resolution, as shown as well.

The selection of RF and DC voltages, V and U, the RF frequency, Ω, and the inscribed radius r_0 between the rods determines the performance of the quadrupole

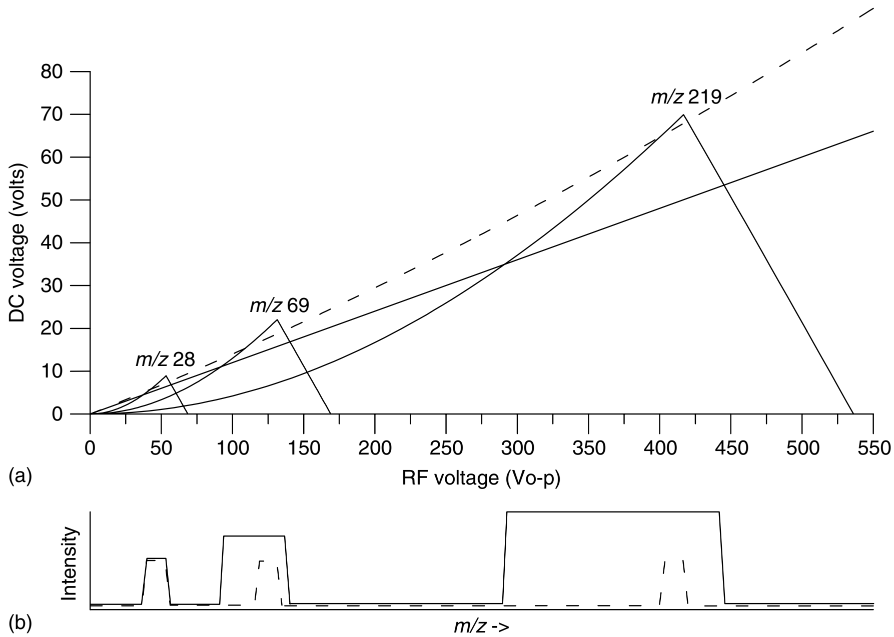


Figure 3 (a) Mathieu stability diagram for ions of *m/z* 28, 69, and 219. Two scan lines are shown, one dotted and a second solid. (b) Simulated mass spectra showing these ions as the RF and DC voltages are scanned along the two scan lines (dotted and solid) shown

mass filter. Typical values are 4 kV RF(0-p), 700 V DC, 1 MHz RF frequency, and 1 cm radius, and yield a mass-to-charge range of ~3000. As can be noted from equation (4), the Mathieu q parameter is proportional to the RF voltage and inversely proportional to (Constantopoulos *et al*., 1999; Pedder, 2002)

$$q_u = \frac{4eV}{m{r_0}^2\Omega^2} \tag{4}$$

the square of the inscribed radius and RF frequency; thus, the mass range of the mass filter can be increased by increasing the RF and DC voltages, or by decreasing the inscribed radius or RF frequency. The mass resolution is a function of the ratio of the RF and DC voltages (the slope of the scan line in Figures 2 and 3); increasing the slope to move the scan line closer to the apex of the stability region (moving from the solid line to the dotted line, for instance) increases the mass resolution, as shown in the spectra in Figure 3(b). Unfortunately, the increase in mass resolution is accompanied by a loss of sensitivity (decrease in transmission of ions through the mass filter). The ultimate resolution that can be achieved is determined not only by how close the scan line approaches the stability region apexes but also by the precision and accuracy of the quadrupole power supplies and the quadrupole rod dimensions.

An important operational mode of the quadrupole is achieved when U, the DC potential, is equal to zero ($a = 0$). This corresponds to a scan line along the x-axis in Figures 2 and 3. This RF-only mode results in transmission of ions of a very

wide range of *m/z* values, and is thus often termed "total ion mode" (Dawson and Douglas, 1999). Note that in this mode, there is a limit to the lowest *m/z* ion that can achieve a stable trajectory. This can be seen as the right-hand edges of the stability regions in Figure 3. This total ion mode is used for the collision cell in tandem quadrupole mass spectrometers, as discussed below; it is also often employed in quadrupoles used to help transmit ions from the ion source region to the quadrupole or other mass analyzer.

3. Tandem mass spectrometry

Most applications of mass spectrometry in proteomics employ mass spectrometry in tandem with one or more other analytical stages (Busch *et al*., 1998; Niessen, 1999). One of those stages is often a separation stage, typically liquid chromatography or capillary electrophoresis (Cole, 1997), leading to LC/MS and CE/MS, as discussed elsewhere in this volume (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5). Another stage may be a second stage of mass spectrometry (the combination of two or more stages of mass spectrometric analysis in series is termed tandem mass spectrometry (McLafferty 1983; Fetterolf and Yost, 1983) or MS/MS, or even MS^n, where n can be ≥ 2). Indeed, a common approach in proteomics investigations is combining a stage of separation with tandem mass spectrometry, that is LC/MS/MS and LC/MS^n. These tandem (sometimes termed hyphenated) methods offer dramatically improved selectivity and information to solve tough analytical problems, as often encountered in proteomics.

Informing power is a figure of merit that provides a means of quantifying the amount of information available in such an analytical procedure. Fetterolf and Yost (1984) demonstrated the use of informing power to illustrate the advantages of hyphenated methods such as LC/MS/MS over single-stage analytical methods. Note that MS/MS is particularly important in LC/MS (compared to combined gas chromatography/mass spectrometry – GC/MS) for two reasons. First, LC typically provides much poorer chromatographic resolution than GC, meaning that LC separations are often incomplete, with coeluting peaks that MS/MS can help resolve. Second, LC/MS employs ionization methods such as ESI and APCI that provide only molecular-type ions and little or no structural information (GC/MS, in contrast, typically employs EI, which provides significant fragmentation and, therefore, structural information with only a single stage of mass spectrometry). Thus, LC/MS/MS is invaluable in providing structural information for the identification of components eluting from the LC column and for detecting those compound with high selectivity.

MS/MS is readily performed "tandem-in-space" (i.e., with two mass analyzers, typically quadrupole mass filters, in sequence). Note that MS/MS can also be performed "tandem-in-time", with two or more stages of mass analysis performed sequentially in time in a single ion trap such as a quadrupole ion trap (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5).

The triple quadrupole mass spectrometer (TQMS, see Figure 4) is the most common MS/MS instrument in use (Yost and Enke, 1979). The operational principles of the triple quadrupole tandem mass spectrometer can be described in a similar

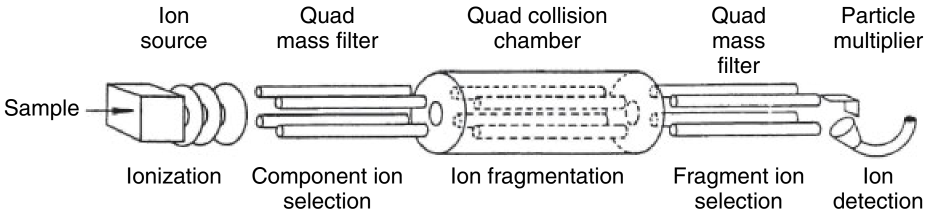


Figure 4 Schematic of a triple quadrupole mass spectrometer (TQMS)

fashion to the single quadrupole. Ions from the ion source enter the first quadrupole mass filter (Q1), which selects ions of a selected *m/z* and transmits them to the second quadrupole. The second quadrupole (Q2) functions as a collision cell in RF-only mode; CID is accomplished by adding a collision gas, typically nitrogen or argon, at a pressure of around 1–10 millitorr (1–10 microbar). The third quadrupole (Q3), another mass filter, provides a means of mass-analyzing the products of CID from the collision cell (Fetterolf and Yost, 1983).

Note that the RF-only quadrupole collision cell in triple quadrupole instruments is often replaced with a higher-order multipole (a hexapole or an octopole, with 6 or 8 rods, respectively) (McCloskey, 1990). The use of an RF-only higher-order multipole provides effective transmission of a wider mass range than an RF-only quadrupole. The stability diagram for a hexapole or octopole (as in Figure 2 for the quadrupole) has less well-defined stability boundaries, so the low-mass cut-off characteristic of an RF-only quadrupole is not an issue with higher-order multipole collision cells. A practical advantage that should be noted is that these multipoles require less demanding tuning procedures for the optimization of voltages and frequencies.

A fundamental understanding of the scan modes associated with the TQMS is essential for understanding the MS and MS/MS capabilities of the instrument. These are summarized in Table 1 (McLafferty, 1983; Fetterolf and Yost, 1983; Busch *et al*., 1998; Dawson and Douglas, 1999).

First, note that the triple quadrupole tandem mass spectrometer can readily be used to perform a single stage of mass spectrometry (MS) by setting two of the three quadrupoles into RF-only (total ion) mode. Either Q1 or Q3 can be used as a mass filter, scanned to obtain a full spectrum, or set to pass a single fixed *m/z* (or a few fixed *m/z* values), a mode termed selected ion monitoring (SIM). When Q1 and Q3 are both used as mass analyzers, there are four possible MS/MS scan modes. In the daughter ion scan mode (also called product ion scan), Q3 is scanned to obtain a spectrum of all the daughter ions generated by CID in q2 of the parent or precursor ion mass-selected with Q1. In the parent ion scan mode (also called precursor ion scan), Q1 is scanned to obtain a spectrum of all the parent ions that upon CID in q2 produce the daughter or product ion mass-selected with Q3. In a neutral loss scan, both Q1 and Q3 are scanned at the same rate but with a fixed difference in mass to detect only those ions that undergo a specific neutral loss in q2 such as a loss of mass 18 (H_2O). An important MS/MS scan mode for quantitating targeted compounds with maximum sensitivity is the selected reaction monitoring mode (SRM) in which one (or a few) selected parent-ion-to-daughter-ion transitions are

Table 1 MS and MS/MS operational modes of a triple quadrupole mass spectrometer, showing the modes for the two mass filters (Quadrupole 1 and Quadrupole 3) and the quadrupole collision cell (q2)

Scan mode	Q1	q2	Q3
MS – full scan[a]	Scan *m/z*	RF-only	RF-only
MS – selected ion monitoring (SIM)[a]	Fixed *m/z*	RF-only	RF-only
MS/MS – daughter ion scan	Fixed *m/z*	RF-only collision gas	Scan *m/z*
MS/MS – parent ion scan	Scan *m/z*	RF-only collision gas	Fixed *m/z*
MS/MS - neutral loss scan	Scan *m/z*	RF-only collision gas	Scan *m/z*[b]
MS/MS – selected reaction monitoring (SRM)	Fixed *m/z*	RF-only collision gas	Fixed *m/z*

[a]Note that the MS scan modes can be performed by mass selection with either Q1 or Q3, with the other one in RF-only (total ion) mode.
[b]Q3 scanned at a difference in mass corresponding to the mass of the selected neutral.

monitored by setting Q1 and Q3 to pass specific *m/z* values. SRM is the MS/MS scan mode analogous to the MS selected ion monitoring scan mode (SIM).

The relationship between the MS/MS scan modes can be explored by examining the plot of the three-dimensional MS/MS data – intensity (on a logarithmic scale) versus parent ion *m/z* versus daughter ion *m/z* – for a single compound, as shown in Figure 5 (Yost and Enke, 1979). Such a three-dimensional data set could be obtained by a series of daughter ion scans of each parent ion formed in the ion source, or by a series of parent ion scans of each daughter ion. It could even be obtained by a series of neutral loss scans for every possible neutral loss (i.e., by interrogating the surface in a series of lines parallel to the front line of the plot). Finally, a series of SRM experiments could measure the intensity of every possible parent ion–daughter ion pair (i.e., every point on the surface).

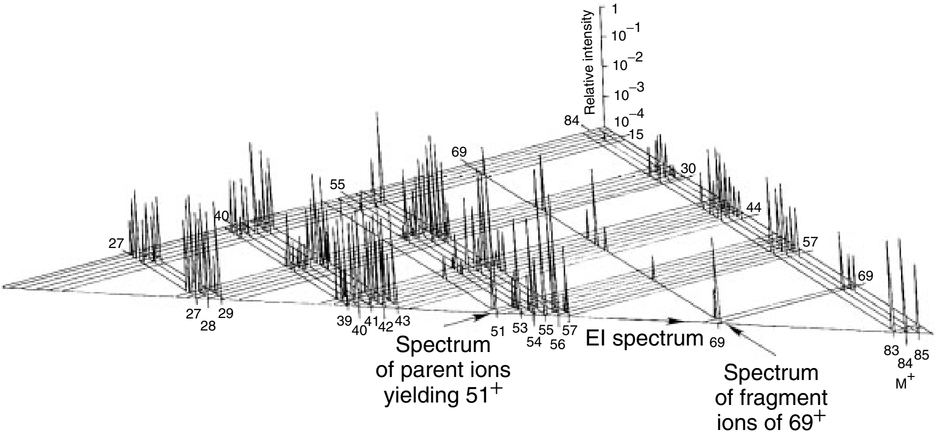


Figure 5 Plot of the three-dimensional dataset generated by electron ionization/MS/MS of cyclohexane. A similar plot of 1-propanol could be found in Fetterolf and Yost (1983)

Finally, note that all of these MS and MS/MS scan modes are readily achieved on a TQMS, and indeed, on most tandem-in-space MS/MS instruments. Note that this is not so on tandem-in-time MS/MS instruments, where the selection of a parent ion precedes (temporally, not spatially) the mass selection of the daughter ion. On tandem-in-time MS/MS instruments such as ion traps, parent scans and neutral loss scans are not available scan modes.

4. Conclusions

The quadrupole mass analyzer (and the RF-only quadrupole and multipole collision cells) enjoy a central role in the success of mass spectrometric methods in proteomics. Development of an understanding of the theoretical and practical aspects of quadrupoles can improve ones ability to utilize single and triple quadrupole mass spectrometers effectively to solve important problems in proteomics.

References

Busch K, Glish G and McLuckey S (1998) *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry*, VCH Publishers.

Cech N and Enke CG (2001) Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, **20**, 362–387.

Cole R (1997) *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation and Applications*, John Wiley & Sons.

Constantopoulos T, Jackson G and Enke C (1999) Effects of salt concentration on analyte response using electrospray ionization mass spectrometry. *Journal of the American Society of Mass Spectrometry*, **10**, 625–634.

Dawson P (1976) *Quadrupole Mass Spectrometry and its Applications*, Elsevier.

Dawson P and Douglas D (1999) Use of quadrupoles in mass spectrometry. *Encyclopedia of Spectroscopy and Spectrometry*, Academic Press.

Fetterolf D and Yost R (1983) Tandem mass spectrometry (MS/MS) instrumentation. *Mass Spectrometry Reviews*, **2**, 1–45.

Fetterolf D and Yost R (1984) Added resolution elements for greater informing power in tandem mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes*, **62**, 33–49.

Kebarle P (2000) A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of Mass Spectrometry*, **35**, 804–817.

King R, Bonfiglio R, Fernandez-Metzler C, Miller-Stein C and Olah T (2000) Mechanistic investigation of ionization suppression in electrospray ionization. *Journal of the American Society of Mass Spectrometry*, **11**, 942–950.

March R and Hughes R (1988) *Quadrupole Storage Mass Spectrometry*, John Wiley & Sons.

McCloskey J (1990) *Methods in Enzymology: Volume 193, Mass Spectrometry*, Academic Press.

McLafferty F (1983) *Tandem Mass Spectrometry*, John Wiley & Sons.

Niessen W (1999) Applications of hyphenated techniques in mass spectrometry. *Encyclopedia of Spectroscopy and Spectrometry*, Academic Press.

Pedder R (2001) Practicle Quadrupole Theory: Graphical Theory. *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, Chicago, IL.

Pedder R (2002) Practicle Quadrupole Theory: Peak Shapes at Various Ion Energies. *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL.

Watson JT (1997) *Introduction to Mass Spectrometry*, *Third Edition*, Lippincott-Raven Publishers.

Yost RA and Enke CG (1979) Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Analytical Chemistry*, **51**, 1251A–1264A.

Short Specialist Review

Quadrupole ion traps and a new era of evolution

Jae C. Schwartz
Thermo Electron Corporation, San Jose, CA, USA

1. Introduction

For the first 30 years of their existence, three-dimensional (3D) quadrupole ion traps remained a specialty tool for a relatively few physicists and even fewer chemists (March and Hughes, 1989). These devices were ideal to study the fundamental properties of ions since the ions remained trapped in space for comparatively long amounts of time. Then in 1983, 30 years after their invention in 1953 (Paul and Steinwedel, 1960), a practical way of obtaining a mass spectrum of the ions contained in the device was discovered (Stafford *et al*., 1984). This subsequently led to the first commercial quadrupole ion trap mass spectrometer in 1984 (Syka, 1995) and put this technology in the hands of hundreds of analytical chemists. The early versions of this mass spectrometer utilized electron impact and chemical ionization, which occurred internally to the ion trap and allowed the analysis of gas phase samples including those eluting from gas chromatographs. In the early 1990s, with the advent of atmospheric pressure ionization techniques along with the ability to successfully inject ions into the trap from an external ion source, quadrupole ion trap mass spectrometers were developed for analytes in the condensed phase (Schwartz and Jardine, 1996; Bier and Schwartz, 1997) often coupled to liquid chromatographs (LC). These LCMS-compatible quadrupole ion traps have now emerged as powerful tools for the qualitative analysis of biological samples (Wagner *et al*., 2003), and so, now find widespread use in all fields of chemistry including biochemistry. The basis for the success of ion traps is due to their fundamental characteristics of performing rapid and efficient MS^n, along with their high sensitivity for producing full scan mass spectra. These two qualities, for example, result in producing sequence information on low amounts (attomoles) of a biological sample that can be processed to automatically identify the peptide or protein (Eng *et al*., 1994). Even with the current success of ion traps, a significant evolution in ion trap technology is currently taking place that promises to provide a significant leap in performance and therefore in the capabilities for biological and nonbiological applications.

2. Fundamental principles of operation

The physical structure of the 3D quadrupole ion trap is shown in Figure 1(a), and consists of three electrodes; a donut-shaped electrode called the "ring electrode" and two identical electrodes called "end-cap" electrodes, which have a single hole for ions to enter and exit the device. The surfaces of all three electrodes have a hyperbolic profile, which is required in order to generate the desired quadrupole electric field. Now, basic physics (Laplace equation of electrostatics) says that it is impossible to create a static electric field that would trap ions in all three dimensions. Consequently, an alternating electric field is used by applying a radio frequency (RF) voltage to the ring electrode. This voltage creates a RF trapping field in all three dimensions. The average effect of the RF trapping field continuously pushes ions to the center of the device. With the assistance of helium gas introduced into the trap, ions occupy a roughly spherical volume in the center of the device. Collisions with helium reduce the kinetic energy of the ions reducing the cloud size, resulting in increased sensitivity and resolution of the resultant mass spectrum. The helium also serves as a collision gas for dissociation of ions when their kinetic energy is purposely increased. As a simplification, trapped ions can be thought to have nearly simple sinusoidal motion while stored, with frequencies of this motion being inversely related to their m/z. Once a population of ions is trapped, the method that produces their mass spectrum consists of scanning the amplitude of the RF voltage over a certain range linearly with time while applying an additional resonance ejection voltage across the end caps (Syka and Louris, 1988). Ions are ejected as the ion frequency of motion matches the applied resonance signal because of the RF amplitude being changed. This results in ions being ejected from the trap through the end-cap apertures in order of their m/z ratio and are detected. Since there is a direct relationship of the m/z of an ion and the RF amplitude at which it will be ejected, acquiring the signal response while the RF amplitude is ramped produces a mass spectrum.

With the understanding of these basic principles, a typical operational sequence can be considered. First, the ions are formed within or are introduced into the trap.

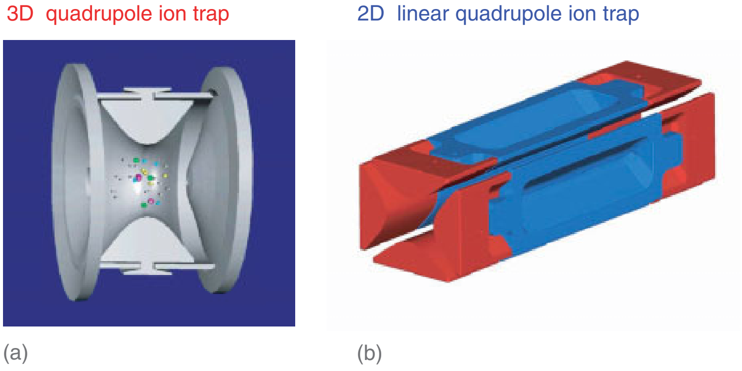


Figure 1 (a) Three-dimensional ion trap structure with ring electrode and two end-cap electrodes. (b) Two-dimensional linear ion trap with four parallel rods cut into three sections

Most modern traps inject ions from an external ion source. This process is somewhat challenging since the ions must penetrate the RF field as they pass through the injection end-cap aperture. As a result, only some fraction of the entering ions is trapped. To compensate, however, simply increasing the accumulation time can assure that the ion trap is full to its capacity. This ability to accumulate ions is a unique feature of trapping-type instruments and helps increase sensitivity and lowers the limits of detection. After a population of ions is trapped, they can then be manipulated in many ways. One example is to isolate an ion of a particular m/z (or range of m/z values) while all others are ejected from the trap. Isolation can be accomplished by many different methods, a common one consisting of applying a multifrequency resonance ejection waveform (Louris and Taylor, 1993). The resulting isolated ions can then be dissociated into fragment ions using a resonance excitation process, which energetically collides the ions with helium atoms (Syka and Louris, 1988). The sequence of isolation and activation of ions can be performed an arbitrary number of times, hence, MS^n, and is another distinctive feature of ion-trapping instruments. When a spectrum of the ions contained in the trap is desired, the RF amplitude is ramped along with the resonance ejection voltage and the ions are ejected to the detector, as was described above. The resultant MS/MS or MS^n spectrum provides characteristic structural information on the isolated precursor ions.

It is inherent in this operational sequence that ion trap mass analyzers have a high duty cycle, and therefore high sensitivity, even for full scan mass spectra. This is due to the fact that the entire spectrum is obtained on the same population of ions whether performing full scan MS or MS^n. In most commercial ion trap systems, the ion accumulation times are automatically controlled to optimize the number of ions analyzed and to minimize space charge effects. As a consequence, as the sample level decreases, the duty cycle actually increases. This is in contrast to beam type mass analyzers, such as quadrupole mass filters that have high duty cycles only when monitoring a single selected ion mass but have very low duty cycles for full scan modes since each m/z scanned requires new ions from the ion source.

3. A new evolutionary era: 2D linear traps

Although 3D quadrupole ion trap mass spectrometers have been extremely successful, their performance has some fundamental limitations. Space charge effects are the Achilles heel of all ion-trapping instruments. By definition, 3D ion traps have RF electric fields containing the ions in all three dimensions. As a consequence, only so much total ion charge can be contained in the ion cloud before space charge effects degrade the quality of the mass spectrum, including mass accuracy and resolution. Thus, one inadequacy in 3D ion traps is that the charge capacity is quite limited ($\sim$1000 charges). A second fundamental limitation is the trapping efficiency of injected ions. As was mentioned above, injected ions must penetrate the RF electric field and, therefore, ions are successfully trapped only during a narrow phase window of the oscillating RF voltage. These limitations argue the need to remove the RF electric field in one of the dimensions of the device, resulting in a 2D or linear ion trap. The removal of the RF field in one

dimension allows the ion cloud to expand from the approximately spherical shape to more of a cylindrical shape of some arbitrary length, resulting in increased charge capacity. In addition, since ions can be injected on the axis without the RF field, the trapping efficiency is significantly increased. A picture of the 2D device is shown in Figure 1(b) and can be contrasted to the 3D trap. This device consists of four parallel rods cut into three sections and has slots in two of the center section rods for the ions to exit. The ions are trapped radially by an RF containment field but are trapped axially by a static electric field. The results of such a device even with a moderate length of less than 2.5 in gives improvements of trapping efficiencies of more than 10-fold and increased storage capacity of more than 20-fold (Schwartz *et al*., 2002). These increases directly translate to increased sensitivity, lower limits of detection, increased dynamic range, and enhanced MS^n performance.

An additional advantage of the 2D linear trap involves higher detection efficiencies. In a 3D trap, since ions are injected through one end cap, but are typically ejected through both end caps equally, normally only one detector can be used. Consequently, not all trapped ions are detected. Since the ions are injected into the 2D trap orthogonal to the axis of detection, two detectors can be utilized to detect twice as many ions as in the 3D trap. This increases the number of detected ions by an additional factor of two and, therefore, a total of 40 times more ions can be detected in each scan.

The resultant performance increase of the 2D linear trap can be demonstrated in a simple example, namely intact proteins. Intact proteins are challenging for ion traps since each protein ion upon electrospray ionization (ESI) will contain many charges and, therefore, will severely limit the actual number of ions that can be analyzed before space charge effects occur. Figure 2(a) shows the ESI mass spectrum of the protein myoglobin having an average molecular weight of 16 950 a.m.u., obtained on a 3D ion trap. A single scan is shown, which includes the various charge states of the protein molecules. In order to identify the molecular weight of the protein, the spectrum containing the different charge states can be deconvoluted to convert the m/z spectrum into a true mass spectrum, shown in the lower spectrum of Figure 2(a). Owing to the limited charge capacity, and that each molecule has an average of 16 charges, it is difficult to identify the molecular weight of the protein because of the low signal-to-noise ratio of the spectrum. Scans can be averaged to increase the quality of the spectrum; however, this takes substantially more time. This is in contrast to the single spectrum obtained on the 2D linear trap shown in Figure 2(b). Owing to the higher trapping efficiency, higher charge capacity, and higher detection efficiency, this single spectrum is comparable to an average of more than 20 scans using a 3D trap. The molecular weight of the protein is clearly identified in the lower deconvoluted mass spectrum.

4. Data-dependent analyses: maximizing the yield of information

Switching from MS to MS/MS or to MS^n has always been relatively simple and fast on ion traps compared to other instruments. Consequently, development of

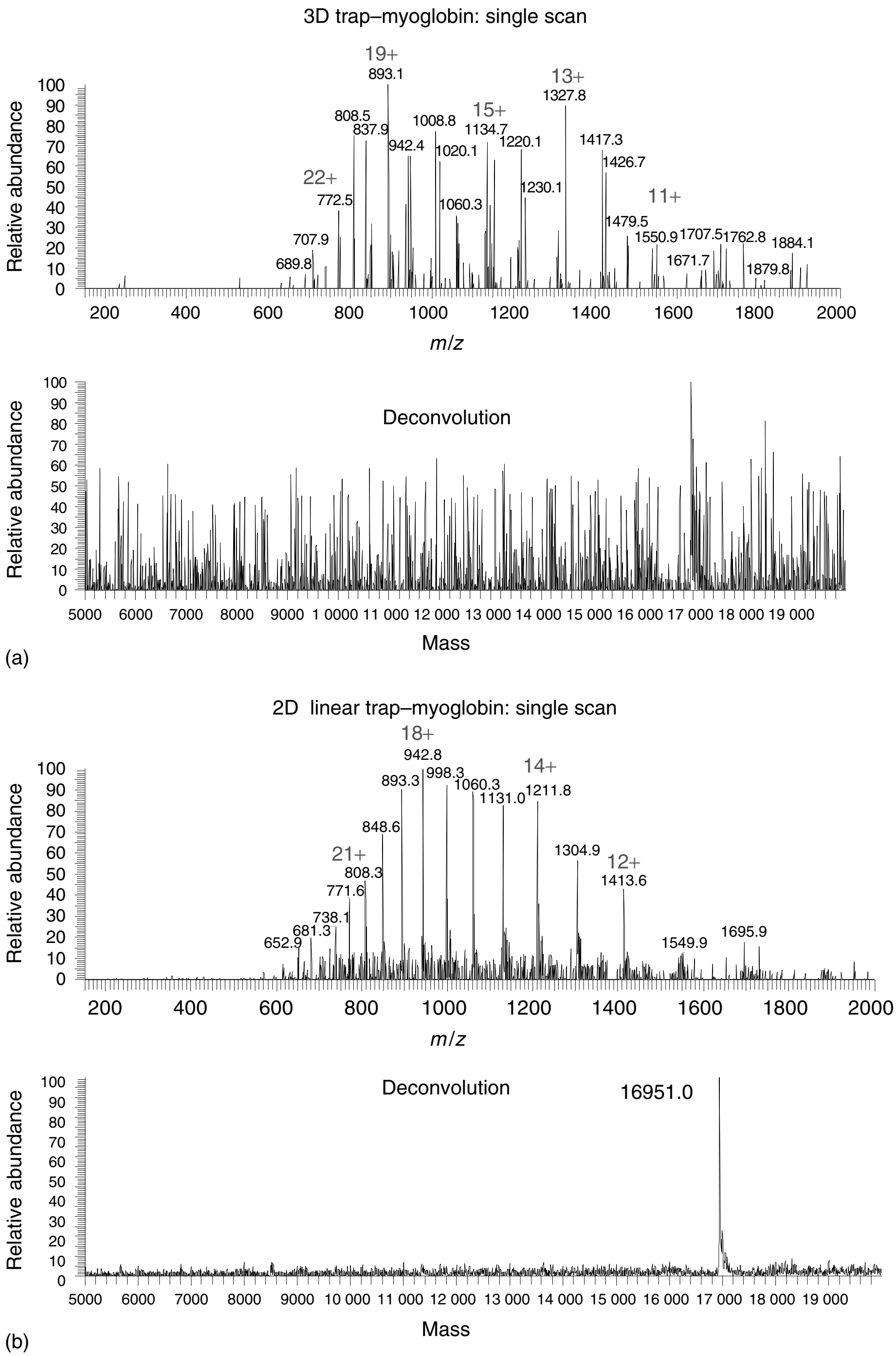


Figure 2 (a) Electrospray ionization mass spectrum of intact myoglobin using a 3D ion trap. (b) Electrospray ionization mass spectrum of intact myoglobin using a 2D linear ion trap

automatic and data-dependent scanning methods has been carried out on these systems and has emerged as "the" technique for yielding maximum information from complex samples (Wenner and Lynn, 2004). A typical method is one where the system first takes a full scan mass spectrum. It then determines the three most intense components in the spectrum and automatically performs MS/MS on each precursor ion, producing structural information for the three selected ions. On the next MS scan, the system can ignore ions already analyzed by MS/MS and search for three new ions in the mass spectrum to perform MS/MS or MS^n on. Many different criteria can be specified to trigger the acquiring of MS/MS spectra besides simply intensity. As just a few examples, the system could require a certain intensity ratio of two ions that exist in the MS scan for use with labeled or tagged analytes, and/or the ions must have a certain charge state, or must be below or above certain intensity thresholds. In one method, the system looks for the loss of phosphoric acid from phosphorylated peptides in the MS/MS spectra. If the loss occurred, then MS^3 is performed on this new ion. This method not only identifies the peptide as a phosphopeptide but can also determine the site of phosphorylation (Zumwalt *et al*., 2003). These data-dependent type of methods can generate a tremendous amount of information. Currently, for peptide and protein identification, the data is processed using two basic approaches. One approach makes use of databases of known peptides and proteins to which the data is compared to in order to produce a likely identification (Mann *et al*., 2001), while a de novo approach attempts to identify the sequence of a sample that may not be known (Zhang and McElvain, 2000). The sensitivity and usefulness of these methods on ion traps is demonstrated in Figure 3(a) and (b). Just 40 attomoles of a myoglobin tryptic digest was injected and separated on a 170-um packed capillary LC column. The mass chromatogram for the peptide at m/z 690 is shown in Figure 3(a) along with the MS spectrum indicating the two charge states of the peptide ion. The MS/MS spectrum of the doubly charged peptide ion obtained automatically by the data-dependent method is shown in Figure 3(b), and yields nearly the complete sequence of the peptide, allowing automatic identification of the peptide and the protein myoglobin using data base searching tools.

Although not fundamental to the technology, other enhanced capabilities have been implemented on the newer 2D linear trap technology (Schwartz *et al*., 2002). The 2D linear trap has the ability to obtain resolutions greater than 30 000 by slowing down the rate of scanning out the ions from the trap. This resolution is sufficient to help unambiguously determine the charge states of ions by the spacing of their isotope peaks in the mass spectrum and also to separate isobaric ions. The overall higher resolutions obtainable can also be traded off for increased scan rates to further reduce analysis time and therefore increase the number of experiments performed in a given time. Another major advantage of this technology, which is certain to promote further developments, is that the mechanical geometry has the advantage of having the back end of the device free from detectors, making this device ideal for coupling to other devices including other mass analyzers in hybrid configurations (*see* Article 10, **Hybrid MS**, Volume 5). One form of linear trap is incorporated as the third quadrupole in a triple quadrupole instrument (Hager, 2002). Another extremely powerful configuration now available combines a 2D linear trap with a Fourier transform mass spectrometer (Horning, *et al*., 2003). This hybrid instrument

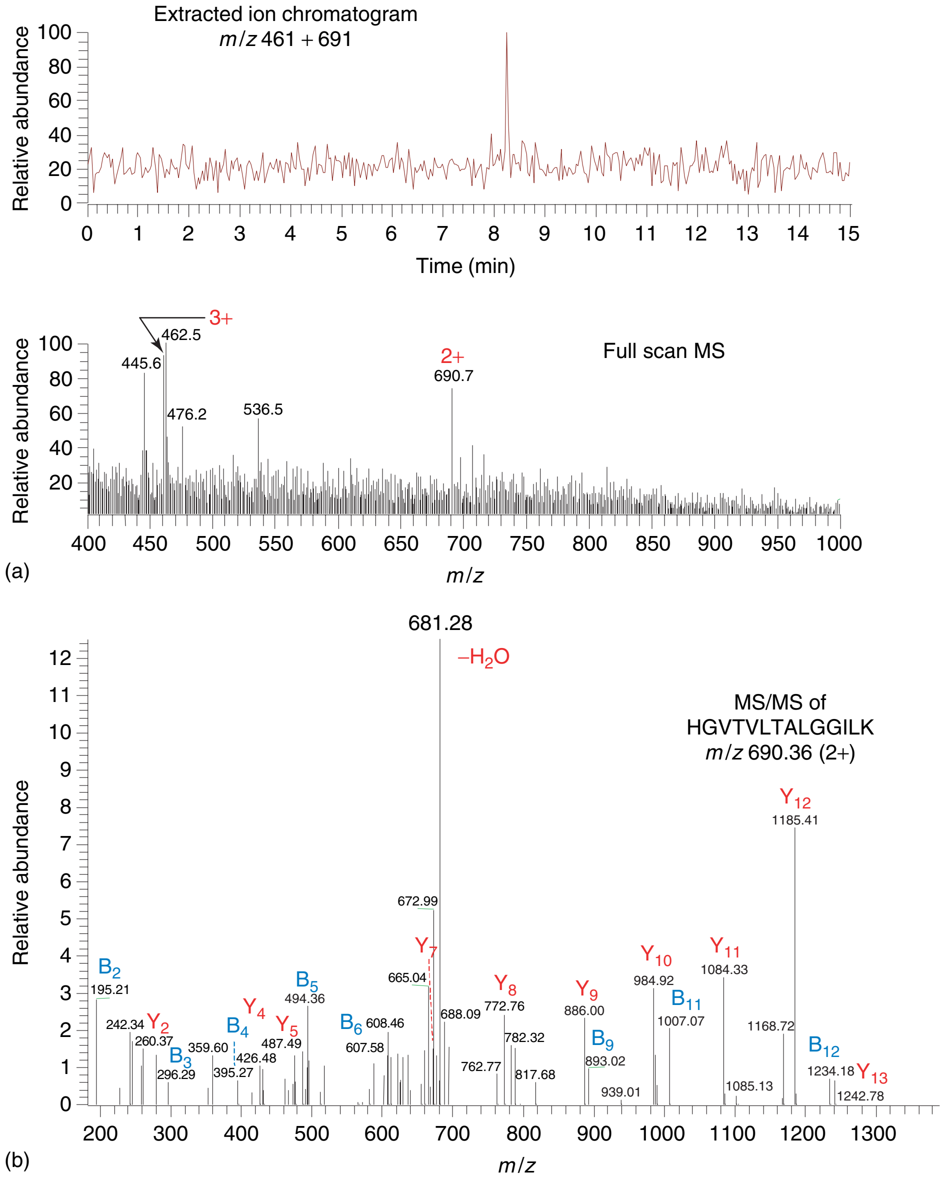


Figure 3 (a) Extracted ion chromatogram showing elution of one peptide from 40 attomoles of myoglobin digest and the mass spectrum at that retention time. (b) The MS/MS spectrum of the peptide showing virtually complete sequence information and allowing automatic peptide and protein identification

not only retains all the functionality of the 2D linear trap but can also give 1-ppm mass accuracies and >100 000 resolution scanning at 1 scan/s. Other hybrids or combinations of devices including ionization techniques can easily be imagined, opening the doors to a vast amount of future capabilities to even further enhance ion traps' usefulness for real-world complex mixture component analysis.

References

Bier ME and Schwartz JC (1997) *Electrospray Ionization*, Richard C (Ed.), John Wiley & Sons: New York, pp. 235–289.

Eng, JK, McCormiack AL and Yates JR III (1994) *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.

March RE and Hughes RJ (1989) *Quadrupole Storage Mass Spectrometry*, Wiley Interscience: New York.

Mann M, Hendrickson RC and Pandey A (2001) *Annual Review of Biochemistry*, **70**, 437–473.

Hager JW (2002) *Rapid Communications in Mass Spectrometry*, **16**, 512–526.

Horning S, Malek R, Wieghaus A, Senko W and Syka JEP (2003) A Hybrid two-Dimensional Quadrupole Ion Trap/Fourier Transform Ion Cyclotron Mass Spectrometer: Accurate Mass and High Resolution at a Chromatography Timescale, *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics*, Montreal, Quebec, Canada, June 8–12.

Louris JN and Taylor DM (1993) *Method and Apparatus for Ejecting Unwanted Ions in an Ion Trap Mass Spectrometer*, US Patent 5,324,939.

Paul W and Steinwedel H (1956) *Apparatus for Separating Charged Particles of Specific Different Charges*, German Patent 944,900; (1960) US Patent 2,939,952

Schwartz JC and Jardine I (1996) *Methods in Enzymology*, Vol. 270, Karger BL, Hancock WS (Eds.), Academic Press Inc.: San Diego, CA, 552–586.

Schwartz JC, Senko MW and Syka, JEP (2002) *Journal of the American Society for Mass Spectrometry*, **13**, 659–669.

Stafford GC, Kelley PE, Syka JEP, Reynolds WE and Todd JFJ (1984) *International Journal of Mass Spectrometry and Ion Processes*, **60**, 85–98.

Syka JEP and Louris JN (1988) US Patent 4,736,101.

Syka JEP (1995) *Practical Aspects of Ion Trap Mass Spectrometry, Volume 1 Fundamentals of Ion Trap Mass Spectrometry*, Vol. 1, March RE and Todd JFJ, (Eds.), CRC Press: Boca Raton, FL, pp. 169–205.

Wagner Y, Sickmann A, Meyer HE and Daum G (2003) *Journal of the American Society for Mass Spectrometry*, **14**, 1003–1011.

Wenner BR and Lynn BC (2004) *Journal of the American Society for Mass Spectrometry*, **15**, 150–157.

Zhang Z and McElvain JS (2000) *Analytical Chemistry*, **72**, 2337–2350.

Zumwalt AC, Choudhary G, Cho D, Hemmenway E and Myulchereest I (2003) Detection of Phosphorylated peptides from Data Dependent MS32 Neutal-Loss Scans using a Linear Ion Trap Mass Spectrometer, *Proceedings of the 51st ASMS Conference on Mass Spectrometery and Allied Topics*, Montreal, Quebec, Canada, June 8–12.

Short Specialist Review

Hybrid MS

R. H. Bateman and J. I. Langridge
Micromass UK Ltd, Manchester, UK

1. Hybrid MS instruments

Tandem mass spectrometry, or MS/MS, has become the preferred technology for many applications in which mass spectrometry plays a part. First, tandem mass spectrometry allows selection and isolation of specific compounds of interest and their subsequent identification. Second, the extra selectivity of MS/MS enables this technology to be used for quantification of target compounds, even in the presence of complex matrices.

There are several different types of mass analyzers that are commercially available. These different mass analyzers have different characteristics that make them more or less suitable for certain applications. Hybrid MS instruments consist of two mass analyzers in series and a means for fragmenting selected ions, usually positioned after the first mass analyzer and before the second. The means for fragmenting ions is most commonly that of promoting collision-induced decomposition (CID) in which ions undergo collisions with gas molecules in a partially enclosed collision cell. In most hybrid MS instrumentation, the processes of mass selection, fragmentation, and product ion mass analysis take place sequentially in space. The first mass analyzer (MS1) may be used to select ions of a target compound, the CID gas collision cell to fragment those ions, and the second mass analyzer (MS2) to analyze the products of the fragmentation process. In some instances, fragmentation can take place in one or other of the mass analyzers. Here one or more of the processes of mass selection, fragmentation, and product ion mass analysis can take place sequentially in time.

Different combinations of mass analyzers also have different characteristics that make them more or less suited for certain applications. It is possible to combine most types of mass analyzers in almost any sequence, and numerous combinations have been constructed. Table 1 lists the available types of mass analyzers and those combinations that are most commonly used for MS/MS applications.

The following is a brief description of those hybrid instruments that are most commonly used for studies in the life sciences, and their principal areas of application.

Table 1

MS1	MS2					
	Magnetic sector	Quadrupole filter	Quadrupole trap	Axial TOF	Orthogonal TOF	FT-MS
Magnetic sector	✓	✓	✓	✓	✓	
Quadrupole filter		✓✓	✓✓		✓✓	✓✓
Quadrupole trap				✓✓	✓✓	✓✓
Axial TOF				✓✓		
Orthogonal TOF						
FT-MS						

✓Hybrid combinations that are or have been available, but are not commonly used in the life sciences.
✓✓Hybrid combinations that are available and are most commonly used for studies in the life sciences.

2. Triple quadrupole (Q-CID-Q)

In the triple quadrupole tandem mass spectrometer, MS1 is a quadrupole mass filter (*see* Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5) followed by a gas collision cell and then by MS2, which is again a quadrupole mass filter. The name "triple quadrupole" is derived from the first such instrument in which an RF quadrupole was used to guide ions through the gas collision cell. The quadrupole mass filter is typically used to transmit ions of a single m/z value. Therefore, to record a full mass spectrum, the quadrupole must be scanned across the full m/z range to sequentially transmit ions of different m/z values. The duty cycle for this process is quite low, and as a consequence, the sensitivity of a quadrupole mass filter used to record a full spectrum is relatively poor. On the other hand, the quadrupole mass filter will have 100% duty cycle when used to transmit ions of a single m/z value. The triple quadrupole used for selected reaction monitoring (SRM), in which it is used to transmit ions of a single precursor in MS1 and ions of a single product in MS2, is very specific and exceptionally sensitive. Therefore, triple quadrupoles have found significant use in the drug development process where they are used in the SRM or MRM (multiple reaction monitoring) modes to quantify target compounds of biological significance.

A common application in peptide and protein analysis for the triple quadrupole makes use of the "precursor ion scanning" mode. In this mode, MS2 is set to transmit a specific characteristic product ion, while MS1 is scanned to sequentially transmit precursor ions into the gas cell for fragmentation. When the specific product ion is detected, then the precursor ion mass transmitted by the quadrupole is recorded. This approach has proved particularly useful in the analysis of protein post-translational modifications, such as phosphorylation and glycosylation.

3. Quadrupole–quadrupole trap (Q-CID-Q/Trap)

A derivation from the triple quadrupole tandem mass spectrometer is the arrangement in which the MS2 quadrupole mass filter is substituted with a linear

quadrupole ion trap (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5). In addition to the normal attributes of the triple quadrupole, the MS2 linear quadrupole ion trap provides some greater flexibility for qualitative applications. This arrangement allows trapping of precursor or fragment ions in the MS2 linear quadrupole ion trap with subsequent mass selective axial ejection, thus improving full scan MS or MS/MS sensitivity over the triple quadrupole mass spectrometer. The linear quadrupole ion trap can also be used to enhance the relative abundance of certain fragment ions and the abundance of multiply charged ions relative to that of singly charged ions.

4. Quadrupole trap–axial time-of-flight (Trap/CID-TOF)

In this tandem MS/MS arrangement, MS1 is a conventional 3D quadrupole ion trap (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) and MS2 is an axial time-of-flight (TOF) mass spectrometer (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5). The quadrupole ion trap is also used as the collision cell for fragmentation of precursor ions selected by the ion trap. The TOF mass spectrometer incorporates a reflectron to improve mass resolution and mass accuracy. This combination has been primarily coupled with MALDI (*see* Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5). The insertion of the quadrupole ion trap between the MALDI source and TOF mass spectrometer decouples these elements so that mass resolution and mass calibration is no longer dependent on variations in the MALDI source conditions. This arrangement allows MS^n experiments to be carried out in the quadrupole ion trap, with the fragments ejected and mass measured using the axial TOF. This provides MS^n capabilities on an instrument that can obtain mass measurement accuracy of better than 10 ppm. The MS^n and accurate mass capabilities of this instrumentation allow its use for peptide and protein identification.

5. Time-of-flight–time-of-flight (TOF-CID-TOF)

In this instrument, a MALDI ion source is coupled with an axial time-of flight mass analyzer (MS1), and can be used to select precursor ions for fragmentation in a gas cell. Fragmentation may also occur by post source decay (PSD). Mass analysis of the fragment ions produced is then accomplished by further acceleration of product ions into a second axial time-of-flight mass analyzer (MS2). The MS2 mass analyzer incorporates a reflectron (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5). This and the use of appropriately timed acceleration fields in one or both TOF mass analyzers provide improved resolution and mass measurement accuracy on the fragment ions than may be obtained in a traditional PSD experiment on a conventional TOF instrument. These instruments are targeted at "high-throughput" protein identification studies.

6. Quadrupole–orthogonal time-of-flight (Q-CID-TOF)

This hybrid MS/MS instrument includes a quadrupole mass filter (MS1), followed by a gas collision cell for fragmentation of selected ions, and finally, an orthogonal

acceleration time-of-flight (TOF) mass spectrometer (MS2) for recording mass spectra. The orthogonal TOF mass spectrometer (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) has a number of advantages over a scanning quadrupole mass filter. The orthogonal TOF will repetitively extract a portion of a continuous ion beam and will record the full mass spectrum. The speed of acquisition, sensitivity, mass resolution (>15 000 FWHM), and mass accuracy (<5 ppm RMS) are all considerably higher than for a scanning quadrupole mass filter. As a consequence, it is normal practice to use the TOF mass spectrometer for acquisition of both precursor ion and product ion mass spectra. Precursor ion spectra are acquired by operating the quadrupole in the RF (nonresolving) mode so that the total ion flux is transported to the orthogonal TOF for mass analysis. Switching to MS/MS only requires switching the quadrupole into the resolving mode at the required precursor ion mass and raising the collision energy appropriate to that for the selected mass and charge of the precursor ion. This hybrid combination does not offer the sensitivity to match that of a triple quadrupole in the SRM or MRM modes of operation. On the other hand, it provides excellent sensitivity, speed, and selectivity that make it particularly well suited for qualitative applications. It is most commonly used for the identification of unknowns from the exact mass measurement of the precursor ion and its product ions. This hybrid configuration is one of the preferred arrangements not only for the identification of peptides and proteins but also for the identification of metabolites and other small molecules of biological significance.

7. Quadrupole trap–orthogonal time-of-flight (Q-CID/trap-TOF)

This hybrid MS/MS configuration is similar to the previously described quadrupole–orthogonal TOF combination other than the linear quadrupole in the gas collision cell is additionally used to trap ions. In addition to the attributes of the quadrupole–orthogonal TOF combination, this arrangement has the capability to increase the duty cycle of the orthogonal TOF over a narrow mass range, albeit at the cost of reduced duty cycle outside that mass range. This is achieved by temporarily trapping ions in the quadrupole gas cell, releasing them, and pulsing the orthogonal TOF accelerating field such as to synchronize with the arrival of ions of interest. This increases the duty cycle, and hence sensitivity, for those ions. This can be advantageous for certain applications in which, for example, a specific daughter ion species is of interest. As with the "precursor ion scanning technique" on a triple quadrupole, this approach can be useful in the analysis of protein post-translational modifications, such as phosphorylation and glycosylation.

8. Quadrupole–Fourier transform ICR-MS (Q-CID-FTMS)

This hybrid MS/MS instrument consists of a quadrupole mass filter as MS1, followed by a gas collision cell for fragmentation of selected ions, and finally, a

Fourier Transform Ion Cyclotron Resonance (FT-ICR) or Fourier Transform Mass Spectrometer (FTMS) as MS2 for recording of mass spectra (*see* Article 5, **FT-ICR**, Volume 5). This has similar characteristics to that of the quadrupole–orthogonal TOF combination, other than that the TOF is replaced with an FTMS. Modern FTMS systems, employing superconducting magnets with field strengths in the region of 7 to 12 Tesla, are capable of very high resolution (>100 000 FWHM) and mass accuracy (<2 ppm RMS). As with the quadrupole–orthogonal TOF combination, the FTMS can be used for recording both precursor ion and product ion mass spectra.

9. Quadrupole trap–Fourier transform ICR-MS (Trap/CID-FTMS)

This hybrid MS/MS combination incorporates a linear quadrupole ion trap (MS1), which also functions as a gas collision cell for fragmentation of selected ions, and an FTMS (MS2) for recording of mass spectra. This configuration has very similar characteristics and advantages to that of the quadrupole mass filter–FTMS combination. However, the linear quadrupole ion trap also allows mass selective radial ejection to separate ion detectors. This offers a number of additional advantages.

First, the linear quadrupole ion trap may be used for recording product ion spectra with higher sensitivity, albeit at reduced mass resolution and mass accuracy.

Second, the inherent dynamic range of the FT-ICR trap is quite limited; it requires an optimum number of ions for optimum mass accuracy. The linear quadrupole ion trap can be used to accumulate approximately the required number of ions before transferring them to the FT-ICR cell. This technique effectively extends the dynamic range for this type of hybrid MS/MS mass spectrometer.

Third, the measurement time required by the FTMS can be quite long if the highest mass resolution is required. The linear quadrupole ion trap may be used to acquire product ion spectra while simultaneously acquiring the highest mass resolution precursor ion spectrum with the FTMS.

As for the conventional 3D quadrupole ion trap, daughter ions with m/z values less than about 25 ∼ 33 % of the precursor ion m/z value are unstable in the linear quadrupole ion trap, and are not recorded. However, the linear quadrupole ion trap can be programmed to acquire MS^n spectra.

Short Specialist Review

Nano-MALDI and Nano-ESI MS

Gyorgy Marko-Varga, Johan Nilsson and Thomas Laurell
Lund University, Lund, Sweden

1. Introduction

In line with the rapid development of miniaturized and chip-integrated approaches to analytical and bioanalytical systems, the proteomics work scheme has also been the focus of dimensional reduction. The driving arguments for miniaturization are seen in the possibilities of improved sample throughput, reduced sample volume requirements, lower reagent costs, and improved limits of detection (Reyes *et al*., 2002; Auroux *et al*., 2002). In recent years, the analytical endpoint in proteomics has to a great extent become dominated by mass spectrometry and database matching. In this time, laser desorption ionization (MALDI) and electrospray ionization (ESI) have become the two major techniques to ionize the target species prior to time-of-flight separation based on mass and charge (James, 2000). The aspects on miniaturization in this field have, to a great extent, taken the course of optimizing the sample handling and processing steps prior to the MS readout (Marko-Varga *et al*., 2004).

2. Miniaturization in MALDI-TOF MS

An inherent bottleneck in MALDI-TOF MS has been the transfer of the proteomic sample from, for example, a microtiter plate format to the mass spectrometry target plate. The MS instrumentation was, and still most are, not equipped with an MS target plate loader that can handle the dimensions of a 96-well format. As the industry sample-handling robotics follows the dimensions of the 96-well standard format, a sample transfer and format change from the well plate to the MALDI-target is a request. Recent developments have demonstrated that miniaturized protocols to deposit proteomic samples may provide improved analytical performance in the proteomic work scheme such as increased sample throughput (Ekström *et al*., 2000; Little *et al*., 1997), continuous or semicontinuous MALDI-readout from micro liquid chromatography separations (Preisler *et al*., 1998; Preisler *et al*., 2002; Miliotis *et al*., 2000; Wall *et al*., 2002).

A general strategy employed in the protocols for sample handling on the MALDI target has followed the findings reporting that reduced sample spot size provides

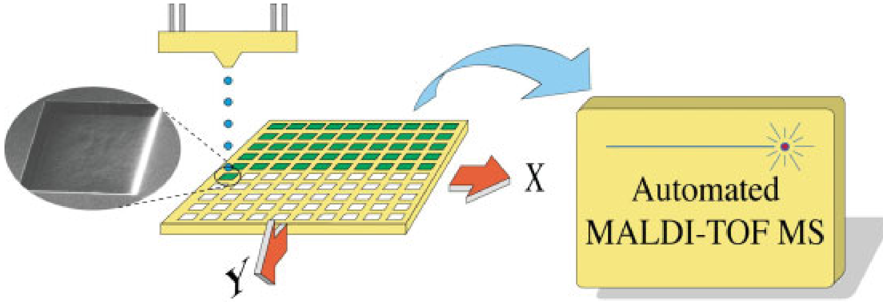


Figure 1 Principal setup for nanovial sample deposition prior to MALDI-TOF MS readout

increased sensitivity in the MS readout (Ekström *et al*., 2000; Gobom *et al*., 1999; Önnerfjord *et al*., 1998). This has also been realized in commercial systems such as the MicroMass, MassPrep^{TM} MALDI-Target, and the so-called anchor point MALDI targets by Bruker Daltonics GmbH, Germany (Schuerenberg *et al*., 2000). Work on further miniaturization and controlling the confinement of the sample to a predefined spot was demonstrated by Ekström *et al*. (2001a) using in-house developed piezoelectric dispensing for controlled deposition of the sample and also silicon microfabricated MALDI target plates to confine the deposited fluid by means of surface tension forces in the so-called on-spot enrichment. Later, this was also demonstrated on disposable polymer nanovial target plates (Marko-Varga *et al*., 2001; Ekström *et al*., 2001b). Figure 1 shows the principal setup for piezoelectric sample deposition and confined sample deposition in nanovial target plates. More recently, a stand-alone system for automatic sample transfer from chip-integrated solid-phase extraction microarrays has been realized (Wallman *et al*., 2003). The sample is transferred from the solid-phase microextraction chip to the microdispenser solely by means of capillary forces (i.e., no external pumping is needed to drive the sample flow through the microchips for subsequent piezoelectric deposition onto a MALDI target plate). The chip-based sample handling protocol provides both sample enrichment and cleanup performed in two steps; (1) microchip solid-phase enrichment and (2) on-spot enrichment, that is, a dual-enrichment mode is obtained that offers a sensitivity increase, superseding conventional MALDI sample preparation protocols (Wallman *et al*., 2004).

3. Miniaturization in ESI-MS

Interfacing chip-integrated protocols to electrospray ionization mass spectrometry (ESI-MS) has been a long-sought goal and several promising approaches have emerged lately. A key issue in this development has been the microfabrication of an electrospray tip that provides both the necessary microfluidic and electrostatic features. The early chip/ESI work demonstrated electrospray directly from the end piece of a microchip with an embedded microchannel (Xue *et al*., 1997; Ramsey and Ramsey, 1997; Kameoka *et al*., 2001). As the flat front surface of the chip side does not provide an optimal zone for building up a reproducible Taylor cone to generate a stable electrospray, much effort has been put into the development of

a microtechnology fabrication process for nanoelectrospray tips. The strategy has therefore been to limit the spreading of the effluent at the ESI-point by extending the microchannel beyond the bulk material of the chip. The insertion of a conventional capillary or regular ESI-tip at the chip termini is a common approach (Figeys and Aebersold, 1998; Bings *et al*., 1999), though not very amenable to low-cost mass fabrication.

With the advent of deep reactive ion etching (Clerc *et al*., 1998), silicon microfabrication could offer engineering tools for out-of-plane high-aspect ratio and high-fidelity microstructures, which basically corresponds to the geometries needed for making a good ESI-tip. Advion BioSciences Ltd., Ithaca, NY, USA, currently offers chip-based ESI-tips fabricated as fine cylindrical pipes in an array format using silicon micromachining (Van Pelt *et al*., 2003a,b). This design has proven to be reproducible and sufficiently robust for automated ESI applications. A similar design was later proposed by the group of Stemme, presenting a slight modification of these tips with a reduced front surface area and thereby claiming improved ESI conditions (Griss *et al*., 2002). Silicon microfabricated devices do not offer single-use components owing to the production costs associated with clean room manufacturing and, consequently, much interest has lately been put into the development of disposable polymer ESI-tips. This development also follows the trends of biochip developments where low-cost polymer chips have gained much attention recently (Ng *et al*., 2002). In line with this, Craighead *et al*., proposed a polymer-based chip-integrated ESI interface based on a triangular polymer lip extending from the microfluidic chip, guiding the sample from the exit of the chip to the point of Taylor cone formation (Kameoka *et al*., 2002). This approach seems like a very reasonable and low-cost solution to the problem of controlling the fluid in the high-voltage field at the chip exit of ESI. An alternative ESI interface, equally interesting from a scientific as well as commercial point of view was demonstrated by Killeen *et al*., Agilent Technologies Inc., Paolo Alto Ca, USA, with a complete system fabricated in laminated polyimide sheets that holds both a micro LC column and upstream interface to conventional rotary valves for sample injection and eluent control (Killeen *et al*., 2003). The chip outlet is defined as a fine laser-cut tip at the end of the separation channel, providing a robust and simple ESI interface. It should be noted that this design offers both on- and off-chip interfacing, which commonly is not the case in regard to lab-on-a-chip systems. This means that conventional high-fidelity analytical techniques can be linked upstream to the chip-integrated sample processing.

4. Conclusions

Ongoing research within miniaturization strategies of mass spectrometry sample handling and processing indicates that we currently are in a transition stage where commercial initiatives at an increasing pace now are entering the arena. The field of lab-on-a-chip, micro total analysis systems, BioMEMS, and nanobiotechnology have now reached a level of maturity where the major interdisciplinary bridging of knowledge is at hand. Also, fundamental material science developments have advanced sufficiently to provide the fruitful environment needed to be able to

successfully address the questions raised in a bioanalytical situation utilizing MEMS and nanotechnology principles and methodologies. It can be anticipated that we are facing a prosperous future with rapid and dramatic improvements in sample preparation, and handling for mass spectrometry on the basis of novel miniaturized concepts, offering both improved sample throughput, sensitivity, and reduced sample volume requirements.

Acknowledgments

The authors wish to acknowledge the support from Knut och Alice Wallenberg Foundation, SWEGENE, The Swedish Foundation for Strategic Research, The Swedish Research Council, Crafoord Foundation, and Carl Trygger Foundation.

References

Auroux PA, Iossifidis D, Reyes DR and Manz A (2002) Micro total analysis systems. 2. Analytical standard operations and applications. *Analytical Chemistry*, **74**(12), 2637–2652.

Bings NH, Wang C, Skinner CD, Colyer CL, Thibault P and Harrison DJ (1999) Microfluidic devises connected to fused-silica capillaries with minimal dead volume. *Analytical Chemistry*, **71**(15), 3292–3296.

Clerc PA, Dellmann L, Gretillat F, Gretillat MA, Indermuhle PF, Jeanneret S, Luginbuhl P, Marxer C, Pfeffer TL, Racine GA, *et al*. (1998) Advanced deep reactive ion etching: a versatile tool for microelectromechanical systems. *Journal of Micromechanics and Microengineering*, **8**(4), 272–278.

Ekström S, Ericsson D, Önnerfjord P, Bengtsson M, Nilsson J, Marko-Varga G and Laurell T (2001a) Signal amplification using "Spot-on-a-chip" technology for the identification of proteins via MALDI-TOF MS. *Analytical Chemistry*, **73**, 214–219.

Ekström S, Nilsson J, Helldin G, Laurell T and Marko-Varga G (2001b) Disposable polymeric high density nanovial arrays for MALDI-TOF Mass Spectrometry - A novel concept, Part II Biological Applications. *Electrophoresis*, **22**, 3984–3992.

Ekström S, Önnerfjord P, Nilsson J, Bengtsson M, Laurell T and Marko-Varga G (2000) Integrated microanalytical technology enabling rapid and automated protein identification. *Analytical Chemistry*, **72**, 286–293.

Figeys D and Aebersold R (1998) Nanoflow solvent gradient delivery from a microfabricated device for protein identifications by electrospray ionization mass spectrometry. *Analytical Chemistry*, **70**, 3721–3727.

Gobom J, Nordhoff E, Mirgorodskaya E, Ekman R and Roepstorff PJ (1999) Sample purication and preparation technique based on nano-scale reversed-phase columns for the sensitive analysis of complex peptide mixtures by matrix-assisted laser desorption/ionization mass spectrometry. *Journal of Mass Spectrometry: JMS*, **34**, 105–116.

Griss P, Melin J, Sjödahl J, Roeraade J and Stemme G (2002) Development of micromachined hollow tips for protein analysis based on nanoelectrospray ionization mass spectrometry. *Journal of Micromechanics and Microengineering*, **12**, 682–687.

James P (Ed.) (2000) *Proteome Research: Mass Spectrometry (Principles and Practice)*, Springer Verlag: ISBN: 3540672559.

Kameoka J, Craighead HG, Zhang H and Henion J (2001) A polymeric microfluidic chip for CE/MS determination of small molecules. *Analytical Chemistry*, **73**(9), 1935–1941.

Kameoka J, Orth R, Ilic B, Czaplewski D, Wachs T and Craighead HG (2002) An electrospray ionization source for integration with microfluidics. *Analytical Chemistry*, **74**(22), 5897–5901.

Killeen K, Yin H, Sobek D, Brennen R and Van de Goor T (2003) GoorChip-LC/MS: HPLC-MS using polymer microfluidics. In *Proceedings of μTAS 2003*, Northrup MA, Jensen K and Harrison DJ (Eds), Transducers Research Foundation Inc, ISBN: 0-9743611-0-0, pp. 481–484.

Little DP, Cornish TJ, Odonnell MJ, Braun A, Cotter RJ and Koster H (1997) MALDI on a chip: Analysis of arrays of low femtomole to subfemtomole quantities of synthetic oligonucleotides and DNA diagnostic products dispensed by a piezoelectric pipet. *Analytical Chemistry*, **69**, 4540–4546.

Marko-Varga G, Ekström S, Helldin G, Nilsson J and Laurell T (2001) Disposable polymeric high density nanovial arrays for MALDI-TOF Mass Spectrometry - A novel concept, Part I–Microstructure development and manufacturing. *Electrophoresis*, **22**, 3978–3983.

Marko-Varga G, Nilsson J and Laurell T (2004) Micro- and nanotechnology for proteomics. In *Proteome Analysis–Interpreting the Genome*, Speicher DW (Ed.), Elsevier, ISBN 0 444 51024 9, pp. 327–365.

Miliotis T, Kjellström S, Nilsson J, Laurell T, Edholm L-E and Marko-Varga G (2000) Capillary liquid chromatography interfaced to matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using an on-line coupled piezoelectric flow-through microdispenser. *Journal of Mass Spectrometry: JMS*, **35**, 369–377.

Ng JMK, Gitlin I, Stroock AD and Whitesides GM (2002) Components for integrated poly(dimethylsiloxane) microfluidic systems. *Electrophoresis*, **23**(20), 3461–3473.

Önnerfjord P, Nilsson J, Wallman L, Laurell T and Marko-Varga G (1998) Picoliter sample preparation in MALDI-TOF MS using a micromachined silicon flow-through dispenser. *Analytical Chemistry*, **70**, 4755–4760.

Preisler J, Foret F and Karger BL (1998) On-line MALDI-TOF MS using a continuous vacuum deposition interface. *Analytical Chemistry*, **70**, 5278–5287.

Preisler J, Hu P, Rejtar T, Moskovets E and Karger BL (2002) Capillary array electrophoresis-MALDI mass spectrometry using a vacuum deposition interface. *Analytical Chemistry*, **74**, 17–25.

Ramsey RS and Ramsey JM (1997) Generating electrospray from microchip devices using electroosmotic pumping. *Analytical Chemistry*, **69**, 1174–1178.

Reyes DR, Iossifidis D, Auroux PA and Manz A (2002) Micro total analysis systems. 1. Introduction, theory, and technology. *Analytical Chemistry*, **74**(12), 2623–2636.

Schuerenberg M, Luebbert C, Eickhoff H, Kalkum M, Lehrach H and Nordhoff E (2000) Prestructured MALDI-MS sample supports. *Analytical Chemistry*, **72**, 3436–3442.

Van Pelt CK, Zhang S, Fung E, Chu IH, Liu TT, Li C, Korfmacher WA and Henion J (2003a) A fully automated nanoelectrospray tandem mass spectrometric method for analysis of Caco-2 samples. *Rapid Communications in Mass Spectrometry: RCM*, **17**(14), 1573–1578.

Van Pelt CK, Zhang S, Kapron J, Huang X and Henion J (2003b) Am chip-based automated nanoelectrospray mass spectrometry. *Lab*, **35**(12), 14–21.

Wall DB, Berger SJ, Finch JW, Cohen SA, Richardson K, Chapman R, Drabble D, Brown J and Gostick D (2002) Continuous sample deposition from reversed-phase liquid chromatography to tracks on a matrix-assisted laser desorption/ionization precoated target for the analysis of protein digests. *Electrophoresis*, **23**, 3193–3204.

Wallman L, Ekström S, Nilsson J, Marko-Varga G and Laurell T (2003) A capillary filling micro-system for solid phase extraction and dispensing of proteomic samples. In *Proceedings of μTAS 2003*, Northrup MA, Jensen K and Harrison DJ (Eds.), Transducers Research Foundation Inc, ISBN: 0-9743611-0-0, pp. 887–890.

Wallman L, Ekström S, Marko-Varga G, Laurell T and Nilsson J (2004) Autonomus protein sample processing on-chip using solid phase microextraction, capillary force pumping, and microdispensing. *Electrophoresis*, **25**, 3778–3787.

Xue Q, Foret F, Dunayevskiy YM, Zavracky PM, McGruer NE and Karger BL (1997) Multichannel microchip electrospray mass spectrometry. *Analytical Chemistry*, **69**, 426–430.

Protein fingerprinting

David Fenyo
GE Healthcare, Piscataway, NJ, USA

1. Introduction

Identification of proteins by searching protein sequence collections with peptide mass fingerprinting data is widely used (Figure 1). The proteins in the sample are first separated to obtain one or a few proteins of interest. These proteins are then digested with a proteolytic enzyme. Mass spectra of the resulting peptide mixtures are acquired. The mass spectra are processed to find the masses of the peptides in the mixture. These measured masses are compared with calculated peptide masses for each protein in a protein sequence collection according to the rules defined by a set of user-defined parameters (Figure 2). A score is calculated for each comparison and the protein sequences in the collection are ranked according to the calculated score. Different search engines calculate the score in different ways (Henzel *et al*., 1993; Mann *et al*., 1993; Pappin *et al*., 1993; Yates *et al*., 1993; James *et al*., 1993; James *et al*., 1994; Wilkins *et al*., 1998; Perkins *et al*., 1999; Clauser *et al*., 1999; Berndt *et al*., 1999; Gras *et al*., 1999; Zhang and Chait, 2000; Gay *et al*., 2002; Eriksson and Fenyo, 2004a; Samuelsson *et al*., 2004; Rognvaldsson *et al*., 2004; Magnin *et al*., 2004; Levander *et al*., 2004). The meaning of these scores is in general not easily understood by the nonexpert user and they are not amenable to automation. Therefore, an additional step is necessary to test the significance of the results and convert the search engine–dependent score into a measurement of the significance of the protein identification result.

The problem of comparing the experimental mass spectra with calculated peptide masses for a protein sequence collection has been solved with different approaches in the different search engines. However, all search engines calculate a score for ranking the proteins in the sequence collection. The best matching protein sequence has the highest probability of being present in the sample that was analyzed and it is therefore ranked highest. The most widely used search engines are Mascot (Perkins *et al*., 1999), ProFound (Zhang and Chait, 2000), and MS-Fit (Clauser *et al*., 1999). For an evaluation of the performance of the different peptide mass fingerprinting algorithms, see Chamrad *et al*. (2004).

This paper will discuss how different search parameters influence the results and how the search engine–dependent scores can be converted into a measurement of the significance of the protein identification result.

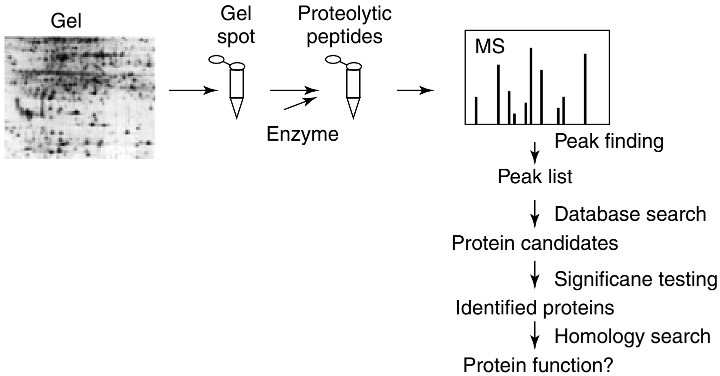


Figure 1 Protein fingerprinting is typically performed by first separating the proteins followed by enzymatic digestion and mass analysis. The raw mass spectra are subsequently analyzed to obtain a list of peptide masses. This peptide mass map is then searched against a protein sequence collection and the significance of each protein candidate is tested

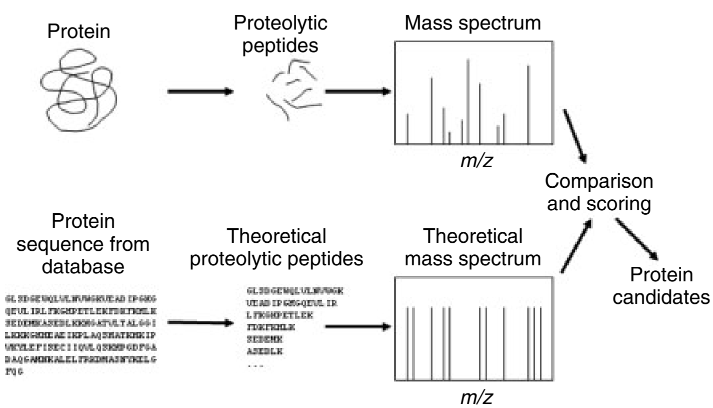


Figure 2 Searching a sequence collection with peptide mass fingerprinting data is performed by mimicking the experiment in silico: each entry in a protein sequence collection is theoretically digested using the same specificity as the enzyme used in the experiment. A theoretical mass spectrum is constructed, compared with the measured mass spectrum. The entries in the protein sequence collection are ranked according to how well they match the experimental data

2. Significance testing

One of the critical steps in protein identification is significance testing (Eriksson and Fenyo, 2004a; Eriksson *et al*., 2000; Eriksson and Fenyo, 2002; Fenyo and Beavis, 2003; Eriksson and Fenyo, 2004b). False identifications are possible due to random matching between the measured and calculated masses. In the result of a search of a sequence collection with peptide mass fingerprinting data, there will always be a highest ranked protein sequence. This protein sequence might correspond to a protein that is in the sample analyzed or simply be a false-positive, that is, get the highest score because of random matching between the calculated and

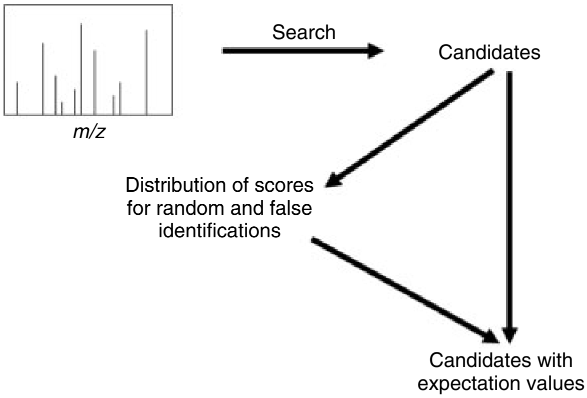


Figure 3 Most protein candidates given by a search engine are due to random matching. The distribution of scores for random and false identifications can therefore be obtained and used to calculate the expectation value for the protein candidates

measured proteolytic peptide masses. The probability that a protein candidate is a false-positive can be estimated by comparing its scores to the distribution of scores for random and false identifications. The distribution of scores for random and false identifications can be obtained by collecting statistics during the search. Figure 3 illustrates the method for using the statistics collected during the search to estimate the significance of the results. A score is calculated for each protein sequence in the collection. For the majority of sequences, the matching with the experimental data is random. An example of the distribution of the scores for proteins in a sequence collection matching a peptide mass fingerprint is shown in Figure 4. Typically, a distribution of scores from randomly matching protein sequences is observed at low scores. This distribution is an extreme value distribution, having a linear tail when plotted on a log–log scale. The expectation value of high-scoring protein sequences is estimated by extrapolation.

3. Search parameters

The different search engines available for peptide mass fingerprinting have a similar set of parameters, including enzyme specificity, sequence collection, modifications, peptide mass tolerance, protein mass, and pI. In cases in which the data quality is very high, a significant result will be obtained with the search parameters set within a wide range. However, in most cases, it is critical to select the parameters carefully when searching a protein sequence collection with peptide mass fingerprinting data. In general, it is recommended that all the information available that can restrict the search be used, for example, if the origin of the sample is known, a lot can be gained by only searching the species of interest and not all known protein sequences from all organisms. Also, information that will increase the number of possible matching peptides in the dataset should be used conservatively, for example, partial modification such as phosphorylation.

It is important to use enzymes with high specificity for peptide mass fingerprinting. The number of incomplete cleavage sites selected will influence the results; a

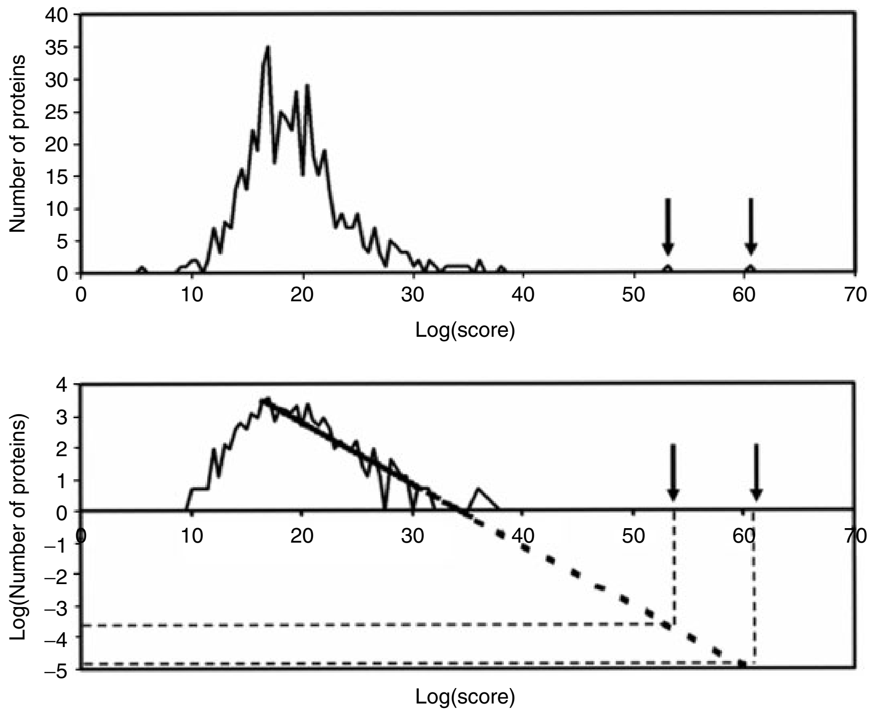


Figure 4 An example of the distribution of scores for random and false identifications for a peptide mass fingerprinting search with ProFound (Zhang and Chait, 2000). The distribution is an extreme value distribution, having a linear tail when plotted on a log–log scale. The expectation value of high-scoring protein sequences is estimated by extrapolation

larger number of incompletes will increase the noise because there will be more possible peptide sequences. This is illustrated in Figure 5 in which the distribution of scores for the random matching shifts to higher scores when the number of possible missed cleavage sites allowed is increased from 1 to 2 and 4. It is therefore recommended that an experimental protocol be used in which the proteins are digested as completely as possible, allowing the search to be performed with a setting of 0 or 1 incomplete cleavage sites.

There are many different sequence collections available for searching. Protein sequence collections are the most common choice for protein identification with peptide mass fingerprinting data. Searching of raw genomic data by translating the entire DNA sequence in all six possible reading frames with peptide mass fingerprinting data can be successful for organisms with small genomes, but is generally not done, because it requires very high quality experimental data. Expressed sequence tag (EST) collections are in general unsuited for searching with peptide mass data, because of their incomplete coverage of the genes. The smaller the sequence collection searched, the higher the significance of the results, provided that the collection contains the sequence of interest (Figure 5), for example, if the origin of the sample is known, it is recommended that only the species of interest be searched.

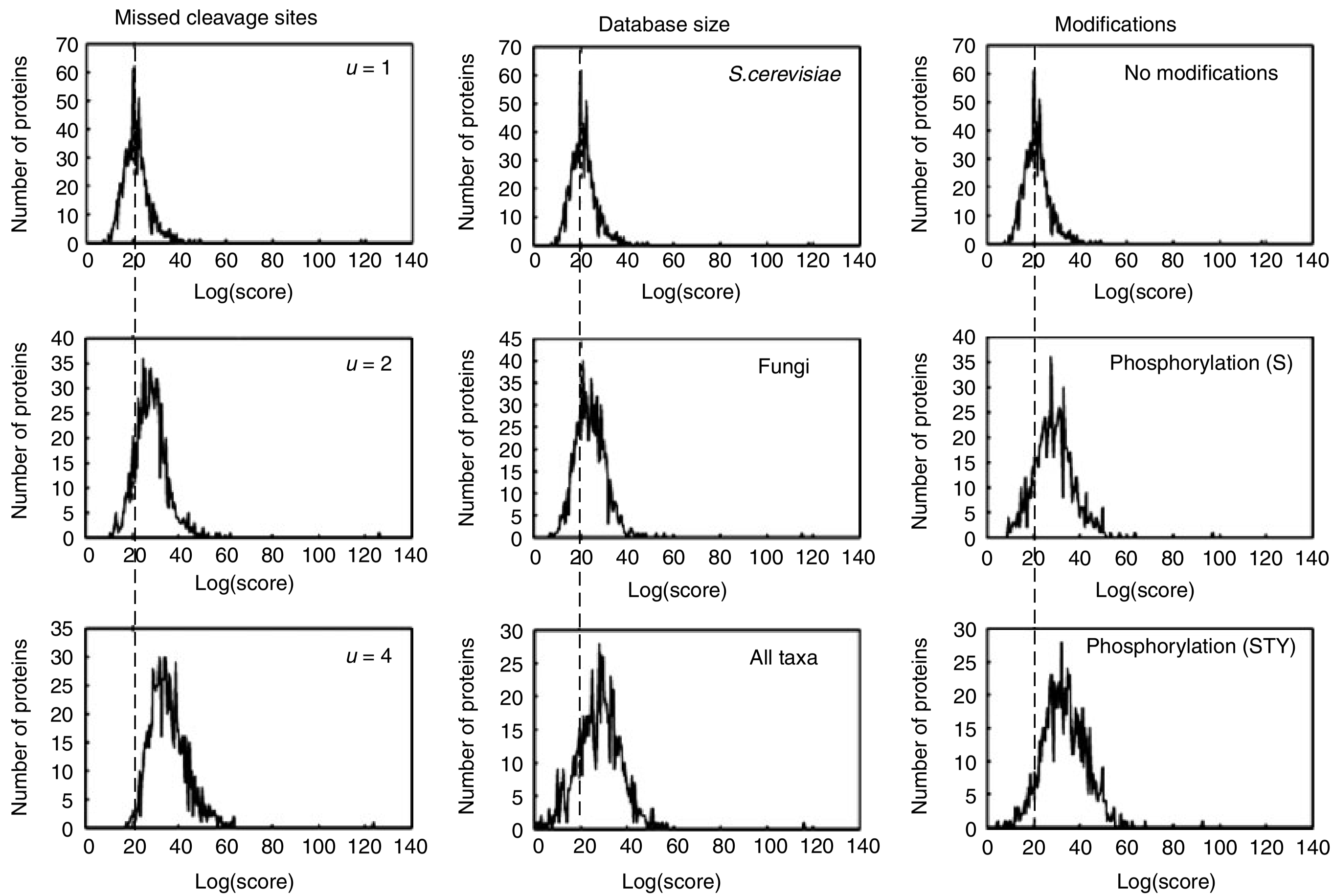


Figure 5 Example showing how the distribution of scores of random matches changes when different parameters are used with the same data set in searches with ProFound (Zhang and Chait, 2000)

Table 1 Example illustrating how the expectation value changes when different parameters (size of sequence collection, missed cleavage sites – *u*, and modifications) are used with the same data set in searches with ProFound (Zhang and Chait, 2000) (See Figure 5)

Sequence collection	E-value	*u*	E-value	Modifications	E-value
S. cerevisiae	4.8E-07	1	4.8E-07	No modifications	4.8E-07
Fungi	8.4E-06	2	1.1E-05	Phosphorylation (S)	2.3E-03
All taxa	2.9E-04	4	6.8E-04	Phosphorylation (STY)	2.1E-02

Proteins are often naturally modified and usually deliberately or unintentionally modified in the sample preparation process. Modifications can be defined as complete (i.e., they are always present on a specific amino acid) or partial (i.e., the modification might or might not be present on an amino acid). Complete modifications (e.g., cystein alkylation) do not increase the search time or change the significance. In contrast, partial modifications (e.g., phosphorylation) increases the search time and in general decreases the significance (Table 1) because the distribution of scores for the random matching shifts to higher scores (see Figure 5). If a peptide contains an amino acid that potentially might be modified, the mass of the unmodified peptide and the masses of the peptide with all possible modifications have to be compared with the measured peptide mass map. Even though a lot of proteins are modified, most proteins have only a few modified amino acids and therefore only a few of the peptides in a peptide mass map will be modified. Searching with partial modifications defined will increase the random background and potentially the number of matching peptides. In most cases, higher significance is achieved when searching without partial modifications defined.

The quality of the results obtained is dependent on the mass tolerance selected (Figure 6). Increasing the mass tolerance will increase the number of both true- and false-matching peptides. The two extreme cases are (1) zero mass tolerance – no peptides match and (2) large mass tolerance – all peptides match. Therefore, the best results are obtained at the mass tolerance that balances the contributions

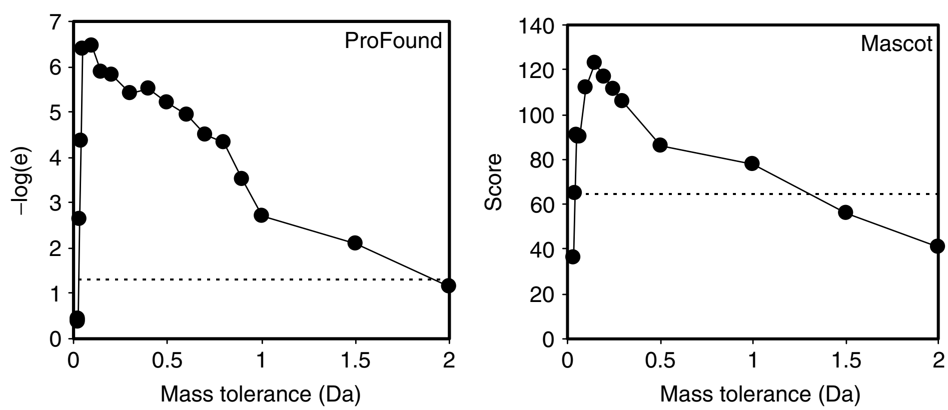


Figure 6 An example showing how the mass tolerance affects the results of the search for ProFound (Zhang and Chait, 2000) and Mascot (Perkins *et al*., 1999). The dotted line shows the probability of 0.05 for the result being false

from true- and false-matching peptides. The best setting for the mass accuracy can differ for different search engines because of the differences in the algorithms, for example, in Figure 6, the best results are obtained with a mass tolerance of 0.15 for Mascot and 0.1 for ProFound.

If the source of the analyte is a spot in 2D gel, information on protein properties such as mass and pI is available. This information can be used to restrict the search by excluding all sequences in the collection being searched that do not match the measured mass and pI and thereby increasing the significance of the results. The tolerance for protein mass and pI should, however, not be set too narrow because the measurement can differ from the calculated because of several reasons: (1) the protein has been processed and only a small domain is observed; (2) the splice variant observed is not in the sequence collection; and (3) the intron–exon boundaries are incorrectly assigned.

4. Summary

For successful protein identification, it is necessary to (1) use a sensitive and selective algorithm; (2) carefully select search parameters; and (3) test the significance of the results. The potential for obtaining a true mass spectrometric protein identification result depends on the choice of algorithm as well as on experimental factors that influence the information content in the mass spectrometric data. Current methods can never definitely prove that a result is true, but an appropriate choice of algorithm can provide a measure of the statistical risk that a result is false – that is, the statistical significance – and guide the practitioner in interpreting the results.

References

Berndt P, Hobohm U and Langen H (1999) Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis*, **20**(18), 3521–3526.

Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J and Bluggel M (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, **4**(3), 619–628.

Clauser KR, Baker P and Burlingame AL (1999) Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **71**(14), 2871–2882.

Eriksson J, Chait BT and Fenyo D (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Analytical Chemistry*, **72**(5), 999–1005.

Eriksson J and Fenyo D (2002) A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis. *Proteomics*, **2**(3), 262–270.

Eriksson J and Fenyo D (2004a) Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of Proteome Research*, **3**(1), 32–36.

Eriksson J and Fenyo D (2004b) The statistical significance of protein identification results as a function of the number of protein sequences searched. *Journal of Proteome Research*, **3**(5), 979–982.

Fenyo D and Beavis RC (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*, **75**(4), 768–774.

Gay S, Binz PA, Hochstrasser DF and Appel RD (2002) Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics*, **2**(10), 1374–1391.

Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, *et al*. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, **20**(18), 3535–3550.

Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C and Watanabe C (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences of the United States of America*, **90**(11), 5011–5015.

James P, Quadroni M, Carafoli E and Gonnet G (1993) Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, **195**(1), 58–64.

James P, Quadroni M, Carafoli E and Gonnet G (1994) Protein identification in DNA databases by peptide mass fingerprinting. *Protein Science*, **3**(8), 1347–1350.

Levander F, Rognvaldsson T, Samuelsson J and James P (2004) Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics*, **4**(9), 2594–2601.

Magnin J, Masselot A, Menzel C and Colinge J (2004) OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting. *Journal of Proteome Research*, **3**(1), 55–60.

Mann M, Hojrup P and Roepstorff P (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, **22**(6), 338–345.

Pappin DDJ, Hojrup P and Bleasby AJ (1993) Rapid identification of proteins by peptide-mass finger printing. *Current Biology*, **3**, 327–332.

Perkins DN, Pappin DJ, Creasy DM and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18), 3551–3567.

Rognvaldsson T, Hakkinen J, Lindberg C, Marko-Varga G, Potthast F and Samuelsson J (2004) Improving automatic peptide mass fingerprint protein identification by combining many peak sets. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, **807**(2), 209–215.

Samuelsson J, Dalevi D, Levander F and Rognvaldsson T (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, **20**(18), 3628–3635.

Wilkins MR, Gasteiger E, Wheeler C, Lindskog I, Sanchez J-C, Bairoch A, Dunn MJ and Hochstrasser DF (1998) Multiple parameter cross-species protein identification using MultiIdent–a world-wide web accessible tool. *Electrophoresis*, **19**(18), 3199–3206.

Yates JR III, Speicher S, Griffin PR and Hunkapiller T (1993) Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry*, **214**(2), 397–408.

Zhang W and Chait BT (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry*, **72**(11), 2482–2489.

Multidimensional liquid chromatography tandem mass spectrometry for biological discovery

Claire M. Delahunty and John R. Yates III
The Scripps Research Institute, LaJolla, CA, USA

1. Introduction

The goal of biological studies is to connect genes and proteins to specific processes. In the past, dissection of biological processes entailed detailed and focused studies of a specific gene or protein. The availability of complete genome sequences has made possible broader studies to dissect the functions of proteins and genes in entire pathways, thus extending the knowledge of the process. Proteomics in an organism is facilitated by the existence of a completed genome sequence, which allows more facile and higher-throughput correlations between gene and protein sequences than has been possible with more traditional protein-sequencing methods. Correlations are derived through the use of mass spectrometry data that reveals the identity of a protein through the molecular weights of peptides in a tryptic map or by fragment ions of individual peptides. Fragmentation of individual peptides is accomplished through the use of tandem mass spectrometers that can select individual ions, induce dissociation, and record the resulting product ions (Hunt *et al*., 1986). The resulting fragment ions represent the amino acid sequence of the peptide. Fragmentation patterns can be matched to sequences in a database, thus identifying the protein from which the peptide was derived. An analytical advantage to this method is the ability to measure peptide mixtures derived from proteolytically digested protein mixtures (Figure 1) (Yates, 1998). This approach represents an efficient and sensitive method to identify the components of protein complexes, protein localization (e.g., proteins located in organelles or subcellular spaces, cellular membranes), and protein modifications (Link *et al*., 1997; Wu *et al*., 2003).

To accommodate complex protein mixtures, separation systems of high resolution are desirable. Multidimensional liquid chromatography (MudPIT) provides high-resolution separations of peptides in an automated, convenient manner. Link *et al*. (1999) developed an integrated form of MudPIT that combines strong cation exchange (SCX) resin with a reversed-phase (RP) to create a biphasic column

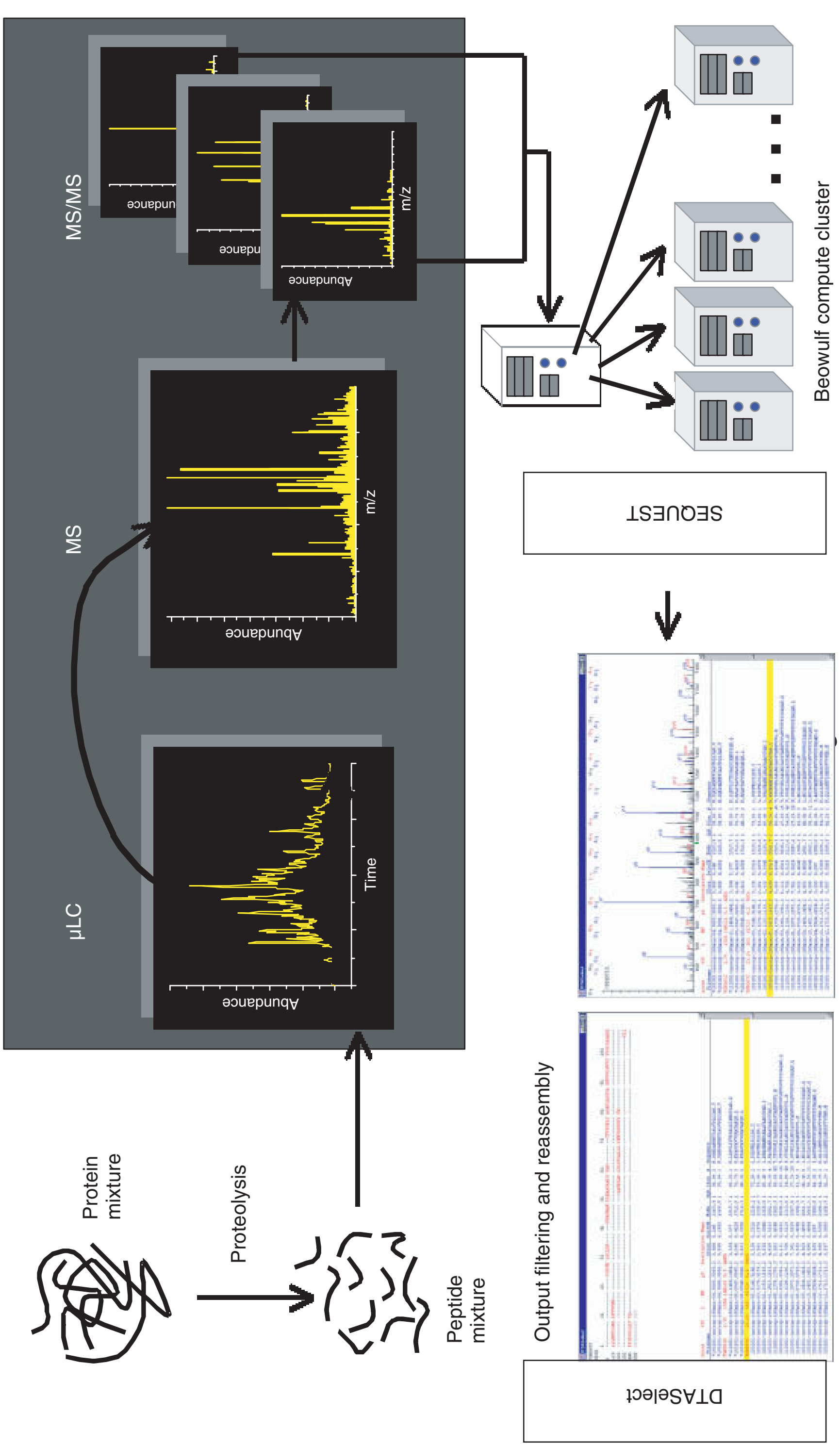


Figure 1 The process of shotgun proteomics involves digesting a protein mixture and then separating the peptides into a tandem mass spectrometer. Peptide ions are individually selected for collision-induced dissociation and then each resulting spectrum is searched through a sequence database. The results are filtered, assembled, and organized to view the results

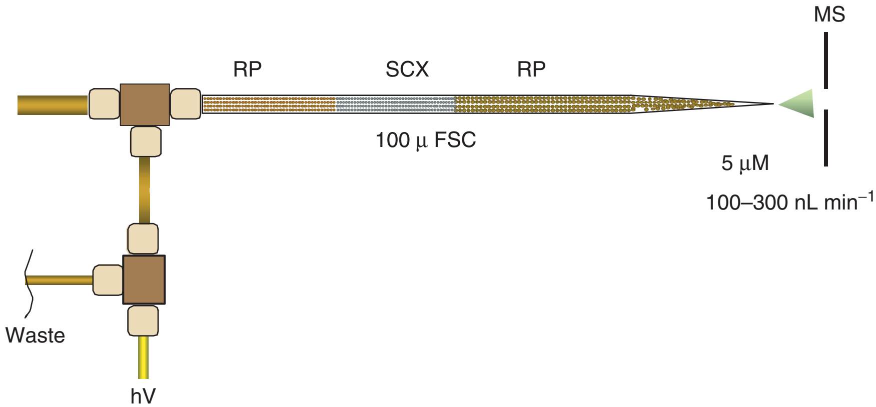


Figure 2 An integrated multidimensional liquid chromatography column to perform high-resolution separations of peptides. The tip of the column is packed with reversed-phase (RP) packing material followed by a layer of strong cation exchange (SCX) resin and then a short layer of RP material. Peptides are electrosprayed out the end of the column directly into a mass spectrometer

(Figure 2). The combination of SCX and RP packing materials creates a bimodal, orthogonal separation by charge and hydrophobicity. Modification to this process was introduced by McDonald *et al*. (2002) to add a third layer of chromatography material to effect an on-column removal of buffer salts that may be present in the sample and to improve retention of peptides on the SCX phase. Washburn *et al*. (2001) demonstrated that this approach was effective for an unbiased analysis of the components of yeast cells by identifying large and small proteins, basic and acidic proteins, and membrane proteins.

2. Analysis of protein complexes

This strategy has been used to improve the analysis of protein complexes and to localize proteins (Figure 3). A recent study used biochemical purification of protein complexes combined with identification using the MudPIT approach (Hazbun *et al*., 2003). The essential hypothetical open reading frames were all epitope-tagged with the Tandem Affinity Purification (TAP) tag that combines a protein A IgG binding domain with the calmodulin-binding peptide (Rigaut *et al*., 1999). The two binding domains are linked together with a short amino acid sequencing containing the tobacco etch virus protease cleavage site. Inclusion of the cleavage site allows protein complexes to be removed from the initial purification step on an IgG column by proteolytic cleavage. The complex is recovered by passing the eluate through a calmodulin column and then released by elution with EDTA (ethylenediamine tetraacetic acid) (Gavin *et al*., 2002). The complex is then digested using trypsin to produce a collection of peptides. Peptides are loaded onto a two-dimensional LC column and separated directly into a tandem mass spectrometer. Tandem mass spectra were automatically collected and then searched against the yeast

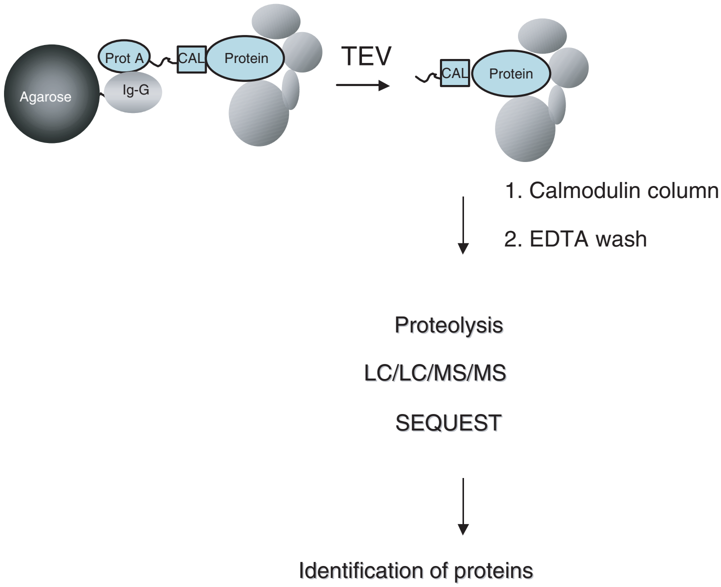


Figure 3 Protein complexes can be analyzed by incorporating an epitope into a protein sequence that contains protein A, the tobacco etch virus protease cleavage site, and the calmodulin-binding peptide. After elution of the complex from the columns, the proteins are digested and analyzed by MudPIT

sequence database to identify the proteins in the mixture. Of the 100 tagged proteins analyzed, complexes were identified for 29 of them. On the basis of the identities of the proteins interacting with the tagged proteins in combination with data from colocalization and yeast two-hybrid experiments, gene ontology annotations could be assigned to the proteins. This study and others have shown MudPIT to be a very effective method to identify proteins in complexes (Graumann *et al*., 2003) as well as to identify sites of modifications in proteins of complexes (Cheeseman *et al*., 2002; MacCoss *et al*., 2002).

3. Analysis of subcellular compartments

A particular challenge for proteomics studies is the identification of proteins in membranes and organelles. These studies are complicated because it is difficult to enrich for membrane or organelle proteins without contamination and because these proteins have limited solubility in buffers compatible with mass spectrometry. By using MudPIT, subtractive analysis approaches can be utilized as well as strategies to digest only the soluble segments of membrane proteins.

Schirmer *et al*. (2003) used a subtractive approach to identify proteins of the nuclear envelope (NE) (Figure 4). Enrichment of the NE is particularly challenging because it is contiguous with the endoplasmic reticulum (ER) and intermixed with mitochondria. An enriched fraction of the microsomal membranes (MM) that includes both ER and mitochondria can be generated. Schirmer *et al*. generated

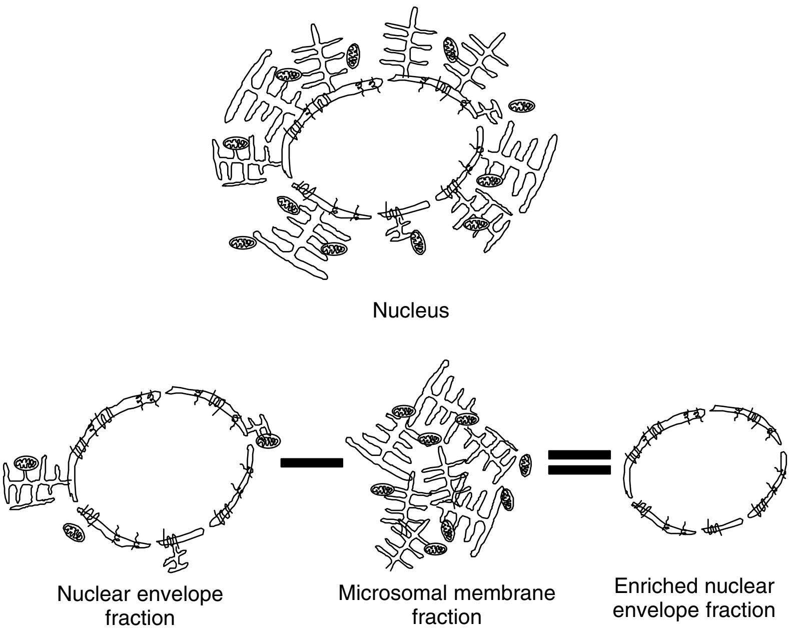


Figure 4 Subtractive proteomics was applied to the nuclear envelope of mammalian cells by enrichment of the microsomal membrane fraction for exhaustive proteome analysis. An enriched nuclear envelope fraction was then analyzed and proteins in common between the two were subtracted from the list of identified proteins

an exhaustive proteome analysis of the microsomal fraction and an enriched NE fraction using MudPIT. Proteins identified in the MM fraction were then subtracted from those identified in the NE fraction. Sixty-seven proteins were identified as hypothetical, integral membrane proteins. This example illustrates a powerful aspect of the shotgun proteomics method enabled by MudPIT. Membrane proteins were readily identified using this approach, since it is not necessary that the proteins be solubilized for digestion only that digestion of the soluble portions of the membrane proteins is achieved. A possible limitation of the approach is potential loss of information from those membrane proteins with minimal regions exposed.

References

Cheeseman IM, Anderson S, Jwa M, Green EM, Kang J, Yates JR III, Chan CS, Drubin DG and Barnes G (2002) Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p. *Cell*, **111**, 163–172.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Graumann, J, Dunipace, LA, Seol, JH, McDonald, WH, Yates, JR III, Wold BJ and Deshaies RJ (2003) Applicability of TAP-MudPIT to pathway proteomics in yeast. *Molecular & Cellular Proteomics*, **3**, 226–237.

Hazbun TR, Malmstrom L, Anderson S, Graczyk BJ, Fox B, Riffle M, Sundin BA, Aranda JD, McDonald WH, Chiu CH, *et al*. (2003) Assigning function to yeast proteins by integration of technologies. *Molecules and Cells*, **12**, 1353–1365.

Hunt DF, Yates JR III, Shabanowitz J, Winston S and Hauer CR (1986) Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 6233–6238.

Link AJ, Carmack E and Yates JR III (1997) A strategy for the identification of proteins localized to subcellular spaces: application to *E. coli* periplasmic proteins. *International Journal of Mass Spectrometry and Ion Processes*, **160**, 303–316.

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM and Yates JR III (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, **17**, 676–682.

MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, *et al*. (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7900–7905.

McDonald, WH, Ohi, R, Miyamoto, DT, Mitchison TJ and Yates JR III (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single dimension LC/MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *International Journal of Mass Spectrrometry*, **219**, 245–251.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Schirmer EC, Florens L, Guan T, Yates JR III and Gerace L (2003) Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science*, **301**, 1380–1382.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Wu CC, MacCoss MJ, Howell KE and Yates JR (2003) A method for the comprehensive proteomic analysis of membrane proteins. *Nature Biotechnology*, **21**, 532–538.

Yates JR III (1998) Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, **33**, 1–19.

Sample preparation for MALDI and electrospray

J. I. Langridge and M. Snel
Micromass UK Ltd, Manchester, UK

1. Introduction

The development of electrospray (ESI) and matrix-assisted laser desorption/ ionization (MALDI) in the early 1990s built upon the advances made by fast atom bombardment (FAB). FAB allowed the analysis of biomolecules previously intractable to mass spectrometric analysis, without derivatization. The subsequent development of MALDI and ESI provided increased sensitivity while significantly increasing the molecular mass of proteins that could be accurately measured. Further refinement of both ionization techniques and the associated mass analyzers has provided modern, state-of-the-art instrumentation that can detect low levels of endogenous biological samples with high accuracy. In particular, the coupling of MALDI and ESI with mass spectrometers capable of MS/MS (*see* Article 10, **Hybrid MS**, Volume 5) has resulted in significant advances in protein identification and characterization. However, one of the limiting factors in this inherently sensitive technology is the preparation of samples for the mass spectrometer. This crucial step often defines the quality of results obtained. For example, one common cause of problems is the incomplete removal of buffers used in the biology laboratory. The presence of detergents, salts, and other common buffer components can adversely affect the performance of ESI and MALDI mass spectrometers. The presence of alkali metals in a sample can cause what should be a single component to appear as multiple adduct peaks containing varying numbers of metal ions.

The following section details the most commonly used sample preparation techniques in use in the laboratory today.

2. Sample preparation for MALDI

Matrix-assisted laser desorption/ionization is a soft ionization technique. In this technique, the analyte molecules are embedded in a matrix consisting of small organic molecules. The sample is irradiated with a laser pulse, which leads to rapid heating of the matrix molecules and results in a transition into the gas phase. Analyte molecules included in the matrix are also brought into the gas phase and are ionized, often through adduct formation. In its ionized form, the analyte can be separated according to its m/z ratio, using a variety of different mass analyzers.

In modern configurations, a time-of-flight (TOF) mass analyzer (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) is most commonly used.

MALDI sample preparation of peptide and protein samples can be extremely straightforward, and as a result, MALDI has rapidly established itself as one of the standard techniques for rapid protein identification. In the years after the introduction of MALDI, the main focus was on the mass determination of intact proteins. The emphasis shifted rapidly to peptide analysis, with the invention of a new technique for protein identification termed *peptide mass fingerprinting* (*PMF*) (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5). In this approach, proteins are enzymatically digested, typically with trypsin, and the resultant peptides mass-measured. The monoisotopic masses obtained from the MALDI PMF experiment are then compared to a protein or nucleotide sequence database. The main sample preparation factors affecting the overall performance of MALDI experiments are choice of matrix, sample purity, solvents used, and the method of applying the matrix and sample. More specialized applications such as phosphopeptide analysis may require more rigorous sample cleanup and sample preparation routines, and these are dealt with later under electrospray sample preparation.

2.1. Sample cleanup

The best results in MALDI analysis are obtained from clean samples. Much can be done in the steps leading up to the MALDI analysis to reduce or even eliminate the need for sample cleanup. Common biological buffers, contaminants, and salts can be tolerated to the levels shown in Table 1. If the buffer/salt levels are too high, these can be reduced by a number of different methods. A widely used technique is binding peptides on reverse-phase material and then washing them with aqueous solutions, followed by elution in high organic. Alternatively, salts may be removed by dialysis. Washing of the sample by adding a droplet of water to a prepared MALDI sample spot and then drawing off the liquid and salt has also been shown to be a useful strategy. It should, however, always be borne in mind that sample cleaning may result in some loss or dilution of sample and if it can be avoided through careful choice of the sample preparation steps, then the best results will be obtained. A common source of sample contamination when a protein sample is being enzymatically digested is keratin. This originates from clothing and human skin, and it is often surprising to researchers when they first start working with MALDI just how sensitive the technique is and how readily keratin can be detected. This contamination can usually be avoided through the use of gloves and clean air enclosures.

2.2. Choice of matrix and solvent

There are numerous ways to prepare a given sample for analysis by MALDI, however, some broad guidelines can be made. The first choice that needs to

Table 1 Common buffers and contaminants and the maximum concentrations tolerated in MALDI and ESI MS

Contaminant	MALDI max. conc.	ESI max. conc.
Phosphate buffer	20 mM	<2 mM
Tris buffer	50 mM	5 mM
Detergents	0.1%	0.1%
SDS	0.01%	0.01%
Alkali metal salts	1 M	–
Glycerol	2%	–
NH_4 bicarbonate	30 mM	50 mM
Guanidine	1 M	10 mM
Sodium azide	1%	–

be made is the choice of matrix. Over the years, a large number of matrices have been suggested, however, two have been used routinely for the analysis of peptides. The most commonly used matrix for peptide mass fingerprinting is α-cyano-4-hydroxycinnamic acid (CHCA). It is relatively easy to obtain fine matrix crystals with this matrix, which tends to give a more homogenous sample surface. A homogenous sample surface makes it easier to find a "hot spot" providing mass spectra containing the analyte of interest. It is also possible to get excellent sensitivity with this matrix. Several different concentration levels and solvents have been suggested for the preparation of CHCA, however, the most commonly used concentration is 10 mg mL^{-1}. Despite this, experiments have shown that for low peptide concentrations, that is, subpicomol on target, matrix concentrations of 2 mg mL^{-1} or lower work better. The solvent mix used depends on the evaporation rate desired; this can be very dependent on ambient humidity and temperature. A mixture of 1:1 acetonitrile:water containing 0.1% trifluoroacetic acid (TFA) generally works well, although the water can also be replaced by ethanol.

For protein analysis, sinapinic acid can be a very effective matrix. This matrix can be used in a dried droplet experiment and in thin-film preparations (see spotting methods). For both of these methods, a sinapinic acid solution of 10 mg mL^{-1} in 6:4 acetonitrile:0.1% TFA is used. The thin-film method additionally calls for a sinapinic acid solution of 10 mg mL^{-1} in acetone or acetonitrile. Sinapinic acid forms very homogenous sample spots with an even analyte distribution.

Another versatile matrix is 2,5-dihydroxybenzoic acid (DHB). This can be used for protein and peptide analysis but is also the matrix of choice for other compounds, for example, oligosaccharides. DHB tends to form large crystals, where the analyte is not evenly distributed and this leads to localized hot spots usually at the tips and edges of the crystals. A wide range of different solvent mixtures and matrix concentrations has been suggested for DHB. A typical example would be 20 mg mL^{-1} in 7:3 0.1% TFA:acetonitrile. Several other matrices and various derivations of the above matrices have also been described in the literature but these are not widely used.

2.3. Spotting methods

Several different spotting methods have been developed for MALDI samples. The methods mostly differ in the way the matrix solution is combined with the analyte solution, and they can have a drastic effect on the quality of the MALDI mass spectrum acquired.

The dried droplet method is a reliable and straightforward method. Roughly equal volumes of matrix and sample solution are mixed. This can be done prior to spotting in the sample tube, followed by spotting of a 00.5–2 μL droplet onto the target plate. This droplet is then allowed to dry either under normal lab conditions or at slightly reduced pressure. Alternatively, the matrix and analyte solutions can be combined directly on the target plate by first depositing a small volume of matrix (0.5–2 μL) and then subsequently adding a droplet of analyte solution. Mixing can then either be left to diffusion or can be aided by aspiration of the matrix/analyte droplet on the target plate using a pipette. Contact of the pipette tip with the sample holder should be avoided, as this can cause crystallization of the matrix.

The sandwich and thin-film methods are slightly more involved in comparison to the dried droplet method. However, these methods can be particularly effective for protein analysis, on-target desalting, and high-sensitivity work. In these methods, a thin film of matrix is deposited on the target plate and allowed to dry. Depositing a small volume of matrix in high organic, for example, acetone, produces the thin film. This thin film consists of very fine crystals. The sample solution added, either combined with matrix or on its own, and the solvent mix in the droplet that is added must be in a high aqueous environment to avoid dissolving the thin film.

A further derivation of this technique is the sandwich method in which a drop of matrix solution is applied after this procedure, covering the analyte with a matrix layer.

3. Sample preparation for electrospray

Electrospray ionization is a soft ionization technique. The sample is introduced into the electrospray ion source as a continuous liquid stream. As such, ESI is often directly coupled to on-line sample cleanup and separation techniques, such as liquid chromatography. This forms the basis of most sample preparation techniques for ESI. LC-ESI MS is compatible with a wide range of solvent compositions and ionizes a diverse range of compounds. It has found particular use in the area of biological mass spectrometry due to the ability of molecular species to accept multiple charges during the electrospray process. As the mass spectrometer separates according to mass/charge (m/z), high-molecular-weight compounds, such as large protein complexes, now appear lower down the m/z range, within the mass range of conventional mass analyzers.

Further refinements of electrospray for biological applications resulted in the development of nanoelectrospray (Wilm and Mann, 1994), where flow rates of the liquid stream maybe as low as a few nanoliters per minute. One of the limiting factors for analysis of proteins and peptides by electrospray or nanoelectrospray

mass spectrometry is sample preparation. The presence of many biological buffers (e.g., HEPES, Urea) interfere with the ionization process and as such they should be avoided, or removed, prior to analysis.

3.1. Purification and concentration by reverse phase

A common strategy in the analysis of protein and peptide samples by mass spectrometry is to improve the quality of the data acquired by using reversed phase (RP) chromatography. This approach allows peptides or proteins to be separated on the basis of their hydrophobicity, and is often used "on-line", coupling liquid chromatography directly to an electrospray MS instrument. This approach has been used widely for the analysis of oligonucleotides, intact proteins, and proteolytic digests of proteins. The RP HPLC separation requires an ion-pairing reagent, and with traditional protein and peptide separation, 0.1–1% trifluoroacetic acid (TFA) is used. The sensitivity of ESI ionization is reduced by approximately a factor of 5 if TFA is present in the liquid stream, and as such alternative stationary phases have been produced by a variety of manufacturers that provide excellent separations using 0.1–1% formic acid as the ion pair. This provides an increase in ESI sensitivity compared to the TFA-based separation. In addition, as ESI is a concentration-dependent ionization technique, the use of capillary and nanoscale chromatography has lead to significant advances in ESI sensitivity.

For peptide analysis, the HPLC is often coupled to a mass spectrometer capable of data-directed switching between the MS and MS/MS modes. Hundreds of MS/MS spectra can be acquired from complex peptide samples in a fully automated fashion, resulting in the structural characterization of many species in a single analysis. For example, protein identification can be achieved via data bank searching of the ESI-MS/MS, providing qualitative information on the proteins that are present and identification of significant numbers of proteins, including low copy number proteins, from a single LC-MS/MS experiment.

The use of microcolumns for ESI sample preparation of in-gel digested samples for nano-ESI was first described by Wilm and Mann (1996). A small amount of reverse-phase material was packed into the tip of a nanospray needle and used only once, a new column being made for each sample. This provides a crude separation using step elution to desalt and concentrate analytes prior to analysis, and this approach has been shown to provide impressive levels of sensitivity. While this approach was first described for ESI, it has also been widely adopted for MALDI (Gobom *et al*., 1999). Disposable pipette tips packed with reverse-phase material have become commercially available from numerous sources and are suited to robotic sample preparation.

3.2. Strong cation exchange (SCX)

If a complex peptide or protein mixture is to be investigated, then a fractionation step prior to analysis via ESI RPLC-MS is advantageous. The use of SCX material provides an orthogonal separation mechanism to RP, separating peptides

or proteins, on the basis of their inherent or carried charge. It is therefore possible to prefractionate the peptides, prior to further separation by reverse-phase on a C18 column. This provides additional separation power, or peak capacity. A further derivation of this approach has seen the use of a biphasic column, where both SCX and RP material are sequentially packed into a nanoscale HPLC column. This allows two-dimensional LC separations to be performed on-line to the mass spectrometer (Link *et al*., 1999).

3.3. Hydrophilic interaction liquid chromatography (HILIC)

HILIC is a versatile, effective alternative to ion exchange and reverse-phase chromatography for the increased retention and separation of polar peptides. The technique permits direct coupling via LC-MS. An additional advantage of HILIC is that it works best where RP works worst – with polar solutes that are poorly retained on RP material. HILIC has been used successfully for the analysis of phosphopeptides, carbohydrates, peptide digests, and polar lipids.

HILIC separates compounds by eluting mostly organic mobile phase over a neutral hydrophilic stationary phase. This results in solutes eluting in order of increasing hydrophilicity – the inverse of reverse phase.

3.4. Purification using graphite columns

The use of reverse-phase material to desalt and concentrate proteolytic digests prior to mass spectrometry is widespread, as discussed previously. However, this does not allow for detection of small or hydrophilic peptides, or peptides altered in hydrophilicity. The use of graphite powder, in a microcolumn format, allows hydrophilic species to be retained efficiently on the graphite (Larsen *et al*., 2002). These can then be eluted and analyzed by mass spectrometry.

3.5. Immobilized metal-affinity chromatography (IMAC)

Despite the ability of mass spectrometry to analyze peptide components in complex mixtures, the detection of endogenous levels of phosphopeptides in a peptide mixture can be problematical. This is partly due to the low ionization efficiency of the phosphorylated species, and also due to the low stoichiometry of phosphorylation. Analysis of phosphorylation sites in complex mixtures can be facilitated through the use of affinity methods, which isolate and enrich phosphopeptides from samples prior to mass spectrometry. A popular method for isolating and enriching involves using trivalent metal ions, such as Fe^{3+}, Ga^{3+}, or Ni^{3+} that are bound to a chromatographic support. This technique is referred to as immobilized metal-affinity chromatography (IMAC). Primarily, IMAC experiments are performed in an off-line manner, with microcolumns (Stensballe *et al*., 2001), although on-line studies have also been reported.

Related articles

Article 2, **Sample preparation for proteomics**, Volume 5; Article 10, **Hybrid MS**, Volume 5; Article 15, **Handling membrane proteins**, Volume 5; Article 19, **Making nanocolumns and tips**, Volume 5

References

Gobom J, Nordhoff E, Mirgorodskaya E, Ekman R and Roepstorff P (1999) Sample purification and preparation technique based on nano-scale reversed-phase columns for the sensitive analysis of complex peptide mixtures by matrix-assisted laser desorption/ionization mass spectrometry. *International Journal of Mass Spectrometry*, **34**(2), 105–116.

Larsen MR, Cordwell SJ and Roepstorff P (2002) Graphite powder as an alternative or supplement to reversed-phase material for desalting and concentration of peptide mixtures prior to matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics*, **2**(9), 1277–1287.

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM and Yates JR III (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, **17**(7), 676–682.

Stensballe A, Andersen S and Jensen ON (2001) Characterization of phosphoproteins from electrophoretic gels by nanoscale Fe(III) affinity chromatography with off-line mass spectrometry analysis. *Proteomics*, **1**(2), 207–222.

Wilm M and Mann M (1994) Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes*, **136**(2–3), 167–180.

Wilm M and Mann M (1996) Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry*, **68**(1), 1–8.

Handling membrane proteins

Robert J. A. Goode
Ludwig Institute for Cancer Research and the Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
The University of Melbourne, Parkville, Victoria, Australia

Richard J. Simpson
Ludwig Institute for Cancer Research and the Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

1. Introduction

In spite of their vital importance to many cellular functions, we remain remarkably ignorant of membrane protein composition. Two major classes of membrane proteins exist: integral membrane proteins, which are embedded in the lipid bilayer, and the so-called peripheral membrane proteins, which are associated noncovalently with integral proteins or membrane lipid head groups (Figure 1).

Membrane protein structural biology lags behind that of the more-soluble class of proteins because only 0.2% of the ~17 000 known protein structures are integral membrane proteins. This is in contrast to a detailed bioinformatic analysis of integral membrane proteins of the helix-bundle class by Wallin and von Heijne (1998), which predicts that 20 to 30% of all open reading frames encode membrane proteins. Integral membrane proteins are also grossly underrepresented in proteomic analyses. These shortcomings have occurred because of the enormous technical challenges associated with their handling and characteristics. Such challenges arise from the low relative abundance of membrane proteins, insolubility caused by the hydrophobicity of their lipid embedded portions, and the variability of their posttranslational modifications, particularly glycosylation.

2. Membrane enrichment

Two-dimensional electrophoretic (2DE) gels are the classical proteomic tool for separation and visualization of complex protein samples (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5); however, the low relative abundance of integral

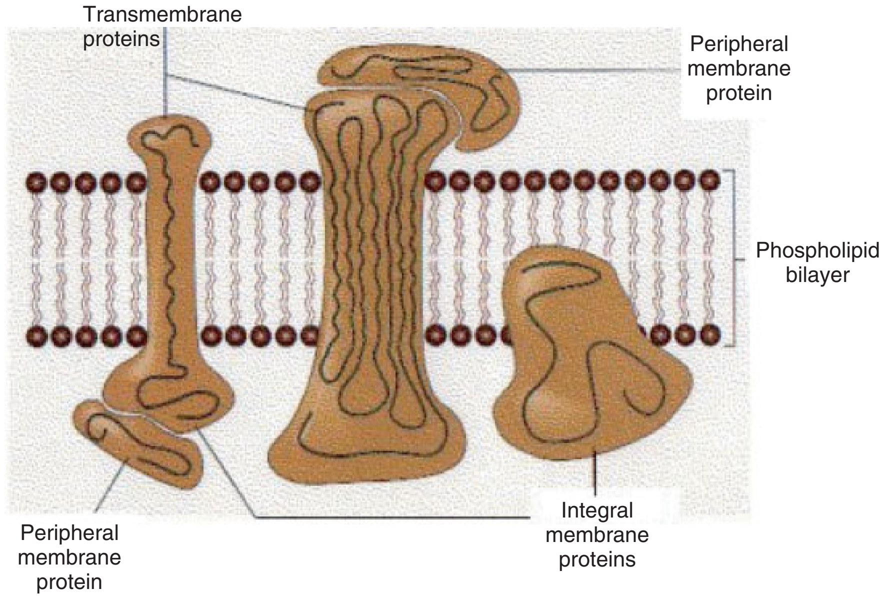


Figure 1 Diagrammatic representation of membrane protein structure. Integral membrane proteins are embedded in the lipid bilayer, while peripheral membrane proteins can be associated with other membrane proteins or with lipid head groups (not shown). (©2004 John W. Kimball. Reproduced by permission of John W. Kimball)

membrane proteins (compared with abundant cytoskeletal, cytoplasmic, and vesicle lumen proteins) makes their detection on 2DE gels near impossible owing to the dynamic range of most visualization methods (around 3–4 orders of magnitude). Therefore, various enrichment strategies have been performed prior to their analysis and identification (for reviews see Thomas and McNamee, 1990; Celis, 1998).

The simplest enrichment method is a microsomal preparation, in which ultracentrifugation is used to pellet the dense cellular membranes after mechanical cell lysis and removal of nuclei and unlysed cells (Simpson, Connolly *et al*., 2000). An advancement of this technique is the use of density gradients to fractionate the various cellular membranes according to their different densities (Pasquali, Fialka *et al*., 1999; Adam, Boyd *et al*., 2003). Sucrose was traditionally used to form the density gradients, although, more recently, Percoll or OptiPrep™ (Iodixanol) have been substituted in order to form nearly iso-osmotic gradients (Ford *et al*., 1994; Pertoft, 2000). Although it is claimed that these methods yield plasma membranes that are free of whole cells, nuclei, and soluble proteins (Thomas and McNamee, 1990), experience suggests that such preparations from epithelial cell lines are heavily contaminated with abundant soluble proteins, such as chaperones, histones, and ribosomal proteins, as judged by detailed mass spectrometry (MS) analysis (Simpson, Connolly *et al*., 2000; Adam, Boyd *et al*., 2003).

Several alternate methods exist for specifically enriching the plasma membrane, including in situ labeling with either a dense bead (e.g., cationic colloidal silica method (Chaney and Jacobson, 1983; Rahbar and Fenselau, 2004)) or an affinity tag

for its subsequent purification on an affinity matrix (e.g., cell surface biotinylation and streptavidin purification (Altin and Pagler, 1995; Shin, Wang *et al*., 2003; Peirce, Wait *et al*., 2004)).

Whilst still embedded in the lipid bilayer, integral membrane proteins can be further enriched by removing many peripheral membrane proteins and nonspecifically bound proteins by washing the membrane pellet with a variety of buffers that interrupt interactions dependent on metal ions (e.g., chelators, EDTA/EGTA) or ionic interactions (e.g., high salt concentrations or high pH buffers, such as 1 M NaCl or sodium carbonate (pH 11), respectively) (Thomas and McNamee, 1990).

3. Solubilization and separation

Solubilization of integral membrane proteins is essential for their subsequent fractionation and characterization, with detergents being the most commonly used method. Although very effective for solubilizing integral membrane proteins, ionic detergents such as sodium dodecyl sulfate (SDS), interfere with isoelectric focusing, the first dimension of 2DE, as well as mass spectrometry. Therefore, other zwitterionic detergents have been tested for both their ability to solubilize integral membrane proteins and their compatibility with 2DE (Henningsen, Gale *et al*., 2002; Luche, Santoni *et al*., 2003; Tastet, Charmont *et al*., 2003; Taylor and Pfeiffer, 2003). However, as no detergent solubilizes all proteins in 2D gels, several other gel-based techniques are available that offer higher solubility at the cost of lower resolving power, such as various one-dimensional PAGE techniques (Simpson, 2003), or the two-dimensional BAC-PAGE (Macfarlane, 1983; Hartinger, Stenius *et al*., 1996) and dSDS-PAGE (Rais, Karas *et al*., 2004).

Chromatographic separation of membrane proteins has been troublesome due to their hydrophobicity, insolubility, and the incompatibility of several chromatographic media with detergents. One exception is hydroxyapatite, which has been used successfully with both ionic and nonionic detergents for membrane protein purification (Fonyo, 1968; Riccio, Aquila *et al*., 1975; Engel, Schagger *et al*., 1980; Valpuesta and Barbon, 1988; Lundahl, Watanabe *et al*., 1992) and as a detergent exchange step prior to 2DE (Wissing, Heim *et al*., 2000). Alternatively, multidimensional liquid chromatography (MDLC), which combines strong cation exchange and reverse-phase liquid chromatography, interfaced with MS has been used to resolve and identify soluble tryptic peptides from membrane proteins (Washburn, Wolters *et al*., 2001). Mass spectrometry is generally impeded by the presence of detergents; however, recently acid-labile surfactants have been shown to improve the recovery of membrane protein–derived peptides in MS analyses (Norris, Porter *et al*., 2003; Yu, Gilar *et al*., 2004) and even positively select for membrane-spanning peptides (Yu, Gilar *et al*., 2004).

4. Dealing with glycosylation

Finally, over 70% of extracellularly exposed proteins are predicted to harbor carbohydrate modifications, which can be highly heterogeneous in composition (Gahmberg and Tolvanen, 1996). This reduces their resolution during separation at both

the protein and peptide level. Therefore, deglycosylation methods have been used to remove carbohydrate moieties either selectively using various enzymes (most commonly PNGase F (EC 3.5.1.52)) or indiscriminately using trifluoromethanesulfonic acid (CAS# 1493-13-6), with the latter being far more efficient (Fryksdale, Jedrzejewski *et al*., 2002).

Related articles

Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5; Article 22, **Two-dimensional gel electrophoresis**, Volume 5; Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5; Article 42, **Membrane-anchored protein complexes**, Volume 5; Article 62, **Glycosylation**, Volume 6; Article 66, **GPI anchors**, Volume 6; Article 74, **Glycoproteomics**, Volume 6; Article 103, **Modeling membrane protein structures**, Volume 6; Article 38, **Transmembrane topology prediction**, Volume 7

References

Adam PJ, Boyd R, Tyson KL, Fletcher GC, Stamps A, Hudson L, Poyser HR, Redpath N, Griffiths M, Steers G, *et al*. (2003) Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *Journal of Biological Chemistry*, **278**(8), 6482–6489.

Altin JG and Pagler EB (1995) A one-step procedure for biotinylation and chemical cross-linking of lymphocyte surface and intracellular membrane-associated molecules. *Analytical Biochemistry*, **224**(1), 382–389.

Celis JE (1998) *Cell biology: A Laboratory Handbook*, Academic Press: San Diego.

Chaney LK and Jacobson BS (1983) Coating cells with colloidal silica for high yield isolation of plasma membrane sheets and identification of transmembrane proteins. *Journal of Biological Chemistry*, **258**(16), 10062–10072.

Engel WD, Schagger H and von Jagow G (1980) Ubiquinol-cytochrome c reductase (EC 1.10.2.2). Isolation in triton X-100 by hydroxyapatite and gel chromatography. Structural and functional properties. *Biochimica et Biophysica Acta*, **592**(2), 211–222.

Fonyo A (1968) Phosphate carrier of rat-liver mitochondria: Its role in phosphate outflow. *Biochemical and Biophysical Research Communications*, **32**(4), 624–628.

Ford T, Graham J and Rickwood D (1994) Iodixanol: A nonionic iso-osmotic centrifugation medium for the formation of self-generated gradients. *Analytical Biochemistry*, **220**(2), 360–366.

Fryksdale BG, Jedrzejewski PT, Wong DL, Gaertner AI and Miller BS (2002) Impact of deglycosylation methods on two-dimensional gel electrophoresis and matrix assisted laser desorption/ionization-time of flight-mass spectrometry for proteomic analysis. *Electrophoresis*, **23**(14), 2184–2193.

Gahmberg CG and Tolvanen M (1996) Why mammalian cell surface proteins are glycoproteins. *Trends in Biochemical Sciences*, **21**(8), 308–311.

Hartinger J, Stenius K, Hogemann D and Jahn R (1996) 16-BAC/SDS-PAGE: A two-dimensional gel electrophoresis system suitable for the separation of integral membrane proteins. *Analytical Biochemistry*, **240**(1), 126–133.

Henningsen R, Gale BL Straub KM and DeNagel DC (2002) Application of zwitterionic detergents to the solubilization of integral membrane proteins for two-dimensional gel electrophoresis and mass spectrometry. *Proteomics*, **2**(11), 1479–1488.

Kimball JW (2004) *Kimball's Biology Pages*. **http://biology-pages.info**.

Luche S, Santoni V and Rabilloud T (2003) Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics*, **3**(3), 249–253.

Lundahl P, Watanabe Y and Takagi T (1992) High-performance hydroxyapatite chromatography of integral membrane proteins and water-soluble proteins in complex with sodium dodecyl sulphate. *Journal of Chromatography*, **604**(1), 95–102.

Macfarlane DE (1983) Use of benzyldimethyl-n-hexadecylammonium chloride ("16-BAC"), a cationic detergent, in an acidic polyacrylamide gel electrophoresis system to detect base labile protein methylation in intact cells. *Analytical Biochemistry*, **132**(2), 231–235.

Norris JL, Porter NA and Caprioli RM (2003) Mass spectrometry of intracellular and membrane proteins using cleavable detergents. *Analytical Chemistry*, **75**(23), 6642–6647.

Pasquali C, Fialka I and Huber LA (1999) Subcellular fractionation, electromigration analysis and mapping of organelles. *Journal of Chromatography. B, Biomedical Sciences and Applications*, **722**(1–2), 89–102.

Peirce MJ, Wait R Begum S, Saklatvala J and Cope AP (2004) Expression profiling of lymphocyte plasma membrane proteins. *Molecular and Cellular Proteomics*, **3**(1), 56–65.

Pertoft H (2000) Fractionation of cells and subcellular particles with Percoll. *Journal of Biochemical and Biophysical Methods*, **44**(1–2), 1–30.

Rahbar AM and Fenselau C (2004) Integration of Jacobson's pellicle method into proteomic strategies for plasma membrane proteins. *Journal of Proteome Research*, **3**(6), 1267–1277.

Rais I, Karas M and Schagger H (2004) Two-dimensional electrophoresis for the isolation of integral membrane proteins and mass spectrometric identification. *Proteomics*, **4**(9), 2567–2571.

Riccio P, Aquila H and Klingenberg M (1975) Purification of the carboxy-atractylate binding protein from mitochondria. *FEBS Letters*, **56**(1), 133–138.

Shin BK, Wang H, Yim AM, Le Naour F, Brichory F, Jang JH, Zhao R, Puravs E, Tra J, Michael CW, *et al*. (2003) Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. *Journal of Biological Chemistry*, **278**(9), 7607–7616.

Simpson RJ, (2003) *Proteins and Proteomics: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Simpson RJ, Connolly LM, Eddes JS, Pereira JJ, Moritz RL and Reid GE (2000) Proteomic analysis of the human colon carcinoma cell line (LIM 1215): Development of a membrane protein database. *Electrophoresis*, **21**(9), 1707–1732.

Tastet C, Charmont S, Chevallet M, Luche S and Rabilloud T (2003) Structure-efficiency relationships of zwitterionic detergents as protein solubilizers in two-dimensional electrophoresis. *Proteomics*, **3**(2), 111–121.

Taylor CM and Pfeiffer SE (2003) Enhanced resolution of glycosylphosphatidylinositol-anchored and transmembrane proteins from the lipid-rich myelin membrane by two-dimensional gel electrophoresis. *Proteomics*, **3**(7), 1303–1312.

Thomas TC and McNamee MG (1990) Purification of membrane proteins. *Methods in Enzymology*, **182**, 499–520.

Valpuesta JM and Barbon PG (1988) A Triton X-100-hydroxyapatite procedure for a rapid purification of bovine heart cytochrome-c oxidase. Characterization of the cytochrome-c oxidase/Triton X-100/phospholipid mixed micelles by laser light scattering. *Biochimica et Biophysica Acta*, **955**(3), 371–375.

Wallin E and von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science*, **7**(4), 1029–1038.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**(3), 242–247.

Wissing J, Heim S, Flohe L, Bilitewski U and Frank R (2000) Enrichment of hydrophobic proteins via Triton X-114 phase partitioning and hydroxyapatite column chromatography for mass spectrometry. *Electrophoresis*, **21**(13), 2589–2593.

Yu YQ, Gilar M and Gebler JC (2004) A complete peptide mapping of membrane proteins: A novel surfactant aiding the enzymatic digestion of bacteriorhodopsin. *Rapid Communications in Mass Spectrometry*, **18**(6), 711–715.

Improvement of sequence coverage in peptide mass fingerprinting

Karin Hjernø and Peter Roepstorff
University of Southern Denmark, Odense, Denmark

1. Introduction

Detailed information regarding the presence of isoforms, splice variants, and posttranslational modifications (PTMs) is essential for the complete understanding of the biological function, location, and properties of a specific protein.

Traditionally, proteomic studies are based on protein separation using two-dimensional (2D) gel electrophoresis followed by identification by mass spectrometry (MS). An alternative strategy based on separation of peptides derived by enzymatic digestion of complex protein mixture using 2D liquid chromatography systems coupled to mass spectrometry has gained extensive use because it allows automation and high throughput. In this approach, each protein is identified on the basis of only a few peptides. However, studies of different isoforms and changes in protein modification require high coverage of the protein sequence and are still best performed after separation of the proteins for example by 2D gels.

Several strategies have been developed for the specific detection of PTMs such as phosphorylation, nitrosylation, glycosylation, and acetylation (Mann and Jensen, 2003; *see also* Article 61, **Posttranslational modification of proteins**, Volume 6). Most of these are designed to identify one specific type of modification by selective enrichment of the modified peptides (for review, see Jensen, 2004).

However, if the aim is to fully characterize a protein and its modifications, a more global strategy for full characterization of proteins is needed. This requires that signals covering all or the majority of the protein sequence are observed, that is, a high sequence coverage is obtained. Although sequence coverage close to 100% might be obtained in peptide mass fingerprinting by Matrix Assisted Laser Desorption/Ionization (MALDI) MS, protein identification is often performed with sequence coverage as low as 10%, and essential peptides may fail to be identified for different reasons. The main cause is selective ion suppression during the mass spectrometric analysis of the complex peptide mixtures. Additional reasons for low sequence coverage are loss of peptides during sample handling and the fact that modified peptides often have a low stochiometri. Consequently, the normal fingerprint procedure has to be carefully optimized before (if ever) a full characterization of the protein can be obtained.

In the following section, we will discuss strategies useful for enhancing the sequence coverage by using complementary information gained by the use of different enzymes for protein cleavage, different methods for enrichment, different matrices, and/or different ionization methods. The strategies described all assume that the proteins are separated, for example, by 2D-PAGE prior to characterization.

2. Enzymatic digestion

Traditionally, trypsin is the enzyme of choice for protein and peptide mass fingerprinting. The masses of tryptic peptides are typically within the range of 700–2500 Dalton (Da), which is optimal for protein identification by peptide mass fingerprinting. However, digestion with a single enzyme will often not provide complete sequence coverage. This is partly due to generation of a number of too large or too small peptides that may not be detected under the chosen conditions.

Especially, lysine-terminated peptides with masses below 1000 Da will often be lost in purification procedure or missing in the spectra owing to suppression effects (see below). However, by using suboptimal digestion conditions, these peptides might be observed as part of larger peptides containing one or more missed cleavage sites.

A more efficient strategy to increase sequence coverage is to combine data from the tryptic digestion with those obtained by digestion with proteolytic enzymes with different specificity. Several examples of successful improvement of the sequence coverage using multiple enzymatic digests are given in the literature (Larsen *et al*., 2001; Choudhary *et al*., 2003; MacCoss *et al*., 2002). Examples of proteases with high specificity are endoproteinases Lys-C, Asp-N, and Glu-C, all of which can be used for in-gel as well as in-solution digestion. The use of specific enzymes can be combined with digestion using less specific enzymes such as subtilisin and chymotrypsin, which generally produce smaller peptides. Inclusion of these enzymes might be an advantage when the specific enzymes produce too large peptides either because of lack of appropriate cleavage sites or because of large modifying groups, which might cause steric hindrance for the enzyme action.

3. Signal suppression effects

Ion suppression is not a fully characterized phenomenon. It is influenced by a combination of presence of impurities, gas-phase affinity of the peptides, choice of matrix in MALDI, and choice of ionization and acquisition method.

3.1. Removal of impurities

The presence of impurities such as salt, buffers, and detergents can result in adducts formation and thereby in reduced signal intensities or can lead to complete suppression of peptide signals. Therefore, desalting/cleanup and concentration of the digest prior to mass spectrometric analysis is crucial, especially when analyzing

low amounts of starting material (e.g., low abundant protein spots). This can be achieved with commercial microcolumns (e.g., ZipTip's from Millipore (Billerica, MA, USA)) or by using small reversed-phase Poros R2 columns made in gel loader tips (Gobom *et al*., 1999). The peptides are bound to the reversed-phase material followed by removal of the contaminants by a washing step and subsequent elution of the peptides directly onto the MALDI target using a matrix solution (Gobom *et al*., 1999).

We have observed that some small hydrophilic and large hydrophobic peptides are lost in this procedure because they are either not retained on the reversed-phase column or not eluted by the matrix solution respectively (Larsen *et al*., 2002; Laugesen and Roepstorff, 2003). The small hydrophilic peptides can be trapped by passing the flow-through from the reverse-phase columns onto a column containing graphite instead of the Poros R2 material, followed again by elution with matrix solution (Larsen *et al*., 2002).

The graphite columns have also been demonstrated to efficiently retain hydrophilic modified peptides such as phosphorylated and glycosylated peptides and thereby improve the chance to identify these modified peptides (Larsen *et al*., 2004).

3.2. Ion suppression/Preferential ionization in MALDI

It is found that tryptic peptides having a C-terminal arginine give higher signal intensity than lysine-terminated peptides, most likely owing to the high proton affinity of the guanidinium group in arginine (Krause *et al*., 1999). Conversion of the lysine residues to the more basic homoarginine by reaction with *O*-methylisourea prior to analysis enhances the signals for lysine-containing peptides, resulting in increased sequence coverage (Brancia *et al*., 2000; Hale *et al*., 2000; Beardsley *et al*., 2000). A recent study by the group of Krause indicates that the presence of phenylalanine, leucine, and proline in the peptide also seems to enhance the desorption/ionization process, resulting in higher signal intensities in a MALDI spectrum, demonstrating that ionization efficiency is a complex and not fully understood phenomenon (Baumgart *et al*., 2004).

3.3. Matrix selection

The choice of matrix also influences the sequence coverage. Thus, different sequence coverages can be obtained when the same digest is analyzed with the three commonly used matrices for peptide/protein fingerprinting: 2,5-dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHCA), and sinapinic acid (SA) (*see* Article 14, **Sample preparation for MALDI and electrospray**, Volume 5). As a consequence, improved sequence coverage can be obtained by combining the results from analysis with different matrices (Gonnet *et al*., 2003; Gobom *et al*., 1999). The use of matrix mixtures has also been reported to increase the sequence coverage compared to the use of a single matrix (Laugesen and Roepstorff, 2003). In addition, these matrix mixtures seem to be more tolerant toward the presence of impurities and thus reduce the need for sample purification.

3.4. Choice of acquisition and ionization method

For a simple protein identification, the MALDI time-of-flight instrument will typically be optimized to yield maximal sensitivity and resolution in the mass range between 700 and 3500 Da, with a highest resolution around 2000 Da. However, by tuning the mass spectrometer (grid voltage and delay time), an improved sensitivity and resolution for the detection of larger peptides at the cost of the smaller can be obtained. Repeated acquisitions with the instrument optimized for different mass regions, therefore, often provide better sequence coverage.

The choice of ionization method also influences which peptides are observed. Thus, our experience, as well as the experience of Kast *et al*. (2003), tells us that only between one-third and one-half of the peptides in a mixture, or even less, are observed in common for Electrospray Ionization (ESI) and MALDI MS. By combining the results from the two ionization methods, nano-ESI and MALDI, considerable improvements of the sequence coverage in protein fingerprinting can be obtained.

3.5. Liquid Chromatography Mass Spectrometry, LCMS

The use of a separation step prior to mass spectrometric analysis overcomes many of the above-mentioned shortcomings when analyzing peptide mixtures. The prepurification step is integrated in the LCMS (Liquid Chromatography Mass Spectrometry) procedure. By separating the individual peptides before reaching the mass spectrometer, suppression effects are reduced or nonexistent and, consequently, sequence coverage close to 100% should be obtainable provided that all the peptides are retained on and subsequently eluted from the chromatographic column, which is not always the case (see above). Until recently, LCMS was only available on ESI instruments. However, off-line LC-MALDI-MS has gained increasing interest, and unpublished data from several groups indicate that this method results in even better sequence coverage than LC-ESI-MS.

4. Interactive data handling

For the description above, it is obvious that complete sequence coverage including observation of peptides present in substochiometric amounts is highly dependent on the complexity of the peptide mixture. Consequently, a prerequisite for complete coverage of the protein sequence including minor variants is that the proteins of interest are isolated prior to analysis (e.g., by 2D gels) and that contaminating proteins such as keratins are eliminated as well as the enzyme used are high quality with minimal autodigestion. Alternatively, exclusion list containing all the masses of known contaminating peptides has to be developed before or during the experiments (using either the program PeakErazor (**http://welcome.to/GPMAW**) or similar software package) to avoid wasting sample material on identifying peptides belonging to contaminants or peptides already identified.

As soon as the protein is identified, interactive data handling needs to be performed. By making an in silico digest of the protein after identification, the nondetected peptides can be examined for their chemical/physical properties such as hydrophobicity/hydrophilicity, acidity/basicity, predicted modifications, and m/z-values. On the basis of this information, a strategy for the detection of these peptides or their potentially modified forms can be designed using the approaches suggested for improving sequence coverage described above.

The interactive data handling can be combined with the concept of hypothesis-driven multistage MS introduced by Kalkum *et al*. (2003) in which they calculate the mass for peptides predicted to be present, but only in trace amounts, for example, possible modified peptides or peptides representing isoforms and splice variants. By performing MS/MS or MS^3 in a MALDI-ion-trap after selecting the appropriate mass-to-charge value, they have been able to identify peptides for which the signals in the mass spectra were suppressed or hidden in the noise.

Off-line LC-MALDI, in contrast to on-line LC-ESI, offers the possibility to "freeze" the sample "in time", which again allows evaluation of the data, use of interactive data handling, and hypothesis-driven multistage MS. However, even without an LC system, a considerably increased sequence coverage can be obtained using interactive data handling combined with the optimization strategies described in this chapter.

5. Conclusion

Here we have presented some approaches that can enhance sequence coverage. The strategy described here is time consuming and not easy to automate. And even using this, complete characterization of a protein might not be possible at a proteomics sensitivity level especially because modified peptides of low stoichiometry might escape detection. The challenge in the future will be to quantify the degree of modification for a specific peptide. This will necessitate strategies for quantitative comparison between two peptides with totally different properties, the modified and nonmodified peptides. We expect that it will also be possible to solve this difficult task in the future.

References

Baumgart S, Lindner Y, Kuhne R, Oberemm A, Wenschuh H and Krause E (2004) The contributions of specific amino acid side chains to signal intensities of peptides in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **18**, 863–868.

Beardsley RL, Karty JA and Reilly JP (2000) Enhancing the intensities of lysine-terminated tryptic peptide ions in MALDI mass spectrometry. *Rapid Communications in Mass Spectrometry*, **14**, 2147–2153.

Brancia FL, Oliver SG and Gaskell SJ (2000) Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Communications in Mass Spectrometry*, **14**, 2070–2073.

Choudhary G, Wu SL, Shieh P and Hancock WS (2003) Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *Journal of Proteome Research*, **2**, 59–67.

Gobom J, Nordhoff E, Mirgorodskaya E, Ekman R and Roepstorff P (1999) Sample purification and preparation technique based on nano-scale reversed-phase columns for the sensitive analysis of complex peptide mixtures by matrix-assisted laser desorption/ionization mass spectrometry. *Journal of Mass Spectrometry*, **34**, 105–116.

Gonnet F, Lemaître G, Waksman G and Tortajada J (2003) MALDI/MS peptide mass finger printing for proteome analysis: identification of hydrophobic proteins attached to eucaryote keratinocyte cytoplasmic membrane using different matrices in concert. *Proteome Science*, **6**, 1–7.

Hale JE, Butler JP, Knierman MD and Becker GW (2000) Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine. *Analytical Biochemistry*, **287**, 110–117.

Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, **8**, 33–41.

Kalkum M, Lyon GJ and Chait BT (2003) Detection of secreted peptides by using hypothesis-driven multistage mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 2795–2800.

Kast J, Parker CE, van der Drift K, Dial JM, Milgram SL, Wilm M, Howell M and Borchers CH (2003) MALDI-directed nano-ESI-MS/MS analysis for protein identification. *Rapid Communications in Mass Spectrometry*, **17**, 1825–1834.

Krause E, Wenschuh H and Jungblut PR (1999) The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Analytical Chemistry*, **71**, 4160–4165.

Larsen MR, Cordwell SJ and Roepstorff P (2002) Graphite powder as an alternative or supplement to reversed-phase material for desalting and concentration of peptide mixtures prior to MALDI mass spectrometry. *Proteomics*, **2**, 1277–1287.

Larsen MR, Graham ME, Robinson PJ and Roepstorff P (2004) Improved detection of hydrophilic phosphopeptides using graphite powder microcolumns and mass spectrometry: evidence for in vivo doubly phosphorylated dynamin I and dynamin III. *Molecular & Cellular Proteomics*, **3**, 456–465.

Larsen MR, Larsen PM, Fey SJ and Roepstorff P (2001) Characterization of stress induced processing of enolase 2 from *Saccharomyces cerevisiae* by 2-D gel electrophoresis and mass spectrometry. *Electrophoresis*, **22**, 566–575.

Laugesen S and Roepstorff P (2003) Combination of two matrices results in improved performance of MALDI MS for peptide mass mapping and protein analysis. *Journal of the American Society for Mass Spectrometry*, **14**, 992–1002.

MacCoss MJ McDonald WH Saraf A Sadygov R Clark JM Tasto JJ Gould KL Wolters D Washburn M and Weiss A (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7900–7905.

Mann M and Jensen ON (2003) Proteomic analysis of post-translational modification. *Nature Biotechnology*, **21**, 255–261.

Tutorial on tandem mass spectrometry database searching

Jimmy K. Eng
Institute for Systems Biology, Seattle, WA, USA

1. Introduction

With recent advances in search software and available databases, a novice can now easily perform a tandem mass spectrometry (MS/MS) sequence database search. Such a search requires an input spectrum, a sequence database to query, and a set of search parameters to guide the search. This tutorial will cover the mechanisms of performing an MS/MS search using the Mascot (Perkins *et al*., 1999) search engine for demonstration purposes.

Figure 1 illustrates the general process of sequence database searching. In general, an acquired tandem mass spectrum of a peptide is compared against theoretical tandem mass spectra of peptides generated in silico from the specified sequence database. The methods by which they are compared, that is, the mathematical scoring algorithms, vary widely between different database search engines such as Mascot, MS-Tag (Clauser *et al*., 1999), or SEQUEST (Eng *et al*., 1994; *see also* Article 3, **Tandem mass spectrometry database searching**, Volume 5). Nevertheless, the different search engines typically produce quite similar results, and the mathematical "behind the scenes" activity does not affect the fact that there are many features that are common and applicable to running a database search on most search engines.

2. Example spectrum to be interpreted

Figure 2 shows a raw, uninterpreted tandem mass spectrum of a peptide that will be used as the input for this exercise. The information derived from the input spectrum is the mass and intensity pairs that represent each peak in the fragmentation spectrum, the mass (1157 a.m.u.) of the precursor ion selected for MS/MS, and the charge state of the precursor ion, if known. The above spectrum was acquired on an ion trap mass spectrometer. The mass spectrometer used for acquisition determines the errors associated with the mass measurements and thus the mass tolerances to be used in the database search. Additionally, the instrument also determines the predominant peptide fragmentation products produced, thus guiding the selection of fragment ion types to consider in the search parameters.

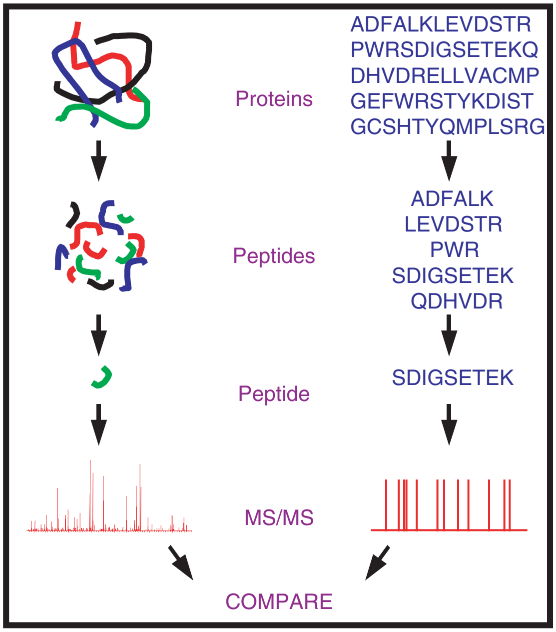


Figure 1 A schematic showing the concept of MS/MS database searching

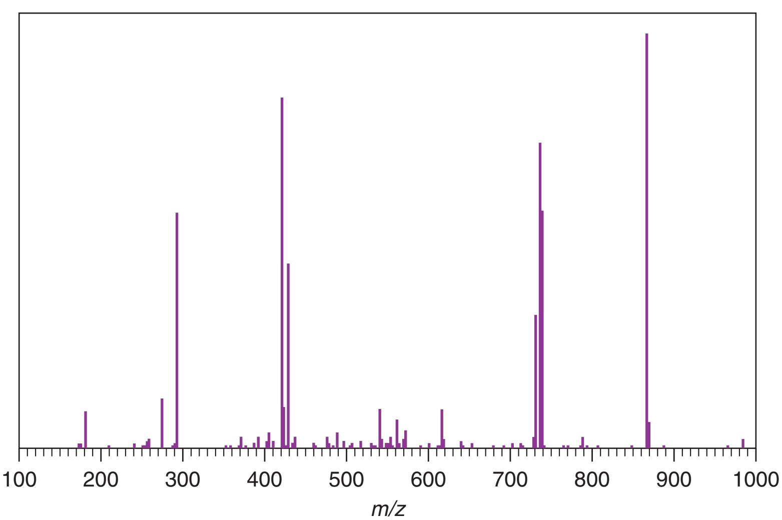


Figure 2 Example uninterpreted tandem mass spectrum. The displayed peak list, precursor ion charge state, and precursor or peptide mass are the input for an MS/MS database search

Any additional knowledge available about the input sample can enhance the potential success of the database search. Specifically, awareness of the organism from which the sample derives permits querying against a species-specific sequence database. Also, because the fragmentation spectra are from peptides, the protease used to digest the intact protein(s) is a valuable information that can be used either as part of the database search parameters, and thus part of the search, or as part of the validation of the search results. Finally, any process that changes the nature of amino acids in the sample such as metabolic labeling or chemical side chain modification, such as cysteine alkylation, must be incorporated into the search parameters (see below).

3. Search parameters

In our current example, there is no suspicion that the peptide is posttranslationally modified; thus, no modifications are specified in the initial query. If processing steps modify the sample, such as cysteine alkylation or metabolic labeling that incorporates an isotopically heavy amino acid, then these modifications should be specified in the search parameters. Modifications can typically be applied to any amino acid and/or the amino or carboxy terminus of a peptide.

Modifications typically are of two types. The first is the static modification that modifies all occurrences of a residue as would be expected with a covalent reaction that goes to completion or a metabolic labeling in which all residues are expected to be replaced. A second type of modification is the variable modification. This type of modification forces the search engine to look for two different forms of an amino acid. A common example is the search for phosphorylation where only a few percent of serine, threonine, or tyrosine residues are actually phosphorylated. In this case, the search is conducted allowing these residues to be either modified (+80 a.m.u.) or unmodified. Note that specifying this type of modification in the search parameters can significantly prolong search times.

The peptide associated with the spectrum in Figure 1 is from *Bos taurus* (cow), so this spectrum should ideally be searched against a database composed of *B. taurus* protein sequences. However, if the sample's species is not known or if the genome of that organism is poorly covered, then a more comprehensive sequence database composed of all species can be substituted, with the hope of finding an identical peptide in a related species. On-line search engines, such as Mascot or MS-Fit, supply a fixed set of sequence databases available for querying against. With a local search engine installation, both publicly available and proprietary sequence databases can be used for the database search.

Figure 3 displays a screen capture of the search parameters defined for this database search. As no cow-specific sequence database is available on the public Mascot search engine, a mammalian subset of the SwissProt (Bairoch and Apweiler, 1997) sequence database is selected. Trypsin is specified in the Enzyme field since trypsin was used to digest the original sample. This specifies that only tryptic peptides of this weight (1157 a.m.u.) in the database will be compared rather than all peptides of this weight. This will dramatically shorten the time required to complete a search. Neither static nor variable modifications are specified. An ESI-TRAP

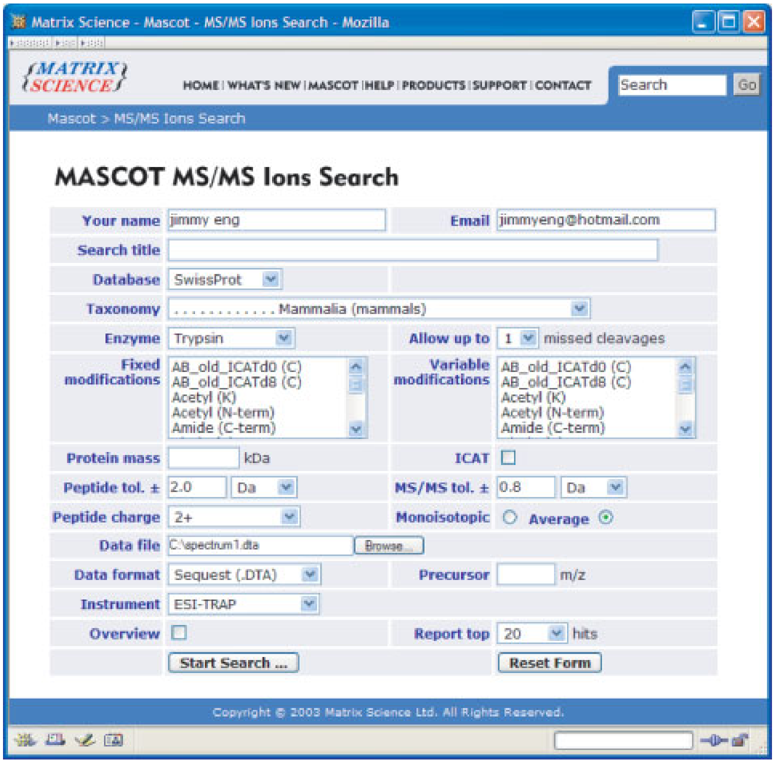


Figure 3 Example Mascot search parameters page. Primary parameter settings are the selection of the sequence database to query, modifications (if any), and acquisition instrument selection

is specified for the Instrument field; the selection of the instrument defines the fragment products to consider in the analysis. Mass tolerances of 2.0 a.m.u. and 0.8 a.m.u. are used for the peptide mass and fragment ion masses respectively. The performance of the mass spectrometer defines these mass tolerances used for the search.

4. Search results

The results of the search are shown in Figure 4. A description of the scoring algorithm and significance determination is beyond the scope of this tutorial but the Mascot search results show no positive identifications because the top ranked peptide, ADSVGKLLTVR received a probability-based MOWSE score of only 24. This score is below the suggested homology and identity significance thresholds of 32 and 41, respectively, for this particular input spectrum. In addition to having a nonsignificant score, the top ranked peptide is from a protein from *Mus musculus* (MCM7_MOUSE), whereas the input sample was from *B. taurus*.

There are many potential explanations and remedies for incorrect identifications from unsuccessful database searches. The correct sequence may not be in the

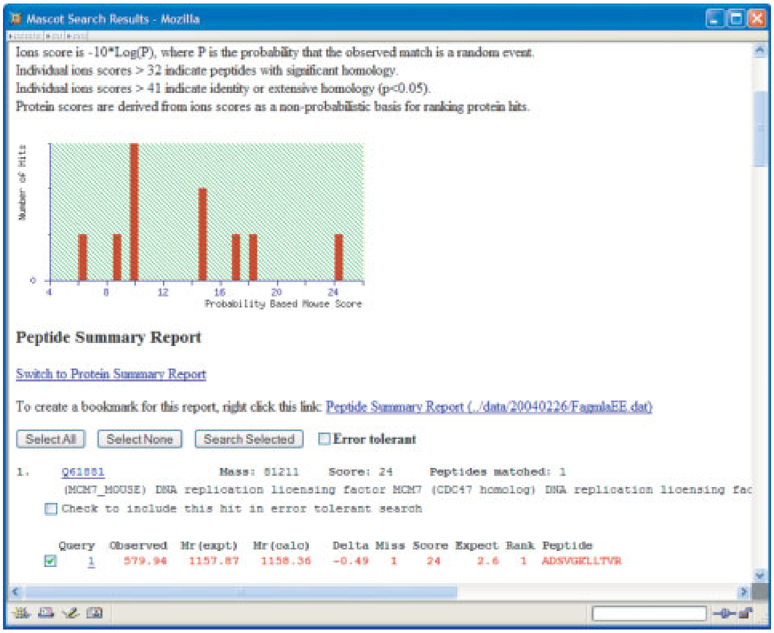


Figure 4 Mascot search results page. The top ranked peptide, ADSVGKLLTVR, was from the wrong organism and had a poor score, indicative of an incorrect identification

database that was queried; searching a larger or different sequence database might resolve the problem. The peptide could be posttranslationally modified, in which case adding common modifications might be considered in a subsequent search. It is possible that mass tolerances have been set too restrictively narrow with the effect of screening out the correct peptide. In this case, the database queried, lack of specified modifications, and narrow mass tolerances were not the culprits for the incorrect identification. The reason the spectrum failed to be identified correctly was due to the specification of the tryptic enzyme constraint. Even though the original sample was processed with trypsin to generate peptides, nontryptic peptides can exist either because of incomplete digestion, contaminating proteolysis, or in-source fragmentation. When removing the enzyme constraint, Mascot identified the peptide YQEPVLGPVR, a fragment tryptic at only the C-termini, from the protein bovine carbonic anhydrase (CASB_BOVIN). The peptide was identified as correct with a score of 72, compared to the suggested homology and identity thresholds of 49 and 61, respectively, as shown in Figure 5.

Figure 6 shows the overlay of calculated versus actual fragment ions for the peptide YQEPVLGPVR. This is a common view when validating search results because the matched spectrum should account for most of the peaks in the MS/MS spectrum (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5). However, this validation process can be difficult and extremely subjective, depending on the quality of the input spectra and the experience of the user. One strategy to facilitate the validation process is to use background information on the input sample as a validation tool rather than a search constraint.

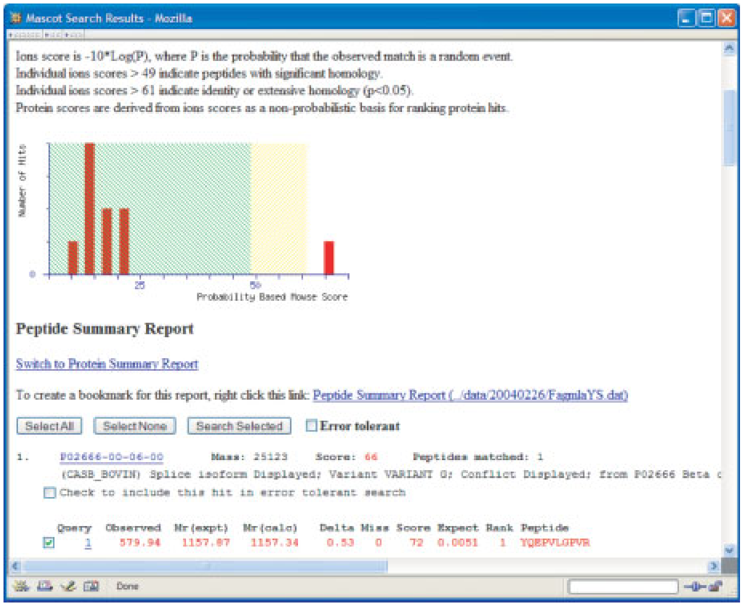


Figure 5 Mascot search results page. After modifying the search parameters to allow for peptides resulting from unexpected cleavage, the top ranked peptide, YQEPVQPVR, had a significant score and was identified from the correct organism

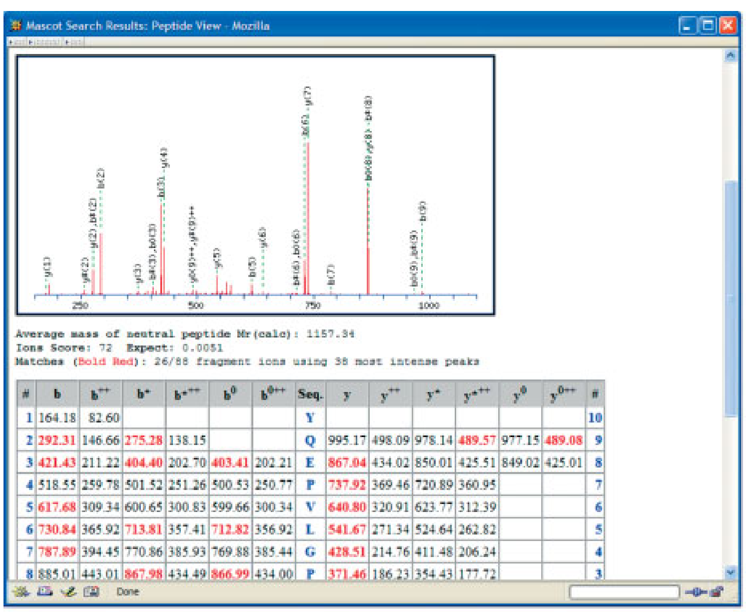


Figure 6 A typical spectral display view shows the extent of the match between input spectrum and identified peptide. Note how all of the major peaks are successfully accounted for

For example, the spectrum was identified out of a much larger database composed of all mammalian sequences as opposed to just searching a (smaller) bovine sequence database. The identification of a bovine protein supports the validity of the identification. Additionally, searching without an enzyme constraint can help the validation process because any peptide string from the sequence database can be analyzed (and not just the supposedly tryptic peptides that were generated in the protein digestion process). For an enzyme unconstrained search, high-scoring search results showing the expected tryptic or partially tryptic termini (as in this case) add confirmatory evidence.

5. Conclusions

Researchers new to the field who are interested in exploring proteomics should not be daunted by the complexity of tandem mass spectrometry. MS/MS database searching is an extremely powerful tool in the analysis of both simple and complex protein samples that couples the explosion of available genome sequence information with improvements in mass spectrometry instrumentation for rapid peptide and protein identification. With the efficiency of modern instrumentation and software, it is extremely easy to acquire and search a significant amount of MS/MS spectra, with the effect of shifting the burden of analysis to the validation process. As shown, the steps in performing an MS/MS database search are simple and straightforward. Likewise, search software has become user-friendlier to the neophyte. However, to achieve optimal results, new entrants to these techniques must educate themselves on the choices they make in both running a search (i.e., how search parameters impact the search results) and validating the search results.

References

Bairoch A and Apweiler R (1997) The Swiss-Prot protein sequence database: its relevance to human molecular medical research. *Journal of Molecular Medicine*, **75**, 312–316.

Clauser KR, Baker P and Burlingame AL (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **14**, 2871–2882.

Eng JK, McCormack AL and Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.

Perkins DN, Pappin DJ, Creasy DM and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Basic Techniques and Approaches

Techniques for ion dissociation (fragmentation) and scanning in MS/MS

Jack Throck Watson
Michigan State University, East Lansing, MI, USA

O. David Sparkman
University of the Pacific, Stockton, CA, USA

1. Introduction

Modern mass spectrometry techniques such as electrospray ionization (ESI) (*see* Article 10, **Hybrid MS**, Volume 5, Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5, and Article 75, **Mass spectrometry**, Volume 6) and atmospheric pressure chemical ionization (APCI) often only produce ions representing the intact molecule, usually protonated molecules (MH^+). Even when the mass of these ions is measured to an accuracy that will allow for a determination of the elemental composition, the identity of the molecule's structure is not necessarily forthcoming. Sometimes structural features of the ion representing the intact molecule can be deduced from the array of fragments formed. Both the mass of the fragments and the mass represented by the neutral loss that occurs when the ion fragments (the dark matter of the mass spectrum) are important in this process. Ions representing the intact analyte molecule produced by ESI and APCI have little or no internal energy to cause fragmentation. Therefore, it is necessary to energize these ions in order to gain the desired information.

Once formed, an ion can be made to dissociate by increasing its internal energy by any of several modes of excitation; these include collision-activated dissociation (CAD) (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5) or collisionally induced dissociation (CID), surface-induced dissociation (again, involving a collision), photo-dissociation, and others. In this introduction, only CAD (same process as CID) will be considered. In CAD or CID, instrumental arrangements are available to control collisions between ions and molecules of a gas, for example, He, Ar, N_2, or any other inert gas. In some cases, the ion-molecule collisions are promoted in a collision cell, which is a relatively small volume within the instrument where the gas is confined, or it can be done in a large primary segment of the mass spectrometer. The extent to which the internal energy of an ion can be increased depends on the kinetic energy of the ion and the

relative masses of the ion and the gaseous molecule involved in the collision; this matter will not be addressed in this chapter.

In CAD, ions of a particular *m/z* value are selected (precursor ions) to collide with the collision gas molecules. Some instrumental provision is also available to determine the *m/z* values of the dissociation products (product ions) of these ion/molecule collisions, that is, there is one mass spectrometric means for selecting a precursor ion, and a second mass spectrometric means for analyzing the secondary ions (the product ions); thus, the overall instrumental technique is called mass spectrometry/mass spectrometry (MS/MS). The technique might consist of two separate mass spectrometers, "MS/MS-in-space", or it might be accomplished in a fixed coordinate of space, in which case, it is called "MS/MS-in-time"; these two modes of operation will be described in detail. There are several different operational or "scan modes" used in MS/MS analyses, for example, product-ion scan, precursor-ion scan, and neutral-loss scan; these different modes of scanning will be described together with the different types of information they provide.

Mass spectrometers determine the mass-to-charge ratio (*m/z*) of an ion (*see* Article 75, **Mass spectrometry**, Volume 6). Because mass is the principal parameter of interest, only ions of unit charge will be considered in this introductory description of MS/MS techniques. The technique of ion–ion collisions or reactions has been introduced recently as a means of reducing the charge state of multiple-charge ions; this technique will not be elaborated on here.

2. Instrumental configurations

2.1. MS/MS-in-space

Perhaps the simplest way to visualize MS/MS is to consider two mass spectrometers connected in tandem, whereby ion current resulting from ions of a specific *m/z* value (or values) selected by the first mass spectrometer passes into a collision cell, where CAD occurs; ion current, resulting from fragmentation (into product ions) as well as from residual precursors ions, passes out of the collision cell into the second mass spectrometer, where it is analyzed according to individual *m/z* values as recorded by a detector at the end of the second instrument. The triple-quadrupole mass spectrometer illustrated in Figure 1 is a manifestation of this MS/MS-in-space technology. The ions are selected in the first quadrupole (Q_1) mass spectrometer

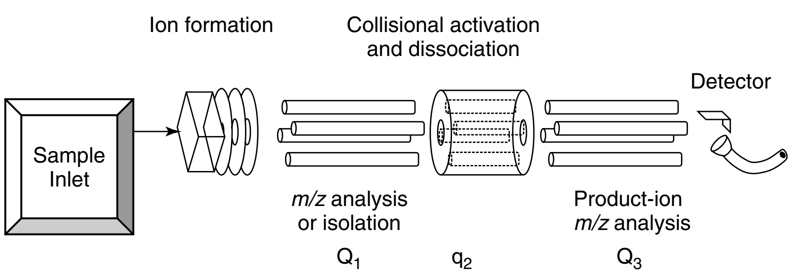


Figure 1 Schematic diagram of a triple-quadrupole mass spectrometer

(this instrument is not scanned), then passed to the second quadrupole (q_2), which is not mass-selective, that is, it is not a mass spectrometer, but rather it provides a means to "refocus" scattered ions in the collision cell where some of the precursor ions are transformed by CAD into product ions; all the ions leave the collision cell and enter the third quadrupole (Q_3) mass spectrometer where they are *m/z*-analyzed. These instruments are called *triple quadrupoles* because the originally developed concept used three quadrupole devices (*see* Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5). In modern instruments that still use a quadrupole *m/z* analyzer for Q_1 and Q_3, the device constituting q_2, the collision cell, is something other than a quadrupole; it is usually a hexapole or octupole device or, as in the case of more recent instrumentation, a lens stack.

2.2. MS/MS-in-time

An increasingly common technique is to perform selection, dissociation, and analysis of secondary ions as a function of time in the same coordinates of space within a mass spectrometer. The quadrupole ion trap (QIT) (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) mass spectrometer has become a popular instrument for conducting MS/MS-in-time; the ion cyclotron resonance mass spectrometer can also be used for this purpose, using technology commonly called FTMS (*see* Article 5, **FT-ICR**, Volume 5). An important limitation of MS/MS-in-time is referred to as the "low-mass cutoff". Product ions that have an *m/z* value of less than approximately one-third that of the precursor ion cannot be stored while CAD of the precursor ion is being carried out. An advantage of MS/MS-in-time over MS/MS-in-space is that only ions with the *m/z* value of the precursor ion are collisionally activated, meaning that secondary fragmentation is less likely to take place than in the collision cell of the MS/MS-in-space configuration, where product ions are also likely to undergo collisional activation and secondary dissociation.

3. Different "ion-scanning modes" for MS/MS with CAD

3.1. Product-ion scan

Probably the most commonly employed technique for MS/MS is one in which a particular precursor ion is subjected to CAD to obtain a product-ion mass spectrum to serve as a "fingerprint"; this MS/MS-in-space mode of operation is carried out essentially as described above. The MS/MS-in-time mode of operation in a quadrupole ion trap begins by transferring all the ions generated by a given sample into the trap, then adjusting the rf/dc parameters of the instrument to eject ions of all *m/z* values except those of the precursor. The normal operating pressure of the QIT is about the same as that of the collision cell in the triple quadrupole (0.13 Pa); therefore, if ions corresponding to the *m/z* value of the precursor ion are then excited, they will undergo collisions with the gas molecules in the trap and dissociate. The second step of the analysis begins by adjusting the rf/dc parameters

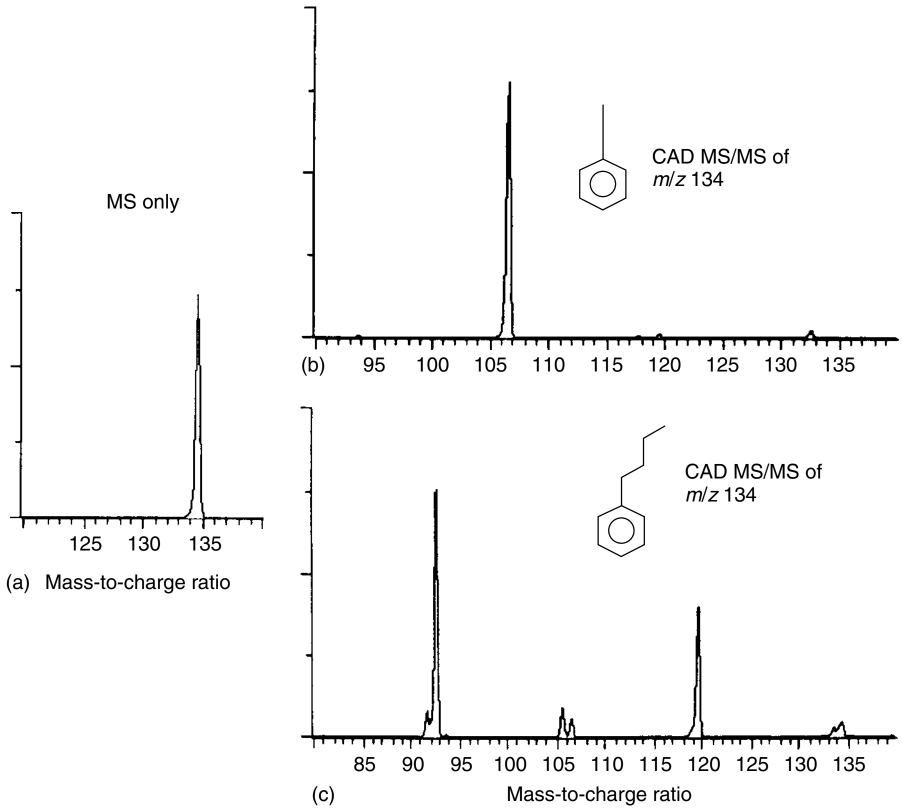


Figure 2 (a) Conventional mass spectrum (no CAD) of either *t*-butyl benzene or *n*-butyl benzene (each has a molecular ion peak at *m/z* 134 and no significant fragment ion peaks). (b) and (c) CAD MS/MS spectra of ion current at *m/z* 134 from *t*-butyl benzene and *n*-butyl benzene, respectively

of the instrument to accommodate ions through a large range of *m/z* values even though the precursor ions are the only ones in the trap; in this way, any nascent product ions produced during CAD of the precursor ions will be trapped as they are produced. The third step consists of a mass-selective ejection scan of the rf/dc parameters to obtain a mass spectrum of the product ions.

The power of the product-ion scan is demonstrated in a comparison of Figure 2(b) and (c), each of which is a product-ion spectrum of a precursor ion of *m/z* 134. Clearly, the pattern of peaks in the two product-ion mass spectra is different. In this way, the product-ion spectrum serves as a fingerprint for each of two compounds, each of which is capable of generating a precursor ion of *m/z* 134. Obviously, the ion of *m/z* 134 has a different structure in each case, which is the basis for the CAD product-ion spectra being different.

3.2. *Precursor-ion scan*

In this scan mode, the second mass spectrometer in the MS/MS-in-space configuration is held at a constant parameter (not scanned), while the first mass spectrometer

is scanned through a specified range of *m/z* values. In this way, an array of precursor ions can be assessed for their capacity to generate a product ion of a specific *m/z* value. An application of a precursor-ion scan is illustrated with the spectrum in Figure 3(b). From previous work, it was known that CAD of protonated peptides containing tyrosine produces the immonium ion of tyrosine at *m/z* 136, that is, the generation of such a product ion can be used as a marker for the presence of tyrosine in a peptide. In this application, the tryptic digest was analyzed by ESI CAD MS/MS by scanning the first mass spectrometer, while holding the second mass spectrometer constant at *m/z* 136; in this mode of operation, the detector produces a signal at a given *m/z* value only when the corresponding precursor ion in the scan of the first mass spectrometer is capable of producing an ion of *m/z* 136 in the collision cell. The complexity of the mixture in this application is represented in Figure 3(a), which is a mass spectrum obtained from analysis of the digest by direct infusion ESI-MS. Notice that the spectrum obtained during ESI-MS/MS with precursor scanning for *m/z* 136 (Figure 3b) is much simpler indicating that only four of the tryptic peptides (T4–T6, T14, and T16) contain tyrosine.

3.3. Neutral-loss scan

Under CAD conditions, some precursor ions are capable of expelling a molecule (the neutral dark matter), such as CO_2, methanol, and so on, in forming the product ion (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5). This phenomenon of neutral loss can be exploited in detecting members of a family of such compounds during the analysis of a complex mixture by CAD MS/MS using the neutral-loss scan. The neutral-loss scan is accomplished by arranging a simultaneous scan of both mass spectrometers (one before, the other following the collision cell, where CAD takes place) offset from one another by a difference in *m/z* values equal to the mass of the molecule expelled from the precursor ions of interest. An application of a neutral-loss scan is illustrated in Figure 3(c). From previous work, it was known that double-protonated peptides containing methionine eliminated CH_3SH (48 Da) during CAD. The spectrum in Figure 3(c) indicates that only two of the tryptic peptides, namely, T2 and T11, among the 20 or so peptides represented in the top spectrum from nonspecific analysis of the digest contain methionine; the data for Figure 3(c) were recorded by scanning Q_1 and Q_3 offset by 24 *m/z* units (mass of 48 Da divided by charge of 2).

It is important to note that the precursor-ion scan and the neutral-loss scan can only be carried out using MS/MS-in-space. These two modes of operation are not possible with MS/MS-in-time. When one of the *m/z* analyzers in an MS/MS-in-space instrument is a nonscanning mass spectrometer, such as a time-of-flight (*see* Article 10, **Hybrid MS**, Volume 5) *m/z* analyzer, the neutral-loss scan mode is not possible because the second analyzer does not scan and, therefore, cannot have a scan offset.

Another important mode of operation for MS/MS is the selected reaction monitoring (SRM) mode. This mode is used for quantitation and specificity. Q_1 is held at a fixed value to allow precursor ions of the specified *m/z* value to pass into the collision cell (q_2). Q_3 is held at a fixed *m/z* value to allow only product

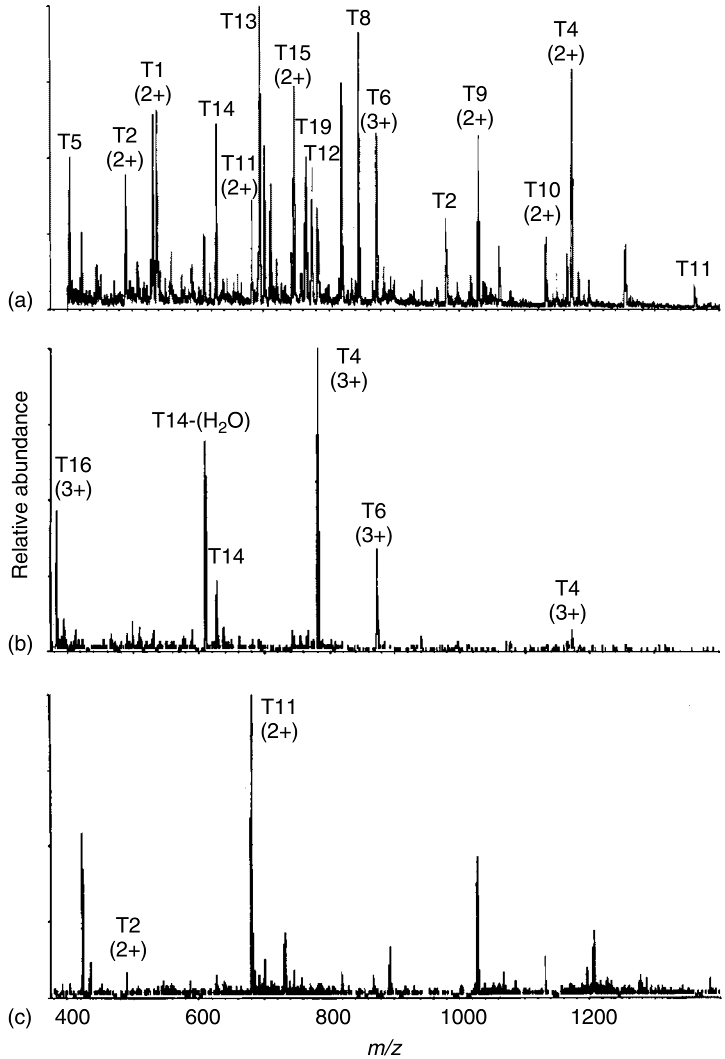


Figure 3 Two examples of results from specialized uses of CAD MS/MS spectra. (a) Conventional electrospray mass spectrum (MS only, no CAD MS/MS) resulting from infusion (no chromatography) of a mixture of tryptic peptides into an electrospray ionization triple-quadrupole mass spectrometer; numbers in parentheses indicate charge state of the corresponding ion. (b) Precusor-ion mass spectrum of the same mixture acquired under conditions allowing detection of only those precursors capable of generating a product ion of *m/z* 136 (immonium ion of tyrosine = 136 Da). (c) Neutral-loss mass spectrum of the same mixture acquired at an offset in *m/z* value between Q_1 and Q_3 to allow detection of only those precursors capable of expelling CH_3SH (48 Da) indicating the presence of methionine in the side chain. (Reprinted by permission of Springer-Verlag, 2001)

ions of a specific mass to reach the detector. In this way, specificity is achieved because only ions that undergo a transition from *m/z* A to *m/z* B in the collision cell will reach the detector; peak area integrated from the product-ion signal provides quantitative information.

4. Summary

CAD can provide characteristic, if not unique, fragmentation (into product ions) of a precursor ion. A variety of scan modes is available with MS/MS instrumentation to permit selective detection of certain compounds depending on the CAD behavior of the precursor ions.

Further reading

Busch KL, Glish GL and McLuckey SA (1988) *Mass Spectrometry/Mass Spectrometry*, VCH Publishers: New York.

De Hoffmann E (1996) Tandem mass spectrometry: a primer. *Journal of Mass Spectrometry: JMS*, **31**, 129–137.

Despeyroux D and Jennings KR (1994) Collision-induced dissociation. In *Biological Mass spectrometry, Present and Future*, Matsuo T, Caprioli RM, Gross ML and Seyama Y (Eds), Wiley: Chichester, pp. 227–238 (ISBN:0471938963).

Gross ML (2000) Charge-remote fragmentation: an account of research on mechanisms and applications. *International Journal of Mass Spectrometry*, **200**, 611–624.

Marcos J, Pascual JA, de la Torre X and Segura J (2002) Fast screening of anabolic steroids and other banned doping substances in human urine by gas chromatography/tandem mass spectrometry. *Journal of Mass Spectrometry: JMS*, **37**, 1059–1073.

McClellan JE, Quarmby ST and Yost RA (2002) Parent and neutral loss monitoring on a quadrupole ion trap mass spectrometer: screening of acylcarnitines in complex mixtures. *Analytical Chemistry*, **74**, 5799–5806.

McLafferty FW (Ed.) (1983) *Tandem Mass Spectrometry*, Wiley: New York.

McLafferty FW (1993) 25 years of MS/MS. *OMS, Organic Mass Spectrometry*, **28**, 1403–1406.

McLuckey SA and Stephenson JL Jr (1998) Ion/ion chemistry of high-mass multiply charged ions. *Mass Spectrometry Reviews*, **17**, 369–407.

Murrell J, Despeyroux D, Lammert SA, Stephenson JL and Goeringer DE (2003) "Fast excitation" cid in a quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry*, **14**, 785–789.

Ospina MP, Powell DH and Yost RA (2003) Internal energy deposition in chemical ionization/tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, **14**, 102–109.

Reid GE and McLuckey SA (2002) 'Top down' protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry: JMS*, **37**, 663–675.

Steen, H and Mann, M (2002) Analysis of bromotryptophan and hydroxyproline modifications by high-resolution, high-accuracy precursor ion scanning utilizing fragment ions with mass-deficient mass tags. *Analytical Chemistry*, **74**, 6230–6236.

Making nanocolumns and tips

Claire Delahunty and John R. Yates III
The Scripps Research Institute, La Jolla, CA, USA

The successful production of nanocolumns and tips is vital to the success of an in-line multidimensional LC-MS/MS analysis (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5). Making nanocolumns for LC/LC-MS/MS is a two-step process: first, fused-silica microcapillary tubing is heated and pulled to produce a column with a tip diameter of 2–5 µm, then the column is packed with chromatographic packing material. The choice and volume of packing materials can vary depending on the complexity of the sample, the volume of the sample, and the goal of the separation, but the optimal diameter of the column tip is the same for all electrospray applications. Typically, the flow rate required for the direct analysis of proteins with femtomole scale sensitivity is 200–300 nL min^{-1}. To achieve this, nanocolumns should have a 50–100 um ID, with a tip ID of approx. 5 um. Columns with smaller IDs can clog easily; those with larger IDs result in less-efficient electrospray.

Nanocolumns may be reproducibly pulled using a commercial CO_2 laser puller (Sutter Instruments, Novato, CA), but prepulled columns can also be purchased (i.e., New Objective, Woburn, MA). By pulling the capillary at the midpoint of its length, two nanocolumns are produced from one 50–54 cm length of capillary tubing. Before placing the capillary in the puller, the polyimide coating must be burned off a 1-in. center section using an alcohol burner. Bunsen burners should not be used since they burn hotter and can seal the capillary. If the column is not properly pulled, packing material will not load smoothly and chromatography will be poor. The parameters for pulling optimal tips vary from instrument to instrument and need to be determined experimentally. Some typical values are listed in Table 1. A detailed discussion of these parameters and their role in capillary construction is given in the Sutter Instrument P-2000 manual.

Packing materials are loaded onto the nanocolumns through the use of a stainless steel pressurization device or "bomb" (The Scripps Research Institute, and Cytopea, Inc., see Figure 1). This bomb is attached to a helium tank with a high-pressure line and valve. The bomb has a removable lid and the lid and the base each has a groove that holds a viton O-ring to ensure a high-pressure seal when the lid is tightened down. The lid is tightened to the base with five bolts. A Swagelock fitting is located in the center of the lid. This fitting holds a Teflon ferrule that allows the capillary to be inserted down into the bomb and into the microfuge tube. The Teflon ferrule can be tightened to hold the capillary in place and provide a high-pressure

Table 1 Typical parameters for pulling a nanocolumn from 100 × 365-μm fused-silica capillary tubing

	Heat	Filament	Velocity	Delay	Pull
1	270	0	30	128	0
2	320	0	40	200	0
3	310	0	30	200	0
4	290	0	20	200	0



Figure 1 Pressure loading device that can be used to pack nanocolumns and to load samples onto nanocolumns

seal. The packing material is suspended in methanol and is placed in a microfuge tube that stands upright in the center of the bomb.

For analyzing simple mixtures of proteins, a one-dimensional nanocolumn can be used. In this case, 10–15 cm of C_{18} reversed-phase (RP) packing material is loaded into the column. However, for more complex mixtures, a biphasic column is employed in which a 7-cm section of RP material is flanked upstream by 3 cm of strong cation exchange (SCX) resin. This configuration allows for multidimensional separation of peptides because peptides are loaded onto and bound to the SCX resin and sequentially "stepped" off with salt pulses, followed by reverse-phase separation of each subset of peptides. Many samples contain competing salts that interfere with the interaction of the peptides with the SCX resin. In these

cases, off-line desalting can be performed using a solid-phase extraction column. Alternatively, peptides can be desalted online using a second phase of reverse-phase packing material directly upstream from the SCX. In this method, peptides can be loaded directly onto the column, desalted in the first cycle of the analysis and subjected to multidimensional separation in the subsequent steps. Comparison of single-dimensional, two-phase, and three-phase LC-MS/MS (McDonald *et al*., 2002) has shown a dramatic increase in the number of peptides identified in the multidimensional runs versus the single dimensional. A larger but less dramatic increase in the number of peptides identified is seen in a three-phase run versus a two-phase. In particular, more hydrophilic peptides are observed in a three-phase column.

Preparation of the nanocolumn:

1. Cut a 50–54 cm length of 100 × 365-µm fused-silica capillary tubing. Hold the center of the tubing over an alcohol burner and burn the polyimide coating off a 5–10 cm section. Remove the charred material by cleaning with a Kimwipe soaked in methanol. The tubing should be clear in this section.
2. Place the length of tubing in the P-200 laser puller so that the clear section is in the mirrored chamber of the puller. In this position, the laser is focused on the center of the tubing and the fused silica can be melted.
3. Select the program on the laser puller that will result in a 3–5 µm ID tip. The heating program should be repeated three times before the tubing is pulled.

Packing the column:

1. Place a small amount (5 mg) of C_{18} RP packing material (5-µm Polaris C_{18}-A or similar) in a microcentrifuge tube and add 1 mL of methanol. Agitate the tube to create a slurry of the packing material.
2. Place the open microcentrifuge tube into the stainless steel bomb and secure the lid by tightening the five screws.
3. Insert the flat end of the pulled capillary column through the ferrule until it reaches the bottom of the microcentrifuge tube. Tighten the ferrule until the capillary does not move when gently tugged.
4. Adjust the pressure on the helium tank to 400–800 psi. Slowly pressurize the bomb by opening the valve on the high-pressure line. If the bomb is pressurized too rapidly, the microcentrifuge tube can rupture.
5. When the pressure is applied to the bomb, packing material should begin to rise and fill the capillary. If packing material stops filling the capillary before 7–10 cm have been packed, release the pressure on the bomb, loosen the ferrule holding the capillary and gently tap the capillary on the bottom of the microcentrifuge tube. Re-tighten the ferrule and repressurize the bomb. Continue packing until the desired amount of packing material is loaded.
6. If a two- or three-phase column is to be made, release the pressure on the bomb and remove the capillary column. Open the bomb and replace the C_{18} packing material with a microcentrifuge tube containing a methanol slurry of SCX (5-µm Partisphere strong cation exchanger, Whatman) packing material.

7. Repressurize the bomb and load 3 cm of SCX packing material. For a three-phase column, repeat the procedure, but load 3 cm of C_{18} packing material.

Reference

McDonald WH, Ohi R, Miyamoto DT, Mitchison TJ and Yates III JR (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *International Journal of Mass Spectrometry*, **219**, 245–251.

Tutorial

Tutorial on protein fingerprinting

Fredrik Levander
Lund University, Lund, Sweden

1. Introduction

Mass spectrometry-based protein identification is one of the core techniques in proteomics. It is based on the comparison of peaks in mass spectra with calculated theoretical masses for proteins as derived from their primary structure. By cutting the protein to identify with a specific endoprotease or chemical reaction, a mixture of peptides is generated. The peptide masses are measured in a mass spectrometer. The cutting procedure is then simulated on all proteins in a protein sequence database and the theoretical mass fingerprints generated are compared to the one acquired on the mass spectrometer (Figure 1). For reviews on protein fingerprinting, also known as peptide mass fingerprinting, see (Henzel *et al*., 2003) and Article 12, **Protein fingerprinting**, Volume 5. Peptide fingerprinting, or peptide fragment fingerprinting, is based on similar principles, but instead of measuring the peptide masses only, the individual peptides are fragmented in the mass spectrometer yielding peptide fragment fingerprints. Peptide fingerprinting is reviewed elsewhere in this encyclopedia (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5), and there is a separate tutorial on the subject (*see* Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5). Nevertheless, there are several points mentioned below that are valid also for peptide fingerprinting.

For satisfactory protein identification it is important to consider the different steps of the workflow since there are several factors that are important to achieve a successful identification. This tutorial will give an overview of the process and some hints on how to succeed with protein fingerprinting.

2. Experimental setup

A working experimental setup is the first key to success and sample preparation is described thoroughly in Article 14, **Sample preparation for MALDI and electrospray**, Volume 5. The most basic prerequisite is that there has to be sufficient protein for the analysis. Trypsin digestion of the sample will work well only if the protein concentration is sufficient. Nowadays this is more critical than the sensitivity of the mass analyzers as the mass spectrometers are usually sensitive enough. An efficient digestion is important for extraction of the protein fragments from gel

Amino acid sequence of intact protein

```
1  MASKEFHIVV ETGIHARPAT LLVHTASKFT SEITLEYKGK SVNLKSIMGV
51 MSLGVGQGAD VTISAEGADA DDAISTIAET MTKEGLAE
```

Mass	**Position**	**Peptide sequence after trypsin digestion**
3712.76	46–83	SIMGVMSLGVGQGADVTISA EGADADDAISTIAETMTK
2626.44	5–28	EFHIVVETGIHARPATLLVH TASK
1230.62	29–38	FTSEITLEYK
560.34	41–45	SVNLK
518.24	84–88	EGLAE
436.22	1–4	MASK
204.13	39–40	GK

Figure 1 Theoretical tryptic digest of a protein. The arginines and lysines are marked in gray in the amino acid sequence

slices. If the protein is still in the gel, there will not be any sample signal, no matter how sensitive the mass spectrometer is. It is even more important for the actual fingerprinting to have as complete cleavage as possible. Some missed cleavages are inevitable, but if two or more have to be considered for each peptide, there will be too many possible masses for efficient identification of most proteins. Efficient sample ionization and sensitive and well-calibrated mass analysis are as important for the analysis. For protein fingerprinting, currently the standard mass spectrometer is of the matrix-assisted laser desorption ionization-time-of-flight (MALDI-ToF) type, *see* Article 7, **Time-of-flight mass spectrometry**, Volume 5.

3. Peak extraction and deisotoping

When the mass spectra have been acquired, the first task is to convert the spectra into peak lists. Mass spectra are a mixture of sample signal and noise and it is not straightforward to do the peak picking. Every peptide will yield a cluster of peaks due to the isotopic distribution, and only the monoisotopic peak should be used for the fingerprinting, see Figure 2. As illustrated in Figure 2 it is often quite easy to select the strongest peaks, but on the lower end of the scale, it is almost impossible to differentiate signal from noise. Peak extraction is normally performed using computer software, which is most often bundled with the mass spectrometer. The peak picking software work quite differently and they also have parameter settings, which will influence the results. If the algorithms are set to be very sensitive they usually extract more background as well as more sample peaks. This can be good in some and bad in other cases, therefore it is worthwhile to search with several peak lists extracted from the same mass spectrum (Rögnvaldsson *et al*., 2004).

4. Peak lists filtering

Mass spectra also contain matrix or solvent peaks depending on the ionization source, and in addition peptide peaks derived from contaminants such as trypsin autolysis peaks and keratins are common in the samples. It is a good idea to remove as many of the contaminant peaks as possible, provided they can be identified. Still, removal of sample peaks is undesirable and it is sometimes worthwhile to search

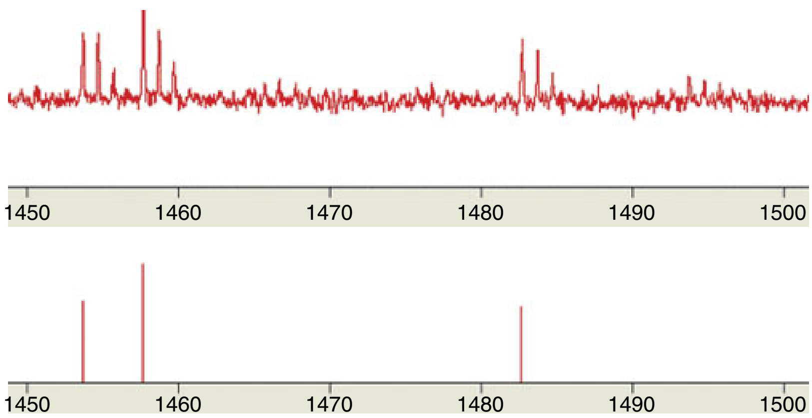


Figure 2 Raw MALDI-ToF spectrum and selected monoisotopic peaks with signal-to-noise threshold set at 2.0. The *x* axis is *m/z* and *y* is the signal intensity

both using the raw peak list and a filtered peak list. Basic filters for trypsin are often included in the database matching programs, but for more extensive filtering one will have to find out what contaminants are often present in the lab by overlaying peak lists. This can be performed manually or automatically (Levander *et al*., 2004).

At the same time as performing the filtering of peak lists, the identified contaminants can also be used to recalibrate the spectra, if this is not automatically included in the database search, as it is in the Aldente search engine, for example, (**www.expasy.org/tools/aldente/**).

5. Database searching

There are numerous different computer programs that perform matching of the peak list with theoretical peak lists derived from the protein database (*see* Article 12, **Protein fingerprinting**, Volume 5). The matching itself is trivial, and the difference between the different algorithms is rather on the scoring level. If all peaks in a spectrum would match all theoretical peaks of a protein, there would not be much ambiguity in the identification process. However, the normal result is that only a fraction of the peaks match, and that many of the theoretical peaks cannot be found in the spectrum or are outside the mass range for which spectra are acquired. There will be some peak matches to many proteins, and the scoring algorithm will try to tell, which is the best protein candidate for the peak list. The algorithms should also account for the possibility that there are several proteins in the sample. It is desirable to have some kind of measure of how likely the protein hit is to be random or correct, and probability-based algorithms should preferentially be used. Mascot from Matrix Science (**www.matrixscience.com**) is probably the most used search engine today, but several other search programs also provide probabilistic scoring.

Irrespective of the choice of database search program, there are several important parameters to set:

The choice of database is most critical. If the protein or a closely related one is not in the database the search will not be successful. If the analyzed protein is from an organism with a sequenced and annotated genome, one can opt to search only that proteome. A problem that can arise is that the protein has not been correctly annotated even if the genome sequence is published. Sometimes searching also in sequences from related organisms can be helpful. It is also possible to find conserved proteins even for organisms that have not been sequenced if the sequence similarity to a known protein is high enough. It is also a good idea to search a large protein database with proteins from several organisms, since it is quite unlikely that a protein from the organism analyzed ends up as a top candidate by chance.

Mass tolerance is an important factor. The mass accuracy of the mass spectrum determines how small the search window can be, that is, which mass error can be tolerated for a peak match. If the mass tolerance window is small, fewer proteins will have to be scored, since there will be fewer random matches than with a large mass error. On the other hand, if the mass window is set too small, one will risk missing hits when the calibration is worse than expected. It is so often that the mass calibration varies quite a lot over a MALDI target plate, and it is necessary to perform several searches with different mass tolerances to find the right protein. Some scoring algorithms utilize the fact that the mass error tends to be linear and makes a correlation of the error which is included in the score (Egelhofer *et al*., 2002, Gasteiger *et al*., 2005).

Amino acid modifications to consider have to be marked.Chemical modification of amino acids will change the mass for a peptide, and if it is not considered, the peptide mass will not be matched. Fixed amino acid modifications are those that can be expected to be true for all (or most) amino acids of one kind, for example, reduction and alkylation of cysteins in 2D gels will impose carbamidomethylation. Variable modifications are those that appear on some amino acids, but not all. Methionine oxidation is a modification that is very frequent and variable methionine oxidation is of a standard setting. There are many natural modifications of proteins, and it could be teasing to set some of these as variable in the search. However, for each variable modification the number of possible masses quickly rises for a peptide, and this makes the search less specific. If a peptide contains a few amino acids that could all be modified independent of each other, the number of combinations rises quickly.

Finally, missed cleavages for the enzyme should be considered, but as for the variable modifications, the number of possible masses quickly rises if many missed cleavages are allowed.

6. Validation of the results

The protein fingerprinting experiment will usually return one or more candidate proteins with scores. The main task is then to determine which hits are true and which are false. To start with one can set a score cutoff in which the false-positive occurrence is within tolerable levels for the particular experiment. A hit that does

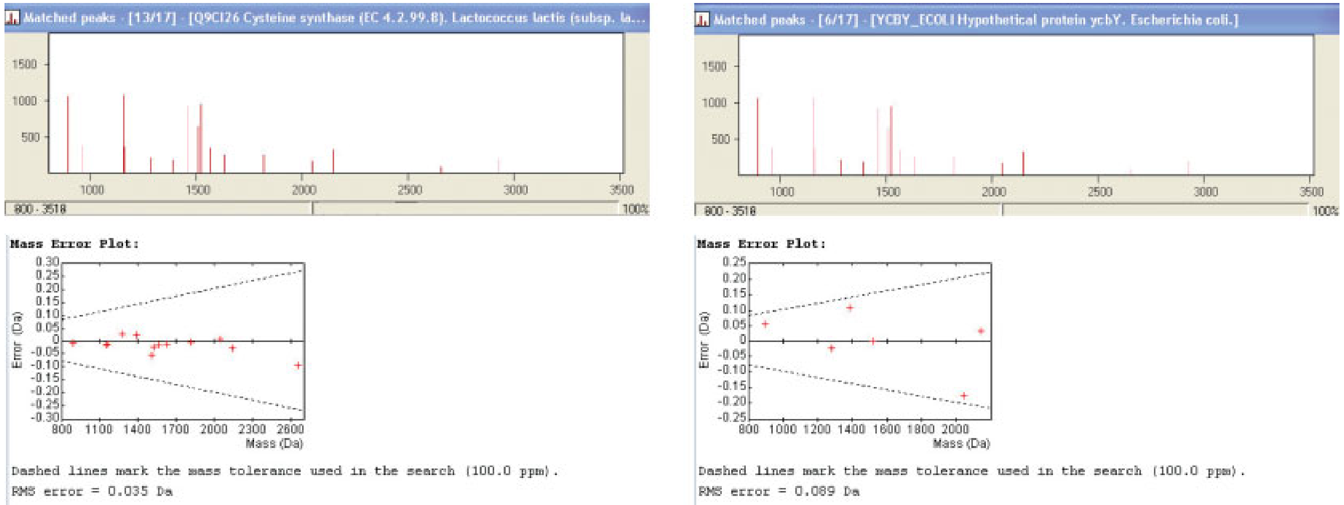


Figure 3 Database matches with a MALDI-ToF peak list. In the spectra the matched peaks are in dark red and the unmatched peaks in light red. To the left is the correct hit. The large spectrum peaks are covered in the match and the mass error is linear. Fifty percent of the protein sequence was covered (not shown). To the right is a false hit. Even though some of the large peaks are covered, the protein coverage is only 11% and the mass error is irregular

not pass the cutoff but has a good score can then be manually inspected with regard to factors that were not included in the scoring of the search program. If the strong peaks in the spectrum match the protein and the mass error is systematic these are good indications of a true protein hit (Figure 3). Tryptic peptides with a *C*-terminal arginine are usually quite strong too. However, for certainty, validation with peptide fingerprinting is often required.

References

Egelhofer V, Gobom J, Seitz H, Giavalisco P, Lehrach H and Nordhoff E (2002) Protein identification by MALDI-TOF-MS peptide mapping: A new strategy. *Analytical Chemistry*, **74**, 1760–1771.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD and Bairoch A (2005) Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, JM (Ed.), Humana Press, New Jersey.

Henzel WJ, Watanabe C and Stults JT (2003) Protein identification: the origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry*, **14**, 931–942.

Levander F, Rögnvaldsson T, Samuelsson J and James P (2004) Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics*, **4**, 2594–2601.

Rögnvaldsson T, Hakkinen J, Lindberg C, Marko-Varga G, Potthast F and Samuelsson J (2004) Improving automatic peptide mass fingerprint protein identification by combining many peak sets. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*, **807**, 209–215.

Introductory Review

Separation-dependent approaches for protein expression profiling

Michael J. Dunn and Stephen R. Pennington
Proteome Research Centre, University College Dublin, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland

1. Introduction

Intense efforts over the last few years have resulted in the availability, at the time of writing (February 2005), of complete genome sequences for 256 organisms (21 archael, 203 bacterial, 32 eukaryotic), including man. This wealth of information is an invaluable resource that will allow comprehensive studies of gene expression that will in turn lead to new insights into cellular functions that determine biologically relevant phenotypes in health and disease. The understanding that one gene can encode more than a single protein has led to a realization that the functional complexity of an organism far exceeds that indicated by its genome sequence alone. While powerful techniques such as DNA microarrays and serial analysis of gene expression (SAGE) make it possible to undertake rapid, global transcriptomic profiling of mRNA expression, processes including alternative mRNA splicing, RNA editing, and co- and posttranslational protein modification make it essential to undertake expression studies at the protein level. The concept of mapping the human complement of protein expression was first proposed more than 25 years ago (Anderson and Anderson, 1982), with the development of a technique in which large numbers of proteins could be separated simultaneously by two-dimensional polyacrylamide gel electrophoresis (2-DE) (O'Farrell, 1975). The term "proteome" was not established until the mid-1990s (Wasinger *et al*., 1995) when it was proposed to define the protein complement of a genome. This article will give an introduction to expression profiling in which the individual proteins in a complex sample are separated prior to semiquantitative analysis, and then identified usually using techniques of mass spectrometry. For an introduction to alternative strategies for expression profiling of unresolved complex protein samples, see Article 21, **Separation-independent approaches for protein expression profiling**, Volume 5.

2. Two-dimensional gel electrophoresis (2-DE)

The technique of two-dimensional gel electrophoresis (2-DE) in which proteins are separated in the first dimension according to their charge properties (isoelectric point, pI) under denaturing conditions, followed by their separation in the second dimension according to their relative molecular mass (M_r) by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was developed more than 25 years ago (O'Farrell, 1975). Nevertheless, it remains the core technology of choice for the majority of applied proteomic projects (Görg *et al*., 2004; *see also* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) due to its ability to separate simultaneously thousands of proteins and to indicate posttranslational modifications that result in alterations in protein pI and/or M_r. Large-format (24 × 21 cm) 2-D gels can routinely separate around 2000 protein spots. Moreover, recent developments including the use of narrow range "zoom" gels (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) and fluorescent dyes that facilitate the multiplex analysis of samples (*see* Article 25, **2D DIGE**, Volume 5 and Article 30, **2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis**, Volume 5) make it possible to achieve greater proteomic coverage combined with more accurate differential expression analysis. Additional advantages of 2-DE are the high-sensitivity visualization of the resulting 2-D separations (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5), compatibility with quantitative computer analysis to detect differentially regulated proteins (Dowsey *et al*., 2003; *see also* Article 26, **Image analysis**, Volume 5), and the relative ease with which proteins from 2-D gels can be identified and characterized by mass spectrometry (*see* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5).

3. Alternatives to 2-DE

Despite the many advantages of 2-DE, there are alternative protein separation strategies. Perhaps the simplest alternative to 2-DE is the use of one-dimensional SDS-PAGE to separate proteins in the sample, on the basis of their M_r followed by protein identification by tandem mass spectrometry (MS/MS), such that several proteins comigrating in a single band can be identified. This method is limited by the complexity of the protein mixture that can be analyzed but is well suited for the analysis of membrane proteins, and has also been successfully applied to the study of protein complexes (Figeys *et al*., 2001). Other approaches avoid the use of gels altogether by combining liquid chromatography (LC) and MS. In these so-called shotgun approaches, a tryptic digest of the sample is separated by one or more dimensions (typically ion-exchange combined with reverse-phase) of LC to reduce the complexity of peptide fractions. These are subsequently introduced (either on- or off-line) into a tandem mass spectrometer for sequence-based identification. For example, the so-called MudPIT approach (Wolters *et al*., 2001) identified around

1500 yeast proteins in a single analysis (Washburn *et al*., 2001). An alternative to this approach that is more robust than multidimensional chromatography, while still allowing complex samples to be analyzed, has been termed *GeLC-MS/MS* (Schirle *et al*., 2003). Here, tryptic digests of protein bands excised from the SDS-PAGE gel are separated by one-dimensional RP-HPLC prior to on- or off-line MS/MS analysis. However, a major limitation of such approaches is that unless combined with some form of stable isotope labeling or "mass tagging", they provide no information on semiquantitative abundance of proteins and are very limited in their ability to detect posttranslational modifications.

The former problem is currently being addressed by the development of a range of MS-based techniques in which stable isotopes are used to differentiate between two or more populations of proteins (Han *et al*., 2001). In general, this approach consists of four steps: (1) differential isotopic labeling of the two (or more) protein mixtures, (2) digestion of the combined labeled samples with a protease such as trypsin or Lys-C, (3) separation of the peptides by multidimensional LC, and (4) semiquantitative analysis and identification of the peptides by MS/MS. Currently, the most widely used method is the isotope-coded affinity tag (ICAT) (Han *et al*., 2001; *see also* Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5), but there are a variety of other approaches involving labeling with stable isotopes at the whole cell, intact protein, or tryptic peptide level (Julka and Regnier, 2004; Ross *et al*., 2004; *see also* Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5). Although these approaches are promising, there are caveats: (1) their quantitative reproducibility needs to be established, (2) the dynamic range of the these techniques may be little better than 2-DE, and (3) there is evidence that they can be complementary to a 2-DE approach in identifying a different subset of proteins from a given sample (Kubota *et al*., 2003).

4. Conclusion

The current array of proteomic techniques makes it possible to characterize global alterations in protein expression associated with the progression of many different biological processes, including human disease. However, there is still no one method that is suitable for the analysis of all samples, and for many projects it is likely that a combination of proteomic platforms, both gel and nongel based, will have to be applied to provide the required depth of proteomic coverage.

Acknowledgments

MJD is the recipient of a Science Foundation Ireland Research Professorship, and is grateful to SFI for the generous support of research in his laboratory. The Proteome Research Centre established by SRP has been supported by grants from the Higher Education Authority in Ireland.

References

Anderson NG and Anderson L (1982) The human protein index. *Clinical Chemistry*, **28**, 739–748.

Dowsey AW, Dunn MJ and Yang GZ (2003) The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, **3**, 1567–1596.

Figeys D, McBroom LD and Moran MF (2001) Mass spectrometry for the study of protein-protein interactions. *Methods*, **24**, 230–239.

Görg A, Weiss W and Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**, 3665–3685.

Han DK, Eng J, Zhou H and Aebersold R (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*, **19**, 946–951.

Julka S and Regnier F (2004) Quantification in proteomics through stable isotope coding: a review. *Journal of Proteome Research*, **3**, 350–363.

Kubota K, Wakabayashi K and Matsuoka T (2003) Proteome analysis of secreted proteins during osteoclast differentiation using two different methods: two-dimensional electrophoresis and isotope-coded affinity tags analysis with two-dimensional chromatography. *Proteomics*, **3**, 616–626.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological Chemistry*, **250**, 4007–4021.

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, *et al*. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, **3**, 1154–1169.

Schirle M, Heurtier MA and Kuster B (2003) Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, **2**, 1297–1305.

Washburn MP, Wolters D and Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL and Humphery-Smith I (1995) Progress with gene-product mapping of the mollicutes: mycoplasma genitalium. *Electrophoresis*, **16**, 1090–1094.

Wolters DA, Washburn MP and Yates JR (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, **73**, 5683–5690.

Introductory Review

Separation-independent approaches for protein expression profiling

Rosalind E. Jenkins
University of Liverpool, Liverpool, UK

Stephen R. Pennington
Proteome Research Centre, University College Dublin, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland

1. Introduction

The completion of the Human Genome Mapping Project (see for example **www.ornl.gov/sci/techresources/Human_Genome/project/progress.shtml**, **www.sanger.ac.uk/HGP/**, **www.tigr.org/tdb/**) plus many others has the potential to accelerate our understanding of the biology of these organisms. One of the most successful tools developed to exploit this new information is the cDNA or oligonucleotide microarray, capable of analyzing the relative level of transcription of thousands of genes simultaneously on a format not much larger than a postage stamp (Duggan *et al*., 1999; Schena *et al*., 1995; *see also* Article 90, **Microarrays: an overview**, Volume 4). Such transcriptomic studies have provided insights into the development of cancer (Liang and Pardee, 2003; Ntzani and Ioannidis, 2003) and infectious diseases (Bryant *et al*., 2004), and aided in the assessment of the toxicity profiles of drugs (Shaw and Morrow, 2003). However, as such methods have become more routine and are in more widespread use, it has become clear that they have to be complemented by the direct analysis of the functional end products of the genes – the proteins. There are several reasons for this: first, the correlation between the level of an mRNA and its encoded protein is frequently poor; second, there are biological samples that are not suitable for mRNA expression analysis, such as plasma and urine; and third, it is not only the expression level of a protein that determines cell fate but also factors such as the subcellular localization, the degree and type of posttranslational modification, and the interaction of the protein with other biomolecules. Currently, analysis of protein expression often involves labor-intensive and low-throughput techniques such as two-dimensional gel electrophoresis (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) and multidimensional liquid chromatography (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological**

discovery, Volume 5), but the field of proteomics would be revolutionized if protein microarrays analogous to DNA microarrays could be developed (*see* Article 24, **Protein arrays**, Volume 5, Article 32, **Arraying proteins and antibodies**, Volume 5, and Article 33, **Basic techniques for the use of reverse phase protein microarrays for signal pathway profiling**, Volume 5). This review describes the basic construction of high-density microarrays and highlights some of the issues arising from the exploitation of the technology for high-throughput protein analysis.

2. Protein versus antibody arrays

Protein arrays are those in which the capture reagent is a peptide or a protein other than an antibody, and one of their main applications is to investigate functional interactions (Figure 1). Protein arrays may be used to investigate potential drug targets or determine substrate binding to enzymes. These arrays may be reversed so that multiple peptides or chemical libraries are immobilized and the array is probed with a single protein of interest. For example, peptides containing consensus phosphorylation sites may be arrayed and incubated with a kinase in the presence of radiolabeled ATP: autoradiography then reveals which peptides have been phosphorylated and provides clues as to the substrate specificity of the kinase (MacBeath and Schreiber, 2000). As a result of progress in genome sequencing, researchers are now able to generate libraries of recombinant proteins that represent a substantial proportion of an organism's proteome and these may be cloned, expressed, and spotted onto an array surface in an automated manner (Schweitzer *et al*., 2003; Walter *et al*., 2000; *see also* Article 24, **Protein arrays**, Volume 5 and Article 32, **Arraying proteins and antibodies**, Volume 5). One application of this approach is to investigate protein–protein interactions on a proteome-wide scale, thereby determining the organism's "interactome". Other applications of the recombinant protein array include screening of manufactured antibodies for specificity and as a diagnostic tool by detecting disease-specific antibodies in human plasma (Lueking *et al*., 2003). There are some limitations to this approach, not least of which is the fact that bacterially expressed proteins are not posttranslationally modified in the same way as those expressed in mammalian cells, and this may affect their interaction with other proteins. However, other sources of protein may be exploited. For instance, protein extracted from a prostate cancer cell line has been fractionated, spotted onto nitrocellulose-coated slides, and used to profile the antibody repertoire in prostate cancer patients (Bouwman *et al*., 2003).

Antibody arrays are those in which the capture reagents immobilized on the array surface are themselves antibodies and they are used as a means of monitoring protein expression changes within the test samples (Figure 2). They are thus designed to perform the same sort of proteomic analysis as comparative two-dimensional gel electrophoresis (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) but in a much higher throughput manner, and are more analogous to DNA microarrays than the protein arrays described above. Protein profiling via antibody microarrays has great potential as a discovery tool, particularly in the field of cancer research. Preliminary studies of oral squamous cell carcinoma (Knezevic *et al*., 2001), prostate

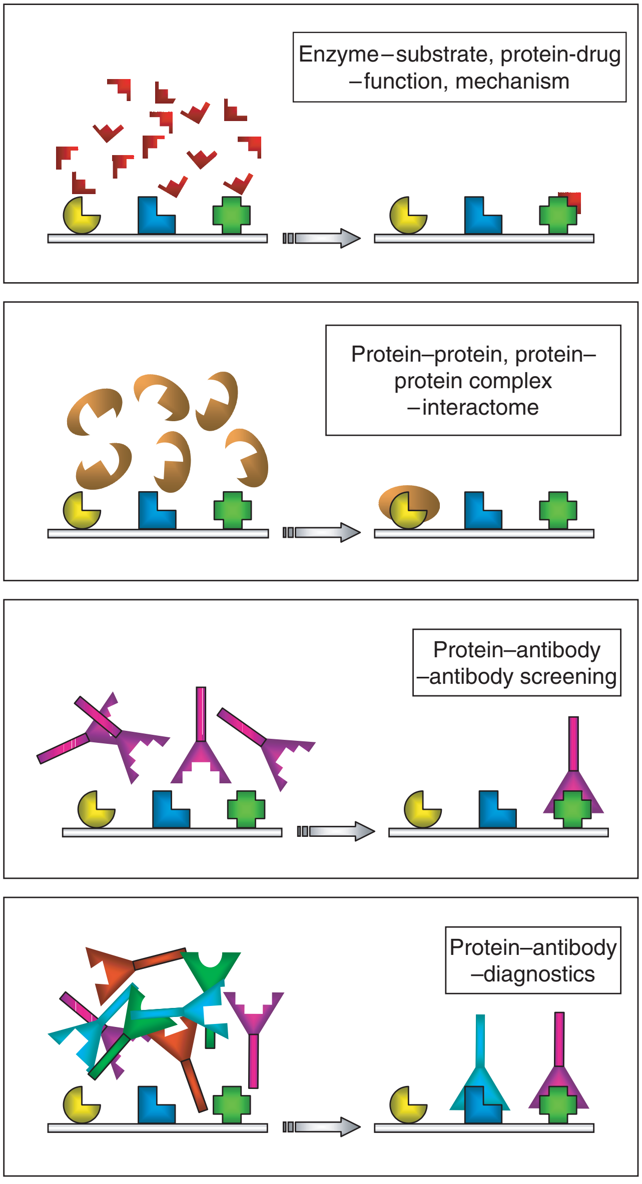


Figure 1 Schematic of protein arrays and their potential applications. A range of proteins other than antibodies are immobilized on a solid support in an addressable format and they are screened for interactions with a variety of samples. They may be probed with complex mixtures of other proteins or biological molecules, or with synthetic moieties such as drugs. This may provide insights into protein function, aid in the development of new drugs, or provide tools for the clinic in the form of diagnostic kits

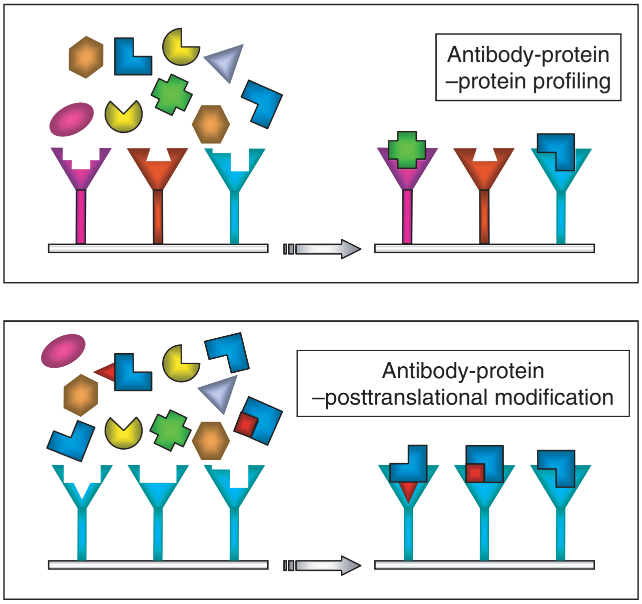


Figure 2 Schematic of antibody arrays and their potential applications. Immobilized antibodies may be used to screen samples such as body fluids or cell lysates in order to assess quantitative changes in protein expression during, for example, disease progression or drug treatment. Through exploitation of the exquisite specificity of antibodies, the posttranslational modifications of proteins may also be determined, shedding light on their activity and the signaling pathways that may be involved in their regulation

cancer (Miller *et al*., 2003), and leukemia (Belov *et al*., 2001) have indicated that the diagnosis and treatment of cancer may be revolutionized as a result of the discoveries made with this technology. However, as the following sections will reveal, there remain many technical difficulties to overcome before protein and antibody arrays become routine tools in the clinical and research environment.

3. Requirements of an array

The basic requirements of an array or chip are apparently simple: a source of specific recognition molecules that can be immobilized in a spatially defined manner, and a method to detect the interaction of the individual molecules in the test mixture with their corresponding recognition molecules on the array. The properties of DNA make it an ideal tool for the specific and quantitative recognition of target nucleic acid sequences. The oligonucleotide or cDNA components of a DNA chip can be readily synthesized in the laboratory, either through plasmid amplification in host cells, PCR, or by direct synthesis onto a solid support (Gao *et al*., 2004; *see also* Article 90, **Microarrays: an overview**, Volume 4). The intermolecular recognition between target and probe is based, in effect, on "linear epitopes" and, thus, retaining a three-dimensional conformation is unnecessary and indeed undesirable. This means that nucleic acid sequences may be immobilized on a solid support

without loss of functionality, and array hybridization may be performed under stringent conditions. Critically, the probes and the targets share their physicochemical properties, which means that every probe will hybridize to their specific target under similar experimental conditions during the screening process. Detection of nucleic acids within the sample is simple because the DNA may be tagged with radioisotopic or fluorescent labels without influencing subsequent interaction with the arrayed DNA. Finally, DNA is relatively resistant to degradation, and the chips, if stored correctly, should have a reasonable shelf life. As will become evident, many of the characteristics of nucleic acids that make them so suitable for analysis by microarray technology do not apply when it comes to the development of protein and antibody arrays.

4. Proteins as elements on an array

Protein and antibody arrays are required to perform rather different analyses, but many of the same principles apply when utilizing protein moieties as the immobilized recognition molecules. The proteins must be generated in reasonable quantities and they must be purified to homogeneity prior to their application to the chip surface. While this can be achieved for a small number of proteins, performing multiple steps on hundreds of proteins in parallel is a major undertaking, especially as each protein may require its own optimized procedures for expression and isolation. Similarly, conventional methods for generating high-quality monoclonal antibodies would be prohibitively time consuming and expensive if they were to be applied to hundreds of protein targets. The proteins that comprise the array must be highly specific, that is, they must be capable of recognizing individual proteins when presented in a complex mixture. Unlike DNA, proteins adopt complex three-dimensional conformations in vivo, and their intermolecular interaction may be perturbed by any of a number of factors that affect the spatial orientation of the molecules. In addition, they are subject to nonspecific interactions when similar epitopes are presented by many different proteins. Finally, the binding of the individual proteins within the sample mixture to the immobilized proteins on the array must occur under similar blocking and washing conditions, not easy to achieve because they are expressing a multitude of physicochemical properties. These are difficult issues to address, but some of the approaches being explored will be described briefly here (Jenkins and Pennington, 2001; Phelan and Nock, 2003; Seong and Choi, 2003).

One approach is to generate libraries of bacteria expressing the proteins of interest via inducible plasmids. The bacteria are cultured in 96-well plate format and many of the steps, from expression and harvesting of the recombinant proteins to spotting on acrylamide-coated slides, are automated. Such an approach has yielded a protein array with over 2400 different human proteins that may be used for screening antibodies for specificity and affinity, and for investigating serum from patients with conditions such as arthritis (Lueking *et al*., 2003). One critical issue with this type of expression platform is that bacteria are unable to perform the full range of posttranslational modifications that are associated with mammalian cells, and thus not all of the bacterially expressed proteins will display the required

intermolecular interactions. Another approach is phage display technology, in which a virus that infects bacteria is genetically manipulated to express a protein of interest on its surface coat (Barbas *et al*., 1991; Clackson *et al*., 1991; Griffiths *et al*., 1994; Huse *et al*., 1989), allowing rapid and potentially parallel manufacture of many proteins. This technology can be used to express proteins or antibody fragments, and the phage can be manipulated to express the proteins fused to tags useful for purification or immobilization. However, the phage must undergo multiple rounds of selection before a protein highly specific for its target is isolated.

One means to avoid the pitfalls associated with proteins as elements on an array is to avoid them altogether. Oligonucleotide aptamers, sequences of nucleic acids that can be engineered to adopt specific three-dimensional conformations, have been used with some success (Cox and Ellington, 2001). All the benefits of working with nucleic acids rather than proteins apply, but the screening process is again laborious, and the proteins most successfully targeted are those that are small and hydrophilic. The approach has been improved recently by the incorporation of a light-reactive nucleotide into the arrayed aptamers, allowing cross-linking to the target protein by UV irradiation and more stringent washing prior to fluorescence detection of bound proteins (Bock *et al*., 2004). Some researchers have begun to exploit the mimetic abilities of inorganic polymers that, once imprinted with the three-dimensional shape of the target protein, can then be used to form the basis of an assay for that protein (Takeuchi *et al*., 1999; Owens *et al*., 1999; Ye and Haupt, 2004). The so-called molecularly imprinted polymers exhibit many useful characteristics, such as stability, resistance to chemical insult, and ability to adopt any conformation required of them. However, because they are not specific for a single protein epitope but recognize the whole protein, they are, in effect, polyclonal and suffer from all the associated problems of cross-reactivity and limited ability to distinguish between alternatively spliced and posttranslationally modified proteins. Nevertheless, it is likely that this technology will continue to be developed and will result in a range of valuable affinity reagents.

5. Immobilization strategies

There are many methods available for the deposition of DNA molecules onto a variety of surfaces without alteration of the molecular properties of the molecules. While some of these may be adapted for the arraying of proteins or antibodies, maintaining the specificity of the capture reagent is much more difficult because the molecular interaction is usually critically dependent on the three-dimensional structure of the protein (*see* Article 32, **Arraying proteins and antibodies**, Volume 5). Irreversible electrostatic binding of proteins to a nitrocellulose-based polymer coated onto glass slides (FAST) allows high density, high-capacity arraying of proteins, but denaturation of the proteins may occur (Stillman and Tonkinson, 2000; Seong and Choi, 2003; Espina *et al*., 2003). Covalent binding directly to substrates such as photolithographically etched nylon or silica surfaces (Blawas and Reichert, 1998; Mooney *et al*., 1996) would probably result in a more stable array, but may also alter the binding properties of the capture reagent. The binding capacity may be increased by modification of the surface of the substrate with

functional groups such as aminosilane (Jenkins and Pennington, 2001; Seong and Choi, 2003). Alternatively, substances such as glutaraldehyde and *N*-succinimidyl-4-maleimidobutyrate may be used as cross-linkers in order to increase accessibility to ligand binding sites by introducing a gap between the substrate surface and the immobilized protein (ShriverLake *et al*., 1997; Seong and Choi, 2003; Blawas and Reichert, 1998; Jenkins and Pennington, 2001; Espina *et al*., 2003).

Immobilization to the solid support may be by passive adsorption, in which the protein to be immobilized is randomly orientated on the array surface: a percentage of the proteins to be immobilized may fortuitously align themselves correctly, but in order to achieve maximum sensitivity of the assay, orientated immobilization is likely to be required (Hodneland *et al*., 2002; Blawas and Reichert, 1998; Peluso *et al*., 2003). Orientation of immunoglobulins is often achieved through the use of bacterial Protein A, G, or L, which display specificity for several regions of mammalian immunoglobulin molecules. These may be engineered to increase their utility in protein arrays: for instance, Protein A has been engineered to express five copies of the immunoglobulin G–binding domain plus a cysteine residue at the C-terminus, the latter to allow strong binding to gold immobilization surfaces (Kanno *et al*., 2000). Orientation of intact antibodies, Fab fragments, or other capture proteins can also be achieved if they are biotinylated and the surface of the array is coated with streptavidin (Peluso *et al*., 2003), but there is a risk that tagging of the proteins may alter their interaction with the protein of interest. Clearly, optimization of the immobilization process is an ongoing task.

6. Detection of interacting proteins

Detection of bound molecules on a DNA microarray is readily achieved, but detection of bound proteins in a rapid and sensitive manner is more problematic (see Haab, 2003 for review and *see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). Direct labeling of the sample proteins with fluorescent dyes is already well established for other proteomics approaches such as difference gel electrophoresis (DIGE) (Unlü *et al*., 1997). This technique has been extended to antibody arrays (Haab *et al*., 2001) but may not be entirely appropriate as labeling with fluorescent tags involves chemical modification of the proteins and may adversely influence the interaction of the proteins with their respective recognition molecule on the array. It is also important that the stoichiometry of labeling is consistent, that is, the ratio of protein molecules to fluorescent tag molecules is reasonably consistent for any individual protein in any complex mixture. To this end, the reaction conditions must be optimized to ensure that as close to complete labeling as possible of amino groups is achieved so that any differences in signal intensity observed between control and test arrays are due to differences in protein level rather than labeling efficiency. However, the greater the degree of protein modification, the higher is the likelihood that the protein–protein interaction will be disrupted. As methods to manufacture targeted antibodies develop, it may be possible to generate a panel of affinity reagents that recognize their cognate protein in its labeled rather than native state, sidestepping problems associated with

protein modification. However, the sensitivity of direct labeling and detection with a fluorophore is such that a large array with low-density features may be required at a time when miniaturized and high-density arrays are the preferred option.

Small tags such as biotin may be used to label the protein mixture to be probed, and those proteins captured on an array may then be detected using a molecule that displays high affinity for biotin, such as streptavidin. Modification with biotin tends to have less of an impact on protein conformation than labeling with large fluorophores and is therefore less likely to disrupt protein–protein interaction. The streptavidin is labeled either with a fluorophore or with an enzyme such as peroxidase, the latter allowing amplification of the signal on incubation with a color-producing substrate. Thus, indirect detection of captured proteins can have the advantage of increased sensitivity compared to direct detection. No modification of the sample proteins would be required if a labeled antibody specific for the proteins of interest was available, but in the case of antibody arrays, this would mean that two highly specific antibodies raised in different hosts would be required. As has already been discussed, generating a single antibody with the required characteristics can be a problem.

While the above methods for detection of the captured moiety are routine and thus have been optimized for similar assays such as Western blotting or enzyme-linked immunosorbant assay (ELISA), a method to detect captured proteins directly and with high sensitivity without the need to modify the sample would be ideal. Biosensors have been designed that measure the amount of protein captured on a sensing layer by detecting changes in the emission of light, heat, electrons, or ions (Nice and Catimel, 1999; Ziegler and Göpel, 1998; Jenkins and Pennington, 2001). For example, in a surface plasmon resonance (SPR) detector, a glass slide is coated with metal and protein immobilized to it via an ester linkage. When the protein of interest is captured, the mass at the glass surface is increased, and this causes a change in the refractive index of the glass that can be related to the amount of protein bound. This technology forms the basis of the BIAcore system (**www.biacore.com**) (Nelson *et al*., 2000; Nice and Catimel, 1999; Karlsson, 2004). The current applications of this system to proteomics include the screening of aptamers and antibodies to be used as capture reagents, confirmation of interactions between proteins and other moieties and, by coupling directly to a mass spectrometer, identification of unknown interaction partners (Karlsson, 2004). However, if a suitable degree of miniaturization could be achieved, detection methods such as this could be applicable to protein and antibody arrays.

7. Conclusion

Despite the many technical difficulties with the production of protein and antibody arrays that have been summarized in this article, the first commercially available tools for protein and antibody arrays are beginning to appear on the market. Thus, slides may be purchased that are suitable for the immobilization of proteins, and a range of strategies may be used to deposit the proteins of interest in a spatially defined manner, allowing researchers to produce their own arrays. Alternatively, some companies offer a service whereby customized antibodies or peptides are

generated for subsequent incorporation into a chip, and others offer premade arrays of about 2000 "functional" proteins. While significant progress has been made, further developments in all aspects of array production will be required before protein and antibody arrays are sufficiently reliable, cheap, and data-rich to elicit a revolution in our understanding of biological systems comparable to that brought about by DNA chip technology.

References

Barbas CF, Kang AS, Lerner RA and Benkovic SJ (1991) Assembly of combinatorial antibody libraries on phage surfaces – the gene-III site. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 7978–7982.

Belov L, de la Vega O, dos Remedios CG, Mulligan SP and Christopherson RI (2001) Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Research*, **61**, 4483–4489.

Blawas AS and Reichert WM (1998) Protein patterning. *Biomaterials*, **19**, 595–609.

Bock C, Coleman M, Collins B, Davis J, Foulds G, Gold L, Greef C, Heil J, Heilig JS, Hicke B, *et al*. (2004) Photoaptamer arrays applied to multiplexed proteomic analysis. *Proteomics*, **4**, 609–618.

Bouwman K, Qiu J, Zhou H, Schotanus M, Mangold LA, Vogt R, Erlandson E, Trenkle J, Partin AW, Misek D, *et al*. (2003) Microarrays of tumor cell derived proteins uncover a distinct pattern of prostate cancer serum immunoreactivity. *Proteomics*, **3**, 2200–2207.

Bryant PA, Venter D, Robins-Browne R and Curtis N (2004) Chips with everything: DNA microarrays in infectious diseases. *The Lancet Infectious Diseases*, **4**, 100–111.

Clackson T, Hoogenboom HR, Griffiths AD and Winter G (1991) Making antibody fragments using phage display libraries. *Nature*, **352**, 624–628.

Cox JC and Ellington AD (2001) Automated selection of anti-protein aptamers. *Bioorganic & Medicinal Chemistry*, **9**, 2525–2531.

Duggan DJ, Bittner M, Chen YD, Meltzer P and Trent JM (1999) Expression profiling using cDNA microarrays. *Nature Genetics*, **21**, 10–14.

Espina V, Mehta AI, Winters ME, Calvert V, Wulfkuhle J, Petricoin EF III and Liotta LA (2003) Protein microarrays: molecular profiling technologies for clinical specimens. *Proteomics*, **3**, 2091–2100.

Gao X, Gulari E and Zhou X (2004) In situ synthesis of oligonucleotide microarrays. *Biopolymers*, **73**, 579–596.

Griffiths AD, Williams SC, Hartley O, Tomlinson IM, Waterhouse P, Crosby WL, Kontermann RE, Jones PT, Low NM, Allison TJ, *et al*. (1994) Isolation of high-affinity human-antibodies directly from large synthetic repertoires. *The EMBO Journal*, **13**, 3245–3260.

Haab BB (2003) Methods and applications of antibody microarrays in cancer research. *Proteomics*, **3**, 2116–2122.

Haab BB, Dunham MJ and Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**, 1–13.

Hodneland CD, Lee YS, Min DH and Mrksich M (2002) Selective immobilization of proteins to self-assembled monolayers presenting active site-directed capture ligands. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5048–5052.

Huse WD, Sastry L, Iverson SA, Kang AS, Altingmees M, Burton DR, Benkovic SJ and Lerner RA (1989) Generation of a large combinatorial library of the immunoglobulin repertoire in phage-lambda. *Science*, **246**, 1275–1281.

Jenkins RE and Pennington SR (2001) Arrays for protein expression profiling: towards a viable alternative to two-dimensional gel electrophoresis? *Proteomics*, **1**, 13–29.

Kanno S, Yanagida Y, Haruyama T, Kobatake E and Aizawa M (2000) Assembling of engineered IgG-binding protein on gold surface for highly oriented antibody immobilization download full text of article. *Journal of Biotechnology*, **76**, 207–214.

Karlsson R (2004) SPR for molecular interaction analysis: a review of emerging application areas. *Journal of Molecular Recognition*, **17**, 151–161.
Knezevic V, Leethanakul C, Bichsel VE, Worth JM, Prabhu VV, Gutkind JS, Liotta LA, Munson PJ, Petricoin EF III and Krizman DB (2001) Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics*, **1**, 1271–1278.
Liang P and Pardee AB (2003) Analysing differential gene expression in cancer. *Nature Reviews. Cancer*, **3**, 869–876.
Lueking A, Possling A, Huber O, Beveridge A, Horn M, Eickhoff H, Schuchardt J, Lehrach H and Cahill DJ (2003) A nonredundant human protein chip for antibody screening and serum profiling. *Molecular & Cellular Proteomics*, **2**, 1342–1349.
MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.
Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Butler EB, Teh BS and Haab BB (2003) Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. *Proteomics*, **3**, 56–63.
Mooney JF, Hunt AJ, McIntosh JR, Liberko CA, Walba DM and Rogers CT (1996) Patterning of functional antibodies and other proteins by photolithography of silane monolayers. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 12287–12291.
Nelson RW, Nedelkov D and Tubbs KA (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. *Electrophoresis*, **21**, 1155–1163.
Nice EC and Catimel B (1999) Instrumental biosensors: new perspectives for the analysis of biomolecular interactions. *BioEssays*, **21**, 339–352.
Ntzani EE and Ioannidis JP (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, **362**, 1439–1444.
Owens PK, Karlsson L, Lutz ESM and Andersson LI (1999) Molecular imprinting for bio-and pharmaceutical analysis. *Trac-Trends in Analytical Chemistry*, **18**, 146–154.
Peluso P, Wilson DS, Do D, Tran H, Venkatasubbaiah M, Quincy D, Heidecker B, Poindexter K, Tolani N, Phelan M, *et al*. (2003) Optimizing antibody immobilization strategies for the construction of protein microarrays. *Analytical Biochemistry*, **312**, 113–124.
Phelan ML and Nock S (2003) Generation of bioreagents for protein chips. *Proteomics*, **3**, 2123–2134.
Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
Schweitzer B, Predki P and Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. *Proteomics*, **3**, 2190–2199.
Seong SY and Choi CY (2003) Current status of protein chip development in terms of fabrication and application. *Proteomics*, **3**, 2176–2189.
Shaw KJ and Morrow BJ (2003) Transcriptional profiling and drug discovery. *Current Opinion in Pharmacology*, **3**, 508–512.
ShriverLake LC, Donner B, Edelstein R, Breslin K, Bhatia SK and Ligler FS (1997) Antibody immobilization using heterobifunctional crosslinkers. *Biosensors & Bioelectronics*, **12**, 1101–1106.
Stillman BA and Tonkinson JL (2000) FAST slides: a novel surface for microarrays. *Biotechniques*, **29**, 630–635.
Takeuchi T, Fukuma D and Matsui J (1999) Combinatorial molecular imprinting: an approach to synthetic polymer receptors. *Analytical Chemistry*, **71**, 285–290.
Unlü M, Morgan ME and Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.
Walter G, Bussow K, Cahill D, Lueking A and Lehrach H (2000) Protein arrays for gene expression and molecular interaction screening. *Current Opinion in Microbiology*, **3**, 298–302.
Ye L and Haupt K (2004) Molecularly imprinted polymers as antibody and receptor mimics for assays, sensors and drug discovery. *Analytical and Bioanalytical Chemistry*, **378**, 1887–1897.
Ziegler C and Göpel W (1998) Biosensor development. *Current Opinion in Chemical Biology*, **2**, 585–591.

Specialist Review

Two-dimensional gel electrophoresis

Angelika Görg and Walter Weiss
Technische Universität München, Freising-Weihenstephan, Germany

Michael J. Dunn
Proteome Research Centre, University College Dublin, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland

1. Introduction

Two-dimensional gel electrophoresis (2DE) with immobilized pH gradients (IPGs) combined with protein identification by mass spectrometry (MS) is currently the workhorse for proteomics. In spite of new promising technologies that have emerged (MudPIT, stable isotope labeling, arrays) (*see* Article 13, **Multi-dimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5, Article 23, **ICAT and other labeling strategies for semi-quantitative LC-based expression profiling**, Volume 5, and Article 24, **Protein arrays**, Volume 5), 2DE is currently the only technique that can be routinely applied for parallel quantitative expression profiling of large sets of complex protein mixtures such as whole-cell lysates.

Whatever technology is used, proteome analysis is technically challenging, because the number of different proteins expressed at a given time under defined biological conditions is likely to be in the range of several thousand for simple prokaryotic organisms up to at least 10 000 in eukaryotic cell extracts. Moreover, current proteomic studies have revealed that the majority of identified proteins are abundant "housekeeping" proteins that are present in numbers of 10^5 to 10^6 copies per cell, whereas proteins such as receptor molecules that are present in much lower concentrations (typically <100 molecules per cell) are usually not detected. Consequently, improved methods for enrichment of low-abundance proteins are required, such as prefractionation procedures, as well as more sensitive detection and quantitation methods.

2DE couples isoelectric focusing (IEF) and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) to resolve proteins according to two independent parameters, that is, isoelectric point (pI) in the first dimension and molecular mass (M_r) in the second. Depending on the gel size and pH gradient used, 2DE can resolve more than 5000 proteins simultaneously ($\sim$2000 proteins

routinely), and can detect <1 ng of protein per spot. Furthermore, it delivers a map of intact proteins, which reflects changes in protein expression level, isoforms or posttranslational modifications. This is in contrast to LC-MS/MS-based methods, which perform analysis on peptides, where M_r and pI information is lost, and where stable isotope labeling is required for quantitative analysis.

The former limitations of carrier ampholyte-based 2DE (O'Farrell, 1975) with respect to reproducibility, resolution, separation of very acidic and/or very basic proteins, and sample loading capacity have been solved by the introduction of IPGs for the first dimension of 2DE (Görg *et al*., 1988). Narrow-overlapping pH gradients provide increased resolution ($\Delta pI = 0.001$) and detection of low-abundance proteins, whereas alkaline proteins up to pH 12 have been separated under steady state conditions (Görg *et al*., 2000; Wildgruber *et al*., 2002; Drews *et al*., 2004).

The analysis of very hydrophobic proteins such as integral membrane proteins remains a challenge for both 2DE and LC-based proteomic approaches. Currently, the best strategy is the combination of SDS-PAGE analysis of membrane fractions in combination with LC-MS/MS. This method has been termed *geLC-MS/MS* (Li *et al*., 2003).

2. 2DE-MS proteomic workflow

The major steps of the 2DE-MS workflow are

- sample preparation/solubilization
- protein separation by 2DE
- protein detection and quantitation
- computer-assisted analysis of 2DE patterns
- protein identification and characterization
- 2D protein database construction.

and will be discussed in the following sections.

Sample preparation should be as simple as possible to increase reproducibility. Although many "standard" protocols for sample preparation have been published, these protocols must be optimized for the type of sample to be analyzed. Some general recommendations, however, can be made: protein modifications during sample preparation must be minimized, because they might result in artifactual spots on 2D gels. In particular, proteolytic enzymes in the sample must be inactivated. Samples containing urea must not be heated to avoid charge heterogeneities caused by carbamylation of the proteins by isocyanate formed in the decomposition of urea (Dunn, 1993). The three fundamental steps in sample preparation are cell disruption, inactivation and or removal of interfering substances, and solubilization of the proteins (*see* Article 2, **Sample preparation for proteomics**, Volume 5).

Cell disruption can be achieved by osmotic lysis, freeze-thaw cycling, detergent lysis, enzymatic lysis of the cell wall, sonication, grinding with (or without) liquid nitrogen, high pressure (e.g., French press), homogenization with glass beads and a bead beater, nitrogen cavitation, or a rotating blade homogenizer. These methods can be used individually or in combination.

During or after cell lysis, *interfering compounds* such as proteolytic enzymes, salts, lipids, nucleic acids, polysaccharides, and plant phenols have to be inactivated or removed. The two most important parameters are *salt* and *proteolysis*. *Proteases* must be inactivated to prevent protein degradation that otherwise may result in artifactual spots and loss of high M_r proteins. Protease inhibitors are usually added, but they may modify proteins and cause charge artifacts. Other remedies are boiling the sample in SDS-buffer (without urea!), or inactivating proteases by low pH (e.g., precipitating with ice-cold trichloroacetic acid (TCA)). It should be kept in mind that it may be rather difficult to completely inactivate all proteases. TCA/acetone precipitation is very useful for minimizing protein degradation and removing interfering compounds, such as salt. Attention has to be paid, however, to protein losses due to incomplete precipitation and/or resolubilization of proteins. Moreover, a completely different set of proteins may be obtained by extraction with lysis buffer, depending on whether there was a preceding TCA precipitation step. On the other hand, this effect can be used for the enrichment of very alkaline proteins (such as ribosomal or nuclear proteins) from total cell lysates (Görg *et al*., 2000).

Salt ions may interfere with electrophoretic separation and should be removed if their concentration is too high (>100 mM). Removal can be achieved by precipitation of proteins with TCA or organic solvents (e.g., cold acetone). One alternative is the use of 2D cleanup kits (e.g., Amersham Biosciences). Another is dilution of the sample below a critical salt concentration followed by application of a larger sample volume onto the IPG gel. The sample is "desalted" in the gel by using low voltage at the beginning of the run.

After cell disruption and/or removal of interfering compounds, the individual polypeptides must be *denatured, solublized, and reduced* to disrupt intra- and intermolecular interactions while maintaining the inherent charge properties. The most popular sample solubilization buffer is based on O'Farrell's lysis buffer and modifications thereof (O'Farrell, 1975) (9 M urea, 2–4% CHAPS, 1% dithiothreitol (DTT), 2% (v/v) carrier ampholytes). Unfortunately, urea lysis buffer is not ideal for the solubilization of all protein classes, particularly for membrane or other highly hydrophobic proteins. Improvement in the solubilization of hydrophobic proteins has come with the use of thiourea and new zwitterionic detergents such as sulfobetaines (Luche *et al*., 2003).

3. Prefractionation procedures

Prefractionation can be used to reduce the complexity of the sample, enrich for certain proteins such as low-copy number proteins or alkaline proteins, and to get some information on the topology of the proteins. This can be accomplished by:

- isolation of specific cell types from a tissue, for example, fluorescence activated cell sorting (FACS), or laser capture micro dissection (*see* Article 6, **Laser-based microdissection approaches and applications**, Volume 5);
- isolation of cell compartments and/or organelles, for example, by sucrose gradient centrifugation, or free flow electrophoresis;

- selective precipitation of certain protein classes (e.g., TCA/acetone precipitation for ribosomal proteins);
- sequential extraction procedures with increasingly powerful solubilizing buffers, for example, aqueous buffers, organic solvents (e.g., ethanol or chloroform/methanol), and detergent-based extraction solutions;
- chromatographic or electrokinetic separation methods, such as column chromatography, affinity purification, electrophoresis in the liquid phase and/or IEF in granulated gels.

4. 2DE with IPGs (IPG-Dalt)

IPGs are based on the use of the bifunctional Immobiline reagents, a series of 10 chemically well-defined acrylamide derivatives with the general structure CH_2=CH–CO–NH–R, where R is either a carboxyl or an amino group. These reagents form a series of buffers with different pK values between pK 1 and >12. Because the reactive end of the molecule is copolymerized with the acrylamide matrix, extremely stable pH gradients are generated, allowing true steady state IEF with increased reproducibility, as has been demonstrated in several inter-laboratory comparisons (Corbett *et al*., 1994; Blomberg *et al*., 1995).

The original protocol of 2DE with immobilized pH gradients (IPG-Dalt) was described by Görg *et al*. (1988, updated in 2000 and 2004), summarizing the critical parameters inherent to isoelectric focusing with IPGs and a number of experimental conditions (Figure 1). The first dimension of IPG-Dalt, IEF, is performed in individual, 3-mm wide and up to 24-cm-long IPG gel strips cast on GelBond PAGfilm (laboratory-made or commercial Immobiline Dry-Strips). Samples can be applied either by cup-loading or by in-gel rehydration. For analytical purposes, typically 100 µg of protein can be loaded on an 18-cm-long, wide pH range gradient, and 500 µg on narrow range IPGs. For micropreparative purposes, five to ten times more protein can be applied. After IEF, the IPG strips are equilibrated with SDS-buffer in the presence of urea, glycerol, DTT, and iodoacetamide, and applied onto horizontal or vertical SDS gels in the second dimension (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5).

The *choice of pH gradient* primarily depends on the sample's protein complexity. *Wide* IPGs, such as IPG 4–7, 4–9, or 3–11, are used to analyze simple proteomes (small genome, organelle, or other subfraction), or to get an overview of a more complex proteome, respectively (Figure 2).

With samples such as total lysates of eukaryotic cells, 2D electrophoresis on a single wide-range pH gradient reveals only a small percentage of the whole proteome. The best remedy, preferably in combination with prefractionation procedures, is to use multiple *narrow overlapping* IPGs ("zoom-in" gels, e.g., IPG 4–5, IPG 4.5–5.5, or 5.0–6.0) and/or extended separation distances to achieve an optimal resolution to avoid multiple proteins in a single spot for unambiguous protein identification and to facilitate the application of higher protein amounts for the detection of minor components (Wildgruber *et al*., 2000; Westbrook *et al*., 2001).

Strongly alkaline proteins such as ribosomal and nuclear proteins with closely related pIs between 10.5 and 11.8 are focused to the steady state by using IPGs

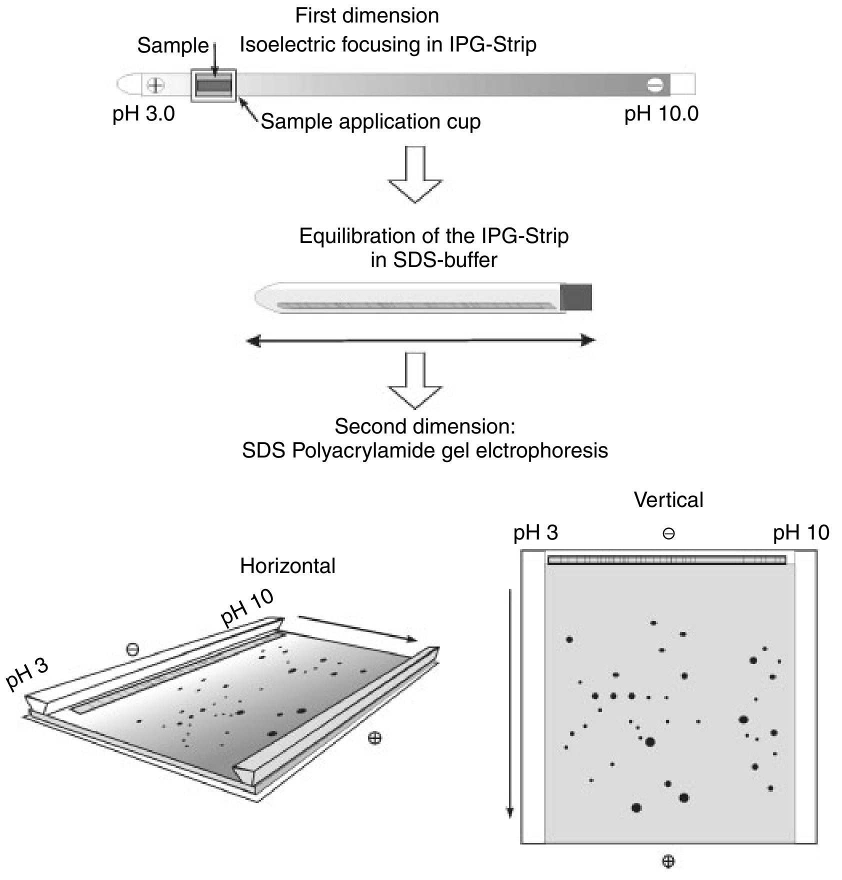


Figure 1 The principle of 2DE with IPGs according to Görg *et al*. 1988, 2000, 2004 (Reproduced by permission of Wiley from Westermeier R and Görg A (2004) Two-dimensional electrophoresis in proteomics. In *Protein Purification, Third Edition*, Janson JC (Ed.), John Wiley & Sons Inc., New York (in press).)

3–12, 6–12, or 9–12 (Wildgruber *et al*., 2002; Drews *et al*., 2004) (Figure 3). For optimized separation, cup-loading at the anode is mandatory, and the use of high voltages (final settings up to 8000 V) is recommended. With narrow gradients between pH 7 and 10, horizontal streaking due to DTT depletion can occur at the basic end. To avoid streaking, the cysteines should be stabilized as mixed disulfides by using hydroxyethyldisulfide reagent (DeStreak^{TM} Amersham Biosciences) in the IPG strip rehydration solution instead of a reductant (Olsson *et al*., 2002).

4.1. *IEF temperature*

Spot positions vary along the pH axis with different applied temperatures (Görg *et al*., 1991). It is thus very important to run the separations at an actively controlled temperature, where 20°C proved to provide the optimal conditions.

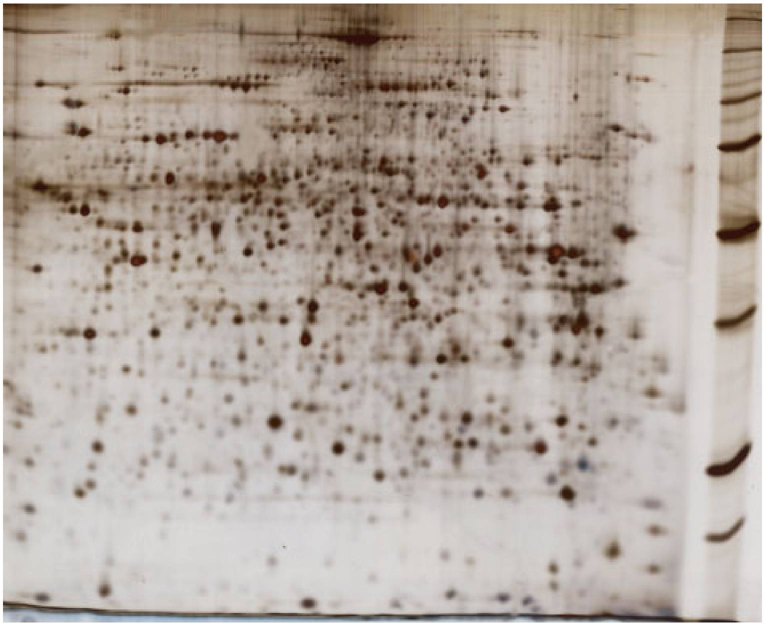


Figure 2 2DE of a TCA-acetone extract of mouse liver proteins, separated by IEF in a 24-cm-long IPG strip containing a nonlinear pH gradient 3–11, followed by SDS-PAGE in a vertical 12.5% gel. Protein detection by silver staining (Reproduced by permission of Wiley from Westermeier R and Görg A (2004) Two-dimensional electrophoresis in proteomics. In *Protein Purification, Third Edition*, Janson JC (Ed.), John Wiley & Sons Inc., New York (in press).)

4.2. *Optimization of focusing conditions*

Settings are usually limited to 50 µA per strip and 150 V to avoid Joule heating, because the conductivity is initially high due to salts. As the run proceeds, the salt ions migrate to the electrodes, resulting in decreased conductivity and allowing high voltages to be applied. Final settings up to 8000 V are particularly used for zoom-in gels and alkaline pH gradients. The longer the IPG strip and the narrower the pH gradient, the more volt-hours are required to achieve steady state separation for high reproducibility (Görg *et al*., 2000; Görg *et al*., 2004).

4.3. *Storage of IPG strips after IEF*

If the second dimension cannot be performed directly after IEF, the IPG strips should be immediately frozen and stored at −70°C between two plastic sheets.

5. Equilibration of IPG gel strips

Before the second-dimension separation, it is essential that the IPG strips are equilibrated to allow the separated proteins to fully interact with SDS. Relatively

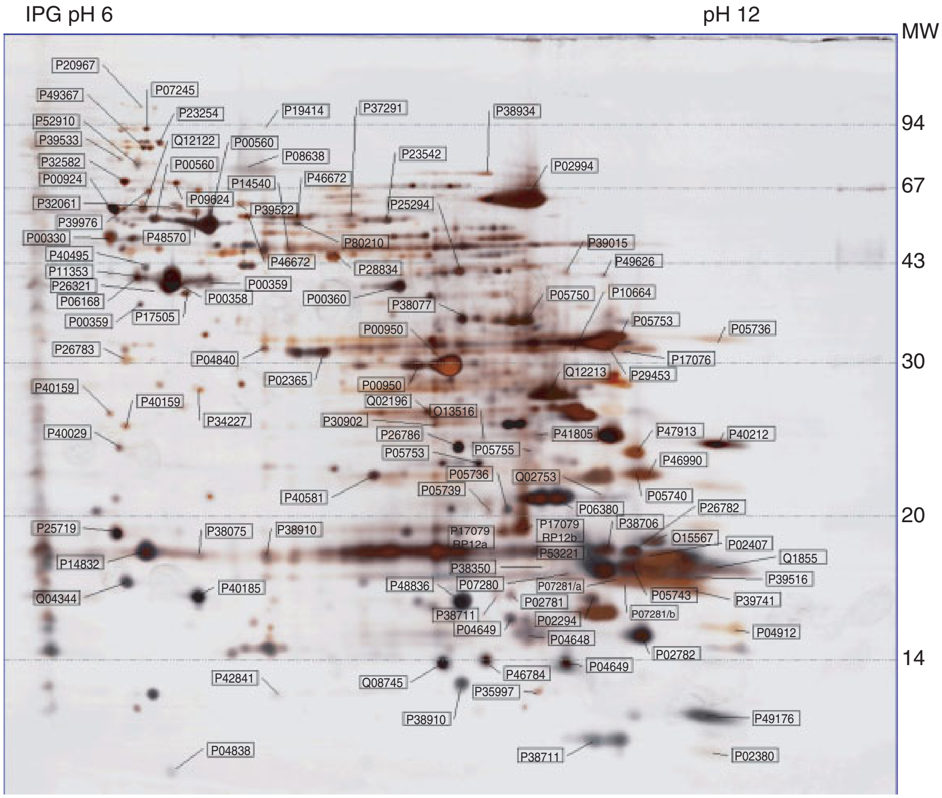


Figure 3 2DE of a TCA-acetone extract of *Saccharomyces cerevisiae* proteins, separated by IEF in an 18-cm-long IPG strip containing a linear pH gradient 6–12, followed by SDS-PAGE in a vertical 15% gel. Protein detection by silver staining showing the 106 mapped and identified spots annotated by Swiss-Prot accession numbers (Reproduced by permission of Wiley-VCH from Wildgruber R, Reil G, Drews O, Parlar H and Görg A (2002) Web-based two-dimensional database of *Saccharomyces cerevisiae* proteins using immobilized pH gradients from pH 6 to pH 12 and matrix-assisted laser desorption/ionization-time of flight mass spectrometry. *Proteomics,* **2**, 727–732.)

long equilibration times (10–15 min), as well as urea and glycerol (to reduce electroendosmotic effects) are required to improve protein transfer from the first to the second dimension. The best protocol by far is to incubate the IPG strips for 10–15 min in the Tris-HCl buffer originally described by Görg *et al*. (1988) (50 mM Tris-HCl (pH 8.8), containing 2% (w/v) SDS, 1% (w/v) dithiothreitol (DTT), 6 M urea and 30% (w/v) glycerol). This is followed by a further 10–15 min equilibration in the same solution containing 4% (w/v) iodoacetamide instead of DTT. The latter step is used to alkylate any free DTT, as otherwise it migrates through the second-dimension SDS-PAGE gel, resulting in an artifact known as *point streaking* that can be observed after silver staining. More importantly, the iodoacetamide alkylates sulfhydryl groups and prevents their reoxidation; this step is highly recommended for subsequent spot identification by mass spectrometry. After equilibration, the IPG strips are applied onto the surface of the second-dimension horizontal or vertical SDS-PAGE gels.

6. Second dimension: SDS-PAGE

SDS-PAGE can be performed on horizontal or vertical systems. Horizontal setups are ideally suited for ready-made gels (e.g., ExcelGel SDS; Amersham Biosciences), whereas vertical systems are preferred for multiple runs in parallel, in particular for large-scale proteome analysis which usually requires simultaneous electrophoresis of batches of second-dimension SDS-PAGE gels for higher throughput and maximal reproducibility. The most commonly used buffers for the second dimension of 2DE are the discontinuous buffer system of Laemmli (1970) and modifications thereof, although for special purposes other buffer systems are employed, such as borate buffers for the separation of highly glycosylated proteins (Patton *et al*., 1991), or Tris-Tricine buffer systems for the separation of low molecular mass (3–30 kDa) polypeptides (Schägger and von Jagow, 1987). Typically, gel sizes of $20 \times 25\,\text{cm}^2$ and a gel thickness of 1.0 mm are recommended. In contrast to horizontal SDS-PAGE systems, it is not necessary to use stacking gels with vertical setups, as the protein zones within the IPG strips are already concentrated and the nonrestrictive, low polyacrylamide concentration IEF gel acts as a stacking gel (Dunn and Görg, 2001).

7. Difference gel electrophoresis (DIGE)

A bottleneck for high-throughput proteomic studies is image analysis. In conventional 2D methodology, protein samples are separated on individual gels, stained, and quantified, followed by image comparison with computer-aided image analysis programs. Because multistep 2DE technology often prohibits different images from being perfectly superimposable, image analysis is frequently very time consuming. To shorten this laborious procedure, Ünlü *et al*. (1997) have developed a method called *fluorescent difference gel electrophoresis* (DIGE), in which two samples are labeled in vitro using two different fluorescent cyanine dyes (CyDyes, Amersham Biosciences) differing in their excitation and emission wavelengths, then mixed before IEF and separated on a single 2D gel. After consecutive excitation with both wavelengths, the images are overlaid and subtracted, whereby only differences (e.g., up- or downregulated, and/or posttranslationally modified proteins) between the two samples are visualized (Figure 4). Owing to the comigration of both samples, methodological variations in spot positions and protein abundance are excluded, and, consequently, image analysis is facilitated considerably (*see* Article 25, **2D DIGE**, Volume 5 and Article 30, **2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis**, Volume 5).

A third cyanine dye is now available, which makes it possible to include an internal standard, which is run on all gels within a series of experiments. This internal standard, typically a pooled mixture of all the samples in the experiment labeled with this dye, is used for normalization of data between gels thereby minimizing experimental variation and increasing the confidence in matching and quantitation of different gels in complex experimental designs (Alban *et al*., 2003). Applications that profit from the DIGE system include the investigation of differential protein

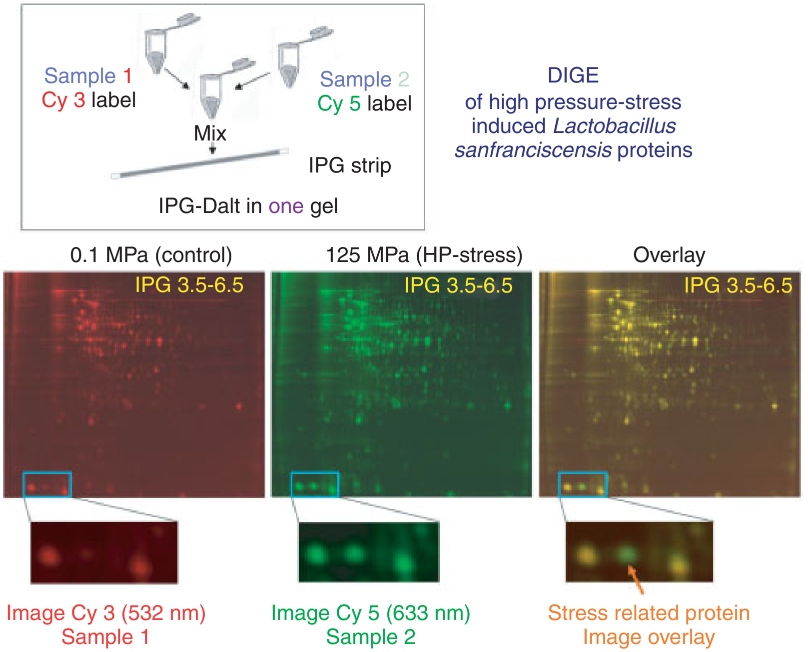


Figure 4 DIGE of high-pressure inducible *Lactobacillus sanfranciscensis* proteins. Samples (Control (at 0.1 Mpa) and high-pressure stressed (at 125 MPa)) were labeled in vitro with two different fluorescent cyanine dyes (Cy3 and Cy5, respectively) differing in their excitation and emission wavelengths. The samples were mixed, and the mixture was separated on a single 2DE gel. After consecutive excitation with both wavelengths, the resultant gel images were overlayed to visualize differences (e.g., up- or downregulated proteins) between the samples (Reproduced by permission of Cold Spring Harbor Laboratory Press from Görg A, Drews O and Weiss W (2004) Separation of proteins using two-dimensional gel electrophoresis. In *Purifying Proteins for Proteomics*, Simpson RJ (Ed.), Cold Spring Harbor Laboratory Press, New York, pp. 391–430.)

expression of samples generated under various prespecified conditions, the comparison of extracts, and the analysis of biological variance. In short, all analyses in which 2D gels need to be compared are simplified and accelerated by this method.

8. Protein visualization

After 2DE, the separated proteins have to be visualized, either by "universal" or by specific staining methods. The most important properties of stains are high sensitivity (low detection limit), high linear dynamic range (for quantitative accuracy), reproducibility, and compatibility with postelectrophoretic protein identification procedures, such as mass spectrometry (Dunn, 1993; *see also* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). Unfortunately, currently no staining method meets all requirements for proteome analysis.

Universal detection methods of proteins on 2D gels include staining with anionic dyes (e.g., Coomassie Blue), negative staining with metal cations (e.g., zinc imidazole), silver staining, fluorescence staining or labeling, and radioactive isotopes, using autoradiography, fluorography, or PhosphorImaging (Patton, 2002). For most of these staining procedures, the resolved polypeptides have to be fixed in solutions such as in ethanol/acetic acid/ H_2O for at least several hours (but usually overnight) before staining to remove any compounds (e.g., carrier ampholytes, detergents) that might interfere with detection.

Specific staining methods for detection of posttranslational modifications (e.g., glycosylation, phosphorylation) are employed either directly in the 2DE gel, or after transfer (blotting) onto an immobilizing membrane. The blotted proteins can be probed with specific antibodies (e.g., against phosphotyrosine residues) or with lectins (against carbohydrate moieties).

9. Image analysis

The major steps of computer-aided image analysis include (1) data acquisition, (2) spot detection and quantitation, (3) pattern matching, and (4) database construction (Dowsey *et al*., 2003; *see also* Article 26, **Image analysis**, Volume 5).

Currently, several 2D image analysis software packages are commercially available. Programs have been continuously improved and enhanced over the years in terms of faster matching algorithms with lesser manual intervention, and with focus on automation and better integration of data from various sources. New 2D software packages have also emerged which offer completely new approaches to image analysis and novel algorithms for more reliable spot detection, quantitation, and matching. Several programs include options such as control of a spot cutting robot, automated import of protein identification results from mass spectrometry, superior annotation flexibility (e.g., protein identity, mass spectrum, intensity/quantity, links to the Internet), and/or multichannel image merging of up to three different images to independent color channels for fast image comparison.

10. Protein identification from 2D gel spots

Mass spectrometry has become the technique of choice for identification of proteins from excised 2D gel spots as these methods are very sensitive, require small amounts of sample (femtomole to attomole concentrations) and have the capacity for high-sample throughput (Aebersold and Mann, 2003; *see also* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5). Recent advances in mass spectrometry also allow the investigation of posttranslational modifications including phosphorylation and glycosylation (Mann and Jensen, 2003; Kalume *et al*., 2003) (*see* Article 61, **Posttranslational modification of proteins**, Volume 6, Article 62, **Glycosylation**, Volume 6, Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6, and Article 73, **Protein phosphorylation analysis – a primer**, Volume 6). Peptide mass fingerprinting (PMF) is typically the primary tool for protein identification (*see* Article 3, **Tandem**

mass spectrometry database searching, Volume 5, Article 7, **Time-of-flight mass spectrometry**, Volume 5, Article 14, **Sample preparation for MALDI and electrospray**, Volume 5, Article 16, **Improvement of sequence coverage in peptide mass fingerprinting**, Volume 5, and Article 75, **Mass spectrometry**, Volume 6). This technique is based on the finding that a set of peptide masses obtained by MS analysis of a protein digest (usually trypsin) provides a characteristic mass fingerprint of that protein. The protein is then identified by comparison of the experimental mass fingerprint with theoretical peptide masses generated in silico using protein and nucleotide sequence databases. This approach proves very effective when trying to identify proteins from species whose genomes are completely sequenced, but is not so reliable for organisms whose genomes have not been completed.

If it proves impossible to identify a protein based on PMF alone, it is then essential to obtain amino acid sequence information. This is most readily accomplished with MALDI-MS with postsource decay (PSD) or chemical assisted fragmentation (CAF), or by using tandem mass spectrometry. MS/MS takes advantage of two-stage MS instruments, either MALDI-TOF-TOF-MS/MS or ESI-MS/MS triple quadrupole, ion-trap or Q-TOF machines (Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5, Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5, and Article 10, **Hybrid MS**, Volume 5) to induce fragmentation of peptide bonds. One approach is to generate a short partial sequence or 'tag', which is used in combination with the mass of the intact parent peptide ion to provide significant additional information for the homology search. A second approach uses a database searching algorithm SEQUEST to match uninterpreted experimental MS/MS spectra with predicted fragment patterns generated in silico from sequences in protein and nucleotide databases (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5).

11. 2D PAGE databases

Currently, enormous efforts are being undertaken to display and analyze with 2D PAGE the proteomes from a large number of organisms, ranging from *organelles* such as mitochondriae, nuclei or ribosomes to simple *prokaroytes* including *Escherichia coli*, *Bacillus subtilis*, *Haemophilus influenzae*, *Mycobacterium tuberculosis,* and *Helicobacter pylori*, to *single-celled eukaryotes* such as the yeast *Saccharomyces cerevisiae* (Figure 3), to *multicellular organisms*, for example, *Caenorhabditis elegans, plants* such as rice (*Oryza sativa*) or *Arabidopsis thaliana*, and *mammalian cells and tissues* including rat and human heart, mouse and human liver, mouse and human brain, different cancer cell lines, HeLa cells, human fibroplasts, human keratinocytes, rat and human serum and so on. Most of these and many other studies in progress are summarized at **www.expasy.org/ch2d/2d-index.html** ("WORLD-2DPAGE Index to 2D PAGE databases"). The Proteomics Standards Initiative (PSI) aims to define community standards for data representation in proteomics to facilitate data comparision, exchange, and verification (**http://psidev.sourceforge.net/**).

12. Concluding remarks

Although we have today a diversity of emerging proteomic platforms, there is still no generally applicable method that can replace 2DE in its ability to simultaneously separate and display several thousand proteins from complex samples such as microorganisms, cells, and tissues. 2DE using IPGs in the first dimension (IPG-Dalt) has proven to be extremely flexible with respect to the requirements of proteome analysis. Although by no means perfect, IPG-Dalt coupled with mass spectrometry remains the core technology for separating and identifying complex protein mixtures in proteomic projects at least for the foreseeable future.

Further reading

Pennington SR and Dunn MJ (Eds.) (2001) *Proteomics: from Protein Sequence to Function*, Border Information and Outreach Service: Oxford.

Simpson RJ (Ed.) (2002) *Proteins and Proteomics*, Cold Spring Harbor Laboratory Press: New York.

Simpson RJ (Ed.) (2004) *Purifying Proteins for Proteomics*, Cold Spring Harbor Laboratory Press: New York.

References

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207 (Review).

Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S and Currie I (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.

Blomberg A, Blomberg L, Norbeck J, Fey SJ, Larsen PM, Roepstorff P, Degand H, Boutry M, Posch A and Görg A (1995) Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis*, **16**, 1935–1945.

Corbett JM, Dunn MJ, Posch A and Görg A (1994) Positional reproducibility of protein spots in two-dimensional polyacrylamide gel electrophoresis using immobilised pH gradient isoelectric focusing in the first dimension: an interlaboratory comparison. *Electrophoresis*, **15**, 1205–1211.

Dowsey AW, Dunn MJ and Yang GZ (2003) The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, **3**, 1567–1596.

Drews O, Reil G, Parlar H and Görg A (2004) Setting up standards and a reference map for the alkaline proteome of the Gram-positive bacterium *Lactococcus lactis*. *Proteomics*, **4**, 1293–1304.

Dunn MJ (1993) *Gel Electrophoresis: Proteins*, Border Information and Outreach Service: Oxford.

Dunn MJ and Görg A (2001) Two-dimensional polyacrylamide gel electrophoresis for proteome analysis. In *Proteomics: from Protein Sequence to Function*, Pennington SR and Dunn MJ (Eds.), Border Information and Outreach Service: Oxford, pp. 43–63.

Görg A, Drews O and Weiss W (2004) Separation of proteins using two-dimensional gel electrophoresis. In *Purifying Proteins for Proteomics*, Simpson RJ (Ed.), Cold Spring Harbor Laboratory Press: New York, pp. 391–430.

Görg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R and Weiss W (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053.

Görg A, Postel W, Friedrich C, Kuick R, Strahler JR and Hanash SM (1991) Temperature-dependent spot positional variability in two-dimensional polypeptide patterns. *Electrophoresis*, **12**, 653–658.

Görg A, Postel W and Günther S (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **9**, 531–546.

Kalume DE, Molina H and Pandey A (2003) Tackling the phosphoproteome: tools and strategies. *Current Opinion in Chemical Biology*, **7**, 64–69 (Review).

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685.

Li J, Steen H and Gygi SP (2003) Potein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response. *Molecular & Cellular Proteomics*, **2**, 1198–1204.

Luche S, Santoni V and Rabilloud T (2003) Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics*, **3**, 249–253.

Mann M and Jensen ON (2003) Proteomic analysis of post-translational modifications. *Nature Biotechnology*, **21**, 255–261 (Review).

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, **250**, 4007–4021.

Olsson I, Larsson K, Palmgren R and Bjellqvist B (2002) Organic disulfides as a means to generate streak-free two-dimensional maps with narrow range basic immobilized pH gradient strips as first dimension. *Proteomics*, **2**, 1630–1632.

Patton WF (2002) Detection technologies in proteome analysis. *Journal of Chromatography B*, **771**, 3–31 (Review).

Patton WF, Chung-Welch N, Lopez MF, Cambria RP, Utterback BL and Skea WM (1991) Tris-tricine and Tris-borate buffer systems provide better estimates of human mesothelial cell intermediate filament protein molecular weights than the standard Tris-glycine system. *Analytical Biochemistry*, **197**, 25–33.

Schägger H and von Jagow G (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analytical Biochemistry*, **166**, 368–379.

Ünlü M, Morgan ME and Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Westbrook JA, Yan JX, Wait R, Welson SY and Dunn MJ (2001) Zooming-in on the proteome: very narrow-range immobilised pH gradients reveal more protein species and isoforms. *Electrophoresis*, **22**, 2865–2871.

Westermeier R and Görg A (2004) Two-dimensional electrophoresis in proteomics. In *Protein Purification*, Third Edition, Janson JC (Ed.), John Wiley & Sons: New York, (In press).

Wildgruber R, Harder A, Obermaier C, Boguth G, Weiss W, Fey SJ, Larsen PM and Görg A (2000) Towards higher resolution: two-dimensional electrophoresis of *Saccharomyces cerevisiae* proteins using overlapping narrow immobilized pH gradients. *Electrophoresis*, **21**, 2610–2616.

Wildgruber R, Reil G, Drews O, Parlar H and Görg A (2002) Web-based two-dimensional database of *Saccharomyces cerevisiae* proteins using immobilized pH gradients from pH 6 to pH 12 and matrix-assisted laser desorption/ionization-time of flight mass spectrometry. *Proteomics*, **2**, 727–732.

Specialist Review

ICAT and other labeling strategies for semiquantitative LC-based expression profiling

Connie Byrne and Gerard Cagney
University College Dublin, Dublin, Ireland

1. Introduction

Comprehensive proteome analysis requires the identification and quantification of all the proteins in an experimental or clinical sample. This is very much an ideal, because such samples often consist of thousands of proteins spanning a range of biophysical and biochemical properties, and that vary in abundance by many orders of magnitude. Fortunately, developments in protein and peptide ionization linked to mass spectrometry (MS) have made the identification of proteins almost routine (*see* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5). However, only a single peptide or protein (or small number) is usually presented to the MS detector at a time. For over two decades, two-dimensional (2D) gels have been the method of choice for separating proteins (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5), but separation strategies based on liquid chromatography (LC) are increasingly common. The LC apparatus may be interfaced directly to the instrument during electrospray experiments (Whitehouse *et al*., 1985). The resolving power of LC is based on factors that include the type of separation media, the solvent flow rate, and the gradient profile. Often, an orthogonal separation step is necessary to separate peptides in complex mixtures. Several configurations have been described (reviewed in Aebersold and Mann, 2003), with two approaches particularly common in recently published large-scale shotgun proteomics experiments. The first involves separating intact proteins using conventional SDS-PAGE gels before digestion and analysis by LC-MS (e.g., Lasonder *et al*., 2002), while another approach involves separation of peptides by ion exchange with LC in a single biphasic capillary column that acts as the electrospray source (Multidimensional Protein Identification Technology, “MUDPIT”; Link *et al*., 1999).

Very large numbers of proteins can be identified in individual samples using these approaches. In fact, experiments identifying thousands of proteins have now been described (e.g., Lasonder *et al*., 2002; Washburn *et al*., 2001; Schirle *et al*., 2003). However, in order to compare the states of the proteome of two or more samples, the quantity as well as the identity of each protein must be determined. Measurement of protein quantity is normally carried out by methods specific for a single protein:

Western blot, ELISA, or other immunological assays. High-throughput methods to estimate the abundance of large numbers of proteins in a single experiment have so far mostly used stable isotope-labeling methods (Hamdan and Righetti, 2002; Regnier *et al*., 2002). Isotopically labeled variants of proteins and peptides are readily detected in mass spectrometers because they vary in mass while having similar chemical properties. Here we describe general experimental strategies for semiquantitative proteomics involving LC-MS, issues to consider when carrying out such experiments, and some examples of the use of these approaches.

Two further comments should be made by way of introduction. First, although matrix-assisted laser desorption ionization (MALDI) MS does not use a direct LC interface to the instrument and as a method is not discussed here, the principles of stable isotope labeling for quantitative proteomics may equally be applied to MALDI-based instruments. Second, differential proteomics experiments attempt to answer the question "What protein levels are different in Sample A compared to Sample B?" However, other factors need to be considered when discussing "expression" proteomics, because this phrase implies some knowledge of the contribution of the synthesis and degradation rates of individual protein to the abundance of those proteins. While we will concern ourselves primarily with technical methods for measuring protein relative abundance, a useful discussion of these concepts can be found in Julka and Regnier (2004).

2. Strategies

There are two ways to experimentally introduce stable isotopes to the proteome (Figure 1). Using the first method, labeled amino acids are incorporated into the media of a cultured experimental system (in vivo). The second type of approach involves incorporating the label at the end of the experiment by chemical or enzymatic means (in vitro). In both cases, two (or more) samples are differentially labeled so that the signal of the corresponding peptides can be compared in the mass spectrometer to estimate their relative abundance. For this reason, an ideal semiquantitative proteomics strategy will have the following features:

- global labeling of all peptides or proteins to completion;
- a mass difference between the sister peptides where the heavier species is beyond the isotope shoulder of the lighter species;
- no detrimental effect of label incorporation on recovery or cleanup steps, mobility in chromatography steps, or ionization during MS;
- simple inexpensive experimental procedure that is easy to carry out on many samples in an identical manner, using any type of experimental model or clinical sample.

2.1. *In vivo methods*

In vivo methods are potentially global in that there is an opportunity for an entire proteome to become labeled with an amino acid containing ^{15}N or ^{13}C (Table 1).

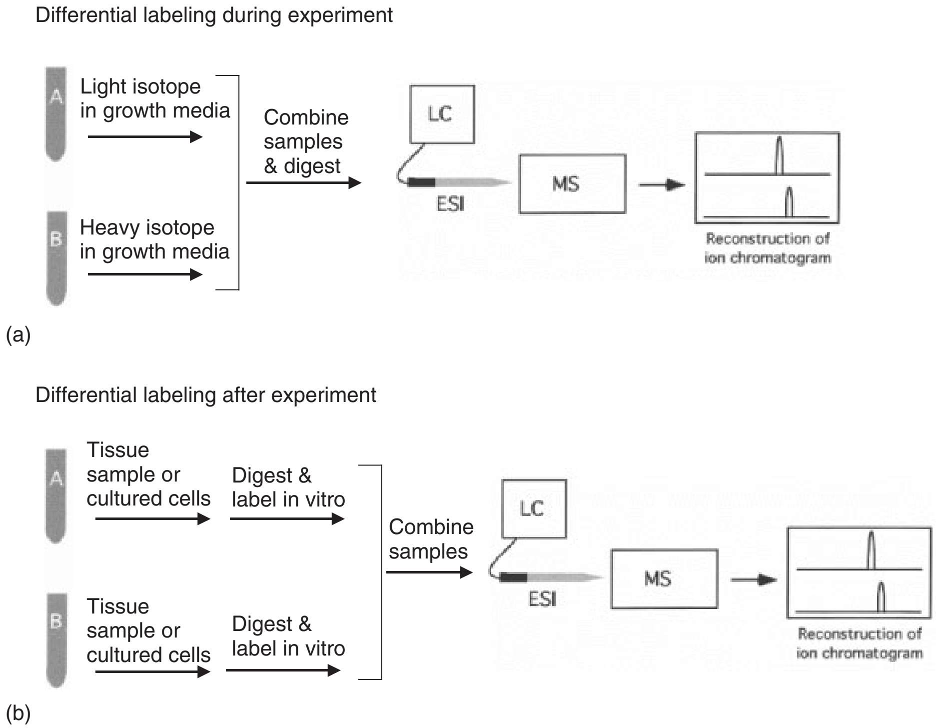


Figure 1 Differential labeling strategies for quantitative proteomics. The label may be incorporated in vivo if the experimental model may be cultured (a). Alternatively, experimental or clinical samples may be labeled in vitro (b)

The organism or cell is grown in culture long enough to allow complete incorporation of the labeled amino acid, and can then be used to compare with a strain grown on normal media. Advantages of this technique are that no chemical labeling or affinity purification steps are required, the method is compatible with virtually all cell culture conditions (including primary cells), and is relatively convenient and inexpensive. The method has been used with microbial organisms (Oda *et al*., 1999; Smith *et al*., 2002; Washburn *et al*., 2002), mammalian cells (Conrads *et al*., 2001), and even small animals fed with labeled yeast or bacteria (Krijgsveld *et al*., 2003).

Initially, ^{15}N media was used to grow microbes (Lahm and Langen, 2000) but a disadvantage of this method is that the exact number of labeled atoms in each peptide can only be inferred once the amino acid sequence is known. Others have used ^{2}H- or ^{13}C-labeled amino acids. The laboratory of Matthias Mann introduced the term SILAC (stable isotope labeling by amino acids in cell culture) in two elegant studies using mammalian cells (Ong *et al*., 2002; Ong *et al*., 2003). ^{13}C-labels were found to be superior to ^{2}H-labeled amino acids because they coelute during reversed-phase chromatography. When "sister" peptides (labeled and unlabeled) do not coelute, it makes relative analysis more complicated (see below). Deuterated amino acids are cheaper than ^{13}C-labeled ones; therefore, work that does not require the high quantitation precision afforded by the ^{13}C6-Arg method can be more economically performed using deuterated amino acids.

Table 1 Labeling technologies for quantitative proteomics

Method	Labeling reagent	Mass difference (Da)	Labeling chemistry	Labeled residue	Labeling step	Comments	References
ICAT	ICAT reagent (8 × ^{2}H)	+8	Thiol	Cysteine	In vitro	Solution phase	Gygi *et al*. (1999)
Cleavable ICAT	Cleavable and solid-phase ICAT reagent (7–9 × ^{13}C)	+7 to +9	Thiol	Cysteine	In vitro	Solid phase	Hansen *et al*. (2003); Zhou *et al*. (2002)
iTRAQ™	iTRAQ™ reagents	+1, +2, +3, +4 for each of the four labels	Amine	Lysine	In vitro	Commercially available (Applied Biosystems)	**www.appliedbiosystems.com**
O18	$H_2{}^{18}O$	+4	Carboxyl	C-terminal carboxylic acid generated by trypsin hydrolysis	In vitro	Drying step required, sometimes reduced efficiency if peptides fail to resuspend	Reynolds *et al*. (2002)
N15	^{15}N	Sequence dependent	In vivo incorporation		In vivo	Restricted to organisms that can be grown in defined media	Lahm and Langen (2000)
SILAC	$^{13}C6$-Arginine	+6	In vivo incorporation	Arginine	In vivo	SILAC	Ong *et al*. (2003)
	D3-Leucine	+3	In vivo incorporation	Leucine	In vivo	SILAC	Ong *et al*. (2002)
MCAT	*o*-methylisourea	+42	Guanidation	C-terminal lysine	In vitro		Krause *et al*. (1999)
Succinimide	($^{2}H_3$)-*N*-acetoxysuccinimide	+42 (arginine labeling) +84 (lysine labeling)	Acylation	Lysine and arginine	In vitro		Ji *et al*. (2000)

2.2. In vitro methods

Labeling using $H_2{}^{18}O$ incorporation is an extremely valuable method for semi-quantitative proteomics. The oxygen atom can be incorporated into peptide C-termini during peptide bond hydrolysis by trypsin, a routine step in proteomics experiments. The enzymology of this reaction has been extensively studied by Catherine Fenselau and others (Yao *et al*., 2001; Reynolds *et al*., 2002; Yao *et al*., 2003) and the incorporation ^{18}O of works for other proteases including chymotrypsin and endoglycosidase Lys C. Two ^{18}O atoms are normally incorporated, generating a 4-Da shift in the mass of the heavier species. This method can also result in improved sequence assignment during peptide sequencing by tandem MS because the ^{18}O atoms label the C terminus of the peptide fragment ions, resulting in specific shifting of the y-series (Schnolzer *et al*., 1996; Shevchenko *et al*., 1997). The method relies on replacement of normal water with ^{18}O-containing water for the labeled sample, in turn requiring a drying step. This can cause problems if the peptides do not resolubilize easily. Other potential problems reported for this method include varying rates of incorporation due to peptide sequence (Schnolzer *et al*., 1996; Stewart *et al*., 2001) and the presence of urea during proteolysis (Regnier *et al*., 2002).

To date, the majority of LC-based expression proteomics studies have used the isotope-coded affinity tag (ICAT) method developed by the Aebersold laboratory (Gygi *et al*., 1999). The ICAT reagent incorporates three moieties: an iodoacteyl group that can react with cysteine side chains under ambient conditions; a linker incorporating eight deuterium atoms; and a biotin group. The biotin allows very efficient recovery of the labeled peptides, while the use of thiol chemistry means that the diversity of the digested proteome is reduced (only peptides containing cysteine are labeled), greatly simplifying analysis by MS and subsequent interpretation. Recently, variant ICAT reagents that incorporate improvements on the original version have been described (Zhou *et al*., 2002; Hansen *et al*., 2003; Li *et al*., 2003; Oda *et al*., 2003). These variants use solid-phase or acid-cleavable reagents to improve peptide recovery and ease of automation, and replace the deuterium label with ^{13}C so that the retention times of heavy/light peptide pairs during reverse-phase chromatography is improved. A side-by-side comparison of the new reagent with the original one showed that it was more efficient and sensitive (Zhou *et al*., 2002). Additional reported advantages of the solid-phase method are that it is faster and simpler, requiring less manual input; it permits more stringent washing conditions to remove noncovalently associated molecules; and owing to the small size and chemical nature of the tag, MS/MS spectra are not complicated by undesirable fragmentation of the label itself.

Several additional stable isotope incorporation strategies have now been described, exploiting different aspects of peptide chemistry. The use of the carboxyl-terminus and cysteine side chains has been described above, but labeling of the amino-terminus, side-chain amino groups, side-chain carboxyl groups, and the side chains of tryptophan has also been used. Excellent review articles have been written describing the merits of these approaches (Julka and Regnier, 2004; Hamdan and Righetti, 2002) and they will only be mentioned briefly here.

Modification of primary amine groups by acylation (sometimes referred to as global internal standard technology, GIST; Chakraborty and Regnier, 2002) is generally easy to carry out. Reaction with ($^2H_3$)-*N*-acetoxysuccinimide generates a +42-Da shift upon binding the N-terminus (Ji *et al*., 2000). The reagent also labels the ε-amino group of lysine, so tryptic peptides ending with lysine will be +84-Da, while those ending in arginine with be +42 Da. In turn, the $^1H_3$- and $^2H_3$-acetate reagents differ by 3 Da. Related methods include modification of terminal amino groups with succinic anhydride (Munchbach *et al*., 2000), 2,4-dinitrofluorobenzene (Chen *et al*., 1999), and phenylisothiocyanate (Mason *et al*., 2003). Methyl esterification of peptide carboxyl groups modifies the side chains of aspartic and glutamic acid residues, as well as the peptide carboxyl terminus, to the corresponding methyl esters. The method was earlier used by mass spectrometrists to facilitate sequencing by tandem MS but reaction with either d0- or d3-methanol can also be used for expression proteomics (Goodlett *et al*., 2001). A strategy to specifically label tryptophan was devised, whereby tryptophan and cysteine residues were reacted with 2,4-dinitrobenzenesulfenyl chloride (containing ^{13}C) followed by reversal of the cysteine labeling by reduction/alkyation (Kuyama *et al*., 2003).

Two other approaches merit mention. The "quaternary amine tag" (QAT) methodology, recently described by Regnier and coworkers (Ren *et al*., 2004), involves reduction of cysteine residues with (3-acrylamidopropyl)trimethylammonium chloride. This procedure has several advantages, including simple chemistry, enhancement of peptide ionization by the tag, and the ability to effectively fractionate complex samples following tagging using cation exchange chromatography. A second approach that has been revisited is guanidination of lysine ε-amino groups. The method was originally used in MALDI MS to enhance signal from lysine-containing peptides, which is normally weaker than peptides containing arginine (Krause *et al*., 1999), but the modification is easy to carry out and adds 42 Da to the modified peptides (Hale *et al*., 2000; Beardsley *et al*., 2000; Brancia *et al*., 2000; Keough *et al*., 2000; Cagney and Emili, 2002).

Recently, a new approach has been described that differs in important respects from the methods described above. Termed tandem mass tagging (TMT), two related reagents are used to label alternative samples (Thompson *et al*., 2003). The TMT reagents have the same overall mass and reactive group (e.g., succinimide ester) so they migrate together during chromatography and have identical peptide reactivities. They have three addition groups: a mass normalization group (e.g., methionine), a cleavage enhancement group (proline), and a sensitization group (e.g., guanidino) which generates the so-called TMT fragment of known mass upon collision-induced dissociation (CID). Relative abundance determination is done by comparing the intensity of the TMT signals – because MS/MS is used for this, signal-to-noise ratio is very high and the sensitivity of this method should be greater than for other methods. Possible shortcomings are the need to carry out MS/MS scans (which limit the number of peptides that may be analyzed per unit time) and the difficulty in making these reagents. A similar set of reagents (iTRAQ) capable of measuring the relative abundance of peptides/proteins from four samples has recently been commercialized by Applied Biosystems.

3. Experimental issues

The power of high-throughput proteomics methods also has limitations. First and importantly, the stable isotope-labeling methods mentioned in this article measure *relative*, not *absolute* abundance. Moreover, they are generally pairwise (i.e., compare two samples with each other in a single experiment). The exceptions include iTRAQ, where up to four differentially labeled reagents have been developed (**www.appliedbiosystems.com**). For pairwise reagents, it is possible to design experiments where, for instance, all samples are compared to a single "reference" sample. Proposals to provide resources for large-scale absolute quantitative proteomics projects have been made recently (Aebersold, 2003).

Second, most of these strategies rely on the experiment and postexperiment handling being exactly equivalent. In cases in which this is in question, it may be desirable to carry out the experiment in both orientations, that is, in duplicate with the heavy/light labels switched in each replicate. The abundance changes can then be averaged. Related to this is the problem of different chromatographic retention times for compounds labeled using deuterium (Zhang *et al*., 2001) or ^{15}N (Conrads *et al*., 2001). This can usually be solved by making an alternative compound using ^{13}C, but sometimes this is not possible, and ^{13}C is much more expensive than deuterium.

Third, the maximum dynamic range that can be measured in a semiquantitative proteomics experiment by MS is inherently limited by the dynamic range of the signal that can be quantified in the mass spectrometer. This relates to another problem: cases where the low-abundance protein is not expressed at all or is expressed at levels that are difficult to distinguish from noise.

Fourth, many artifacts can arise in isotope-labeling experiments. For instance, the proteins or peptides may be incompletely labeled or a subclass of peptide may not be labeled at all. Examples include the ICAT reagent that does not modify cysteine-free peptides (but may be present twice in a peptide containing two cysteines) or ^{18}O-labeling during trypsin hydrolysis, where only a single ^{18}O atom may be incorporated instead of two. Correction for these factors may need to be carried out at the data interpretation stages. In some cases, the labeled peptide may be close in mass to the unlabeled peptide, (e.g., ^{18}O-labeling), resulting in overlap with the isotope shoulder of the light peptide. Again, these factors can be corrected for, but must be considered at the experiment design stage and especially when drawing conclusions at the end of the experiment.

Fifth, on the basis of relative abundance measurements for peptides derived from the same protein (which in general should be present in equimolar amounts), the measurement error for the methods described above are roughly in the range 10–50% (e.g., Shiio *et al*., 2002). This is large, so if the change in expression of an individual protein needs to be known with more precision, additional experiments need to be done. A complicating factor is the potential presence of protein modifications, for instance, cleavage products or phosphorylated peptides. When comparing signal from unmodified peptides derived from two samples, it is normally assumed that this represents all such peptides in the experiment. However, if a fraction of the peptides is actually modified, they may not be considered when determining relative abundance because they appear at a different *m/z* position in the mass spectra, thus overestimating the abundance of the protein in one sample.

However, the data is normally precise enough to establish that a protein is either *up*- or *downregulated*, and when this type of data is available for tens or hundreds of proteins, phenomena at the level of the cell rather than at the merely protein level may become apparent. The study of proteomic effects of Myc oncoprotein induction in rat fibroblasts was an early and impressive illustration of this principle (Shiio *et al*., 2002).

4. Conclusion/Future prospects

By combining stable isotope labeling and peptide reactive chemistries with mass spectrometry, a number of effective strategies of differential proteomics are now available. It is not clear if any of these approaches will become dominant, although the ICAT strategy has been used in a number of important pioneering differential proteomics studies. The choice of strategy will depend on factors such as type of sample (in vitro, clinical), the complexity of the sample, and mass spectrometry facilities. Unlike other differential gene expression studies, where similar reagents have been used for the majority of studies, the diversity of LC-based quantitative proteomics methods may create problems for those attempting to compare results from studies using different methodologies.

Further reading

Cahill DJ and Nordhoff E (2003) Protein arrays and their role in proteomics. *Advances in Biochemical Engineering/Biotechnology*, **83**, 177–187.

References

Aebersold R (2003) Constellations in a cellular universe. *Nature*, **422**(6928), 115–116.

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**(6928), 198–207.

Beardsley RL, Karty JA and Reilly JP (2000) Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **14**(23), 2147–2153.

Brancia FL, Oliver SG and Gaskell SJ (2000) Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Communications in Mass Spectrometry*, **14**(21), 2070–2073.

Cagney G and Emili A (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nature Biotechnology*, **20**(2), 163–170.

Chakraborty A and Regnier FE (2002) Global internal standard technology for comparative proteomics. *Journal of Chromatography. A*, **949**(1–2), 173–184.

Chen X, Chen YH and Anderson VE (1999) Protein cross-links: universal isolation and characterization by isotopic derivatization and electrospray ionization mass spectrometry. *Analytical Biochemistry*, **273**(2), 192–203.

Conrads TP, Alving K, Veenstra TD, Belov ME, Anderson GA, Anderson DJ, Lipton MS, Pasa-Tolic L, Udseth HR, Chrisler WB, *et al*. (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Analytical Chemistry*, **73**, 2132–2139.

Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R and Kolker E (2001) Differential stable isotope labelling of peptides for quantitation and de novo sequence derivation. *Rapid Communications in Mass Spectrometry*, **15**(14), 1214–1221.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Hale JE, Butler JP, Knierman MD and Becker GW (2000) Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine. *Analytical Biochemistry*, **287**(1), 110–117.

Hamdan M and Righetti PG (2002) Modern strategies for protein quantification in proteome analysis: advantages and limitations. *Mass Spectrometry Reviews*, **21**(4), 287–302.

Hansen KC, Schmitt-Ulms G, Chalkley RJ, Hirsch J, Baldwin MA and Burlingame AL (2003) Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-Isotope-coded affinity tag and multidimensional chromatography. *Molecular & Cellular Proteomics*, **2**(5), 299–314.

Ji J, Chakraborty A, Geng M, Zhang X, Amini A, Bina M and Regnier F (2000) Strategy for qualitative and quantitative analysis in proteomics based on signature peptides. *Journal of Chromatography. B, Biomedical Sciences and Applications*, **745**(1), 197–210.

Julka S and Regnier F (2004) Quantification in proteomics through stable isotope coding: a review. *Journal of Proteome Research*, **3**, 350–363.

Keough T, Lacey MP and Youngquist RS (2000) Derivatization procedures to facilitate de novo sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **14**(24), 2348–2356.

Krause E, Wenschuh H and Jungblut PR (1999) The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Analytical Chemistry*, **71**(19), 4160–4165.

Krijgsveld J, Ketting RF, Mahmoudi T, Johansen J, Artal-Sanz M, Verrijzer CP, Plasterk RH and Heck AJ (2003) Metabolic labelling of C. *Elegans and D. Melanogaster for Quantitative Proteomics. Nature Biotechnology*, **21**(8), 927–931.

Kuyama H, Watanabe M, Toda C, Ando E, Tanaka K and Nishimura O (2003) An approach to quantitative proteome analysis by labelling tryptophan residues. *Rapid Communications in Mass Spectrometry*, **17**(14), 1642–1650.

Lahm HW and Langen H (2000) Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis*, **21**, 2105–2114.

Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, *et al*. (2002) Analysis of the P*lasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*, **419**(6906), 537–542.

Li J, Steen H and Gygi SP (2003) Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: The yeast salinity stress response. *Molecular & Cellular Proteomics*, **2**(11), 1198–1204.

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM and Yates JR III (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, **17**(7), 676–682.

Mason DE and Liebler DC (2003) Quantitative analysis of modified proteins by LC-MS/MS of peptides labeled with phenyl isocyanate. *Journal of Proteome Research*, **2**(3), 265–272.

Munchbach M, Quadroni M, Miotto G and James P (2000) Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labelling of peptides with a fragmentation-directing moiety. *Analytical Chemistry*, **72**(17), 4047–4057.

Oda Y, Huang K, Cross FR, Cowburn D and Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6591–6596.

Oda Y, Owa T, Sato T, Boucher B, Daniels S, Yamanaka H, Shinohara Y, Yokoi A, Kuromitsu J and Nagasu T (2003) Quantitative chemical proteomics for identifying candidate drug targets. *Analytical Chemistry*, **75**(9), 2159–2165.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A and Mann M (2002) Stable isotope labelling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, **1**(5), 376–386.

Ong SE, Kratchmarova I and Mann M (2003) Properties of 13C-substituted arginine in stable isotope labelling by amino acids in cell culture (SILAC). *Journal of Proteome Research*, **2**(2), 173–181.

Regnier FE, Riggs L, Zhang R, Xiong L, Liu P, Chakraborty A, Seeley E, Sioma C and Thompson RA (2002) Comparative proteomics based on stable isotope labelling and affinity selection. *Journal of Mass Spectrometry*, **37**(2), 133–145.

Ren D, Julka S, Inerowicz HD and Regnier FE (2004) Enrichment of cysteine-containing peptides from tryptic digests using a quaternary amine tag. *Analytical Chemistry*, **76**(15), 4522–4530.

Reynolds KJ, Yao X and Fenselau C (2002) Proteolytic 18O labelling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *Proteome Research*, **1**(1), 27–33.

Schirle M, Heurtier M-A and Kuster B (2003) Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, **2**, 1297–1305.

Schnolzer M, Jedrzejewski P and Lehmann WD (1996) Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. *Electrophoresis*, **17**(5), 945–953.

Shevchenko A, Chernushevich I, Ens W, Standing KG, Thomson B, Wilm M and Mann M (1997) Rapid ‘de novo’ peptide sequencing by a combination of nanoelectrospray, isotopic labelling and a quadrupole/time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry*, **11**(9), 1015–1024.

Shiio Y, Donohoe S, Yi EC, Goodlett DR, Aebersold R and Eisenman RN (2002) Quantitative proteomic analysis of Myc oncoprotein function. *The EMBO Journal*, **21**(19), 5088–5096.

Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra TD and Udseth HR (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**(5), 513–523.

Stewart II, Thomson T and Figeys D (2001) 18O labelling: a tool for proteomics. *Rapid Communications in Mass Spectrometry*, **15**(24), 2456–2465.

Thompson A, Schaefer J, Kuhn K, Kiene S, Schwarz J, Schmidt G, Johnstone R, Neumann T and Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, **75**, 1895–1904.

Washburn MP, Ulaszek R, Deciu C, Schieltz DM and Yates JR III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Analytical Chemistry*, **74**(7), 1650–1657.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**(3), 242–247.

Whitehouse CM, Dreyer RN, Yamashita M and Fenn JB (1985) Related Articles, Links Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, **57**(3), 675–679.

Yao X, Afonso C and Fenselau C (2003) Dissection of proteolytic 18O labelling: endoprotease-catalyzed 16O-to-18O exchange of truncated peptide substrates. *Journal of Proteome Research*, **2**(2), 147–152.

Yao X, Freas A, Ramirez J, Demirev PA and Fenselau C (2001) Proteolytic 18O labelling for comparative proteomics: model studies with two serotypes of adenovirus. *Analytical Chemistry*, **73**(13), 2836–2842.

Zhang R, Sioma CS, Wang S and Regnier FE (2001) Fractionation of isotopically labeled peptides in quantitative proteomics. *Analytical Chemistry*, **73**(21), 5142–5149.

Zhou H, Ranish JA, Watts JD and Aebersold R (2002) Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology*, **20**(5), 512–515.

Specialist Review

Protein arrays

Derek Murphy and Dolores J. Cahill
Centre for Human Proteomics, Royal College of Surgeons in Ireland, Dublin, Ireland

1. Introduction

High-density DNA microarray technology has played a key role in the analysis of whole genomes and their gene expression patterns. The ability to study many thousands of individual genes, using oligonucleotide or cDNA arrays, is now very widespread, with its uses ranging from the profiling of gene expression patterns in whole organisms or tissues to the comparison of healthy and pathological samples. This technology together with the sequencing of the human genome has produced vast amounts of gene expression profiling data and, importantly, new bioinformatics tools to handle these data. Such data reveal important information regarding gene expression, however, the function of most genes lies at the protein level. Whether studying, for example, development in a particular organism or disease processes, knowledge of alternations in protein levels, protein structures (including modifications), protein–protein interactions, and so on, is crucial to elucidate with such complex biological phenomena. However, such knowledge is more difficult to attain, even if only one protein is being studied, due to the enormous complexity of the protein world. Alternative splicing, proteolytic events, and posttranslational modifications (e.g., glycosylation, acetylation, phosphorylation) are just some of the events that can occur to gene products, resulting in many more proteins than the genes coding for them. Further, proteins can form complexes with other proteins, cofactors, DNA/RNA, and so on.

Traditional methods for the analysis of proteomes include two-dimensional gel electrophoresis or chromatography, which, when combined with mass spectrometry, enables large-scale separation and identification of proteins, including many of their modifications (Melton, 2004). To complement the functional analysis of proteins on a large scale, proteins can be studied in array formats. Theoretically, such an array would contain functionally active proteins, in all their modified states, immobilized on a surface at high density or in solution in nanowells. However, protein activity is dependent on a wide range of factors (posttranslational modifications, cellular localization, pH, presence/absence of cofactors, etc.), which makes the production of a protein chip containing the whole proteome a daunting task. Current protein and antibody arrays represent the first steps toward this goal.

2. Generation of content for protein arrays

The first requirement toward the generation of protein arrays is a source of large numbers of recombinant proteins. A number of strategies are currently employed to provide sources of thousands of proteins for the generation of protein arrays. One approach is high-throughput cloning or amplification using PCR of defined open reading frames coding for the proteins of interest (Kersten *et al*., 2003; Reboul *et al*., 2003). The successful implementation of such an approach relies on the availability of sequence data and its correct annotation in the databases. In particular, the definition of the open reading frames of alternative splice variants of one protein remains difficult with such an approach. Also, previously uncharacterized proteins will be absent, limiting this approach as a discovery tool. For these reasons, this approach has proved most valuable in the production of chips containing proteins from well-characterized organisms, such as *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (Schweitzer *et al*., 2003).

Another approach is the use of protein expression libraries, which are used for a "shotgun" approach to the generation of recombinant proteins (Bussow *et al*., 1998). Such libraries are generated using mRNA isolated directly from the cell (Lueking *et al*., 2000; Bussow *et al*., 2000). In this approach, mRNA is isolated from the tissue or organism of interest and directionally subcloned into a protein expression vector suitable for heterologous expression in either a bacterial (e.g., *Escherichia coli*) or eukaryotic organism (e.g., yeast) – some solutions exist for the rapid movement of coding region from one organism to another, such as the GATEWAY system (Life Technologies) (Walhout *et al*., 2000) or dual expression vectors (Lueking *et al*., 2000). Not only does this approach circumvent the cloning of individual open reading frames, readily permitting the expression of tens of thousands of proteins, but also this strategy automatically includes splice variants and previously uncharacterized gene products, that is, there is no selection based on presently available databases and annotations.

Both strategies generally use a system by which the expression of the recombinant proteins is controlled by an inducible promoter. Thus, the time and duration of protein expression can be tightly regulated. One common and well-characterized system for expression in bacteria is the IPTG-inducible LacZ promoter. Because recombinant proteins are being expressed in a foreign environment, for example, human proteins expressed in *E. coli*, measures can be taken to improve expression levels. For example, plasmids coding for tRNAs, which are responsible for the translation of codons rarely found in bacteria but more commonly in mammals, can be introduced into the system, for example, the *argU* gene on the pSE111 vector that codes for a rare arginine tRNA (Bussow *et al*., 1998; Brinkmann *et al*., 1989). Other issues can also be addressed such as the expression of proteins that are toxic to the host organism. This remains a challenge; however, a number of systems have been developed to minimize basal level expression of the toxic recombinant protein in the host before induction. One such system for expression in *E. coli* is the pLysS vector (Stratagene), which is a low-copy-number plasmid that carries an expression cassette from which the T7 lysozyme gene is expressed at low levels. This T7 lysozyme binds to T7 RNA polymerase and inhibits transcription by this

enzyme. This approach has been used successfully for high-throughput expression of mammalian proteins in bacteria (Ding *et al*., 2002).

Apart from just the question of expressing the recombinant proteins of choice in a particular host system, there is also the question of tagging these proteins. A large number of tags exist, the coding region of which can readily be introduced into vectors. Depending on the approach taken, many options are available, including N-terminal and/or C-terminal tagging, multiple different tags are also an option, in particular, in conjunction with proteolytic sites for subsequent removal of the tag, if required. These tags can vary in size from the short His_6 tag to the 30-kDa GFP tag. Such tags are required for the detection of the expressed recombinant proteins and, in conjunction with an appropriate affinity separation method, for their purification (more steps of purification can be introduced using multiple tags). For a review of different expression systems, see Braun and LaBaer (2003).

In one example, a human fetal brain cDNA library was subcloned into a bacterial expression vector, permitting controlled IPTG-inducible expression of His6-tagged recombinant proteins. The *E. coli* clones of this library were then arrayed in high density onto PVDF membranes, grown overnight, and recombinant protein expression induced for a controlled length of time. Those clones expressing a human protein are readily detected by means of an anti-His_6 antibody. In this fashion, a protein array containing 10–12 000 different human proteins has been generated (Bussow *et al*., 1998). Approximately 66% of the proteins expressed in this library are, according to present annotation, full length. Also, such proteins have been shown to be readily expressed and purified in high throughput using available automated platforms (Braun *et al*., 2002; Lueking *et al*., 2003). It is such approaches that today allow the generation of protein microarrays.

3. Protein microarrays: toward the proteome on a chip

Once a source of proteins has been established, consideration must be given to the format that a protein microarray can have. For example, there are presently a number of possible surfaces available for the manufacture of such chips. In general, the available surfaces can be divided into two major categories: flat (planar) and 3D (mostly gel-like) surfaces (Angenendt *et al*., 2002; Angenendt *et al*., 2003). Planar surfaces have been primarily developed in the area of cDNA arrays. In general, the glass surface of these chips are treated to produce a thin layer of a particular chemical group, for example, an aldehyde group or poly-L-Lysine (Haab *et al*., 2001). The proteins are bound to the surface of these chips by either covalent bonds or simple electrical charge. Another type of planar chip surface available is plastic polymer-coated slides, such as the MaxiSorb slides from Nunc – an approach used in ELISAs for many years. The simple adoption of technology from the area of cDNA arrays brings with it a number of major concerns when used with proteins. First, unlike DNA, the surface charge of proteins is highly variable and the use of simple uniform electrostatic interaction for the immobilization of different proteins results in large variation in the amount of protein bound. Second, the structural conformation of proteins deposited onto a planar surface cannot be expected to very closely mimic "native" proteins, but would be more similar to

proteins present on membranes, such as in traditional Western blotting or dot blot experiments. Third, there is no control over the orientation of the proteins on the surface, which may result in, for example, the inaccessibility of an active site.

The development of the gel-like 3D chip surface was partly driven in an attempt to minimize the denaturation of the immobilized proteins in the arrays. A number of such surfaces exist based around polyacrylamide and agarose coated on a glass surface, which provide a hydrophilic environment for the proteins. Such surfaces allow the user to adjust various conditions, such as pH and salt concentrations, by incubating the chips in the appropriate buffer. Using a "homemade" polyacrylamide surface, a glass chip containing 2413 nonredundant purified human fusion proteins, arrayed at a density of up to 1600 proteins cm^{-2}, has been successfully employed in antibody binding studies, including the screening of human sera samples (Lueking *et al*., 2003), indicating the maintenance of a degree of structural conformation of the proteins involved. There are also non-gel-like 3D surfaces available, such as the FAST slides from Schleicher and Schuell, which have a nitrocellulose surface. Like any 3D surface, this surface will allow a much higher concentration of proteins per spot. One advantage of the nongel 3D surface is the increased shelf life of these slides.

However, these surface solutions still leave the problem of controlling the orientation of the proteins on the surface of the slides, which is necessary to maximize the activity of these proteins. For example, in order to maximize the binding of antibodies arranged on a surface to their epitopes, it would clearly be an advantage if the heavy chain were attached to, or close to, the surface and the antigen-binding site as far away from the surface as possible. One approach has been the use of affinity tags, for example, a nickel-coated slide has a natural and specific affinity to His_6-tagged recombinant proteins. This approach was used to array 5800 yeast proteins that were screened for their ability to interact with calmodulins and phospholipids (Zhu *et al*., 2001). Similarly, successful orientation of antibodies and antibody Fab fragments was achieved by biotinylating such antibodies/Fab fragments and arraying them on a streptavidin-coated surface (Peluso *et al*., 2003). A further development of surface chemistry involves the use of a polyethylene glycol layer (PEG) (Angenendt *et al*., 2003) or dendrimers (Benters *et al*., 2002; Benters *et al*., 2001). These approaches involve the coupling of the proteins to epoxy groups, which act as spacers preventing direct protein-surface contact and, thus, eliminate the need for blocking reagents to reduce background binding. One further development of this approach has been to link chelating iminodiacetic acid groups to PEG, which in turn can be bound by Cu^{2+} ions and so provide a highly specific binding site for His_6-tagged proteins (Cha *et al*., 2004). The authors demonstrated one additional, and potentially very important, use for such technology and that is the elimination of the need for prepurification of tagged recombinant proteins for arrays. A number of studies have been carried out comparing the different surfaces available for protein array work, including antibody arrays that assess background noise, sensitivity/detection limits, reproducibility, and storage for a variety of experimental designs (Angenendt *et al*., 2002; Angenendt *et al*., 2003).

4. Microfluidic chips

While the development of 3D surfaces and spacers goes some way to address the problems faced when looking at protein–protein interactions on a chip, many experiments that involve interactions of proteins in a functional state will prove difficult to perform on these chips, which is obviously a major drawback of the current protein arrays. Ideally, we would like to look at protein interactions, where the proteins are in their native state and are functional, that is, in conditions as close as possible to those in nature. This may be solved by developing a microfluidic chip, which is a series of enclosed microchannels within a chip format, such as silicon, plastic, or glass. The potential of microfluidic chips would include the ability to maintain proteins in their functional conformations and to therefore perform interactions such as protein–protein, protein–peptide, protein–compound, protein–DNA, protein–ligand, and protein–antibody interactions in solution. Also, the area of enzyme studies would greatly profit from such a system. In fact, the first steps have been taken to use a "lab-on-a-chip" to study some of the reactions in the glycolytic pathway of yeast (Young *et al*., 2003). Using this system, enzymatic reactions in volumes as low as 6.3–8 nL could be studied (Dietrich *et al*., 2004). The ability to use such small volumes in microfluidic chips is very important due to the high cost of proteins, antibodies, compound libraries, peptides, and even the difficulties in obtaining enough patient samples to screen in this format, when screening in high throughput, can be prohibitive. Such technology would also allow us to fully exploit the libraries of proteins that currently exist.

Recent developments show that microchannels can be generated in plastic, where the microchannels have a diameter of $100\,\mu\text{m} \times 100\,\mu\text{m}$, can be centimeters in length, and can be fabricated using glass, silicon, and plastic (Guber *et al*., 2004). However, there are current technological challenges, which need to be addressed, such as loading, sample positioning and manipulation of samples in these chips, handling nanoliter volumes, and scanning the chip.

5. Applications of protein arrays

Theoretically, present protein arrays provide a tool to examine interactions with proteins, whether they be with other proteins (including antibodies), peptides, DNA/RNA, or chemical compounds on a large scale. In one proof-of-principle experiment, a small number of well-defined protein–protein interactions, including an interaction dependent on a small molecule, were demonstrated in a microarray format (MacBeath and Schreiber, 2000). While still somewhat in its infancy, a number of studies have been completed in which high-throughput screening of protein microarrays has proven successful. As mentioned above, calmodulin- and phospholipids-interacting proteins have been identified by screening almost 6000 yeast proteins, generated by expressing previously annotated open reading frames (Zhu *et al*., 2001). A human protein expression library, involving in situ expression of tens of thousands of recombinant proteins on large membranes (Bussow *et al*., 1998; Lueking *et al*., 1999), has been successfully screened to demonstrate antibody–protein interactions. This same library has also provided a

source of purified recombinant proteins for microarrays, which were demonstrated to be a successful platform for the large-scale study of protein–antibody interactions (Lueking *et al*., 2003), an important step toward large-scale studies of antibody specificity. By screening thousands of different proteins from the relevant organism with a particular antibody, it is possible to determine which proteins contain antigens recognized by the antibody in question. In another study (Michaud *et al*., 2003), 11 polyclonal and monoclonal antibodies were screened against 5000 yeast proteins and the results demonstrated the cross-reactivities of many of the antibodies screened.

One further development of the study of protein–antibody interactions is using high-content protein arrays to determine the target of antibodies previously identified as potentially interesting markers in disease. Our group is presently working to identify the antigens of antibodies, initially identified by immunohistochemistry as interesting markers in certain types of cancer (Figure 1 shows a pipeline for this work). Each antibody can be screened against an appropriate protein expression library, arrayed on either PVDF membranes or, as purified proteins, arrayed on glass chips. Unlike previous approaches, which could at best identify potential epitopes of a particular antibody, this approach identifies the actual target protein. In addition, potential cross-reacting proteins can also be identified using this method, which is important information for assessing an antibody in disease-screening procedures. Similarly, protein arrays could provide a useful screening technique to be introduced into any antibody production process. For example, during the production of monoclonal antibodies in mice immunized with a particular antigen,

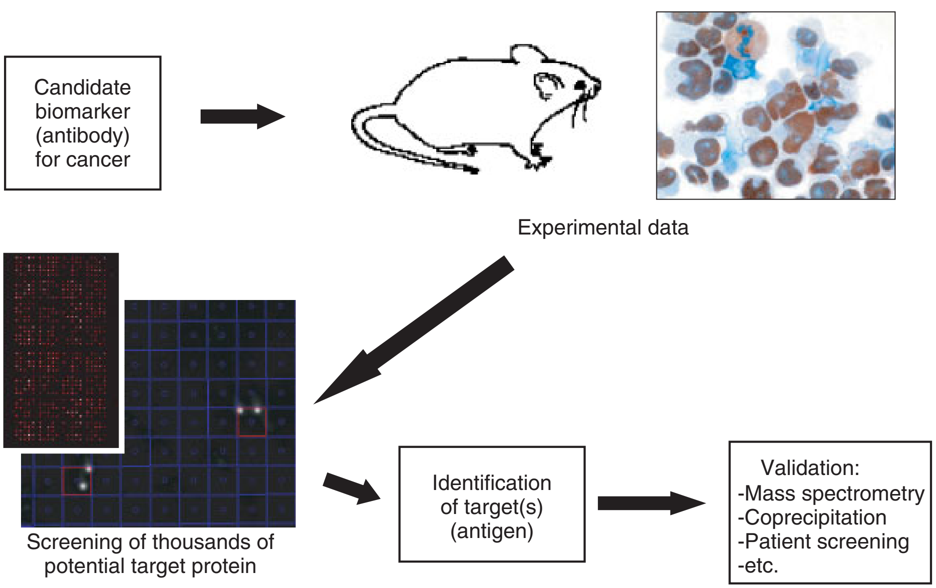


Figure 1 Pipeline showing integration of protein array technology into cancer marker discovery and characterization. Antibodies that are potentially useful as cancer markers are produced and tested, for example, using immunohistochemistry. Protein array technology can then be introduced to confirm/discover the protein target bound by the antibody, and assess the specificity/cross-reactivity

the antibody-producing B cells are fused with myeloma cells to form hybridomas, which can then be cultured for large-scale expression of the antibody. During this process, the supernatant from each hybridoma clone is screened for the ability of the antibody present to bind to the target antigen. By introducing a screen of a large-content protein array at this point, the antibody could be tested not only for its ability to bind the known antigen, but valuable data on specificity/cross-reactivity could also be generated.

Just as it is possible to characterize the binding of a single antibody using protein arrays, it is also possible to characterize many antibodies present in a single sample, such as to profile the antibody repertoire in serum or plasma. One immediate application is the use of "allergen arrays" to screen for the presence of particular IgE molecules in a patient sample. The traditional approach involves the use of simple extracts from potential allergens. Such extracts, containing both allergens and nonallergens, are commonly used in skin prick tests to determine the possible source of an allergic reaction in the patient. Using modern arraying technology and recombinant allergens (e.g., pollen and fungus proteins) relatively large arrays have been produced for screening purposes (Hiller *et al*., 2002; Deinhofer *et al*., 2004; Jahn-Schmid *et al*., 2003; Wiltshire *et al*., 2000). These arrays can also readily include nonprotein allergens such as latex, and so on. Such arrays are readily miniaturized, permitting the screening of very low volumes of a patient's blood, and are also more accurate in identifying the precise source of the allergic response.

In a similar fashion, protein arrays can be used to profile antibodies present in the blood of patients with autoimmune diseases. Initial autoantigen arrays consisted of almost 200 proteins, peptides, and other biomolecules (including several forms of dsDNA and ssDNA), which were known autoantigens to several well-characterized autoimmune diseases, including rheumatoid arthritis and systemic lupus erythematosus (Robinson *et al*., 2002). Another approach is to use large arrayed libraries of recombinant proteins as a potential autoantigen array. In a proof-of-principle experiment, a protein array chip containing almost 2500 purified recombinant human proteins was used to profile the autoantibodies present in small volume samples from patients with alopecia and rheumatoid arthritis (Lueking *et al*., 2003). This approach permitted the identification of previously known autoantigens and also previously uncharacterized protein autoantigens. Initial results from screening a large recombinant mouse protein array also indicate the usefulness of this approach to characterize autoimmune disease in a mouse model system for systemic lupus erythematosus (C. Gutjahr, 2005). Protein arrays can contribute enormously to our understanding of the general mechanisms involved in autoimmunity, as well as provide a platform on which to develop the technology for more complete diagnostics of the various autoimmune diseases.

Presently, the use of protein array screening to identify protein–protein (nonantibody) interactions is still very much in its infancy and limited to specialized laboratories. Many proteins are difficult to study in solution (e.g., membrane proteins) or may require the presence of various cofactors. Another approach to screening "difficult" proteins is the use of peptide screening.

Figure 2 shows an outline of the approach taken by our group to identify proteins that interact with the cytoplasmic tail of a membrane protein, in this

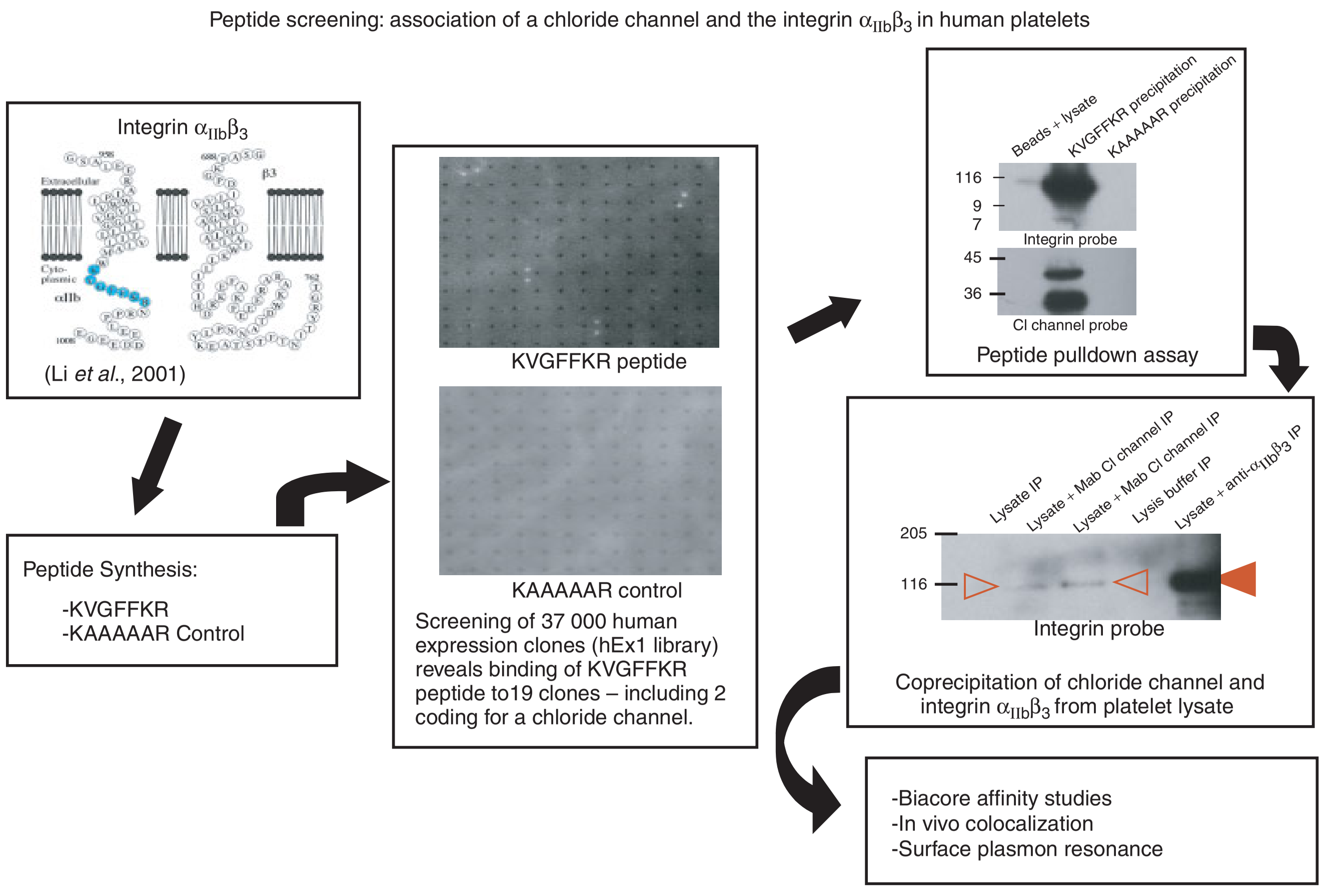


Figure 2 Example of the identification of protein–protein interactions using protein arrays. In order to identify proteins that potentially interact with the cytoplasmic tail of the platelet integrin $\alpha_{IIb}\beta_3$, a labeled peptide of this region was generated and a protein array library with over 37 000 clones was screened. One of the proteins identified, the chloride channel ICln, was further characterized and confirmed as an interaction partner of the integrin in biological systems (Larkin *et al*., 2004) (Reproduced by permission of American Society for Biochemistry & Molecular Biology)

case a platelet integrin (Larkin *et al*., 2004). The particular conserved α-integrin cytoplasmic motif, KVGFFKR, had previously been shown to play a critical role in the regulation of activation of the platelet integrin $\alpha_{IIb}\beta_3$ (Stephens *et al*., 1998). In order to discover the molecular mechanisms involved in this regulation, it is necessary to discover what proteins are interacting with this integrin, and more specifically with this region of the cytoplasmic tail. In order to overcome the difficulties associated with working with entire membrane proteins, a tagged peptide (biotin-KVGFFKR) was synthesized corresponding to the region of interest. This peptide was then screened against a high-density array of 37 000 *E. coli* clones expressing recombinant human proteins (Bussow *et al*., 1998; Lueking *et al*., 1999), and 19 clones, coding for 13 different proteins, were identified as binding the labeled peptide. Of these 19 clones, a strong binding could be shown between this labeled peptide and purified proteins isolated from three clones. Of these clones, one codes for a protein that could not be shown to be present in platelets, and two code for a putative chloride channel, ICln, shown for the first time to be present in platelets. A number of experiments were carried out, including peptide pulldown assays and coprecipitation experiments, to confirm the interaction between ICln and integrin $\alpha_{IIb}\beta_3$ (see Figure 2 and Larkin *et al*., 2004). Such an experiment reveals the enormous potential of a protein array approach, not only in identifying novel protein interactions but also toward teasing apart biological pathways in general.

Applications of protein array technology such as target identification and characterization, target validation, diagnostic marker identification and validation, preclinical study monitoring, and patient typing seem to be feasible. We have recently reviewed the value of recent combinations of efforts in genomics, proteomics, and biochip technology and their impact on the overall drug development process (Huels *et al*., 2002). For the first time, tools are available to study disturbances within biological systems, such as disease or drug treatment, on the gene and protein expression levels.

Proteins, as targets, dominate pharmaceutical R&D with ligand–receptor interactions and enzymes as the vast majority, comprising ~45% and 28%, respectively, of the targets (Drews, 2000). Additionally, many therapeutic proteins, especially humanized antibodies, are in clinical development. The ultimate tool for high-throughput and significant screening would be to test new leads or new targets in a highly parallel manner. Some examples for applications in this direction already exist. Recently, an immunosensor array has been developed that enables the simultaneous detection of clinical analytes (Rowe *et al*., 1999). Here, capture antibodies and analytes were arrayed on microscope slides using flow chambers in a cross-wise fashion. This current format is low-density (6×6 pattern) but has high-throughput potential, as it involves automated image analysis and microfluidics; it is already becoming one of the future formats for enzyme activity testing and other assays (Cohen *et al*., 1999). In another study, small sets of active enzymes were immobilized in a hydrophilic gel matrix. Enzymatic cleavage of the substrate could be detected and inhibitors blocked the reaction (Arenkov *et al*., 2000). More recently, an enzyme array that is suitable for assays of enzyme inhibition has been reported (Park and Clark, 2002). Initial publications in the area of receptor–ligand interaction studies in a microarray format have shown that the interaction of immobilized compounds and proteins in solutions can be determined (Zhu *et al*.,

2001; MacBeath and Schreiber, 2000; Mangold *et al*., 1999). This technology allows high-throughput screening of ligand–receptor interactions with small sample volumes.

The multiparallel possibilities of protein array applications have the potential to not only allow the optimization of preclinical, toxicological, and clinical studies through better selection and stratification of individuals but also to effect how diagnostics are used in drug development.

References

Angenendt P, Glokler J, Murphy D, Lehrach H and Cahill DJ (2002) Toward optimized antibody microarrays: A comparison of current microarray support materials. *Analytical Biochemistry*, **309**(2), 253–260.

Angenendt P, Glokler J, Sobek J, Lehrach H and Cahill DJ (2003) Next generation of protein microarray support materials: evaluation for protein and antibody microarray applications. *Journal Of Chromatography. A*, **1009**(1–2), 97–104.

Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V and Mirzabekov A (2000) Protein microchips: Use for immunoassay and enzymatic reactions. *Analytical Biochemistry*, **278**(2), 123–131.

Benters R, Niemeyer CM, Drutschmann D, Blohm D and Wohrle D (2002) DNA microarrays with PAMAM dendritic linker systems. *Nucleic Acids Research*, **30**(2), E10.

Benters R, Niemeyer CM and Wohrle D (2001) Dendrimer-activated solid supports for nucleic acid and protein microarrays. *Chembiochem*, **2**(9), 686–694.

Braun P, Hu Y, Shen B, Halleck A, Koundinya M, Harlow E and LaBaer J (2002) Proteome-scale purification of human proteins from bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(5), 2654–2659.

Braun P and LaBaer J (2003) High throughput protein production for functional proteomics. *Trends in Biotechnology*, **21**(9), 383–388.

Brinkmann U, Mattes RE and Buckel P (1989) High-level expression of recombinant genes in Escherichia coli is dependent on the availability of the DNAY gene product. *Gene*, **85**(1), 109–114.

Bussow K, Cahill D, Nietfeld W, Bancroft D, Scherzinger E, Lehrach H and Walter G (1998) A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Research*, **26**(21), 5007–5008.

Bussow K, Nordhoff E, Lübbert C, Lehrach H and Walter G (2000) A human cDNA library for high-throughput protein expression screening. *Genomics*, **65**(1), 1–8.

Cha T, Guo A, Jun Y, Pei D and Zhu XY (2004) Immobilization of oriented protein molecules on poly(ethylene glycol)-coated Si(111). *Proteomics*, **4**(7), 1965–1976.

Cohen CB, Chin-Dixon E, Jeong S and Nikiforov TT (1999) A microchip-based enzyme assay for protein kinase A. *Analytical Biochemistry*, **273**(1), 89–97.

Deinhofer K, Sevcik H, Balic N, Harwanegg C, Hiller R, Rumpold H, Mueller MW and Spitzauer S (2004) Microarrayed allergens for IgE profiling. *Methods*, **32**(3), 249–254.

Dietrich HR, Knoll J, Van Den Doel LR, Van Dedem GW, Daran-Lapujade PA, Van Vliet LJ, Moerman R, Pronk JT and Young IT (2004) Nanoarrays: A method for performing enzymatic assays. *Analytical Chemistry*, **76**(14), 4112–4117.

Ding HT, Ren H, Chen Q, Fang G, Li LF, Li R, Wang Z, Jia XY, Liang YH, Hu MH, *et al*. (2002) Parallel cloning, expression, purification and crystallization of human proteins for structural genomics. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 12), 2102–2108.

Drews J (2000) Drug discovery: A historical perspective. *Science*, **287**(5460), 1960–1964.

Guber A, Heckele M, Herrmanna D, Muslijaa A, Sailea V, Eichhornb L, Gietzeltb T, Hoffmannc W, Hauserd PC, Tanyanyiwad J, *et al*. (2004) Microfluidic lab-on-a-chip systems based on polymers – fabrication and application. *Chemical Engineering Journal*, **101**(1–3), 447–453.

Gutjahr C, Murphy D, Lueking A, Koenig A, Janitz M, O'Brien J, Korn B, Horn S, Lehrach H and Cahill DJ (2005) Mouse protein arrays from a T(H)1 cell cDNA library for antibody screening and serum profiling. *Genomics*, **85**(3), 285–296.

Haab BB, Dunham MJ and Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**(2), research0004.1–research0004.13.

Hiller R, Laffer S, Harwanegg C, Huber M, Schmidt WM, Twardosz A, Barletta B, Becker WM, Blaser K, Breiteneder H, *et al*. (2002) Microarrayed allergen molecules: Diagnostic gatekeepers for allergy treatment. *The FASEB Journal: Official Publication of The Federation of American Societies for Experimental Biology*, **16**(3), 414–416.

Huels C, Muellner S, Meyer HE and Cahill DJ (2002) The impact of protein biochips and microarrays on the drug development process. *Drug Discovery Today*, **7**(18 Suppl), S119–S124.

Jahn-Schmid B, Harwanegg C, Hiller R, Bohle B, Ebner C, Scheiner O and Mueller MW (2003) Allergen microarray: Comparison of microarray using recombinant allergens with conventional diagnostic methods to detect allergen-specific serum immunoglobulin E. *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology*, **33**(10), 1443–1449.

Kersten B, Feilner T, Kramer A, Wehrmeyer S, Possling A, Witt I, Zanor MI, Stracke R, Lueking A, Kreutzberger J, *et al*. (2003) Generation of Arabidopsis protein chips for antibody and serum screening. *Plant Molecular Biology*, **52**(5), 999–1010.

Larkin D, Murphy D, Reilly DF, Cahill M, Sattler E, Harriott P, Cahill DJ and Moran N (2004) ICln, a novel integrin alphaIIbbeta3-associated protein, functionally regulates platelet activation. *The Journal of Biological Chemistry*, **279**(26), 27286–27293.

Lueking A, Horn M, Eickhoff H, Bussow K, Lehrach H and Walter G (1999) Protein microarrays for gene expression and antibody screening. *Analytical Biochemistry*, **270**(1), 103–111.

Lueking A, Holz C, Gotthold C, Lehrach H and Cahill D (2000) A system for dual protein expression in *Pichia pastoris* and *Escherichia coli*. *Protein Expression and Purification*, **20**(3), 372–378.

Lueking A, Possling A, Huber O, Beveridge A, Horn M, Eickhoff H, Schuchardt J, Lehrach H and Cahill DJ (2003) A nonredundant human protein chip for antibody screening and serum profiling. *Molecular & Cellular Proteomics*, **2**(12), 1342–1349.

MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**(5485), 1760–1763.

Mangold U, Dax CI, Saar K, Schwab W, Kirschbaum B and Mullner S (1999) Identification and characterization of potential new therapeutic targets in inflammatory and autoimmune diseases. *European Journal of Biochemistry*, **266**(3), 1184–1191.

Melton L (2004) Proteomics in multiplex. *Nature*, **429**(6987), 101–107.

Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, Snyder M, Predki PF and Schweitzer BI (2003) Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnology*, **21**(12), 1509–1512.

Park CB and Clark DS (2002) Sol-gel encapsulated enzyme arrays for high-throughput screening of biocatalytic activity. *Biotechnology and Bioengineering*, **78**(2), 229–235.

Peluso P, Wilson DS, Do D, Tran H, Venkatasubbaiah M, Quincy D, Heidecker B, Poindexter K, Tolani N, Phelan M, *et al*. (2003) Optimizing antibody immobilization strategies for the construction of protein microarrays. *Analytical Biochemistry*, **312**(2), 113–124.

Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, *et al*. (2003) C. elegans ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics*, **34**(1), 35–41.

Robinson WH, DiGennaro C, Hueber W, Haab BB, Kamachi M, Dean EJ, Fournel S, Fong D, Genovese MC, de Vegvar HE, *et al*. (2002) Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Medicine*, **8**(3), 295–301.

Rowe CA, Scruggs SB, Feldstein MJ, Golden JP and Ligler FS (1999) An array immunosensor for simultaneous detection of clinical analytes. *Analytical Chemistry*, **71**(2), 433–439.

Schweitzer B, Predki P and Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. *Proteomics*, **3**(11), 2190–2199.

Stephens G, O'Luanaigh N, Reilly D, Harriott P, Walker B, Fitzgerald D and Moran N (1998) A sequence within the cytoplasmic tail of GpIIb independently activates platelet aggregation and thromboxane synthesis. *The Journal Of Biological Chemistry*, **273**(32), 20317–20322.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N and Vidal M (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**(5450), 116–122.

Wiltshire S, O-Malley S, Lambert J, Kukanskis K, Edgar D, Kingsmore SF and Schweitzer B (2000) Detection of multiple allergen-specific IgEs on microarrays by immunoassay with rolling circle amplification. *Clinical Chemistry*, **46**(12), 1990–1993.

Young IT, Moerman R, Van Den Doel LR, Iordanov V, Kroon A, Dietrich HR, Van Dedem GW, Bossche A, Gray BL, Sarro L, *et al*. (2003) Monitoring enzymatic reactions in nanolitre wells. *Journal of Microscopy*, **212**(Pt 3), 254–263.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**(5537), 2101–2105.

2D DIGE

Kathryn S. Lilley
University of Cambridge, Cambridge, UK

1. Introduction

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) has been widely used over the past four decades to resolve several thousand proteins in a single sample. This has enabled the identification of the major proteins in a tissue or subcellular fraction by mass spectrometric methods. In addition, 2D PAGE has been used to compare relative abundances of proteins in related samples, such as between mutant and wild type organisms or control and diseased tissues, allowing the response of classes of proteins to be determined. To date, the majority of comparative protein-profiling studies have produced qualitative data, which have enabled the investigator to determine whether or not a particular protein shows an increase or decrease in expression. This provides no measure of the extent of this change in expression, therefore, it is unsuitable for the clustered data analysis needed for an insight into functionality. Quantitative proteomics allows coexpression patterns to be studied, and proteins showing similar expression trends may be assigned membership of the same functional groups; however, quantitation has been hampered by several factors. Firstly, silver staining, being more sensitive than Coomassie staining methods, has been widely used for high-sensitivity protein visualization on 2D PAGE, but it is unsuitable for quantitative analysis as it has a limited dynamic range. The most sensitive silver staining methods are also incompatible with protein-identification methods based on mass spectrometry. More recently, the Sypro family of postelectrophoretic fluorescent stains (Molecular Probes, Eugene, Oregon, USA) have emerged as alternatives, offering a better dynamic range than silver staining and ease of use (Malone *et al*., 2001). Another problem is the irreproducibility of 2D gels; no two gels run identically meaning that corresponding spots between two gels have to be matched prior to quantification. Finally, normalization has proved challenging, especially in the case of silver staining where staining is protein dependent. These aforementioned factors all add variability to the system that make it unsuitable for the accurate quantitation.

Difference gel electrophoresis (DIGE) circumvents many of the issues associated with traditional 2D PAGE, such as gel-to-gel variation and limited dynamic range, and allows more accurate and sensitive quantitative proteomics studies. This minireview is an overview of this technique, describing its strengths and limitations.

The DIGE technique was first described some time ago by Jon Minden's laboratory (Ünlü *et al*., 1997) and is now available as a technique from Amersham

Biosciences. This technique relies on preelectrophoretic labeling of samples with one of three spectrally resolvable fluorescent CyDyes (Cy2, Cy3, and Cy5), allowing multiplexing of samples into the same gel. There are currently two types of CyDye labeling chemistries available from Amersham Biosciences.

2. Minimal labeling

The most established chemistry is employed in the "minimal labeling" method, which has been available from this supplier since July 2002. Here, the CyDyes are supplied with an *N*-hydroxy succinimidyl ester group that reacts with the epsilon amino group of lysine side chains. Labeling reactions are engineered such that the stoichiometry of protein to fluor results in only 2–5% of the total number of lysine residues being labeled. It is imperative to keep a low dye:protein ratio to avoid multiple dye additions, as these would result in multiple spots being resolved in the second dimension of the DIGE gel. The typically high lysine content of most proteins makes it challenging to force the labeling reaction to saturation without using excessive amounts of reagent. The fluors carry an intrinsic charge of +1, such that the pI of the protein is preserved upon labeling. The three fluors are also mass matched, each labeling event adding approximately 500 Da to the mass of the protein. Labeling with CyDye DIGE Fluors is very sensitive with a detection limit of around 500 pg of a single protein and a linear response in protein concentration over at least five orders of magnitude. In comparison, the limit of detection with silver stain is in the region of 1ng of protein with a dynamic range of no more than two orders of magnitude (Lilley *et al*., 2002; Tonge *et al*., 2001). The labeling system is compatible with the downstream processing commonly used to identify proteins, which involves the generation of tryptic peptides. Trypsin cleaves the peptide bonds on the C-terminal side of lysine and arginine residues, but, as so few lysine residues are modified by dye labeling, peptide generation is largely unhindered. It is also unlikely that a peptide modified with a CyDye DIGE Fluor would be extracted from a gel piece, hence, interference by the fluors in peptide mass fingerprinting and de novo sequencing mass spectrometric techniques is minimal. A drawback of this minimal labeling system is the fact that the majority of the protein within a sample remains unlabeled and, in the case of smaller molecular weight species, the labeled portion of the protein may migrate to a slightly different position on a 2D gel. To ensure that the maximum amount of protein is excised for downstream processing, minimally labeled DIGE gels are often poststained with a total protein stain such as SyproRuby.

3. Saturation labeling

The second more recent chemistry was released by Amersham Biosciences in July 2003 and is designed for use in situations where sample abundance is limited. This differs from the original *N*-hydroxy succinimidyl chemistry in that CyDyes with no intrinsic charge are supplied with a thiol-reactive maleimide group. These "saturation" dyes are utilized in such a way to bring about labeling of every

cysteine residue within a protein. The saturation labeling is much more sensitive as more fluorophor is introduced into each protein species, Shaw *et al*. reporting an order of magnitude increase in sensitivity over the original minimal dyes (Shaw *et al*., 2003). While the added sensitivity that these dyes provide is desirable, their use is technically more challenging. The reaction conditions have to be carefully optimized for each type of sample to ensure complete reduction of cysteine residues and a protein:dye ratio sufficient for stoichiometric labeling. Substoichiometric labeling will lead to multiple spots in the second dimension, whereas the use of too much dye may lead to unwanted addition reactions with lysine residues, resulting in the formation of charge trains in the first dimension. It is also impossible to compare the 2D spot maps between samples labeled with the two chemistries. Proteins containing multiple cysteine residues may appear as larger molecular weight species when labeled with the saturation dyes. For studies where identification by mass spectrometric techniques is required, a preparative gel with increased protein loading will be required to produce a 2D protein map. In situations where the amount of sample is very small, this technique at best can be considered of diagnostic value. The use of these saturation dyes is not well established, and the maleimide dyes are currently only available for Cy3 and Cy5.

4. Experimental design

DIGE is a particularly powerful approach to study the changes in protein abundance in experimental studies involving comparison of multiple samples. In such studies involving the use of minimal dyes, it is desirable to label samples with either Cy3 or Cy5 minimal dyes, whereas Cy2 is used to label a pooled sample comprising equal amounts of each of the samples within the study, and acts as an internal standard (see Figure 1). This ensures that all proteins present in the samples are represented, allowing both inter- and intra-gel matching. Variation in spot volumes due to gel-specific experimental factors, for example protein loss during sample entry into the immobilized pH gradient strip, will be the same for each sample within a single gel. Consequently, the relative amount of a protein in a gel in one sample compared with another will be unaffected. The spot volumes are normalized for dye discrepancy, arising from differences in laser intensities, fluorescence, and filter transmittance, using a method based on the assumption that the majority of protein spots have not changed in expression level (Alban *et al*., 2003). The spot volumes from the labeled samples are compared to the internal standard giving standardized abundances. This allows the variation in spot running success to be taken into consideration. For the analysis, software developed for the DIGE system (DeCyder™ by Amersham Bioscience, Sweden) is typically used. This software has a codetection algorithm that simultaneously detects labeled protein spots from images that arise from the same gel and increases accuracy in the quantification of standardized abundance (Alban *et al*., 2003). The standardized abundances can then be compared across groups to detect changes in protein expression. In the case of the saturation dyes, where a Cy2 label is not available, the internal standard is labeled with one of the dyes and individual samples appear on separate gels labeled

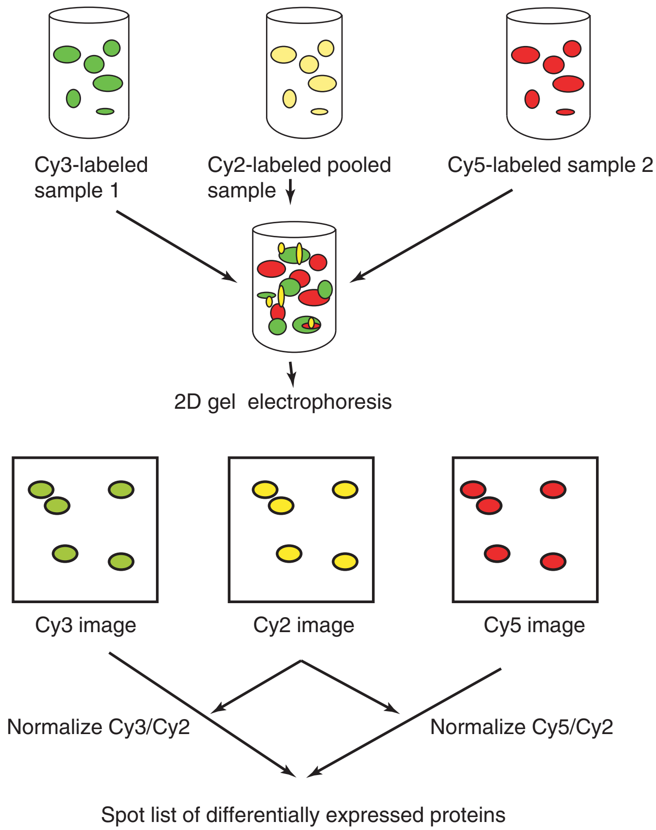


Figure 1 Schematic outline of a 2D DIGE study using an internal pooled standard constructed from equal amounts of all the samples in the study, labeled with Cy2. Samples 1 and 2 are labeled with either Cy3 or Cy5. Each 2D gel performed within the study will have the sample Cy2 standard, combined with the Cy3 and Cy5 labeled samples prior to electrophoresis. The spot intensities from samples 1 and 2 can be normalized using the corresponding Cy2 spot intensities. This approach allows the measurement of more subtle protein expressional differences with increased statistical confidence

with the other dyes. This approach reduces the extent of multiplexing and increases the number of gels within a comparative experiment (Shaw *et al*., 2003).

The design of a protein-profiling analysis experiment using DIGE is crucial to the amount of statistical significance that can be placed on the data. Consideration must be given to methods employed to assess both biological and experimental noise within the system being studied and ample biological and technical replicates must be processed. It is essential to measure the experimental variation in the DIGE process for any new set of samples. This can be achieved by running sets of gels where the same sample is labeled with all dyes to be used, loaded onto one gel, and fully analyzed. This will also give an indication of the inherent error in the system and suggest the threshold of significance or a fold change above which true changes in expression can be measured. In the case of the minimal dyes, a system bias at low spot volumes is observed owing to the different fluorescence characteristics

of acrylamide at the different wavelengths of excitation for each of Cy2, Cy3, and Cy5. This system bias can be greatly reduced by employing a dye swap approach; several replicate gels are run where each sample appears with the opposite labeling, that is, gel 1 would contain pooled sample in Cy2 channel, wild type in Cy3 channel, and mutant in Cy5; gel 2 would contain pooled sample in Cy2 channel, mutant in the Cy3 channel, and wild type in Cy5 channel (Karp *et al*., 2004). The multigel approach allows many data points to be collected for each group to be compared. Spots of interest can be selected by looking for significant change across the groups, for example, with a univariate statistical test, for example, Student's t-test or analysis of variance (ANOVA). These give a probability score (p) for each spot. This score indicates the probability that the groups are the same, consequently a spot with a low score, for example, $p < 0.05$, represents significant difference in relative abundance. The number of replicates required in a study depends on the amount of variation in the system being investigated and on how small the changes in expression are that you wish to measure at a given confidence level. Increasing the number of replicates will increase confidence in smaller changes in expression.

5. Drawbacks of the system

Regardless of the benefits of DIGE, the 2D PAGE process itself has some limitations. For global expression analysis, every protein should be resolved as a discrete detectable spot, The following groups of proteins, however, are often poorly represented: those with extreme isoelectric points (pI) or molecular weight; hydrophobic proteins; lower abundance proteins. It has been calculated that somewhere in the region of 90% of the total protein of a typical cell is made up of only 10% of the 10 000–20 000 possible different species, and hence many low abundance proteins may not be detectable (Zuo *et al*., 2001). Co-migration is also an issue, with proteins of similar pI and denatured molecular weight becoming focused at the same position of the gel. This makes it impossible to accurately determine the relative abundance of an individual protein within a mixed spot. There continues to be improvements to the 2D PAGE technique, however. Enhanced resolution of protein species can be achieved by the use of narrow range immobilized pH gradient (IPG) strips and/or prefractionation of sample; these greatly improve the chance of identification and assignment of function to scarce species (Tonella *et al*., 2001). Membrane proteins remain a problem. The use of more rigorous detergents such as amidosulphobetaine 14 (ASB14) has increased the number of membrane-associated proteins, which can be resolved by 2D PAGE (Santoni *et al*., 2000). In the case of studies involving integral membrane proteins, a 2D PAGE approach should be avoided and differential isotopic labeling strategies involving protein digestion in solution digestion to peptides employed (Li *et al*., 2003).

One of the main criticisms of DIGE is the financial outlay necessary to install the system in a laboratory. An appropriate scanning system and dedicated software are required and the cost of CyDyes necessary for large scale studies is not trivial.

6. Applications of DIGE

To date, the DIGE technology has been used with great success to study a variety of systems, allowing the detection of more subtle changes in protein expression than conventional methods where separate samples are loaded onto each gel (Gade *et al*., 2003); these include breast cancer cells (Gharbi *et al*., 2002), cat brain (Bergh *et al*., 2003), esophageal cancer cells (Zhou *et al*., 2002), yeast (Hu *et al*., 2003), chloroplast (Kubis *et al*., 2003), GPI-anchored proteins (Borner *et al*., 2003), murine mitochondria (Kernec *et al*., 2001), mouse brain (Skynner *et al*., 2002), and rat heart (Sakai *et al*., 2003), although as yet there is a dearth of publications where the internal standard system is used. Knowles *et al*. (2003) have described a significant increase in the accuracy of determination of differential protein expression using the Cy2-labeled internal standard approach. They compared the relative abundances of proteins in cerebral cortex from wild-type mice and neurokinin 1 receptor knockout mice to elucidate molecular pathways involving this protein. They also compared relative abundances and significance values for differentially expressed spots derived from gels incorporating the pooled Cy2-labeled standard with values derived from the same gels, but without normalizing spot volumes to the corresponding pooled standard. They demonstrated that virtually all differentially expressed spots gave lower significance levels and a higher incidence of false–positives when derived without using the pooled standard for normalization. The authors reported being able to measure as little as 10% change in abundance with 95% confidence ($p < 0.05$).

In conclusion, 2D DIGE is the most powerful 2D PAGE-based approach for widespread protein-profiling studies by virtue of the ability to multiplex and link samples across numerous different gels in a study using an internal standard. This approach also gives information about more subtle changes in protein expression than conventional 2D PAGE. The advent of the saturation dyes has increased the sensitivity of this system. The use of 2D DIGE, however, will not result in a global analysis of a proteome as membrane proteins and proteins with extremes of pI and molecular weight will be poorly represented. In this respect, other quantitative techniques such as differential labeling with stable isotope can be considered to be complementary.

Further reading

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **17**, 994–999.

References

Alban A, David SO, Björkesten L, Andersson C, Sloge E, Lewis S and Currie I (2003) A novel experimental design for comparative two-dimensional gel analysis: Two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3** (1), 36–44.

Bergh GVD, Clerens S, Vandesande F and Arckens L (2003) Reversed-phase high-performance liquid chromatography prefractionation prior to two-dimensional difference gel electrophoresis and mass spectrometry identifies new differentially expressed proteins between striate cortex of kitten and adult cat. *Electrophoresis*, **24** (9), 1471–1481.

Borner GHH, Lilley KS, Stevens TJ and Dupree P (2003) Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis. *Plant Physiology*, **132** (2), 568–577.

Gade D, Thiermann J, Markowsky D and Rabus R (2003) Evaluation of two-dimensional difference gel electrophoresis for protein profiling. Soluble proteins of the marine bacterium Pirellula sp. strain 1. *Journal of Molecular Microbiology and Biotechnology*, **5** (4), 240–251.

Gharbi S, Gaffney P, Yang A, Zvelebil MJ, Cramer R, Waterfield MD and Timms JF (2002) Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Molecular and Cellular Proteomics*, **1** (2), 91–98.

Hu Y, Wang G, Chen GYJ, Fu X and Yao SQ (2003) Proteome analysis of *Saccharomyces cerevisiae* under metal stress by two-dimensional differential gel electrophoresis. *Electrophoresis*, **24** (9), 1458–1470.

Karp, N, Kreil, D and Lilley KS (2004) Determining a significant change in protein expression with DeCyderTM during a pair-wise comparison using two-dimensional difference gel electrophoresis. *Proteomics*, **4**(5), 1421–1432.

Kernec F, Unlu M, Labeikovsky W, Minden JS and Koretsky AP (2001) Changes in the mitochondrial proteome from mouse hearts deficient in creatine kinase. *Physiological Genomics*, **6** (2), 117–128.

Knowles MR, Cervino S, Skynner HA, Hunt SP, de Felipe C, Salim K, Meneses-Lorente G, McAllister G and Guest PC (2003) Multiplex proteomic analysis by two-dimensional differential in-gel electrophoresis. *Proteomics*, **3** (7), 1162–1171.

Kubis S, Baldwin A, Patel R, Razzaq A, Dupree P, Lilley K, Kurth J, Leister D and Jarvis P (2003) The Arabidopsis ppi1 mutant is specifically defective in the expression, chloroplast import, and accumulation of photosynthetic proteins. *Plant Cell*, **15** (8), 1859–1871.

Li J, Steen H and Gygi SP (2003) Protein Profiling with Cleavable Isotope-coded Affinity Tag (cICAT) Reagents: The Yeast Salinity Stress Response. *Molecular and Cellular Proteomics*, **2** (11), 1198–1204.

Lilley KS, Razzaq A and Dupree P (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Current Opinion in Chemical Biology*, **6** (1), 46–50.

Malone J, Radabaugh M, Leimgruber RM and Gerstenecker GS (2001) Practical aspects of fluorescent staining for proteomics applications. *Electrophoresis*, **22**, 919–932.

Sakai J, Ishikawa H, Kojima S, Satoh H, Yamamoto S and Kanaoka M (2003) Proteomic analysis of rat heart in ischemia and ischemia- reperfusion using fluorescence two-dimensional difference gel electrophoresis. *Proteomics*, **3** (7), 1318–1324.

Santoni V, Molloy MP and Rabilloud T (2000) Membrane proteins and proteomics: Un amour impossible? *Electrophoresis*, **21**, 1054–1070.

Shaw J, Rowlinson R, Nickson J, Stone T, Sweet A, Williams K and Tonge R (2003) Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. *Proteomics*, **3** (7), 1181–1195.

Skynner HA, Rosahl TW, Knowles MR, Salim K, Reid L, Cothliff R, McAllister G and Guest PC (2002) Alterations of stress related proteins in genetically altered mice revealed by two-dimensional differential in-gel electrophoresis analysis. *Proteomics*, **2** (8), 1018–1025.

Tonella L, Hoogland C, Binz P-A, Appel RD, Hochstrasser DF and Sanchez J-C (2001) New perspectives in the *Escherichia coli* proteome investigation. *Proteomics*, **1**, 409–423.

Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, Pognan F, Hawkins E, Currie I and Davison M (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics*, **1** (3), 377–396.

Ünlü M, Morgan ME and Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18** (11), 2071–2077.

Zhou H, Ranish JA, Watts JD and Aebersold R (2002) Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology*, **20** (5), 512–515.
Zuo, Z, Echan, L, Hembach P, Tang HY, Speicher KD, Santoli D and Speicher DW (2001) Towards global analysis of mammalian proteomes using sample prefractionation prior to narrow pH range two-dimensional gels and using one-dimensional gels for insoluble large proteins. *Electrophoresis*, **22** 1603–1615.

Short Specialist Review

Image analysis

Andrew W. Dowsey and Guang-Zhong Yang
Imperial College, London, UK

Michael J. Dunn
Proteome Research Centre, University College Dublin, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland

1. Introduction

Quantitative analysis of 2D gel electrophoresis (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) is faced with a number of challenges. Although at first glance the resolution of 2-DE seems impressive, it is still not sufficient compared to the enormous diversity of cellular proteins. Comigrating proteins in the same spot are not uncommon (Figure 1a). In general, 2-DE has a maximum dynamic range of 10^4. At this value, the scarcest proteins require an expert eye to discriminate valid spots from noise (Figure 1c). In addition, spots tend to have symmetric diffusion in the pI dimension but often severe tails in the M_r dimension (Figure 1b). Another major issue is geometric deformation due to casting, polymerization, and the running procedure, making it difficult to quantify differential expression (Figure 1d). Other issues include contrast variations (Figure 1e) due to stain exposure, sample loading errors and protein losses during processing.

For a biochemist to analyze a single pair of gels for differential protein expression, bearing in mind the thousands of candidate proteins, it requires several hours of expert analysis. There is a great economic argument and substantial efficiency gains to be had in eliminating this bottleneck with the use of large-scale automated processing. In general, the image processing pipeline for 2-DE includes:

- Preprocessing of the gel images
- Spot segmentation, modeling, and expression quantification
- Image registration to correct for geometric and intensity deformation
- Identification of differential expression
- Interpretation, either manually through visualization or with image data mining
- Creation of federated 2-DE databases.

2. Preprocessing

The purpose of preprocessing is the removal of image artifacts to improve the general quality of the images. The techniques used are not significantly different

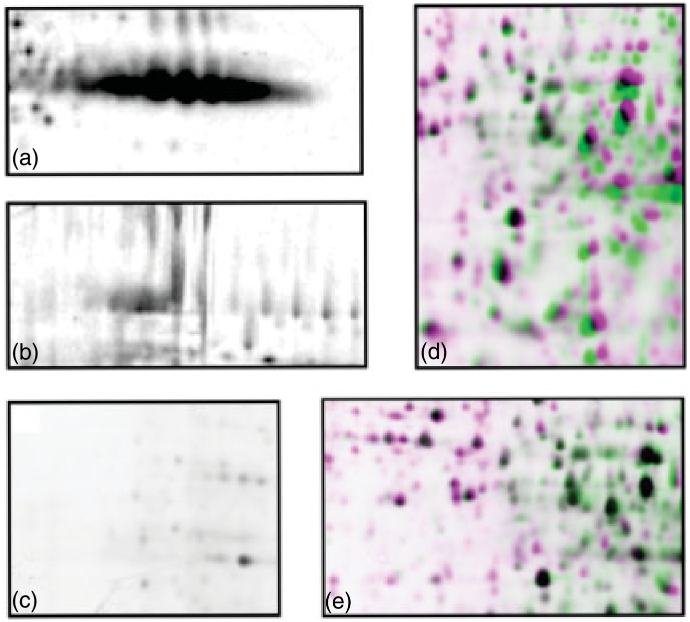


Figure 1 Common problems for computational image analysis as illustrated on silver stained gels: (a) Comigrating spots forming a complex region. (b) Streaking and smearing. (c) Weak spots. Two superimposed gels, one green and one magenta, show (d) nonlinear local distortions between the gels, and after spatial registration, (e) regional intensity inhomogeneities

from other disciplines in digital imaging. For example, in order to correct for uneven background intensity, mathematical morphology is used. The method treats the gel as a topographic relief, where the protein spots are depressions, and variability is estimated by opening the image with a spherical structuring element "the rolling ball", which is larger than the largest spot but with more curvature than the background. To correct for current leakage, which causes a global "frown" in gels whose sides are not fully isolated, model-based processing techniques are used (Gustafsson *et al*., 2002). The parameters are estimated by fitting a straight line to the cathode and a low-degree polynomial to the gel front. The space of gel front curves is searched for the optimal least squares fit.

3. Spot modeling

Traditional methods for differential expression analysis involve explicit modeling of the protein spots. Spot detection concerns the individual resolution of each spot into spot centers, intensities, and geometric properties. Currently, the most popular segmentation technique is the watershed transform (WST). A watershed is the boundary of a region (catchment basin) in a landscape where all water drains to a common point (Figure 2a). The most popular algorithmic approach by Vincent and Soille follows the immersion principle. Imagine piercing all local minima and slowly submerging the landscape in water. Where catchment basins merge

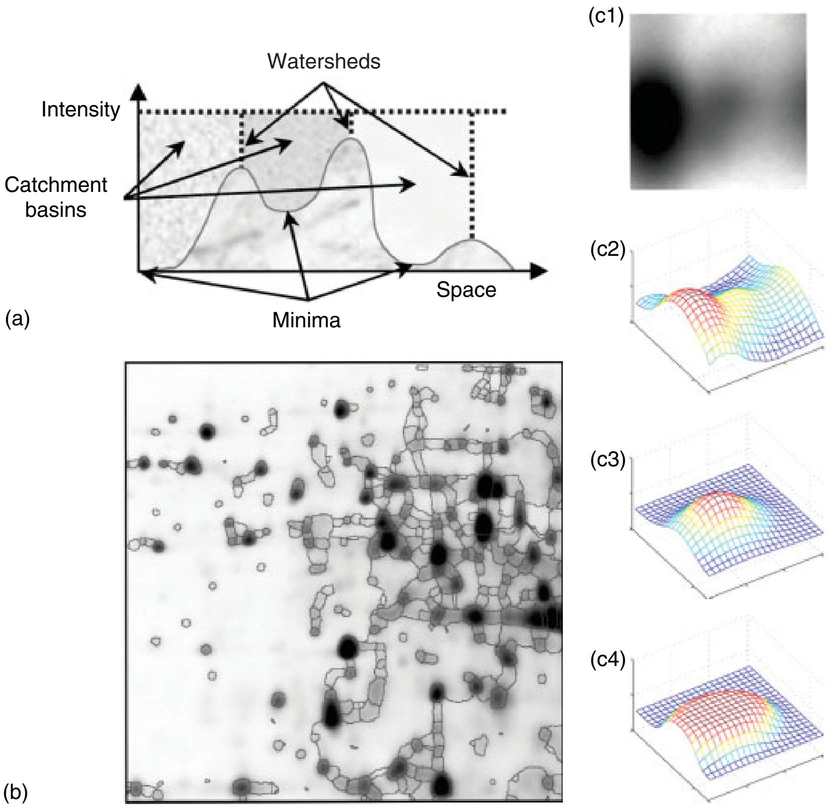


Figure 2 (a) Cross section of a region processed by the watershed transform (WST), treated by the immersion principle of Vincent and Soille. (b) Oversegmentation by the WST. Then the shoulder spot seen in image (c1), with 3D representation (c2), is fitted by both (c3) a 2D Gaussian, and (c4) the diffusion model

dams are built and watershed points are labeled. Noise cause oversegmentation (Figure 2b), which can be minimized through extra restrictions before (marker controlled watersheds) or after (region merging) segmentation.

Spots can overlap in a way that only one segment is found (Figure 2c1, 2c2), so greater prior information of spot shape and its parametric composition is required. Traditionally, this involves optimizing the parameters of a two-dimensional Gaussian (Figure 2c3) to minimize the squared residuals. When saturation effects occur, the spot can no longer be modeled by a Gaussian, and a simplified diffusion process (Bettens *et al*., 1997) is used instead (Figure 2c4). Recently, statistical point distribution models generated from training data have been used (Rogers *et al*., 2003). The principle modes of variation of points placed on the boundary of each spot are captured using PCA. In heavily saturated regions, gradient information cannot be used, and thus a linear programming solution with elliptical elements has been proposed (Efrat *et al*., 2002).

After spot detection, characteristic information about each spot is extracted both to quantify the protein expression and to aid in the spot matching process.

Typical measurements include integrated optical density (VOL: the sum of the pixel intensities), the scaled volume (SV: the volume of a spot normalized with respect to background), and pI/M_r interpolation.

4. Image registration

Image registration represents a class of techniques to align and normalize two images "optimally". Geometric and intensity distortion attributed to experimental uncertainty (i.e., protein diffusion in 2-DE) is undesirable and therefore should be eliminated. On the other hand, intrinsic difference (i.e., changes in protein expression) must be kept. The source image I_s is brought into alignment with the reference image I_r, constrained by the space of allowable transformations (mappings) on I_s, M, and guided by the similarity measure function *sim* – which is at a maximum when the alignment is optimal.

Point pattern matching (PPM) is the standard feature-based approach used today, which finds the closest spot correspondence between a point-pattern (source spot list) and a target point set (reference spot list). Owing to differential expression and inaccuracies in spot detection, these methods need be robust to outliers and allow nonlinear distortions. A wide variety of PPM methods have been developed that include iterative closest point (ICP), the alignment method, bipartite graph matching (geometric hashing), expectation maximization, dynamic programming, and robust pattern matching (RPM). Only RPM methods interleave image warping directly. An issue with feature matching is the number of candidate transforms that need to be evaluated, due to a combinational explosion when each arc (drawn between each two points) in the point pattern is mapped onto every arc in the target point set. To reduce the search space, CAROL (Efrat *et al*., 2002) removes arcs by Delaunay triangulation of pattern subsets. The pattern is transformed by each candidate and the one with the shortest Hausdorff distance (the arc length of the worst spot match) is selected. The positions of subset matches are then used to locally transform the sample image before a global matching takes place.

It was not until recently that the processing power became cheap enough for pixel-based registration methods. Z3 (Smilansky, 2001) is a hybrid registration strategy where the two gel images are covered by small rectangles each containing a cluster of spots. As each sample rectangle is matched to the best reference candidate, a function constraining candidate selection is updated. This function defines a global transformation, which increases in complexity as more rectangles are matched. Once all rectangles have been processed, the sample image is warped by the final transform. The algorithm repeats on the transformed image until no further warping occurs.

Later, a direct registration algorithm with no recourse to detecting spots was proposed (Veeser *et al*., 2001). The control points of a bilinear or third-order B-Spline piecewise mapping are sought with the (BFGS) Broyden-Fletcher-Goldfarb-Shannon optimizer and the closed form derivative of a cross-correlation similarity measure. Convergence occurs as the gel images have smooth gradients, and therefore, the cross correlation with respect to the control points is smooth and continuous. To avoid local minima, a multiresolution approach is used,

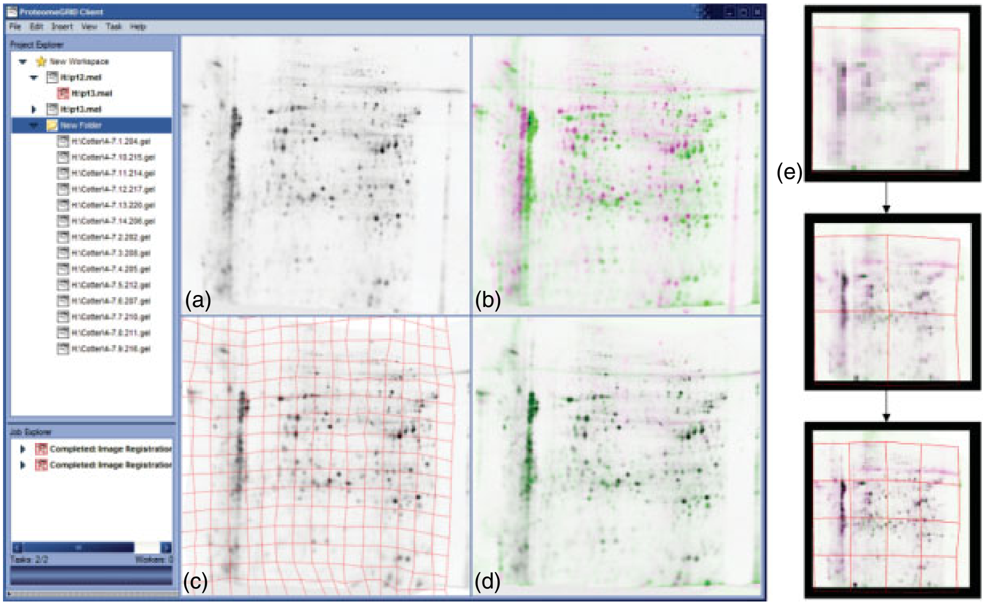


Figure 3 MIR image registration in the ProteomeGRID Client. (a) Control gel. (b) The control gel (magenta) overlaid on a treated sample gel (green). The geometric distortions are clearly evident. (c) The deformation mesh calculated by MIR. (d) The control gel overlaid on the transformed sample gel. (e) MIR searches for the global maximum and avoids local maxima by iteratively refining a course mesh on progressively higher resolution images

where the gel images increase in detail from 64×64 pixels to 1024×1024 as the mesh increases from 1×1 pieces to 16×16 (Figure 3).

Thus far, no software package attempts to correct for contrast variations. If we assume the inhomogeneities can be modeled by a smoothly varying multiplicative "bias field", which attenuates or amplifies intensity at each point, it can be corrected for without affecting true differential protein expression, which causes far more abrupt changes in intensity. Dowsey *et al*. (2003) demonstrated the shape-from-orientation approach as suitable for 2D gel analysis. An implicit thin plate spline surface is fitted to sparse orientation constraints (the ratio in integrated optical density IOD between spot matches) in a regularization framework.

5. Differential expression analysis

Without bias field correction, mutation detection requires robust statistical methods. Traditional methods are based on Student's T-test, the Mann–Whitney test, or Analysis of Variance (ANOVA). PCA can be used as an implicit bias field correction. The first component is interpreted as the bias field; the next few contain the most common volume changes over the set; and the rest are dominated by single spots and therefore indicate noise. However, if a particular model of activity is known, constraint analysis can be used to search the gel for spots satisfying the model constraints. Posttranslational modification, putative point mutations, and putative precursor product-pairs are examples of models that can be detected.

Once the differential expression is found, classification analyses are used to find intrinsic trends and relationships, that is, the intermixed reactions cells have to altered conditions. By utilizing prior results, classification also becomes a means to identify the unknown origin of tissue samples or the progression of disease or the efficacy of the treatment. Among the different techniques proposed, ChiClust (Harris *et al*., 2002) uses Ward's minimum variance method, where the sum of the squared distances from each member to the group centroid is used as an indicator of dispersion. Groups are joined only if the increase in this sum is less for that pair of groups than for any other. For time course experiments, Vohradsky (1997) first compared cluster analysis and cluster analyses to classify the stages of *Streptomyces coelicolor* growth, and then compared spot profile classification, PCA, cluster analysis, and artificial neural networks (ANN) in the detection of transition phase.

6. The statistical expression analysis (SEA) approach for large-scale analyses

With traditional techniques, symbolic representation of spots is determined at the very early stages of the processing pipeline. Although the early use of abstraction greatly simplifies data handling and management, once such a description is reached intrinsic errors, dependant on the spot modeling and matching algorithms used, are persistent throughout the subsequent processing steps. Current technologies cannot therefore provide a coherent and integratable workflow without extensive user interaction. In practice, if the analysis was modeled statistically and performed simultaneously between sets of technical replicates, small insignificant expression changes over one pair could become significant when reinforced by the same consistent changes in the others. This permits the creation of a database of probabilistic baselines, or "norms", which characterize confidence levels for protein expression under different experimental settings. Such evidence will build up a statistical formation model of 2-DE, which would then be mined to discover intrinsic trends, increase the acuity of new results and provide valuable feedback to 2-DE scientists on the sensitivity of their experiments.

The use of SEA offers the potential for developing an automated high-throughput proteomics pipeline, but the computational demands that image mining techniques bring requires the utilization of Grid technology and the development of specific imaging middleware, to ensure its success. The Grid, through the Globus Open Services Grid Architecture, offers seamless computational cooperation between thousands of commodity PCs and specialist equipment, such as automated spot cutters and mass spectrometers. The ProteomeGRID initiative (**http://www.proteomegrid.org/**) is realizing this ambition, and its *proTurbo* clustering technology (Dowsey *et al*., 2004) can already be employed on shared office and laboratory workstations to harvest spare capacity.

Further reading

All techniques illustrated in this section, and a comprehensive review of the role of bioinformatics in 2-DE on which this text is based, is available in Dowsey *et al*. (2003).

References

Bettens E, Scheunders P, Van Dyck D, Moens L and Van Osta P (1997) Computer analysis of two-dimensional electrophoresis gels: a new segmentation and modeling algorithm. *Electrophoresis*, **18**(5), 792–798.

Dowsey AW, Dunn MJ and Yang GZ (2003) The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, **3**(8), 1567–1596.

Dowsey AW, Dunn MJ and Yang GZ (2004) ProteomeGRID: towards a high-throughput proteomics pipeline through opportunistic cluster image computing for two-dimensional gel electrophoresis. *Proteomics*, **4**(12), 3800–3812.

Efrat A, Hoffmann F, Kriegel K, Schultz C and Wenk C (2002) Geometric algorithms for the analysis of 2D-electrophoresis gels. *Journal of Computational Biology*, **9**(2), 299–315.

Gustafsson JS, Blomberg A and Rudemo M (2002) Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern. *Electrophoresis*, **23**(11), 1731–1744.

Harris RA, Yang A, Stein RC, Lucy K, Brusten L, Herath A, Parekh R, Waterfield MD, O'Hare MJ, Neville MA, *et al*. (2002) Cluster analysis of an extensive human breast cancer cell line protein expression map database. *Proteomics*, **2**(2), 212–223.

Rogers M, Graham J and Tonge RP (2003) Statistical models of shape for the analysis of protein spots in 2-D electrophoresis gel images. *Proteomics*, **3**(6), 887–896.

Smilansky Z (2001) Automatic registration for images of two-dimensional protein gels. *Electrophoresis*, **22**(9), 1616–1626.

Veeser S, Dunn MJ and Yang GZ (2001) Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, **1**(7), 856–870.

Vohradsky J (1997) Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. *Electrophoresis*, **18**(15), 2749–2754.

Short Specialist Review

Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices

Wayne F. Patton

Perkin-Elmer Life Sciences, Boston, MA, USA

1. Introduction

The marriage of advanced imaging instrumentation and small-molecule detection reagents has been crucial to inventing new capabilities in electrophoretic detection, especially in the realm of nonradioactive posttranslational modification analysis. Fluorescent total-protein stains, such as SYPRO Ruby dye (Molecular Probes), Deep Purple dye (Amersham Life Sciences), and the cyanine dyes of the DIGE analysis system (Amersham Life Sciences, *see* Article 30, **2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis**, Volume 5), provide detection limits matching silver staining but offer significantly better dynamic range of quantitation and better compatibility with protein characterization techniques, such as mass spectrometry (Patton, 2002; Mackintosh *et al*., 2003). Accurate measurement of total protein amounts is fundamental to evaluating posttranslational modification changes, since it provides a denominator term, serving as a baseline for monitoring potential alterations; that is, the total protein measurement ensures that different samples are present in comparable amounts on the gel and aids in distinguishing between changes in the levels of posttranslational modification per unit protein compared with changes in protein expression levels without concomitant changes in the modification level. Typically, total protein staining is performed after detecting specific posttranslational modification.

2. Acquiring images

The introduction of fluorescent dye technology has played a pivotal role in spurring development of advanced imaging instrumentation, such as CCD camera-based multiwavelength imaging platforms, as well as gel scanners with multiple laser excitation sources (reviewed in Patton, 2000a,b). For example, the ProXPRESS 2D Proteomic Imaging System (Perkin-Elmer) is a sensitive imaging instrument

that enables the use of a wide range of fluorescent and colored dyes owing to its CCD camera and multiwavelength emission and excitation capabilities. The instrument's high-pressure xenon arc lamp provides broadband wavelength coverage and requires modest power, allowing visualization of the wide range of dyes commonly encountered in proteomics investigations, including Coomassie Blue, Amido Black, silver, colloidal gold, and the variety of fluorescent dyes now available. While the spatial resolution of conventional fixed CCD camera-imaging systems is typically inferior to laser-based gel scanners and photographic film, this problem is circumvented with the ProXPRESS instrument by mechanically scanning the CCD camera over the gel or blot and collecting multiple images that are subsequently automatically reconstructed into a complete image. Thus, the system readily delivers the same 50-μm spatial resolution obtained with high-end laser scanners, such as the FLA-5000 fluorescent imager (Fuji Corporation). By acquiring images in succession, as many as four different fluorescent labels may be viewed from any single gel.

3. Detecting posttranslational modifications

The rapid and comprehensive elucidation of protein posttranslational modifications is certainly one of the more important challenges facing the field of proteomics. Three prominent protein posttranslational modifications, glycosylation, phosphorylation, and S-nitrosylation are now readily detectable after polyacrylamide gel electrophoresis using commercially available kits and instrumentation. The basic principles of the detection schemes are summarized below.

3.1. *Glycosylation*

Alterations in glycosylation profiles arising in human cancer are known to contribute in part to the malignant phenotype observed downstream of oncogenic events (*see* Article 62, **Glycosylation**, Volume 6). A sensitive green-fluorescent glycoprotein-specific stain, Pro-Q Emerald 300 dye (Molecular Probes), detects glycoproteins directly in polyacrylamide gels or on polyvinylidene difluoride (PVDF) membranes. Pro-Q Emerald 300 dye is conjugated to glycoproteins by a periodic acid Schiff's base (PAS) mechanism using room-temperature reaction conditions. As little as 300 pg of α_1-acid glycoprotein (40% carbohydrate) may be detected in gels after staining with the dye, and a 500- to 1000-fold linear dynamic range of detection is obtained. A UV transilluminator-based imaging system is required to detect the stain, but a related stain, Pro-Q Emerald 488 dye, allows the detection of glycoproteins using visible light excitation sources. GlycoProfile III fluorescent glycoprotein detection kit (Sigma-Aldrich) also operates by a PAS-staining mechanism and is suitable for fluorescence-based detection of glycoproteins. The dye can detect roughly 150 ng of a glycoprotein in gels. Additionally, chemiluminescence-based approaches employing PAS labeling with digoxigenin hydrazide followed by immunodetection with alkaline phosphatase-conjugated antidigoxigenin antibody conjugated (DIG Glycan Detection Kit, Roche Molecular Biochemicals), or PAS

labeling with biotin hydrazide followed by detection with horseradish peroxidase-conjugated streptavidin (ECL Glycoprotein Detection kit Amersham Life Sciences) detect glycoproteins on electroblots with high sensitivity.

3.2. Phosphorylation

Reversible protein phosphorylation plays a critical regulatory role in the circuitry of biological signal transduction, and disregulation of this process is a hallmark of carcinogenesis (*see* Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6). Pro-Q Diamond phosphoprotein stain (Molecular Probes) detects phosphoserine-, phosphothreonine-, and phosphotyrosine-containing proteins on SDS-polyacrylamide gels, isoelectric focusing gels, 2-D gels, electroblots, and protein microarrays by a mechanism that combines a chelating fluorophore and a transition metal ion (Schulenberg *et al*., 2003). The staining is rapid, simple to perform, readily reversible, and fully compatible with modern microchemical analysis procedures, such as MALDI-TOF mass spectrometry. Pro-Q Diamond dye can detect as little as 8 ng of pepsin, a monophosphorylated protein. The linear response of the fluorescent dye allows rigorous quantitation of phosphorylation changes over a 500–1000-fold concentration range. Alternatively, phosphoproteins may be selectively detected in gels through alkaline hydrolysis of serine or threonine phosphate esters, precipitation of the released inorganic phosphate with calcium, formation of an insoluble phosphomolybdate complex, and then visualization of the complex with methyl green dye, as recently commercialized in the GelCode phosphoprotein detection kit (Pierce Chemical Company). Detection by the staining method is not particularly sensitive, however, with microgram to milligram amounts of phosphoprotein required to obtain a discernible signal. Phosphotyrosine residues cannot be detected with the stain.

3.3. S-nitrosylation

In an analogous manner as with protein phosphorylation, protein S-nitrosylation is thought to represent a key mechanism for the reversible posttranslational regulation of protein activity and, consequently, cellular function (*see* Article 68, **S-nitrosylation and thiolation**, Volume 6). Until relatively recently, the detection of protein S-nitrosylation in cells and tissues has been limited by a lack of appropriate detection techniques that permit identification of this labile posttranslational modification on target proteins. The Biotin Switch method was devised to facilitate the detection of protein S-nitrosylation after SDS-polyacrylamide gel electrophoresis and electroblotting (reviewed in Patton, 2002). The basic approach has been commercialized as the NitroGlo nitrosylation detection kit (Perkin-Elmer). The NitroGlo kit provides an optimized procedure for detecting protein S-nitrosylation that has been validated using known S-nitrosylated proteins, such as H-ras and creatine phosphokinase. Following chemical blocking of free cysteine residues, S-nitrosylated cysteine residues are reduced to free thiols. These free thiols are then modified with a biotinylating reagent. Thus, only the cysteine residues that had

nitroso modifications are tagged with a biotin label. Gel electrophoresis may then be performed to separate the various proteins in the sample. Alternatively, the labeled proteins may be selectively enriched using streptavidin affinity chromatography prior to gel electrophoresis. Western blotting, using chemiluminescence reagents, such as the Western Lightning kit (Perkin-Elmer), and subsequent imaging of the blots, generates a fingerprint pattern of the S-nitrosylated proteins in the specimen.

In conclusion, small-molecule-based detection molecules, in combination with advanced analytical imaging devices, now permit the routine monitoring of common protein posttranslational modifications in polyacrylamide gels and on electroblot membranes, without resorting to the use of hazardous radiolabeling approaches. Though these probes do not provide the same level of detection sensitivity as radiolabeling, they are safer and more convenient, allowing measurement of native posttranslational modification levels, as opposed to changes in levels associated with a particular labeling incubation time period employed in an experiment.

References

Mackintosh J, Choi H, Bae S, Veal D, Bell P, Ferrari B, Van Dyk D, Verrills N, Paik Y and Karuso P (2003) A fluorescent natural product for ultra sensitive detection of proteins in one-dimensional and two-dimensional gel electrophoresis. *Proteomics*, **3**, 2273–2288.

Patton W (2000a) A thousand points of light; The application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis*, **21**, 1123–1144.

Patton W (2000b) Making blind robots see; The synergy between fluorescent dyes and imaging devices in automated proteomics. *BioTechniques*, **28**, 944–957.

Patton W (2002) Detection technologies in proteome analysis. *Journal of Chromatography. B, Biomedical Applications*, **771**, 3–31.

Schulenberg B, Aggeler B, Beechem J, Capaldi R and Patton W (2003) Analysis of steady-state protein phosphorylation in mitochondria using a novel fluorescent phosphosensor dye. *The Journal of Biological Chemistry*, **278**, 27251–27255.

Short Specialist Review

Real-time measurements of protein dynamics

Caoimhín G. Concannon, Heiko Duessmann and Jochen H. M. Prehn
Royal College of Surgeons in Ireland, Dublin, Ireland

1. Introduction

In recent years, a variety of imaging-based techniques has been developed, enabling more advanced studies of protein function and has opened new windows into our understanding of key cellular functions ranging from gene expression, intracellular signaling pathways, and second messenger cascades. Central to these advancements has been the development of new imaging technologies, sophisticated computer software for image acquisition and analysis, but most importantly the discovery and cloning of the green fluorescent protein (GFP) from the jellyfish *Aequorea* (*see* Article 58, **Immunofluorescent labeling and fluorescent dyes**, Volume 5). The versatility of GFP-like proteins as intracellular molecular sensors stems from their intrinsic ability to form chromophores within live organisms, tissues, and cells without the requirement for any additional cofactors or substrates with the exception of molecular oxygen. Since these proteins are genetically encoded, this property can therefore be manipulated to create genetic fusions to any variety of protein sequences that can then be targeted to appropriate subcellular sites, enabling biochemical processes to be examined in real time using live cell imaging (Lippincott-Schwartz and Patterson, 2003). Molecular biology techniques have further aided the visualization of protein dynamics with the generation of mutants of GFP with blue, cyan, and yellow emission spectra (right up to the 529 nm emission maximum of yellow fluorescent protein, YFP) (Zhang *et al*., 2002). In addition, the discovery of "GFP-like proteins" from Anthozoa (coral) animals has further expanded the variety of colors available for multicolor labeling of different proteins. The most important of these are the red fluorescent proteins (RFPs) that possess emission maxima at wavelengths greater than 570 nm, an area of the spectrum where eukaryotic cells display reduced autofluorescence. DsRed, in particular, has several positive attributes such as impressive brightness and increased pH stability over GFP. However, it does have the problem that it is an obligate tetramer, which can limit its applications in terms of in vivo protein labeling. In this article, we will discuss several techniques that have been derived from GFP technology for the measurement of protein location and movement within living cells.

2. Visualizing protein dynamics during apoptotic cell death

We have utilized GFP technology to visualize the translocation of proteins involved in a form of cell death termed apoptosis. Central to the initiation of apoptosis is the release of several proapoptotic proteins, such as cytochrome *c*, from mitochondria. When released into the cytosol, cytochrome *c* forms part of a complex termed the *apoptosome*, which is involved in the activation of caspases, a group of proteolytic enzymes responsible for the execution of apoptosis. Using GFP-tagged cytochrome *c* and time lapse confocal microscopy, we can quantitatively visualize the translocation of cytochrome *c* from mitochondria by monitoring the redistribution of the fluorescent signal of GFP following the addition of an apoptotic stimulus (Luetjens *et al*., 2001). Cytochrome-*c*-GFP release during apoptosis is a surprisingly

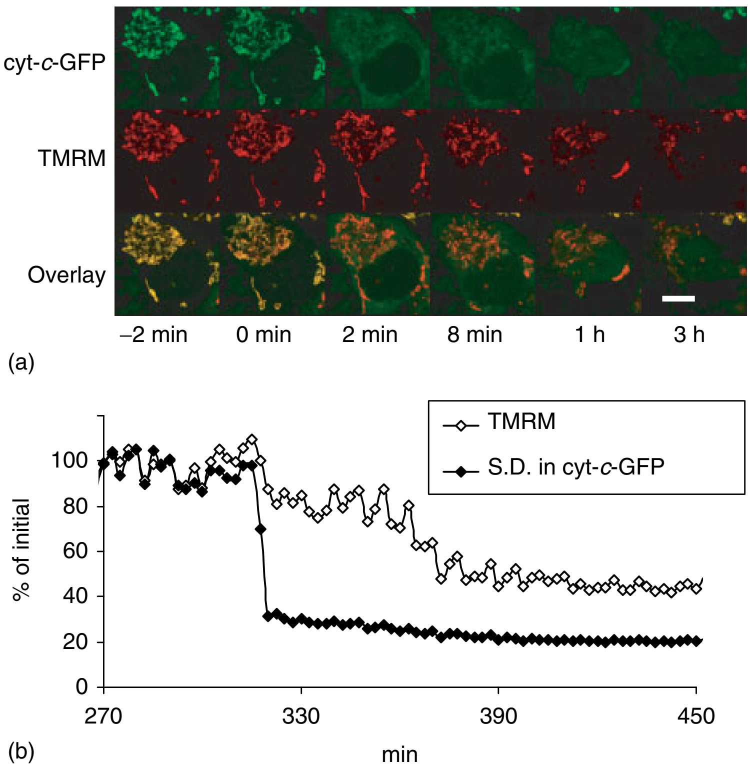


Figure 1 The release of cytochrome-*c*-GFP triggers a decrease in mitochondrial TMRM uptake. (a) HeLa D98 cells stably expressing cytochrome-*c*-GFP were equilibrated with 30-nM TMRM and treated with 3-µM STS. Cytochrome-*c*-GFP release was followed by a reduction in mitochondrial TMRM uptake. Bar, 10 µm. (b) Individual trace of the cell in (a). The release of cytochrome *c*-GFP was detected as a reduction in the standard deviation of the GFP pixel intensities of single cells. Changes in TMRM uptake were calculated from single cells by determining their average pixel intensity in the TMRM-sensitive channel. The onset of the decrease in mitochondrial TMRM uptake occurred simultaneously to mitochondrial cytochrome-*c*-GFP release, suggesting an irreversible mitochondrial membrane potential depolarization during the period monitored

rapid and coordinated process that can be quantitatively described by measuring the standard deviation of the average fluorescent intensity per pixel. In nonapoptotic cells, cytochrome *c*-GFP is compartmentalized within mitochondria, which contributes to a high standard deviation of pixel intensities (Figure 1). Upon its redistribution to the cytosol, its localization becomes more homogeneous throughout the cell, leading to a decrease in the standard deviation of pixel intensities. Measurements of protein dynamics can be combined with measurements of cellular functions using a variety of fluorescent probes such as potential- or redox-sensitive dyes (Dussmann *et al*., 2003). For example, the release of cytochrome-*c*-GFP is accompanied by a decrease in mitochondrial membrane potential that can be detected using the cationic probe TMRM (Figure 1). Studies of protein dynamics can also be combined with studies of protease activation or protein interactions detected by fluorescence resonance energy transfer (FRET) (*see* Article 51, **FRET-based reporters for intracellular enzyme activity**, Volume 5) technology. For example, we have delineated the temporal relationships between release of proapoptotic factors from mitochondria and caspase activation on a single cell level during apoptosis by combining GFP and FRET imaging (Rehm *et al*., 2003).

3. Fluorescence recovery after photobleaching (FRAP) and photoactivation

The visualization of the lateral mobility and dynamics of proteins within living cells with fluorescence recovery after photobleaching (FRAP) is facilitated with the introduction of improved scanning and laser light intensity regulation in point scanning confocal microscopes (*see* Article 53, **Photobleaching (FRAP/FLIP) and dynamic imaging**, Volume 5). Using this technique, a defined region of a cell is extinguished of fluorescence by a brief but intense excitation pulse, which is sufficient to irreversibly inactivate the fluorescent proteins within that defined area. The influx of new fluorescent proteins into the photobleached region is monitored over time with low-intensity laser light until a recovery plateau is achieved. This recovery of fluorescence results from protein diffusion or transport into the region of interest, and its quantification can provide valuable information about the kinetic properties of a particular protein within that compartment of a cell (Lippincott-Schwartz *et al*., 2003). If the protein is freely able to move within that compartment and the volume of the compartment can be assumed to be infinite compared to the bleached area, the fluorescence will recover to the levels of the prebleach values. Usually, a FRAP curve is fitted to the fluorescence recovery kinetics postbleaching. From this curve, it is possible to calculate the diffusion constant/transport rate, as well as the fraction of the protein that is mobile in that compartment. This information in combination with computer modeling can help resolve cellular pathways and has been successfully used to analyze protein transport pathways as well as measuring the dynamics of nuclear proteins.

Another method to visualize the movements of proteins within their native environment is to utilize the phenomenon of photoactivation. This method involves the use of GFP variants whose fluorescence is switched on or significantly increased

following activation with light of particular wavelengths. Currently, there are three readily available photoactivable fluorescent proteins – Kaede, KFP1, and photoactivatable-GFP (PA-GFP). Native PA-GFP has a major absorbance peak at 399 nm and an emission maxima at ~517 nm. However, following photoactivation at ~405 nm, a major new absorbance peak is observed at 504 nm, resulting in ~100-fold increase in its fluorescence at 517 nm. This technique has the distinct advantage of allowing the fluorescence of a protein to be switched on. Therefore, newly synthesized proteins will not fluoresce unless photoactivated and are unobserved, helping monitor the dynamics of proteins at particular time points. In this manner, using time lapse microscopy, it is possible to track both the movement and life span of proteins within cells. A disadvantage of the use of most photoactivatable fluorescent proteins is the requirement of intense light for photoactivation with relatively high photon energy levels. This requirement of short wavelength light from the visible spectrum might have detrimental effects on cells and is undesirable for live cell imaging. However, the development of new fluorescent proteins photoactivatable at longer wavelengths has great promise for the future use of photoactivation for measuring protein dynamics.

References

Dussmann H, Kogel D, Rehm M and Prehn JH (2003) Mitochondrial membrane permeabilization and superoxide production during apoptosis. A single-cell analysis. *The Journal of Biological Chemistry*, **278**, 12645–12649.

Lippincott-Schwartz J, Altan-Bonnet N and Patterson GH (2003) Photobleaching and photoactivation: Following protein dynamics in living cells. *Nature Cell Biology*, **5**(Suppl 1), S7–S14.

Lippincott-Schwartz J and Patterson GH (2003) Development and use of fluorescent protein markers in living cells. *Science*, **300**, 87–91.

Luetjens CM, Kogel D, Reimertz C, Dussmann H, Renz A, Schulze-Osthoff K, Nieminen AL, Poppe M and Prehn JH (2001) Multiple kinetics of mitochondrial cytochrome release in drug-induced apoptosis. *Molecular Pharmacology*, **60**, 1008–1019.

Rehm M, Dussmann H and Prehn JH (2003) Real-time single cell analysis of Smac/DIABLO release during apoptosis. *The Journal of Cell Biology*, **162**, 1031–1043.

Zhang J, Campbell RE, Ting AY and Tsien RY (2002) Creating new fluorescent probes for cell biology. *Nature Reviews. Molecular Cell Biology*, **3**, 906–918.

Basic Techniques and Approaches

Two-dimensional gel electrophoresis (2-DE)

Emma McGregor
Proteome Sciences plc, London, UK

Michael J. Dunn
Proteome Research Centre, University College Dublin, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland

1. Introduction

The human genome contains fewer open reading frames (around 30 000 ORFs) encoding functional proteins than was generally predicted and, like all other completed genomes, contains many "novel" genes with no ascribed functions. Moreover, it is now apparent that one gene does not encode a single protein, because of processes such as alternative mRNA splicing, RNA editing, and posttranslational protein modification. Therefore, the functional complexity of an organism far exceeds that indicated by its genome sequence alone.

The concept of mapping the human complement of protein expression was first proposed more than 25 years ago (Anderson and Anderson, 1982), with the development of a technique in which large numbers of proteins could be separated simultaneously by two-dimensional polyacrylamide gel electrophoresis (2-DE) (Klose, 1975; O'Farrell, 1975), the term "proteome" not being established until the mid-nineties, defining the protein complement of a genome. Further details can be found in recent reviews (Dunn and Gorg, 2001; Rabilloud, 2002).

2. Two-dimensional gel electrophoresis (2-DE)

2-DE involves the separation of solubilized proteins in the first dimension according to their charge properties (isoelectric point, p*I*) by isoelectric focusing (IEF) under denaturing conditions, followed by their separation in the second dimension by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), according to their relative molecular mass (M_r). As the charge and mass properties of proteins are essentially independent parameters, this orthogonal combination of charge (p*I*) and size (M_r) separations results in the sample proteins being distributed across the two-dimensional gel profile (Figure 1). Achieving optimal separation of protein mixtures relies greatly on how the sample is handled and processed prior to 2-DE.

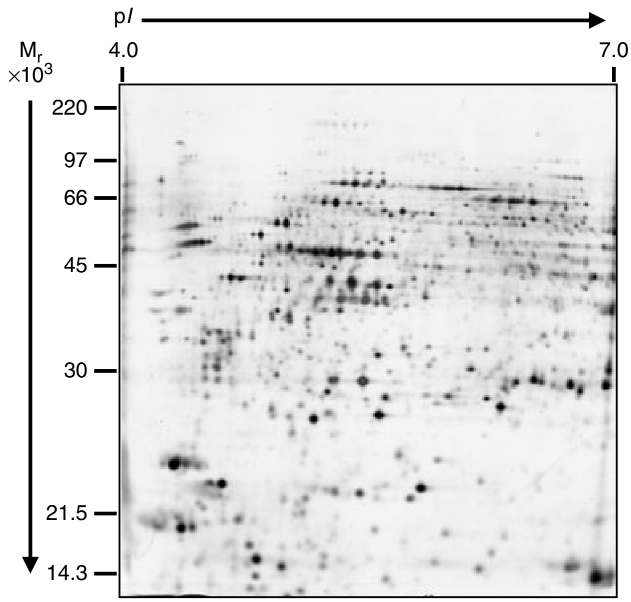


Figure 1 A two-dimensional electrophoresis (2-DE) separation of proteins solubilized from brain grey matter. The first dimension comprised an 18-cm nonlinear pH 4–7 immobilized pH gradient (IPG) subjected to isoelectric focusing. The second dimension was a 21-cm 12% SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) gel. Proteins were detected by silver staining. The nonlinear pH range of the first-dimension IPG strip is indicated along the top of the gel, acidic pH to the left. The M_r (relative molecular mass) scale can be used to estimate the molecular weights of the separated proteins

2.1. Sample preparation

Sample preparation is perhaps the most important step in a proteomics experiment as artifacts introduced at this stage can often be magnified with the potential to impair the validity of the results (*see* Article 2, **Sample preparation for proteomics**, Volume 5). No single method for sample preparation can be applied universally owing to the diverse nature of samples that are analyzed by 2-DE, but some general considerations can be mentioned. The high-resolution capacity of 2-DE is exquisitely able to detect subtle posttranslational modifications such as phosphorylation, but it will also readily reveal artifactual modifications such as protein carbamylation that can be induced by heating of samples in the presence of urea. In addition, proteases present within samples can readily result in artifactual spots, so that samples should be subjected to minimal handling and kept cold at all times, and it is possible to add cocktails of protease inhibitors.

2.2. Solubilization

Ideally, the solubilization procedure for 2-DE would result in the disruption of all noncovalently bound protein complexes and aggregates into a solution of individual

polypeptides. If this is not achieved, persistent protein complexes in the sample are likely to result in new spots in the 2-D profile, with a concomitant reduction in the intensity of those spots representing the single polypeptides. The method of solubilization must also allow the removal of substances, such as salts, lipids, polysaccharides, and nucleic acids as these can interfere with the 2-DE separation. Finally, the sample proteins must remain soluble during the 2-DE process. For these reasons, sample solubility could be viewed as one of, if not the most, critical factors for successful protein separation by 2-DE (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5).

2.3. Isoelectric focusing (IEF)

Immobilized pH gradients (IPG) (Amersham Biosciences), generated using immobiline reagents, are now widely used, replacing the synthetic carrier ampholytes (SCA) previously used to generate the pH gradients required for IEF. The use of IPGs can achieve highly reproducible protein separations and have high resolution.

The IPG IEF dimension (first dimension) of 2-DE is performed on individual gel strips, 3–5-mm wide, cast on a plastic support. IEF is generally carried out using either an IPGphor (Amersham Biosciences) or Multiphor (Amersham Biosciences) apparatus. By applying a low voltage during sample application by in-gel rehydration (Rabilloud *et al*., 1994), improved protein entry, especially of high molecular weight proteins has been reported using the IPGphor (Gorg *et al*., 2000). However, an increased number of detectable protein spots in 2-D gels has been reported following IEF using the Multiphor compared to the same samples subjected to IEF using the IPGphor (Choe and Lee, 2000). Recently, the IPGphor was used for sample loading under a low voltage and the strips were then transferred to the Multiphor for focusing, with an increased sample entry compared to using the Multiphor alone being reported (Craven *et al*., 2002).

The most common IPGs used are those in the pH ranges of 3–10, linear, and nonlinear. However, intermediate (e.g., pH 4–7, 6–9) and narrow (e.g., pH 4.0–5.0, 4.5–5.5, 5.0–6.0, 5.5–6.7) range IPG IEF gels are now available and have the capability of “pulling apart” the protein profile, thus increasing resolution in particular regions of a gel. This “zoom gel” approach results in enhanced proteomic coverage (Westbrook *et al*., 2001).

2.4. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Following steady-state IEF, strips are equilibrated and then applied to the surface of either vertical or horizontal slab SDS-PAGE gels (the second dimension). Using standard format SDS-gels (20 × 20 cm) with 18 cm IPG strips, it is possible to routinely separate 2000 proteins from whole-cell and tissue extracts. Resolution can be significantly enhanced (separation of 5000–10 000 proteins) using large format (40 × 30 cm) 2-D gels. However, gels of this size are very rarely used owing to the handling problems associated with such large gels. The longest commercial IPG

IEF gels have a length of 24 cm. In contrast, mini-gels (7 × 7 cm) can be run using 7-cm IPG strips. While these gels will only separate a few hundred proteins, they can be very useful for rapid screening purposes. Second-dimension SDS-PAGE is usually carried out using apparatus capable of running simultaneously multiple large-format 2-D gels (e.g., Ettan DALT 2, 12 gels, Amersham Biosciences; Protean Plus Dodeca Cell, 12 gels, Bio-Rad). While such equipment allows large numbers of 2-D protein separations to be performed, the procedure is still very time consuming and labor intensive.

2.5. Protein detection and visualization

Following separation by electrophoresis, proteins must be visualized at high sensitivity (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). Appropriate detection methods should therefore ideally combine the properties of a high dynamic range (i.e., the ability to detect proteins present in the gel at a wide range of relative abundance), linearity of staining response (to facilitate rigorous quantitative analysis), and if possible, compatibility with subsequent protein identification by mass spectrometry.

Staining by Coomassie brilliant blue (CBB) has, for many years, been a standard method for protein detection following gel electrophoresis, but its limited sensitivity (around 100-ng protein) stimulated the development of a more sensitive (around 10-ng protein) method utilizing CBB in a colloidal form (Neuhoff *et al*., 1988). Since its first description in 1979 (Switzer, III *et al*., 1979), silver staining has often been the method of choice for protein detection on 2-D gels due to its high sensitivity (around 0.1-ng protein). Detection methods based on the postelectrophoretic staining of proteins with fluorescent compounds have the potential of increased sensitivity combined with an extended dynamic range for improved quantitation (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). The most commonly used reagents are the SYPRO series of dyes from Molecular Probes (Patton, 2000). However, more specialized fluorescent detection compounds are available, for example, Pro-Q Diamond and Pro-Q Emerald, used for the detection of phosphorylated and glycosylated protein isoforms respectively (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5), and the fluorescent cyanine dyes, Cy2, Cy3, and Cy5 used in the staining procedure known as two-dimensional difference gel electrophoresis (DIGE) (Unlu *et al*., 1997; *see also* Article 25, **2D DIGE**, Volume 5 and Article 30, **2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis**, Volume 5).

2.6. Protein identification

Mass spectrometry is used for routine protein identification in proteomics experiments as the associated methods are very sensitive, require small amounts of sample

(femtomole to attomole concentrations), and have the capacity for high sample throughput (Patterson and Aebersold, 1995; Yates, 1998; *see also* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5). Peptide mass fingerprinting (PMF) is typically the primary tool for protein identification (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5). It is based on a set of peptide masses obtained by MS analysis of a protein digest (usually using trypsin) that provides a characteristic mass fingerprint of that protein (*see* Article 16, **Improvement of sequence coverage in peptide mass fingerprinting**, Volume 5). Identification of the protein is then achieved by comparison of the experimental mass fingerprint with theoretical peptide masses generated in silico using protein and nucleotide sequence databases. This approach proves very effective when trying to identify proteins from species whose genomes are completely sequenced, but is not so reliable for organisms whose genomes have not been completed. If it proves impossible to identify a protein based on PMF alone, it is then essential to obtain amino acid sequence information. This is most readily accomplished using tandem mass spectrometry (MS/MS), using either MALDI-MS with post-source decay (PSD) or chemical assisted fragmentation (CAF), MALDI-TOF-TOF-MS/MS or ESI-MS/MS triple-quadropole, ion-trap, or Q-TOF instruments (*see* Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5, Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5, Article 10, **Hybrid MS**, Volume 5, and Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5) to induce fragmentation of peptide bonds.

3. Conclusion

The current array of proteomic techniques makes it possible to characterize global alterations in protein expression associated with the progression of many different human diseases; however, there is still no one method that can replace 2-DE and its ability to simultaneously separate and display several thousands of proteins from complex mixtures. Although there are disadvantages associated with 2-DE, for example, separation of hydrophobic and/or membrane proteins, 2-DE nevertheless remains the core technique of choice for the separation of complex protein mixtures in the majority of current proteomics projects.

Acknowledgments

We thank Dr. David Cotter and Kyla Pennington for the use of the gel image presented in Figure 1.

References

Anderson NG and Anderson L (1982) The human protein index. *Clinical Chemistry*, **28**, 739–748.

Choe LH and Lee KH (2000) A comparison of three commercially available isoelectric focusing units for proteome analysis: the multiphor, the IPGphor and the protean IEF cell. *Electrophoresis*, **21**, 993–1000.

Craven RA, Jackson DH, Selby PJ and Banks RE (2002) Increased protein entry together with improved focussing using a combined IPGphor/Multiphor approach. *Proteomics*, **2**, 1061–1063.

Dunn MJ and Gorg A (2001) Two-dimensional polyacrylamide gel electrophoresis for proteome analysis. In *Proteomics, From Protein Sequence to Function*, Pennington SR and Dunn MJ (Eds.), BIOS Scientific Publishers Ltd: Oxford, pp. 43–63.

Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R and Weiss W (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053.

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, **26**, 231–243.

Neuhoff V, Arold N, Taube D and Ehrhardt W (1988) Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis*, **9**, 255–262.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, **250**, 4007–4021.

Patterson SD and Aebersold R (1995) Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis*, **16**, 1791–1814.

Patton WF (2000) A thousand points of light: the application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis*, **21**, 1123–1144.

Rabilloud T, Valette C and Lawrence JJ (1994) Sample application by in-gel rehydration improves the resolution of two-dimensional electrophoresis with immobilized pH gradients in the first dimension. *Electrophoresis*, **15**, 1552–1558.

Rabilloud T (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics*, **2**, 3–10.

Switzer RC III, Merril CR and Shifrin S (1979) A highly sensitive silver stain for detecting proteins and peptides in polyacrylamide gels. *Analytical Biochemistry*, **98**, 231–237.

Unlu M, Morgan ME and Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Westbrook JA, Yan JX, Wait R, Welson SY and Dunn MJ (2001) Zooming-in on the proteome: very narrow-range immobilised pH gradients reveal more protein species and isoforms. *Electrophoresis*, **22**, 2865–2871.

Yates JR III (1998) Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, **33**, 1–19.

2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis

Edward Hawkins and Stephen O. David
GE Healthcare, Chalfont St., Giles, UK

1. 2-D DIGE

Two-dimensional polyacrylamide gel electrophoresis (2-DE) was first introduced in 1975 to separate complex protein mixtures (O'Farrell, 1975; Klose, 1975). It is a powerful technique allowing the simultaneous visualization of large portions of the Proteome (approximately 3000 spots per gel). Unfortunately, the issues of lack of reproducibility, sensitivity, and quantitative capabilities of conventional 2-DE have limited the use of this technique as a quantitative tool (Alban *et al*., 2003).

A group at the Carnegie Mellon University in Pittsburg first described a method that enabled more than one sample to be separated in a single two-dimensional (2-D) polyacrylamide gel, known as 2-D Difference Gel Electrophoresis (DIGE) (Unlu *et al*., 1997).

Two-dimensional difference gel electrophoresis (2-D DIGE) builds upon the 2-D gel electrophoresis technique to provide highly accurate quantitative capabilities that is not possible to easily achieve with traditional 2-DE.

2-D DIGE enables up to three protein extracts to be separated on the same 2-D gel. This is made possible through the labeling of each extract using spectrally resolvable, size- and charge-matched fluorescent dyes known as CyDye™ DIGE Fluors (Alban *et al*., 2003). The samples are labeled prior to electrophoresis, using either Cy2, Cy3, or Cy5, mixed together, and separated using conventional 2-DE. Gels are then scanned using a Fluorescence Imager, such as Typhoon variable mode imager, to visualize all samples on the gel and image analysis is then performed using specialized 2-D software known as DeCyder™ 2-D Differential Analysis Software.

2. Internal Standard: a reference sample in 2-DE

Multiplexing of samples in 2-D DIGE adds to 2-DE, the advantages of a reference sample commonly seen in microarray studies. This reference sample known as

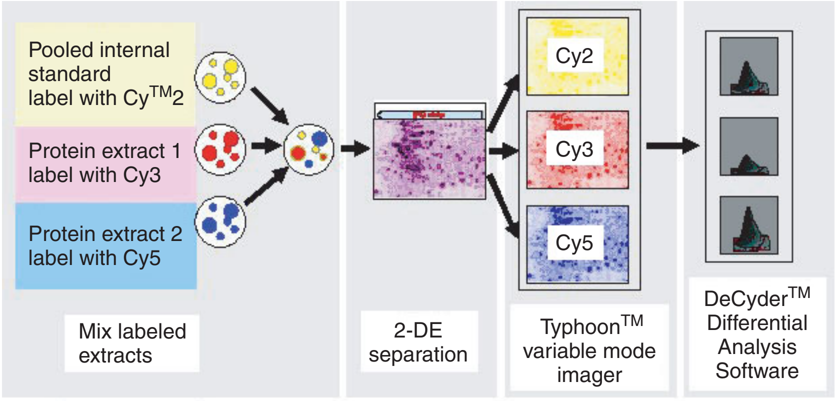


Figure 1 Principle of 2-D DIGE

the "Internal Standard" in 2-D DIGE is made of equal quantities of each of the biological samples being studied. Electrophoresis of the "Internal Standard" on each gel (see Figure 1) in the experiment along with the individual biological samples means that, the abundance of each protein spot on a gel can be measured relative (i.e., semiquantitatively as a ratio) to its corresponding spot in the internal standard present on the same gel.

Measurement of protein abundance as a "Standardized Ratio" instead of an "Absolute volume" greatly reduces the influence of experimental variation on the measured abundance of a protein spot on a 2-D gel. This reduction in experimental variation allows the detection and accurate quantitation of real biological expression changes free from the "noise" normally associated with experimental variation. This was demonstrated by Knowles *et al*. (2003) in a study that enabled them to accurately measure changes in protein expression as low as 10%.

Ettan™ DIGE is a complete system that has been optimized to fully benefit from the advantages provided by the 2-D DIGE technique.

3. Workflow

3.1. Labeling

The CyDye™ DIGE Fluors are Cyanine dyes and are supplied in two forms, CyDye™ DIGE Fluor minimal dyes (Cy2, Cy3, and Cy5) and CyDye™ DIGE Fluor saturation dyes (Cy3 and Cy5).

The "minimal" dyes react at pH 8.5 to specifically label the epsilon amino group of lysine residues. The ratio of dye to protein is kept very low (typically 50 µg of protein to 400 pmoles of dye) to ensure that the only proteins visualized are those labeled with a single dye molecule. This approach, referred to as "minimal labeling", ensures that the pattern seen on a 2-D gel is the same as that seen for a poststaining technique. The advantage of labeling lysine is that it is an abundant amino acid and is well represented in the total protein population (approximately

7.2% of all vertebrate proteins contain lysine). However, if this amino acid is not present in a protein, then this protein will not be visualized. This labeling approach requires approximately 50 µg of each protein sample to be loaded unto each gel and is suitable for situations where sufficient protein sample is available (e.g., Alban *et al*., 2003; Tonge *et al*., 2001; Zhou *et al*., 2002).

The "saturation" dyes react at pH 8.0 to specifically label the thiol group of cysteine residues, after an initial reduction step. The ratio of dye to protein is much higher (typically 5 µg to 4 nmoles of dye) to ensure that all the available cysteine groups are labeled. This approach is referred to as "saturation labeling". The lower proportion of proteins that contains cysteines (approximately 3.3% of all vertebrate proteins contain cysteines) in comparison to lysines in the total protein population means that some proteins visualized using the "minimal" dyes will not be seen using the "saturation" dyes. For this reason, the "minimal" dyes are used for general applications, while the "saturation" dyes are used for more specialized applications. The "saturation labeling" approach requires only approximately 5 µg of each protein sample to be loaded unto each gel and is ideal in situations in which only a small amount of protein is available. For example, samples obtained using techniques such as Laser Capture Microdissection. (Kondo *et al*., 2003)

3.2. Separation and imaging

Following separation of protein extracts on a standard two-dimensional (2-D) polyacrylamide gel, a Fluorescence scanner, for example the Typhoon™ 9000 series scanner, is used to scan the gels. The scanner has the ability to spectrally distinguish between the three CyDye Fluors, through the use of a combination of several lasers and filters. This enables the different samples on the same gel to be visualized individually.

3.3. Image analysis

The digital images generated by the imager can then be analyzed using an Image analysis software package, for example DeCyder™, that calculates the amount of fluorescence present in each spot. The DeCyder software utilizes proprietary algorithms to perform the unique form of image analysis required to accurately analyze the images generated using this technique, compared to standard 2-D image analysis. One such algorithm known as codetection generates identical spot detection on all images from the same gel. This method of detection decreases image analysis variation and increases accuracy in a manner that is not possible using standard image analysis algorithms. In addition, the presence of an internal standard on each gel enables gel-to-gel matching of all samples to be achieved through the matching of only the internal standard present on each gel.

Images from each individual gel are codetected in the DIA module of DeCyder (Figure 2). Gel-to-gel matching in the BVA module of DeCyder is achieved by choosing one of the internal standard images as a master image and then matching only the internal standard sample present on each gel to the master. After matching, proteins that are differentially expressed can then be identified.

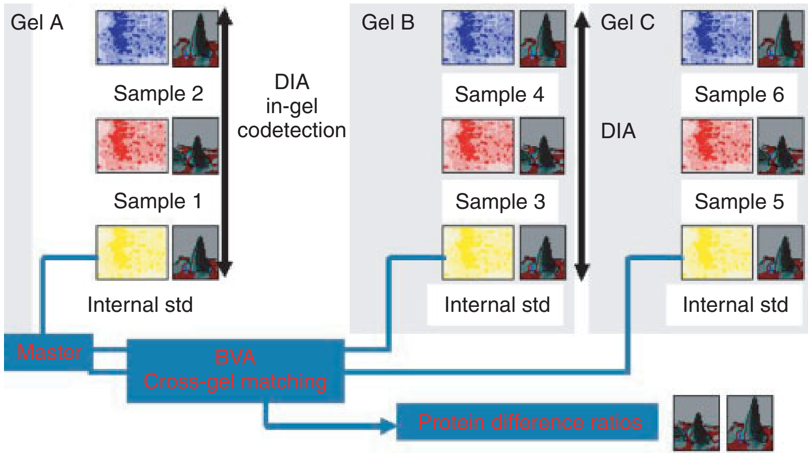


Figure 2 Codetection and matching

3.4. *Spot handling*

Following user confirmation of proteins of interest, a pick list of spots to be identified by mass spectrometry (MS) is then exported from the image analysis software to a spot excision robot. Spots can then be excised directly from a poststained analytical gel (if the gel contains sufficient protein for MS analysis) or from a separate post stained preparative gel. Picked spots are then digested and spotted for MS.

3.5. *2-D DIGE: a tool for proteomics*

Profiling proteomics involves the identification of proteins that are differentially expressed between biological samples or the identification of the protein complement in a sample. The sensitivity, broad dynamic range, reproducibility, and accuracy of quantitation obtained with 2-D DIGE makes this an ideal tool for proteome profiling.

Several proteomes have been successfully studied using this technique including those of bacteria, yeast, plants, fruit fly, insect, mouse and rat liver, rat kidney, rat heart, rat lung, cat brain, mouse and rat brain, guinea pig brain, human brain, and human cancer cells.

Further reading

Lilley KS and Friedman DB (2004) All about DIGE: quantification technology for differential-display 2D-gel proteomics. *Expert Review of Proteomics*, **1**, 401–409.

Website (2005) **http://www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm**.

Website (2005) **http://www1.amershambiosciences.com/aptrix/upp01077.nsf/content/uk_homepage**.

References

Alban A, David SO, Bjorksten L, Andersson C, Sloge E, Lewis S and Currie I (2003) A novel experimental design for comparative two-dimensional gel analysis: Two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.

Klose J (1975) Protein mapping by combined isoelectric focussing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutation in mammals. *Humangenetik*, **26**, 231–243.

Knowles MR, Cervino S, Skynner HA, Hunt SP, De Felipe C, Salim K, Meneses-Lorente G, McAllister G and Guest PC (2003) Multiplex proteomic analysis by two-dimensional differential in-gel electrophoresis. *Proteomics*, **3**, 1162–1171.

Kondo T, Seike M, Mori Y, Fujii K, Yamada T and Hirohashi S (2003) Application of sensitive fluorescent dyes in linkage of laser microdissection and two-dimensional gel electrophoresis as a cancer proteomic study tool. *Proteomics*, **3**, 1758–1766.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological Chemistry*, **250**, 4007–4021.

Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, Pognan F, Hawkins E, Currie I and Davison M (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics*, **1**, 377–396.

Unlu M, Morgan ME and Minden JS (1997) A single method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Zhou G, Li H, DeCamp D, Chen S, Shu H, Gong Y, Flaig M, Gillespie JW, Hu N, Taylor PR, *et al*. (2002) 2D, Differential In-gel Electrophoresis for the Identification of Esophageal Scans Cell Cancer-specific Protein Markers. *Molecular & Cellular Proteomics*, **1**, 117–123.

MS-based methods for identification of 2-DE-resolved proteins

Achim Treumann
Royal College of Surgeons in Ireland, Dublin, Ireland

Christopher Gerner
Medical University of Vienna, Vienna, Austria

1. Introduction

When O'Farrell (1975) and Klose (1975) published the first reports of separating and displaying complex protein mixtures on two-dimensional (2-D) gels (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5) on the basis of isoelectric points and electrophoretic mobility, the research community rapidly adopted two-dimensional electrophoresis (2-DE) as the method of choice for the analysis of the protein content of tissues, cells, and organelles.

However, identification of the proteins in these 2-D maps posed a significant problem in the early years of 2-DE. The advent of new ionization techniques that were capable of generating ions of nonvolatile analytes and transferring them into the gas phase for mass spectrometric (MS) analysis invigorated the 2-DE research field. Matrix-assisted laser desorption ionization (MALDI) coupled to time-of-flight (TOF) detection (MALDI-TOF) (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) and electrospray ionization (ESI) (*see* Article 18, **Techniques for ion dissociation (fragmentation) and scanning in MS/MS**, Volume 5) opened up the possibility of identifying and characterizing many proteins in a rapid and sensitive manner.

2. Peptide mass fingerprinting

2.1. Experimental

The most frequently used approach to protein identification from 2-D gels takes advantage of the relatively high purity of proteins in individual spots that are cut

out from 2-D gels. Peptide mass fingerprinting (PMF) is based on the realization that the masses of a number of peptides that are generated from one specific protein through sequence-specific proteolysis (frequently using trypsin, an enzyme that cleaves proteins after Arg and Lys) provide sufficient information for the unambiguous identification of this protein.

The following workflow is typical for PMF-based protein identification.

- A spot of interest is cut out from a gel and subjected *in-gel* to sequence-specific proteolysis.
- The resulting peptides are eluted from the gel and analyzed by mass spectrometry, most frequently by MALDI-TOF MS.
- These masses are then compared to datasets produced by in silico digestion of large numbers of proteins in a database. A score is calculated to assess the probability that the experimental dataset is derived from particular proteins in the database.

For high-throughput analyses, it is important that all of these steps can now be automated, significantly increasing the reproducibility and the speed of analysis.

MALDI is the ionization method of choice for PMF, due to its relatively higher tolerance of low-level contaminants when compared to ESI, to its high level of automation, and to its general combination with very sensitive and accurate TOF-detectors. The increased mass accuracy restricts the number of isobaric peptides in a database, and this greatly improves the stringency of the database search (Clauser *et al*., 1999).

2.2. *Data interpretation (search engines and algorithms)*

A large number of different search engines (**Mascot**, **Protein Prospector (MS-Fit)**, **ProFound**, **MassSearch**, **PeptIdent**) are available to perform the search of PMF data against a variety of different databases. Many of these provide free access through the Internet.

Generally, however, it is advisable to purchase a license for an in-house version of these search engines, as these will allow customization to the user's needs.

3. Peptide fragmentation

3.1. *Experimental*

Fragment ion spectra of peptides (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5 and Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5) provide sequence information and are as a consequence far more informative than the mass of a peptide. Sequence from one single peptide can provide sufficient information for the unambiguous identification of the protein this peptide is derived from. As the accessibility of a sequenced genome is an essential prerequisite for the identification of proteins via

PMF, fragmentation spectra are the only way to identify proteins from organisms where ordered genomic databases are not yet available.

Whereas the initial stages of the workflow for protein identification by peptide fragmentation are identical to those for peptide mass fingerprinting (cutting spots out of a gel, *in-gel* proteolytic digest and elution of the peptide mixture from the gel), there are various ways of performing the subsequent MS analysis as follows.

- Post source decay (PSD) on a MALDI-TOF mass spectrometer. Because of limited sensitivity and difficult data analysis, this method is rarely used.
- LC-ESI-MS with collision-induced dissociation (CID) on an iontrap or a hybrid mass spectrometer (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5 and Article 10, **Hybrid MS**, Volume 5). When used for single spots from 2-DE, LC-ESI-MSMS is generally more time consuming than the acquisition of a mass fingerprint and it requires separate instrumentation from that used for PMF. However, it can be automated and it yields high-quality spectral information on peptides that can be chosen automatically during acquisition. Annotations of a majority of the spots in 2-DE gels have been performed using exclusively LC-coupled CID on hybrid mass spectrometers (e.g., O'Neill *et al*., 2002).
- Nanospray ESI-MS (*see* Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5 and Article 19, **Making nanocolumns and tips**, Volume 5). A separate aliquot of the digest is infused into a CID-capable mass spectrometer at very low flow rates (typically nl min^{-1}) through a metal-coated class capillary. Recently, a commercial device has been introduced that automates sample introduction at nanoflow rates, bypassing the difficult and time-consuming preparation and positioning of sample-filled glass capillaries (Van Pelt *et al*., 2002).
- MALDI-TOFTOF MS. Recently, new MALDI-TOF mass spectrometers have reached the market that permit the rapid and sensitive acquisition of peptide fragmentation spectra. For the purpose of protein identification from 2-DE spots, these mass spectrometers are the instrumentation of choice as peptide mass fingerprints, and peptide sequence information can be obtained unattended in one run on a single instrument.

3.2. Data interpretation (search engines and algorithms)

De novo interpretation of peptide fragmentation spectra is a lengthy and difficult task for which as yet no reliable computer algorithms have been developed. It is always necessary to check the output of computer-generated de novo sequences manually, which makes this unsuitable for high-throughput studies.

Several approaches have been developed to match peptide fragmentation spectra to peptide sequences from protein databases (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5 and Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5). Independent of the choice of search engine, it is still advisable to check assigned spectra of intermediate quality manually to be confident of excluding false-positives.

4. Gel staining and MS response

Protein spots that are to be excised from a 2-D gel for the purpose of identification need to be visualized before the excision and digestion (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). It is essential that the staining of the gel is compatible with MS analysis and that it is sensitive enough to allow for the identification of a large number of proteins from a single gel.

Despite its relatively low sensitivity, colloidal Coomassie Brilliant Blue (CBB) staining is still one of the most popular staining methods in 2-DE-based proteomics studies. This is due to the simplicity of the staining procedure and to its excellent compatibility with MS-based protein identification. Any spots that are visible by CBB staining should be identifiable by MS-based analysis.

Silver-based staining is about 1 order of magnitude more sensitive than CBB staining, and MS-compatible silver staining methods have been developed (Shevchenko *et al*., 1996) and are in use in several proteomic laboratories around the world.

Fluorescence-based stains (Berggren *et al*., 1999; Rabilloud *et al*., 2001) provide excellent sensitivity and linearity for protein staining in gels, contributing to their superiority over silver-based staining methods, while also being compatible with MS-based protein identification.

5. Conclusions

Recent advances in sample preparation and analysis are addressing some of the drawbacks of 2-DE coupled with MS-based protein identification, in particular, the selectivity of 2-DE due to solubilization issues in the isoelectrophoretic dimension (leading to an underrepresentation of large, small, very acidic, very basic, and very hydrophobic proteins). The limited dynamic range of 2-DE-based experiments has been improved through careful experimental design and purification of subproteomes before analysis and through increased sensitivity of new mass spectrometers that have entered the market.

Proteome analysis through the 2-DE-based approach remains popular despite the development of several competing chromatography or chip-based technologies. This is due to the unsurpassed capability of 2-DE to display a large number of proteins from very complex mixtures, allowing for sample comparison and relative quantification. It is very likely that 2-DE coupled with MS-based protein identification will remain one of the mainstays of proteome analysis for expression profiling and for functional proteomics in years to come.

Further reading

Aebersold R and Goodlett DR (2001) Mass spectrometry in proteomics. *Chemical Reviews*, **101**, 269–295.

Ong SE and Pandey A (2001) An evaluation of the use of two-dimensional gel electrophoresis in proteomics. *Biomolecular Engineering*, **18**, 195–205.

References

Berggren K, Steinberg TH, Lauber WM, Carroll JA, Lopez MF, Chernokalskaya E, Zieske L, Diwu ZJ, Haugland RP and Patton WF (1999) A luminescent ruthenium complex for ultrasensitive detection of proteins immobilized on membrane supports *Analytical Biochemistry*, **276**, 129–143.

Clauser KR, Baker P and Burlingame AL (1999) Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS MS and database searching. *Analytical Chemistry*, **71**, 2871–2882.

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, **26**, 231–243.

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *The Journal of Biological Chemistry*, **250**, 4007–4021.

O'Neill EE, Brock CJ, von Kriegsheim AF, Pearce AC, Dwek RA, Watson SP and Hebestreit HF (2002) Towards complete analysis of the platelet proteome. *Proteomics*, **2**, 288–305.

Rabilloud T, Strub JM, Luche S, van Dorsselaer A and Lunardi J (2001) Comparison between Sypro Ruby and ruthenium II tris (bathophenanthroline disulfonate) as fluorescent stains for protein detection in gels. *Proteomics*, **1**, 699–704.

Shevchenko A, Wilm M, Vorm O and Mann M (1996) Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Analytical Chemistry*, **68**, 850–858.

Van Pelt C, Zhang S and Henion J (2002)) Characterization of a fully automated nanoelectrospray system with mass spectrometric detection for proteomic analyses. *Journal of Biomolecular Techniques*, **13**, 72–84.

Arraying proteins and antibodies

David S. Wilson and Steffen Nock
Absalus Inc., Mountain View, CA, USA

1. Solid and suspension arrays

Two fundamentally different types of protein arrays (*see* Article 24, **Protein arrays**, Volume 5 and Article 100, **Protein microarrays as an emerging tool for proteomics**, Volume 4) are currently in use: (1) the classical two-dimensional (2D) array format and (2) encoded particle-based systems. Both types of arrays consist of distinct features onto which small amounts of certain proteins are immobilized. The distinction between 2D and liquid arrays consists in the manner in which the different features are "addressed" such that the identity of the immobilized protein can be deciphered. The features in 2D arrays are addressed spatially according to the arrangement used in dispensing the various molecules at known locations on a planar surface (*see* Article 90, **Microarrays: an overview**, Volume 4). The features in the liquid arrays, on the other hand, lose their spatial addresses since they consist of individual particles that, after being derivatized by proteins, are mixed together to form a suspension in a buffer. The addressing of the features in liquid arrays is achieved by encoding the particles optically in such a way that they can be "read" after being mixed together. The most common way to encode beads relies on the incorporation of different ratios of two or more fluorophores (Kettman *et al*., 1998). A flow cytometer is used to read particles one at a time, thus decoding the identity of each particle (and thereby the identity of the immobilized protein) while simultaneously reading the signal that is the object of the assay. The most popular suspension array platform is marketed by Luminex (Austin, TX). Some aspects of the different systems are compared in Table 1. Instruments and reagents are available commercially to allow researchers to construct their own 2D or suspension arrays, and ready-made arrays (*see* Article 33, **Basic techniques for the use of reverse phase protein microarrays for signal pathway profiling**, Volume 5) of both types are also available for certain important classes of proteins (cytokines, phospho proteins etc.).

2. Methods of protein immobilization

Proteins are very fragile molecules that are easily denatured at liquid–solid and liquid–air interfaces. All of these issues are significantly mitigated by using suspension arrays, but even in such cases, the nature of the substrates used for

Table 1 Comparison of multiplexed protein analysis platforms

	Two-dimensional arrays	Suspended particle arrays
Assay type	Direct capture, sandwich assay, functional assays	Direct capture, sandwich assay
Array construction method	Microspotting of all specificities onto each array	Batch incubation of particles followed by particle mixing
Detection (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5)	Fluorescence, radioactivity, SPR, Mass spectrometry	Fluorescence
Multiplexing	High	Medium
Sample throughput	Low	High
Protein concentration required to saturate the surface	High ($\sim$1 mg ml^{-1})	May be done under dilute conditions
Storage	Dry	In buffer

protein immobilization will influence the protein density on the surfaces and the fraction of immobilized protein that is active.

Three fundamentally different approaches to protein immobilization have been described.

1. *Physical adsorption:* The most common and easiest method is to immobilize the protein via multiple noncovalent, heterogeneous linkages to the surface. The adherent forces are based on nonspecific adsorption onto hydrophobic surfaces such as plastics or gold, or onto charged surfaces such as naked or amine-derivatized glass, or nitrocellulose. When proteins come into contact with such surfaces, they spontaneously coat them to form a protein film. In such films, each protein is attached via multiple weak interactions (hydrophobic, electrostatic and van der Waals), but since the interaction surface on each protein is large, the sum of these weak interactions results in binding energies that are orders of magnitude stronger than typical "specific" protein–protein interactions, for example. The main disadvantage of these approaches is that, in order for a protein to make such extensive contact with the surface, it is often denatured. Generally, 75–90% of proteins attached to surfaces in this way are thought to be inactive (Butler *et al*., 1993).
2. *Random chemical attachment:* The second most popular and easy to use method is based on random covalent attachment of proteins. Typically, the surface is derivatized with an amino-reactive functionality such as *N*-hydroxy-succinimide, which then couples with the side chain amino group of lysines on proteins. Lysine is a common amino acid and its amino group is almost always on the surface of a folded protein. The number and positions of the attachment sites are heterogeneous, as is the orientation of the immobilized proteins. As a result, some fraction of the proteins will be inactive due to denaturation arising from strain caused by the multiple attachment sites or to modification of the actual binding or catalytic site. Random chemical attachment is the method

used for most of the suspension arrays, such as the Luminex system, and has also been used extensively for planar arrays (MacBeath and Schreiber, 2000). In some cases, the protein is attached indirectly to the surface through random biotinylation of the protein and use of streptavidin-coated surfaces (Peluso *et al*., 2003).

3. *Site-specific attachment:* The most elegant method for immobilizing proteins is by single-point attachment to an appropriate side chain on the protein surface. The most commonly used approach is to take advantage of a unique cysteine residue that is either naturally present in a protein or is introduced into a recombinant protein by genetic engineering. Surfaces can be derivatized by a number of thiol-reactive functionalities such as maleimide or haloacetamides. Under appropriate conditions, the coupling can occur almost exclusively through the unique thiol group, such that the immobilized proteins are homogenously oriented with respect to the surface. Several studies have demonstrated the superior performance of proteins immobilized by this type of method (see Peluso *et al*., 2003 and references therein). Recombinant proteins can also be specifically oriented onto surfaces by incorporating affinity tags such as the hexahistidine tag (Zhu *et al*., 2001). Site-specific chemical or enzymatic biotinylation can be used to orient proteins on streptavidin-coated surfaces (Peluso *et al*., 2003).

The choice of methods for immobilization depends highly on the application, the number and types of proteins that will be immobilized, and the sensitivity required for the assay. The larger the number of proteins to be immobilized, the more work is involved and hence simplicity is highly valued. The size and stability of the proteins is also very relevant. Unstable or small proteins are particularly susceptible to inactivation by immobilization through physical adsorption or random chemical attachment. For larger proteins, the likelihood that a particular protein will have its binding site, as opposed to another region of the protein, chemically or structurally modified is less likely.

Antibodies are very stable and large (150 kDa) molecules and hence activity is generally well retained upon random chemical attachment to surfaces. Noncovalent adsorption may also be used, but generally only 10–25% of the antibodies retain their activity under these conditions (Butler *et al*., 1993). For the best performance, antibodies can be immobilized in an oriented fashion through their carbohydrate groups on the Fc region, or by a more complex and less robust method that involves proteolytic processing and selective reduction of certain disulfide bonds (Peluso *et al*., 2003). While it is possible to engineer single cysteines into recombinant antibodies expressed in *Escherichia coli*, the production yields are typically too low to justify the added expense. Recombinant antibody fragments known as single-chain Fvs are notoriously unstable compared to larger fragments or intact antibodies.

3. Manufacturing arrays

Most of the devices for printing 2D arrays were originally developed for DNA spotting (*see* Article 91, **Creating and hybridizing spotted DNA arrays**,

Volume 4) and are not ideal for printing proteins. Printing methods are of two types – contact and noncontact. In contact printing, a tiny needle or capillary takes up a small amount of protein solution and contacts the array surface, thus transferring a sub-nanoliter volume (Haab *et al*., 2001). Contact printing can sometimes damage the substrate surface. Noncontact printing can be achieved with ink-jet technology (*see* Article 95, **Integrating genotypic, molecular profiling, and clinical data to elucidate common human diseases**, Volume 4), which allows picoliter volumes of protein solution to be applied to a surface without direct physical contact between the printing device and the array (Roda *et al*., 2000). A piezo-based printing device is currently marketed by Perkin-Elmer Life Sciences (Foster City, CA).

Both contact and noncontact-based printing, however, generally results in protein dehydration during the long serial printing process. This problem can be ameliorated by spotting the proteins in a humidified chamber and adding glycerol to the protein solution in order to prevent evaporation (MacBeath and Schreiber, 2000). Another option is to use hydrogel surfaces, which have the added advantage of supporting a higher density of protein per unit area on the array, due to the third dimension. This approach is reviewed by Mirzabekov and Kolchinsky (2002), and a commercial slide-based polyacrylamide is currently marketed by Perkin-Elmer Life Sciences (Wellesley, MA), which claims that the slides can immobilize antibodies at a density and activity capable of binding 13.5 pmol cm^{-2} of antigen. This density is five- to sixfold higher than can be obtained on optimized flat surfaces (Peluso *et al*., 2003). Using this novel hydrogel surface, a higher percentage of the immobilized protein was active than on flat polylysine-coated slides, and this translated to a higher signal-to-noise ratio (Miller *et al*., 2003).

The numerous difficulties in manufacturing reproducible, highly active 2D arrays (inconsistency in drop size and shape, protein denaturation at the liquid–air interphase, dehydration etc.) are irrelevant for the manufacturing of suspension arrays (Carson and Vignali, 1999). In the latter case, each protein solution is mixed with a single type of particle carrying a unique optical code. Protein-conjugated beads can be stored for over 1 year in suspension, so there are no dehydration issues. The different types of beads are then mixed together to create the suspension array. In the near future, there will be commercially available optically encoded magnetic beads, which will allow for high-throughput separations to be integrated into multiplexed protein detection assays.

References

Butler JE, Ni L, Brown WR, Joshi KS, Chang J, Rosenberg B and Voss EW (1993) The immunochemistry of sandwich ELISAs – VI. Greater than 90% of monoclonal and 75% of polyclonal anti-fluorescyl capture antibodies (CAbs) are denatured by passive adsorption. *Molecular Immunology*, **30**, 1165–1175.

Carson RT and Vignali DA (1999) Simultaneous quantitation of 15 cytokines using a multiplexed flow cytometric assay. *Journal of Immunological Methods*, **227**, 41–52.

Haab BB, Dunham MJ and Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**, RESEARCH0004.

Kettman JR, Davies T, Chandler D, Oliver KG and Fulton RJ (1998) Classification and properties of 64 multiplexed microsphere sets. *Cytometry*, **33**, 234–243.

MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Butler EB, Teh BS and Haab BB (2003) Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers. *Proteomics*, **3**, 56–63.

Mirzabekov A and Kolchinsky A (2002) Emerging array-based technologies in proteomics. *Current Opinion in Chemical Biology*, **6**, 70–75.

Peluso P, Wilson DS, Do D, Tran H, Venkatasubbaiah M, Quincy D, Heidecker B, Poindexter K, Tolani N, Phelan M, *et al*. (2003) Optimizing antibody immobilization strategies for the construction of protein microarrays. *Analytical Biochemistry*, **312**, 113–124.

Roda A, Guardigli M, Russo C, Pasini P and Baraldini M (2000) Protein microdeposition using a conventional ink-jet printer. *Biotechniques*, **28**, 492–496.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Basic techniques for the use of reverse phase protein microarrays for signal pathway profiling

Virginia Espina, Julia Wulfkuhle and Lance A. Liotta
National Cancer Institute, Bethesda, MD, USA

Valerie S. Calvert and Emanuel F. Petricoin III
Center for Biologics Evaluation and Research, Bethesda, MD, USA

1. Introduction

At a functional level, cancer is both a proteomic and a genomic disease. A future in which the functional state of a large portion of the protein signal pathways within a patient's biopsy is analyzed could be realized within the next few years (Liotta *et al*., 2001; Petricoin *et al*., 2002; *see also* Article 96, **The promise of gene signatures in cancer diagnosis and prognosis**, Volume 4 and Article 100, **Protein microarrays as an emerging tool for proteomics**, Volume 4). A molecular network map of the state of key signaling pathways within a patient's tumor cells will become the starting point for a personalized and targeted therapy tailored to the individual tumor's molecular defect. Armed with the "wiring diagram" of the deranged cancer cell, it could be feasible to develop and administer a rationally selected combination therapy that targets multiple interdependent points along a pathogenic pathway, or targeting separate pathways. Following rebiopsy or molecular imaging, the effect of the treatment could be monitored in real time. In fact, an overarching goal for many investigators is to develop a "circuit map" of the cellular protein network (Blume-Jensen and Hunter, 2001; Bowden *et al*., 1999; Celis and Gromov, 2003; Hunter, 2000; Jeong *et al*., 2000; Charboneau, 2002; Chen *et al*., 2002; Cutler, 2003; Delehanty and Ligler, 2002; Ideker *et al*., 2001; Wilson and Nock, 2003; Zhu and Snyder, 2001; Zhu and Snyder, 2003; Huels *et al*., 2002; Humphery-Smith *et al*., 2002; MacBeath, 2002; MacBeath and Schreiber, 2000; Miller *et al*., 2003; Knezevic *et al*., 2001; Paweletz *et al*., 2001; Grubb *et al*., 2003; Wulfkuhle *et al*., 2003; Nishizuka *et al*., 2003a,b; Liotta *et al*., 2003; Roberts *et al*., 2004). Protein microarrays are especially well suited

for this approach since they represent an emerging technology that can detect many simultaneous posttranslational events in an extensively parallel mode, which is not possible using gene transcript profiling. Thus, protein microarrays offer an extremely attractive and complementary technology platform to existing expression profiling platforms (*see* Article 61, **Posttranslational modification of proteins**, Volume 6).

2. Classes of protein array technology

In general, protein microarray formats fall into two major classes: (1) forward phase arrays (FPA) and (2) reverse phase arrays (RPA), depending on whether the analyte to be measured is captured out of a solution phase or prebound to the solid phase (Figure 1) (*see* Article 24, **Protein arrays**, Volume 5). Using the FPA format, a capture or bait molecule, usually an antibody, is immobilized onto a substratum (e.g., derivatized glass, nitrocellulose, hydrogel). Each spot contains one specific and distinct immobilized antibody or bait protein. Using the FPA format, many hundreds to thousands of distinct bait molecules can be arrayed on a slide and each array is incubated with one test sample (e.g., a cellular lysate from one treatment condition). Conversely, the RPA format immobilizes an individual test sample in each array spot, such that an array is composed of hundreds of different patient samples or cellular lysates. In the RPA format, each array is incubated with one detection protein (e.g., antibody) and a single analyte endpoint is measured and directly compared across multiple samples.

3. Technical considerations

In contrast to gene microarray construction, the probes used (e.g., antibodies, aptamers, ligands, drugs) for protein microarrays cannot be directly manufactured with a priori predictable affinity and specificity. The availability of high-quality, specific antibodies or suitable protein-binding ligands is the single most limiting factor, and the starting point, for the successful utilization of protein microarray technology (Templin *et al*., 2002). Certainly, before using either RPA or FPA format, antibody specificity must be *thoroughly* validated. Each antibody to be used on an RPA is first analyzed on Western blot with suitable positive and negative controls and then a complex biological sample similar to that applied and analyzed on the array is analyzed. A significant challenge for protein microarrays in which a specific phosphorylation site is to be measured is the requirement for antibodies that are specific to the modification or activation state of the target protein. Fortunately, many sets of high-quality modification state–specific antibodies are commercially available, with many new phospho-specific antibodies becoming available monthly.

Most gene array platforms use planar silica substrata to immobilize the probes (Miller *et al*., 2002), and protein microarrays have used the same substrata and spotting equipment. While this approach may be successful for high abundant or recombinant analytes, a planar substratum may not be able to attain sufficient surface area per spot for femtomolar detection of analyte proteins in biologically

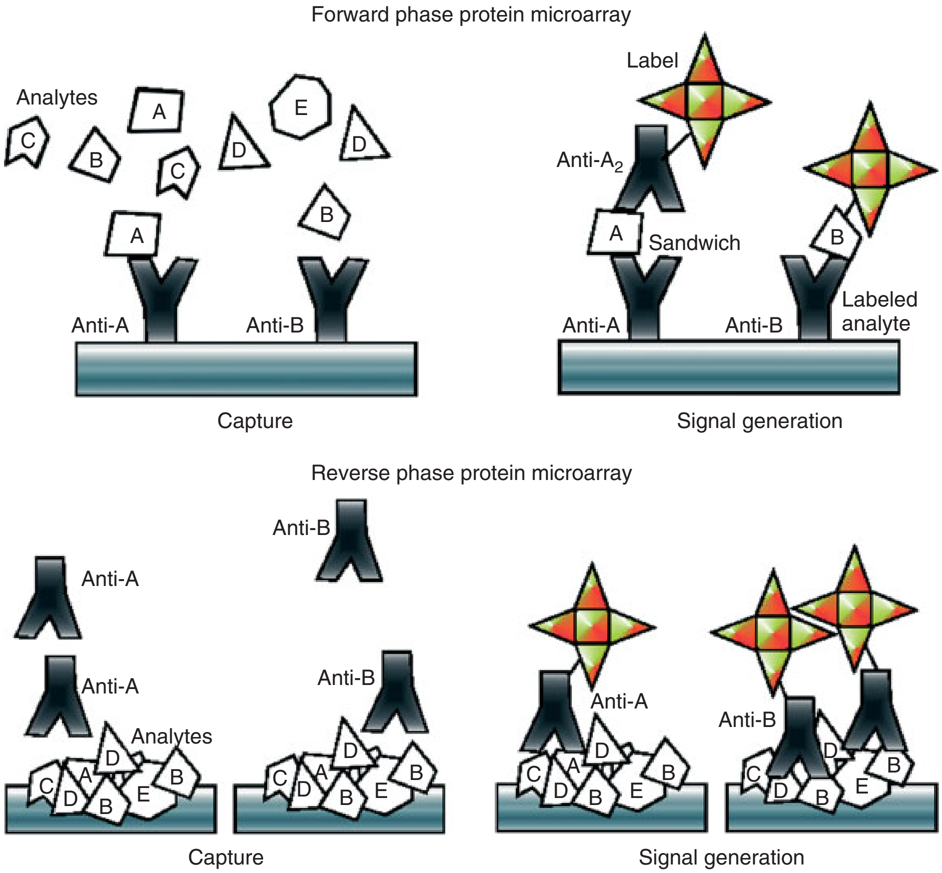


Figure 1 Classes of protein microarray platforms. Forward phase arrays (a) immobilize a bait molecule such as an antibody designed to capture specific analytes with a mixture of test sample proteins. The bound analytes are detected by a second sandwich antibody, or by labeling the analyte directly (a, right). Reverse phase arrays immobilize the test sample analytes on the solid phase. An analyte-specific ligand (e.g., antibody; b, left) is applied in solution phase. Bound antibodies are detected by secondary tagging and signal amplification (b, right)

relevant samples. Optimal substrata for protein microarrays must have high binding capacity, high surface area, and intrinsically low background signal (Seong, 2002). The choice of the substratum will dictate the immobilization chemistries employed. This, in turn, will impact on the native state, or on the appropriate orientation of the immobilized protein bait, or on the capture molecule.

4. Reverse phase arrays (RPA) for tissue analysis

We have developed RPAs as a format that appears to address many of the analytical challenges of detecting small levels of analytes (e.g., phosphorylated AKT) from very small amounts of input material (e.g., patient biopsy specimen) (Charboneau, 2002; Paweletz *et al*., 2001; Grubb *et al*., 2003; Wulfkuhle *et al*., 2003; Nishizuka

et al., 2003a,b; Liotta *et al*., 2003; Roberts *et al*., 2004; Zha *et al*., 2004) (Figure 2). This format has been successfully applied to analyze the state of the pro-survival, apoptosis and mitogenesis pathways within microdissected premalignant lesions, laser capture microdissected breast cancer, metastatic colorectal cancer, ovarian cancer, prostate cancer, follicular lymphoma, and lung epithelia exposed to different particulate toxicants (Charboneau, 2002; Paweletz *et al*., 2001; Grubb *et al*., 2003; Wulfkuhle *et al*., 2003; Nishizuka *et al*., 2003a,b; Liotta *et al*., 2003; Roberts *et al*., 2004; Zha *et al*., 2004). Each spot within an RPA contains an immobilized bait zone of solubilized and immobilized cellular lysate that measures only a few hundred microns in diameter. The high sensitivity of RPAs is due in part because the detection probe (e.g., antibody) can be tagged and the signal amplified, independent from the immobilized analyte protein. Third-generation amplification chemistries now available can be taken advantage of for highly sensitive detection

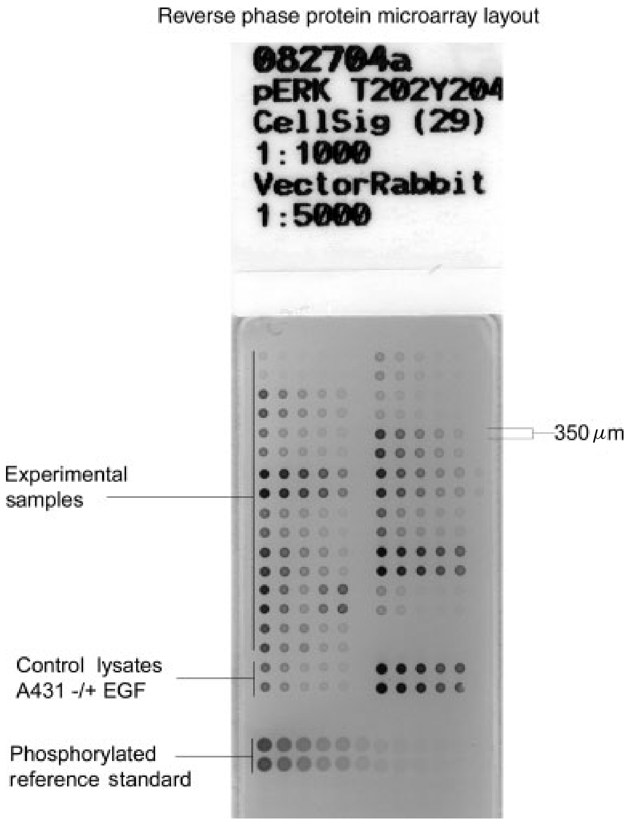


Figure 2 Idealized reverse phase array format. Duplicate samples are printed in dilution curves representing undiluted, 1:2, 1:4, 1:8, and 1:16 dilutions. The sixth spot represents a negative control, consisting of extraction buffer without sample. Each set of duplicate spots represents a patient sample before or after treatment, or microdissected normal, premalignant or stromal tissue cells. Positive and negative control reference lysates are printed on each array for monitoring assay performance. A reference standard, used as a conversion for experimental intensity values, is also printed on each array

(Hunyady *et al*., 1996; Bobrow *et al*., 1989; Bobrow *et al*., 1991; King *et al*., 1997). For example, coupling the detection antibody with highly sensitive tyramide-based avidin/biotin signal amplification systems can yield detection sensitivities down to fewer than 1000–5000 molecules/spot. A biopsy of 10 000 cells can yield 100 RPA arrays. Each array can be probed with a different antibody. Using commercially available automated equipment, RPAs exhibit excellent within-run and between-run analytical precision (3–10% c.v.) (Liotta *et al*., 2003). Moreover, RPAs do not require direct labeling of the sample analyte, which often destroys the epitope that is being detected, and do not utilize a two-site antibody sandwich, which often limits the repertoire of analytes that can be effectively measured. Therefore, there is no experimental variability introduced due to labeling yield, efficiency, or epitope masking. Even subtle differences in an analyte can be measured because each sample is exposed for the same amount of time to the same concentration of primary and secondary antibody and amplification reagents. RPA platforms can utilize reliable commercially available automated stainers designed for immunohistochemistry. Since each sample on the RPA format is applied in a "miniature dilution curve" on the array (Figure 2), in principle, a calibration curve is developed for each antibody, for each sample, and for each analyte concentration (Figure 3). This dynamic range analysis provides an improved means of matching the antibody concentration with the analyte concentration so that the linear range of each analyte measurement is ensured. The RPAs can be detected using a variety of protein array labels and amplification chemistries including fluorescent, radioactive, luminescent, and colorimetric read-outs (King *et al*., 1997; Kukar *et al*., 2002; Morozov *et al*., 2002; Schweitzer *et al*., 2002; Wiese, 2003).

5. Concluding remarks

RPAs may now provide a technological opportunity to elucidate and profile the ongoing protein phosphorylation levels in clinical specimens. The technical attributes of this type of array are amenable to small quantities of material, such as those obtained by laser capture microdissection. Using this format, many dozens of phospho-specific endpoints can be semiquantitatively measured using only a few thousand cells. In the future, we can envisage a time when cancer patient management is an exercise in patient-specific targeting of signaling networks using the information gleaned from these arrays. The information within the patient-specific profile could then be used to select an optimal suite of therapies best suited to the individual.

References

Blume-Jensen P and Hunter T (2001) Oncogenic kinase signalling. *Nature*, **411**, 355–365.

Bobrow MN, Harris TD, Shaughnessy KJ and Litt GJ (1989) Catalyzed reporter deposition, a novel method of signal amplification. Application to immunoassays. *Journal of Immunological Methods*, **125**, 279–285.

Bobrow MN, Shaughnessy KJ and Litt GJ (1991) Catalyzed reporter deposition, a novel method of signal amplification. II. Application to membrane immunoassays. *Journal of Immunological Methods*, **137**, 103–112.

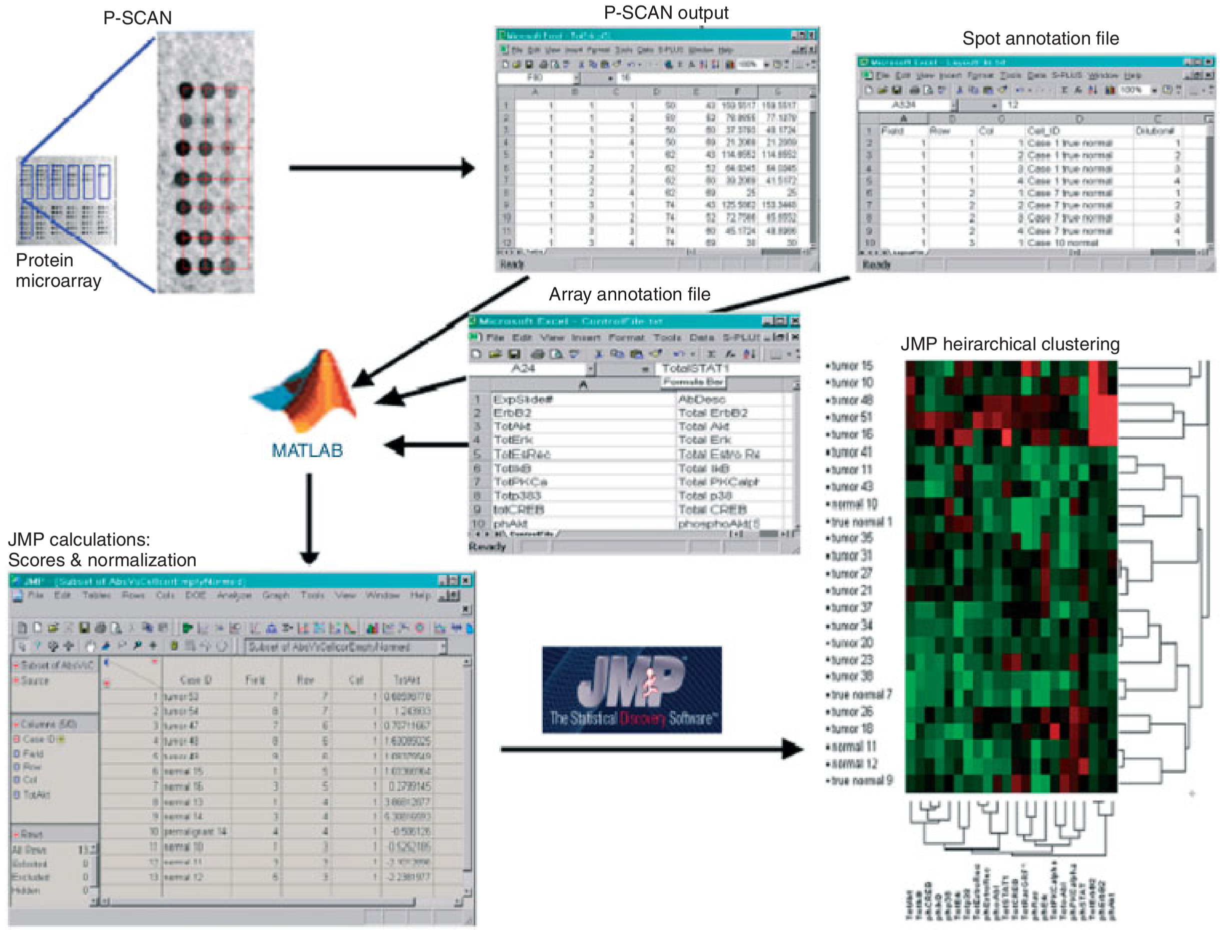


Figure 3 Phosphoproteomic molecular network mapping using reverse phase protein microarrays. Multiple phosphorylation kinase substrates are measured concomitantly and then subjected to bioinformatics analysis. Intensity values can be queried by unsupervised hierarchical clustering tools, which can reveal unpredicted linkages in both pathway phosphorylations as well as patient-specific linkages such as outcome and response to therapy

Bowden ET, Barth M, Thomas D, Glazer RI and Mueller SC (1999) An invasion-related complex of cortactin, paxillin and PKCmu associates with invadopodia at sites of extracellular matrix degradation. *Oncogene*, **18**, 4440–4449.

Celis JE and Gromov P (2003) Proteomics in translational cancer research: toward an integrated approach. *Cancer Cell*, **3**, 9–15.

Charboneau L (2002) Utility of reverse phase protein arrays: applications to signaling pathways and human body arrays. *Briefings in Functional Genomics and Proteomics*, **1**, 305–315.

Chen G, Gharib TG, Huang CC, Thomas DG, Shedden KA, Taylor JM, Kardia SL, Misek DE, Giordano TJ, Iannettoni MD, *et al*. (2002) Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. *Clinical Cancer Research*, **8**, 2298–2305.

Cutler P (2003) Protein arrays: the current state-of-the-art. *Proteomics*, **3**, 3–18.

Delehanty JB and Ligler FS (2002) A microarray immunoassay for simultaneous detection of proteins and bacteria. *Analytical Chemistry*, **74**, 5681–5687.

Grubb RL, Calvert VS, Paweletz CP, Phillips JL, Linehan WM, Gillespie JW, Emmert-Buck MR, Liotta L and Petricoin EF (2003) Signal pathway profiling of prostate cancer using reverse phase protein arrays. *Proteomics*, **3**(11), 2142–2146.

Huels C, Muellner S, Meyer HE and Cahill DJ (2002) The impact of protein biochips and microarrays on the drug development process. *Drug Discovery Today*, **7**, S119–S124.

Humphery-Smith I, Wischerhoff E and Hashimoto R (2002) Protein arrays for assessment of target selectivity. *Drug Discovery World*, **4**, 17–27.

Hunter T (2000) Signaling-2000 and beyond. *Cell*, **100**, 113–127.

Hunyady B, Krempels K, Harta G and Mezey E (1996) Immunohistochemical signal amplification by catalyzed reporter deposition and its application in double immunostaining. *The Journal of Histochemistry and Cytochemistry*, **44**, 1353–1362.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Jeong H, Tombor B, Albert R, Oltvai ZN and Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

King G, Payne S, Walker F and Murray GI (1997) A highly sensitive detection method for immunohistochemistry using biotinylated tyramine. *The Journal of Pathology*, **183**, 237–241.

Knezevic V, Leethanakul C, Bichsel VE, Worth JM, Prabhu VV, Gutkind JS, Liotta LA, Munson PJ, Petricoin EF III and Krizman DB (2001) Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics*, **1**, 1271–1278.

Kukar T, Eckenrode S, Gu Y, Lian W, Megginson M, She JX and Wu D (2002) Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins. *Analytical Biochemistry*, **306**, 50–54.

Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J and Petricoin EF (2003) Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell*, **3**(4), 317–325.

Liotta LA, Kohn EC and Petricoin EF (2001) Clinical proteomics: personalized molecular medicine. *The Journal of the American Medical Association*, **286**, 2211–2214.

MacBeath G (2002) Protein microarrays and proteomics. *Nature Genetics*, **32**(Suppl), 526–532.

MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Miller LD, Long PM, Wong L, Mukherjee S, McShane LM and Liu ET (2002) Optimal gene expression analysis by microarrays. *Cancer Cell*, **2**, 353–361.

Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Butler EB, Teh BS and Haab BB (2003) Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. *Proteomics*, **3**, 56–63.

Morozov VN, Gavryushkin AV and Deev AA (2002) Direct detection of isotopically labeled metabolites bound to a protein microarray using a charge-coupled device. *Journal of Biochemical and Biophysical Methods*, **51**, 57–67.

Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Munson PJ, *et al*. (2003a) Diagnostic markers that distinguish colon and

ovarian adenocarcinomas: Identification by genomic, proteomic, and tissue array profiling. *Cancer Research*, **63**(17), 5243–5250.

Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Munson PJ, *et al*. (2003b) Proteomic profiling of the NCI60 cancer cell lines using new high-density 'reverse-phase' lysate microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(24), 14229–14234.

Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, Petricoin IE and Liotta LA (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Petricoin EF, Zoon KC, Kohn EC, Barrett JC and Liotta LA (2002) Clinical proteomics: translating benchside promise into bedside reality. *Nature Reviews Drug Discovery*, **1**, 683–695.

Roberts E, Charboneau L, Espina V, Liotta L, Petricoin E and Dreher K (2004) Application of laser capture microdissection and protein microarray technologies in the molecular analysis of airway injury following pollution particle exposure. *Journal of Toxicology and Environmental Health Part A*, **67**(11), 851–861.

Seong SY (2002) Microimmunoassay using a protein chip: optimizing conditions for protein immobilization. *Clinical and Diagnostic Laboratory Immunology*, **9**, 927–930.

Schweitzer B, Roberts S, Grimwade B, Shao W, Wang M, Fu Q, Shu Q, Laroche I, Zhou Z, Tchernev VT, *et al*. (2002) Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nature Biotechnology*, **20**, 359–365.

Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF and Joos TO (2002) Protein microarray technology. *Trends in Biotechnology*, **20**, 160–166.

Wiese R (2003) Analysis of several fluorescent detector molecules for protein microarray use. *Luminescence*, **18**, 25–30.

Wilson DS and Nock S (2003) Recent developments in protein microarray technology. *Angewandte Chemie (International Ed. in English)*, **42**, 494–500.

Wulfkuhle JD, Aquino JA, Calvert VS, Fishman DA, Coukos G, Liotta LA and Petricoin EF (2003) Signal pathway profiling of ovarian cancer from human tissue specimens using reverse-phase microarrays. *Proteomics*, **3**(11), 2085–2090.

Zha H, Raffeld M, Charboneau L, Pittaluga S, Kwak LW, Petricoin E, Liotta LA and Jaffe ES (2004) Similarities of prosurvival signals in Bcl-2-positive and Bcl-2-negative follicular lymphomas identified by reverse phase protein microarray. *Journal of Laboratory Investigation*, **84**(2), 235–244.

Zhu H and Snyder M (2001) Protein arrays and microarrays. *Current Opinion in Chemical Biology*, **5**, 40–45.

Zhu H and Snyder M (2003) Protein chip technology. *Current Opinion in Chemical Biology*, **7**, 55–63.

Introductory Review

Protein interactions in cellular signaling

Tony Pawson
Samuel Lunenfeld Research Institute, Toronto, ON, Canada

1. Interaction domains

Protein–protein interactions are frequently mediated by specific domains, that are typically composed of 30–150 amino acids, and have a modular design such that their N- and C-termini are juxtaposed in space (Kuriyan and Cowburn, 1997; Pawson and Nash, 2000). As a consequence, an interaction domain can be readily incorporated into a host polypeptide in a fashion that leaves its ligand-binding surface available to engage a target protein. Interaction domains often recognize short peptide motifs, and the binding properties of a particular domain can therefore be defined by examining its binding sites on physiological partners and by its ability to select specific sequences from degenerate peptide libraries (Songyang *et al.*, 1993). Understanding this lexicon of protein–protein interactions makes it possible to predict the complement of interaction domains and recognition motifs present in a given protein, and thus its potential to associate with other polypeptides, based purely on its primary amino acid sequence (Yaffe *et al.*, 2001; *see also* Article 45, **Computational methods for the prediction of protein interaction partners**, Volume 5). Although protein interactions are frequently much more complex, since they are dependent on protein conformation, this simple concept provides a useful bridge between genomic and proteomic data and a basis for subsequent experimentation (*see* Article 46, **Functional classification of proteins based on protein interaction data**, Volume 5).

A critical feature of dynamic cellular responses is that protein interactions are carefully regulated. Molecular interactions control critical activities such as cell proliferation, differentiation, survival, and adhesion, all of which are essential for normal embryonic development and the maintenance of adult tissues. It is therefore important that the relevant complexes only assemble in response to appropriate signals. This regulation is commonly achieved in two ways. First, many interaction domains only bind with significant affinity to their targets once these have been modified by posttranslational modifications, such as phosphorylation (*see* Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6), and the interaction is terminated once the binding partner is dephosphorylated. In addition, interaction domains often undergo intramolecular interactions that block

their ability to bind an external ligand; these internal interactions must be broken before a productive multiprotein complex can be formed (Groemping *et al*., 2003).

Most regulatory proteins possess multiple domains, and therefore have the potential to bind several different ligands. Such proteins can be composed exclusively of interaction domains, and therefore act to nucleate the formation of larger complexes (Pawson and Scott, 1997). They can also possess enzymatic domains, which are regulated and targeted to their substrates by their associated interaction domains. In either case, proteins can be viewed as being organized in a cassette-like fashion from several distinct domains. An additional point is that the same interaction and catalytic domains are found repeatedly in many different proteins, and are often present in hundreds of copies in the human proteome. Thus, a significant means through which functional diversity has been achieved during the course of evolution appears to be through the arrangement of interaction and catalytic domains into novel combinations, with new biological properties.

2. SH2 domains and phosphotyrosine signaling

The Src homology 2 (SH2) domain is the prototypic interaction module (Sadowski *et al*., 1986), and illustrates several of these points. SH2 domains are a common feature of an otherwise diverse set of cytoplasmic proteins that mediate the biological effects of growth factor receptors with tyrosine kinase activity, through their ability to directly recognize phosphotyrosine (pTyr)-containing sites. The Src SH2 domain, for example, binds phosphopeptide motifs with the consensus sequence pTyr-Glu-Glu-Ile, which adopt an extended conformation when associated with the SH2 domain (Waksman *et al*., 1993). The pTyr residue occupies a deep, positively charged binding pocket that is conserved among all SH2 domains. About half of the binding energy for the SH2-phosphopeptide interaction comes from recognition of the phosphorylated tyrosine, and as a consequence, the SH2 domain has negligible affinity for unphosphorylated sites. The peptide residues C-terminal to the pTyr occupy a second binding surface on the Src SH2 domain, which includes a hydrophobic pocket for the side chain of the Ile in the +3 position relative to the pTyr. This latter binding surface is rather variable among different SH2 domains, and as a consequence, each SH2 domain binds preferentially to a distinct phosphopeptide motif, depending on the nature of the amino acids following the pTyr. This selectivity imparts an element of specificity in tyrosine kinase signaling, which can be enhanced through the cooperative actions of multiple interaction domains. For example, proteins with two tandem SH2 domains, such as phospholipase-Cγ or the ZAP-70 intracellular tyrosine kinase, bind with increased affinity and specificity to motifs containing two adjacent pTyr motifs, as compared to a single SH2 domain-phosphopeptide interaction (Ottinger *et al*., 1998).

The full-length c-Src protein-tyrosine kinase associates with the plasma membrane through an N-terminal myristate group, and possesses SH3, SH2, and kinase domains. SH3 domains also mediate protein–protein interactions, most commonly by binding proline-rich sequences that adopt a polyproline type II helix (Waksman *et al*., 1993). In its autoinhibited state, Src is phosphorylated within its C-terminal tail by another tyrosine kinase (Csk), inducing an intramolecular

interaction between the SH2 domain and the phosphorylated C-terminus. This in turn promotes the association of the SH3 domain with the linker region between the kinase and SH2 domains. As a consequence of these intramolecular contacts, not only is the kinase domain locked in an inactive configuration but the SH2 and SH3 domains are also prevented from engaging another protein. Once this autoinhibitory embrace is disrupted, for example, by dephosphorylation of the tail, the SH3 and SH2 domains are released, and can guide the freshly activated kinase to its cellular substrates. Consistent with the importance of these intramolecular interactions in controlling enzymatic activity, deletion of the C-terminal tail renders Src oncogenic, as seen for the v-Src transforming protein (Martin, 2001).

Upon activation by the relevant growth factor, or following oncogenic mutation, receptor tyrosine kinases (RTK) cluster and become autophosphorylated at residues that potentiate kinase activity, usually by inducing a conformational change within the kinase domain. RTKs are also autophosphorylated at sites that lie in noncatalytic regions of the receptor, for example, in the C-terminal tail or in the region between the plasma membrane and the kinase domain. These phosphorylated motifs, in turn, serve as docking sites for the SH2 domains of cytoplasmic signaling proteins; the nature of the amino acids flanking each pTyr residue dictate which SH2-containing proteins will bind preferentially to that site, and thus determine the signaling output of a given receptor (Heldin *et al*., 1998; Schlessinger, 2000). The recruitment of an SH2-containing protein to an RTK can stimulate its function, and thus the activity of cytoplasmic signaling pathways, in a variety of ways. These include (1) relocalization to the plasma membrane and consequent juxtaposition next to its substrates; (2) conformational change induced by SH2 domain binding to the receptor, resulting in increased enzymatic activity; or (3) enhanced tyrosine phosphorylation once bound to the receptor, which can lead either to association with other proteins or to altered enzymatic activity (or both). Of interest, recent work has indicated that a subset of SH2 domains, as found in the protein APS, bind selectively to the phosphorylated activation segment of the kinase domain, and thereby directly regulate receptor kinase activity (Hu *et al*., 2003).

Among the proteins that interact with RTKs are scaffolding proteins such as IRS-1, Shc, or FRS2. These bind receptors through pTyr-binding (PTB) domains that recognize Asn-Pro-X-pTyr motifs. PTB domains are therefore similar to SH2 domains in the sense that they bind specific motifs in a pTyr-dependent fashion but, interestingly, are quite different from SH2 domains both in their structure and in their mode of phosphopeptide recognition (Forman-Kay and Pawson, 1999). This makes the point that multiple different domains can evolve to interact with related sequences. Once bound to the autophosphorylated insulin receptor, IRS-1 is phosphorylated at numerous tyrosine sites, which subsequently recruit SH2 domain proteins such as phosphatidylinositol (PI) 3′-kinase and the adaptor Grb2 (White, 1998). Thus, signaling pathways frequently proceed through an ordered series of modular protein–protein interactions.

Proteins with SH2 domains have diverse functions, and therefore allow RTKs to couple to a wide range of signaling pathways. These functions include the control of phospholipid metabolism, regulation of Ras-like GTPases, cytoskeletal organization, protein phosphorylation, and gene expression, amongst others. One possible explanation for the widespread use of interaction modules such as SH2 domains

is to facilitate the evolution of new signaling pathways. For example, existing evidence suggests that tyrosine kinases evolved to mediate cellular communication, and were therefore important for the development of multicellular animals. The simultaneous appearance of SH2 domains may have allowed preexisting proteins to immediately bind pTyr through the acquisition of an SH2 domain, and thus to serve as cytoplasmic targets of RTKs.

3. Protein–protein interactions: a common mechanism for signal transduction

Variations of the scheme outlined above for RTKs are employed by a wide range of cell receptors. As two of many examples, activation of the Fas cell-surface receptor activates a protease cascade that induces apoptotic cell death. Once oligomerized by its ligand, Fas signals through a cytoplasmic interaction domain, termed a *death domain* (Aravind *et al*., 1999; Strasser *et al*., 2000). The Fas death domain forms a heterodimer with the death domain of an adaptor protein, Fadd, which in turn binds through a death-effector domain to Caspase 8, a protease which initiates the apoptotic pathway. Although different in detail, this pathway is reminiscent of an RTK, which recruits the SH2 domain of the Grb2 adaptor. Grb2 binds through its SH3 domains to Sos, a guanine nucleotide exchange factor that stimulates the Ras GTPase to adopt an active, GTP-bound conformation capable of recruiting targets such as the Raf protein serine/threonine kinase (Pawson *et al*., 2002). In both cases, the activated receptor binds a modular adaptor, which then stimulates a signaling pathway, leading to a cellular response. This paradigm is also exploited by signaling pathways that emanate from intracellular cues. For example, double-stranded breaks in DNA recruit proteins that induce DNA-repair and cell cycle arrest. This DNA damage response involves the phosphorylation of signaling proteins on Ser/Thr residues, which are then recognized by interaction domains such as FHA or BRCT (as found in the tumor suppressor protein BRCA1) (Clapperton *et al*., 2004; Li *et al*., 2002; Yaffe and Elia, 2001).

4. Summary

Protein interactions, especially those mediated by dedicated interaction domains, can regulate a variety of biochemical processes that are critical for dynamic cellular organization, as follows:

1. Interaction domains can control protein subcellular localization. This is particularly evident for domains that bind to specific phospholipids, such as PH domains that can recognize phosphoinositides (PI), including PI-3,4,5-P_3 and PI-4,5-P_2. These domains can therefore target signaling proteins to selected membranes enriched in the relevant phospholipid.
2. Interaction domains can control dynamic cellular responses through their recognition of posttranslational protein modifications, such as phosphorylation, methylation, acetylation, hydroxylation, or ubiquitination (Pawson and Nash, 2003).

3. Distinct domains can be joined to create adaptor proteins, which link otherwise distinct proteins into a common signaling pathway (*see* Article 42, **Membrane-anchored protein complexes**, Volume 5).
4. As an extension of this concept, modular protein interactions can be employed in combination to yield multiprotein signaling complexes capable of sophisticated regulation (*see* Article 41, **Investigating protein–protein interactions in multisubunit proteins: the case of eukaryotic RNA polymerases**, Volume 5 and Article 42, **Membrane-anchored protein complexes**, Volume 5). An example involves PAR-6, which contains both PDZ and PB1 interaction domains, and mediates diverse interactions with protein kinases, scaffolding proteins, and GTPases. PAR-6 is the central component of a complex that controls numerous aspects of cell polarity (Macara, 2004).
5. Interaction domains, once joined to catalytic domains, can regulate both their enzymatic activity and the ability to select specific substrates.
6. The combinatorial use of protein–protein interactions can yield extended networks of proteins that are responsible for cellular organization (Giot *et al*., 2003; *see also* Article 38, **The *C. elegans* interactome project**, Volume 5, Article 39, **The yeast interactome**, Volume 5, and Article 44, **Protein interaction databases**, Volume 5).
7. Aberrant protein–protein interactions can result in the deregulation of signaling pathways, and thus in diseases such as cancer.

Further reading

Sicheri F, Moarefi I and Kuriyan J (1997) Crystal structure of the Src family tyrosine kinase Hck. *Nature*, **385**, 602–609.

Tong AH, *et al*. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

References

Aravind L, Dixit VM and Koonin EV (1999) The domains of death: evolution of the apoptosis machinery. *Trends in Biochemical Sciences*, **24**, 47–53.

Clapperton JA, Manke IA, Lowery DM, Ho T, Haire LF, Yaffe MB and Smerdon SJ (2004) Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. *Nature Structural & Molecular Biology*, **11**, 512–518.

Forman-Kay JD and Pawson T (1999) Diversity in protein recognition by PTB domains. *Current Opinion in Structural Biology*, **9**, 690–695.

Giot L, *et al*. (2003) A protein interaction Map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Groemping Y, Lapouge K, Smerdon SJ and Rittinger K (2003) Molecular basis of phosphorylation-induced activation of the NADPH oxidase. *Cell*, **113**, 343–355.

Heldin C-H, Ostman A and Ronnstrand L (1998) Signal transduction via platelet-derived growth factor receptors. *Biochimica et Biophysia Acta*, **1378**, F79–F113.

Hu J, Liu J, Ghirlando R, Saltiel AR and Hubbard SR (2003) Structural basis for recruitment of the adaptor protein APS to the activated insulin receptor. *Molecular Cell*, **12**, 1379–1389.

Kuriyan J and Cowburn D (1997) Modular peptide recognition domains in eukaryotic signaling. *Annual Review of Biophysics and Biomolecular Structure*, **26**, 259–288.

Li J, Williams BL, Haire LF, Goldberg M, Wilker E, Durocher D, Yaffe MB, Jackson SP and Smerdon SJ (2002) Structural and functional versatility of the FHA domain in DNA damage signaling b the tumor suppressor kinase Chk2. *Molecular Cell*, **9**, 1045–1054.

Macara IG (2004) Parsing the polarity code. *Nature Reviews. Molecular Cell Biology*, **5**, 220–231.

Martin GS (2001) The hunting of the Src. *Nature Reviews. Molecular Cell Biology*, **2**, 467–475.

Ottinger EA, Botfield MC and Shoelson SE (1998) Tandem SH2 domains confer high specificity in tyrosine kinase signaling. *The Journal of Biological Chemistry*, **273**, 729–735.

Pawson T and Nash P (2000) Protein–protein interactions define specificity in signal transduction. *Genes & Development*, **14**, 1027–1047.

Pawson T and Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Pawson T, Raina M and Nash P (2002) Interaction domains: from simple binding events to complex cellular behaviour. *FEBS Letters*, **513**, 2–10.

Pawson T and Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075–2080.

Sadowski I, Stone JC and Pawson T (1986) A non-catalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus $P130^{gag-fps}$. *Molecular and Cellular Biology*, **6**, 4396–4408.

Schlessinger J (2000) Cell signaling by receptor tyrosine kinases. *Cell*, **103**, 211–225.

Songyang Z, Shoelson SE, Chadhuri M, Gish G, Pawson T, King F, Roberts T, Ratnofsky S, Schaffhausen B and Cantley LC (1993) Identification of phosphotyrosine peptide motifs which bind to SH2 domains. *Cell*, **72**, 767–778.

Strasser A, O'Connor L and Dixit VM (2000) Apoptosis signaling. *Annual Review of Biochemistry*, **69**, 217–245.

Waksman G, Shoelson S, Pant N, Cowburn D and Kuriyan J (1993) Binding of a high affinity phosphotyrosyl peptide in the *src* SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell*, **72**, 779–790.

White MF (1998) The IRS-signaling system: a network of docking proteins that mediate insulin and cytokine action. *Recent Progress in Hormone Research*, **53**, 119–138.

Yaffe MB and Elia AE (2001) Phosphoserine/threonine-binding domains. *Current Opinion in Cell Biology*, **13**, 131–138.

Yaffe MB, Laparc GG, Lai J, Obata T, Volinia S and Cantley LC (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnology*, **19**, 348–353.

Introductory Review

Structural biology of protein complexes

Gabriel Waksman
Birkbeck and University College, London, UK

The formidable advances in protein sciences in recent years have highlighted the importance of protein–protein interactions in biology. Before the proteomics revolution, we knew that proteins were capable of interacting with each other and that protein function was regulated by interacting partners. However, the extent and degree of the protein–protein interaction network was not realized. It is now believed that not only a majority of proteins in an eukaryotic cell is involved in complex formation at some point in the life of the cell but also that each protein may have on an average 6–8 interacting partners (Tong *et al*., 2004).

The study of the structure of protein–protein complexes predates the proteomics revolution. The pioneering work on antigen–antibody and protease–protease inhibitor complexes has provided insight into protein–protein interfaces and their properties (Ruhlmann *et al*., 1972; Amit *et al*., 1986). However, more recently, the structure of larger complexes that function as molecular machines has been determined, shedding light into important cellular functions such as transcription (the structure of the RNA polymerase II core complex (Cramer *et al*., 2001)), translation (the structure of the ribosome (Ban *et al*., 2000; Wimberly *et al*., 2000)), replication (the structure of the γ-complex in bacteria (Jeruzalmi *et al*., 2001)), or the cytoskeleton (the structure of the Arp2/3 complex (Robinson *et al*., 2001)), to cite only a few. It is not the purpose of this short specialist review to make an exhaustive list of all protein–protein complexes, the structure of which has been determined to date, but instead to provide general concepts on the common and distinctive features of protein–protein interfaces and their roles.

Protein–protein interactions can be classified into approximately three subtypes, depending on their stability and the mode of interactions. Although any one protein may be involved in interactions with many others, they may form stable interactions only with a few. These form core-complexes, which are stable, can be purified, and are amenable to structural studies. A number of core-complex structures have been determined, and these structures have been instrumental in understanding how these core-complexes carry out their functions (see list above). A second category of interactions are transient, and the proteins involved in transient associations are often regulatory proteins, the role of which may be to confer short-lived, physiologically regulated, properties to other proteins or to core-complexes. The complexes that proteins form transiently

may be unstable and difficult to purify. Successes in determining the structure of transient complexes have been dependent on the affinity of the various constituents participating in complex formation. One historical breakthrough in this regard has been the determination of the structure of the first antigen–antibody complex (Amit *et al*., 1986), which defined the architecture and chemistry of protein–protein interactions in this versatile structural framework. Finally, protein–protein interactions in core-complexes or transient ones may be mediated by specialized, small, domains dedicated to protein–protein recognition (Pawson and Scott, 1997). In this review, interactions mediated by specialized domains are treated separately as their functions and roles have been studied in great details.

Protein–protein interfaces have been analyzed thoroughly in studies that have emphasized their high degree of versatility and plasticity (Jones and Thornton, 1996; Lo Conte *et al*., 1999). This is, in part, due to the fact that protein–protein interactions encompass a wide range of affinities. However, even within a particular range of affinities, the structural features underpinning binding may vary. For example, protease inhibitors appear to use main chain–main chain interactions, whereas antigen–antibody interaction is mediated by side chain–side chain interactions (Jackson, 1999). Side chain–side chain interactions may be more likely to determine specificity. In contrast, serine proteases must bind tightly to their target proteases. This may be best achieved using constrained “main chain–main chain” conformation, and means that the inhibitor will be highly committed to the enzyme. Similar observations corresponding to similar requirements have been observed in the interaction of pilus subunit with bacterial chaperones (Choudhury *et al*., 1999; Sauer *et al*., 1999). Proteins may be able to use the same template for interactions with different proteins or with different parts of the same protein. For example, in the growth hormone/growth hormone receptor complex (GH/GHR), two receptor molecules bind to different parts of the same ligand (de Vos *et al*., 1992). This complex structure and the subsequent site-directed mutagenesis studies have defined an important concept in protein–protein interaction, that of “hot spot” (Clackson and Wells, 1995). In the protein–protein interface observed in the GH/GHR complex, although a myriad of hydrogen bonds, van der Waals contacts, and electrostatic interactions is observed, only a limited few of these interactions have been shown to play an important role in binding. Similar observations have been made in other systems (Dall’Acqua *et al*., 1996; Bradshaw *et al*., 1999). One important property of protein–protein interfaces is the shape complementarity between the two regions coming together in the interactions. Remarkably, such complementarity is very often mediated by water molecules, judiciously placed at the interface to fill in holes and increase contacts (Bhat *et al*., 1994; Lubman and Waksman, 2003). The role of water in both the structural and thermodynamic basis of protein–protein interactions is essential and yet very poorly understood.

One remarkable feature of protein–protein interactions is that they are often mediated by small domains that specifically bind small sequence motifs on proteins. The Src-Homology 2 (SH2) domain was the first such domain to be recognized (Sadowski *et al*., 1986). SH2 domains are involved in the building up of large complexes at and around signaling receptors. SH2 domains bind specifically

sequences containing phosphorylated tyrosines and are able to discriminate between tyrosine-phosphorylated sites by exercising some preferences for residues located C-terminally relative to the phosphotyrosine (Bradshaw and Waksman, 2002). Since the discovery of SH2 domains, a large number of protein domains with specialized roles in protein–protein interactions have been found (see **http://www.mshri.on.ca/pawson/domains.html** for an exhaustive list of such domains). SH3 and WW domains specifically recognize and bind sequence motifs containing prolines (Musacchio, 2002). Proline-rich motifs are among the most common motifs identified, and thus SH3 and WW domains play major roles in protein–protein interactions. PDZ domains are essential for the integrity of the postsynaptic density, a large protein complex formed around glutamate and NMDA receptors in the nervous system (Sattler and Tymianski, 2001). Finally, bromodomains are similar to SH2 domains in that both bind (and thus induce recruitment of proteins that contain them) to sites of protein modifications. Bromodomains, unlike SH2 domains, which bind to tyrosine-phosphorylated sites, bind specifically acetylated lysines, and thus play important roles in chromatin remodeling during transcription and replication (Dhalluin *et al*., 1999).

There is no unifying theme among the structures of the protein–protein interaction domains listed above. However, as their structures have been characterized, intense efforts have been devoted to designing specific binding inhibitors capable of disrupting protein–protein interaction mediated by these modules. For example, the SH2 domain of the Src kinase has been targeted for molecular design and binding competitors able to inhibit osteoclast function have been found (Sawyer *et al*., 2002). As the major phenotype in Src knockout mice is a thickening of the bones, it is hoped that an Src SH2 domain–binding inhibitor could be used to combat osteoporosis, a devastating disorder in elderly women.

Using peptides or peptide mimics to disrupt protein–protein interfaces is not a novel idea but this approach has benefited from structural information. Notably, the molecular design of rigid peptido-mimetics is believed to greatly enhance the potential of peptides as therapeutics by not only locking the peptide in a defined binding conformation but also by preventing or slowing degradation (Patani and LaVoie, 1996). However, peptides or peptide-based compounds do not readily cross the cell barriers, and thus are not as effective as hoped. Recently, a peptide derived from the NR2 chain of the NMDA receptor known to interact with the second PDZ domain of PSD95, an essential component of the postsynaptic density, was made effective in reducing cerebral infarction in rats subjected to transient focal cerebral ischemia by fusing it to the HIV1-Tat translocator peptide (Aarts *et al*., 2002). Thus, the use of translocator peptides may be a promising avenue of research for the delivery of therapeutic peptides or proteins (Becker-Hapak *et al*., 2001).

In the next few years, we will see more of the core-complexes being tackled for structural studies. The methods to overexpress and purify them in large quantities are in place and it may just be a matter of time before their structures are determined. It will be more difficult to tackle the transient complexes involving proteins exhibiting only low affinity for each other. These may require the development of novel cross-linking or tethering methods, capable of capturing the complex as it forms.

References

Aarts M, Liu L, Besshoh S, Arundine M, Gurd JW, Want YT, Salter MW and Tymianski M (2002) Treatment of ischemic brain damage by perturbing NMDA receptor-PSD-95 protein interactions. *Science*, **298**, 846–850.

Amit AG, Mariuzza RA, Phillips SE and Poljak RJ (1986) Three-dimensional structure of an antigen-antibody complex at 2.8 A resolution. *Science*, **233** (4765), 747–753.

Ban N, Nissen P, Hansen J, Moore PB and Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289** (5481), 905–920.

Becker-Hapak M, McAllister SS and Dowdy SF (2001) TAT-mediated protein transduction into mammalian cells. *Methods (San Diego, Calif.)*, **24** (3), 247–256.

Bhat TN, Bentley GA, Boulot G, Greene MI, Tello D, Dall'Acqua W, Souchon H, Schwarz FP, Mariuzza RA and Poljak RJ (1994) Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 1089–1093.

Bradshaw JM, Mitaxov V and Waksman G (1999) Investigation of Phosphotyrosine Recognition by the SH2 Domain of the Src Kinase. *Journal of Molecular Biology*, **293** (4), 971–985.

Bradshaw JM and Waksman G (2002) Molecular recognition by SH2 domains. *Advances in protein chemistry*, **61**, 161–210.

Choudhury D, Thompson A, Stojanoff V, Langermann S, Pinkner J, Hultgren SJ and Knight SD (1999) X-ray structure of the FimC-FimH chaperone-adhesin complex from uropathogenic Escherichia coli [see comments]. *Science*, **285** (5430), 1061–1066.

Clackson T and Wells JA (1995) A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science*, **267**, 383–386.

Cramer P, Bushnell DA and Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, **292** (5523), 1863–1876. Epub 2001 Apr 19.

Dall'Acqua W, Goldman ER, Eisenstein E and Mariuzza RA (1996) A Mutational Analysis of the Binding of Two Different Proteins to the Same Antibody. *Biochemistry*, **35**, 9667–9676.

de Vos AM, Ultsch M and Kossiakoff AA (1992) Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, **255** (5042), 306–312.

Dhalluin C, Carlson JE, Zeng L, He C, Aggarwal AK and Zhou M-M (1999) Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, **399**, 491–496.

Jackson RM (1999) Comparison of protein-protein interactions in serine-protease and antibody-antigen complexes: implications for the protein docking problem. *Protein Science*, **8**, 603–613.

Jeruzalmi D, O'Donnell M and Kuriyan J (2001) Crystal structure of the processivity clamp loader gamma (gamma) complex of E. coli DNA polymerase III. *Cell*, **106** (4), 429–441.

Jones S and Thornton JM (1996) Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **93** (1), 13–20.

Lo Conte L, Chothia C and Janin J (1999) The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, **285** (5), 2177–2198.

Lubman OY and Waksman G (2003) Structural and thermodynamic basis for the interaction of the Src SH2 domain with the activated form of the PDGF beta-receptor. *Journal of Molecular Biology*, **328** (3), 655–668.

Musacchio A (2002) How SH3 domains recognize proline. *Advances in Protein Chemistry*, **61**, 211–268.

Patani GA and La Voie EJ (1996) Bioisosterism: A Rational Approach in Drug Design. *Chemical Reviews*, **96** (8), 3147–3176.

Pawson T and Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075–2080.

Robinson RC, Turbedsky K, Kaiser DA, Marchand JB, Higgs HN, Choe S and Pollard TD (2001) Crystal structure of Arp2/3 complex. *Science*, **294** (5547), 1679–1684.

Ruhlmann A, Schramm HJ, Kukla D and Huber R (1972) Pancreatic trypsin inhibitor (Kunitz). II. Complexes with proteinases. *Cold Spring Harbor Symposia on Quantitative Biology*, **36**, 148–150.

Sadowski I, Stone JC and Pawson T (1986) A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of fujinami sarcoma virus p130$^{gag-fps}$. *Molecular and Cellular Biology*, **6** (12), 4396–4408.

Sattler R and Tymianski M (2001) Molecular mechanisms of glutamate receptor-mediated excitotoxic neuronal cell death. *Molecular Neurobiology*, **24** (1-3), 107–129.

Sauer FG, Fütterer K, Pinkner JS, Dobson KW, Hultgren SJ and Waksman G (1999) Structural Basis of Chaperone Function and Pilus Biogenesis. *Science*, **285**, 1058–1061.

Sawyer T, Bohacek RS, Dalgarno D, Eyermann CJ, Kawahata N, Metcalf III CA, Shakespeare W, Sundaramoorthi R, Wang Y and Yang MG (2002) Src Homology 2 inhibitors: peptidomimetic and nonpeptide. *Mini Reviews in Medicinal Chemistry*, **2**, 475–488.

Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Wimberly BT, Brodersen DE, Clemons WM, Jr., Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T and Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. *Nature*, **407** (6802), 327–339.

Specialist Review

Biochemistry of protein complexes

Bertrand Séraphin
Centre for Molecular Genetics, Gif sur Yvette, France

1. Introduction

Proteins are quantitatively more abundant than RNA and DNA in average bacterial or eukaryotic cells. However, this abundance is contrasted by the extreme variation in the relative abundance of individual proteins: some being represented by several million of molecules in a given cell (such as hemoglobin in erythrocytes), while for others only a few molecules may be present at a given time. (This can go down to even less than one protein per cell when averaged on time.) This skewed distribution diversity is obviously not encountered for genomic DNA but is already partly apparent at the mRNA level.

Chemically, proteins are also more diverse than nucleic acids (*see* Article 96, **Fundamentals of protein structure and function**, Volume 6). This results from their assembly from over 20 amino acids that have significantly more diverse chemical and physical properties than the four bases present in DNA or RNA. Amino acids diversity can be further expended by numerous reversible or permanent posttranslational modifications. These amino acid properties direct the folding of proteins in a wide variety of tri-dimensional structures that contrast with the monotonous helical structure of B-type double stranded DNA. (Again, at this level, RNA appears to have intermediate properties between DNA and proteins by its number of possible natural structures.) The shape taken by proteins upon folding allow them to present an extreme diversity of surfaces that have been exploited by biological organisms to govern recognition and interaction of proteins with virtually any other type of chemicals: from water to small enzymatic substrates, from membrane to antigens, from DNA to proteins themselves. It is thus not surprising to observe that most proteins in a given cell do not exist as monomers but associate rather with other proteins to form large assemblies. Understanding cell function requires deciphering the network created through protein–protein interactions. This task is, however, difficult because, as indicated above, some proteins are poorly abundant while other polypeptides are embedded into biological membranes and thus difficult to handle. This is further complicated by the fact that some interactions are transient. Thus, it is not surprising that many techniques have been developed to analyze protein interactions. Some rely on indirect observation of the consequences

of protein interaction (e.g., two-hybrid assays (Fields and Song, 1989), fluorescence resonance energy transfer (FRET) (Wouters *et al*., 2001; *see also* Article 43, **Energy transfer–based approaches to study G protein–coupled receptor dimerization and activation**, Volume 5 etc.), while others detect directly the interaction between the protein factors (e.g., GST pull downs (Kaelin *et al*., 1991), coimmunoprecipitation (Lane and Crawford, 1979), or biochemical purifications (Deutscher, 1990)). In this review, I will focus on the biochemical methods that have been developed for the analysis of protein complexes.

2. A variety of complexes

The term "complex" needs to be defined as it encompasses many different meanings. I will refer here to complexes as a group of associated proteins that can be characterized biochemically. This definition has several implications: while a complex contains several polypeptides, they may be identical (homomeric complex) or they may represent at least two different sequences (heteromeric complex). Furthermore, the operational definition of complexes as entities that can be purified biochemically does not make any assumption about the stability, size, or biological function of the complex. Thus, a complex displaying an extreme stability in vitro may represent a transient intermediate in a biochemical pathway in vivo. In contrast, a very important biological complex may be too unstable to withstand conditions encountered during biochemical purification. In the same vein, a complex obtained following extensive purification may only represent a small unit resulting from the breakdown of a large molecular machine that carries the real biological activity in vivo. Further complication results from the heterogeneity in protein levels and affinity: the mass-action law indicates, indeed, that it may be possible to recover significant amount of a low-affinity complex formed by highly abundant proteins, while an intrinsically more stable assembly formed by extremely low abundance factors may dissociate under the same conditions preventing its purification. It is also clear that additional specific assays are required to test the function of protein complexes. However, while the practical definition of complexes as biochemically purifiable units suggests that the corresponding polypeptides interact in vivo, this needs to be ascertained and the physiological relevance of the interaction needs to be demonstrated as well. Indeed, it is not uncommon to find (highly abundant) proteins associated nonspecifically with a protein of interest. These "contaminants" may result from artifactual interactions induced by the experimentalist during protein extraction and purification (e.g., aggregation promoted by partial denaturation) to natural interactions that are not related to the process of interest (e.g., interaction of a protein with a chaperone may be relevant to complex assembly but not to the biological function of the protein itself). In this context, it is worth remembering that many protein interactions are transient and dynamic. Protein extraction and purification will block this dynamic in a certain number of states, resulting in the accumulation of proteins in a specific set of complexes. This distribution may differ significantly from the natural distribution of the corresponding factor. An additional bias may be introduced during complex purification: while two entities may be present in equimolecular amount, one may be efficiently recovered while the

other will be recovered in low amount. This bias, which often results from intrinsic chemical and/or physical property of the different entities, may further distort the relation between the observed complexes and the "real" biological situation present in vivo. Despite these limitations, biochemical analysis remains an essential tool for protein complexes analysis. Indeed, this strategy allows for a direct observation of protein complexes and thus determination of a wide variety of parameters that cannot be determined in many other types of experiments. For example, biochemical approaches permit the determination of complex stoichiometry, reveal the presence of posttranslational modification, isoenzyme or splice variant, allow the possibility to perform activity tests and/or determine affinity constants, and so on. Biochemical purification of a complex may even lead to the determination of the complex structure. Biochemistry of complexes benefits from the high synergy involved in complex formation in vivo. Many experiments have indeed revealed that pairwise interactions between subunits of a complex are often weak but that the whole assembly is stabilized through the synergistic network of associations occurring when all subunits are present simultaneously. This feature, that may be essential for the correct assembly of protein complexes in vivo, may be responsible for a high number of false-negative results in studies of protein interaction relying on simple binary tests (such as FRET studies and standard two-hybrid assays). However, as stated above, the mere finding of several proteins associated together is far from proving that this complex exists in vivo and additional analyses are, in each case, required to demonstrate the validity and functional significance of the proposed interactions.

3. Defining a target and a protein source

At least, two entry points are required for the biochemical analysis of protein complexes. First, the knowledge of at least one characteristic "target" of the complex to be analyzed is needed. In the early days of biochemistry, this "target" was often a biochemical activity. With the development of molecular biology and the availability of numerous genomic and cDNA sequences, this "target" can also be a predicted protein sequence. In other cases, antibodies directed against one or several polypeptides have been used. The characteristic "target" will be essential for the development of an appropriate purification strategy and may be the basis for specific activity assay. Beside the essential knowledge that defines the "target", one requires a source of proteins. The two major possibilities are the original host or recombinant proteins. Natural sources present the obvious advantage of being biologically relevant. Thus, protein complexes recovered from the original host will certainly provide significant information. Furthermore, no assumption needs to be made on the nature and the number of subunits present in the target complex when it is isolated from the natural host. The main problems encountered in analyzing protein complexes recovered from such a source are the purity and the limiting quantity of the starting biological material. If it is indeed easy to get large amounts of a homogeneous population of a cultivable unicellular organism, it may be difficult to recover sufficient material if the target protein is expressed in a few cells from a complex multicellular plant or animal. A similar problem is encountered if the proteins of interest are present in a single-cell organism

that cannot be grown in the laboratory and/or that is only present in a complex community of species. One can sometimes bypass the problem encountered with a multicellular organism as homogenous cell lines may provide a suitable source for biochemical purification of specific complexes. In the latter case, however, the results may represent the bona fide biological situation of the cell line but may only have a partial significance for the whole organism. Indeed, in the cell line, the complex structure and/or function may be altered in the absence of cues emanating from neighboring cells and/or hormonal signals originating from distant tissues. Additional complications related to the quantity and homogeneity of the starting biological material may be encountered if the complex is only transiently expressed or assembled. For example, it is easily understandable that difficulties will arise if low levels of the target complex are only transiently expressed during a very short time window of the cell cycle even if large amounts of nonsynchronized cells are easily produced. A further level of complication may arise if the protein is simultaneously present in different complexes in the starting biological material and/or if variants (e.g., as a result of alternative splicing) of the target factor do coexist. Most of the problems linked to the quantity and homogeneity of the starting biological material encountered with natural source of proteins can be bypassed if the complex is produced in a recombinant form. This strategy requires, however, the previous knowledge of the complex composition, the availability of clones containing the corresponding coding sequences, and the possibility to express efficiently the various subunits. In addition, effective complex assembly should occur in vivo if the subunits are coexpressed or in vitro if individual proteins are produced. Beside the requirement for the knowledge of the protein complex composition, additional difficulties encountered with the production of recombinant complexes often result from non- or low-expression of one or more subunit, lack of posttranslational modification(s), and absence of suitable condition for complex reassembly. If successful, this approach provides, however, large quantities of homogenous material perfectly suited for studies such as structure determination. In addition, the recombinant factors can easily be modified by standard genetic engineering methods and the consequences of these modifications on complex structure and/or function can easily been assayed.

4. Protein complex purification

Once a source of proteins and the target have been defined, a strategy for the purification of the complex can be developed. The strategy will have to be compatible with both the target definition and the protein source. When protein complexes need to be recovered from natural host cells, classical biochemical purification (Deutscher, 1990) appears to be the method of choice when limited information is available on the target, for example, when only its enzymatic activity is known. Biochemical characterization of the complex is then often a tedious process in which various methods such as differential centrifugation, ion exchange chromatography, size exclusion chromatography, or differential precipitation are tested for their ability to enrich fractions of the starting extract for the desired activity while reducing protein complexity (Deutscher, 1990). Combining several

enrichment steps results in the final biochemical purification of the target complex. This strategy has been used to a large extent in the early days of biochemistry and still remains a method of choice. Abundant examples can be found in textbooks with some well-known cases such as the *E. coli* DNA polymerase (Lehman *et al*., 1958; Richardson *et al*., 1964) or the yeast nuclear pore (Rout *et al*., 2000; *see also* Article 41, **Investigating protein–protein interactions in multisubunit proteins: the case of eukaryotic RNA polymerases**, Volume 5).

In theory, any complex should be analyzable following this strategy as long as resources in starting material, time, and money are not limiting. However, this approach suffers from several problems, making its widespread use difficult. The main problem resides in the definition of the purification steps required to recover the complex of interest and their relatively low specificity in the enrichment for the desired target. Selection of the purification steps is often performed by a trial-and-error process, which is tedious and time consuming. Furthermore, the purification efficiency of each step of the procedure is often poor. Thus, many successive steps have often to be combined to result in the recovery of a pure complex. Additional problems may be encountered if the complex is labile. Another complication may arise if the assay used to follow complex purification indicates the presence of the final product of a cascade of reactions. Indeed, the various enzymes will most likely fractionate apart during purification, disrupting the assay even if all activities are intact (e.g., Kramer and Utans, 1991). Two alternatives are then possible: either one has to combine fractions to restore the activity defined in the original assay or one has to define specific assays for each of the partial reactions. It is easy to imagine that inextricable situations may arise if a large number of fractions are required to reconstitute an activity and if, simultaneously, no partial assays are available. Despite these potential pitfalls, biochemical purification is the method of choice to purify a complex when limited information is available on its composition.

In another extreme situation, one may have little information on the activity of a complex but some specific knowledge on its composition. This can be the sequence of one of the protein subunit or an antibody directed against one of its component. It is then tempting to use these data to design an efficient strategy to recover the target complex. If an antibody is available, it can be used in an assay for protein complex purification through the probing of fraction for the presence of the polypeptide of interest (e.g., by western blotting or ELISA) (see Miller *et al*., 2001 for example). However, this reagent can often be used in a more efficient manner as a direct tool for protein purification. In the latter case, antibodies serve as highly specific affinity reagents to fish the target complex out of a crude extract (e.g., Gygi *et al*., 1999). If this immunoaffinity purification strategy is chosen, one has to make sure that relatively large amounts of highly specific antibodies are available. The main advantage of an immunoaffinity procedure is the extremely high and specific enrichment for the target complex that can be achieved in a single step. If the complex is relatively abundant and the antibodies of high quality (e.g., monoclonals), one can even expect in the best cases to visualize the various subunits of the complex on a protein gel after a single purification step. If the target complex is of low abundance, immunoaffinity purification has then to be combined with classical biochemical purification steps. The drawbacks of the method are of various orders. Obviously, the requirement for substantial amount of high quality

antibodies is often limiting. However, the high affinity of antibodies, which make them so efficient for purification, may also become problematic. Indeed, it is often difficult to elute intact complexes from an immunoaffinity column because the harsh conditions required to split the antibody-antigen interaction often induce the simultaneous breaking up of the target complex. This can be avoided, however, when antibodies have been raised against a small peptide that can itself be used as a gentle elution reagent. Nevertheless, if immunoaffinity has to be combined with additional steps during complex purification, it often has to be used last, given the likelihood that the complex will dissociate during elution. Furthermore, one has to consider that the breaking up of the complex after immunoaffinity chromatography may prevent some types of downstream analyses (e.g., activity assays). Obviously, if antibodies are not available but other proteins or chemicals specifically interacting with a subunit of the target complex are known (such as a protein partner produced in large amount as a recombinant factor or an inactive substrate mimic), the latter ones can be used instead for affinity purification.

The efficiency and specificity of affinity purification can also be harnessed if the sequence of a subunit of the target protein complex is known. For this purpose, the sequence coding for a protein domain with defined affinity properties (often called a *tag*) is grafted to the sequence coding for polypeptide of interest through genetic engineering. The corresponding fusion polypeptide, endowed with the binding properties of this new domain, can then be purified efficiently and specifically on the appropriate affinity media (Terpe, 2003). This strategy, which has been extensively used for the expression of single polypeptides, may also be used for protein complexes as long as the other subunits are coexpressed with the tagged one. Thus, if the various subunits of a complex are known, they may all be expressed as recombinant forms (Tan, 2001). However, if only a single subunit of the target complex is known, one has to express the corresponding tagged version in the natural host to allow for complex assembly. The requirement for a host allowing expression of an affinity-tagged protein may thus limit the application of this strategy. The main advantages of affinity tags over antibodies are their constant characteristics. Many tags have thus been developed and one can predict with relatively good accuracy the binding characteristics of the resulting fusion proteins for the corresponding affinity media (Terpe, 2003). Also, gentle elution conditions have been developed for many of them, allowing the recovery of folded proteins or intact complexes rather than denatured partners. Another advantage of tags is the easiness with which one can fuse them to the target protein through standard cloning techniques. A main drawback, however, is the possibility that the tag will interfere with the function and/or the interaction pattern of the fused polypeptide. Despite this problem, tags have proven to be widely useful in the analysis of protein complexes (e.g., Rigaut *et al*., 1999). As for antibodies, tags were found to be sufficient for the recovery of moderately abundant protein complexes in single affinity-purification steps. For low-abundance complexes, strategies using two different affinity tags in a row have been developed (Rigaut *et al*., 1999). In these cases, the two affinity tags could be fused to the same bait protein. Alternatively, two independent baits belonging to the same complex may be each fused to one of the affinity tag (Puig *et al*., 2001; Caspary *et al*., 1999). The latter strategy ensures that the purified material does not contain mostly the free form of an abundant

subunit of a complex of interest. Overall, affinity purification with tags appears as a method of choice for complex purification if the sequence of one subunit of the complex is known and if an appropriate expression system is available.

If all the subunits of a complex are known, the possibility to reconstitute the complex from recombinant subunits is appealing, particularly if large amounts of complexes are required for structural or functional studies. Two opposite strategies are then possible. On the one hand, one can overexpress simultaneously all the subunits in a single system and rely on this system to reassemble the complex (Tan, 2001). On the other hand, each of the subunits may be expressed independently. While this is often technically simpler, suitable conditions allowing the reassembly of the complex in vitro needs then to be developed (e.g., Tan *et al*., 1996). Obviously, a combination of these two strategies may be used when appropriate. Various hosts have been used for the overproduction of recombinant complexes including *E. coli* (Tan, 2001) and baculoviruses (Lawrie *et al*., 2001). The main advantages of the former are the production costs, its rapidity and easy handling, and the wide knowledge accumulated on this host over the years. Furthermore, vectors allowing the coexpression of multisubunit complexes are now available. Baculoviruses are more costly and time consuming but allow for the easy testing of various combinations of coexpressed subunits as those are simply generated by coinfection. In all cases, complexes are often produced at high level and thus necessitate a single (affinity-) purification step before their use. Like for single protein overexpression, recombinant protein complex production may suffer from the potential absence of protein modification and processing that naturally occur in the natural host. However, the main limitation may reside in the number of subunits that this strategy is able to handle, either due to the difficulties associated with overproducing simultaneously many subunits in a single host or because of the problem encountered when reassociating independently produced subunits in vitro. Up to now, only relatively small assemblies have been produced in a recombinant form and larger complexes such as the yeast proteasome (Groll *et al*., 1997) or archeal/bacterial ribosomes (e.g., Wimberly *et al*., 2000) need to be isolated from their natural hosts. However, given the increased interest for protein complexes, one can anticipate that improved strategies for the production of such assemblies will appear in the near future.

5. Biochemical purification of complexes: what for?

Different reasons may lead to the purification of a protein complex. First, one may want to know the partners of a given protein as those may modulate its function or affect its cellular location. The recent developments in mass spectrometry (reviewed by Aebersold and Mann, 2003; *see also* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5) facilitate this task nowadays by providing a rapid and sensitive manner to identify low amounts of protein. Combined with efficient affinity tags and rapid purification protocols, this technology has allowed large-scale analyses of protein partners in the model eukaryotic species *Saccharomyces cerevisiae* (Gavin *et al*., 2002; Ho *et al*., 2002) as well as in human cells (Bouwmeester *et al*., 2004). These studies have revealed that most

proteins assemble in protein complexes. Data could be compared for yeast for which many large-scale experiments have been performed. Surprisingly, slightly different results were obtained using different tags (von Mering *et al*., 2002) and the data were also poorly overlapping with the pairwise interaction network obtained using the two-hybrid method (Uetz *et al*., 2000). These differences are difficult to explain for the time being. One can expect, however, that accumulating data will attribute these discrepancies to methodological problems and/or difference in the various experiments (strains, growth conditions, etc.). These studies clearly indicate, however, that the functional units of cells are protein complexes rather than independent polypeptides. A second application of the biochemical purification of protein complexes is to learn about the activity and function of an entity once its composition is known. The purified complex may be used in all kinds of biochemical and biophysical assays to get some insights into its biological role. Among them, one may envisage to test for interaction between complexes and/or for posttranslational modifications modulating their activity. Last but not least, a biochemically purified complex of defined composition may be used to get some information on the architectural organization of complexes. This includes simple information on subunit stoichiometry as well as protein–protein contacts and extends to detailed data encompassing the protein complex structure obtained by electron microscopy or X-ray crystallography (Aloy *et al*., 2004; *see also* Article 35, **Structural biology of protein complexes**, Volume 5 and Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6). Finally, while the list of strategies used to analyze biochemically purified complexes includes a large number of methods, one may anticipate further developments in this area in the near future, given the implication of such assemblies in all cellular functions.

Acknowledgments

I thank my colleagues for encouragement and support. Work in my laboratory was supported by La Ligue contre le Cancer (Équipe Labelisée 2001), CNRS (Programme PGP), the Human Science Frontier Program and the French Research Ministry (ACI BCMS).

References

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, *et al* (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.

Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, *et al* (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nature cell biology*, **6**, 97–105.

Caspary F, Shevchenko A, Wilm M and Seraphin B (1999) Partial purification of the yeast U2 snRNP reveals a novel yeast pre-mRNA splicing factor required for pre-spliceosome assembly. *The EMBO Journal*, **18**, 3463–3474.

Deutscher M (1990) *Guide to Protein Purification*, Academic Press: San Diego.

Fields S and Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD and Huber R (1997) Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature*, **386**, 463–471.

Gygi SP, Han DK, Gingras AC, Sonenberg N and Aebersold R (1999) Protein analysis by mass spectrometry and sequence database searching: tools for cancer research in the post-genomic era. *Electrophoresis*, **20**, 310–319.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Kaelin WG Jr., Pallas DC, DeCaprio JA, Kaye FJ and Livingston DM (1991) Identification of cellular proteins that can interact specifically with the T/E1A-binding region of the retinoblastoma gene product. *Cell*, **64**, 521–532.

Kramer A and Utans U (1991) Three protein factors (SF1, SF3 and U2AF) function in pre-splicing complex formation in addition to snRNPs. *The EMBO Journal*, **10**, 1503–1509.

Lane DP and Crawford LV (1979) T antigen is bound to a host protein in SV40-transformed cells. *Nature*, **278**, 261–263.

Lawrie AM, Tito P, Hernandez H, Brown NR, Robinson CV, Endicott JA, Noble ME and Johnson LN (2001) Xenopus phospho-CDK7/cyclin H expressed in baculoviral-infected insect cells. *Protein Expression and Purification*, **23**, 252–260.

Lehman IR, Bessman MJ, Simms ES and Kornberg A (1958) Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. *The Journal of Biological Chemistry*, **233**, 163–170.

Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF and Shilatifard A (2001) COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12902–12907.

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M and Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**, 218–229.

Richardson CC, Schildkraut CL, Aposhian HV and Kornberg A (1964) Enzymatic Synthesis of Deoxyribonucleic Acid. Xiv. Further Purification and Properties of Deoxyribonucleic Acid Polymerase of *Escherichia coli*. *The Journal of Biological Chemistry*, **239**, 222–232.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Rout MP, Aitchison JD, Suprapto A, Hjertaas K, Zhao Y and Chait BT (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *Journal of Cell Biology*, **148**, 635–651.

Tan S (2001) A modular polycistronic expression system for overexpressing protein complexes in *Escherichia coli*. *Protein Expression and Purification*, **21**, 224–234.

Tan S, Hunziker Y, Sargent DF and Richmond TJ (1996) Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature*, **381**, 127–151.

Terpe K (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, **60**, 523–533.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T and Ramakrishnan Jr V (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.

Wouters FS, Verveer PJ and Bastiaens PI (2001) Imaging biochemistry inside cells. *Trends in Cell Biology*, **11**, 203–211.

Specialist Review

Inferring gene function and biochemical networks from protein interactions

Stephen W. Michnick
Université de Montréal, Montréal, QC, Canada

1. Introduction

The emerging offspring of the genomics revolution variously called proteomics, functional genomics, chemical genetics, or systems biology can be attributed an overall aim: as only a fraction of gene functions can be inferred from primary gene sequences, we need to develop strategies to define gene function that are not conducted at the level of a classical gene-by-gene approach but that aim at characterizing the totality of genes or large subsets thereof. The question then is by what approaches do we meaningfully ascribe function to genes? Furthermore, the emerging fields of systems biology and chemical genetics seek a deeper appreciation of the biochemical organization of living cells and the molecular schemes that all living things share as well as those things that make individual cells and organisms unique. What we seek are conceptual and experimental approaches that will tell us how, when, where, and under what circumstances mRNA message and proteins are processed, become modified to be activated or inactivated, how they are destroyed, and, central to all of these, what other molecules do they interact with.

In this review, I will discuss three issues. First, I will provide some historical, logical, and mechanistic context to explain how the measurement of protein–protein interactions has come to be a crucial component of efforts to define gene function and the organization of biochemical networks. Second, I will explain a conceptual basis for using protein interaction data, along with other approaches such as genome-wide gene expression analysis and genetics to define the organization of biochemical networks. I will specifically discuss how one particular experimental approach, protein fragment complementation assays (PCA), can allow us to go beyond the static representations of protein interactions that we now use, to dynamic studies of protein interactions as a tool for inferring the organization of information flow in biochemical networks. Finally, I will provide some perspectives on where protein interaction networks (PINs) take us, with a discussion of their utility in chemical genetic analyses; the use of organic molecules to both probe biochemical processes and protein interactions as probes for the actions of such molecules on biochemical networks.

2. Protein–protein interactions and gene function

Certainly, an overall goal of genomics is to ascribe function to gene products on a large scale. Even among very well studied organisms such as the budding yeast *Saccharomyces cerevisiae* and enteric bacterium *Escherichia coli*, a significant proportion of their genes cannot be even classified as participating in some general cellular function. Why is this the case and what must be done to remedy the situation for so many genes? It is useful here to take a historical perspective in explaining how techniques have evolved to better address gene function. In the 1980s, there emerged a new breed of biologist, whom we came to call "gene jocks". These were rare individuals who had mastered what were then the new methods of molecular cloning and their utility in identifying the genes underlying biological processes. Specifically, they devised many ingenious strategies to simultaneously screen cDNA libraries using a protein- or enzyme-specific assay that allow for both selection of positive clones and validation of their biological relevance with the same assay (Figure 1a) (see Grimm, 2004 for review) (Aruffo and Seed, 1987; D'Andrea *et al*., 1989; Lin *et al*., 1992; Sako *et al*., 1993). As clever as these screening designs could be, a major barrier to utilizing such approaches on a large scale is that one needs a very specific, sensitive, simple, and scalable assay to directly determine whether a cDNA codes for a sought-after activity. All of this changed following the publication of Fields and Song (1989) on the yeast two-hybrid expression-cloning strategy (Figure 1a). The genius of this approach is that it is a simple, robust and scalable assay, but must important, the detection of interactions of bait proteins of known functions with expressed proteins in a cDNA library is completely general, thus removing the need to develop a specific assay for every gene. The elegance of the two-hybrid expression-cloning approach has made it among the most widely used of techniques and also the first approach used to achieve genome-wide screening for protein:protein interactions in several organisms (Fields and Song, 1989; Drees, 1999; Vidal and Legrain, 1999; Ito *et al*., 2000; Uetz *et al*., 2000; Walhout *et al*., 2000; Ito *et al*., 2001). Completely different, but in vitro strategies to map out protein–protein interactions on a large scale are based on large-scale protein purification followed by mass-spectroscopic analysis (Gavin *et al*., 2002; Ho *et al*., 2002). However, there is a crucial technical distinction between large-scale in vitro approaches versus in vivo techniques such as the two-hybrid system. With in vivo methods, one avoids the complex problem of defining conditions under which proteins can be extracted from cell lysates. For example, extracted proteins may be denatured, exist in a form in the cell that cannot be easily reproduced in vitro, or individual protein complexes may be weakly associated and therefore undetectable in dilute solutions.

A purely protein interaction–based screening approach is limited in that the assays themselves do not provide any immediate information that would allow one to decide whether a cDNA product is likely to be involved in a specific cellular function and so while two-hybrid assays achieve generality, specificity is lost. Ideally, it would be advantageous to combine the generality of two-hybrid screening with the specificity of functional assay-based assays as done in classical expression-cloning approaches. This has been the subject of research in my lab for the last 10 years and, specifically, the development of what we call protein fragment

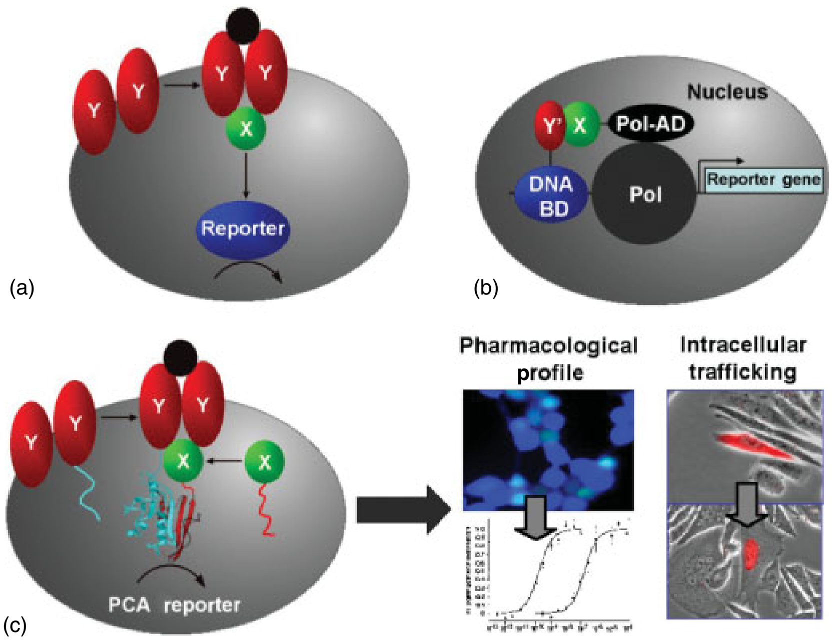


Figure 1 Alternative cell-based expression-cloning strategies. In all examples, a hypothetical intracellular effector protein (X) of a known membrane-associated hormone receptor (Y) is sought by screening a cDNA library. (a) Classical expression cloning strategy. A cell line that expresses receptor Y is transfected with plasmids harboring a cDNA library. An assay is devised to detect whether the potential effector protein cDNA is expressed from a library plasmid, consisting of a reporter enzyme (Reporter), which is activated to produce a detectable product if the receptor is bound to hormone (black sphere). Clonal cells showing reporter activation are isolated and the cDNA-harboring plasmid extracted and sequenced to identify the effector protein X. Note that reporter activation by hormone-induced receptor activation when X is expressed serves as a "first-pass" validation of X being a biologically relevant effector. (b) A yeast two-hybrid expression-cloning strategy. Here it is assumed that effector protein X interacts with an intracellular domain of the receptor Y called Y'. Receptor domain Y' is fused to the DNA binding domain, while a cDNA library is fused to the activation domain of a yeast transcription factor (e.g., Gal4) and expressed in the *S. cerevisiae* nucleus. If plasmid harboring X is expressed simultaneously with Y', the transcription factor is reconstituted, resulting in transcription and expression of a reporter gene (e.g., beta-galactosidase, selectable metabolic marker, etc.). cDNA from positive clones is sequenced to identify putative effector protein X. First-pass validation requires some biological assay (e.g., system described in (a)). (c) A PCA-based expression-cloning strategy. Like the two-hybrid approach, it is assumed that effector protein interacts with the receptor. Unlike the classical approach, no specific assay for receptor activation is required. Receptor Y is expressed as a fusion with one complementary fragment (cyan), while cDNA library is constructed in plasmids as fusion to the complementary fragment (red) of a PCA reporter protein. If X is coded in the cDNA, it interacts with Y, allowing for folding and reconstitution of the PCA reporter protein. (1c, right) First-pass validation is performed by examining response of the interaction of X with Y to hormone stimulation. For example, hormone dose-response of PCA reporter versus or other biologically relevant changes, such as cellular translocation of the X-Y complex (right), are measured

complementation assays (PCA) and their applications to expression cloning. In the general PCA strategy, an enzyme is *rationally* dissected into two fragments and the fragments are fused to two proteins that are thought to bind to each other. Folding of the reporter enzyme from its fragments is catalyzed by the binding of the test proteins to each other, and is detected as reconstitution of enzyme activity (Figure 1c). The PCA strategy takes advantage of the spontaneous all-or-none nature of protein folding and the design of fragments follows from basic concepts of protein engineering. The all-or-none folding of reporter protein from fragments means that the generation of signal by complementation has an enormous dynamic range over a very narrow range conditions, unlike fluorescence resonance energy transfer techniques (Zhang *et al*., 2002). We demonstrated the principle of PCA and it has been generalized to a number of enzymes including dihydrofolate reductase, glycinamide ribonucleotide transformylase, aminoglycoside kinase, hygromycin B kinase, TEM ß-lactamase, green fluorescent protein (GFP), and firefly and renilla luciferases (Pelletier *et al*., 1998; Pelletier *et al*., 1999; Remy *et al*., 1999; Ghosh *et al*., 2000; Michnick *et al*., 2000; Remy and Michnick, 2001; Remy *et al*., 2001; Galarneau *et al*., 2002; Hu *et al*., 2002; Paulmurugan *et al*., 2002; Spotts *et al*., 2002; Wehrman *et al*., 2002; Paulmurugan and Gambhir, 2003). The application of PCAs to library screening has been demonstrated in a systematic model and to screening of cDNA libraries. The cDNA library screening strategy consists of two steps (Figure 1c): First, a large-scale screening of “bait” proteins (X) fused to one PCA reporter fragment against a cDNA library fused to the complementary reporter fragment. In the second step, positive hits from the screen are then directly functionally “validated” by testing for perturbations of the interaction, as measured by PCA, by agents that act on the biochemical network in which the bait protein is known to participate. We have described a cDNA library screening strategy using a GFP-based PCA in which we successfully identified novel substrates and regulators of the serine/threonine protein kinase PKB/Akt (Remy and Michnick, 2004a; Remy *et al*., 2004).

The PCA strategy has these unique features:

- Molecular interactions, including complete cellular networks, are detected directly – not through secondary events such as transcription activation.
- Genes are expressed in the relevant cellular context, reflecting the native state of the protein, with the correct posttranslational modifications.
- Events induced by hormones or growth factors can be detected, providing target validation by linking specific interactions to specific networks. For example, using insulin-sensitive cells, molecular interactions stimulated by insulin can be detected and quantitated.
- The subcellular location of protein interactions can be determined, whether in the membrane, cytoplasm, or nucleus and the movement of protein complexes can be visualized.

In addition to the specific capabilities of PCA described above, are special features of this approach that make it appropriate for genomic screening of molecular interactions, including: (1) PCAs are not a single assay but a series of assays; an assay can be chosen because it works in a specific cell type appropriate

for studying interactions of some class of proteins; (2) PCAs are inexpensive, requiring no specialized reagents beyond those necessary for a particular assay and off-the-shelf materials and technology; (3) PCAs can be automated and high-throughput screening could be done; (4) PCAs are designed at the level of the atomic structure of the enzymes used; because of this, there is additional flexibility in designing the probe fragments to control the sensitivity and stringencies of the assays; (5) PCAs can be based on enzymes for which the detection of protein–protein interactions can be determined differently including by dominant selection or production of a fluorescent or colored product.

3. Mapping the biochemical machinery of cells

The potential application of genome-wide approaches to cellular biology is becoming a reality. For example, drawing upon systematic genome-wide analyses of *S. cerevisiae*, several groups have suggested that biochemical networks may be organized as modules of physically or genetically linked genes that contribute to a particular cellular function (Hartwell *et al*., 1999; Winzeler *et al*., 1999; Ihmels *et al*., 2002; Snel *et al*., 2002; Rives and Galitski, 2003; Tong *et al*., 2004; Yao *et al*., 2004; Yook *et al*., 2004), and it has been proposed that such analyses could be used to identify cellular processes to which specific genes are linked and upon which cellular perturbations act (Hartwell *et al*., 1997; Marton *et al*., 1998; Giaever *et al*., 1999; Hardwick *et al*., 1999; Hughes *et al*., 2000; Kuruvilla *et al*., 2002; Giaever *et al*., 2004; Lum *et al*., 2004; Parsons *et al*., 2004).

Why is it that the systematic collection of protein–protein interactions is so important to the goal of delineating the biochemical networks that underlie cellular processes? The reason is a fundamental fact about the nature of biochemical networks. Biochemical processes are mediated by dynamic noncovalently associated multienzyme complexes that permit the most efficient chemical machinery, in which the substrates and products of a series of steps are transferred from one active site to another over minimal distance, with minimal diffusional loss of unstable intermediates, and in chemical environments suited to stabilizing reactive intermediates (Perham, 1975; Pawson and Nash, 2000; Vidal, 2001). Biochemical "networks" are complexes of dynamically assembling and disassembling protein complexes and, therefore, a meaningful representation of a biochemical network in a living cell would be first, a step-by-step analysis of individual protein–protein interactions and second, analysis of the dynamics of interactions in response to perturbations that impinge upon the network under study and the temporal and spatial distribution of these interactions.

How can protein interaction networks (henceforth, PINs) help us to study these dynamic assembly events? It is first instructive to examine how biochemical networks have been studied at a genome-wide scale in gene transcription experiments. The advent of DNA microarray technologies has changed the manner in which we view biochemical networks (Chu *et al*., 1998; Holstege *et al*., 1998; Spellman *et al*., 1998; Hughes *et al*., 2000; Roberts *et al*., 2000; Ideker *et al*., 2001). A strategy for mapping biochemical networks on the basis of expression profiles depends on two components: first, the ability to monitor the expression

of genes on a genome-wide level and second, ways to systematically perturb the biochemical networks underlying the expression of genes in a systematic way (Figure 2). The key practical issue is whether a set of specific perturbations is possible. In yeast, this is relatively simple thanks to the availability of a systematic set of gene knockouts, while for other organisms, siRNAs are coming to be the most common approach to gene perturbation (McManus and Sharp, 2002). The practical monitoring of changes in expression of complete genomes or large subsets of genes has allowed researchers to begin to scrutinize in some detail the evolution of genetic programs and sometimes, by inference, the underlying biochemical networks that control these programs. The most well described analyses of gene expression have been performed for the bakers yeast *S. cerevisiae*, for which the entire genome

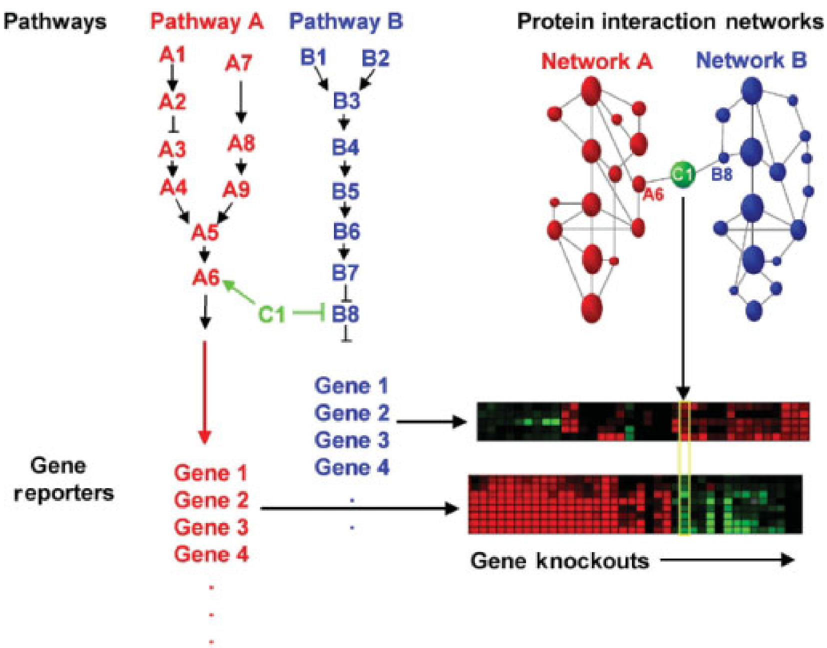


Figure 2 Combining gene expression profiles and protein-interaction networks to determine the organization of biochemical pathways. Two related pathways (A and B) are thought to be organized as depicted (upper left). The role of each protein as an activator (arrows) or inhibitor (T-bars) of the pathway is indicated. Interactions of the component proteins of these pathways (upper right) suggest a modular organization in which an additional protein (C1) has been identified, linking the two pathways. In order to confirm the organization of these pathways and role of each protein, the expression of reporter genes that are activated by the individual pathways are monitored (lower right). Relative expression levels are determined for wild-type versus cells in which individual pathway component proteins have been knocked out. The position in the pathway and role of the novel pathway integrator C1 are determined. First, the position of C1 in the pathways is determined by which proteins in the two pathways C1 interact with (A6 and B8). Second, the role of C1 is determined by comparing reporter gene expression in wild-type and cells in which C1 is knocked out (lower right, yellow rectangle). Knocking out C1 results in an increase in expression of pathway B reporter genes (red squares) and a decrease in expression of pathway A reporter genes (green squares). Expression profile results are consistent with an inhibitory role of C1 on pathway B and an activating role on pathway A as depicted

has been sequenced and microarrays representing all predicted genes have been available for some time (Lashkari *et al*., 1997). Complete analyses of the cell cycle and responses of the organism to a number of perturbations have been explored. It is possible to conceive of mapping the output of a genetic program (gene expression changes) back to the organization of the biochemical networks underlying these changes. This transcriptocentric approach to mapping networks does not help us to side-step an obvious problem: If you do not know the details of the underlying machinery leading to specific outputs of a system, then any new insights from analyzing outputs remain inferences that must still be tested directly. It is like saying that you might be able to figure out how a radio works by monitoring the audio output and systematically removing components while changing the tuning and other settings. This is where PINs become an important component of a more detailed analysis of biochemical networks. PINs have already been used extensively as a strategy to identify genes involved in a wide range of biological processes from the development of specific organs to the identification of pathogenic proteins (Walhout *et al*., 2000; Rain *et al*., 2001). However, it was Ideker and colleagues who first described how a PIN can be used to guide a systematic analysis of a biochemical network, using a unique metabolic response, the switch from glucose to galactose as carbon source in yeast as an example (Ideker *et al*., 2001). In their strategy, a PIN is constructed for all known interactions for proteins, which are either known to interact or might be inferred to interact on the basis of common changes in expression in response to the glucose-galactose switch. Combining these data with examinations of changes in protein expression and knockouts of individual genes, they were able to both reproduce what was known about the glucose-galactose response, and to identify some proteins whose role in the response was not known. Thus, the PIN constructed for the glucose-galactose response genes served as a template for generating and testing hypotheses about the organization of a biochemical network underlying a simple but important biological adaptation. Another elegant example of where a PIN was used as the framework for the detailed analysis of the biochemical networks underlying a process was done by Yao and colleagues on how proteins involved in the vertebrate aryl receptor response could be organized into functional classes of interacting components (Yao *et al*., 2004).

In the examples described above, PINs are static entities; that is, they reflect simply existing knowledge from a variety of sources of which proteins interact with each other. It would be of tremendous value if the interactions themselves could be studied as dynamic entities. Specifically, if we could understand where, when, and how interactions are occurring in a living cell, we could use these sources of information not merely as a framework for constructing a static map of biochemical networks but a dynamic description of information flow through networks.

In a PCA-based biochemical network mapping project, specific PCAs for proteins that interact at various strategic points (for example at different points in a pathway hierarchy) in a network serve as "sentinels" for the state of the pathway under different conditions. First, cells containing PCA sentinels are treated with agents (chemical inhibitors siRNAs, hormones, etc.) that would be thought to perturb the biochemical network under study. A change in the PCA sentinels reporter signal would then reveal that the relationship is between the point of action of

the perturbing agent (say, some enzyme in a subnetwork or "module") and the sentinels. So for instance, if an enzyme were inhibited with a small molecule and the sentinel signal decreased, we could hypothesize that this enzyme must be somehow positively coupled to the function of the sentinel proteins. A series of perturbations within individual modules in the network would result in a pattern of responses or "pharmacological profile", as detected by PCA, which should be consistent with the response of the pathway under study. Second, interactions of protein components of a network should take place in specific subcellular compartments or locations consistent with the function of the network. The combined pharmacological profiles and subcellular interaction patterns serve then to describe a biochemical network (Remy and Michnick, 2001; Remy and Michnick, 2004b,c; Remy *et al*., 2004).

As much as we want to know how the components of biochemical networks interact with each other, we also want to know how agents (e.g., siRNAs, gene knockouts, small organic molecules) that perturb these networks exert their effects. For example, even if we know that a small organic molecule acts very specifically on some target protein in vitro, this in no way assures that the consequences of the actions of the molecule on the target in the cell can be predicted. Three examples illustrate how this PCA sentinel approach might be applied to determine actions of a small molecule on the cell. In a first example, a small molecule has been shown to act on a specific target. We thus verify the predicted outcome on the network in which the target protein participates, by testing the compound against a series of PCA sentinels that report on the state of that network and also on a series of control networks. If the molecule does what we predict, only the target network is affected, however, if sentinels in other networks report perturbations, then the specificity of action of the molecule must be reevaluated. In a second example, a novel compound has been shown to cause some profound phenotype in a cell and we want to quickly link the observed phenotype to effects on known networks that would likely be those affected. The same strategy is used as above, but in this case we hypothesize that in order to obtain the observed pleiotropic phenotype, more than one network must be affected. A third example is one in which the specific target of a compound is known, but paradoxically, predicted effects on cell do not occur because of feed back from a branch in the biochemical network that compensates for the inhibition of a target. There are numerous examples of such paradoxes (Hall-Jackson *et al*., 1999). In these cases, a "two-hit" strategy of knocking out both the target and another target in the compensating branch (as monitored by the PCA sentinels signals) might produce a profound phenotype in the cell, while hitting either individual target would do nothing.

The strategies described above turn chemical genetics to a new focus that we call "network-based targeting", as opposed to protein targeting that is usually practiced. There are some considerable advantages of network targeting that can be summarized as follows: (1) optimal target(s) for a network of interest can be identified in simple cell-based assay; (2) the evaluation of off-pathway effects can be studied with the same assays; (3) the assays are simple and performed in intact cells and thus necessitate neither the production of recombinant proteins nor protein purification, as is required for the development of large scale in vitro based assays, often an intractable problem; (4) finally, this strategy enables the analysis of combinations of perturbations that may impact multiple targets simultaneously.

4. Conclusions and perspectives

I have attempted here to provide an overview of how the construction of PINs can and will continue to guide efforts to provide the basis for gene function assignment and for studying biochemical networks. Static PINs as provided by two-hybrid and mass-spectroscopic approaches will have an enduring place in these efforts, while PCA-based approaches will allow us to "place" each gene product at its relevant point in a network and define the dynamic organization of the network. Specific chemical and other perturbations could be used to generate a functional profile, including dominant-negative forms of enzymes, receptor- or enzyme-specific peptides, or siRNAs. The ability to monitor the network in living cells containing all of the components of the network studied can reveal hidden connections and will allow for a full investigation of the global organization of biochemical networks in cells and to address fundamental questions about how biochemical machineries are organized (Hartwell *et al*., 1999). Finally, I discussed how network mapping could be used in combination with other genome-wide approaches to increase the quantity and quality of information about the actions of small molecules on the living cells and the intricate networks that make up their chemical machinery.

References

Aruffo A and Seed B (1987) Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. *Proceedings of the National Academy of Sciences of the United States of America*, **84**(23), 8573–8577.

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, and Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**(5389), 699–705.

D'Andrea AD, Lodish HF and Wong GG (1989) Expression cloning of the murine erythropoietin receptor. *Cell*, **57**, 277–285.

Drees BL (1999) Progress and variations in two-hybrid and three-hybrid technologies. *Current Opinion in Chemical Biology*, **3**(1), 64–70.

Fields S and Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Galarneau A, Primeau M, Trudeau LE and Michnick SW (2002) beta-Lactamase protein fragment complementation assays as in vivo and in. *Nature Biotechnology*, **20**(6), 619–622.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Ghosh I, Hamilton AD and Regan L (2000) Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *Journal of the American Chemical Society*, **122**(23), 5658–5659.

Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP and Davis RW (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(3), 793–798.

Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A and Davis RW (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics*, **21**(3), 278–283.

Grimm S (2004) The art and design of genetic screens: mammalian culture cells. *Nature Reviews. Genetics*, **5**(3), 179–189.

Hall-Jackson CA, Eyers PA, Cohen P, Goedert M, Boyle FT, Hewitt N, Plant H and Hedge P (1999) Paradoxical activation of Raf by a novel Raf inhibitor. *Chemistry & Biology*, **6**(8), 559–568.

Hardwick JS, Kuruvilla FG, Tong JK, Shamji AF and Schreiber SL (1999) Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(26), 14866–14870.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature*, **402**(Suppl 6761), C47–C52.

Hartwell LH, Szankasi P, Roberts CJ, Murray AW and Friend SH (1997) Integrating genetic approaches into the discovery of anticancer drugs. *Science*, **278**(5340), 1064–1068.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**(6868), 180–183.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES and Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**(5), 717–728.

Hu CD, Chinenov Y and Kerppola TK (2002) Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Molecular Cell*, **9**(4), 789–798.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai E, He YD, *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**(1), 109–126.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**(5518), 929–934.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y and Barkai N (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**(4), 370–377.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, Zhang J, Ma Y, Taylor SS and Tsien RY (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, SakakiY, Giaever G, *et al*. (2000) Toward a protein-a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(3), 1143–1147.

Kuruvilla FG, Shamji AF, Sternson SM, Hergenrother PJ and Schreiber SL (2002) Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature*, **416**(6881), 653–657.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO and Davis RW (1997) Yeast microarrays for genome-wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(24), 13057–13062.

Lin HY, Wang XF, Ng-Eaton E, Weinberg RA and Lodish HF (1992) Expression cloning of the TGF-beta type II receptor, a functional transmembrane serine/threonine kinase [published erratum appears in *Cell* 1992 Sep 18;**70**(6):following 1068]. *Cell*, **68**(4), 775–785.

Lum PY, Armour CD, Stepaniants SB, Cavet G, Wolf MK, Butler JS, Hinshaw JC, GarnierP, Prestwich GD, Leonardson A, *et al*. (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**(1), 121–137.

Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, BurchardJ, Slade D, Dai H, *et al*. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, **4**(11), 1293–1301.

McManus MT and Sharp PA (2002) Gene silencing in mammals by small interfering RNAs. *Nature Reviews. Genetics*, **3**(10), 737–747.

Michnick SW, Remy I, C-Valois F-X, V-Belisle A and Pelletier JN (2000) Detection of protein–protein interactions by protein fragment complementation strategies. In *Methods in*

Enzymology, Vol. 328, Abelson JN, Emr SD and Thorner J (Eds.), Academic Press: New York, pp. 208–230.

Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR and Boone C (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature Biotechnology*, **22**(1), 62–69.

Paulmurugan R and Gambhir SS (2003) Monitoring protein-protein interactions using split synthetic renilla luciferase protein-fragment-assisted complementation. *Analytical Chemistry*, **75**(7), 1584–1589.

Paulmurugan R, Umezawa Y and Gambhir SS (2002) Noninvasive imaging of protein-protein interactions in living subjects by using reporter protein complementation and reconstitution strategies. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(24), 15608–15613.

Pawson T and Nash P (2000) Protein-protein interactions define specificity in signal transduction. *Genes & Development*, **14**(9), 1027–1047.

Pelletier JN, Arndt KM, Plöckthun A and Michnick SW (1999) An *in vivo* competition strategy for the selection of optimized protein-protein interactions. *Nature Biotechnology*, **17**, 683–690.

Pelletier JN, Remy I and Michnick SW (1998) Protein-fragment complementation assays: a general strategy for the *in vivo* detection of protein-protein interactions. *Journal of Biomolecular Techniques*, **10**, 32–39.

Perham RN (1975) Self-assembly of biological macromolecules. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **272**(915), 123–136.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of Helicobacter pylori. *Nature*, **409**(6817), 211–215.

Remy I and Michnick SW (2001) Visualization of biochemical networks in living cells. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(14), 7678–7683.

Remy I and Michnick SW (2004a) A cDNA library functional screening strategy based on fluorescent protein complementation assays to identify novel components of signaling pathways. *Methods*, **32**(4), 381–388.

Remy I and Michnick SW (2004b) Mapping biochemical networks with protein-fragment complementation assays. *Methods in Molecular Biology*, **261**, 411–426.

Remy I and Michnick SW (2004c) Regulation of apoptosis by the Ft1 protein, a new modulator of protein kinase B/Akt. *Molecular and Cellular Biology*, **24**(4), 1493–1504.

Remy I, Montmarquette A and Michnick SW (2004) PKB/Akt modulates TGF-beta signalling through a direct interaction with Smad3. *Nature Cell Biology*, **6**(4), 358–365.

Remy I, Pelletier JN, Galarneau A and Michnick SW (2001) Protein interactions and library screening with protein fragment complementation strategies. In *Protein-protein Interactions: A Molecular Cloning Manual*, Golemis EA (Ed.), Cold Spring Harbor Laboratory Press: Plainview, pp. 449–475.

Remy I, Wilson IA and Michnick SW (1999) Erythropoietin receptor activation by a ligand-induced conformation change. *Science*, **283**, 990–993.

Rives AW and Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(3), 1128–1133.

Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, *et al*. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**(5454), 873–880.

Sako D, Chang XJ, Barone KM, Vachino G, White HM, Shaw G, Veldman GM, Bean KM, Ahern TJ, Furie B, *et al*. (1993) Expression cloning of a functional glycoprotein ligand for P-selectin. *Cell*, **75**(6), 1179–1186.

Snel B, Bork P and Huynen MA (2002) The identification of functional modules from the genomic association of genes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(9), 5890–5895.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast

Saccharomyces cerevisiae by microarray hybridization. *Molecular and Cellular Biology*, **9**(12), 3273–3297.

Spotts JM, Dolmetsch RE and Greenberg ME (2002) Time-lapse imaging of a dynamic phosphorylation-dependent protein-protein interaction in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(23), 15142–15147.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**(5659), 808–813.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Vidal M (2001) A biological atlas of functional maps. *Cell*, **104**(3), 333–339.

Vidal M and Legrain P (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Research*, **27**(4), 919–929.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, and Vidal M (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**(5450), 116–122.

Wehrman T, Kleaveland B, Her JH, Balint RF and Blau HM (2002) Protein-protein interactions monitored in mammalian cells via complementation of beta-lactamase enzyme fragments. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(6), 3469–3474.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al*. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**(5429), 901–906.

Yao G, Craven M, Drinkwater N and Bradfield CA (2004) Interaction networks in yeast define and enumerate the signaling steps of the vertebrate aryl hydrocarbon receptor. *PLoS Biology*, **2**(3), E65.

Yook SH, Oltvai ZN and Barabasi AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics*, **4**(4), 928–942.

Zhang J, Campbell RE, Ting AY and Tsien RY (2002) Creating new fluorescent probes for cell biology. *Nature Reviews. Molecular Cell Biology*, **3**(12), 906–918.

Specialist Review

The *C. elegans* interactome project

David E. Hill, Michael E. Cusick and Marc Vidal
Dana-Farber Cancer Institute, Harvard Medical School, Center for Cancer Systems Biology, Boston, MA, USA

1. Introduction

Molecular biology has uncovered fundamental mechanisms of biology and of human disease, but essential questions remain unanswered. Particularly, since gene products usually mediate their function in association with other gene products, biological processes need to be regarded as complex cellular networks of interconnected components. It is increasingly apparent that such complex networks have properties of their own, sometimes referred to as "emergent properties". Those properties might relate back to intriguing novel biological phenomena (*see* Article 118, **Data collection and analysis in systems biology**, Volume 6).

Thus, to understand both normal biological processes of interest as well as molecular disease mechanisms, one needs to consider systems approaches that allow global analyses of the properties, dynamics, and functions of the molecular networks in which gene products act (Ideker *et al*., 2001; Vidal, 2001). It is becoming clear that even when we will have precise structural and functional information on most gene products, we still will not be able to fully understand biological phenomena and disease mechanisms. Again, this is because the systems formed by large numbers of individual genes, proteins, RNAs, and metabolites tend to exhibit properties that are not observable by reductionism, which consists in studying genes one at a time. Such a paradigm shift is particularly important to consider in light of novel therapeutic strategies. For a thorough understanding of drug action, and to maximize our chances of finding novel robust therapies, we probably will need to pay more attention to the complexity of cellular networks (Kitano, 2002).

One of the first steps in the development of global approaches to comprehend molecular networks is the systematic mapping of macromolecular interactions and biochemical reactions. Since many gene products mediate their function or are regulated through physical protein–protein interactions, we need to develop ways to comprehensively map protein–protein interaction, or "interactome", networks (*see* Article 39, **The yeast interactome**, Volume 5). Ultimately, our goal should be to generate a high-quality "human interactome map" that, together with other functional genomic and proteomic (or "omic") information, will serve as a backbone

for the drawing of global functional wiring diagrams (Vidal, 2001; Ge *et al*., 2003). Given how the human genome project (HGP) (*see* Article 24, **The Human Genome Project**, Volume 3) benefited from the sequencing and annotation of the genome of various model organisms, it was proposed several years back that the launching of a human interactome project would also require the development of interactome projects for the proteome of selected model organisms (Walhout *et al*., 1998).

Using the nematode *Caenorhabditis elegans* as a model system, we have developed several key solutions to the challenges inherent to the making of interactome maps for metazoan organisms. So far, we have generated a *C. elegans* interactome map that contains ∼5500 potential interactions, referred to below as Worm Interactome Version 5 (WI5) (Walhout *et al*., 2000b; Davy *et al*., 2001; Matthews *et al*., 2001; Boulton *et al*., 2002; Walhout *et al*., 2002; Reboul *et al*., 2003; Li *et al*., 2004; Tewari *et al*., 2004). This dataset, together with a recently released interactome map for *Drosophila melanogaster*, represents the first interactome map for metazoan organisms.

Although already helpful, the protein interaction data in WI5 is far from complete and needs improvements. We will discuss how to improve interactome maps on three fundamental aspects of the worm interactome map: completeness, sensitivity, and specificity.

2. Yeast interactome maps

The yeast *Saccharromyces cerevisiae* has been used to develop "unicellular" eukaryotic interactome maps (*see* Article 39, **The yeast interactome**, Volume 5). Large-scale proteome-wide experiments, including high-throughput yeast two-hybrid (HT-Y2H) screens (Uetz *et al*., 2000; Ito *et al*., 2001), affinity purifications followed by mass spectrometry (AP-MS) (Gavin *et al*., 2002; Ho *et al*., 2002), and computational predictions (Dandekar *et al*., 1998; Marcotte *et al*., 1999; Pellegrini *et al*., 1999), together with interactions collected in the MIPS database (Mewes *et al*., 2002; *see also* Article 45, **Computational methods for the prediction of protein interaction partners**, Volume 5) have provided ∼28 000 protein–protein interactions, 2493 of which were identified by at least two of these four sources (constituting the basis of a "filtered yeast interactome" or FYI dataset (Han *et al*., 2004)).

These yeast interactome maps have provided hypothetical functions for hundreds of uncharacterized genes by following the "guilt-by-association" approach, whereby proteins that have similar biological functions tend to cluster together in interactome networks (Schwikowski *et al*., 2000; *see also* Article 46, **Functional classification of proteins based on protein interaction data**, Volume 5). Although the map for the yeast interactome network is far from complete and could be improved in sensitivity and specificity (Grigoriev, 2003), some of its "emergent" global topological properties have already become evident. For example, the yeast interactome network is scale-free, that is, most proteins interact with few partners, while few proteins, the "hubs", interact with many partners (Jeong *et al*., 2001). The small-world nature of protein networks also allows inference of some protein interactions from others (Goldberg and Roth, 2003). Finally, the yeast interactome

data correlate significantly with other functional genomic approaches such as expression (transcriptome) and phenotypic (phenome) profiling (Ge *et al.*, 2001; Grigoriev, 2001; Begley *et al.*, 2002; Jansen *et al.*, 2002; Kemmeren *et al.*, 2002), suggesting that a global understanding of the yeast proteome might arise from the integration of several systematic "-omic" approaches (Vidal, 2001; Ge *et al.*, 2003).

3. *C. elegans* as a model "-omic" organism

Caenorhabditis elegans is a useful multicellular model organism because: (1) genetic manipulations are efficient (Riddle *et al.*, 1997), (2) transgenic animals are conveniently generated (Praitis *et al.*, 2001), (3) its cellular lineage is invariant (all cells are named) and has been more extensively studied than in any other organism (Sulston and Horvitz, 1977; Sulston *et al.*, 1983), (4) its genome is completely sequenced (The C. elegans Sequencing Consortium, 1998), and (5) there is often functional conservation between worm genes and their human orthologs (Pennisi, 1998; Walhout *et al.*, 1998; Tewari *et al.*, 2004). Hence, many biological models obtained from studying the worm can be directly used to predict gene and network functions in human molecular networks. Examples include the discovery of the Ras pathway (Sternberg and Han, 1998; Walhout *et al.*, 2000b), and the discovery of programmed cell death in *C. elegans,* especially the roles of the worm orthologs of caspase (CED-3) and Bcl2 (CED-9) (Yuan and Horvitz, 1990; Hengartner and Horvitz, 1994; Yuan and Horvitz, 2004).

The *C. elegans* EST project has identified more than 12 000 expressed genes to date (Hanazawa *et al.*, 2001). In addition, we initiated genome-wide experimental tests to address the quality of both the gene number estimate and the predictions of exon structure (Reboul *et al.*, 2001; Reboul *et al.*, 2003; Lamesch *et al.*, 2004). Finally, the genomic information is freely available on a widely used database, Wormbase (**http://www.wormbase.org**), in which the genetic data, the physical maps, the genome sequence, and the EST sequences are stored and organized in a user-friendly manner (Stein *et al.*, 2001).

Several laboratories have initiated large-scale projects aimed at describing loss-of-function phenotypes for every single predicted *C. elegans* open reading frame (ORF) using RNAi technologies (Fraser *et al.*, 2000; Gonczy *et al.*, 2000; Piano *et al.*, 2000; Maeda *et al.*, 2001; Kamath *et al.*, 2003; Rual *et al.*, 2004a). A gene expression map, or "topomap", has been deduced from DNA microarray experiments involving ~17 000 genes across more than 550 different conditions (Kim *et al.*, 2001). New technologies are also being developed to solve the 3D structure of many *C. elegans* proteins (Chance *et al.*, 2002; Luan *et al.*, 2004). Additionally, the "localizome" project is under way to determine in what cell and at what stage of development each ORF is expressed (Hope *et al.*, 1996; Dupuy *et al.*, 2004). Finally, an HT-Y2H protein interaction mapping project is ongoing in our laboratory (Walhout *et al.*, 2000b; Davy *et al.*, 2001; Boulton *et al.*, 2002; Walhout *et al.*, 2002; Reboul *et al.*, 2003; Li *et al.*, 2004; Tewari *et al.*, 2004). Importantly, we are in the process of learning how to integrate such large-scale high-throughput datasets with other types of functional genomics data (Boulton *et al.*, 2002; Walhout *et al.*, 2002).

4. Gateway cloning of ORFeomes for functional proteomics

An important challenge in the development of systems approaches is to be able to express nearly all predicted proteins identified as potentially implicated in a biological process of interest, under different conditions and in various hosts. This in turn allows the development of various large-scale functional genomic and proteomic approaches. For example, HT-Y2H (Walhout *et al*., 2000c) and AP-MS (Gavin *et al*., 2002; Ho *et al*., 2002) strategies, as well as proteome chips or reverse transfection strategies (Zhu *et al*., 2001; Ziauddin and Sabatini, 2001), require large numbers of protein-encoding ORFs to be cloned precisely into multiple expression vectors.

Randomly picked cDNA clones in the context of various "transcriptome" projects can rarely be used directly for protein expression, because either their 5′-end is not cloned in the appropriate reading frame, or their 3′-end is not compatible with the expression of C-terminal fusion proteins, or both. Additionally, most cDNAs produced in transcriptome projects are *not* available in vectors that allow transfer of the protein-encoding sequences to a variety of expression vectors by automated, high-throughput methods. Furthermore, standard methods of adding suitable restriction sites to the ends of either DNA fragments or PCR products and then employing restriction digestions and ligations to subclone desired DNA fragments are not readily compatible with high-throughput systems.

To address these problems, an alternative strategy, Gateway recombinational cloning, for manipulating DNA from any source for primary or secondary cloning purposes and utilizing in vitro reactions amenable to high-throughput settings was developed several years ago (Hartley *et al*., 2000; Walhout *et al*., 2000c). Gateway bypasses the need for traditional enzymatic digestion, purification, and ligation of DNA molecules, and is particularly suited for cloning large numbers of DNA fragments, such as ORFs. The Gateway systems employs a two-step strategy, whereby the initial step is the generation of "Entry clones", which constitute a standardized resource that can be shared with the scientific community. Any DNA, such as ORFs, obtained as an "Entry clone" can then subsequently be transferred to any Gateway-compatible expression vector, termed a *Destination clone*, without reducing the quality of the original resource, and virtually any standard expression vector is readily modified into a Destination Vector by inserting a 1.8-kb cassette that contains all the necessary Gateway sequences (Hartley *et al*., 2000; Walhout *et al*., 2000c). The power of the Gateway cloning system is that a collection of ORFs "stored" as Gateway Entry clones can be transferred in parallel to one or more Destination Vectors in a simple in vitro reaction that can take place in 96-well or 384-well plates.

5. The *C. elegans* ORFeome project

We implemented this Gateway cloning to generate a collection of cloned ORFs for further functional genomics and proteomics studies (Reboul *et al*., 2001; Reboul *et al*., 2003; Rual *et al*., (2004b)). The model system chosen was the nematode *C. elegans* because of the relative simplicity of its genome (e.g., small introns and

short intergenic sequences) and the high quality of its genome sequence (The *C. elegans* Sequencing Consortium, 1998). On the basis of gene annotations and gene prediction tools, the *C. elegans* genome is expected to contain >19 000 protein-encoding genes. PCR reactions were performed with 19 477 pairs of Gateway-tailed ORF-specific primers using as templates a highly representative cDNA library generated from all stages of *C. elegans* development (Walhout *et al*., 2000b).

After Gateway cloning, the resulting PCR products into the Donor vector pDONR201, pools of transformants for each ORF Entry clone were subjected to ORF sequence tags (OSTs) analysis to verify both gene identity and the presence of at least one splicing event. Such ORF pooling strategy allows the maintenance of a diversity of splice variants in each Entry clone (i.e., between the Start and Stop primers) while providing experimental verification of exon–intron structure for each cloned ORF. In addition, the pooling strategy facilitated the throughput needed for the completion of the project. All ~12 000 successfully cloned and sequenced ORFs were then rearrayed to generate "version 1.1" of the *C. elegans* ORFeome (Reboul *et al*., 2003) with pertinent data on each cloned ORF available via a searchable website at **http://worfdb.dfci.harvard.edu** (Vaglio *et al*., 2003).

6. Genome annotation and ORFeome resources

Genome annotation is still a very dynamic process, as more experimental data is obtained to verify both gene structure and expression. The *C. elegans* community has actively maintained a database that regularly updates the gene annotation (Stein *et al*., 2001). Originally, version 1.1 of the *C. elegans* ORFeome was obtained using the August 1999 WormBase genome annotation version WS9, in which ~50% of ORFs were accurately predicted, leaving over one-half of all ORFs needing improved annotation (Reboul *et al*., 2003). One consequence of this was that a significant fraction of the ORFs could not be amplified owing to mispredictions leading to erroneous primer design (Reboul *et al*., 2001; Reboul *et al*., 2003). Recently, we utilized an updated version, WS100 (May 2003), to redesign primers for 3000 newly predicted or repredicted ORFs, and obtained ~2000 of these corrected ORFs. Through this iterative process, we now have ~13 000 ORFs in ORFeome, Version "3.1" which includes the 11 000 original cloned ORFs plus 2000 corrected or newly predicted ORFs (Lamesch *et al*., 2004). As the *C. elegans* genome annotations keep improving, for example, from the comparison with the genome sequence of related nematodes such as *C. briggsae* (Stein *et al*., 2003), better ORFeomes will become available and, consequently, we will approach a finishing point of our ORFeome resource.

Gateway cloning of ORFs is not restricted to eukaryotic cDNAs. Bacterial genomes are also amenable to using Gateway for generating ORFeome collections. However, while exon–intron structure is not an issue for bacterial genomes, there are still significant challenges in gene prediction, namely, in the start site for individual ORFs. For example, construction of a *Brucella melitensis* ORFeome required manual curation to identify and correct the start site of translation for over one-quarter of the 3198 predicted ORFs (Dricot *et al*., 2004). On the basis of both gene prediction tools and manual curation, an ORFeome cloning project

using Gateway was able to obtain 3091 ORFs from this organism (Dricot *et al*., 2004).

7. Improved yeast two-hybrid for large-scale protein interaction mapping

With the availability of Gateway-cloned *C. elegans* ORFs, we have carried out protein–protein interaction mapping studies using a highly reliable, high-throughput yeast two-hybrid (HT-Y2H) system (Walhout *et al*., 2000a; Walhout and Vidal, 2001). The canonical and most commonly employed Y2H system is based on the reconstitution of a transcription factor (e.g., Gal4) (Fields and Song, 1989). DNA sequences coding for proteins of interest X and Y are separately fused to sequences coding for the Gal4 DNA binding (DB) and activation domains (AD), respectively, to create sequences coding for "DB-X" and "AD-Y" hybrid proteins. HT-Y2H screening has been widely used for protein interaction mapping because it manipulates DNA not proteins, making it amenable to both standardization and high throughput.

Earlier versions of the Y2H were inconsistent because they displayed (1) high numbers of false-positives, thereby generating a high background, (2) no direct link between physical interaction and function, and (3) limited automation. In response to these issues, an improved version of the Y2H that significantly lowers the rate of false-positives and is amenable to automation has been developed (Vidal, 1997; Walhout and Vidal, 2001; Vidalain *et al*., 2004) and utilized for our worm protein–protein interaction, or "interactome", mapping project (Li *et al*., 2004).

8. The *C. elegans* interactome project: modular screens

Our *C. elegans* interactome map started with medium-throughput Y2H interactome maps generated for distinct biological processes, or "modules": vulval development, protein degradation, DNA damage response, germ-line formation, and dauer formation (Walhout *et al*., 2000c; Davy *et al*., 2001; Boulton *et al*., 2002; Walhout *et al*., 2002; Tewari *et al*., 2004). Our first attempt at interactome mapping involved the orthologs of the pRb and Ras pathways in the worm, both pathways critical for vulval development (Walhout *et al*., 2000c). To generate this map, ~30 ORFs were fused to DB and used to screen a worm AD-cDNA library (Walhout *et al*., 2000c). Approximately 150 potential interactions were identified involving 18 characterized (and named) genes and ~100 uncharacterized ORFs. More than 50% of interactions previously reported in the Ras and pRb pathways were found, and interesting patterns were observed, including clusters of interactions involving interactions conserved in other species ("interologs") and potential global networks of interactions. This analysis contributed to the understanding of Ras and pRb pathways in *C. elegans* as well as in human. Subsequent modular maps examined the 26 S proteasome complex (Davy *et al*., 2001), proteins implicated in the DNA damage response (DDR) and cell-cycle checkpoint (Boulton *et al*., 2002), proteins

suspected to function in the development of the *C. elegans* germline (Walhout *et al*., 2002), and proteins known to function during larval stages in response to environmental stress that results in a development arrest decision leading to dauer formation instead of normal hermaphrodite adults (Tewari *et al*., 2004). These module-scale networks were shown to (1) contain valuable information to formulate biological hypotheses for large numbers of uncharacterized proteins, (2) provide new potential links between proteins already known to be involved in certain biological processes, and (3) point to unexpected functional relationships between seemingly distinct biological processes (*see* Article 46, **Functional classification of proteins based on protein interaction data**, Volume 5).

9. The *C. elegans* interactome project: multicellularity and large-scale screens

To investigate the fraction of the worm interactome most likely important to multicellularity, we selected a set of ~3000 predicted proteins as Y2H baits, on the basis of structural and functional criteria that relate to multicellularity (Li *et al*., 2004). The proteins selected for this analysis were grouped in three classes: ~350 proteins with a clear ortholog in unicellular yeast ("Ancient"), 2200 proteins with a clear ortholog in multicellular organisms including *Drosophila*, *Arabidopsis*, or humans but not in yeast ("Multicellular"), and 400 proteins with no obvious ortholog outside of *C. elegans* ("Worm"). Out of these ~3000 selected baits, ~2000 were present in the ORFeome version 1.1. Of these, a few autoactivated the Y2H *GAL1::HIS3* reporter gene as DNA binding fusion baits (DB-X), or conferred toxicity to yeast cells, and so were eliminated from the population of baits used for subsequent Y2H analysis.

The remaining DB-X baits were screened against two different Gal4 activation domain (AD-Y) libraries, each with distinct yet complementary advantages. The AD-ORFeome 1.0 library (generated by pooling all ~11 000 cloned ORFs from the ORFeome resource (Reboul *et al*., 2001; Reboul *et al*., 2003)) is normalized (each ORF present in equal abundance) but is limited because many biologically relevant protein–protein interactions cannot be detected by Y2H assay when strictly using full-length proteins (Legrain and Selig, 2000; Uetz *et al*., 2000). In contrast, AD-wrmcDNA (Walhout *et al*., 2000b) is *not* normalized, but it contains multiple alternative AD-Y junction sequences for most genes, and so encodes multiple domain arrangements for each protein represented. The combined AD-ORFeome 1.0 and AD-wrmcDNA libraries contain an estimated 14 000 AD-Y fusions (collapsing splice variants into single genes) (Reboul *et al*., 2003).

Specificity of the Y2H can be maximized by applying stringent experimental and bioinformatics criteria (Vidalain *et al*., 2004). First, three different Gal4-responsive promoters are used to minimize nonspecific promoter activation (Vidal, 1997) and AD-Y clones are only considered if they activate at least two of the three promoters. Second, expressing both DB-X and AD-Y from single-copy, not multicopy, plasmids reduces spurious effects due to overexpression of bait and/or prey. Third, all interaction pairs are retested in fresh yeast cells to eliminate

spontaneously arising auto-activators and to confirm the reproducibility of the interaction. Finally, Y2H interactor-encoding cDNAs or ORFs are PCR-amplified directly from yeast colonies and sequenced, giving rise to Interaction Sequence Tags (ISTs) (Walhout *et al*., 2000b). For each IST, the AD-Y reading frame is verified to avoid recovering out-of-frame peptides. For the *C. elegans* Interactome (Li *et al*., 2004), employing ∼2000 baits, ∼16 000 ISTs were initially obtained. The interactions between bait and prey were then subdivided into three confidence classes: those found at least three times independently and for which at least one IST revealed an in-frame AD-Y junction (Subset "Core-1"); those found fewer than three times but passed the retest and for which at least one IST revealed an in-frame AD-Y junction (Subset "Core-2"); and all other Y2H interactions found in our screens (Subset "Non-Core"). By these criteria, and combined with the interactions found in our modular screens and through interologs, the current interactome map is a network of 5500 high-quality interactions (Li *et al*., 2004).

10. Properties of the *C. elegans* interactome network

The current *S. cerevisiae* interactome is a scale-free network, a topology that might relate to genetic robustness (Jeong *et al*., 2001; Gu *et al*., 2003). We used the WI5 Core dataset to investigate the topology of the *C. elegans* interactome. As bait proteins have more interactors than prey proteins on average, the analysis on degree distribution should be conducted separately for bait and prey proteins to avoid bias. When the proportion of *C. elegans* proteins that have k interactors ($P(k)$) is plotted against k on a log–log scale, a linear curve is obtained, similar to that seen for *S. cerevisiae* (Li *et al*., 2004). Thus, the worm interactome also exhibits a "scale-free", or power law, degree distribution. In addition, the worm interactome network exhibits small-world properties. These observations extend previous findings that many other biological networks also exhibit small-world and scale-free properties (Wagner, 2001; *see also* Article 27, **Noncoding RNAs in mammals**, Volume 3).

This WI5 Y2H dataset also allows us to analyze whether or not evolutionary recent proteins tend to preferentially interact with each other rather than with ancient proteins. Subdividing the nodes of the network into "ancient", "multicellular", and "worm" proteins as defined above, the three groups seem to connect equally well with each other, suggesting that new cellular functions rely on a combination of both evolutionary new and ancient components (Li *et al*., 2004; *see also* Article 37, **Functional analysis of genes**, Volume 3).

11. Accuracy of the *C. elegans* interactome map

ISTs identified in any Y2H study can arise due to the nature of Y2H itself rather than a bona fide interaction between two polypeptides. These "biological false-positives" can appear to be real-positives after retesting in yeast but are shown to be false-positives when tested in an entirely different assay system. Thus, even with stringent criteria for identifying putative interacting proteins via

Y2H, it is better to experimentally assess the overall quality of Y2H results in an alternative assay. To experimentally assess the overall quality of our HT-Y2H scheme, a representative sample of interaction pairs from each of the three datasets, Core-1, Core-2, and Non-Core, was tested in an in vivo coaffinity purification (co-AP) Glutathione-S-Transferase (GST) pull-down assay. Interactions detected in a completely different protein interaction assay are unlikely to be technical false-positives. ORF pairs corresponding to Y2H interactions were retrieved from our ORFeome 1.1 resource (Reboul *et al*., 2003) and Gateway-transferred into two mammalian expression vectors, in frame with sequences encoding either the GST or the Myc Tag, for DB-X baits (GST-X) and AD-Y preys (Myc-Y), respectively (Xu *et al*., 2003). In cotransfection experiments for which both GST-X and Myc-Y fusion proteins were expressed at detectable levels, the GST pull-down success rates were 82% for Core-1, 59% for Core-2, and 35% for Non-Core.

Since co-AP GST pull-down assays have a false-negative rate of their own owing to technical limitations different from Y2H (we only used a single binding condition, and proteins have been expressed from a single set of vectors), these results demonstrate that our datasets contain a large proportion of highly reliable interactions and confirm their relative quality. To our knowledge, this is the first time that the overall quality of a proteome-wide Y2H interactome mapping dataset was experimentally validated using an independent protein interaction assay.

Finally, until there is a precise and dynamic view of the whole proteome of an organism of interest, it will be virtually impossible to demonstrate that any two proteins found by Y2H do not interact in vivo. Thus, biological false-positives are extremely difficult to define and detect. However, increasing evidence for biologically genuine Y2H interactions can be obtained by accumulating experimental evidence emerging from other functional genomic or proteomic approaches (Vidal, 2001; Ge *et al*., 2003).

12. Overlap between interactome and transcriptome or phenome

Previous studies related interactome data to genome-wide expression (transcriptome) and phenotypic profiling (phenome) data in *S. cerevisiae* (Ge *et al*., 2001). However, would such correlations between different functional genomic datasets also hold for a multicellular organism, and could they help formulate biological hypotheses? To investigate this, we attempted to integrate our *C. elegans* interactome data with currently available transcriptome and phenome datasets.

Pearson correlation coefficients (PCCs), a metric used to assess a linear relationship between variables in two independent datasets, were calculated from the expression profiles described in the *C. elegans* transcriptome "Topomap" dataset (Kim *et al*., 2001) then compared between gene pairs involved in experimentally derived interactions in WI5 and gene pairs from randomized datasets. Approximately, 150 Core Y2H interactions ($\sim$10%) correspond to gene pairs with significantly higher PCCs than expected from random (p-value < 0.05). Those pairs can be considered as "more biologically likely" since two independent orthogonal approaches point to a functional relationship between the corresponding genes.

The remaining pairs are considered as "without corroborating evidence" by transcriptome analysis. Lack of coexpression does not suggest that the corresponding interactions are erroneous. Indeed, 75% of "literature" pairs, taken from published low-throughput interactions and hence considered biologically relevant, do not correlate with transcriptome data.

These results underscore some of the complexities inherent in attempting to overlap functional genomic data sets derived from metazoans. Unlike unicellular organisms, in metazoans, biological processes may occur differentially not just in single cells but also across different organs or tissues.

13. Functional and directional wiring diagrams of protein networks

Although combined functional genomic and proteomic maps provide biological hypotheses of gene function for many as yet uncharacterized gene products predicted from genome sequencing projects, they fail to provide a functional and directional model for protein networks. In other words, interactome maps do not indicate the functional consequences of the physical interactions detected. To address these issues, it will be necessary to carry out a systematic perturbation approach that consists of the generation of double knockouts (by RNAi knockdown, for example) of all pairwise combinations for a subgroup of genes suspected to be involved in a particular biological process.

Using *C. elegans* dauer formation, a developmental arrest decision during larval development in response to environmental stress, as an example, we tested the technologies required to generate functional and directional wiring diagrams of gene networks. After combining interactome, transcriptome, and phenome maps of the dauer regulatory network, starting from proteins known to be involved in the process, all pairwise combinations of double perturbations were created. We scored synthetic genetic interactions (i.e., double knockout gives rise to a more severe phenotype than either one alone) and epistatic relationships (two single knockouts giving rise to unlike phenotypes and the double gives rise to only one of these two phenotypes). We detected many such genetic interactions, and by mapping them into a wiring diagram, we are able to generate novel hypotheses about the dauer network (Tewari *et al*., 2004).

Thus, interactome maps can be used as a scaffold for the discovery of functional relationships between components of a particular biological process.

14. Toward a comprehensive analysis of biological networks

The development of technologies such as Gateway and HT-Y2H for manipulating large numbers of genes on the basis of the information obtained from genome sequencing and gene annotation efforts will enable scientists to begin asking biological questions related to how networks control developmental processes in normal and disease circumstances. However, this will be attainable only through

concerted efforts to create ORFeome collections for a wide variety of organisms, especially humans, and then utilize those ORFeomes for functional studies such as Y2H. The construction of the *C. elegans* ORFeome and its utilization for interactome mapping has provided a glimpse into how these efforts can be conducted for human studies and the challenges inherent in both creating and using such resources. Integration of information from multiple experimental platforms will be necessary to develop models of biological networks that can be both predictive and open to experimental manipulation.

References

Begley TJ, Rosenbach AS, Ideker T and Samson LD (2002) Damage recovery pathways in *Saccharomyces cerevisiae* revealed by genomic phenotyping and interactome mapping. *Molecular Cancer Research*, **1**(2), 103–112.

Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, Hill DE and Vidal M (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science*, **295**(5552), 127–131.

Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, *et al*. (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Science*, **11**(4), 723–738.

Dandekar T, Snel B, Huynen M and Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, **23**(9), 324–328.

Davy A, Bello P, Thierry-Mieg N, Vaglio P, Hitti J, Doucette-Stamm L, Thierry-Mieg D, Reboul J, Boulton S, Walhout AJ, *et al*. (2001) A protein-protein interaction map of the *Caenorhabditis elegans* 26 S proteasome. *EMBO Reports*, **2**(9), 821–828.

Dricot A, Rual J-F, Lamesch J-F, Bertin N, Dupuy D, Hao T, Lambert T, Hallez T, Delroisse J-M, Vandenhaute J, *et al*. (2004) Cloning of the *brucella melitensis* ORFeome version 1.1. *Genome Research*, **14**(10b), 2201–2206.

Dupuy D, Li Q, Deplancke B, Boxem M, Hao T, Lamesch P, Hope IA, Hill DE, Walhout AJM and Vidal M (2004) The *C. elegans* promoterome 1.0: a gateway to large-scale expression mapping. *Genome Research*, **14**(10b), 2169–2175.

Fields S and Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M and Ahringer J (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature*, **408**(6810), 325–330.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Ge H, Liu Z, Church GM and Vidal M (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics*, **29**(4), 482–486.

Ge H, Walhout AJ and Vidal M (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics*, **19**(10), 551–560.

Goldberg DS and Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4372–4376.

Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, *et al*. (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature*, **408**(6810), 331–336.

Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, **29**(17), 3513–3519.

Grigoriev A (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, **31**(14), 4157–4161.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW and Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**(6918), 63–66.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, WalhoutAJ, Cusick ME, Roth FP, *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hanazawa M, Mochii M, Ueno N, Kohara Y and Iino Y (2001) Use of cDNA subtraction and RNA interference screens in combination reveals genes required for germ-line development in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(15), 8686–8691.

Hartley JL, Temple GF and Brasch MA (2000) DNA cloning using in vitro site-specific recombination. *Genome Research*, **10**(11), 1788–1795.

Hengartner MO and Horvitz HR (1994) *C. elegans* cell survival gene *ced-9* encodes a functional homolog of the mammalian proto-oncogene bcl-2. *Cell*, **76**(4), 665–676.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**(6868), 180–183.

Hope IA, Albertson DG, Martinelli SD, Lynch AS, Sonnhammer E and Durbin R (1996) The *C. elegans* expression pattern database: a beginning. *Trends in Genetics*, **12**, 370–371.

Ideker T, Galitski T and Hood L (2001) A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, **2**, 343–372.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Jansen R, Greenbaum D and Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Research*, **12**(1), 37–46.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, *et al*. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**(6920), 231–237.

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A and Holstege FC (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, **9**(5), 1133–1143.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN and Davidson GS (2001) A gene expression map for *Caenorhabditis elegans*. *Science*, **293**(5537), 2087–2092.

Kitano H (2002) Computational systems biology. *Nature*, **420**(6912), 206–210.

Lamesch P, Milstein S, Hao T, Rosenberg J, Li N, Sequerra R, Bosak S, Doucette-Stamm L, Vandenhaute J, Hill DE, *et al*. (2004) *C. elegans* ORFeome Version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Research*, **14**(10b), 2064–2069.

Legrain P and Selig L (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Letters*, **480**(1), 32–36.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**(5657), 540–543.

Luan C-H, Qiu S, Finley JB, Carson JB, Gray RJ, Huang W, Johnson D, Tsao J, Reboul J, Vaglio P, *et al*. (2004) High-throughput expression of *C. elegans* proteins. *Genome Research*, **14**(10b), 2102–2110.

Maeda I, Kohara Y, Yamamoto M and Sugimoto A (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Current Biology*, **11**, 171–176.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**(6757), 83–86.

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S and Vidal M (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, **11**(12), 2120–2126.

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Weil B (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, **30**(1), 31–34.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), 4285–4288.

Pennisi E (1998) Worming secrets from the *C. elegans* genome. *Science*, **282**(5396), 1972–1974.

Piano F, Schetter AJ, Mangone M, Stein L and Kemphues KJ (2000) RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Current Biology*, **10**(24), 1619–1622.

Praitis V, Casey E, Collar D and Austin J (2001) Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics*, **157**(3), 1217–1226.

Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, *et al*. (2003) *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics*, **34**(1), 35–41.

Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J, *et al*. (2001) Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genetics*, **27**(3), 332–336.

Riddle DL, Blumenthal T, Meyer BJ and Priess JR (1997) *C. elegans II*, Cold Spring Harbor Laboratory Press: Plainview, NY.

Rual J-F, Ceron J, Koreth J, Hao T, Nicot A-S, Hirozane-Kishikawa T, Vandenhaute J, Orkin SH, Hill DE, van den Heuvel S, *et al*. (2004a). Genome-wide analysis using the *C. elegans* ORFeome-RNAi library. *Genome Research*, **14**(10b), 2162–2168.

Rual JF, Hill DE and Vidal M (2004b) ORFeome projects: gateway between genomics and omics. *Current Opinion in Chemical Biology*, **8**(1), 20–25.

Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**(12), 1257–1261.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al*. (2003) The genome sequence of caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biology*, **1**(2), E45.

Stein L, Sternberg P, Durbin R, Thierry-Mieg J and Spieth J (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research*, **29**(1), 82–86.

Sternberg PW and Han M (1998) Genetics of RAS signaling in *C. elegans*. *Trends Genet*, **14**(11), 466–472.

Sulston JE and Horvitz HR (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology*, **56**(1), 110–156.

Sulston JE, Schierenberg E, White JG and Thomson JN (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, **100**(1), 64–119.

Tewari M, Hu PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, *et al*. (2004) Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. *Molecular Cell*, **13**(4), 469–482.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology *Science*, **282**(5396), 2012–2018.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**(6770), 623–627.

Vaglio P, Lamesch P, Reboul J, Rual JF, Martinez M, Hill D and Vidal M (2003) WorfDB: the *Caenorhabditis elegans* ORFeome database. *Nucleic Acids Research*, **31**(1), 237–240.

Vidal M (1997) The reverse two-hybrid system. In *The Yeast Two-Hybrid System*, Bartels P and Fields S (Eds.), Oxford University Press: New York, pp. 109–147.

Vidal M (2001) A biological atlas of functional maps. *Cell*, **104**(3), 333–339.

Vidalain PO, Boxem M, Ge H, Li S and Vidal M (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, **32**(4), 363–370.

Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, **18**(7), 1283–1292.

Walhout AJ, Boulton SJ and Vidal M (2000a) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast*, **17**(2), 88–94.

Walhout M, Endoh H, Thierry-Mieg N, Wong W and Vidal M (1998) A model of elegance. *American Journal of Human Genetics*, **63**(4), 955–961.

Walhout AJ, Reboul J, Shtanko O, Bertin N, Vaglio P, Ge H, Lee H, Doucette-Stamm L, Gunsalus KC, Schetter AJ, *et al*. (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Current Biology*, **12**(22), 1952–1958.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N and Vidal M (2000b) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**(5450), 116–122.

Walhout AJ, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S and Vidal M (2000c) GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods in Enzymology*, **328**, 575–592.

Walhout AJ and Vidal M (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, **24**(3), 297–306.

Xu L, Wei Y, Reboul J, Vaglio P, Shin TH, Vidal M, Elledge SJ and Harper JW (2003) BTB proteins are substrate-specific adaptors in an SCF-like modular ubiquitin ligase containing CUL-3. *Nature*, **425**(6955), 316–321.

Yuan JY and Horvitz HR (1990) The *Caenorhabditis elegans* genes *ced-3* and *ced-4* act cell autonomously to cause programmed cell death. *Developmental Biology*, **138**(1), 33–41.

Yuan J and Horvitz HR (2004) A first insight into the molecular mechanisms of apoptosis. *Cell*, **116**(Suppl 2), S53–S56.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**(5537), 2101–2105.

Ziauddin J and Sabatini DM (2001) Microarrays of cells expressing defined cDNAs. *Nature*, **411**(6833), 107–110.

Specialist Review

The yeast interactome

Peter Uetz
Institute of Genetics (ITG), Forschungszentrum Karlsruhe, Karlsruhe, Germany

Andrei Grigoriev
GPC Biotech, Martinsried, Germany

1. Introduction

The genome of the budding yeast *Saccharomyces cerevisiae* was the first eukaryotic genome to be sequenced. Therefore, yeast was also the first eukaryote whose complete set of proteins became known in 1996 (Goffeau *et al*., 1997). This led to the first systematic attempts to study protein interactions (the "interactome") and function at the end of the twentieth century. As a consequence, the yeast proteome and interactome are probably the best known of all eukaryotes, including human. As a guide to understanding a cell, the knowledge of its "interactome" is absolutely necessary as protein interactions are critical to all processes in biology. Here we summarize current knowledge of the yeast interactome and some of its implications for other model systems.

1.1. *Types of interaction data: sources and classification*

Historically, that is, until the yeast genome sequence became available, proteins and their interactions were studied one by one as they were identified in genetic screens or by other methods. However, the availability of the genome sequence allowed the study of all yeast proteins in a systematic way and as a consequence, the majority of our current knowledge of the yeast interactome stems mostly from a few large-scale studies. These studies included the yeast *two-hybrid* assay (Figure 1) and the purification of protein complexes and their analysis by *mass spectrometry* (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5 and Article 36, **Biochemistry of protein complexes**, Volume 5). More recently, *genetic screens* also yielded a large number of interactions, but it remains unclear how many of those represent physical interactions (see below for details). Nevertheless, detailed "small-scale" studies remain extremely important as they still provide the only means for careful functional analysis and the placement of interaction data in a biological context.

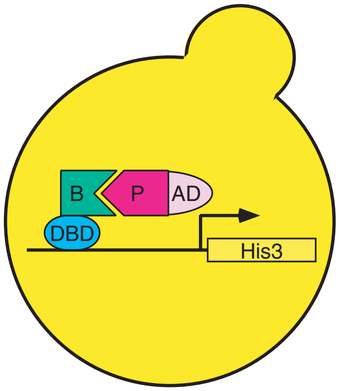


Figure 1 *The two-hybrid system* is based on the expression of two fusion (i.e., two hybrid) proteins in a cell. One of the fusion proteins contains a DNA-binding domain (DBD), which can bind to the promoter of a reporter gene (here: His3), and a "B" protein ("bait"). The second fusion protein consists of a transcription activation domain and an ORF protein (open reading frame, any protein). If B interacts with ORF, a transcription factor is formed, which can switch on the reporter gene. This enables the cell to grow on a histidine-free medium (the endogenous His3 gene has to be deleted). Hence, a growing yeast colony indicates that the two expressed proteins interact. If they do not interact, no colony is formed

2. Methods for interaction analysis and their data

2.1. *Two-hybrid interactions*

The two-hybrid system uses two hybrid proteins, which reconstitute a transcription factor when they interact. This transcription factor can switch on a reporter gene (Figure 1). Two-hybrid interactions are usually binary although occasionally a third protein may mediate interactions between the two hybrids. For yeast, there are about 7000 two-hybrid protein interactions available in databases derived from small and large-scale screens (Table 1).

The first large-scale study (Uetz *et al*., 2000) included two independent screens: one tested 192 baits for interactions with all 6000 yeast proteins and found 281 reproducible interactions for 87 proteins. In the second screen, all ~6000 activation domain constructs were pooled and tested against each DNA-binding domain fusion (Figure 1), resulting in 692 interactions involving 817 proteins.

Ito *et al*. (2001) presented the largest two-hybrid interaction screen to date, which detected 4549 interactions among 3278 proteins. Of these, 841 were found four or more times and they represented a "core" subset of more reliable interactions.

2.2. *Protein complex purification and analysis by mass spectrometry*

Often, proteins interact within stable protein complexes (*see* Article 35, **Structural biology of protein complexes**, Volume 5 and Article 36, **Biochemistry of protein**

Table 1 Large-scale and examples of smaller studies of protein–protein interactions in yeast

Method (proteins)	Interactions	References
Y2H	~4500	Ito *et al*. (2001)
Y2H	~1000	Uetz *et al*. (2000)
Y2H (cell polarity)	191	Drees *et al*. (2001)
Y2H (RNA splicing)	170	Fromont-Racine *et al*. (1997)
Essential unknown proteins	271	Hazbun *et al*., 2003
Y2H + PD (SH3 domain)	394 + 233[a]	Tong *et al*. (2002)
Y2H (PX domain)	75	Vollert and Uetz (2004)
Purification + MS	~18 000[b]	Gavin *et al*. (2002)
Purification + MS	~33 000[b]	Ho *et al*. (2002)
Synthetic lethality	~4000	Tong *et al*. (2004)
Combined large- and small-scale collections	~15 700	From DIP database
	~12 700	From MINT database
	~15 500	From MIPS database

[a]59 interactions were common to both approaches and therefore considered as highly likely. PD = phage display; MS = mass spectrometry.
[b]Counting all-against-all possible interactions between complex members in filtered datasets (matrix model), actual numbers of protein–protein interactions cannot be determined from the data.

complexes, Volume 5). These complexes can be purified and their components identified. In contrast to two-hybrid data, this provides no or little information on which subunit interacts with which other subunit. Two-hybrid data, on the other hand, do not tell which interactions lead to a stable complex (Figure 2).

Two large-scale studies identified protein complexes in yeast. Gavin *et al*. (2002) analyzed 1739 genes, including 1143 human orthologs, and identified 232 distinct multiprotein complexes using tandem affinity purification (TAP) (Figure 3). This set of proteins was not representative as it contained many highly conserved proteins and so we cannot conclusively extrapolate these numbers to the whole yeast proteome. However, these results suggest that there may be no more than 400 or 500 complexes in yeast, although many of them cannot be precisely defined as some of their subunits may vary.

A more conclusive number may become available through the analysis by Emili *et al*. (personal communication), who have continued the analysis of Ho *et al*. and purified a total of 4455 TAP-tagged baits of which more than 3400 resulted in successful purifications. However, an analysis of this dataset was not available at the time of this writing.

Ho *et al*. (2002) chose an initial set of 725 bait proteins and identified the proteins associated with these baits using a Flag epitope tag in the high-throughput mass spectrometric protein complex identification (HMS-PCI). After removal of potential false positives, this filtered data set contained 1578 different interacting proteins.

Two models are used to map complexes to pairwise interactions. The *spoke model* assumes that the bait protein interacts with all other proteins in the complex. The *matrix model* assumes that every protein in a complex interacts with every other protein, that is, there are $N(N-1)/2$ interactions in a complex containing

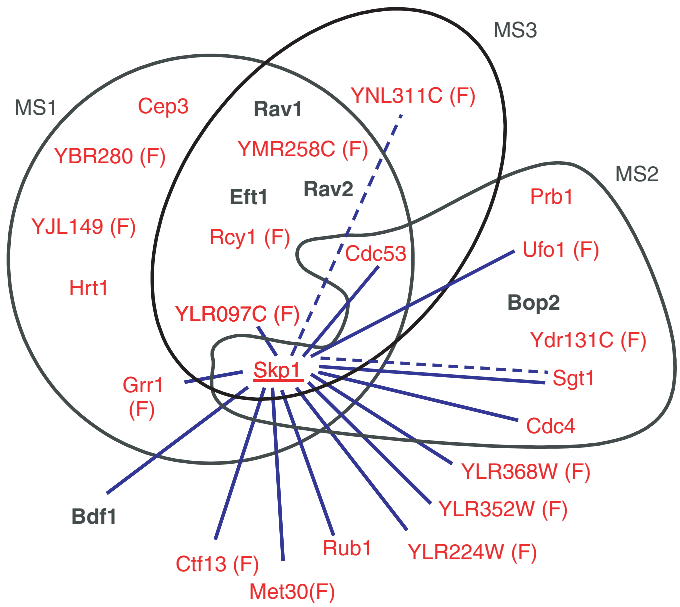


Figure 2 *Two-hybrid versus protein complex data.* Skp1 is a protein involved in ubiquitin-mediated protein degradation and has been used as bait for two-hybrid screens and epitope-tagged for mass spectrometry analysis. The purified complexes of Skp1 from three independent MS studies (MS1–3) and the binary interactions from two Y2H studies (solid and broken blue lines) are compared. Despite the differences in the data sets, most of the interactions found seem to be plausible: All proteins colored red are known to be involved in protein degradation. Skp1 is directed to its target proteins via so-called F-box proteins, which contain a short peptide motif, the F-box (F). For details, see Titz *et al*. (2004)

N proteins. These models represent theoretical minimum and maximum numbers of interactions and are both correct only if a complex is a heterodimer. The real topology of the interactions in a complex is most likely somewhere in between these extremes. When compared to a literature benchmark, one of the matrix models had turned out to be three times more incorrect than the spoke (Bader and Hogue, 2002), although other studies have not found any significant difference (Cornell *et al*., 2004).

A graphical representation of complexes yields a fairly intricate display of interconnections between individual bait-target groups (Figure 4).

2.3. *Genetic interactions and synthetic lethality*

Single mutations often do not affect the viability of a cell significantly. However, if two or more such mutations are combined, this may have a much more severe effect than expected from the individual mutants or may even cause lethality. Genes that exhibit such a synthetic phenotype are said to interact genetically. In many cases, the products of these genes are components of a complex, a pathway, or are otherwise functionally related (e.g., compensate each other’s function). Therefore,

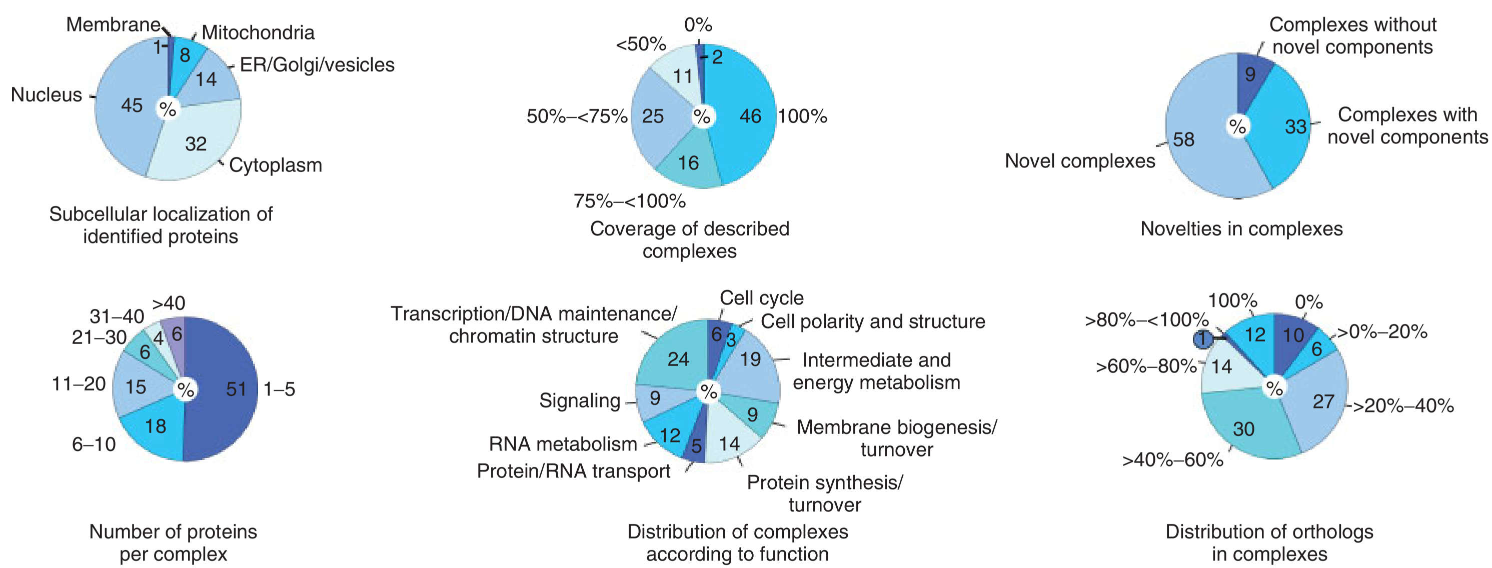


Figure 3 *Statistics of proteins and complexes.* Numbers inside pie charts represent the percentages of total proteins (a) and complexes (b–f) based on the data published by Gavin *et al*. (2002). Outer labels show partitioning of the data according to the chart function. (b) indicates the coverage of protein complexes when compared to the known complexes, as cataloged in the Yeast Protein Database (YPD). For example, 46% of the complexes contained all known (100%) components of these complexes. (f) shows the conservation of proteins in complexes as defined by the presence of orthologs. For example, 27% of the complexes contained 20–40% human orthologs and 10% of the complexes did not contain any human orthologs. (Reprinted with permission from Nature (Gavin *et al*.) ©2002 Macmillan Magazines Limited)

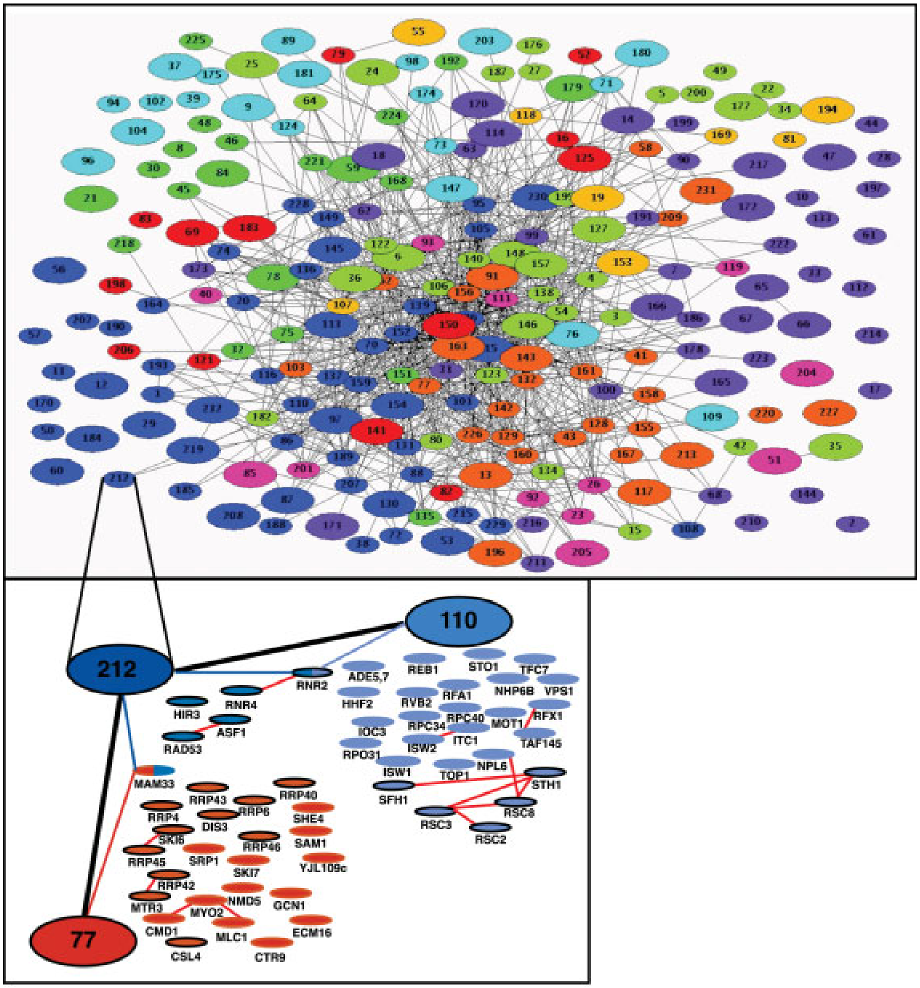


Figure 4 *The protein complex network, and grouping of connected complexes.* Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing the distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes are color coded: red, cell cycle; dark green, signaling; dark blue, transcription, DNA maintenance, chromatin structure; pink, protein and RNA transport; orange, RNA metabolism; light green, protein synthesis and turnover; brown, cell polarity and structure; violet, intermediate and energy metabolism; light blue, membrane biogenesis and traffic. The lower panel is an example of a complex (yeast TAP-C212) linked to two other complexes (yeast TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. Red lines indicate physical interactions as listed in YPD22. (Reprinted with permission from Nature (Gavin *et al.*) ©2002 Macmillan Magazines Limited)

such combined mutations often lead to an additive or even stronger effect. However, the proteins do not need to interact physically in order to show such a synthetic phenotype.

Tong *et al*. (2004) presented the largest analysis of synthetic lethal interactions so far. Their 132 SGA (i.e., "synthetic genetic analysis") screens focused on query genes involved in actin-based cell polarity, cell wall biosynthesis, microtubule-based chromosome segregation, and DNA synthesis and repair, detecting 4000 interactions amongst 1000 genes. It suggests that the yeast synthetic genetic network may contain up to 100 000 interactions although 24 000 appear more realistic based on the biased choice of genes studied. Approximately 20% of the query genes (not included in the 132 query genes mentioned above) showed no genetic interactions. The frequency of false negatives was estimated to be in the range of 20–40%.

2.4. *Other physical methods*

Protein interactions have been identified using a number of different methods including coimmunoprecipitation, protein arrays (*see* Article 24, **Protein arrays**, Volume 5), FRET (*see* Article 51, **FRET-based reporters for intracellular enzyme activity**, Volume 5), or structural biology (Section (7)). However, none of these methods has been used on a large scale to identify protein interactions although such projects are under way, especially in structural genomics (*see* Article 97, **Structural genomics – expanding protein structural universe**, Volume 6) and using protein arrays (Zhu *et al*., 2001; *see also* Article 24, **Protein arrays**, Volume 5).

3. Data quality

Many interactions from large-scale studies are not as rigorously quality controlled as from small-scale experiments. This is true for both two-hybrid and mass spectrometry analyses. Various studies estimated that two-hybrid data may have 50–90% false positives (Mrowka, 2001; Sprinzak *et al*., 2003), while mass spectrometry datasets may have >30% false positives (Gavin *et al*., 2002). The reason for two-hybrid false positives is not well understood but it somehow derives from unspecific activation of the reporter genes used. Mass spectrometry may result in false positives if the complexes are not highly purified. On the other hand, stringent purification may result in false negatives as subunits are lost. However, smaller-scale screens also detected interactions which are not in agreement with structural data (Edwards *et al*., 2002), indicating a need in the integration of diverse types of data to confirm interactions by other means.

Other approaches to assessment of interaction datasets have included analysis of orthologs (Deane *et al*., 2002; Pagel *et al*., 2004), gene expression (Kemmeren *et al*., 2002), and network topology (Saito *et al*., 2002; Goldberg and Roth, 2003). A comprehensive dataset of putative yeast interactions was collected by von Mering *et al*. (2002) from a large number of sources including two-hybrid and complex pull-down results, smaller-scale studies and in silico predictions (including conserved gene neighborhood, gene fusion events and co-occurrence of genes as well as correlated mRNA expression). The gene pairs collected have been assigned

confidence on the basis of the number of sources of supporting evidence, with the high-confidence set supported by at least two sources including about 2500 interactions between some 1000 proteins.

Interestingly, homologous protein interactions found in two species appear to be more reliable than interactions found by two independent methods (Lehner and Fraser, 2004).

4. Comparison of various interaction datasets

Various high-throughput projects produced surprisingly different results. Biologists need to be aware of these differences as they indicate important limitations of different approaches (Cornell *et al*., 2004).

4.1. Mass spectrometry data

In mass spectrometry approaches, of 115 common baits used by TAP (Gavin) and HMS-PCI (Ho) only 47 retrieved complexes containing common proteins. In 33 other complexes with the same bait, there was no other common member. As a result, there is considerable disparity between the size and content of complexes generated by TAP and HMS–PCI.

The TAP approach was also more successful at identifying reverse interactions than HMS–PCI. In 39% of instances where B is used as a bait, it identified A, compared to 19% for HMS–PCI.

4.2. Two-hybrid data

There are 220 bait proteins in common between the two large-scale screens. For these baits, the Ito dataset contains 871 interactions, while that of Uetz contains 430; 164 interactions were common to both datasets. The Ito dataset contains many highly connected proteins (HCPs) several of which interact with more than 100 preys. Of the 2161 HCP interactions, only 67 (3.1%) have been verified by multiple datasets. Most studies therefore agree that such interactions consist mostly of false positives. Uetz *et al*. (2000) filtered out such HCPs as false positives.

In addition, in the vast majority of cases, unverified interactions were identified with the HCP as the bait protein. In only 221 of the 2161 interactions is the HCP the identified protein.

4.3. Transient versus stable interactions

Aloy and Russell (2002) noticed that mass spectrometry studies appear to identify preferentially stable interactions while two-hybrid studies tend to prefer transient interactions. Transient interactions also occur more frequently in HMS–PCI data than in TAP data, which may be due to a more rigorous washing regime than in the HMS–PCI procedure.

5. Integrating protein–protein interactions with other datasets

The first attempt at integrating the yeast proteome data was the study of a relationship between gene expression and protein interactions (Grigoriev, 2001).

Protein pairs encoded by coexpressed genes were found to interact with each other more frequently than random protein pairs. This approach allows one to evaluate interaction datasets using large-scale gene expression profiling as benchmark. For example, the MIPS collection, with interactions collected from literature and often supported by additional evidence, has shown by far the best agreement with gene expression profiles than large-scale interaction screens. Similarly, the "core" subset from Ito *et al*. (2001) has shown better agreement than the rest of their interactions. These findings were supported by a number of follow-up papers (Ge *et al*., 2001; Jansen *et al*., 2002; Kemmeren *et al*., 2002; Deane *et al*., 2002). Cornell *et al*. (2004) compared protein pairs co-occurring in multiple TAP/HMS–PCI complexes with the pairs detected only once and found the former to display a much stronger correlation with gene expression data.

Note that correlation of mRNA expression is not necessary for protein interaction, that is, proteins whose expression is not correlated strongly may still be bona fide interactors as long as they coexist in sufficient proximity within a cell.

Integration of various datasets remains a tremendous challenge for bioinformatics, especially when interactions have to be displayed graphically together with expression data, homology, structures, or localization in a cell (Uetz *et al*., 2002).

6. The yeast interaction network

When the results of large- and small-scale two-hybrid screens were put together, the resulting network surprisingly connected almost all yeast proteins (Figure 5); only a few proteins resulted in smaller networks that were not linked to the large component (Schwikowski *et al*., 2000). Such interaction networks have led to a flurry of bioinformatic studies, exploring the topology and biological significance of its properties. Here, we summarize only a few interesting results and refer the reader to the literature for more detailed analysis (Barabasi and Oltvai, 2004; *see also* Article 111, **Functional inference from probabilistic protein interaction networks**, Volume 6).

6.1. Computational identification of protein complexes in networks

Several studies have demonstrated that complexes can be found in networks by searching for local clusters of interacting proteins (Bader and Hogue, 2003; Krause *et al*., 2003; Spirin and Mirny, 2003; Bu *et al*., 2003). Given that many proteins are shared by several complexes, these findings blur the distinction between defined complexes and whole proteome interaction networks and emphasize the dynamics of protein interactions within a cell.

6.2. Modularity of the protein interaction network in yeast

Early topological analyses of metabolic networks indicated that they can be considered as modular. The same seems to be true for protein interaction networks

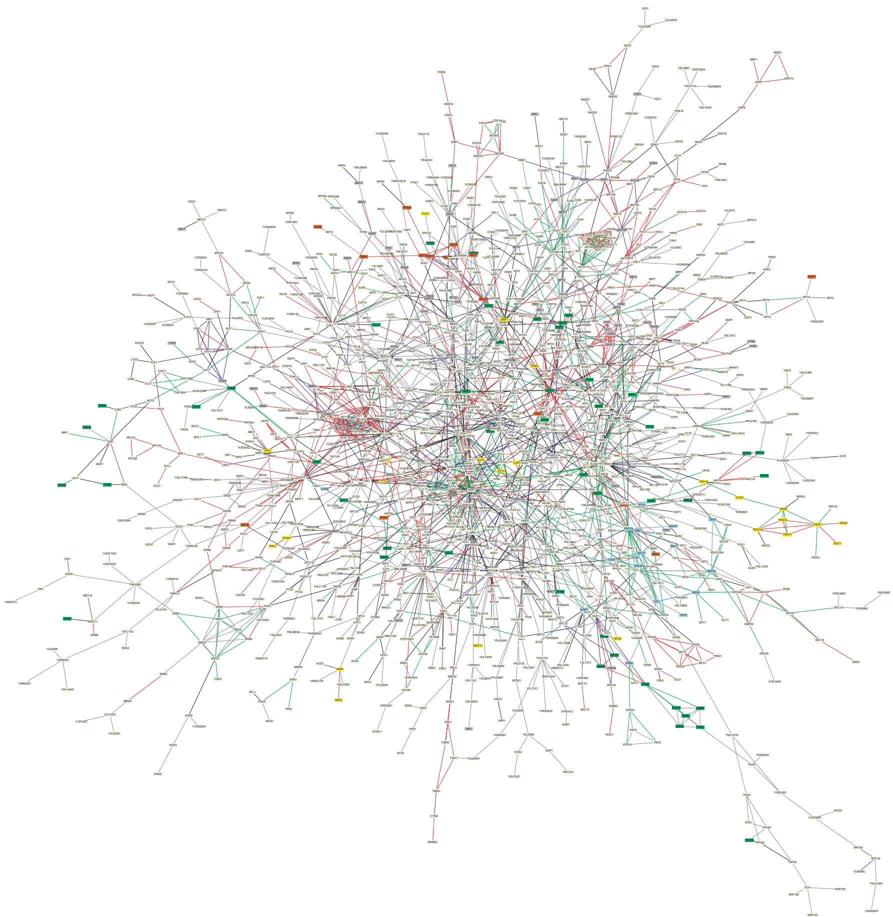


Figure 5 *An interaction map of the yeast proteome assembled from published interactions.* The map contains 1548 proteins and 2358 interactions. Proteins are colored according to their functional role as defined by the Yeast Protein Database (YPD); proteins involved in membrane fusion (blue), chromatin structure (gray), cell structure (green), lipid metabolism (yellow), and cytokinesis (red). After Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1257–1261. (Reproduced by permission of Nature Publishing Group)

(Figure 6). Han *et al*. (2004) identified two classes of highly connected proteins (hubs): *party hubs* (that interact mostly within a protein complex) and *date hubs* (interacting with other hubs, mainly transiently but not in stable complexes).

Averaging Pearson correlation coefficients (PCCs) of mRNA expression between hubs and their interaction partners produced a bimodal distribution for expression data related to various biological functions such as "stress response" and the "cell cycle" (Figure 6b). Other expression groups did not show such a clear bimodal distribution. After excluding ribosomal hub proteins, these two peaks of the distribution contained 91 date (lower PCCs) and 108 party hubs (higher PCCs) in the high-confidence interaction dataset of von Mering *et al*. (2002).

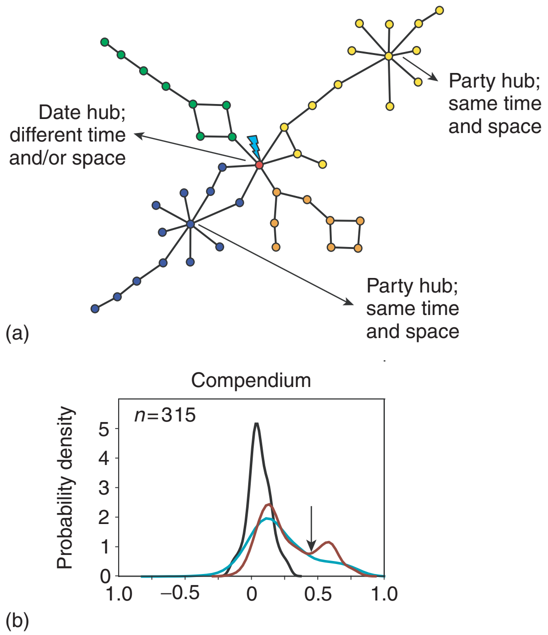


Figure 6 *Date and party hubs*. Probability densities of the average Pearson correlation coefficients (PCCs) were calculated from a global expression profiling compendium. The number n refers to the number of data points for each gene for each condition. No bimodal distribution is observed with the average PCCs of nonhub proteins (cyan curve) or for hubs in randomized networks (black curve). (Reprinted with permission from Nature (Han *et al*.) ©2004 Macmillan Magazines Limited)

Interestingly, changes in the network connectivity resulting from removal of the party hubs (involved in complexes) were similar to those resulting from removal of random proteins, whereas deletion of date hubs produced much more serious changes in network connectivity. Finally, date and party hubs are threefold more likely to be essential than nonhubs, but their individual knockout affects cellular viability to the same extent.

7. Which proteins do interact (or do not?)

As expected, most proteins interact with functionally related and colocalized proteins. In the interaction network (Figure 5), proteins of the same function and cellular location tend to cluster together, with 60–75% of the interactions occurring between proteins with a common functional assignment/same subcellular compartment. When proteins are grouped by functional role or subcellular compartment, interesting patterns of crosstalk between groups can be seen (Figure 7).

However, the tendency to interact at all is different among different functional classes. Signaling proteins or proteins within protein complexes need to interact with multiple proteins while enzymes tend to interact only with few proteins if at all (they rather interact with low-molecular weight substrates). With respect

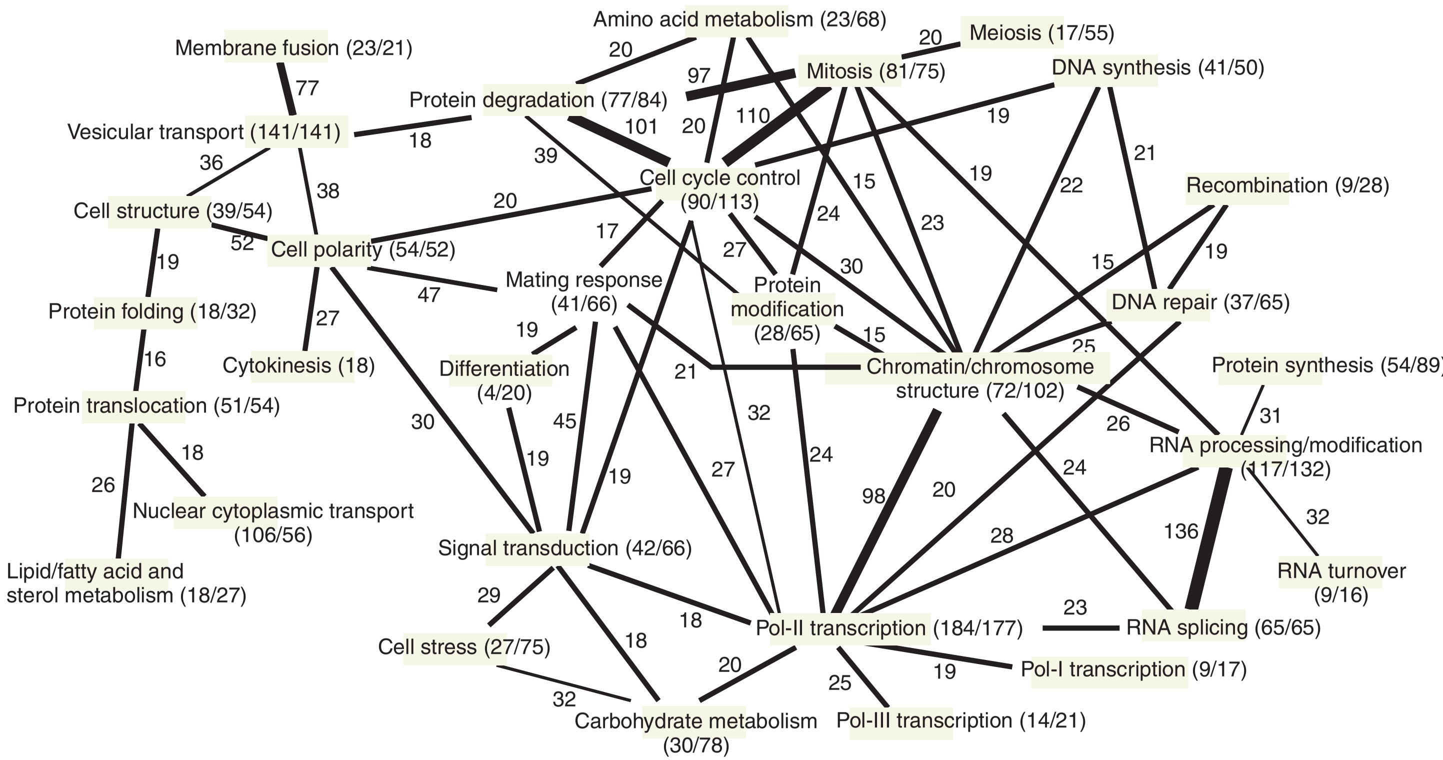


Figure 7 *Interactions between functional groups*. Numbers in parentheses indicate, first, the number of interactions within a group, and second, the number of proteins in a group. Numbers near connecting lines indicate the number of interactions between proteins of the two connected groups. For example, there are 77 interactions between the 21 proteins involved in membrane fusion and the 141 proteins involved in vesicular transport (upper left corner); 23 protein interactions connect the 21 proteins involved in membrane fusion. Only connections with 15 or more interactions are included here. Note that only proteins with known function are shown (many of these have several functions). The sum of all interactions in this diagram is therefore smaller than the number of all interactions. After Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1257–1261. (Reproduced by permission of Nature Publishing Group)

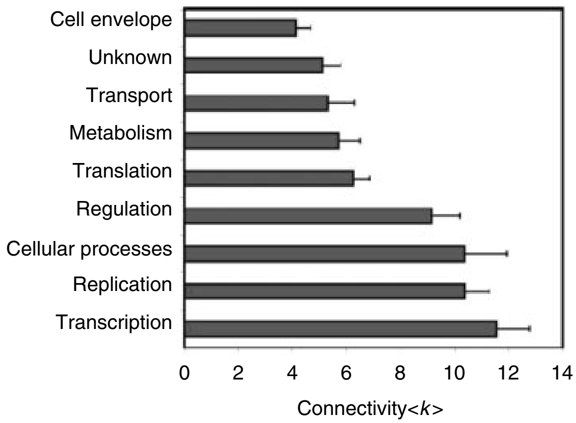


Figure 8 *Average connectivity levels k of functional classes.* Error was estimated from 100 random samplings of each functional class with 50 proteins each. After Kunin V, Pereira-Leal JB and Ouzounis CA (2004) Functional evolution of the yeast protein interaction network. *Molecular Biology and Evolution*, **21**, 1171–1176. (Reproduced by permission of Oxford University Press)

to *function*, proteins involved in the yeast cell envelope appear as the least connected, followed by unknown proteins, and proteins involved in transport and metabolism. At the other extreme, proteins involved in transcription, replication, cellular processes, and regulatory functions have, on average, almost twice as many binding partners. Interestingly, proteins of unknown function ("Unknown" functional class) are close to the lower end of connectivity, a trend that has been observed already in some of the initial high-throughput interaction screens (Uetz *et al*., 2000).

Figure 8 summarizes the average connectivity of proteins of various functional classes.

8. Evolution and conservation of protein complexes and interactions

Gavin *et al*. (2002) showed that many protein complexes are conserved between yeast and human, even if the proteins have only limited sequence homology. Teichmann (2002) showed that homologous proteins involved in such complexes share 46% sequence identity between *S. cerevisiae* and *S. pombe*, while proteins not known to be involved in complexes share 38% identical amino acids. This is not a huge difference and false-positive interactions have been used to explain why highly connected proteins are not subject to more evolutionary constraints (Wagner, 2001; Hahn *et al*., 2004).

Pagel *et al*. (2004) detected a correlation between the number of interaction partners of a yeast protein and its likelihood of having an ortholog in other ascomycota species.

Surprisingly, the proteins of oldest origin (i.e., proteins that are conserved across most species) do not display the highest connectivity (Kunin *et al*., 2004). This

is in contrast to the prediction of the preferential attachment model that oldest proteins are expected to display the highest levels of connectivity (Albert and Barabasi, 2000). The majority of the most highly connected proteins are found to have emerged during the eukaryotic radiation. Although Kunin *et al*. (2004) observed that the most recent proteins tend to be of lower connectivity, they failed to detect the steady increase of connectivity with the protein age.

Structurally, Aloy *et al*. (2003) found that close homologs (30–40% or higher sequence identity) almost invariably interact the same way, that is, with the same interaction surfaces. This similarity cutoff correlates well with the conservation of function that appears to be almost complete above a threshold of about 40% amino acid identity.

9. How many protein interactions are there in yeast?

The fact that neither of the two studies by Uetz *et al*. and Ito *et al*. recapitulated more than 13% of the published interactions detected by the community of yeast biologists indicates that there are many more interactions remaining to be discovered.

Using approaches ranging from an educated guess to probabilistic modeling of large-scale two-hybrid screens, various authors estimated the lower and upper boundaries for the number of protein interactions in yeast to be around 10 000 and 30 000 interactions (Figure 9).

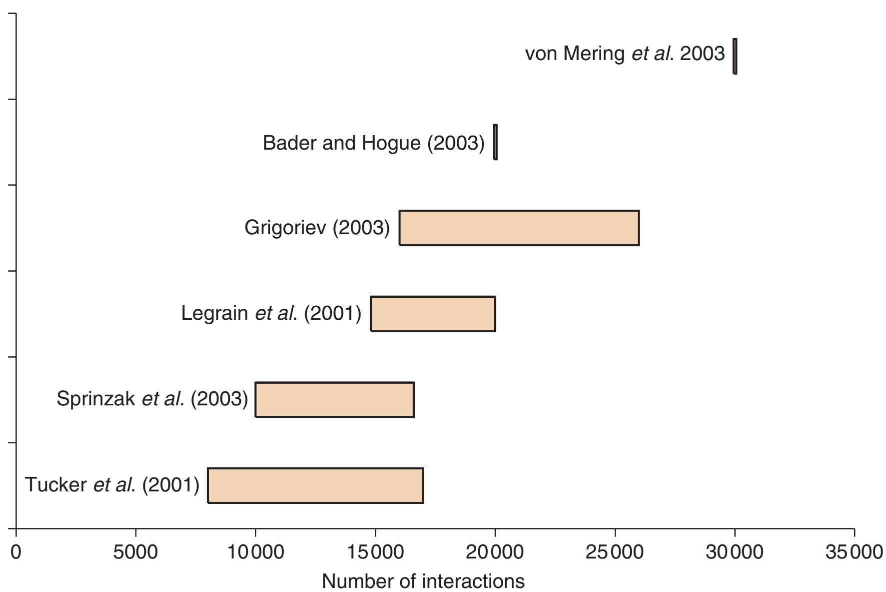


Figure 9 *Estimates of the size of the yeast interaction network, excluding homotypic interactions.* Numbers and ranges are as in the original publications (von Mering, 2002; Bader and Hogue, 2003; Grigoriev, 2003; Legrain *et al*., 2001; Sprinzak *et al*., 2003), not taking into account the changes in the estimated number of the yeast genes (for details see Grigoriev, 2004)

10. Biological relevance of protein interactions

Many protein interactions are absolutely essential. However, it remains unclear which fraction of interactions are really necessary on a global scale. Jeong *et al*. (2001) investigated an yeast interaction network with 1870 proteins as nodes, connected by 2240 direct physical interactions, and found that the scale-free nature of the network makes it relatively error-tolerant, at least when the diameter of the network is measured (i.e., the average distance between two random proteins). Single random mutations in the genome of *S. cerevisiae*, modeled by the removal of randomly selected yeast proteins, do not affect the overall topology of the network. By contrast, when the most connected proteins are computationally eliminated, the network diameter increases rapidly. The likelihood that removal of a protein will prove lethal correlates with the number of interactions the protein has. For example, although proteins with five or fewer links constitute about 93% of the total number of proteins, Jeong *et al*. found that only about 21% of them are essential. By contrast, only some 0.7% of the yeast proteins with known phenotypic profiles have more than 15 links, but single deletion of 62% or so of these proves lethal.

Related articles

Article 34, **Protein interactions in cellular signaling**, Volume 5; Article 35, **Structural biology of protein complexes**, Volume 5; Article 36, **Biochemistry of protein complexes**, Volume 5; Article 37, **Inferring gene function and biochemical networks from protein interactions**, Volume 5; Article 38, **The *C. elegans* interactome project**, Volume 5; Article 41, **Investigating protein–protein interactions in multisubunit proteins: the case of eukaryotic RNA polymerases**, Volume 5; Article 42, **Membrane-anchored protein complexes**, Volume 5; Article 44, **Protein interaction databases**, Volume 5; Article 45, **Computational methods for the prediction of protein interaction partners**, Volume 5; Article 46, **Functional classification of proteins based on protein interaction data**, Volume 5; Article 107, **Integrative approaches to biology in the twenty-first century**, Volume 6; Article 111, **Functional inference from probabilistic protein interaction networks**, Volume 6; Article 118, **Data collection and analysis in systems biology**, Volume 6

Acknowledgments

We thank Pierre Legrain, Nils Johnsson, and members of the Uetz lab for critical reading of the manuscript and Andrew Emili for providing unpublished information. Peter Uetz has been supported by DFG grant UE 50/2-1/2.

Further reading

Barabasi AL and Bonabeau E (2003) Scale-free networks. *Scientific American*, **288**, 60–69.

Dorogovtsev SN and Mendes JFF (2003) *Evolution of Networks. From Biological Nets to the Internet and WWW*, Oxford University Press: Oxford.

Golemis EA (Ed.) (2001) *Protein-Protein Interactions: A Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press: Cold Spring Harbor.

Fields S and Bartel PL (Eds.) (1997) *The Yeast Two-Hybrid System*, Oxford University Press: Oxford.

Kleanthous C (Ed.) (2001) *Protein-Protein Recognition*, Oxford University Press: Oxford.

References

Albert R and Barabasi AL (2000) Topology of evolving networks: local events and universality. *Physical Review Letters*, **85**, 5234–5237.

Aloy P and Russell RB (2002) The third dimension for protein interactions and complexes. *Trends in Biochemical Sciences*, **27**, 633–638.

Aloy P, Ceulemans H, Stark A and Russell RB (2003) The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, **332**, 989–998.

Bader GD and Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, **20**, 991–997.

Bader GD and Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Barabasi AL and Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nature Reviews. Genetics*, **5**, 101–113.

Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, *et al*. (2003) Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, **31**, 2443–2450.

Cornell M, Paton NW and Oliver SG (2004) A critical and integrated view of the yeast interactome. *Comparative and Functional Genomics*, **5**, 382–402.

Deane CM, Salwinski L, Xenarios I and Eisenberg D (2002) Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Molecular and Cellular Proteomics*, **1**, 349–356.

Drees BL, Sundin B, Brazeau E, Caviston JP, Chen GC, Guo W, Kozminski KG, Lau MW, Moskow JJ, Tong A, *et al*. (2001) A protein interaction map for cell polarity development. *The Journal of Cell Biology*, **154**, 549–571.

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J and Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*, **18**, 529–536.

Fromont-Racine M, Rain JC and Legrain P (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genetics*, **16**, 277–282.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ge H, Liu Z, Church GM and Vidal M (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics*, **29**, 482–486.

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M and Alexandraki D (1997) The yeast genome directory. *Nature*, **387**, 5–105.

Goldberg DS and Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4372–4376.

Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, **29**, 3513–3519.

Grigoriev A (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, **31**, 4157–4161.

Grigoriev A (2004) Understanding the yeast proteome: A bioinformatics perspective. *Expert Reviews in Proteomics*, **1**, 133–145.

Hahn MW, Conant GC and Wagner AJ, (2004) Molecular evolution in large genetic networks: does connectivity equal constraint? *Journal of Molecular Evolution*, **58**, 203–211.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hazbun TR, Malmstrom L, Anderson S, Graczyk BJ, Fox B, Riffle M, Sundin BA, Aranda JD, McDonald WH, Chiu CH, *et al*. (2003) Assigning function to yeast proteins by integration of technologies. *Molecular Cell*, **12**, 1353–1365.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574.

Jansen R, Greenbaum D and Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Research*, **12**, 37–46.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A and Holstege FC (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, **9**, 1133–1143.

Krause R, von Mehring C and Bork P (2003) A comprehensive set of protein complexes in yeast: Mining large scale protein–protein interaction screens. *Bioinformatics*, **19**, 1901–1908.

Kunin V, Pereira-Leal JB and Ouzounis CA (2004) Functional evolution of the yeast protein interaction network. *Molecular Biology and Evolution*, **21**, 1171–1176.

Legrain P, Wojcik J and Gauthier JM (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends in Genetics*, **17**, 346–352.

Lehner B and Fraser AG (2004) A first-draft human protein-interaction map. *Genome Biology*, **5**, R63.

Mrowka R, Patzak A and Herzel H (2001) Is there a bias in proteome research? *Genome Research*, **11**, 1971–1973.

Pagel P, Mewes HW and Frishman D (2004) Conservation of protein-protein interactions – lessons from ascomycota. *Trends in Genetics*, **20**, 72–76.

Saito R, Suzuki H and Hayashizaki Y (2002) Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, **30**, 1163–1168.

Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1257–1261.

Spirin V and Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12123–12128.

Sprinzak E, Sattath S and Margalit H (2003) How reliable are experimental protein–protein interaction data? *Journal of Molecular Biology*, **327**, 919–923.

Teichmann SA (2002) The constraints protein-protein interactions place on sequence divergence. *Journal of Molecular Biology*, **324**, 399–407.

Titz B, Schlesner M and Uetz P (2004) What do we learn from high-throughput protein interaction data and networks? *Expert Reviews in Proteomics*, **1**, 89–99.

Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, *et al*. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Uetz P, Ideker T and Schwikowski B (2002) Visualization and integration of protein-protein interactions. In *Protein-Protein Interactions – A Molecular Cloning Manual*, Golemis E (Ed.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor.

Vollert CS and Uetz P (2004) The PX domain protein interaction network in yeast. *Molecular and Cellular Proteomics*, **3**, 1–12.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, **18**, 1283–1292.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Short Specialist Review

Human signaling pathways analyzed by protein interaction mapping

Frédéric Colland
Hybrigenics SA, Paris, France

Pierre Legrain
CEA Saclay, Gif-sur-Yvette, France

The generation of protein interaction maps can provide initial insight into cellular mechanisms. The study of such maps can be used as a means of suggesting the functions of hundreds of previously uncharacterized genes (*see* Article 39, **The yeast interactome**, Volume 5). Most protein interaction maps have been developed with the yeast two-hybrid system (Legrain *et al*., 2001), the complex pull-down approach involving the coimmunoprecipitation of proteins followed by mass spectrometry (Gavin *et al*., 2002; *see also* Article 36, **Biochemistry of protein complexes**, Volume 5) or the phage display method (Tong *et al*., 2002). However, these systems present a challenging technical problem in that the resulting datasets contain large numbers of irrelevant interactions due to the use of high-throughput methods. Biological validation of protein interaction networks is therefore required, by integrating heterogeneous data, such as those relating to expression profiling, genetic data, phenotypic analyses, protein localization, or published functional annotations. However, only limited amounts of information are currently available for most of the proteins encoded in the human genome, suggesting that functional annotations will remain limited or uncertain for a large proportion of protein networks.

The putative functions of proteins involved in the interaction network can be validated by means of cellular assays. Such studies have been performed in *Caenorhabditis elegans*, with the establishment of protein interaction maps for DNA damage response (Boulton *et al*., 2002) and DAF7/TGFβ signal transduction (Tewari *et al*., 2004), followed by functional validation (*see* Article 38, **The *C. elegans* interactome project**, Volume 5). However, to date, only specific studies focusing on a very limited set of proteins and interactions have been carried out for mammalian proteins (Boeda *et al*., 2002). A key limitation is the difficulty involved in handling the large number of interactions identified, together with the numerous specific functional analyses required to demonstrate the biological function of each newly identified protein and/or interaction. Two recent studies combined protein interaction

mapping with functional validation assays. The first dealt with the TNF-α/NF-κB signal transduction pathway (Bouwmeester *et al*., 2004) and the second with the human Smad signaling pathway (Colland *et al*., 2004). In these two studies, protein interaction mapping, carried out with different technologies, led to the generation of large datasets, which were explored with the help of the functional annotations available in public databases. The global validation of protein networks was achieved by identifying well-known proteins involved in these pathways. A subset of proteins, as yet poorly annotated or functionally unrelated to these pathways, was then selected for study in cellular assays, with loss-of-function and overexpression analyses, to confirm the involvement of these proteins in these pathways.

TNF-α triggers a signaling cascade that converges on activation of the transcription factor NF-κB. Bouwmeester and coworkers used a complex pull-down technique derived from the tandem affinity purification method developed for yeast protein complex analyses (Rigaut *et al*., 1999) to identify proteins that copurified with 32 known and candidate components of this pathway. An initial dataset of 680 protein associations was constructed. It should be noted that the complex pull-down method does not identify individual protein–protein contacts but instead identifies groups of proteins in stable association. In this dataset, they identified 70% of the previously reported interactions with the original 32 proteins; the success rate for the detection of known interactions was therefore fairly high. Thorough statistical analysis of the dataset led to the selection of a subset of 131 for which interactions were predicted to occur with a high level of confidence. Eighty proteins in this dataset were not well characterized or were unrelated to the TNF-α/NF-κB signal transduction pathway and may therefore correspond to previously unidentified components of this pathway. Twenty-eight proteins were chosen for further functional characterization in an RNAi gene expression perturbation assay, with phenotypic effects monitored on an NF-κB-dependent luciferase reporter readout. Ten of these 28 candidates displayed a decrease in luciferase activity, consistent with a role in the modulation of TNF-α/NF-κB signal transduction.

In an independent study, Colland and coworkers investigated the human Smad signaling pathway (Colland *et al*., 2004). Members of the TGFβ superfamily (e.g., TGFβ, activins, and bone morphogenetic proteins (BMPs)) are secreted signaling molecules that regulate many biological processes, including cell growth, differentiation, and morphogenesis. The disruption of components of the TGFβ superfamily pathways is associated with human diseases including fibrosis, inflammatory disorders, and cancer. The effects of these pleiotropic molecules are transduced via various kinase receptors and Smad transcription factors, which regulate target gene expression in collaboration with other transcriptional partners (Massague, 1998).

Protein–protein interactions involving 23 Smad and related proteins were screened with the yeast two-hybrid system. Protein complexes were not characterized but individual contacts between proteins were identified (Figure 1). A complex network of 755 interactions involving 591 proteins was generated. A dedicated navigation tool for protein interaction maps was used to explore and analyze such complex interaction databases (the PIMRider®, (Rain *et al*., 2001)). The network includes 18 known Smad-connected proteins and 179 poorly or unannotated proteins. The yeast two-hybrid method used in this study makes it possible to

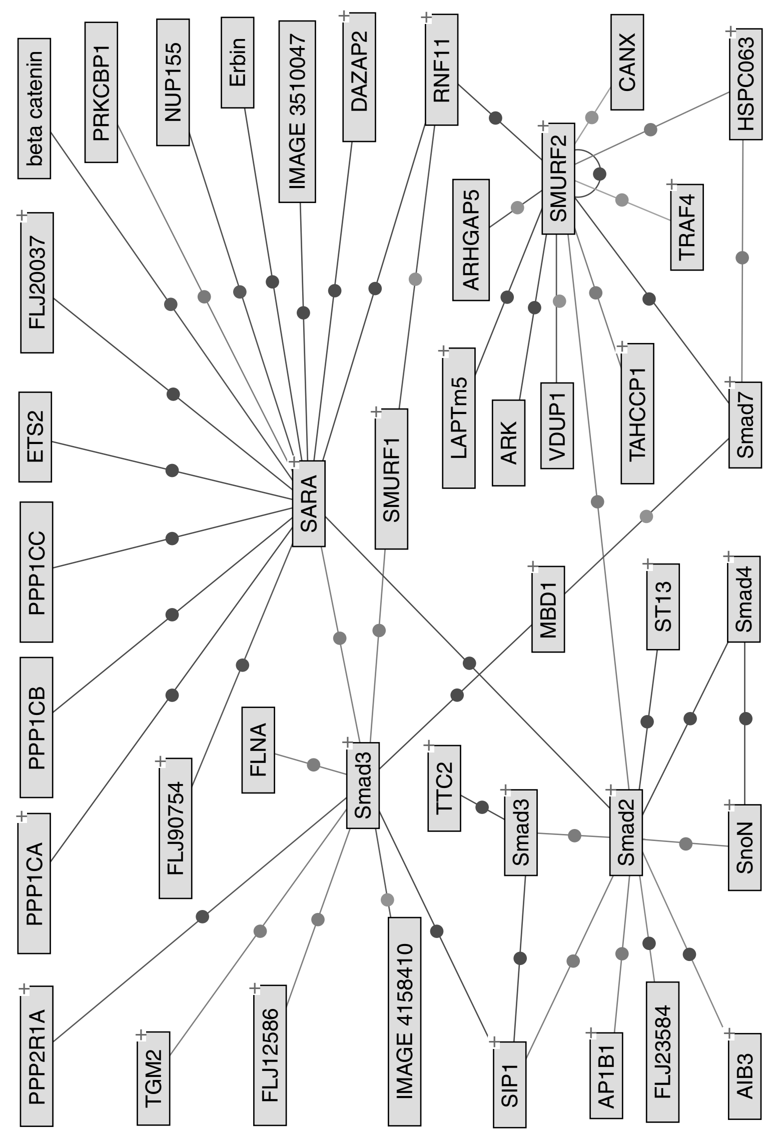


Figure 1 Partial protein interaction network for the Smad pathway

identify the fragment of the protein that contains the interacting domain ("selected interacting domain" or SID®). These SIDs and the functional domains of each given protein are displayed in PIMRider, facilitating thorough searches for correlations between characterized interaction motifs and selected interacting domains. Each protein–protein interaction pair is also characterized by a confidence score based on a statistical model, reflecting the reliability of identified interactions in the context of the full dataset.

Fourteen proteins were selected on the basis of interaction scores and functional domain annotations. These proteins were tested for their involvement in the Smad pathway by loss-of-function and overexpression studies in mammalian cells. Using Smad-specific reporter genes and Smad-dependent endogenous target genes as readout, eight of these 14 proteins were found to be associated with the Smad-related signal transduction. Four of these eight proteins (PPP1CA, ZNF8, MAN1, and RNF11) have independently been shown to be involved in the Smad pathway by other groups (Bennett and Alphey, 2002; Jiao *et al*., 2002; Raju *et al*., 2003; Subramaniam *et al*., 2003). The other four proteins (KIAA1196, HYPA, LMO4, and LAPTm5) have unknown functions or are poorly characterized, and biological functions could be suggested for these proteins. For example, the *LAPTm5* gene is upregulated by TGFβ and the LAPTm5 protein acts as a negative regulator of the TGFβ pathway. These findings, combined with the LAPTm5 interaction data, suggest that LAPTm5 may act as a smurf2 receptor in the lysosomal membrane and may target some TGFβ signaling components to the lysosomal compartment for degradation (Colland, 2004).

These two studies demonstrate that a large-scale protein–protein interaction map, combined with functional validation in mammalian cells, can be used to identify new components of signal transduction pathways. Clearly, such studies cover only a limited number of the potential interactions between human proteins. However, questions may be raised concerning the quality of datasets generated in studies aiming to cover the complete proteomes of complex organisms and the experimental methods currently available for the functional validation of such datasets. Focused studies, such as those summarized here, involve tens of proteins and generate protein interaction maps including hundreds of interactions. In-depth analyses of these datasets have led to the selection of tens of putative new components of signaling pathways. About half these candidates turn out to be functionally related to the expected pathways. Protein interaction maps generated at the whole-genome scale have recently been published for complex organisms such as the nematode *C. elegans* (Li *et al*., 2004; *see also* Article 38, **The *C. elegans* interactome project**, Volume 5) and the fruit fly *Drosophila melanogaster* (Giot *et al*., 2003). These two exhaustive studies generated genome-wide protein interaction maps connecting 3000 proteins via 5500 interactions for *C. elegans* and 5000 proteins via 5000 interactions for *Drosophila*. Given that these genomes each encode more than 10 000 proteins, these maps provide a much lower level of coverage than those based on a more focused set of proteins. With such a low density of information, associated with a presumably high frequency of false–positives, poor annotations of large parts of the networks are likely to preclude the design of further functional validation experiments, which are required to confirm the annotation made for each protein.

In conclusion, extensive functional proteomic studies, combining large-scale protein interaction mapping with functional assays, focusing on specific cellular functions or processes should considerably increase our knowledge about many human genes that are still poorly annotated, even with currently available methods. In the long term, the availability of a common format for the storage of all interaction datasets should facilitate the thorough exploration and comparison of protein networks (Hermjakob *et al*., 2004). Ultimately, a unified dataset of protein interactions will be constructed for the genome-wide exploration of proteomes.

Acknowledgments

We thank all Hybrigenics staff for their contribution.

References

Bennett D and Alphey L (2002) PP1 binds Sara and negatively regulates Dpp signaling in *Drosophila melanogaster*. *Nature Genetics*, **31**, 419–423.

Boeda B, El-Amraoui A, Bahloul A, Goodyear G, Daviet L, Blanchard S, Perfettini I, Fath KR, Shorte S, Reiners J, *et al*. (2002) Myosin VIIa, harmonin and cadherin 23, three Usher I gene products that cooperate to shape the sensory hair cell bundle. *EMBO Journal*, **21**, 6689–6699.

Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, Hill DE and Vidal M (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science*, **295**, 127–131.

Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, *et al*. (2004) A physical and functional map of the human TNF-alpha/NF-kappaB signal transduction pathway. *Nature Cell Biology*, **6**, 97–105.

Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, Hamburger A, Meil A, Wojcik J, Legrain P and Gauthier JM (2004) Functional proteomic mapping of a human signaling pathway. *Genome Research*, **14**, 1324–1332.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al*. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**, 177–183.

Jiao K, Zhou Y and Hogan BL (2002) Identification of mZnf8, a mouse Kruppel-like transcriptional repressor, as a novel nuclear interaction partner of Smad1. *Molecular and Cellular Biology*, **22**, 7633–7644.

Legrain P, Wojcik J and Gauthier JM (2001) Protein-protein interaction maps: a lead towards cellular functions. *Trends in Genetics*, **17**, 346–352.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.

Massague J (1998) TGF-beta signal transduction. *Annual Review of Biochemistry*, **67**, 753–791.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.

Raju GP, Dimova N, Klein PS and Huang HC (2003) SANE, a novel LEM domain protein, regulates bone morphogenetic protein signaling through interaction with Smad1. *Journal of Biological Chemistry*, **278**, 428–437.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Subramaniam V, Li H, Wong M, Kitching R, Attisano L, Wrana J, Zubovits J, Burger AM and Seth A (2003) The RING-H2 protein RNF11 is overexpressed in breast cancer and is a target of Smurf2 E3 ligase. *British Journal of Cancer*, **89**, 1538–1544.

Tewari M, Hu PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, *et al*. (2004) Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. *Molecular Cell*, **13**, 469–482.

Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, *et al*. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Short Specialist Review

Investigating protein–protein interactions in multisubunit proteins: the case of eukaryotic RNA polymerases

Benjamin Guglielmi, Cécile Zaros, Pierre Thuriaux and Michel Werner

Service de Biochimie et Génétique Moléculaire, Gif-sur-Yvette, France

The development of protein purification based on fast and nondisruptive affinity purification methods rather than on conventional chromatography has provided growing evidence that many proteins may exist in vast heteromultimeric associations with molecular weights reaching 1 MDa or above (*see* Article 99, **Large complexes by X-ray methods**, Volume 6 and Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6). Meanwhile, mass spectrometry (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5) coupled with genome sequence knowledge and to the precise annotation of model organisms such as *Saccharomyces cerevisiae* (*see* Article 43, **Functional genomics in** ***Saccharomyces cerevisiae***, Volume 3, Article 97, **Seven years of yeast microarray analysis**, Volume 4, and Article 39, **The yeast interactome**, Volume 5) has considerably accelerated the identification of polypeptides belonging to such complexes. Thus, the Tandem Affinity Purification method (Rigaut *et al*., 1999), in which two different affinity tags (a calmodulin-binding peptide and *Staphylococcus aureus* protein A) allow the fast purification of multiprotein complexes at near-physiological ionic strengths, preserves structures that are typically disrupted by conventional purification procedures. When applied to the yeast proteome (Gavin *et al*., 2002; Ho *et al*., 2002), this approach suggests that the eukaryotic cell may harbor several hundreds of these complexes (*see* Article 35, **Structural biology of protein complexes**, Volume 5).

Some of these structures have a remarkably stable subunit composition, like the mRNA polyadenylation complex where tandem affinity purification from different polypeptides consistently yielded the same set of 20 different subunits. Others have a much more dynamic organization. Consider, for example, a molecule of DNA-dependent RNA polymerase (Pol) II that successively finds a promoter, passes by an early phase of “abortive” transcription (probably without moving along its DNA), then moves along its template to synthesize RNA in a processive way, and

finally terminates transcription before it possibly recycles onto the same promoter. During these *avatars* of the Transcription Cycle, the same molecule is associated with a nearly 1-MDa Mediator, with the General Transcription Factors TFIIB, TFIIE, TFIIF, TFIIH, with more or less well-defined elongation factors such as TFIIS, Spt4, Spt5, and some others, with the capping enzyme and perhaps with the polyadenylation machine mentioned above (Woychik and Hampsey, 2002). The cumulated mass of this ensemble of more than 70 polypeptides would be close to the size of ribosomes (4.2 MDa in eukaryotes), but it is rather unlikely that a huge "transcriptosome" of that size exists as a stable structure. Indeed, when yeast Pol II subunits were purified by tandem affinity purification, they consistently yielded a 19 polypeptide complex of about 0.75 MDa with the 12-subunit Pol II, the 3-subunit General Transcription Factor TFIIF, two elongation factors (Spt4 and Spt5), and the mRNA cap methylase, which is probably the elongating form of RNA polymerase II. Initiation Factors such as TFIIB, TFIIE, TFIIS, and TFIIH did not copurify with Pol II in tandem affinity purification, although there is good evidence that they *directly* bind Pol II. There was also no trace of the Mediator, although the latter can stably bind Pol II in vitro, thus forming edifices of about 1.5 MDa that can be visualized by electron microscopy.

Knowing the structure of such a huge and more or less transient complex at an atomic level of resolution is undoubtedly one of the most important challenges of modern protein chemistry (*see* Article 99, **Large complexes by X-ray methods**, Volume 6 and Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6). Short of unforeseen technical breakthroughs, however, this is unlikely to be done rapidly. In the very well studied case of the yeast 12-subunit Pol II complex (Cramer *et al*., 2000; Cramer *et al*., 2001; Armache, 2003; Bushnell, 2003), it took more than 15 years to obtain the (nearly) complete atomic structure of the 12-subunit yeast Pol II, starting from electron microscopic studies that provided the first low-resolution images of RNA polymerases I and II (Edwards *et al*., 1990; Schultz *et al*., 1990). These images paved the way for a very recent atomic reconstitution of Pol II associated with the mononomeric initiation factor TFIIB and elongation factor TFIIS. In most cases, however, cell biologists still largely rely on generic methods for investigating protein complexes in terms of protein–protein interactions and for obtaining at least a broad picture of their general organization. These methods will never replace the precise knowledge brought by crystallographic structures, but are extremely important for obtaining an overall view of the subunit organization of such complexes and of their modular structure.

Protein–protein interaction mapping is based on the idea that individual polypeptides have autonomous folds and that their interactions can therefore be reconstituted when presenting to each other two interacting partners in physiological conditions. The aim, therefore, is to detect the formation of stable heterodimers by suitable biochemical or genetic tests. The biochemical approach is based on protein "pull-down" assays, where one partner polypeptide is fused to an affinity tag (e.g., glutathione-S-transferase or polyhistidine) and bound to a resin (e.g., glutathione Sepharose or Ni-NTA agarose), the putative partner protein being chromatographed through the affinity resin. The retention of a given partner is detected by antibodies directly raised against that partner itself or against suitable epitopes fused to it.

The setup can vary widely. In one form, the binding test is done between two purified proteins. This method is very sensitive but its specificity is questionable since the concentrations of the purified proteins can be in far excess over those that are found in the cell. Indeed, the fact that proteins tend to bind to each other by fairly unspecific interactions is a major source of false-positive. Testing interactions between proteins expressed from whole-cell extracts, typically of *Escherichia coli* or insect viruses (baculoviruses), has the advantage that the protein used as "bait" is challenged for nonspecific interactions by the whole-cell-free extract itself. False-positive interactions are therefore likely to be less frequent, as long as the two partners are not massively overproduced from the host cell.

In the genetic two-hybrid test (Fields and Song, 1989), one polypeptide is typically fused to the Gal4 DNA binding domain (G_{BD}) and tested against putative partners fused to the Gal4 activation domain (G_{AD}). If able to interact, they will form an active heterodimeric Gal4 activator, inducing the transcription of suitable yeast reporter genes in vivo. An interesting development of the two-hybrid method is readily applicable to organisms with compact genomes such as yeast, where it is possible to screen a given bait protein for its ability to interact with a library of small genomic fragments (around 700 bp; Fromont-Racine *et al*., 1997). Since intergenic distances are short and since introns are rare in *S. cerevisiae*, this essentially amounts to testing a random library of coding sequence fragments. In a typical experiment, 10^7 clones can be tested, usually yielding several dozens of interactants. This approach provides an excellent indicator of specificity, as distinct but overlapping fragments corresponding to the same partner should be isolated in the same screen. Comparing the limits of these fragments then allows to rather precisely delineate the interacting domain. This approach has been used to investigate the organization of complex edifices such as the yeast 17-subunit Pol III (Flores *et al*., 1999).

The atomic structure of Pol II, where the 12 subunits are connected by 16 pairwise interactions, offers a rare opportunity for comparing the predictive power of these two approaches and leaves little doubt as to the superiority of the two-hybrid approach, especially when the latter is performed in a library of fragments. GST pull-down assays have been performed with the 12 subunits of the human Pol II, and was also extended to the fission yeast (Kimura and Ishihama, 2000). This can be readily compared to the yeast structural data, since there is considerable homology between the two RNA polymerases and since several of the human subunits can functionally replace their yeast counterparts in vivo. In the human enzyme, the method correctly detected eight interactions, missed eight others, and had seven false-positives (Edwards *et al*., 2002; Figure 1). False-positives were therefore a major problem, especially with some small subunits that were (falsely) predicted to interact with many partners. For example, the Rpb5 subunit interacted with five partners, including Rpb1, in the pull-down experiments, while it only binds the latter in the crystal structure. The *S. pombe* Rpb3 subunit was also seen to interact with six partners, of which only two (Rpb2 and Rpb11) actually did in the structure.

The analysis of yeast RNA polymerase III (Flores *et al*., 1999) suggests that the two-hybrid approach may be much less prone to false-positives. RNA polymerase III is a very complex enzyme that contains no less than 17 subunits (Siaut *et al*.,

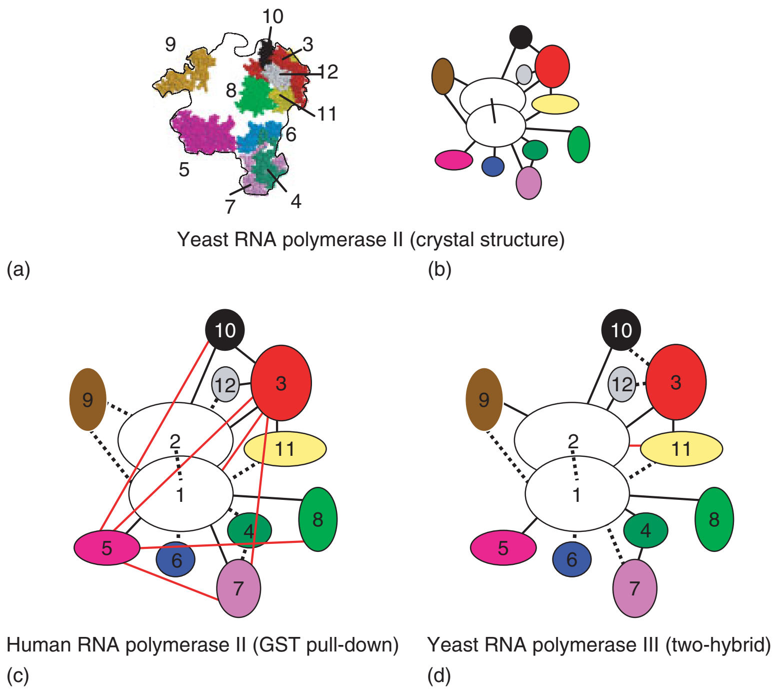


Figure 1 Organization of the subunits in yeast and human RNA polymerases (a) Projection of the RNA polymerase II structure (*S. cerevisiae*). The location of the subunits is indicated within the contour of the model (according to Cramer *et al*., 2000, 2001). Rpb1 and Rpb2 were omitted for clarity. Color code: Rpb3: red; Rpb4: dark green; Rpb5: magenta; Rpb6: blue; Rpb7: violet; Rpb8: light green; Rpb9: orange; Rpb10: black; Rpb11: yellow; Rpb12: gray. (b) Protein interactions in RNA polymerase II, according to the crystallographic analysis (Cramer *et al*., 2000, 2001). The connections between the Pol II subunits are indicated by the black lines. The color code is the same as in (a). (c) Protein interactions in human RNA polymerase II, according to the GST pull-down experiments (Acker *et al*., 1997). The red lines represent false-positive interactions. Dashed lines correspond to interactions that remained undetected. (d) Protein interactions in yeast RNA polymerase III, according to the two-hybrid assay (Flores *et al*., 1999). The code is as in (c)

2003). A core of 12 subunits is homologous (7) or even identical (5) to the 12-subunit crystal structure of Pol II (Armache, 2003; Bushnell, 2003). Nine of the 16 interactions predicted by the Pol II structure were found in the two-hybrid analysis of Pol III, with only one false-positive (Rpb2-Rpb11). Moreover, the domains of interactions predicted by this analysis matched the ones found in the Pol II structure. The five Pol III–specific subunits were also included in this two-hybrid screening, which allocated them to two distinct groups, one formed by the Rpc53 and Rpc37 subunits and one formed by Rpc31, Rpc34, and Rpc82. The case of Rpc37 was quite striking, since this polypeptide was first identified as a specific partner of Rpc53 before being proven to be a bona fide subunit of Pol III (Flores *et al*., 1999). The other three subunits interacted with each other, and are also connected to the 12-subunit core (via a Rpc31-Rpc17 interaction) and to the TFIIIB initiation factor (via Rpc34-Brf1 and Rpc17-Brf1 interactions). Brf1 is a conserved polypeptide that has close homology to the TFIIB initiation factor of Pol

II, and is one of the two polypeptides forming the Pol III initiation factor TFIIIB. The physiological relevance of these interactions was supported by genetic studies where amino acid replacements were selected to impair the two-hybrid interaction and were then shown to result in transcription defects (Andrau *et al*., 1999; Brun *et al*., 1997; Ferri *et al*., 2000) or dissociation of the corresponding subcomplexes (Werner *et al*., 2002). Moreover, the effect of such mutations was in some cases corrected by the overexpression of other components of the same complex (Briand *et al*., 2001).

Most cell biologists would agree that the association of proteins with each other has been a major playground of evolution, improving their biological performance in the extremely crowded environment of living cells. In the eukaryotic cell, one could therefore view the nucleoplasm, the cytosol, and also the matrix compartment of organelles as highly organized communities of multiprotein complexes, rather than as crowded swimming pools for individualistic proteins. However, a long time will pass before structural data provide a general view of how most of these complexes are organized in terms of subunit interactions. In the meantime, there is ample room for generic methods such as the two-hybrid screening approach discussed above and even the somewhat less reliable protein pull-down approach. If carefully applied, these methods can provide important insights into the organization of multiprotein complexes, especially when supported by the in vivo or in vitro analysis of yeast mutants defective in the corresponding protein–protein interactions.

References

Acker J, de Graaf M, Cheynel I, Khazak V, Kedinger C and Vigneron M (1997) Interactions between the human RNA polymerase II subunits. *Journal of Biological Chemistry*, **272**, 16815–16821.

Andrau J-C, Sentenac A and Werner M (1999) Mutagenesis of yeast TFIIIB70 reveals C-terminal residues critical for interaction with TBP and C34. *Journal of Molecular Biology*, **288**, 511–520.

Armache KJ, Kettenberger H and Cramer P (2003) Architecture of initiation-competent 12-subunit RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6964–6968.

Briand JF, Navarro F, Rematier P, Boschiero C, Labarre S, Werner M, Shpakovski GV and Thuriaux P (2001) Partners of Rpb8p, a small subunit shared by yeast RNA polymerases I, II, and III. *Molecular and Cellular Biology*, **21**, 6056–6065.

Brun I, Sentenac A and Werner M (1997) Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO Journal*, **16**, 5730–5741.

Bushnell DA and Kornberg, RD (2003) Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6969–6973.

Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AM, David PR and Kornberg RD (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, **288**, 640–649.

Cramer P, Bushnell DA and Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science*, **292**, 1863–1876.

Edwards AM, Darst SA, Feaver WJ, Thompson NE, Burgess RR and Kornberg RD (1990) Purification and lipid-layer crystallization of yeast RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2122–2126.

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J and Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*, **18**, 529–536.

Ferri ML, Peyroche G, Siaut M, Lefebvre O, Carles C, Conesa C and Sentenac A (2000) A novel subunit of yeast RNA polymerase III interacts with the TFIIB- related domain of TFIIIB70. *Molecular and Cellular Biology*, **20**, 488–495.

Fields S and Song O-K (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Flores A, Briand J-F, Gadal O, Andrau J-C, Rubbi L, Van Mullem V, Boschiero C, Goussot M, Marck C, Carles C, *et al*. (1999) A protein-protein interaction map of yeast RNA polymerase III. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 7815–7820.

Fromont-Racine M, Rain J-C and Legrain P (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genetics*, **16**, 277–282.

Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Kimura M and Ishihama A (2000) Involvement of multiple subunit-subunit contacts in the assembly of RNA polymerase II. *Nucleic Acids Research*, **28**, 952–959.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Schultz P, Célia H, Riva M, Darst SA, Colin P, Kornberg RD, Sentenac A and Oudet P (1990) Structural study of the yeast RNA polymerase A: electron microscopy of lipid-bound molecules and two-dimensional crystals. *Journal of Molecular Biology*, **216**, 353–362.

Siaut M, Zaros C, Levivier E, Ferri ML, Werner M, Callebaut I, Thuriaux P, Sentenac A and Conesa C (2003) A Rpb4/Rpb7 like complex in yeast RNA Polymerase III contains the orthologue of mammalian CGRP-RCP. *Molecular and Cellular Biology*, **23**, 195–205.

Werner M, Hermann-Le Denmat S, Treich I, Sentenac A and Thuriaux P (1992) Effect of mutations in a zinc binding domain of yeast RNA polymerase C (III) on enzyme function and subunit association. *Molecular and Cellular Biology*, **12**, 1087–1095.

Woychik NA and Hampsey M (2002) The RNA Polymerase II Machinery. Structure Illuminates Function. *Cell*, **108**, 453–463.

Short Specialist Review

Membrane-anchored protein complexes

Igor Stagljar
University of Zurich-Irchel, Zurich, Switzerland

1. Introduction

The recent advances in analyses of completely sequenced genomes of numerous model organisms, including the human genome, have revealed that approximately one-third of all predicted gene products of a given organism are likely to be associated with membranes (Auerbach *et al*., 2002). These integral and membrane-associated proteins execute a variety of essential cellular tasks. For example, transmembrane receptors bind specific signaling molecules such as hormones, growth factors, and neurotransmitters, and initiate specific cellular responses. Ion channels, transporters, and pumps mediate the exchange of membrane-impermeable molecules between cellular compartments, and between a cell and its extracellular environment. Cell-adhesion molecules enable many animal cells to adhere tightly and specifically with cells of the same or similar type ensuring their segregation into distinct tissues (Fetchko *et al*., 2003). Because of their exposure to the extracellular space, membrane proteins are also of considerable pharmacological importance. Namely, 50% of currently known drug targets (∼500) are either membrane receptors or ion channels. Altogether, the central role of membrane proteins in many cellular processes, their direct link to human disease, and their accessibility toward drugs makes an understanding of membrane protein function desirable.

However, the study of membrane proteins has historically been problematic. The hydrophobic nature of membrane proteins often results in insoluble proteins, which makes protein isolation difficult and therefore hinders the determination of protein complex composition and protein function. Recently, several yeast genetic techniques have made the characterization of interactions among membrane proteins more feasible (Reviewed in (Auerbach *et al*., 2002)). By characterizing protein interactions, researchers can link unknown proteins to proteins of known function, thus providing insight into the unknown protein function. Also, since protein interactions are central in regulating protein activity, knowledge of protein interactions provides understanding of both normal and mutant protein activity.

This review provides a brief overview of the available genetic methods for detecting protein interactions among membrane proteins.

2. Technologies for detecting membrane protein interactions

Generally speaking, the approaches for dissecting protein–protein interactions between membrane proteins can be divided into biochemical and genetic methods (Stagljar, 2003). Biochemical methods establish the state of interacting proteins by directly working with proteins in order to determine the composition of protein complexes, whereas genetic methods indirectly determine protein interactions on the basis of outputs produced through the manipulation of endogenous or exogenous gene networks. Traditionally, biochemical methods such as copurification or coimmunoprecipitation have been used to investigate the composition of protein complexes (Adams *et al*., 2002; Einarson and Orlinich, 2002; Pedrazzi and Stagljar, 2004). However, these methods are tedious and time consuming, requiring extensive optimization for each given protein pair and are therefore unsuitable for the simultaneous application to the tens of thousands of uncharacterized proteins predicted from genome sequences.

Aside from biochemical approaches, expression systems that are based on the detection of protein–protein interactions in vivo have become a popular tool because they require little individual optimization and are well suited for screenings in a high-throughput format. Among genetic assays, the YTH system takes a special position (Fields and Song, 1989). The organism used for the YTH system, the unicellular eukaryote *Saccharomyces cerevisiae*, is well suited to high-throughput screenings, is thoroughly characterized (completely sequenced genome), easy to handle, grows quickly, is well suited for the expression of heterologous proteins, and contains a multitude of selective markers available for nutritional selection (e.g., *HIS3, URA3, TRP1, LEU2*), drug susceptibility (e.g., *URA3*, cyh^{R}), drug resistance (e.g., *kanMX, natMX*), and colors (e.g., LacZ, GFP, GusA) (Melese and Hieter, 2002).

Despite the fact that YTH is a powerful and versatile technique to identify new partners, this method is limited because it cannot be applied to all proteins. For example, since in the YTH system one reconstitutes and monitors TF activity, all assays must be performed within the nucleus. This situation presents a problem when studying membrane proteins, since, in order to evaluate membrane protein interactions by the YTH method, the membrane protein must be portioned into fragments and then relocalized to the nucleus. This limitation not only affects the presentation of the protein of interest, but also limits the protein interacting partners, which can be detected by the YTH system as some membrane proteins cannot be effectively expressed in the nuclear milieu. In addition, membrane proteins are often co- and posttranslationaly modified (they are glycosylated and/or acquire disulfide bridges), and these modifications are not expected to occur in the nucleus (Thaminy and Stagljar, 2002). To overcome these problems, several yeast-based strategies have been designed to study protein interactions in their natural environment – membrane (Stagljar and Fields, 2002).

2.1. G-protein fusion assay

The G-protein fusion assay utilizes the G-protein coupled yeast mating pathway receptor to study protein interactions among membrane proteins (Ehrhard *et al*.,

2000). G-protein coupled receptors are composed of three subunits, G_α, G_β, and G_γ (Figure 1). When the G_α subunit of the G-protein coupled receptor binds a ligand (either mating factor a or α), GTP is exchanged for GDP on G_α, which leads to the dissociation of the trimeric receptor complex into G_α and $G_{\beta\gamma}$. Once the $G_{\beta\gamma}$ subunits are released from G_α, $G_{\beta\gamma}$ is no longer tethered to the membrane, which enables it to act on various downstream effectors thus activating the mating cascade.

In order to detect membrane protein interactions, the yeast-mating pathway G-protein signaling process was modified in the following manner. An unmodified integral membrane protein-Y serves as a "bait" and a "prey" X is fused to the $G_{\beta\gamma}$ subunit of the G-protein coupled receptor ($G_{\beta\gamma}$-X) (Figure 1). In the presence of ligand, if no interaction between $G_{\beta\gamma}$-X and the integral membrane protein Y occurs, the activation of the mating cascade is stimulated and is either detected by an indirect transcriptional readout or by a characteristic cell morphology change, that is, Schmoo formation, induced in the presence of the mating factor ligand (Figure 1b). However, if protein $G_{\beta\gamma}$-X interacts with Y, the Y integral membrane protein continues to sequester $G_{\beta\gamma}$-X at the membrane, which blocks the downstream mating cascade (Figure 1c).

The G-protein fusion assay has been shown to detect known interactions between the binding partners syntaxin 1a and neuronal Sec1 (nSec1), and between the fibroblast-derived growth factor receptor 3 (FGFR3) and SNT-1 (Ehrhard *et al*., 2000). In addition, mutations of nSec1 that fail to bind syntaxin 1a were isolated using the G-protein fusion assay (Ehrhard *et al*., 2000).

2.2. *Reverse Ras recruitment system (reverse RRS)*

In yeast, activation of Ras via the exchange of bound GDP for GTP is mediated by the Cdc25 guanyl exchange factor and results in cell growth. Temperature-sensitive mutations of *cdc25* (allele- *cdc25-2*) grow normally at the permissive temperature of 23°C but fail to grow at the restrictive temperature of 36°C. A constitutively active cytoplasmic version of a mammalian Ras (mRas), which lacks the carboxy-terminal domain "CAAX box" required for normal plasma membrane association, is unable to complement the growth phenotype of the *cdc25-2* allele at 36°C since mRas is not at the plasma membrane (PM). However, if the constitutively active cytoplasmic mRas is associated with the PM via protein–protein interactions, the growth phenotype of the *cdc25-2* mutant yeast at 36°C is complemented. In the reverse RRS, an unmodified integral membrane protein Y "bait" is expressed together with a "prey" protein X fused to mRas (X-mRas) within a *cdc25-2* mutant yeast strain (Hubsman *et al*., 2001) (Figure 2). If Y and X-mRas fail to interact, no growth of *cdc25-2* mutants will be observed at 36°C since X-mRas is not targeted to the PM (Figure 2a). However, if Y and X-mRas interact, X-mRas is recruited to the PM, thus allowing the growth of *cdc25-2* mutant cells at the restrictive temperature of 36°C (Figure 2b).

The reverse RRS has been used to isolate two novel interacting partners of the small GTPase Chp (Hubsman *et al*., 2001). Also, through the use of promoters expressing different levels of the X-mRas prey, Köhler and Müller were able to

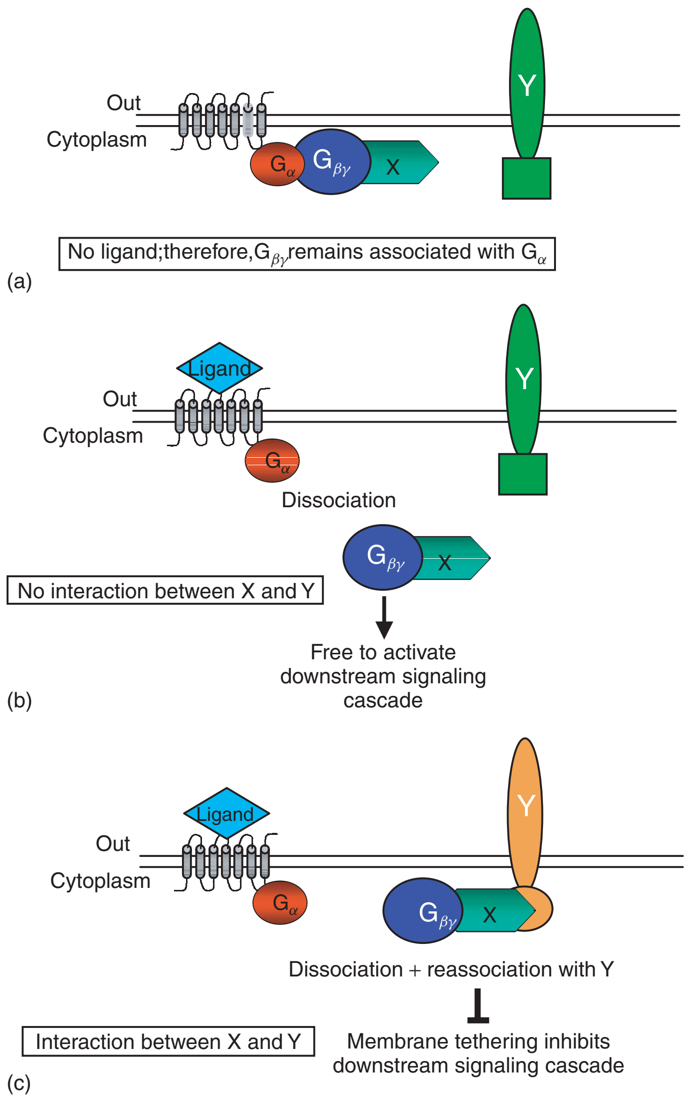


Figure 1 G-protein fusion assay: The G-protein fusion assay can test for interactions between a native integral membrane protein Y and a protein X, which is fused to the $G_{\beta\gamma}$ subunit of the G-protein coupled receptor ($G_{\beta\gamma}$-X). (a) When no ligand is present, the $G_{\beta\gamma}$ subunit of the G-protein coupled receptor ($G_{\beta\gamma}$-X) remains associated with the G_α subunit of the G-protein coupled receptor and no mating cascade events are activated. (b) If no interaction occurs between Y and $G_{\beta\gamma}$-X, upon ligand binding of the receptor, the $G_{\beta\gamma}$-X subunit dissociates from G_α and therefore is free to activate downstream signaling events. (c) If an interaction occurs between Y and $G_{\beta\gamma}$-X, the integral membrane protein Y sequesters $G_{\beta\gamma}$-X at the membrane, which in effect inhibits downstream signaling events although in the presence of ligand

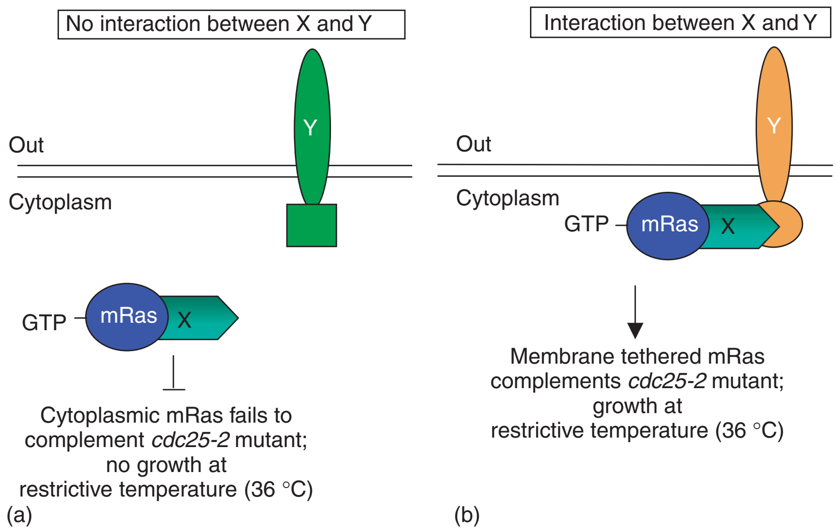


Figure 2 Reverse Ras recruitment system: The Reverse Ras recruitment system tests for interactions between a cytoplasmic constitutively active mRas, which is fused to protein X (X-mRas), and a membrane associated protein Y. (a) If no interaction occurs between proteins X-mRas and Y, protein X-mRas remains cytoplasmic and is unable to complement the growth phenotype of the mutant yeast strain *cdc25-2*. (b) However, if X-mRas and Y interact, X-mRas is then tethered to the membrane, which enables X-mRas to complement the growth phenotype of the mutant yeast strain *cdc25-2*

detect interactions between two membrane-associated proteins (Kohler and Muller, 2003). In this case, high to moderate expression levels of the membrane-associated prey, the adaptor protein growth factor receptor binding protein 2 (Grb2) fused to a myristoylation signal and mRas, supported only slight growth of the *cdc25-2* mutant yeast. Upon addition of an interacting bait, substantial increases in cell growth were observed presumably from the enhancement or stabilization of the X-mRas membrane localization, or perhaps clustering of the Ras facilitated further signaling (Kohler and Muller, 2003). Either way, for certain cases the limitation of using membrane associated proteins as preys, which normally would result in cell growth in the absence of protein interactions, has been eliminated. However, it is still possible that the Grb2 bait used by Köhler and Müller is a special case since Grb2 is a member of the Ras signaling pathway. Furthermore, the question still remains whether the RRS can be used to examine integral membrane proteins as baits.

2.3. Split-ubiquitin assay systems

The split-ubiquitin system provides another approach to study membrane protein interactions (Stagljar *et al*., 1998; Wittke *et al*., 1999). Ubiquitin (Ub) is a small, highly conserved protein that is attached to lysine residues of proteins in order to tag the proteins for proteasomal degradation (Hershko and Ciechanover, 1992). Ubiquitin-tagged proteins are recognized by ubiquitin-specific proteases (UBPs)

that cleave after the C-terminal (Gly 76) residue of Ub and the first residue of the target protein, allowing release of the protein for degradation by the 26 S proteasome. Johnsson and Varshavsky found that native ubiquitin can be split into an N-terminal (Nub) and a C-terminal (Cub) half (Johnsson and Varshavsky, 1994). The two halves retain a basic affinity for each other and spontaneously reassemble to form quasi-native ubiquitin. If a reporter protein is fused to the C-terminus of Cub, it will be cleaved off by UBPs upon assembly of the Nub and Cub moieties (Figure 3a). A point mutation in the N-terminal domain of ubiquitin (NubG) abolishes the affinity of the two halves for each other, such that NubG and Cub fail to refold into split-ubiquitin when coexpressed in yeast (Figure 3b). However, if the two ubiquitin halves are fused to the interacting proteins X and Y, this interaction brings the NubG and Cub moieties close enough together to reconstitute quasi-native Ub, resulting in the release of the reporter protein by the UBPs (Figure 3c).

The versatility of transcriptional activation of reporter genes in yeast was used to convert the split-ubiquitin system into a genetic assay for the in vivo detection of membrane protein interactions (Stagljar *et al*., 1998). In this membrane-based YTH assay (Figure 4), an artificial transcription factor (TF) consisting of the bacterial LexA protein and the *Herpes simplex* VP16 transactivator protein is fused to the Cub moiety. An integral membrane protein (X) is expressed as a fusion to the Cub-LexA-VP16 reporter cassette, with this cassette attached either to the N- or C-terminus of the transmembrane protein, depending on the orientation in the membrane of this protein. The second protein under investigation (Y), either another transmembrane protein or a cytoplasmic protein, is expressed as a fusion to NubG. If interaction between the X and Y proteins occurs, a split-ubiquitin molecule can be reconstituted, leading to the proteolytic release of the transcription factor to activate a reporter gene (Stagljar *et al*., 1998). Thus, the reassociation event initiated by the protein interaction is converted into a transcriptional output that can be easily

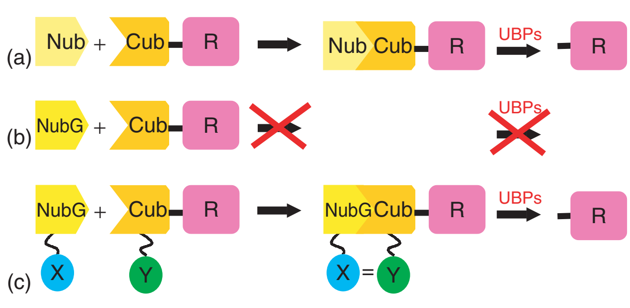


Figure 3 Split-ubiquitin: (a) Ubiquitin can be expressed as an N-terminal (Nub) half as well as a C-terminal (Cub) half that is fused to a reporter protein. The two halves retain affinity for each other and spontaneously reassemble to form the so-called split-ubiquitin. (b) A point mutation in the N-terminal half of ubiquitin (NubG) completely abolishes the affinity of the two halves for each other, and as the separate NubG and Cub parts are not recognized by ubiquitin-specific proteases (UBPs), no detectable cleavage of the attached reporter takes place. (c) NubG and Cub are fused to the interacting proteins X and Y. The X–Y interaction brings the NubG and Cub domains close enough together to reconstitute ubiquitin, resulting in the release of the reporter protein by the action of the UBPs

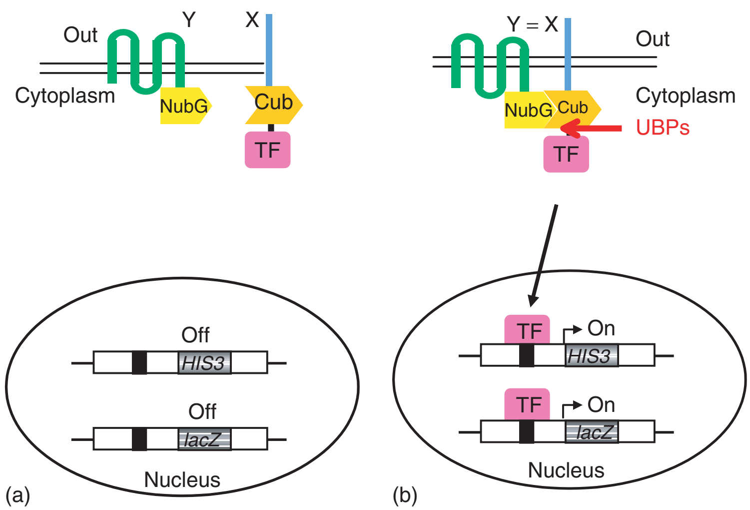


Figure 4 The split-ubiquitin membrane YTH system (MbYTH): (a) A membrane protein of interest X is fused to Cub followed by an artificial transcription factor (TF), while another membrane (or cytoplasmic) protein Y is fused to the NubG domain (Y-NubG). Co-expression of X-Cub-TF with a noninteracting Y-NubG does not lead to the formation of split-ubiquitin nor cleavage by UBPs. (b) On interaction of the X and Y proteins, ubiquitin reconstitution occurs, leading to proteolytic cleavage and the subsequent release of the transcription factor. This factor activates reporter genes to result in cells that are histidine prototrophs and that turn blue in a β-galactosidase assay

detected. This assay has been used to investigate the influence of mutations on the assembly of fragments of presenilin (a protein implicated in the Alzheimer's disease) (Cervantes *et al*., 2001), to characterize the interaction between the yeast α1,2-mannosidase Mns1p and Rer1p (Massaad and Herscovics, 2001) and Wbp1p and Sss1p in the endoplasmic reticulum (Scheper *et al*., 2003), and to study intra- and intermolecular interactions between plant sucrose transporters (Reinders *et al*., 2002). This approach has been successfully adapted for prey library screening and has identified three novel interacting partners of mammalian ErbB3, a receptor tyrosine kinase involved in regulation of proliferation and differentiation of many tissue types (Thaminy *et al*., 2003), and BAP31, a human polytopic integral membrane protein of the endoplasmic reticulum involved in apoptosis (Wang *et al*., 2003). Thus, MbYTH allows rapid and sensitive characterization of proteins associated with a particular full-length transmembrane protein of interest and is generally applicable to most transmembrane signaling proteins.

In another split-ubiquitin-based approach, Johnsson, and colleagues (Wittke *et al*., 1999) have fused a destabilized version of the yeast Ura3 protein, termed rUra3, to the Cub moiety. An integral membrane protein X is expressed as a fusion to the Cub-rUra3 cassette. If cells expressing the fusion protein are grown on a medium containing the compound 5-fluoro-orotic acid (5-FOA), they die because the rUra3 protein converts 5-FOA into a toxic product. However, if the cells coexpress an interacting protein Y fused to NubG, the Cub, and NubG moieties can be forced into close proximity by the X–Y interaction and associate to form

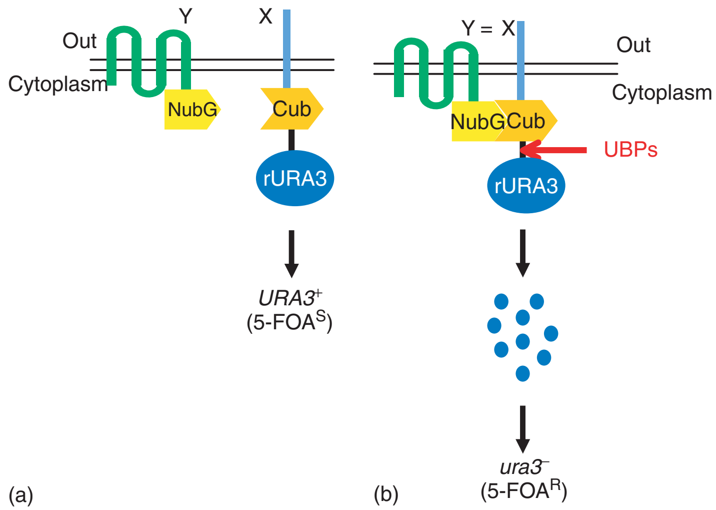


Figure 5 The rUra3-based membrane YTH system: (a) A membrane protein (X) under investigation is expressed as a fusion to the Cub domain, which is fused to a destabilized version of the Ura3 protein (rUra3). The NubG domain is linked to the membrane protein Y. If X and Y do not interact, there is no ubiquitin reconstitution and thus no UBP-mediated cleavage, resulting in yeast cells that contain Ura3 activity and thus die (5-FOA^S) on medium containing 5-fluoro-orotic acid (5-FOA), a toxic metabolite of the Ura3 enzyme. (b) If the X and Y proteins interact, the Cub and NubG domains are brought into close proximity, where they reconstitute an active ubiquitin. Cleavage by UBPs then releases rUra3 from the Cub fusion. The cleaved rUra3 is targeted for rapid destruction by the enzymes of the N-end rule to yield cells that are uracil auxotrophs *(ura3)* and 5-FOA resistant (5-FOA^R)

split-ubiquitin. This association, in turn, leads to UBP-mediated cleavage at the C-terminus of Cub and the release of the rUra3 fusion protein into the cytosol. Since the newly created N-terminal residue of the rUra3 protein is destabilizing in the N-end rule pathway of protein degradation (Varshavsky, 1997), the entire fusion protein is degraded by the 26 S proteasome, leading to cells that can grow on medium containing 5-FOA (Wittke *et al*., 1999). In this way, cells expressing two interacting proteins can be identified by their ability to survive selection on 5-FOA plates (Figure 5). The rUra3-based split-ubiquitin method was used to map the interactions between several *S. cerevisiae* integral membrane proteins (Wittke *et al*., 1999) and to analyze changes in protein conformation and stability of the *S. cerevisiae* protein Sec62, a component of the translocation machinery in the membrane of the endoplasmic reticulum (Raquet *et al*., 2001). In addition, the system has recently been used to systematically test the pairwise interactions between peroxins, proteins that play a role in protein import in the peroxisomes (Eckert and Johnsson, 2003).

3. Conclusions

Membrane-associated proteins perform a critical role in many essential cellular processes and thus represent one of the most important class of drug targets;

currently, they account for ~50% of all known pharmaceutical drug targets. However, analysis of membrane protein interactions is notoriously difficult because of the hydrophobic nature of such proteins, which makes them unsuitable bait candidates to use in most of the interactive proteomic technologies. The development of TAP technology linked to mass spectrometry analyses as well as the YTH system has provided a valuable means to identify proteins that physically interact in vivo (Fields and Sternglanz, 1994; Rigaut *et al*., 1999). Furthermore, in the past four years, the scientific community has witnessed the development of several genetic assays in yeast that are capable of detecting membrane protein interactions (Ehrhard *et al*., 2000; Hubsman *et al*., 2001; Stagljar *et al*., 1998; Urech *et al*., 2003; Wittke *et al*., 1999). In addition, genetic assays based on complementation of proteins or protein fragments have also been developed in organisms other than yeast that allow the monitoring of membrane protein interactions in real time, including ones amenable to use in mammalian cells (Blakely *et al*., 2000; Remy and Michnick, 1999; Rossi *et al*., 1997). Altogether, these assays represent an important step forward toward the elucidation of physical protein–protein interactions, since prior to their development the interactions of membrane proteins were difficult to study using the then existing biochemical and genetic methods. In the near future, a logical extension of the use of various such interactive proteomic technologies for membrane proteins will be their adaptation in a high-throughput format. This approach, in combination with automated screens, should help in elucidating membrane protein interactions on a genome-wide scale. Clearly, such approach will be highly instructive in understanding the pathways underlying many human diseases and will most likely reveal new targets for their therapy. Lastly, such knowledge can be used to provide predictors of certain diseases and may thus be used to develop biomarkers as fingerprints for particular disorders or cellular responses.

Acknowledgments

Research projects in my group are financed by grants from the Stiftung für medizinische Forschung, Novartis Foundation, Olga Mayenfish Foundation, Gebert Rüf Foundation, EU Grant HPRN-CT-2002-00240, the Swiss Cancer League (OCS-01310-02-2003), the Swiss National Science Foundation (Nrs. 31-58798.99 and 3100A0-100256/1), and the Zürcher Krebsliga.

References

Adams PD, Seeholzer S and Ohh M (2002) Identification of associated proteins by co-immunoprecipitation. In *Protein-protein Interactions (A Molecular Cloning Manual)*, Golemis DE (Ed.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 53–74.

Auerbach D, Thaminy S, Hottiger MO and Stagljar I (2002) The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics*, **2**, 611–623.

Blakely BT, Rossi FM, Tillotson B, Palmer M, Estelles A and Blau HM (2000) Epidermal growth factor receptor dimerization monitored in live cells. *Nature Biotechnology*, **18**, 218–222.

Cervantes S, Gonzalez-Duarte R and Marfany G (2001) Homodimerization of presenilin N-terminal fragments is affected by mutations linked to Alzheimer's disease. *FEBS Letters*, **505**, 81–86.

Eckert JH and Johnsson N (2003) Pex10p links the ubiquitin conjugating enzyme Pex4p to the protein import machinery of the peroxisome. *Journal of Cell Science*, **116**, 3623–3634.

Ehrhard KN, Jacoby JJ, Fu XY, Jahn R and Dohlman HG (2000) Use of G-protein fusions to monitor integral membrane protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1075–1079.

Einarson MB and Orlinich JR (2002) Identification of protein-protein interactions with glutathione-S-transferase fusion proteins. In *Protein-protein Interactions (A Molecular Cloning Manual)*, Golemis DE (Ed.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 37–57.

Fetchko M, Auerbach D and Stagljar I (2003) Yeast genetic methods for the detection of membrane protein interactions: potential use in drug discovery. *BioDrugs*, **1**, 413–424.

Fields S and Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Fields S and Sternglanz R (1994) The two-hybrid system: an assay for protein-protein interactions. *Trends in Genetics*, **10**, 286–292.

Hershko A and Ciechanover A (1992) The ubiquitin system for protein degradation. *Annual Review of Biochemistry*, **61**, 761–807.

Hubsman M, Yudkovsky G and Aronheim A (2001) A novel approach for the identification of protein-protein interaction with integral membrane proteins. *Nucleic Acids Research*, **29**, E18.

Johnsson N and Varshavsky A (1994) Split ubiquitin as a sensor of protein interactions in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 10340–10344.

Kohler F and Muller KM (2003) Adaptation of the Ras-recruitment system to the analysis of interactions between membrane-associated proteins. *Nucleic Acids Research*, **31**, e28.

Massaad MJ and Herscovics A (2001) Interaction of the endoplasmic reticulum alpha1,2-mannosidase Mns1p with Rer1p using the split-ubiquitin system. *Journal of Cell Science*, **114**, 4629–4635.

Melese T and Hieter P (2002) From genetics and genomics to drug discovery: yeast rises to the challenge. *Trends in Pharmacological Sciences*, **23**, 544–547.

Pedrazzi G and Stagljar I (2004) Protein-protein interactions. *Methods in Molecular Biology (Clifton, N.J.)*, **241**, 269–283.

Raquet X, Eckert JH, Muller S and Johnsson N (2001) Detection of altered protein conformations in living cells. *Journal of Molecular Biology*, **305**, 927–938.

Reinders A, Schultze W, Thaminy S, Stagljar I, Frommer WB and Ward JM (2002) Intra- and intermolecular interactions in sucrose transporters at the plasma membrane detected by the split-ubiquitin system and functional assays. *Structure (Cambridge)*, **10**, 763–772.

Remy I and Michnick SW (1999) Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5394–5399.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032.

Rossi F, Charlton CA and Blau HM (1997) Monitoring protein-protein interactions in intact eukaryotic cells by ß-galactosidase complementation. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 8405–8410.

Scheper W, Thaminy S, Kais S, Stagljar I and Römisch K (2003) Coordination of N-glycosylation and protein translocation across the endoplasmic reticulum membrane by Sss1 protein. *Journal of Biological Chemistry*, **278**, 37998–37903.

Stagljar I (2003) Finding partner(s): emerging protein Interaction technologies applied to signaling networks. *Science STKE*, **2003**, pe56.

Stagljar I and Fields S (2002) Analysis of membrane protein interactions using yeast-based technologies. *Trends in Biochemical Sciences*, **27**, 559–563.

Stagljar I, Korostensky C, Johnsson N and te Heesen S (1998) A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5187–5192.

Thaminy S, Auerbach D, Arnoldo A and Stagljar I (2003) Identification of novel ErbB3-interacting factors using the split-ubiquitin membrane yeast two-hybrid system. *Genome Research*, **13**, 1744–1753.

Thaminy S and Stagljar I (2002) The membrane-based yeast two-hybrid system. In *Protein-protein Interactions (A Molecular Cloning Manual)*, Golemis DE (Ed.), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, pp. 395–405.

Urech DM, Lichtlen P and Barberis A (2003) Cell growth selection system to detect extracellular and transmembrane protein interactions. *Biochimica et Biophysica Acta*, **1622**, 117–127.

Varshavsky A (1997) The N-end rule pathway of protein degradation. *Genes Cells*, **2**, 13–28.

Wang B, Nguyen M, Breckenridge DG, Stojanovic M, Clemons PA, Kuppig S and Shore GC (2003) Uncleaved BAP31 in association with A4 protein at the endoplasmic reticulum is an inhibitor of Fas-initiated release of cytochrome c from mitochondria. *Journal of Biological Chemistry*, **278**, 14461–14468.

Wittke S, Lewke N, Muller S and Johnsson N (1999) Probing the molecular environment of membrane proteins in vivo. *Molecular Biology of the Cell*, **10**, 2519–2530.

Short Specialist Review

Energy transfer–based approaches to study G protein–coupled receptor dimerization and activation

Ralf Jockers and Stefano Marullo
Institut Cochin, Paris, France

Michel Bouvier
Université de Montréal, Montréal, QC, Canada

1. Introduction

Because they play a key role in cell communication, G protein–coupled receptors (GPCRs) constitute a very active area of research and represent the largest group of potential drug targets for medicinal chemistry.

During the past decades, research on GPCRs has taken advantage of several technical advances, the energy transfer–based techniques (ETTs) being one of the most fruitful. These techniques allow the detection of protein–protein interactions and their dynamic variation at "physiological" expression levels in living cells.

ETTs are based on the nonradiative transfer of energy between a donor and an acceptor molecule. Because the efficacy of energy transfer (ET) varies inversely with the 6th power of the distance between the donor and the acceptor, ET from the donor will result in the emission of light by the acceptor only if the two molecules are in close proximity (10–100 Å). The detection of ET between proteins fused to energy donor and acceptor, respectively, is therefore a reflection of molecular interaction between the proteins of interest. A second important parameter, which influences the amount of ET to the acceptor, is its relative orientation to the donor. Accordingly, a variation of ET can also reflect conformational changes within the molecule, to which the donor or the acceptor is appended. Several variants of ET approaches have been used successfully to study GPCR biology. Fluorescence Resonance Energy Transfer (FRET) is the method of choice when the principal aim of the study is to identify the subcellular compartment in which the interaction between two proteins occurs. FRET, between an excited fluorescent donor and a fluorescent acceptor, can be measured with several methods, which have been recently reviewed elsewhere (Eidne *et al*., 2002). Bioluminescence Resonance Energy Transfer (BRET) occurring between luciferase, in the presence

of its substrate, and a fluorescent protein as acceptor is particularly well adapted for high-throughput tests (Boute *et al*., 2002).

2. ETTs to probe GPCR self-association and conformational changes induced by agonists

One of the most striking recent conceptual changes in the field of GPCRs has been the demonstration that these receptors, classically considered to function as monomers, are actually organized as homo- or hetero-oligomers (Table 1) (reviewed in Breitwieser, 2004 and Kroeger *et al*., 2003). Although the precise biological signification of the phenomenon is still a matter of debate, ETTs certainly played a major role in establishing that GPCR oligomerization is constitutive in living cells and that it occurs early in the biosynthetic pathway at “physiological” concentrations of receptors (Ayoub *et al*., 2002; Issafras *et al*., 2002; Terrillon *et al*., 2003). ETTs were also used to identify molecular domains involved in the oligomerization process in intact cells (Overton *et al*., 2003; Salahpour *et al*., 2004) and data obtained with ETTs contributed significantly to a model, which postulates that GPCR homo-oligomerization may be a prerequisite for their targeting from the endoplasmic reticulum to the plasma membrane (Salahpour *et al*., 2004).

A large number of studies have investigated the role of agonist activation on GPCR oligomer formation using ETTs. In most cases, GPCR activation did not result in significant changes of ET, consistent with the notion that constitutive oligomerization is independent of the activation state of the receptor (Ayoub *et al*., 2002). However, in apparent contradiction with the hypothesis above, the activation of some receptors was clearly associated with marked changes of ET (Eidne *et al*., 2002). Although some authors have interpreted these changes as evidence for agonist modulation of dimerization, studies addressing specifically this issue have concluded that agonist-dependent changes in ET most likely reflect internal conformational changes of the receptors (Ayoub *et al*., 2004). The possibility to observe ET changes after agonist challenge may depend on both the amplitude of conformational changes induced by the agonist in a given receptor and on the design of the ET experiment itself.

Conformational changes occurring upon receptor activation were directly probed by FRET experiments in which the donor and the acceptor were inserted at different positions within a single receptor protomer (Vilardaga *et al*., 2003). Different activation and deactivation kinetics of various GPCR (from millisecond to second) were measured in the presence of their agonists or antagonists, and correlations could be established between the timing of the hormonal effects elicited by the receptors and the rapidity of their conformational switch.

3. ETTs to probe signaling events downstream of GPCR activation

During the past several years, ETTs in living cells have challenged some aspects of one of the most solid models in biochemistry, the mechanism of heterotrimeric

Table 1 Energy transfer–based techniques to study GPCR biology: Key references

Biological phenomenon	References	Receptor	Major findings
GPCR dimerization	Overton *et al*. (2000)	Yeast pheromone receptor	Constitutive GPCR oligomerization shown by FRET
	Rocheville (2000)	Somatostatin and dopamine receptors	GPCR oligomerization shown by agonist-induced change of FRET
	Angers *et al*. (2000)	β2-adrenergic receptor	Constitutive GPCR oligomerization shown by BRET
	McVey *et al*. (2001)	β2-adrenergic and δ-opioid receptors	Monitoring receptor oligomerization by time-resolved fluorescence resonance energy transfer
	Mercier *et al*. (2002)	β1- and β2-adrenergic receptor	Self-association affinity of GPCR protomers measured by BRET
Identification of dimerization domains	Overton and Blumer (2002)	Yeast pheromone receptor	Characterization of the receptor dimerizaiton interface
	Overton *et al*. (2003)	Yeast pheromone receptor	Motifs involved in receptor dimerization
	Salahpour *et al*. (2004)	β2-adrenergic receptor	TM6 participates in the receptor dimerization interface
Detection of agonist-induced conformation changes	Vilardaga *et al*. (2003)	Parathyroid hormone and α2-adrenergic receptors	Rapid agonist-induced conformational changes monitored by intramolecular FRET
	Ayoub *et al*. (2002)	MT2 and MT1/2 melatonin receptors	Ligand-induced changes of BRET within preexisting dimers
Monitoring of ligand binding to GPCRs	Turcatti *et al*. (1996)	Tachykinin NK2 receptor	FRET between fluorescence-labeled GPCRs and fluorescent agonists
	Palanche *et al*. (2001)	Tachykinin NK2 receptor	Binding kinetics of a fluorescent agonist to a GFP-tagged GPCR, monitored by FRET
	Ayoub *et al*. (2004)	MT2 and MT1/2 melatonin receptors	EC_{50} values for ligand-induced changes of BRET correlate with ligand affinities
G protein activation	Janetopoulos *et al*. (2001)	Dictyostelium G protein	Monitoring of G_s protein activation by FRET
	Bunemann *et al*. (2003)	Mammalian G_i protein	Monitoring of G_i protein activation kinetics
β-arrestin recruitment	Angers *et al*. (2000)	β2-adrenergic receptor and β-arrestin	Agonist-induced translocation of β-arrestin to GPCRs

G protein activation. Receptor-mediated activation of G proteins was visualized in *Dictyostelium discoideum* by monitoring FRET between the $G\alpha_s$ and the $G\gamma$ subunits fused to energy donor and acceptor fluorescent proteins. According to the established model, receptor stimulation led to a decrease of FRET signal that was

interpreted as evidence for rapid dissociation of the subunits. Surprisingly, however, during continuous stimulation the G_s protein remained in its activated state despite the desensitization of the physiological response, suggesting that whether or not receptors are phosphorylated they catalyze the G protein cycle (Janetopoulos *et al*., 2001). A very similar strategy was followed by a different group, who studied G_i activation by GPCRs in mammalian cells. However, FRET changes were dependent on the specific position of the FRET acceptor on the $G\gamma$ subunit, both agonist-promoted increase and decrease in the signal being observed. This finding led the authors to suggest that the changes in FRET reflected conformational rearrangements of the heterotrimer rather than a true dissociation of the subunits (Bunemann *et al*., 2003). The detection of a persistent G protein heterotrimer during activation clearly questions our understanding of how G proteins propagate signals through either their $G\alpha$ or their $G\beta\gamma$ subunits.

ETTs are particularly well adapted to monitor the translocation dynamics of signaling and/or regulatory proteins to activated receptors. For example, the translocation of ß-arrestins to GPCRs has been monitored by BRET (Angers *et al*., 2000), and was proposed as a tool for the identification of orphan receptor ligands (Bertrand *et al*., 2002). In addition, although the ET-based monitoring of ß-arrestin translocation confirmed a concept that was previously delineated with other approaches, it constitutes a proof of concept of the potential interest of ETTs to investigate the dynamics of protein complex formation throughout GPCR-dependent signaling pathways.

References

Angers S, Salahpour A, Joly E, Hilairet S, Chelsky D, Dennis M and Bouvier M (2000) Detection of beta 2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3684–3689.

Ayoub MA, Couturier C, Lucas-Meunier E, Angers S, Fossier P, Bouvier M and Jockers R (2002) Monitoring of ligand-independent dimerization and ligand-induced conformational changes of melatonin receptors in living cells by bioluminescence resonance energy transfer. *The Journal of Biological Chemistry*, **277**, 21522–21528.

Ayoub MA, Levoye A, Delagrange P and Jockers R (2004) Preferential formation of MT1/MT2 melatonin receptor heterodimers with distinct ligand interaction properties compared to MT2 homodimers. *Molecular Pharmacology*, **66**(2), 312–321.

Bertrand L, Parent S, Caron M, Legault M, Joly E, Angers S, Bouvier M, Brown M, Houle B and Menard L (2002) The BRET2/arrestin assay in stable recombinant cells: a platform to screen for compounds that interact with G protein-coupled receptors (GPCRS). *Journal of Receptor and Signal Transduction Research*, **22**, 533–541.

Boute N, Jockers R and Issad T (2002) The use of resonance energy transfer in high-throughput screening: BRET versus FRET. *Trends in Pharmacological Sciences*, **23**, 351–354.

Breitwieser GE (2004) G protein-coupled receptor oligomerization: implications for G protein activation and cell signaling. *Circulation Research*, **94**, 17–27.

Bunemann M, Frank M and Lohse MJ (2003) Gi protein activation in intact cells involves subunit rearrangement rather than dissociation. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 16077–16082.

Eidne KA, Kroeger KM and Hanyaloglu AC (2002) Applications of novel resonance energy transfer techniques to study dynamic hormone receptor interactions in living cells. *Trends in Endocrinology and Metabolism*, **13**, 415–421.

Issafras H, Angers S, Bulenger S, Blanpain C, Parmentier M, Labbe-Jullie C, Bouvier M and Marullo S (2002) Constitutive agonist-independent CCR5 oligomerization and antibody-mediated clustering occurring at physiological levels of receptors. *The Journal of Biological Chemistry*, **277**, 34666–34673.

Janetopoulos C, Jin T and Devreotes P (2001) Receptor-mediated activation of heterotrimeric G-proteins in living cells. *Science*, **291**, 2408–2411.

Kroeger KM, Pfleger KD and Eidne KA (2003) G-protein coupled receptor oligomerization in neuroendocrine pathways. *Frontiers in Neuroendocrinology*, **24**, 254–278.

McVey MD, Ramsay D, Kellett E, Rees S, Wilson S, Pope AJ and Milligan G (2001) Monitoring receptor oligomerization using time-resolved fluorescence resonance energy transfer and bioluminescence resonance energy transfer. The human delta-opioid receptor displays constitutive oligomerization at the cell surface, which is not regulated by receptor occupancy. *The Journal of Biological Chemistry*, **276**, 14092–14099.

Mercier JF, Salahpour A, Angers S, Breit A and Bouvier M (2002) Quantitative assessment of beta 1 and beta 2-adrenergic receptor homo and hetero-dimerization by bioluminescence resonance energy transfer. *The Journal of Biological Chemistry*, **277**, 44925–44931.

Overton MC and Blumer KJ (2000) G-protein-coupled receptors function as oligomers in vivo. *Current Biology*, **10**, 341–344.

Overton MC and Blumer KJ (2002) Use of fluorescence resonance energy transfer to analyze oligomerization of G-protein-coupled receptors expressed in yeast. *Methods*, **27**, 324–332.

Overton MC, Chinault SL and Blumer KJ (2003) Oligomerization, biogenesis, and signaling is promoted by a glycophorin A-like dimerization motif in transmembrane domain 1 of a yeast G protein-coupled receptor. *The Journal of Biological Chemistry*, **278**, 49369–49377.

Palanche T, Ilien B, Zoffmann S, Reck MP, Bucher B, Edelstein SJ and Galzi JL, (2001) The neurokinin A receptor activates calcium and cAMP responses through distinct conformational states. *The Journal of Biological Chemistry*, **276**, 34853–34861.

Rocheville M, Lange DC, Kumar U, Patel SC, Patel RC and Patel YC (2000) Receptors for dopamin and somatostatin: formation of hetero-oligomers with enhanced functional activity. *Science*, **288**, 154–157.

Salahpour A, Angers S, Mercier JF, Lagace M, Marullo S and Bouvier M (2004) Homodimerization of the beta 2-adrenergic receptor as a pre-requisite for cell surface targeting. *The Journal of Biological Chemistry*, **279**(32), 33390–33397.

Turcatti G, Nemeth K, Edgerton MD, Meseth U, Talabot F, Peitsch M, Knowles J, Vogel H and Chollet A (1996) Probing the structure and function of the tachykinin neurokinin-2 receptor through biosynthetic incorporation of fluorescent amino acids at specific sites. *The Journal of Biological Chemistry*, **271**, 19991–19998.

Terrillon S, Durroux T, Mouillac B, Breit A, Ayoub MA, Taulan M, Jockers R, Barberis C and Bouvier M (2003) Oxytocin and vasopressin V1a and V2 receptors form constitutive homo-and heterodimers during biosynthesis. *Molecular Endocrinology*, **17**, 677–691.

Vilardaga JP, Bunemann M, Krasel C, Castro M and Lohse MJ (2003) Measurement of the millisecond activation switch of G protein-coupled receptors in living cells. *Nature Biotechnology*, **21**, 807–812.

Short Specialist Review

Protein interaction databases

Henning Hermjakob and Rolf Apweiler
European Bioinformatics Institute, Cambridge, UK

Protein interactions occur in many different forms: Proteins interact with and modify each other as elements of signaling chains; they form large molecular machines with more than 100 protein and nonprotein elements, for example, the spliceosome; or, they assemble to provide structural elements of the cell. Proteins interact not only with each other but also with DNA, RNA, and other molecule types. Types of interaction may vary from direct physical contact to very indirect interactions, for example, participation in the same signaling pathway. In many contexts, protein states, in particular, posttranslational modifications, are essential for the function of protein assemblies.

The kind of data produced by experimental technologies to elucidate protein interaction data is similarly diverse, from direct, atom-level structural information by NMR (Pellecchia *et al*., 2002) or X ray via evidence for direct physical contact, for example, by yeast two hybrid assays (Phizicky *et al*., 2003) and information on participation in the same complex by tandem affinity purification (Rigaut *et al*., 1999) to indirect information, for example, colocalization.

With the introduction of technologies for high-throughput interaction screens (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002), the number of interactions published in a single publication now varies over four orders of magnitude, from single interactions to more than 20 000 (Giot *et al*., 2003).

The biological and medical importance of protein interactions, high-throughput technologies, and the broadness of data types and experimental technologies have led to the creation of a significant number of protein interaction data collections, from simple HTML tables providing the results of a particular experiment to large databases providing complex tools. We subsequently introduce some criteria for the comparison of protein interaction databases, and describe some major, publicly accessible protein interaction databases on the basis of these criteria (Table 1), focusing on databases for experimentally, not computationally, derived datasets.

Confidence information: High-throughput experiments are often considered less reliable than detailed, low-throughput experiments. To support the user in estimating the reliability of individual interactions, the authors of large-scale interaction data sets often provide quality indications, for example, the grouping into reliability classes (Li *et al*., 2004) or numerical values (Giot *et al*., 2003). In addition, comparative analysis of different interaction data sets provides quality indications. In the simplest case, this is the number of times an interaction has been observed in independent experiments, but quality assessments are also based, for example, on the comparison against a reference data set considered to be correct (von Mering

Table 1 Contents and features of publicly available protein interaction databases

Acronym	BIND	CYGD	DIP	GRID
Name	Biomolecular Interaction Network Database	Comprehensive Yeast Genome Database	Database of Interacting Proteins	General Repository for Interaction Datasets
Description	The Biomolecular Interaction Network Database (BIND) is a collection of records documenting molecular interactions. The contents of BIND include high-throughput data submissions and hand-curated information gathered from the scientific literature.	The MIPS Comprehensive Yeast Genome Database (CYGD) aims to present information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, the budding yeast Saccharomyces cerevisiae.	The DIP database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein–protein interactions.	The GRID is a database of genetic and physical interactions developed in the Tyers Group at the Samuel Lunenfeld Research Institute at Mount Sinai Hospital.
Contents	94 368 interactions (including genetic interactions)	15 488 (9103 physical and 6385 genetic)	44 349 protein interactions	54 000 genetic and protein interactions
Species range	All	*Saccharomyces cerevisiae*	All (107 organisms)	Yeast, Fly, Worm, Human, Mouse, Zebrafish, *S. pombe*, and Rat
Search	Simple search box, and field-specific search interface for complex queries. Search by GI, gene name, PubMed id, protein description. Blast search. Returns 21 interactions for ”lsm7”.	Simple search box. Search by ORF/gene name or PubMed id. Cross-referenced from well-annotated Yeast Genome Database.	Four search forms. Search by description, prosite motif or publication. Blast search. Search by ”lsm7” returned no result, recommended way is to search UniProt, then use the link there, or the sequence.	Simple search box. Search by GI number, ORF/Gene name. Search for ”lsm7” returns 20 interaction partners.
Visualization	Java WebStart application. Network viewer with unusual ”ontoglyphs” to visualize protein properties. Support for Cytoscape in download files.	The CYGD site was being updated at the time of writing, the visualization system could not be accessed.	Graph view with interaction reliability projected on edge colors.	”Osprey” application. Requires local installation, provides powerful graph exploration and integration of user-supplied data. No connection to GRID search, Osprey search only on current dataset.

Confidence information	Indirect, number of PubMed abstracts per interaction.	Not available.	Interaction confidence classification according to three different methods. Grouping of interactions into reliable "core" and less reliable "noncore".	Indirect, list of publications.
Data availability	Free to academic and commercial users.	Free to academic users. License required for commercial use. No redistribution.	Free to academic end users, no redistribution allowed. Flat file download after registration. License required for commercial use.	Free to academic end users, no redistribution allowed. Flat file download after registration. License required for commercial use.
Data structure	Highly detailed, complex data structure. Based on NCBI data types. Published and available from website.	Database structure unpublished.	Relatively simple relational table structure. Overview described in Salwinski *et al*. (2004).	Tab-delimited format.
Download formats	Various full and subset files in ASN.1, Fasta, XML (BIND-specific) format	Tab-delimited files.	XIN (DIP-specific) XML format, PSI-MI XML, Fasta	Tab-delimited format.
Software availability	Mainly C/C++, source code available.	Not available.	Not available.	Not available.
Documentation	Well-documented, including data submission and curation manuals.	Project descriptions.	Usage and search guide documents.	No Grid documentation, but message-based support forum. On-line manual for Osprey.
URL	**http://www.blueprint.org/bind/bind.php**	**http://mips.gsf.de**	**http://dip.doe-mbi.ucla.edu/dip/Main.cgi**	**http://biodata.mshri.on.ca/grid/servlet/Index**
Reference	Bader *et al*. (2003)	Mewes *et al*. (2004)	Salwinski *et al*. (2004)	Breitkreutz *et al*. (2003)
Notes			Separate satellite database providing information on protein state.	
Visited on	July 2, 2004	July 5, 2004	July 5, 2004	July 5, 2004

(*continued overleaf*)

Table 1 (*continued*)

Acronym	HPRD	Hybrigenics	IntAct	MINT
Name	Human Protein Reference Database	Hybrigenics S.A.	Open Source Database of Molecular Interactions	Molecular Interactions Database
Description	The Human Protein Reference Database represents a centralized platform to visually depict and integrate information pertaining to domain architecture, posttranslational modifications, interaction networks, and disease association for each protein in the human proteome.	PIMRider® is Hybrigenics' functional proteomics software platform, dedicated to the exploration of protein pathways. Based on reliable Protein Interaction Maps (PIM®), PIMRider® leads to the unraveling of biological functions.	IntAct provides an open source database and toolkit for the storage, presentation, and analysis of protein interactions. IntAct data is curated from large and small-scale experiments.	MINT is a relational database designed to store interactions between biological molecules. Presently, MINT focuses on experimentally verified protein interactions with special emphasis on proteomes from mammalian organisms.
Contents	15 944 protein interactions	4200 protein interactions	37 680 protein interactions	42 534 interactions
Species range	Human	Human, *Drosophila*, *H. pylori*	All (120 species)	All
Search	Multifield search form. Search by protein name, gene name, posttranslational modification, GO, and domain annotation of proteins.	Simple search box. Search by gene name/description.	Simple search box. Search by gene name, UniProt, InterPro, GO, PubMed, model organism database accession numbers.	Multifield search form. Search by UniProt, InterPro, PDB, GO, PubMed identifiers, gene names, UniProt keywords.
Visualization	Visualization of interacting domains. No graph-oriented visualization.	PIMRider® visualization tool. Java-based for local installation. Graph view, linked to view or interacting domains. Filtering on confidence score (PBS).	Graph visualization, highlighting of graph nodes according to GO annotation.	Java Applet Mint viewer with visualization of number of times an interaction has been observed.
Confidence information	Indirect, by indication of experiment type.	Full integration of PBS score Rain *et al*. (2001)	Indirect, by number of experiments.	Indirect, by number of experiments.

Data availability	Free to academic users. File download after registration. License required for commercial use.	TGF-Beta, Drosophila datasets free for all users after registration. H.pylori dataset free to academic users. File download after registration. License required for commercial use.	Free to academic and commercial users.	File download after registration. No statement on commercial use of redistribution restrictions.
Data structure	Moderately complex, no detail documentation, only autogenerated UML diagram.	Not documented.	Moderately complex, UML diagram and relational schema available on the web.	Data structure follows the PSI-MI standard.
Download formats	HRPD-specific XML format, PSI-MI XML.	PSI-MI XML.	PSI-MI XML, GO formatted controlled vocabularies.	PSI-MI XML.
Software availability	Open source availability announced, but not yet there.	PIMWalker tool (PIMRider with reduced features, but capable of visualizing PSI-MI files) available for noncommercial use after registration. Contains partial source code.	Open source, well-documented and freely available, with installation instructions.	Not available.
Documentation	Extensive FAQ documents.	Compact, but comprehensive PIMRider on-line manual.	Detailed user manual.	On-line documentation.
URL	**http://www.hprd.org/**	**http://pim.hybrigenics.com/**	**http://www.ebi.ac.uk/intact**	**http://mint.bio.uniroma2.it/mint/**
Reference	Peri *et al*. (2004)	Rain *et al*. (2001)	Hermjakob *et al*. (2004b)	Zanzoni *et al*. (2002)
Notes			Dynamic download of interaction networks in PSI-MI format, for example, for Cytoscape support.	
Visited on	July 5, 2004	July 6, 2004	July 5, 2004	July 6, 2004

et al., 2002), shared subcellular location and cellular role annotation of interacting proteins (Sprinzak *et al*., 2003), comparison to RNA expression profiles, or similar interactions in paralogous sequences (Deane *et al*., 2002).

Standards: The number of interactions in the yeast *Saccharomyces cerevisiae* interactome has been estimated to be about 10 000–26 000 (Sprinzak *et al*., 2003; Grigoriev, 2003). Taking into account all relevant species and the change of protein interactions depending on protein and cellular state, the number of interactions is nearly unlimited. As it is unlikely that any given database will be able to collect all available protein interaction data, collecting data from several, potentially specialized databases is essential to assemble a reasonably complete picture of the currently available protein interaction data in a given domain. In addition to providing a complete picture, such collections also provide the basis for interaction data reliability assessments through comparative analysis. However, such collection may be difficult and labor-intensive due to different data formats and annotations used by different databases. To improve this situation, major interaction data providers, among them BIND (Biomolecular Interaction Network Database), DIP (Database of Interacting Proteins), Hybrigenics, HPRD (Human Protein Reference Database), IntAct, MIPS, and MINT, in the framework of the HUPO Proteomics Standards Initiative (**http://psidev.sf.net**) (*see* Article 61, **Data standardization and the HUPO proteomics standards initiative**, Volume 7) have jointly developed the PSI-MI XML format, a community standard for the representation of protein interaction data. PSI-MI 1.0 provides a basic exchange format for protein interaction data (Hermjakob *et al*., 2004a); level 2.0 is under development (Orchard *et al*., 2004) and will provide additional features and an extension to additional molecule types, in particular, RNA and DNA. In addition to a standard format, PSI-MI provides a set of interaction-specific controlled vocabularies to standardize not only the format but also the contents of protein interaction data, for example, to select all data derived by a certain technology.

Visualization: In scientific publications, protein interactions are often visualized as groups of adjacent shapes with textual annotation. While such representations are intuitive, they are difficult to generate automatically. For automatically generated visualization, nearly all tools are based on the abstraction of proteins as nodes and interactions as edges in a graph. This representation allows the application of well-established graph layout algorithms to interaction networks, and provides the basis for additional analysis types, for example, interaction distance, or the identification of clusters of highly interconnected proteins. Although mostly based on this basic abstraction, tools differ significantly in the technical implementation, user interface, and, in particular, in methods to project additional information onto such interaction networks, for example, interaction confidence or Gene Ontology (Harris *et al*., 2004) terms annotated to the interacting proteins. In addition to the tools provided directly by databases, additional tools for graph-based interaction network analysis are provided by commercial and academic organizations, for example, Cytoscape (Shannon *et al*., 2003). Figure 1(a–e) shows the display of lsm7 (yeast) interactions in some visualization systems. Default settings have been used as much as possible.

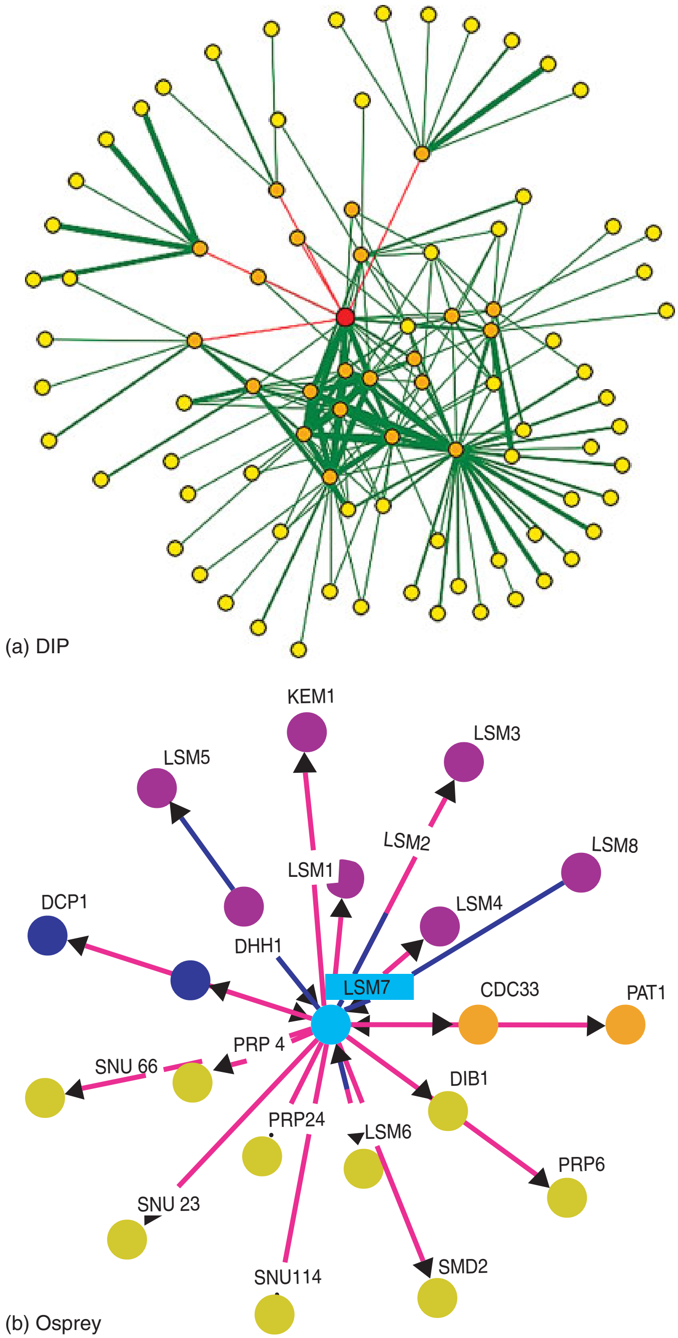
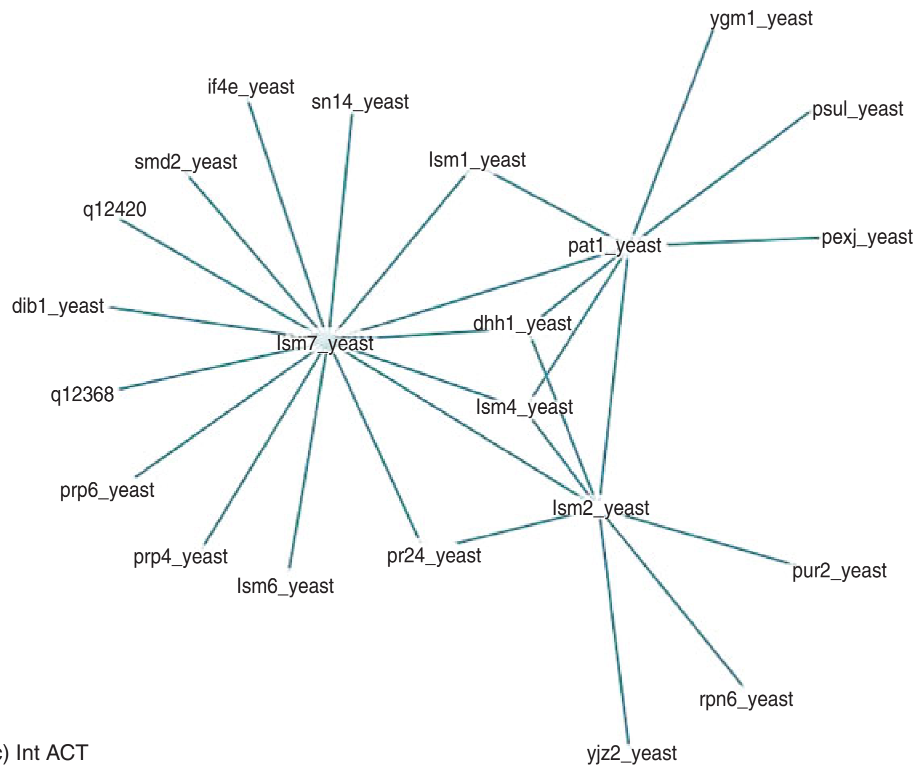
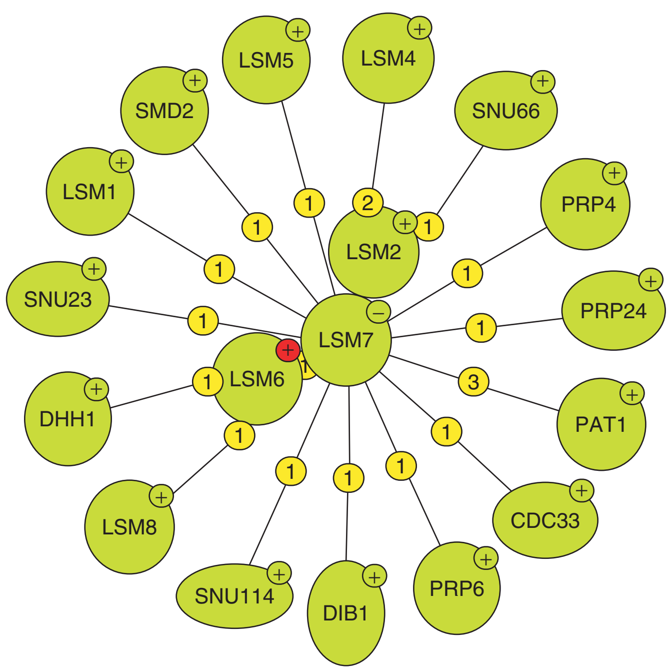


Figure 1 (a) Visualization of lsm7 (yeast) interactions in DIP. (b) Visualization of lsm7 (yeast) interactions in Osprey. (c) Visualization of lsm7 (yeast) interactions in IntAct. (d) Visualization of lsm7 (yeast) interactions in MINT. (e) Visualization of lsm7 (yeast) interactions in BIND

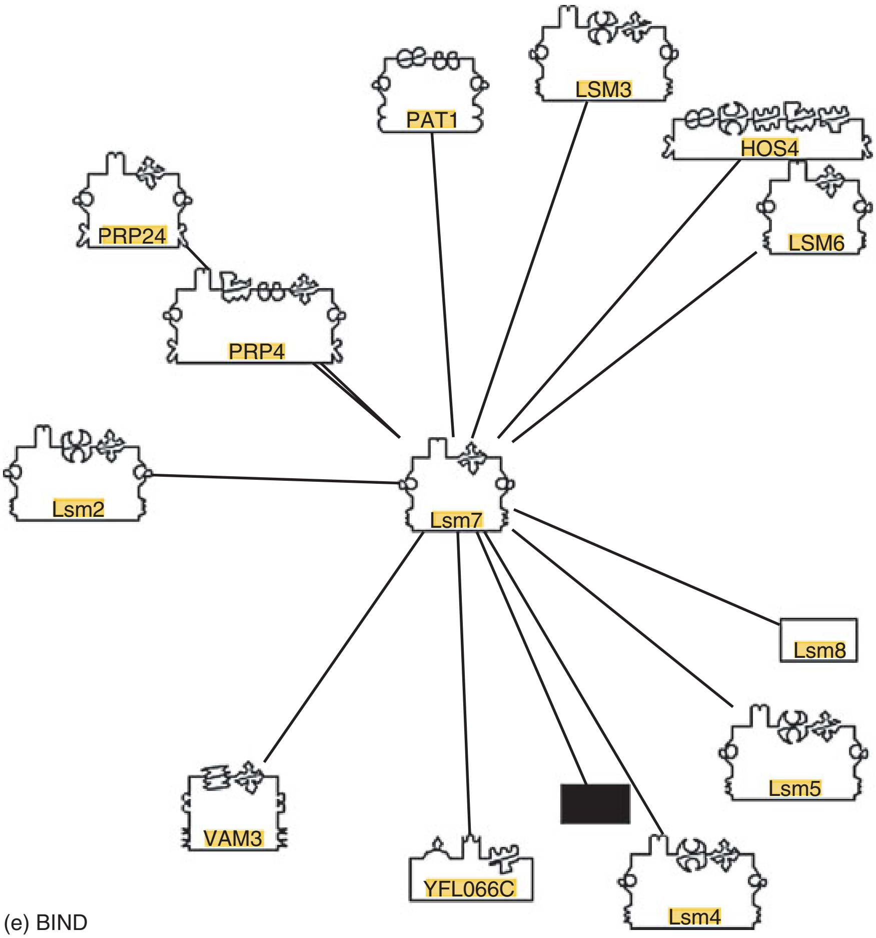


(c) Int ACT



(d) MINT

Figure 1 *(continued)*



(e) BIND

Figure 1 (*continued*)

The basic knowledge about protein interactions is currently increasing at a rapid pace, and a broad array of databases provides large data collections and powerful analysis tools, but the representation of protein interaction facts in near-textbook quality, taking into account interaction details such as protein state and dissociation constants, remains a challenge to the scientific community, both in terms of available tools and available data. Unlike other domains of molecular biology, in particular, DNA and protein sequences as well as macromolecular structures, protein interactions share the problem of data fragmentation in proteomics. Systematic deposition of data in public databases, and exchange of such data between databases, is only emerging. However, the standardization of interaction data through the PSI-MI standard, an increasing collaboration between protein interaction databases, and publicly available, open source analysis tools pave the way to public, user-friendly, and well-accessible protein interaction data resources, reflecting the huge biological significance of protein interactions.

References

Bader GD, Betel D and Hogue CW (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Research*, **31**(1), 248–250.

Breitkreutz BJ, Stark C and Tyers M (2003) The GRID: the general repository for interaction datasets. *Genome Biology*, **4**(3), R23.

Deane CM, Salwinski L, Xenarios I and Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics: MCP*, **1**(5), 349–356.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM and Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B and Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**(5651), 1727–1736.

Grigoriev A (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, **31**(14), 4157–4161.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, Database issue, D258–D261.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, OrchardS, Sarkans U, von Mering C, *et al*. (2004a) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**(2), 177–183.

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, VingronM, Roechert B, Roepstorff P, Valencia A, *et al*. (2004b) IntAct: an open source molecular interaction database. *Nucleic Acids Research*, **32**, Database issue, D452–D455.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**(6868), 180–183.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**(5657), 540–543.

Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, PagelP, Strack N, Stumpflen V, *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, **32**, Database issue, D41–D44.

Orchard S, Taylor CF, Hermjakob H, Zhu W, Julian RK Jr and Apweiler R (2004). Advances in the development of common interchange standards for proteomic data. *Proteomics*, **4**(8), 2363–2365.

Pellecchia M, Sem DS and Wuthrich K (2002) NMR in drug discovery. *Nature Reviews. Drug Discovery*, **1**(3), 211–219.

Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, **32**, Database issue, D497–D501.

Phizicky E, Bastiaens PI, Zhu H, Snyder M and Fields S (2003) Protein analysis on a proteomic scale. *Nature*, **422**(6928), 208–215.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**(6817), 211–215.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**(10), 1030–1032.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Research*, **32**, Database issue, D449–D451.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11), 2498–2504.

Sprinzak E, Sattath S and Margalit H (2003) How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, **327**(5), 919–923.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**(6770), 623–627.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G (2002) MINT: a molecular INTeraction database. *FEBS Letters*, **513**(1), 135–140.

Short Specialist Review

Computational methods for the prediction of protein interaction partners

Alfonso Valencia and David Juan
CNB-CSIC, Protein Design Group, Madrid, Spain

1. Genome- and sequence-based methods

1.1. Phylogenetic profiles

The Phylogenetic profiles method (Pellegrini *et al*., 1999) rests on the very intuitive concept of co-occurrence of related proteins in various genomes. This co-occurrence can be studied by comparing profiles of the YES/NO presence profiles (known as *phylogenetic profiles*) of a given protein pair in a set of complete genomes from/of different organisms (*see* Article 41, **Phylogenetic profiling**, Volume 7 and Article 48, **Connecting genes by comparative genomics**, Volume 7). The basis for assuming a relation between the corresponding proteins grows stronger in direct proportion to the similarity of these profiles. The main advantage of this method is its simplicity, but one of its main drawbacks is the difficulty of distinguishing between functional and speciation/lateral transferences and major speciation events (*see* Article 66, **Methods for detecting horizontal transfer of genes**, Volume 4 and Article 40, **The domains of life and their evolutionary implications**, Volume 7), that is, gram-positive versus gram-negative specific protein-driven profiles. Other problems are related to the method's inability to deal with complex cases such as multidomain proteins, ortholog replacement of a protein by other proteins, or relationships dependent on the organism's environment. However, regardless of the drawbacks and limitations, the idea is still powerful enough to detect interesting sets of potential functional relations.

1.2. Conservation of gene neighboring

This method (Dandekar *et al*., 1998) is based on the well-known organization of bacterial genomes in operons as basic units of transcription coordination of functional-related genes. These organisms tend to cluster in their genomes sets of genes whose expression needs to be concerted. The clearest case is operon

organization. In this method, the order to filter out those cases due to chance conservation of the neighborhood in a number of organisms is translated into statistical evidence of the required functional relation. This method is obviously very powerful in predicting functional links between bacterial genes. However, its validity for eukaryotic organisms remains unclear. Although the relationship established between two genes by this method is fundamentally functional, the strength of this relationship is to point up its usefulness for retrieving sets of potential physically interacting proteins, thus providing an interesting way of drastically reducing the number of proteins to be explored in a particular experiment.

1.3. Gene fusion

Another set of methods focuses on finding out those genes that appear fused in some genome composition (Enright *et al*., 1999; Marcotte *et al*., 1999). The principle is that some organisms have chosen to take advantage of the functional advantages provided by these fusion events, probably related to the organization of protein domains in a single polypeptide. This indicates their potential functional relation. This type of fusion event is apparently more frequent in metabolic proteins (Tsoka and Ouzounis, 2000). An advantage of this method is that it can be applied to eukaryotic genomes, where fusion events in large multidomain proteins are possibly more frequent than in bacterial genomes, even if the extent to which this is true remains still to be quantified. It is also unclear exactly what biological or biochemical advantage is derived from the protein domain fusion.

1.4. Correlated mutations (in silico two hybrid)

This method and the one that follows it (MirrorTree) increase the level of detail in the study of a protein–protein interaction. These two methods use multiple sequence alignments as input information instead of genome composition or organization. Multiple sequence alignments are efficient tools for exploring evolutionary relationships, and they provide a powerful handle on protein evolutionary history.

In the case of the in silico two-hybrid method (Pazos and Valencia, 2002), the relations between a particular pair of proteins are deduced from the detection of correlated mutation changes between the corresponding multiple sequence alignments of these proteins. These interprotein correlated mutations are supposed to be the consequence of coadaptation clues of strong functional relationships and are specifically related to a physical interaction. From a methodological point of view, the signal of correlation is weak, and, in practice, it is difficult to distinguish when/whether/why/how inter- and intraprotein-correlated mutations are detected. Thus, an interaction index is calculated on the basis of the corresponding distributions of possible intra- and interprotein interaction pairs.

1.5. Similarity of phylogenetic trees (MirrorTree)

The second method using multiple sequence alignment information is based on the observed similarity of the gene trees of related proteins (Pazos and Valencia,

2001). This relationship was observed in a number of cases and first quantified by Goh *et al*. (2000). The statistical demonstration of this principle was obtained by analyzing a large set of interacting proteins (Pazos and Valencia, 2001). The interpretation of these observations points to the coevolution of interacting proteins as an important factor in protein evolution that resulted in the detectable similarity of their gene trees. In practice, the method compares the distance matrices of the corresponding sequences instead of the deduced gene trees. These matrices are implicit representations of phylogenetic trees. The similarity between distance matrices is evaluated by using a correlation formulation for the pair of matrices (Pazos and Valencia, 2001).

Both the in silico two-hybrid and the MirrorTree methods use family alignments as a starting point. In both cases, their comparison requires a reduction to sequences of organisms present in both protein families. This reduction enables exploration of the evolution of protein pairs in a comparable framework. However, it also implies a high amount of available sequence information.

Each of these five methods represents a different level of detail in the use of genomic information. The first three of them (phylogenetic profiles, gene neighboring, and gene fusion) work on genes as a whole, exploring global genomic relationships between them (simultaneous presence, genome location, and domain composition). The other two, In silico two-hybrid and MirrorTree, can be classified at a different level of complexity, since they use more detailed information exploring protein sequences and not just whole gene relationships. The increasing level of descriptive detail represents not only some advantages but also some drawbacks. The negative points are related to the reduction of coverage regarding the additional constraints imposed by constructions of the multiple sequence alignments and to the noise introduced by the bias in the distribution of sequences in the corresponding multiple sequence alignments. On the positive side, these methods offer the additional possibility of getting more direct insights into the physical relation between proteins and protein residues, beyond the general functional relations discovered with the genome-based methods.

1.6. From the prediction of interaction partners to the prediction of interacting regions

All these computational methods for the prediction of protein interactions are reasonable alternatives for the prediction of protein interaction partners and a good source of information about the global structure of the interaction network. Together with these general predictions, the problem in molecular biology is how to translate these predictions to the molecular details of the corresponding interactions.

Often, the question of which protein is interacting with a given query protein is followed by questions about the possible region of interaction between the predicted interaction partners. Alternative experimental approaches could be the solution to the structure of the predicted protein complex or the systematic mutagenesis of the potential surface residues (Alanine scanning approach). Of the computational methods described here, those based on structural models, as well as the in silico two-hybrid and MirrorTree methods, provide competitive alternatives. The

in silico two-hybrid and MirrorTree methods have been used for the prediction of interacting surfaces in a number of cases, including preliminary approaches for their combination with physical docking methods (for an overview, *see* Valencia and Pazos, 2003).

1.7. Comparison of genome- and sequence-based method

The first three genome-based methods have been extensively tested against large collections of known proteins extracted from different experimental sources (von Mering *et al*., 2002). In this case, it was clear that the capacity of the computational methods for predicting functional interactions is greater than other methods based on the similarity of expression profiles or experimental yeast-two-hybrid approaches, and only slightly lower than that obtained with biochemical complex pull-down approaches. The analysis of the type of interactions detected by the various experimental and computational methods reveals the specialization of the methods in the detection of interactions with different functions and interaction types (Hoffmann and Valencia, 2003).

We carried out a comparison of the five computational methods on a large set of *Escherichia coli* protein, using the information from 44 fully sequenced bacterial genomes to generate the corresponding family alignments and genome comparisons. The main conclusion of this study is that the five methods have a similar capacity for predicting functional relations in terms of the number of predicting interactions and the accuracy of the predictions. All the methods show the typical sensibility-specificity trade-off, and their coverage is related to their underlying principles. Thus, for example, phylogenetic profiles provide a higher number of interesting predictions than gene fusion (which are less commonly detected). In turn, it includes a larger number of false positives, particularly physical interactions, since its foundation is more closely related to general functional interaction than to physical interaction. This situation is also visible for those methods using multiple sequence alignments, where the increase in specificity in the detection of physical interactions comes with a reduction in the number of predictions.

From these comparative studies, it is clear that the accuracy of the predictions increases when they are validated by more than one method. In the comparative study of the five methods, 37 protein pairs were predicted as interacting by three or more methods and 83.7% of them were found to correspond to proteins of a common biochemical pathway (KEGG and EcoCyc databases) or components of known complexes (DIP database), 56.7% being identified as interacting proteins by automatic exploration of Medline abstracts. These confirmation levels justify the capability of these computational methods to provide reliable predictions, which makes them valuable sources of information that can be combined with high-throughput experimental approaches.

2. Structure-based method

In addition to these sequence- and genome-based prediction methods, a set of alternative approaches uses additional structural information (*see* Article 35, **Structural**

biology of protein complexes, Volume 5). Aloy and Russell (2002) have described a method to derive protein–protein interactions based on homology with complexes of known structure. This method evaluates the probability of homologous complexes by using an empirical interaction potential derived from a known complex. A related approach uses threading (*see* Article 72, **Threading for protein-fold recognition**, Volume 7) to build models of unrelated sequence proteins (Lu *et al*., 2002). In this case, the prediction of interactions is based on the fitness of the protein sequences in the corresponding complexes evaluated with the corresponding threading potentials. In principle, this method aims for a higher sensitivity at the expense of a lower specificity than the homology-based method. In any case, the applicability of these approaches is still limited by the relatively small number of known protein complexes.

2.1. Method based on the statistics of domain composition

Finally, a different set of approaches for the prediction of interactions uses domain composition as the source of information about protein interactions. The first of these methods was developed by Gómez *et al*. (2001), who used the composition of domains of large proteins (the network properties defined by experimentally established protein interactions to assign a probability for each possible interaction) based on attraction/repulsion potentials calculated from known domain–domain interactions, and a global comparison of the structure of the predicted interaction network with the structure of known networks, that is, scale-free structure. A related approach (Sprinzak and Margalit, 2001) represents each protein as a list of descriptive domains (signatures) extracted from the InterPro database of domains. The statistical significance of these signatures is contrasted with a model of distribution of domains in the database.

These domain-composition methods have similar strengths and drawbacks. They are able to provide highly reliable protein interaction predictions, but only for well-known proteins and/or organisms since they depend critically on the availability of high quality and detailed information.

3. Limitations of the computational methods for the prediction of protein interactions

The interesting possibilities opened up by these computational methods should not make us forget their obvious limitations. The interactions extracted by the sequence- and genome-based methods does not directly reflect the complexity of the network of interactions in biological systems (*see* Article 111, **Functional inference from probabilistic protein interaction networks**, Volume 6). The current predictions are still not able to account directly for the distribution of the interactions in space (e.g., subcellular and tissue localization) and in time (e.g., order of the interactions). Just what comprises the relation between the predictions and the nature of the interactions from transient to stable protein interactions also remains unclear. The computational methods are unable to reproduce directly the consequences of

posttranslational modifications in the regulation of protein interactions (e.g., the many interactions specifically controlled by phosphorylation). Finally, the extent to which the computational predictions based on properties of bacterial genomes (such as composition and organization) can be extrapolated to the more complex organization, transcriptional controls, and interactions of eukaryotes is still an open question.

References

Aloy P and Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5896–5901.

Dandekar T, Snel B, Huynen M and Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, **23**, 324–328.

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Goh CS, Bogan AA, Joachimiak M, Walther D and Cohen FE (2000) Co-evolution of proteins with their interaction partners *Journal of Molecular Biology*, **299**, 283–293.

Gómez SM, Lo SH and Rzhetsky A (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics*, **159**, 1291–1298.

Hoffmann R and Valencia A (2003) Protein interaction: same network, different hubs. *Trends in Genetics*, **19**, 681–683.

Lu L, Lu H and Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Pazos F and Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**, 609–614.

Pazos F and Valencia A (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function and Genetics*, **47**, 219–227.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Sprinzak E and Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, **311**, 681–692.

Tsoka S and Ouzounis CA (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genetics*, **26**, 141–142.

Valencia A and Pazos F (2003) In *Structural Bioinformatics*, Bourne PH and Weissig H (Ed.), Wiley-Liss: Hoboken, pp. 411–426.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Short Specialist Review

Functional classification of proteins based on protein interaction data

Christine Brun, Anaïs Baudot and Bernard Jacq
CNRS Université de la Méditerranée, Marseille, France

Deciphering gene/protein function on a large scale for a better understanding of cell functioning and organism development is one of the biggest challenges in biology. For this purpose, approaches have been designed following both methodological progress and thinking advancement. In this respect, computational methods played a constant role overtime, evolving with the way biologists were apprehending gene/protein function.

Since the seventies, biologists have been comparing protein sequences using alignment methods (Needleman and Wunsch, 1970; Smith and Waterman, 1981). They progressively introduced useful measures such as the identity and similarity percentages, the z-scores, the Blast scores, and so on (*see* Article 93, **Detecting protein homology using NCBI tools**, Volume 8). Later, methods enabling the comparison of secondary and tertiary protein structures had also been devised (*see* Article 75, **Protein structure comparison**, Volume 7). The main use of such comparison methods is intended to get new insight into protein function by inferring functional relationships between proteins, thereby making possible the inference of function from a protein of known function to a protein of unknown function. The underlying hypothesis upon which such a functional inference is based is the evolutionary conservation of protein sequences and structures in order to perform a conserved function. But the relationship linking sequence, structure, and function is not always straightforward and although these approaches usually lead to useful and testable hypotheses, they remain a risky exercise supporting in some cases wrong conclusions (Devos and Valencia, 2000). For instance, first, a slight change in protein sequence may not be taken into account in the analysis while leading to important functional changes or second, two protein domains with no primary sequence similarity may wrongly be proposed to be functionally unrelated whereas they in fact share the same 3D structure and therefore the same function (Grossman and Laimins, 1996).

In addition, as we previously discussed (Jacq, 2001), the function of a gene/protein is a complex notion that can be defined at several integrated levels of complexity (molecule, cell, tissue, organism levels, etc.). Generally, sequence and structure analyses solely reveal the possible molecular function(s) of proteins when

domains of known function are identified in their sequences. Consequently, the functional knowledge granted by the previously described approaches only concern the biochemical role of proteins, without informing us about the particular cellular, physiological, or developmental process(es) in which it is exerted.

In order to both obtain a more contextual vision of gene/protein function and to be able to make functional predictions, computational methods relying upon genome organization have been developed. The domain fusion or Rosetta Stone method establishes that two proteins from a given organism are functionally related when they exist as a single fused polypeptide in another proteome (Enright *et al*., 1999; Marcotte *et al*., 1999a). Other methods have also been grounded on the facts that genes repeatedly found as neighbours on chromosomes in different organisms may encode functionally related proteins (Pellegrini *et al*., 1999) and that the phylogenetic coinheritance of proteins in several proteomes suggests their functional link (Dandekar *et al*., 1998; Overbeek *et al*., 1999; Tamames *et al*., 1997). Although these methods and their combinations (Marcotte *et al*., 1999b) were used to predict the function for a number of proteins, they still suffer from limitations essentially related to the fact that they work better when applied to completely sequenced genomes and are more appropriate to prokaryotic genome organization compared to the eukaryotic one. In addition, they are only valid for a small number of proteins and solely permit a "functional linkage" to be proposed between proteins sometimes without specifying the cellular process(es) in which the linked proteins are involved.

It then appears that providing new computational methods that enable the decoding of the cellular, the physiological, and the developmental function of gene/proteins on a large scale would not only widen the field of investigations, but more importantly would bring a necessary novel, comprehensive, and integrated understanding of gene function. Because protein action is seldom isolated but rather exerted in concert with other proteins, molecular interactions between proteins are essential actors for all biological processes in all organisms. Having access to the list of protein partners with which any given protein interacts, recapitulates the essential aspects of its cellular function, and it also provides a kind of condensed "functional identity card" for proteins. Interactions thus represent the raw material onto which new methods for protein functional descriptions could be grounded.

Recent years have seen the introduction of many different high-throughput methods, such as the DNA microarrays (*see* Article 90, **Microarrays: an overview**, Volume 4) and large-scale two-hybrid screens. Protein–protein interaction maps are now available for three eukaryotic model organisms: the budding yeast (*see* Article 39, **The yeast interactome**, Volume 5), the worm (*see* Article 38, **The *C. elegans* interactome project**, Volume 5), and the fly (Formstecher *et al*., 2005; Giot *et al*., 2003). They form large intricate networks allowing a renewed vision of the cell functioning as an integrated system. However, they need to be analyzed in detail in order to extract and reveal the functional information they contain. Various methods of biological network analysis have been proposed so far. They may, for instance, allow the identification of functional modules after network clustering (Rives and Galitski, 2003), or assigning a function to proteins of unknown function on the basis of the functional annotations of their neighbors (Vazquez *et al*., 2003).

Another way of analyzing the interaction network is to functionally compare proteins at the cellular level. As stated above, this approach would represent a useful complement to sequence comparison methods, which addresses function at the molecular level. We thus propose a new bioinformatic method named PRODISTIN (Protein distance based on interactions) (Brun *et al*., 2003) and allowing a functional classification of the proteins according to the identity of their interacting partners. The central idea in this interaction-based functional clustering is not to compare proteins themselves but instead to compare the list of their interaction partners, assuming that the more two proteins share interacting partners, the more they should be functionally related. Let us consider three proteins A, B, C, each of them establishing 30 specific interactions (experimentally determined) with other protein partners. If A and C, B and C, and A and B have respectively 25, 13, and 2 common interactors, it seems intuitively reasonable to conclude that A and C are highly functionally related, that B and C share at least some functional features and that A and B are probably not functionally (or only marginally) related. In order to translate this rather simple hypothesis into a mathematical formalism, we decided to calculate the Czekanowski–Dice distance between the proteins forming the network. This distance, which is intended to provide a direct measurement of the functional relationships between proteins (belonging to the same multiprotein complex, the same pathway or more broadly, the same cellular process(es)), corresponds to:

$$D(i, j) = \frac{\#(\mathrm{Int}(i)\Delta\mathrm{Int}(j))}{\#(\mathrm{Int}(i) \cup \mathrm{Int}(j)) + \#(\mathrm{Int}(i) \cap \mathrm{Int}(j))} \tag{1}$$

in which i and j denote two proteins, Int(i) and Int(j) are the lists of their interactors, and Δ the symmetrical difference between the two sets. A key advantage of such a distance is that it increases the weight of the shared interactors by giving more weight to the similarities than to the differences and authorizes the use of a tree representation as an output of functional similarities.

In practice, starting from a list of binary protein–protein interactions (*see* Article 44, **Protein interaction databases**, Volume 5), the PRODISTIN method consists of three different and successive bioinformatic steps: first, the functional distance is calculated between all possible pairs of proteins in the network; second, all distance values are clustered using a neighbor joining algorithm, leading to a classification tree; third, the tree is visualized and subdivided into formal classes. The PRODISTIN classes, which allow a powerful functional interpretation of the tree, are delimited according to tree topology and protein functional annotations (such as Gene Ontology (GO) terms (Ashburner *et al*., 2000)). We define them as subtrees containing at least three proteins sharing the same functional annotations and accounting for at least 50% of the class members.

Figure 1 shows a classification tree containing 602 yeast proteins, result of the application of PRODISTIN to 2946 protein–protein interactions involving 2139 proteins, that is, 38% of the *Saccharomyces cerevisiae* proteome. A detailed analysis of this PRODISTIN tree permitted an integrated analysis of yeast cellular processes and their crosstalks (Brun *et al*., 2003). Indeed, the PRODISTIN method efficiently clusters proteins involved in the same cellular process(es). On the basis of the belonging of a protein to a PRODISTIN class devoted to a particular cellular

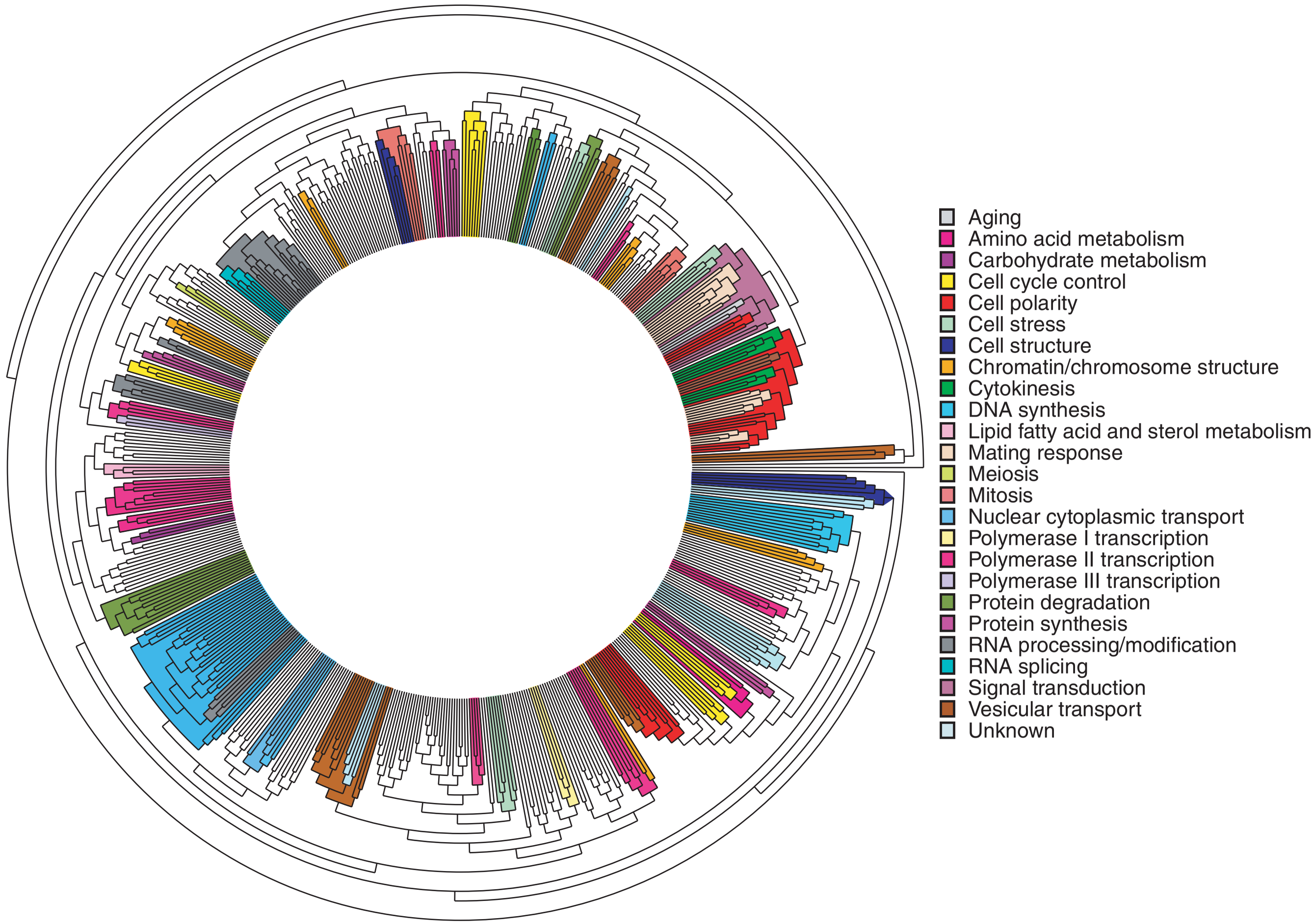


Figure 1 A functional classification tree for 602 yeast proteins computed with the PRODISTIN method. PRODISTIN classes on the circular classification tree have been colored according to their corresponding "cellular role". Protein names have been omitted for clarity

process, the classification enables to propose the involvement of the protein in this process, regardless of the current knowledge about its function. Doing so, we proposed a cellular function for 45% of the otherwise uncharacterized protein present in the tree and the involvement of proteins of known function in other functions (Brun *et al*. 2003).

As predicting the function at a cellular level differs from predicting the function at the molecular level, classifying proteins functionally according to their interaction subnetwork differs from classifying proteins in a structural or a sequence-based manner (*see* Article 82, **Structure comparison and protein structure classifications**, Volume 6, Article 91, **Classification of proteins by sequence signatures**, Volume 6, and Article 92, **Classification of proteins by clustering techniques**, Volume 6). For instance, what should we expect from both types of classification while investigating the evolutionary fate of duplicated genes in yeast? Taking into account that duplicated genes are mainly annotated as having identical molecular function according to GO (Baudot *et al*., 2004), no major differences in classification should be expected using structure- and sequence-based classifications. Conversely, the PRODISTIN functional classification appeared to represent a valuable tool when studying the evolution of the function of the yeast duplicated genes, since a new type of information emerged from its use in an evolutionary perspective (Baudot *et al*., 2004). Indeed, comparing the surrounding subnetwork for paralogous proteins allows us to differentiate several types of paralogue pairs according to their classification features. Three different behaviors of the pairs of paralogues regarding the PRODISTIN classification were identified, leading to the establishment of a scale of functional divergence for the duplicated genes based on the protein–protein network analysis, independently of sequence similarity. From the less to the more divergent in cellular function, either paralogues belong to the same functional class, or to different classes devoted to the same cellular function, or finally to different classes devoted to different functions. Comparing these results with the functional information carried by GO annotations and sequence comparison, it appeared that interaction network analysis reveals functional subtleties, which are not discernible by other means.

As a conclusion, this first use of the functional PRODISTIN classification for yeast proteins in order to address a specific question relative to gene/protein function validates the approach and more broadly, emphasizes the importance of interaction networks data and their analysis in deciphering cell functioning. Considering the cellular function of genes/proteins in the context of a molecular network, as in the aforementioned study of the evolution of the function of duplicated genes, would undoubtedly represent an important issue while approaching several other important biological problems. Just to cite a few, it is likely that questions such as the functional aspects of horizontal transfer, the integration of different signaling pathways or the functional relationship linking orthologous gene/proteins from model organisms for which interaction maps are available will largely benefit from applying the PRODISTIN method. Furthermore, as protein profiling methods will progressively allow a detailed description of the different proteomes from different cell types of the same organism, comparisons of the different protein networks encoded by a single genome will soon become possible. As far as the human proteomes are concerned, it is likely that the new vision of considering diseases

as perturbations of specific molecular networks, which can be studied by network analysis methods such as PRODISTIN will offer new perspectives in understanding both their molecular and their phenotypical aspects.

Related articles

Article 90, **Microarrays: an overview**, Volume 4; Article 38, **The *C. elegans* interactome project**, Volume 5; Article 39, **The yeast interactome**, Volume 5; Article 44, **Protein interaction databases**, Volume 5; Article 82, **Structure comparison and protein structure classifications**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 92, **Classification of proteins by clustering techniques**, Volume 6; Article 75, **Protein structure comparison**, Volume 7; Article 93, **Detecting protein homology using NCBI tools**, Volume 8

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Baudot A, Jacq B and Brun C (2004) A scale of functional divergence for yeast duplicated genes revealed from the analysis of the protein-protein interaction network. *Genome Biology*, **5**, R76.

Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A and Jacq B (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, **5**, R6.

Dandekar T, Snel B, Huynen M and Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Science*, **23**, 324–328.

Devos D and Valencia A (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Formstecher E, Aresta S, Collura V, Hamberger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C, *et al*. (2005). Protein interaction mapping: a drosophila case study. *Genome Research*, in press.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Grossman SR and Laimins LA (1996) EBNA1 and E2: a new paradigm for origin-binding proteins? *Trends in Microbiology*, **4**, 87–89.

Jacq B (2001) Protein function from the perspective of molecular interactions and genetic networks. *Briefings in Bioinformatics*, **2**, 38–50.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–2901.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Rives AW and Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1128–1133.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Tamames J, Casari G, Ouzounis C and Valencia A (1997) Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution*, **44**, 66–73.

Vazquez A, Flammini A, Maritan A and Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, **21**, 697–700.

Short Specialist Review

Analyzing proteomic, genomic and transcriptomic elemental compositions to uncover the intimate evolution of biopolymers

Peggy Baudouin-Cornu
Service de Biochimie et de Génétique Moléculaire CEA/Saclay, F-91191 Gif-sur-Yvette Cedex, France

Jason G. Bragg
University of New Mexico, Department of Biology, Albuquerque, NM, US MSC03 2020

To grow and reproduce, organisms must acquire the constituents of biopolymers from their environments. A number of the elemental constituents of biopolymers, including phosphorus, carbon, hydrogen, nitrogen, and sulfur, are subjected to biogeochemical cycles, and may vary in their abundance in both space and time. Therefore, organisms must be adapted to survive both persistent conditions of their habitats and substantial perturbations in the availability of nutrients. Several studies have now revealed that adaptive biases in the elemental composition of biopolymers may provide one mechanism for dealing with these perturbations. They also demonstrate that nutritional constraints may shape the evolution of biopolymers at a more intimate scale than amino acid (*see* Article 96, **Fundamentals of protein structure and function**, Volume 6) or base composition (*see* Article 47, **The mouse genome sequence**, Volume 3, Article 7, **Genetic signatures of natural selection**, Volume 1). Availability of more and more data on genomes (*see* Article 1, **Eukaryotic genomics**, Volume 3, Article 2, **Genome sequencing of microbial species**, Volume 3, Article 65, **Environmental shotgun sequencing**, Volume 4), proteomes (*see* Article 34, **Large-scale protein annotation**, Volume 7, Article 94, **Expression and localization of proteins in mammalian cells**, Volume 4) and transcriptomes (*see* Article 79, **Technologies for systematic analysis of eukaryotic transcriptomes**, Volume 4, Article 81, **Using ESTs for genome annotation – predicting the transcriptome**, Volume 4) now provides an unprecedented opportunity to study adaptive imprints in the atomic composition of biopolymers, and the factors driving these adaptations.

In their seminal 1989 paper, Mazel and Marlière reported such an imprint in the cyanobacterium *Calothrix* sp. PCC 701 light-harvesting phycobilisome (Mazel and Marlière, 1989). They determined that *Calothrix* expresses a phycobilisome specifically depleted in sulfur-containing amino acids when grown in a sulfur-limited medium. Since phycobilisomes can account for up to 60% of the soluble proteins of cyanobacteria, they proposed that this differential expression allowed a significant sparing in the quantity of sulfur atoms required for protein synthesis. Existence in *Calothrix* of an operon encoding a phycobilisome specifically depleted in sulfur atoms was seen as evidence that nutrient availability could have influenced biopolymer evolution (Mazel and Marlière, 1989).

More recently, a comparable mechanism was reported in the yeast *Saccharomyces cerevisiae* (Fauchon *et al*., 2002). When *S. cerevisiae* is exposed to cadmium, the synthesis of glutathione, a tripeptidic thiol critical for cadmium detoxification, is strongly induced (Vido *et al*., 2001). Fauchon *et al*. revealed that this is accompanied by a deep transcriptional and translational reorganization, leading to the transient expression of a set of proteins containing 30% fewer sulfur atoms than the set expressed in the absence of cadmium. They proposed that this allows significantly more sulfur to be dedicated to glutathione synthesis at a time when it is critical for the cell (Fauchon *et al*., 2002).

In the two above-mentioned examples, biases in atomic composition of subsets of proteins were apparently selected to achieve a significant sparing in an element, at the scale of a whole cell. Other mechanisms could select for biases in the elemental composition of subsets of proteins. For example, Baudouin-Cornu *et al*. (2001) showed that in the yeast *S. cerevisiae*, and the bacterium *Escherichia coli*, proteins used for assimilating sulfur and carbon tend to contain fewer sulfur and carbon atoms, respectively, relative to the rest of the proteomes of these organisms (Baudouin-Cornu *et al*., 2001). The reduced use of an element in a small number of moderately expressed proteins likely has little impact on the overall use of this element by a cell. However, when an element is scarce in the growth medium, amino acids containing this element may be less abundant in the cell, as shown for the two sulfur-containing amino acids in *S. cerevisiae* (Lafaye *et al*., 2005). This would result in a decrease in the corresponding aminoacyl-tRNAs, as observed in the case of amino acid starvations (Dittmar *et al*., 2005), and may reduce the rate of translation of proteins, according to their content of this element. Baudouin-Cornu *et al*. proposed that the biases they reported were selected during transient episodes of scarcity in sulfur or carbon, and allow *S. cerevisiae* and *E. coli* to maintain functional sulfur or carbon assimilatory pathways when their environment is scarce in sulfur or carbon (Baudouin-Cornu *et al*., 2001). Taken together, these studies (Mazel and Marlière, 1989; Baudouin-Cornu *et al*., 2001; Fauchon *et al*., 2002) support the conclusion that systematic biases in the atomic composition of subsets of proteins can result from adaptation to transient changes in the environment.

However, for organisms growing under the normal conditions of their habitat (i.e., in the absence of a severe, transient disruption in nutrient supply), the quantities of different elements required for synthesizing biopolymers are likely influenced by atomic composition of whole genomes, proteomes or total RNA (hereafter, “RNAome”), rather than specific subsets of proteins. Therefore, at this scale, biopolymers may be expected to reflect adaptation to persistent, rather than

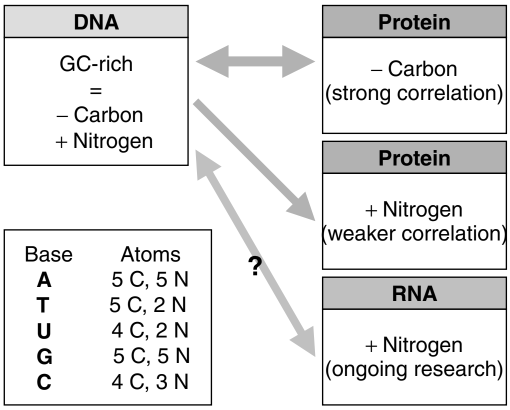


Figure 1 Schematic representation of the relationships between carbon and nitrogen contents of the three main biopolymer classes. The carbon (C) and nitrogen (N) compositions of adenine (A), thymine (T), uracil (U), guanine (G) and cytosine (C) are indicated in the lower left-hand corner. Statistics on the correlations between genomic GC contents and proteomic nitrogen and carbon contents can be found in Baudouin-Cornu *et al*., 2004 and Bragg and Hyder, 2004. See text for more details

transient, environmental features. A small number of studies have now attempted to evaluate the extent of variation among organisms in the elemental composition of biopolymers at the scale of whole genomes and proteomes, and to identify factors (e.g., traits of species) that are associated with this variation.

Analysis of the elemental composition of genomes (double stranded DNA) is straightforward, since genomic carbon and nitrogen content are each related directly to guanine and cytosine (GC) content. Each GC pair has 8 nitrogen atoms and 9 carbon atoms, while each adenine and thymine (AT) pair has 7 nitrogen atoms and 10 carbon atoms. Therefore, a GC-rich genome contains more nitrogen and less carbon per base pair than an AT-rich genome (*see* Figure 1). This observation, taken together with wide variation in GC content among bacteria (ca. 25% to >70%), prompted McEwan *et al*. to test the association between bacterial genomic GC content and the ability to fix atmospheric nitrogen (McEwan *et al*., 1998). They found that within aerobic genera, nitrogen-fixing bacteria had higher genomic GC content, and therefore higher genomic nitrogen content per base pair, than nonfixing species (McEwan *et al*., 1998). They suggested that this may represent an adaptive association between nitrogen expenditure in genomes (and potentially in RNAomes, see below), and nitrogen fixing. However, different types of biopolymers are represented in different abundances and may therefore have different consequences for nutrient use. According to Neidhardt and Umbarger's compilation, 3% of the dry weight of a bacterial cell is accounted for by DNA, 55% by proteins, and 20% by RNA (Neidhardt and Umbarger, 1996). Therefore, in terms of nutritional constraints, the elemental composition of genomes may be less important than the elemental composition of proteomes or RNAomes.

Predictions of proteomes from genome sequences (*see* Article 16, **Searching for genes and biologically related signals in DNA sequences**, Volume 7, Article 25, **Gene finding using multiple related species: a classification**

approach, Volume 7), have allowed interesting insights into proteome elemental compositions of unicellular organisms, especially prokaryotes (*see* Article 13, **Prokaryotic gene identification *in silico***, Volume 7). For example, Bragg and Hyder found that proteome nitrogen and carbon content were correlated positively and negatively (respectively) with genome GC content. This means that the carbon and nitrogen contents of genomes are correlated positively with the carbon and nitrogen contents of proteomes, respectively (Bragg and Hyder, 2004). These correlations are due to the structure of the genetic code (Baudouin-Cornu *et al*., 2004; Bragg and Hyder, 2004). Additionally, predicted whole proteomes allow more subtle studies of proteomic elemental composition than just the comparison of means. Using the quantile representation proposed by Karlin *et al*. to compare amino acid compositions (Karlin *et al*., 1992), Baudouin-Cornu *et al*. found the same negative correlation between genome GC contents and proteome carbon contents, but also showed that the quantile distributions of proteome carbon contents were stochastically ordered, whereas quantile distributions of proteome nitrogen contents were not (Baudouin-Cornu *et al*., 2004). This suggests that mean values of proteomic carbon content likely provide a reliable indication of relative proteomic carbon use among species, whereas for nitrogen, mean proteomic values are less likely to provide a reliable indication of relative proteomic nitrogen use. That is, the expression levels of proteins with different nitrogen content (within proteomes) may have a relatively larger role in determining average protein nitrogen use among organisms (e.g., in comparison to carbon). This highlights the potential usefulness of considering protein expression levels explicitly in studies of proteomic elemental composition, as expression data become increasingly available.

Although both studies of proteomic carbon and nitrogen composition led to interesting observations (Baudouin-Cornu *et al*., 2004; Bragg and Hyder, 2004), they did not identify links between the elemental composition of proteomes and the environments in which different organisms live. A recent study of proteome sulfur contents has revealed such relationships (Bragg *et al*., in press). In particular, we observed a tendency for species living at high temperature to have lower proteomic sulfur use. To our knowledge, this is the first example of a simple relationship between an environmental feature and the quantity of a specific element used in proteins, at the level of whole proteomes. However, as genome sequences and lifestyle data become available for a growing number of microorganisms, we anticipate that more relationships will be revealed between environmental factors and the atomic composition of proteomes or RNAomes.

Variation in the nitrogen and carbon content of RNAomes among organisms has not (to our knowledge) been considered explicitly, despite the observation that RNAs may account for 20% of cellular dry mass (Neidhardt and Umbarger, 1996). The bases of RNA, A, C, G, and U (uracil), contain 5, 3, 5, and 2 nitrogen atoms, respectively, while A and G each contain 5 carbon atoms, and C and U each contain 4 carbon atoms. The nitrogen and carbon content of RNAomes are not related exactly to GC content (as in genomes), because (1) RNAs in cells are typically single-stranded, and (2) different RNA molecules may vary greatly in their abundance. However, if the parity G = C and A = U holds approximately (Chargaff's second parity rule, see Forsdyke and Mortimer (2000) for a review), and GC-rich genomes encode GC-rich RNAomes, it is likely that organisms with high

genomic nitrogen content (high GC content), would have nitrogen rich RNAomes (McEwan *et al*., 1998). This suggests that there is a correlation among organisms from relatively low nitrogen content in the three main biopolymers (and low GC content), to relatively high nitrogen content (and high GC content) in these biopolymers (Figure 1). This prediction hints that comparing RNAome elemental composition among species would reveal interesting features and deserves to be undertaken. In particular, ribosomal RNAs (rRNAs) constitute up to 80% of the total RNAs in a cell, and may therefore have an inordinate influence in the elemental composition of total RNA. An analysis of the elemental composition of rRNA sequences among different organisms may thus provide useful insights into variation in nitrogen expenditure in RNA.

Future studies of the atomic composition of biopolymers hold much promise for our understanding of both biopolymer evolution, and the nutrient requirements of organisms. Such studies should continue to examine variation in atomic composition both among subsets of biopolymers within organisms, and among organisms. For example, they might include analyses aimed at testing whether multicellular organisms have biases in the atomic composition of biopolymers that are expressed during specific stages of development, or in specific types of cells. Among organisms, it may be enlightening to study the atomic composition of biopolymers by testing hypotheses concerning traits that influence the access of different organisms to specific nutrients.

Acknowledgments

JGB was supported by an NSF Biocomplexity fellowship.

References

Baudouin-Cornu P, Schuerer K, Marlière P and Thomas D (2004) Intimate evolution of proteins. Proteome atomic content correlates with genome base composition. *The Journal of Biological Chemistry*, **279**, 5421–5428.

Baudouin-Cornu P, Surdin-Kerjan Y, Marlière P and Thomas D (2001) Molecular evolution of protein atomic composition. *Science*, **293**, 297–300.

Bragg JG and Hyder CL (2004) Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proceedings. Biological sciences/The Royal Society*, **271**(Suppl 5), S374–S377.

Bragg JG, Thomas D and Baudouin-Cornu P Variation among species in proteomic sulphur content is related to environmental conditions. *Proceedings. Biological sciences/The Royal Society*, **http://www.journals.royalsoc.ac.uk/openurl.asp?genre=article&id=doi:10.1098/rspb.2005.3441** (in press).

Dittmar KA, Sørensen MA, Elf J, Ehrenberg M and Pan T (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Reports*, **6**, 151–157.

Fauchon M, Lagniel G, Aude JC, Lombardia L, Soularue P, Petat C, Marguerie G, Sentenac A, Werner M and Labarre J (2002) Sulfur sparing in the yeast proteome in response to sulfur demand. *Molecular Cell*, **9**, 713–723.

Forsdyke DR and Mortimer JR (2000) Chargaff's legacy. *Gene*, **261**, 127–137.

Karlin S, Blaisdell BE and Bucher P (1992) Quantile distributions of amino acid usage in protein classes. *Protein Engineering*, **5**, 729–738.

Lafaye A, Junot C, Pereira Y, Lagniel G, Tabet JC, Ezan E and Labarre J (2005) Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *The Journal of Biological Chemistry*, **280**, 24723–24730.

Mazel D and Marlière P (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature*, **341**, 245–248.

McEwan CE, Gatherer D and McEwan NR (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*, **128**, 173–178.

Neidhardt FC and Umbarger HE (1996) Chemical composition of *Escherichia coli*. In *Escherichia coli and Salmonella typhymurium: Cellular and Molecular Biology*, Neidhardt FC, Curtis RI, Ingraham JL, Lin ECC, Low, KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M and Umbarger, HE (Eds), American Society for Microbiology Press: Washington, DC, pp. 13–28.

Vido K, Spector D, Lagniel G, Lopez S, Toledano MB and Labarre J (2001) A proteome analysis of the cadmium response in Saccharomyces cerevisiae. *The Journal of Biological Chemistry*, **276**, 8469–8474.

Short Specialist Review

Topology of protein interaction networks and cell physiology

Marie-Claude Marsolier-Kergoat
Service de Biochimie et de Génétique Moléculaire, CEA/Saclay, Gif-sur-Yvette, France

Cell physiology involves physical interactions between many types of molecules. The protein interaction network of a cell can be defined as the set of the nonoriented links connecting proteins (the network vertices) that have been experimentally found to interact. The accumulation of small-scale studies and the development of high-throughput technologies have recently led to a first description of these networks in several organisms like the bacteria *Helicobacter pylori* (Rain *et al*., 2001) and *Escherichia coli* (Butland *et al*., 2005), the yeast *Saccharomyces cerevisiae* (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002) (*see* Article 39, **The yeast interactome**, Volume 5), and the metazoans *Drosophila melanogaster* (Giot *et al*., 2003; Formstecher *et al*., 2005) and *Caenorhabditis elegans* (Li *et al*., 2004) (*see* Article 38, **The *C. elegans* interactome project**, Volume 5). Numerous analyses have reported the existence of correlations between topological characteristics of proteins in the interaction networks and their evolutionary conservation or other functional features (Jeong *et al*., 2001; Fraser *et al*., 2002, 2003; Maslov and Sneppen 2002; Barabasi and Oltvai, 2004). However, these correlations have remained largely controversial. Thus, the potential correlation between the number of links (or degree) of a protein and its evolutionary rate, initially reported by Fraser and collaborators (Fraser *et al*., 2002, 2003), has been subsequently questioned by several reanalyses (Bloom and Adami, 2003; Jordan *et al*., 2003; Hahn *et al*., 2004), one of them suggesting that this correlation is a mere artifact resulting from the effect of protein abundance (Bloom and Adami, 2003). In this article, I will address two related questions: Locally, are the topological characteristics of a protein correlated with a basic, functional feature like its essentiality? Globally, can the overall structure of protein interaction networks be interpreted in terms of cell physiology? Another article in this section deals with the prediction of gene function and biochemical networks from protein interactions (*see* Article 37, **Inferring gene function and biochemical networks from protein interactions**, Volume 5) and complements this approach.

1. Correlations between the essentiality of a protein and its topological characteristics in interaction networks

A protein is called essential when the deletion of the corresponding gene is lethal even under optimal growth conditions. Protein essentiality has so far only been

systematically assessed in *S. cerevisiae* (Giaever *et al*., 2002) and in *C. elegans* (Kamath *et al*., 2003), and the discussion about the correlations between protein essentiality and protein network topology has been mostly restricted to *S. cerevisiae* due to the wealth of available data for this model organism. It was quickly pointed out that protein interaction networks, like many other biological, technical or social networks, exhibit high variability in their vertex degree and consist of a majority of low-connected vertices and a minority of highly connected ones (Jeong *et al*., 2001). The degree distribution of the protein interaction networks was originally claimed to follow a power law (Jeong *et al*., 2001). However, subsequent studies have shown that this conclusion was not justified (Przulj *et al*., 2004; Tanaka *et al*., 2005; see also Keller, 2005), so that it appears more appropriate to simply describe the protein network degree distribution as large. Networks with a large degree distribution are characterized by the fact that their structure (as assessed by the size of their largest cluster of vertices or by their diameter, the average length of the shortest paths between any two vertices) is little affected by random vertex removal but highly sensitive to the removal of their most connected vertices (Albert *et al*., 2000). Equating gene deletion to vertex removal and lethality to network failure allowed to propose that the minority of highly connected vertices in a protein interaction network should correspond to essential proteins. This suggestion was initially made by Barabasi and collaborators (Jeong *et al*., 2001), and since then others have striven to demonstrate a correlation between the essentiality of a protein and its degree in interaction networks (Wuchty, 2004; Yu *et al*., 2004, for a review see Barabasi and Oltvai, 2004). However, we have shown that this correlation is mainly due to biases in the protein interaction databases (*see* Article 44, **Protein interaction databases**, Volume 5) used by these authors (Coulomb *et al*., 2005). Among other things we have demonstrated that, for *S. cerevisiae*, essential proteins are the objects of almost twice as many research articles than nonessential ones and we have found the predictable positive correlation between the number of articles referring to a protein and the number of its interactants indicated in databases like Database of Interacting Proteins (DIP, Xenarios *et al*., 2000) that record the results of small-scale studies (Coulomb *et al*., 2005 and unpublished data). Analysis of an a priori unbiased data set of yeast protein interactions yielded at best a marginal correlation between protein degree and essentiality (Coulomb *et al*., 2005). Similar results (unpublished data) have also been found using data from systematic studies on *C. elegans* (Kamath *et al*., 2003; Li *et al*., 2004). Other correlations have also been reported between the essentiality of a protein and the average degree of its neighbors or its clustering coefficient (the ratio between the number of links connecting the k neighbors of a given vertex and $k(k-1)/2$, the number of all possible links between these k neighbors) (Maslov and Sneppen, 2002; Yu *et al*., 2004). However, we failed to observe these correlations using different, a priori unbiased databases (Coulomb *et al*., 2005).

2. Overall structure of protein interaction networks and cell physiology

The large degree distribution of protein interaction networks has been suggested to account for cell robustness against mutation (Jeong *et al*., 2001; Barabasi

and Oltvai, 2004). The hypothesis that only the removal of highly connected proteins is lethal would have elegantly explained the fact that only 18.7% of *S. cerevisiae* proteins are essential (Giaever *et al*., 2002). However, this hypothesis is inconsistent with the observation that the essentiality of a protein is poorly, if ever, related to its degree in interaction networks.

Negative degree correlation, that is a propensity for high-degree vertices to attach to low-degree vertices, was another global feature reported for protein interaction networks and proposed to have a physiological signification (Maslov and Sneppen, 2002; Newman 2002; 2003). The systematic suppression of links between highly connected vertices was suggested to decrease the probability of cross-talk between different functional pathways and to increase mutational robustness (Maslov and Sneppen, 2002). However, we and others found that this degree correlation was assessed using severely flawed data sets derived from two-hybrid experiments (Aloy and Russell 2002; Coulomb *et al*., 2005), and using a priori unbiased databases failed to reveal any positive or negative degree correlation in protein networks (Coulomb *et al*., 2005).

3. Conclusion

General connections between the topology of protein interaction networks and cell physiology are thus difficult to point out. The local, topological characteristics of proteins are at most weakly related to their essentiality and the global features of protein interaction networks, like their large distribution of degrees or their lack of degree correlation, do not lend themselves to obvious physiological interpretation. The phenomena that determine the structure of protein networks thus remain debated. It has been shown that many structural characteristics of experimentally observed coexpression and interaction networks could be generated with standard evolution processes involving gene duplication and addition and deletion of links (Wagner, 2003; Amoutzias *et al*., 2004; van Noort *et al*., 2004), which suggests that the structure of these networks could simply be determined by construction processes. Alternatively, Shakhnovich and coworkers have studied a physical model of protein interaction strictly based on desolvation (nonspecific interactions) and have shown that the networks constituted by these connections devoid of biological signification exhibit many of the structural characteristics of the experimentally determined protein interaction networks (Deeds *et al*., 2006). So, besides their physiological interpretation, even the very structure of protein interaction networks remains a puzzling question.

References

Albert R, Jeong H and Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature*, **406**(6794), 378–382.

Aloy P and Russell RB (2002) Potential artefacts in protein-interaction networks. *FEBS Letters*, **530**(1-3), 253–254.

Amoutzias GD, Robertson DL, Oliver SG and Bornberg-Bauer E (2004) Convergent networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, **5**(3), 274–279.

Barabasi AL and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101–113.

Bloom JD and Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evolutionary Biology*, **3**, 21.

Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, *et al*. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, **433**(7025), 531–537.

Coulomb S, Bauer M, Bernard D and Marsolier-Kergoat MC (2005) Gene essentiality and the topology of protein interaction networks. *Proceedings. Biological Sciences*, **272**(1573), 1721–1725.

Deeds EJ, Ashenberg O and Shakhnovich EI (2006) A simple physical model for scaling in protein-protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(2), 311–316.

Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C, *et al*. (2005) Protein interaction mapping: a Drosophila case study. *Genome Research*, **15**(3), 376–384.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science*, **296**(5568), 750–752.

Fraser HB, Wall DP and Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evolutionary Biology*, **3**, 11.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**(6896), 387–391.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**(5651), 1727–1736.

Hahn MW, Conant GC and Wagner A (2004) Molecular evolution in large genetic networks: does connectivity equal constraint? *Journal of Molecular Evolution*, **58**(2), 203–211.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**(6868), 180–183.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

Jordan IK, Wolf YI and Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evolutionary Biology*, **3**, 1.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, *et al*. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**(6920), 231–237.

Keller EF (2005) Revisiting "scale-free" networks. *BioEssays*, **27**(10), 1060–1068.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**(5657), 540–543.

Maslov S and Sneppen K (2002) Specificity and stability in topology of protein networks. *Science*, **296**(5569), 910–913.

Newman ME (2002) Assortative mixing in networks. *Physical Review Letters*, **89**(20), 208701.

Newman ME (2003) Mixing patterns in networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, **67**(2 Pt 2), 026126.

Przulj N, Corneil DG and Jurisica I (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**(18), 3508–3515.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**(6817), 211–215.

Tanaka R, Yi TM and Doyle J (2005) Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, **579**(23), 5140–5144.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**(6770), 623–627.

van Noort V, Snel B and Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, **5**(3), 280–284.

Wagner A (2003) How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**(1514), 457–466.

Wuchty S (2004) Evolution and topology in the yeast protein interaction network. *Genome Research*, **14**(7), 1310–1314.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM and Eisenberg D (2000) DIP: the database of interacting proteins. *Nucleic Acids Research*, **28**(1), 289–291.

Yu H, Greenbaum D, Xin Lu H, Zhu X and Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends in Genetics*, **20**(6), 227–231.

Bioluminescence resonance energy transfer

Ralf Jockers and Stefano Marullo
Institut Cochin, Paris, France

1. Introduction

The elucidation of complex networks of interaction among proteins is one of the most important and challenging tasks of postgenomic biology. Among the available approaches to study protein–protein interaction in living cells, Resonance Energy Transfer (RET)-based techniques have become increasingly popular over the past few years. RET consists of nonirradiative energy transfer (ET) between a donor and an acceptor. Because the efficacy of ET varies inversely with the 6th power of the distance between the donor and acceptor molecules, ET from the donor results in the emission of light by the acceptor only if the two molecules are in close proximity (10–100 Å). Therefore, the detection of ET between two proteins fused, respectively, to an energy donor and an acceptor, often reflects the existence of molecular interaction between the proteins of interest. In contrast, the lack of RET does not exclude the possibility of interaction between two proteins, for example, when the conformations of the two interacting partners keep the acceptor too distant from the donor. A second important parameter, which may hamper the amount of ET from donor to acceptor, is the relative orientation of the donor to acceptor molecules.

Among existing RET approaches, Fluorescence Resonance Energy Transfer (FRET) is the method of choice when the principal aim of the study is to identify the subcellular compartment where the interaction between two proteins occurs. In FRET experiments, the cyan variant (CFP) of the green fluorescent GFP and its yellow variant (YFP) are widely used as energy donor and acceptor, respectively. Other combinations of fluorescent proteins are possible, such as GFP together with fluorescent proteins emitting in the red part of the spectrum. Various approaches to measure FRET exist and have been compared in recent reviews (Jares-Erijman and Jovin, 2003). They all require the external excitation of the donor molecule.

In 1999, a novel ET technique, the Bioluminescence Resonance Energy Transfer (BRET), has been developed (Xu *et al*., 1999), on the basis of a slightly different principle inspired by a natural phenomenon observed in glowing marine organisms. In the presence of its substrate, coelenterazine, *Renilla* luciferase (Rluc, the luminescent energy donor) transfers some energy to a GFP variant (the energy

acceptor). No excitation of the donor is required in this case and the substrate, which is membrane permeable, can be added to the supernatant of cultured cells. Various types of coelenterazines have been developed, which are characterized by different emission wavelengths and intensities as well as by the duration of the light emission. Two main acceptor fluorescent proteins are used in BRET experiments, the YFP and a blue-shifted fluorescent protein (GFP2). Each acceptor works with a different coelenterazine emitting at specific wavelengths.

BRET protocols have been designed to monitor and quantify both constitutive and regulated molecular interactions in intact cells. A practical guideline on how to perform BRET experiments has been published recently (Issad and Jockers, 2005). Here we will discuss the advantages and limits of BRET-based techniques and illustrate possible applications by describing some specific assays.

2. Strengths and weaknesses of BRET-based assays

BRET is a technique of choice for many applications in cell biology and molecular pharmacology to study constitutive and regulated interactions as well as the dynamics of intra- and intermolecular conformational changes. The experimental conditions for BRET assays are very flexible since the experiments can be performed with purified proteins, subcellular fractions, and permeabilized or intact cells. The opportunity of using intact cells limits the number of false-positive interactions, which can be observed in assays where the integrity of subcellular compartments is lost. Moreover, compared to the yeast two-hybrid system, BRET can be performed in mammalian cells in which the native subcellular localization and posttranslational modifications of the interacting partners are preserved. BRET values are reliable since the measured signals correspond to average values reflecting variable concentrations of interacting partners in large cell populations. As detailed below, BRET saturation experiments may provide quantitative data on the propensity of two interacting partners to interact with each other. Finally, another important feature of BRET assays is the possibility to quantify the amount of expressed energy donor and energy acceptor independently in each individual experiment. This control is important to differentiate specific BRET signals from artificial interactions caused by overexpression of the proteins of interest used in the assay.

Cell autofluorescence and photobleaching, which may cause problems in FRET assays, are minimized in BRET experiments, since the excitation of the energy donor is not obtained through external irradiation. On average, the BRET assays have been estimated to be approximately 10 times more sensitive than the FRET assays in a microplate format (Arai, Ueda *et al*., 2000; Arai *et al*., 2001).

Although BRET measurements require instrumentation with specific features, several BRET readers are now available from different suppliers. BRET readers must be capable of measuring emitted light quasi-simultaneously at two different wavelengths. In addition, these readers are generally driven by a software that calculates BRET ratios in real time. The experimental setup is easy and fully compatible with measurements in 96- or 384-well microplates, permitting to test multiple different experimental conditions very rapidly. Therefore, BRET assays

are equally well suited for academic research and companies, which need to develop high-throughput tests, for drug screening (Boute *et al*., 2002).

Despite all the advantages listed above, some limitations of the BRET technique exist nevertheless. Wild-type proteins cannot be used in BRET assays, as the fusion to a donor or to an acceptor is an inherent feature of the approach. Consequently, the biological properties of fusion proteins should be tested carefully and compared to the wild-type counterpart to be sure that function and/or subcellular localization is not markedly modified. Fusion proteins need to be expressed in heterologous models, which may be difficult to obtain in some cases, although modern transfection reagents and viral vectors allow heterologous protein expression in almost any cell type. Some chemicals may inhibit luciferase activity, whereas others may quench the BRET signal, owing to strong absorption between 400 and 600 nm. Pilot experiments should be performed to detect potential interferences in the assay buffer. Despite several successful attempts of BRET measurements on subcellular fractions obtained by cell fractionation, a suitable microscope-associated detection system has not been developed so far to directly visualize the subcellular localization of BRET in living cells. The limiting factor for achieving this goal is the low intensity of the light generated by the *Renilla* luciferase. Therefore, the development of more sensitive light detectors is necessary to overcome this limitation. Thus, FRET remains, so far, the method of choice when the principal aim of the study is to identify the subcellular compartment in which the interaction between two proteins occurs.

3. Applications of BRET

Protein oligomerization and the regulated movement of proteins between two compartments in living cells have been extensively investigated using different approaches including BRET; they will be discussed in more detail to illustrate the different facets of BRET assays.

3.1. Monitoring protein oligomerization by BRET

The biotechnological use of BRET was first described by Xu *et al*. (1999) who studied the dimerization of bacterial transcription factors of the Kai B family involved in the regulation of the circadian rhythm. Subsequently, BRET has been widely used to study the oligomerization of membrane proteins, in particular, G protein–coupled receptors (GPCR) (reviewed in Kroeger and Eidne, 2004). The BRET approach provided a significant input in the demonstration that constitutive receptor homo-oligomers of GPCRs exist in living cells at physiological receptor expression levels (Issafras *et al*., 2002). Recently, more sophisticated quantitative BRET assays have been developed in order to quantify the relative abundance of receptor oligomers versus monomers, to determine the relative affinity between different protomers engaged in the formation of homo- and heterodimers and to interpret agonist-induced changes in BRET signals (Ayoub *et al*., 2002; Mercier, *et al*., 2002; Couturier and Jockers, 2003). Although these quantitative assays

have been mostly developed to study GPCR oligomerization, the method may be extrapolated to any type of protein that can potentially undergo oligomerization.

The most popular quantitative assay developed so far is the BRET donor saturation assay (Mercier *et al*., 2002; Couturier and Jockers, 2003). Several independent transfections are performed with a constant amount of BRET donor and increasing quantities of BRET acceptor. As the amount of acceptor increases, the BRET signal rises as a hyperbolic function and reaches an asymptote ($BRET_{max}$), which corresponds to the saturation of all available donor molecules by acceptor molecules. Assuming that the association of the fusion proteins occurs at the equilibrium, the amount of acceptor required to achieve half-maximal BRET ($BRET_{50}$), for a given amount of donor, reflects the relative affinity of the two partners. The relative affinities of several interaction partners have been determined using this method, which allowed to compare the propensity of receptor isoforms to be engaged in homodimers versus heterodimers (Mercier *et al*., 2002; Terrillon *et al*., 2003; Ayoub *et al*., 2004; Breit *et al*., 2004; Ramsay *et al*., 2004).

Quantitative assays may be used in some cases to estimate the proportion of the dimeric versus monomeric fraction of a protein (Ayoub *et al*., 2002; Mercier *et al*., 2002; Couturier and Jockers, 2003). If a free equilibrium governs the association of the monomers, whether they are fused to the BRET donor or to the acceptor, one can predict that in the case of a 1:1 molecular ratio of the two BRET partners, 50% of formed dimers (those containing one donor and one acceptor) would produce BRET, whereas dimers that contain only BRET donors or acceptors, would each represent 25% of total dimers. Accordingly, if the proteins of interest only exist as dimers under these conditions, the expected BRET value should correspond approximately to the $BRET_{50}$, which is 50% of the maximal BRET value ($BRET_{max}$), the value reached upon complete saturation of the BRET donor by the acceptor. If the observed BRET value falls significantly below 50% of the $BRET_{max}$, the hypothesis above is not confirmed and a significant proportion of free monomers is likely present. This type of experiment requires that a precise quantitation of BRET partners is available, for example, using binding assays, which correlate fluorescence and luminescence values with real concentrations of BRET partners.

BRET donor saturation assays also proved to be useful for the interpretation of the BRET signal changes induced by agonists (Couturier and Jockers, 2003; Lacasa *et al*., 2005; Percherancier *et al*., 2005). As mentioned above, BRET efficacy varies with both the distance between the donor and the acceptor and the relative orientation of the acceptor to the donor. Consequently, a change of BRET signal may be indicative of dissociation or association between BRET partners, but may also reflect conformational changes of the molecules, to which the donor or the acceptor is fused. In order to discriminate between these two possibilities, $BRET_{50}$ values are compared in cells incubated in the absence or presence of agonists. Conformational changes within BRET partners are revealed by identical $BRET_{50}$ and different $BRET_{max}$ values. In contrast, a significant variation of $BRET_{50}$ values, which reflects a change in the relative affinity of the partners, indicates ligand-induced association or dissociation of BRET partners.

Conformational changes within preformed receptor dimers have been documented for several receptors. EC_{50} values (the concentration of ligand, which

induces the half-maximal change of the BRET signal) have been determined for multiple ligands and for given ratios of BRET donor and acceptor pairs. Cumulated data obtained with the insulin receptor (a receptor with tyrosine kinase activity, Boute *et al*., 2001), the leptin receptor (belonging to the cytokine receptor family, Couturier and Jockers, 2003), and several GPCRs (Angers, *et al*., 2000; Ayoub *et al*., 2004; Percherancier *et al*., 2005) indicate that EC_{50} values are correlated with the affinity of the ligands (K_i). Owing to the unique feature of the BRET experiments, which is to focus the analysis on one given combination of receptors (those respectively fused to a BRET donor and a BRET acceptor), BRET can be used to determine the specific pharmacological profile of receptor heterodimers (Ayoub *et al*., 2004; Percherancier *et al*., 2005). Such information is difficult to obtain using classical radioligand competition binding assays, as ligands can potentially bind simultaneously to all available receptor species: monomers, homodimers, and heterodimers.

3.2. Monitoring protein translocation by BRET

BRET assays were successfully used to monitor regulated protein translocation in living cells. The first example of this approach is the recruitment of the GPCR-associated regulatory and signaling protein ß-arrestin to agonist-stimulated ß2-adrenergic receptors (Angers *et al*., 2000). Subsequent experiments with other GPCRs were reported to produce similar results. As expected, ß-arrestin recruitment to receptor was strictly agonist dependent and the observed EC_{50} values correlated with agonist affinities for corresponding receptors (Hanyaloglu *et al*., 2002; Terrillon *et al*., 2003). As the interaction with ß-arrestin is a general feature of most receptors of the GPCR superfamily, most of the orphan GPCRs (more than 100 remain to be characterized) should be capable of recruiting ß-arrestin in case of activation. Accordingly, a BRET-based screening assay for agonists of orphan GPCRs has been proposed (Bertrand *et al*., 2002) on the basis of receptor-ß-arrestin interaction.

During translocation to activated GPCR, ß-arrestins are ubiquitinated by the Mdm2 ubiquitin ligase (Shenoy *et al*., 2001). This phenomenon is believed to be critical for receptor endocytosis. The stability of ß-arrestin ubiquitination depends on the receptor that has initiated ß-arrestin translocation. Sustained ubiquitination of ß-arrestin, in turn, stabilizes the ß-arrestin-receptor association and affect the fate of internalized receptors (Shenoy and Lefkowitz, 2003). Receptors stably associated with ß-arrestins are principally sorted to lysosomes (degradative pathways), whereas receptors that interact weakly with ß-arrestin are mostly recycled to the plasma membrane. Recently, a BRET-based assay has been developed to study the dynamic of ß-arrestin ubiquitination (Perroy *et al*., 2004). The BRET donor (Rluc) was fused to the N-terminus of ß-arrestin and the BRET acceptor (GFP2) was fused to the N-terminus of Lys48Ala Lys63Ala ubiquitin, a ubiquitin mutant that cannot form polyubiquitin chains. The ubiquitination of Rluc-ß-arrestin by GFP2-ubiquitin results in a BRET signal, which is dynamically regulated by GPCR activation. On the basis of the stability of the BRET signal, the two classes of GPCR described above may be distinguished.

BRET assays have also been employed to monitor the interaction of an enzyme with its substrate, as shown for the tyrosine phosphatase PTP1B, which dephosphorylates tyrosine phosphorylated insulin receptors (Boute *et al*., 2003). Agonist stimulation of the insulin receptor induces receptor autophosphorylation on tyrosine residues and its subsequent internalization. The receptor is then dephosphorylated by PTP1-B. To stabilize the transient interaction between the receptor and the enzyme, a substrate-trapping mutant of PTP1B was used (D181A PTP1B). A BRET signal could be detected, upon insulin stimulation, in intact cells expressing an insulin receptor fused upstream of the luciferase, and the YFP-PTP1B D181A mutant.

In conclusion, there is no theoretical limit to the types of constitutive or regulated protein–protein interactions that can be studied by BRET, provided that the necessity of fusing BRET donors and acceptors to the proteins of interest does not perturb their biological properties. False-positive results, due to protein overexpression, can generally be avoided, because of the high sensitivity of the method, by simply scaling down the amount of transfected cDNAs until the amount of expressed fusion proteins becomes comparable to physiological concentrations. As discussed above, molecular interactions may not be detectable by BRET because the distance or the orientation of BRET donor and acceptor are not appropriate or when the interaction between the proteins of interest is too transient. In many cases, different positioning of BRET donors or acceptors or the creation of appropriate mutants can help overcome these limitations. We have no doubt that the BRET technology is just at its beginning and has a promising future.

References

Angers S, Salahpour A, Joly E, Hilairet S, Chelsky D, Dennis M and Bouvier M (2000) Detection of ß2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfert (BRET). *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3684–3689.

Arai R, Nakagawa H, Tsumoto K, Mahoney W, Kumagai I, Ueda H and Nagamune T (2001) Demonstration of a homogeneous noncompetitive immunoassay based on bioluminescence resonance energy transfer. *Analytical Biochemistry*, **289**, 77–81.

Arai R, Ueda H, Tsumoto K, Mahoney WC, Kumagai I and Nagamune T (2000) Fluorolabeling of antibody variable domains with green fluorescent protein variants: application to an energy transfer-based homogeneous immunoassay. *Protein Engineering*, **13**, 369–376.

Ayoub MA, Couturier C, Lucas-Meunier E, Angers S, Fossier P, Bouvier M and Jockers R (2002) Monitoring of ligand-independent dimerization and ligand-induced conformational changes of melatonin receptors in living cells by bioluminescence resonance energy transfer. *The Journal of Biological Chemistry*, **277**, 21522–21528.

Ayoub MA, Levoye A, Delagrange P and Jockers R (2004) Preferential formation of MT1/MT2 melatonin receptor heterodimers with distinct ligand interaction properties compared with MT2 homodimers. *Molecular Pharmacology*, **66**, 312–321.

Bertrand L, Parent S, Caron M, Legault M, Joly E, Angers S, Bouvier M, Brown M, Houle B and Menard L (2002) The BRET2/arrestin assay in stable recombinant cells: a platform to screen for compounds that interact with G protein-coupled receptors (GPCRS). *Journal of Receptor and Signal Transduction Research*, **22**, 533–541.

Boute N, Boubekeur S, Lacasa D and Issad T (2003) Dynamics of the interaction between the insulin receptor and protein tyrosine-phosphatase 1B in living cells. *EMBO Reports*, **4**, 313–319.

Boute N, Jockers R and Issad T (2002) The use of resonance energy transfer in high-throughput screening: BRET versus FRET. *Trends in Pharmacological Sciences*, **23**, 351–354.

Boute N, Pernet K and Issad T (2001) Monitoring the activation state of the insulin receptor using bioluminescence resonance energy transfer. *Molecular Pharmacology*, **60**, 640–645.

Breit A, Lagace M and Bouvier M (2004) Hetero-oligomerization between ß2- and ß3-adrenergic receptors generates a beta-adrenergic signaling unit with distinct functional properties. *The Journal of Biological Chemistry*, **279**, 28756–28765.

Couturier C and Jockers R (2003) Activation of leptin receptor by a ligand-induced conformational change of constitutive receptor dimers. *The Journal of Biological Chemistry*, **278**, 26604–26611.

Hanyaloglu AC, Seeber RM, Kohout TA, Lefkowitz RJ and Eidne KA (2002) Homo- and hetero-oligomerization of thyrotropin-releasing hormone (TRH) receptor subtypes. Differential regulation of beta-arrestins 1 and 2. *The Journal of Biological Chemistry*, **277**, 50422–50430.

Issad T and Jockers R (2005) *Bioluminescence Resonance Energy Transfer (BRET) To Monitor Protein-protein Interactions*, Humana Press: Totowa.

Issafras H, Angers S, Bulenger S, Blanpain C, Parmentier M, Labbe-Jullie C, Bouvier M and Marullo S (2002) Constitutive agonist-independent CCR5 oligomerization and antibody-mediated clustering occurring at physiological levels of receptors. *The Journal of Biological Chemistry*, **277**, 34666–34673.

Jares-Erijman EA and Jovin TM (2003) FRET imaging. *Nature Biotechnology*, **21**, 1387–1395.

Kroeger KM and Eidne KA (2004) Study of G-protein-coupled receptor-protein interactions by bioluminescence resonance energy transfer. *Methods in Molecular Biology*, **259**, 323–333.

Lacasa D, Boute N and Issad T (2005) Interaction of the insulin receptor with the receptor-like protein tyrosine-phosphatases PTPα and PTPϵ in living cells. *Molecular Pharmacology*, **67**, 1206–1213.

Mercier JF, Salahpour A, Angers S, Breit A and Bouvier M (2002) Quantitative assessment of ß1 and ß2-adrenergic receptor homo and hetero-dimerization by bioluminescence resonance energy transfer. *The Journal of Biological Chemistry*, **277**, 44925–44931.

Percherancier Y, Berchiche Y, Slight I, Volkmer-Engert R, Tamamura H, Fujii N, Bouvier M and Heveker N (2005) Bioluminescence resonance energy transfer reveals ligand-induced conformational changes in CXCR4 homo- and heterodimers. *The Journal of Biological Chemistry*, **280**, 9895–9903.

Perroy J, Pontier S, Charest PG, Aubry M and Bouvier M (2004) Real-time monitoring of ubiquitination in living cells by BRET. *Nature Methods*, **1**, 203–208.

Ramsay D, Carr IC, Pediani J, Lopez-Gimenez JF, Thurlow R, Fidock M and Milligan G (2004) High-affinity interactions between human alpha1A-adrenoceptor C-terminal splice variants produce homo- and heterodimers but do not generate the alpha1L-adrenoceptor. *Molecular Pharmacology*, **66**, 228–239.

Shenoy SK and Lefkowitz RJ (2003) Trafficking patterns of beta-arrestin and G protein-coupled receptors determined by the kinetics of beta-arrestin deubiquitination. *The Journal of Biological Chemistry*, **278**, 14498–14506.

Shenoy SK, McDonald PH, Kohout TA and Lefkowitz RJ (2001) Regulation of receptor fate by ubiquitination of activated beta 2-adrenergic receptor and beta-arrestin. *Science*, **294**, 1307–1313.

Terrillon S, Durroux T, Mouillac B, Breit A, Ayoub MA, Taulan M, Jockers R, Barberis C and Bouvier M (2003) Oxytocin and vasopressin V1a and V2 receptors form constitutive homo- and heterodimers during biosynthesis. *Molecular Endocrinology*, **17**, 677–691.

Xu Y, Piston DW and Johnson CH (1999) A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 151–156.

Introductory Review

Probing cellular function with bioluminescence imaging

Guy A. Rutter, Gabriela da Silva Xavier, Takashi Tsuboi and Kathryn J. Mitchell

University of Bristol, Bristol, UK

John W. Hanrahan

McGill University, Montreal QC, Canada

1. Introduction

In this chapter, we describe luminescence-based methodologies that have been developed to monitor three key aspects of cellular biology in the context of single living mammalian cells or cell populations. Each of these methods is highly amenable to studying the function of proteins introduced into cells through: (1) co-overexpression of the wild-type protein with the reporters described, using a mammalian expression vector, adenoviral gene delivery (Andreolas *et al*., 2002), or transgenesis (Ma *et al*., 2004); (2) inactivation of the protein through RNA interference (da Silva Xavier *et al*., 2004; *see also* Article 60, **siRNA approaches in cell biology**, Volume 5); or (3) introduction of inactivating antibodies into single cells (da Silva Xavier, 2000b, 2004; Andreolas *et al*., 2002). As such, these are powerful techniques for the analysis of gene function.

2. Use of recombinant targeted luciferases for dynamic measurements of intracellular free ATP concentration

The maintenance of intracellular ATP, the cell's energy currency, is a fundamental concern for all living organisms. The use of firefly luciferase, targeted to discrete intracellular domains, provides an extremely sensitive method of monitoring free [ATP] dynamics at the subcellular level. Firefly (*Photinus pyralis*) luciferase uses ATP, luciferin, and oxygen as substrates to produce light with a quantum yield of 0.9:

$$\text{Luciferin} + \text{ATP} + \text{Mg}^{2+} + \text{O}_2 \rightarrow \text{oxyluciferin} + \text{AMP} + \text{PP}_i + \text{CO}_2 + \text{light} \quad (1)$$

Luciferase has previously been employed to measure intracellular ATP concentrations in single cardiac myocytes (Bowers *et al*., 1992) and hepatocytes (Koop and Cobbold, 1993), but only after microinjection of the purified protein. By expressing the recombinant protein from plasmids introduced by microinjection (Kennedy *et al*., 1999), transfection (Tsuboi *et al*., 2003), or the use of adenoviral vectors (Ainscow and Rutter, 2001), we have developed a single cell photon-counting method to image ATP concentrations dynamically and at the subcellular level in single cells (Kennedy *et al*., 1999). Importantly, the addition of targeting sequences to luciferase cDNA allows [ATP] to be imaged in a variety of subcellular domains, including the cytosol, mitochondria, and the subplasma membrane region (Kennedy *et al*., 1999). Given the feeble photon production of luciferase (a few photons per second per cell versus thousands of photons for fluorescent probes), imaging [ATP] at the single cell level requires highly sensitive cameras. There are, broadly speaking, two common types of photon-counting systems. The first involves the coupling of a charged multichannel plate coupled to a charge-coupled device (CCD). Here, a multichannel plate intensifier amplifies the signal up to 10^9-fold, thus giving a high signal-to-noise ratio. Various systems exist on the market (e.g., see websites for Photek, **www.photek.co.uk**, Hamamatsu-Photonics, **www.hamamatsu.com**, and Roper Scientific, **www.roperscientific.com**). Our imaging system consists of an Olympus IX-70 inverted microscope (×10 air objective, 0.4 numerical aperture) and a three-stage intensified CCD camera (Photek ICCD316; Photek, Lewes, E. Sussex, UK). We have used this system, in combination with recombinant targeted luciferases, to monitor changes in localized free [ATP] in a number of cell types (Ainscow *et al*., 2000; Ainscow and Rutter, 2001; Jouaville *et al*., 1999; Kennedy *et al*., 1999; Tsuboi *et al*., 2003) (Figure 1).

However, this type of system gives low spatial resolution. Back-thinned liquid crystal display (LCD) detectors have a lower signal-to-noise ratio as the signal is not amplified prior to detection. Signal-to-noise ratio can be improved through the use of better cooling systems, whereas cooled back-thinned integrating CCD cameras are preferred for imaging tissues (see below). However, at the time of writing, such systems only provide adequate sensitivity to detect changes in single cells over rather lengthy time frames (>10 s.data point^{-1}).

3. Using luciferase bioluminescence to measure release of intracellular ATP

Extracellular ATP is an important autocrine and paracrine signaling molecule. When released from damaged cells, it binds to purinergic (P2X) receptors on sensory neurons that signal pain. Thus, the activation of isolated nociceptors by cell extracts is abolished if the extracts are pretreated with apyrase to degrade ATP, or if P2X receptors on the neuron are blocked using the receptor antagonist, suramin (Cook and McCleskey, 2002). In addition to this pathological role, physiological ATP release mediates neurotransmission, vascular tone, cellular volume, responses to distension and infection, and many other physiological processes. There is thus great interest in measuring its extracellular concentration. Although systemic

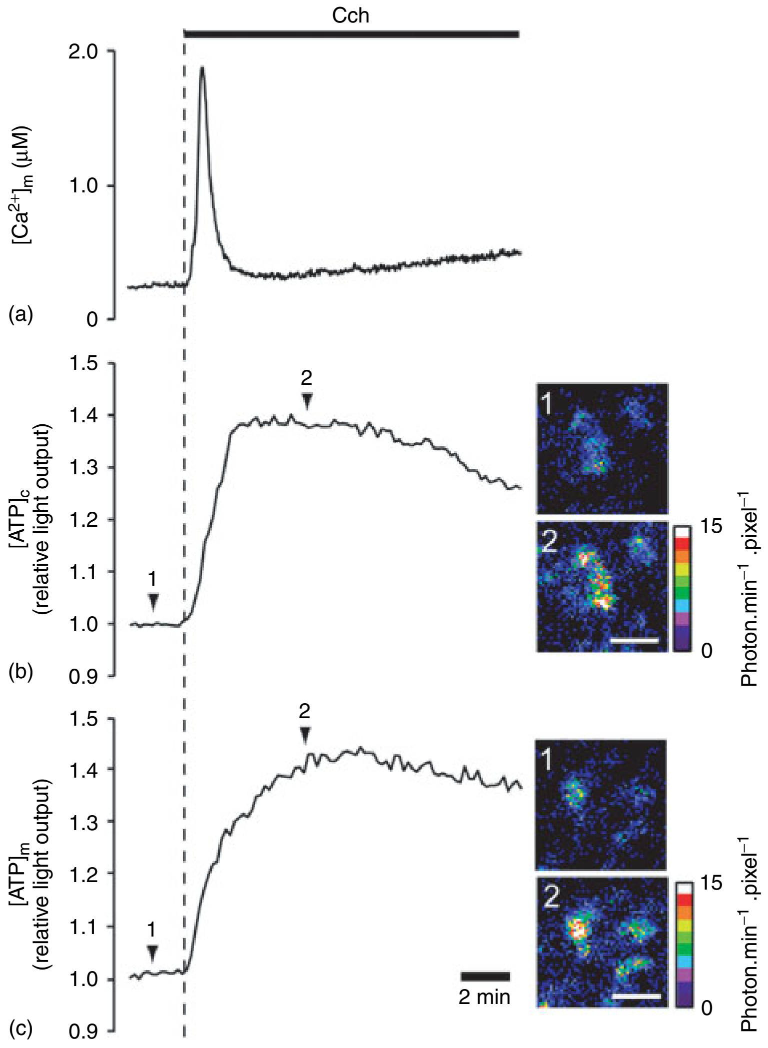


Figure 1 Effect of carbachol (Cch, 100 μM) on (a) mitochondrial $[Ca^{2+}]$, (b) cytosolic [ATP], and (c) mitochondrial [ATP] increases in pancreatic MIN6 β-cells. Typical traces from eight separate experiments are shown. The images in (b) and (c) show the monitored luminescence changes in a typical field of cells, imaged during 60 s, starting at the corresponding time points (1 and 2). The scale bar represents 50 μm

plasma [ATP] ranges are thought to range between 0.4 and 1.4 μM, interstitial ATP levels within tissues depend on local rates of release and degradation and can be much higher. For example, [ATP] is estimated to reach 68 μM at the surface of retinal astrocytes during mechanical stimulation (Newman, 2001), 25 μM at the surface of pancreatic beta cells (Hazama *et al*., 1998), and 14 μM beneath the macula densa during renal tubulo-glomerular feedback (Bell *et al*., 2003).

Luciferase bioluminescence has historically been the most widely used method for studying ATP release.

The most straightforward approach involves collecting samples of extracellular fluid and analyzing them in a luminometer. This approach is quantitative and readily yields an apparent ATP efflux rate. Problems with experimental interventions that could interfere with the luciferase reaction are also minimized. For example, gadolinium (Boudreault and Grygorczyk, 2002), which is often used to block stretch-activated cation channels, as well as suramin (Newman, 2001) are both luciferase inhibitors that can be removed before assay. However, collecting and analyzing ATP samples can be problematic as release by many cells is exquisitely mechano-sensitive. Simply tilting a culture dish gently during collection of samples from cultured airway epithelial cells can increase extracellular [ATP] by 10-fold (Grygorczyk and Hanrahan, 1997). This sensitivity is probably somewhat exaggerated in cultured cells since they are not subjected to constant mechanical stimulation as they would be in vivo. Nevertheless, mechanically induced ATP release is observed with freshly isolated tissues (e.g., Ferguson *et al*., 1997), and care must be taken to avoid physical manipulations and solution turbulence. An approach we have found useful involves culturing cells on porous capillaries in a hollow-fiber bioreactor, so that ATP release can be continuously collected in a flow-through system that provides control over shear stress (Guyot and Hanrahan, 2002).

An alternative to the sample-and-assay approach involves adding luciferase and luciferin to the bath and monitoring bioluminescence in situ. If all ATP were consumed by luciferase, release could be estimated simply by counting the photons generated. However, it is usually more practical to use conditions in which luminescence is proportional to the steady state [ATP] and to assume that a negligible fraction of released ATP is consumed by the luciferase reaction. A staircaselike signal is then obtained that could, in principle, be differentiated to give information about the rate of release (Taylor *et al*., 1998). One interesting method is that of Jans *et al*. (2002), who used a stopped-flow approach to determine ATP efflux rates.

A major advantage of extracellular luciferase for monitoring ATP release in situ comes from the spatial information it provides when combined with low light-level microscopy (e.g., Wang *et al*., 2000). Indeed, spontaneous bursts of ATP release are observed in some conditions (Arcuino *et al*., 2002), which may be mediated by connexin hemichannels (Cotrina *et al*., 1998). On the other hand, vesicular ATP release was observed when intracellular calcium was elevated in astrocytes by the glutamergic agonist AMPA (α-amino-3-methyl-4-isoxazolepropionic acid) and during stimulation by the protein kinase C activator PMA (phorbol 12-myristate 13-acetate) (Coco *et al*., 2003).

ATP at the cell surface is most relevant for extracellular signaling since it bathes the purinoceptors. A novel application of the extracellular luciferin-luciferase reaction has been used to monitor ATP release by isolated pancreatic acini (Sørensen and Novak, 2001). Rather than measuring the bioluminescence signal, the decrease in fluorescence between 510 and 550 nm caused by oxidation of luciferin was used to monitor luciferase activity and, thus, [ATP].

4. Imaging free intracellular Ca^{2+} concentration

Changes in intracellular free Ca^{2+} concentration play a crucial role in the control of cellular function (Rutter, 1998a; Berridge *et al*., 2003). Since its identification and characterization (Shimomura *et al*., 1962), the Ca^{2+}-sensitive bioluminescent protein aequorin, from the jellyfish *Aequorea victoria*, has provided an invaluable tool for the dynamic measurement of intracellular Ca^{2+} ion concentration. Aequorin exists in its native form in a complex with its coenzyme, coelenterazine, and reports changes in Ca^{2+} concentration via an oxidation reaction that produces blue light (Head *et al*., 2000). Microinjection of purified aequorin has been extensively used to monitor Ca^{2+} signaling in a wide spectrum of cell types, including bacteria, arthropods, plants, and mammalian cells. A key technological advance occurred following the cloning of aequorin (Inouye *et al*., 1985), which allowed the generation of recombinant aequorin constructs. Furthermore, the generation of aequorin mutants with decreased Ca^{2+} sensitivity (Kendall, *et al*., 1992) and the ready availability of synthetic coelenterazine analogs have resulted in the generation of several organelle-targeted aequorins able to monitor lumenal Ca^{2+} concentration changes with high specificity. Applied for the first time to monitor free Ca^{2+} concentration in the mitochondria of endothelial (Rizzuto *et al*., 1992) and pancreatic β- (Rutter *et al*., 1993) cells, recombinant aequorin was subsequently used to measure Ca^{2+} concentrations in the sarco- and endoplasmic reticulum (Montero *et al*., 1995), nucleus (Brini *et al*., 1993), Golgi apparatus (Pinton *et al*., 1998), dense core secretory vesicle surface (Pouli *et al*., 1998) and interior (Mitchell, *et al*., 2001), as well as the bacterial periplasm (Jones *et al*., 2002).

The use of aequorin does have disadvantages compared to the use of fluorescent Ca^{2+} probes (Miyawaki *et al*., 1997). Thus, the very limited amount of emitted light enforces severe limitations on single cell or suborganellar imaging (Rutter *et al*., 1996). Moreover, when measuring [Ca^{2+}] in compartments with relatively high steady state levels ($>10\,\mu M$), reconstitution of aequorin with coenzyme requires previous complete depletion of Ca^{2+}, a procedure that may have deleterious effects on the cell. Irreversible destruction of the aequorin complex, combined with the high rate of consumption at steady state, also restrict experiment duration. On the positive side, aequorin can be used to measure Ca^{2+} concentration over a wide range (100 nM to $>500\,\mu M$), and calibration of the emitted light output is made easy by the use of a simple algorithm (Rutter *et al*., 1993). The major advantage of using aequorin, however, is that it is much less severely inhibited at low pH values (<6.0) (Blinks, 1989; Rizzuto *et al*., 1994) than green fluorescent protein (GFP)-based Ca^{2+}-sensitive indicators, for example, cameleons (Miyawaki *et al*., 1997; Llopis *et al*., 1998; Baird *et al*., 1999; Emmanouilidou *et al*., 1999), which detect changes in local Ca^{2+} concentrations using fluorescence resonance energy transfer (FRET) between a cyan-emitting mutant of GFP and an enhanced yellow-emitting GFP, linked by the calmodulin-binding peptide M13. This fundamental property of aequorin becomes crucial when measuring Ca^{2+} concentrations within acidic organelles, such as dense core secretory vesicles whose pH are approximately 5.5

(Urbe *et al*., 1997). Indeed, using a chimaeric-aequorin targeted to the secretory vesicle matrix (Mitchell *et al*., 2001), dense core vesicles have recently been revealed as an important intracellular Ca^{2+}-store involved in ryanodine receptor and nicotinic acid adenine dinucleotide phosphate (NAADP)-mediated intracellular Ca^{2+} signaling and the regulation of insulin secretion (Mitchell *et al*., 2003).

5. Measuring gene expression in single cells and whole organisms using luminescent probes

Bioluminescent reporter genes have also proven to be invaluable in the study of regulated gene expression. Common genetic reporters include bacterial luciferases (bacterial *lux* gene products), firefly (*Photinus*) luciferase, β-galactosidase, β-glucuronidase, alkaline phosphatase, green fluorescent protein, β-lactamase, and human growth hormone (Rutter *et al*., 1998b).

The use of firefly luciferase as a reporter has increased in popularity because of the development of a “constant light” luciferase assay, which allows rapid, quantitative dynamic imaging of gene expression. Commonly, firefly luciferase is coupled with the use of β-galactosidase or β-glucuronidase to allow the expression of the regulated gene to be normalized to that of one which is constitutively expressed. Since changes in free ATP concentration can potentially interfere with measurements with luciferase (see above), it is usually necessary to assess if the experiment results in a net change in ATP content; usually, these changes are small ($<15\%$) and can frequently be neglected in the face of manyfold changes in the activity of a promoter.

Our own dual reporter system involves the promoter of interest fused upstream of the firefly luciferase gene and the cytomegalovirus promoter fused upstream of the *Rotylenchulus reniformis* gene. The reporter gene-bearing plasmids are introduced into cells by transient or stable transfection, or via microinjection. The signal output from firefly and *R. reniformis* luciferases are assessed using a sensitive photon-counting camera (see above) upon addition of luciferin and coelenterazine, respectively (Kennedy *et al*., 1997; da Silva Xavier, *et al*., 2000a; da Silva Xavier *et al*., 2000b; Rafiq *et al*., 1998; Campbell *et al*., 1999; Wu *et al*., 1999; Rafiq *et al*., 2000; Palmer *et al*., 2002; Andreolas *et al*., 2002; Stirland *et al*., 2003).

The use of bioluminescent reporters offers some key advantages over that of fluorescent reporters based on GFP (Chalfie *et al*., 1994; Heim and Tsien, 1996). In the case of GFP, the lack of linearity between protein concentration and fluorescence intensity, as well as interference from cellular autofluorescence (Rutter *et al*., 1998b) significantly limits sensitivity. Moreover, the maturation time of GFP is typically between 2 and 4 h (Siemering *et al*., 1996) such that it cannot easily be used for the monitoring of oscillatory changes in gene expression. This problem is even more acute for the use of β-lactamase and probes based on FRET (Zlokarnik *et al*., 1998), where the accumulation of a cleaved product is involved. At present, measurements of such dynamic changes can only readily be achieved at the single cell level using luciferase, as demonstrated in both mammalian cells (Stirland *et al*., 2003) and plants (Millar *et al*., 1995).

Luciferase-based technologies can also be applied to whole organisms. Imaging in whole plants (Riggs and Chrispeels, 1987) and in rodents (DiLella *et al*., 1988) was first described in the late 1980s. The technology available for the transfer into and imaging of reporter genes in whole organisms has improved considerably in recent years (Contag and Bachman, 2002; Greer and Szalay, 2002). For this application, cooled CCDs are more suitable than image-intensified CCDs, as the former have higher quantum efficiency (approximately 85%) (Rice *et al*., 2001). Luminescent reporters can now be introduced into whole organisms through the use of replication-deficient adenoviruses, while transgenic organisms expressing reporter genes have also been produced to study specific processes within the whole organism. A recent elegant study by Yamaguchi *et al*. (2001) showed that real time in vivo gene expression could be performed over a prolonged period of time to monitor the oscillatory expression of luciferase in living mouse brain.

6. Conclusions

Bioluminescent proteins are now able to provide a means to measure multiple aspects of cellular behavior. The engineering of probes already known, and the discovery of new variants, should allow us to measure an even larger range of cell functions in the future.

Acknowledgments

Supported by grants to G.A.R from the Wellcome Trust, Medical Research Council, Juvenile Diabetes Research Fund International, and Human Frontiers Science Program and to J.W.H. from the Canadian Institutes of Health Research, Canadian Cystic Fibrosis Foundation, and US National Institutes of Health.

Further reading

Kennedy ED, Rizzuto R, Theler JM, Pralong WF, Bastianutto C, Pozzan T and Wollheim CB (1996) Glucose-stimulated insulin secretion correlates with changes in mitochondrial and cytosolic Ca^{2+} in aequorin-expressing INS-1 cells. *Journal of Clinical Investigation*, **98**, 2524–2538.

Maechler P, Wang H and Wollheim CB (1998) Continuous monitoring of ATP levels in living insulin secreting cells expressing cytosolic firefly luciferase. *FEBS Letters*, **422**, 328–332.

McCormack JG, Halestrap AP and Denton RM (1990) Role of calcium ions in regulation of mammalian intramitochondrial metabolism. *Physiological Reviews*, **70**, 391–425.

References

Ainscow EK and Rutter GA (2001) Mitochondrial priming modifies Ca2+ oscillations and insulin secretion in pancreatic islets. *Biochemical Journal*, **353**, 175–180.

Ainscow EK, Zhao C and Rutter GA (2000) Acute overexpression of lactate dehydrogenase-A perturbs beta-cell mitochondrial metabolism and insulin secretion. *Diabetes*, **49**, 1149–1155.

Andreolas C, da Silva Xavier G, Diraison F, Zhou C, Varadi A, Lopez-Casillas F, Ferré P, Foufelle F and Rutter GA (2002) Stimulation of acetyl-CoA carboxylase (ACC1) transcription by glucose requires insulin release and sterol regulatory element binding protein (SREBP1 c) function in pancreatic MIN6 β-cells. *Diabetes*, **51**, 2536–2545.

Arcuino G, Lin JHC, Takano T, Liu C, Jiang L, Gao Q, Kang J and Nedergaard M (2002) Intercellular calcium signaling mediated by point-source burst release of ATP. *Proceedings of the National Academy of Sciences, USA*, **99**, 9840–9845.

Baird GS, Zacharias DA and Tsien RY (1999) Circular permutation and receptor insertion within green fluorescent proteins. *Proceedings of the National Academy of Sciences, USA*, **96**, 11241–11246.

Bell PD, Lapointe J-Y, Sabirov R, Hayashi S, Peti-Peterdi J, Manabe K-I, Kovacs G and Okada Y (2003) Macula densa cell signaling involves ATP release through a maxi anion channel. *Proceedings of the National Academy of Sciences, USA*, **100**, 4322–4432.

Berridge MJ, Bootman MD and Roderick HL (2003) Calcium signalling: dynamics, homeostasis and remodelling. *Nature Reviews in Molecular and Cellular Biology*, **4**, 517–529.

Blinks JR (1989) Use of calcium-regulated photoproteins as intracellular Ca^{2+} indicators. *Methods in Enzymology*, **172**, 164–203.

Boudreault F and Grygorczyk R (2002) Cell swelling-induced ATP release and gadolinium-sensitive channels. *American Journal of Physiology - Cell Physiology*, **282**, C219–C226.

Bowers KC, Allshire AP and Cobbold PH (1992) Bioluminescent measurement in single cardiomyocytes of sudden cytosolic ATP depletion coincident with rigor. *Journal of Molecular and Cellular Cardiology*, **24**, 213–218.

Brini M, Murgia M, Pasti L, Picard D, Pozzan T and Rizzuto R (1993) Nuclear Ca^{2+} concentration measured with specifically targeted recombinant aequorin. *EMBO Journal*, **12**, 4813–4819.

Campbell SC, Cragg H, Elrick LJ, Macfarlane WM, Shennan KIJ and Docherty K (1999) Inhibitory effect of PAX4 on the human insulin and islet amyloid polypeptide (IAPP) promoters. *FEBS Letters*, **46**, 53–57.

Chalfie M, Tu Y, Euskirchen G, Ward WW and Prasher DC (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802–805.

Coco S, Calegari F, Pravettoni E, Pozzi D, Taverna E, Rosa P, Matteoli M and Verderio C (2003) Storage and release of ATP from astrocytes in culture. *Journal of Biological Chemistry*, **278**, 1354–1362.

Contag CH and Bachmann MH (2002) Advances in in vivo bioluminescence imaging of gene expression. *Annual Reviews in Biomedical Engineering*, **4**, 235–260.

Cook SP and McCleskey EW (2002) Cell damage excites nociceptors through release of cytosolic ATP. *Pain*, **95**, 41–47.

Cotrina ML, Lin JHC, Alves-Rodrigues A, Liu S, Li J, Azmi-Ghadimi H, Kang J, Naus CCG and Nedergaard M (1998) Connexins regulate calcium signaling by controlling ATP release. *Proceedings of the National Academy of Sciences, USA*, **95**, 15735–15740.

da Silva Xavier G, Leclerc I, Salt I, Doiron B, Hardie DG, Kahn A and Rutter GA (2000b) Role of AMP-activated protein kinase in the regulation by glucose of islet β-cell gene expression. *Proceedings of the National Academy of Sciences, USA*, **97**, 4023–4028.

da Silva Xavier G, Rutter J and Rutter GA (2004) Involvement of PAS kinase in the stimulation of preproinsulin and pancreatic duodenum homeobox-1 gene expression by glucose. *Proceedings of the National Academy of Sciences, USA*, **101**(22), 8319–8324.

da Silva Xavier G, Varadi A, Ainscow EK and Rutter GA (2000a) Regulation of gene expression by glucose in pancreatic β-cells (MIN6) via insulin secretion and activation of phosphatidyl inositol 3′ kinase. *Journal of Biological Chemistry*, **275**, 36269–36277.

DiLella AG, Hope DA, Chen H, Trumbauer M, Schwartz RJ and Smith RG (1988) Utility of firefly luciferase as a reporter gene for promoter activity in transgenic mice. *Nucleic Acids Research*, **16**, 4159.

Emmanouilidou E, Teschemacher AG, Pouli AE, Nicholls LI, Seward EP and Rutter GA (1999) Imaging Ca^{2+} concentration changes at the secretory vesicle surface with a recombinant targeted cameleon. *Current Biology*, **9**, 915–918.

Ferguson DR, Kennedy I and Burton TJ (1997) ATP is released from rabbit urinary bladder epithelial cells by hydrostatic pressure changes - a possible sensory mechanism. *Journal of Physiology*, **505**, 503–511.

Greer LF 3rd and Szalay AA (2002) Imaging of light emission from the expression of luciferases in living cells and organisms: a review. *Luminescence*, **17**, 43–74.

Grygorczyk R and Hanrahan JW (1997) CFTR-independent ATP release from epithelial cells triggered by mechanical stimuli. *American Journal of Physiology - Cell Physiology*, **272**, C1058–C1066.

Guyot A and Hanrahan JW (2002) ATP release from human airway epithelial cells studied using a capillary cell culture system. *Journal of Physiology*, **545**, 199–206.

Hazama A, Hayashi S and Okada Y (1998) Cell surface measurements of ATP release from single pancreatic b cells using a novel biosensor technique. *Pflügers Archives*, **437**, 31–35.

Head JF, Inouye S, Teranishi K and Shimomura O (2000) The crystal structure of the photoprotein aequorin at 2.3 A resolution. *Nature*, **405**, 372–376.

Heim R and Tsien RY (1996) Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Current Biology*, **6**, 178–182.

Inouye S, Noguchi M, Sakaki Y, Takagi Y, Miyata T, Iwanaga S, Miyata T and Tsuji FI (1985) Cloning and sequence analysis of cDNA for the luminescent protein aequorin. *Proceedings of the National Academy of Sciences, USA*, **82**, 3154–3158.

Jans D, Srinivas SP, Waelkens E, Segal A, Larivière E, Simaels J and Van Driessche W (2002) Hypotonic treatment evokes biphasic ATP release across the basolateral membrane of cultured renal epithelia (A6). *Journal of Physiology*, **545**, 543–555.

Jones HE, Holland IB and Campbell AK (2002) Direct measurement of free Ca^{2+} shows different regulation of Ca^{2+} between the periplasm and the cytosol of Escherichia coli. *Cell Calcium*, **32**, 183–192.

Jouaville LS, Pinton P, Bastianutto C, Rutter GA and Rizzuto R (1999) Regulation of mitochondrial ATP synthesis by calcium: evidence for a long-term metabolic priming. *Proceedings of the National Academy of Sciences, USA*, **96**, 13807–13812.

Kendall JM, Sala-Newby G, Ghalaut V, Dormer RL and Campbell AK (1992) Engineering the Ca^{2+}-activated photoprotein aequorin with reduced affinity for calcium. *Biochemical and Biophysical Research Communications*, **187**, 1091–1097.

Kennedy HJ, Pouli AE, Ainscow EK, Jouaville LS, Rizzuto R and Rutter GA (1999) Glucose generates sub-plasma membrane ATP microdomains in single islet beta-cells. Potential role for strategically located mitochondria. *Journal of Biological Chemistry*, **274**, 13281–13291.

Kennedy HJ, Viollet B, Rafiq I, Kahn A and Rutter GA (1997) Upstream stimulatory factor-2 (USF2) activity is required for glucose stimulation of L-pyruvate kinase promoter activity in single living islet β-cells. *Journal of Biological Chemistry*, **272**, 20636–20640.

Koop A and Cobbold PH (1993) Continuous bioluminescent monitoring of cytoplasmic ATP in single isolated rat hepatocytes during metabolic poisoning. *Biochemical Journal*, **295**, 165–170.

Llopis J, McCaffery JM, Miyawaki A, Farquhar MG and Tsien RY (1998) Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. *Proceedings of the National Academy of Sciences, USA*, **95**, 6803–6808.

Ma D, Shield JPH, Dean D, Leclerc I, Rutter GA and Kelsey G (2004) Defective glucose homeostasis in transgenic mice over-expressing the human transient neonatal diabetes mellitus (TNDM) locus. *Journal of Clinical Investigation*, **114**, 339–348.

Millar AJ, Carre IA, Strayer CA, Chua NH and Kay SA (1995) Circadian clock mutants in Arabidopsis identified by luciferase imaging. *Science*, **267**, 1161–1163.

Mitchell KJ, Lai FA and Rutter GA (2003) Ryanodine receptor type I and nicotinic acid adenine dinucleotide phosphate (NAADP) receptors mediate Ca^{2+} release from insulin-containing vesicles in living pancreatic beta-cells (MIN6). *Journal of Biological Chemistry*, **278**, 11057–11064.

Mitchell KJ, Pinton P, Varadi A, Tacchetti C, Ainscow EK, Pozzan T, Rizzuto R and Rutter GA (2001) Dense core secretory vesicles revealed as a dynamic $Ca(^{2+})$ store in neuroendocrine cells with a vesicle-associated membrane protein aequorin chimaera. *Journal of Cell Biology*, **155**, 41–51.

Miyawaki A, Llopis J, Heim R, McCaffery JM, Adams JA, Ikura M and Tsien RY (1997) Fluorescent indicators for Ca^{2+} based on green fluorescent proteins and calmodulin. *Nature*, **388**, 882–887.

Montero M, Brini M, Marsault R, Alvarez J, Sitia R, Pozzan T and Rizzuto R (1995) Monitoring dynamic changes in free Ca^{2+} concentration in the endoplasmic reticulum of intact cells. *EMBO Journal*, **14**, 5467–5475.

Newman EA (2001) Propagation of intercellular calcium waves in retinal astrocytes and Müller cells. *Journal of Neuroscience*, **21**, 2215–2223.

Palmer G, Rutter GA and Tavaré J (2002) Insulin stimulated fatty acid synthase gene expression does not require increased sterol-response-element-binding protein-1 transcription in primary adipocytes. *Biochemical and Biophysical Research Communications*, **291**, 439–443.

Pinton P, Pozzan T and Rizzuto R (1998) The Golgi apparatus is an inositol 1,4,5-trisphosphate-sensitive Ca^{2+} store, with functional properties distinct from those of the endoplasmic reticulum. *EMBO Journal*, **17**, 5298–5308.

Pouli AE, Karagenc N, Wasmeier C, Hutton JC, Bright N, Arden S, Schofield JG and Rutter GA (1998) A phogrin-aequorin chimaera to image free Ca^{2+} in the vicinity of secretory granules. *Biochemical Journal*, **330**, 1399–1404.

Rafiq I, Da Silva Xavier G, Hooper S and Rutter GA (2000) Glucose-stimulated preproinsulin gene expression and nuclear translocation of pancreatic duodenum homeobox-1 (PDX-1), require activation of phosphatidyl inositol 3-kinase but not p38 MAPK/SAPK2. *Journal of Biological Chemistry*, **275**, 15977–15984.

Rafiq I, Kennedy H and Rutter GA (1998) Glucose-dependent translocation of insulin promoter factor-1 (IPF-1) between the nuclear periphery and the nucleoplasm of single MIN6 β-cells. *Journal of Biological Chemistry*, **273**, 23241–23247.

Rice BW, Cable MD and Nelson MB (2001) In vivo imaging of light-emitting probes. *Journal of Biomedical Optics*, **6**, 432–440.

Riggs CD and Chrispeels MJ (1987) Luciferase reporter gene cassettes for plant gene expression studies. *Nucleic Acids Research*, **15**, 8115.

Rizzuto R, Brini M and Pozzan T (1994) Targeting recombinant aequorin to specific intracellular organelles. *Methods in Cell Biology*, **40**, 339–358.

Rizzuto R, Simpson AW, Brini M and Pozzan T (1992) Rapid changes of mitochondrial Ca^{2+} revealed by specifically targeted recombinant aequorin. *Nature*, **358**, 325–327.

Rutter GA, Burnett P, Rizzuto R, Brini M, Pozzan T, Tavaré JM and Denton RM (1996) Digital imaging of intramitochondrial Ca^{2+} with recombinant targeted aequorin: significance for the regulation of pyruvate dehydrogenase activity. *Proceedings of the National Academy of Sciences, USA*, **93**, 5489–5494.

Rutter GA, Fasolato C and Rizzuto R (1998a) Calcium and organelles: a two-sided story. *Biochemical and Biophysical Research Communications*, **253**, 549–557.

Rutter GA, Kennedy HJ, Wood CD, White MR and Tavaré JM (1998b) Real-time imaging of gene expression in single living cells. *Chemistry and Biology*, **5**, R285–R290.

Rutter GA, Theler JM, Murgia M, Wollheim CB, Pozzan T and Rizzuto R (1993) Stimulated Ca^{2+} influx raises mitochondrial free Ca^{2+} to supramicromolar levels in a pancreatic beta-cell line. Possible role in glucose and agonist-induced insulin secretion. *Journal of Biological Chemistry*, **268**, 22385–22390.

Shimomura O, Johnson FH and Saiga Y (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *Journal of Cell and Comparative Physiology*, **59**, 223–239.

Siemering KR, Golbik R, Sever R and Haseloff J (1996) Mutations that suppress the thermosensitivity of green fluorescent protein. *Current Biology*, **6**, 1653–1663.

Sørensen CE and Novak I (2001) Visualization of ATP release in pancreatic acini in response to cholinergic stimulus. *Journal of Biological Chemistry*, **276**, 32925–32932.

Stirland JA, Seymour ZC, Windeatt S, Norris AJ, Stanley P, Castro MG, Loudon AS, White MR and Davis JR (2003) Real-time imaging of gene promoter activity using an adenoviral reporter construct demonstrates transcriptional dynamics in normal anterior pituitary cells. *Journal of Endocrinology*, **178**, 61–69.

Taylor AL, Kudlow BA, Marrs KL, Gruenert DC, Guggino WB and Schwiebert EM (1998) Bioluminescence detection of ATP release mechanisms in epithelia. *American Journal of Physiology - Cell Physiology*, **275**, C1391–C1406.

Tsuboi T, da Silva Xavier G, Holz GG, Jouaville LS, Thomas AP and Rutter GA (2003) Glucagon-like peptide-1 mobilizes intracellular Ca^{2+} and stimulates mitochondrial ATP synthesis in pancreatic MIN6 beta-cells. *Biochemical Journal*, **369**, 287–299.

Urbe S, Tooze SA and Barr FA (1997) Formation of secretory vesicles in the biosynthetic pathway. *Biochimica Biophysica Acta*, **1358**, 6–22.

Wang Z, Haydon PG and Yeung ES (2000) Direct observation of calcium-independent intercellular ATP signaling in astrocytes. *Analytical Chemistry*, **72**, 2001–2007.

Wu H, MacFarlane WM, Tadayyon M, Arch JR, James RF and Docherty K (1999) Insulin stimulates pancreatic-duodenal homoeobox factor-1 (PDX1) DNA-binding activity and insulin promoter activity in pancreatic beta cells. *Biochemical Journal*, **344**, 813–818.

Yamaguchi S, Kobayashi M, Mitsui S, Ishida Y, van der Horst GT, Suzuki M, Shibata S and Okamura H (2001) View of a mouse clock gene ticking. *Nature*, **409**, 684.

Zlokarnik G, Negulescu PA, Knapp TE, Mere L, Burres N, Feng L, Whitney M, Roemer K and Tsien RY (1998) Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science*, **279**, 84–88.

Small molecule fluorescent probes for protein labeling and their application to cell-based imaging (including FlAsH etc.)

Tony Cass

Institute of Biomedical Engineering, Imperial College London, London, UK

1. Site-specific labeling with bis(arsenicals)

The familiar fluorescent labeling chemistries that have been extensively employed over many years are nowhere near sufficiently specific to be used to label a single protein in the complex milieu of a cell. The broad reactivity of groups such as *N*-hydroxysuccinimide (amines) or iodoacetamide (thiol) precludes their use in specific labeling of a singular cellular component. To achieve this, both a specific labeling reagent and a specific target sequence are necessary. The latter can be readily fused to any given protein so allowing a targeted labeling by the former. In addition to its specificity, the labeling reagent should be cell permeant and ideally nonfluorescent until it has reacted with the target sequence. Griffin *et al*. (1998) described a fluorescein-derived bis(arsenical) that reacted with a tetracysteine motif (CCXXCC) in a proposed α-helix. The chelate effect gives this compound a high (<nM) affinity for the cognate peptide so avoiding the labeling of other cellular cysteine-containing proteins and the potential toxicity arising from there. To further reduce toxicity, FlAsH is added as an ethanedithiol conjugate (FlAsH-EDT_2, Figure 1a) that undergoes dithiol exchange with the target peptides. In a subsequent publication (Adams *et al*., 2002), a more detailed analysis of the reaction of FlAsH and a resofufin analog (ReAsH, Figure 1b) with tetracysteine peptides is described. The apparent K_d values are in the pM range for optimum peptide sequence, where the latter was found to contain CCPGCC. The presence of the proline glycine sequence suggests that a hairpin rather than α-helix is the preferred conformation.

Nakanishi and coworkers in two papers (Nakanishi *et al*., 2001; Nakanishi *et al*., 2004) described similar bis(arsensical) derivatives of Nile Red (BArNile-EDT_2, Figure 1c) that could be used to label tetracysteine sequences fused to proteins in live cells. The fluorescence of Nile Red is sensitive to the hydrophobicity of its local environment and this was exploited to measure calcium-dependent conformational

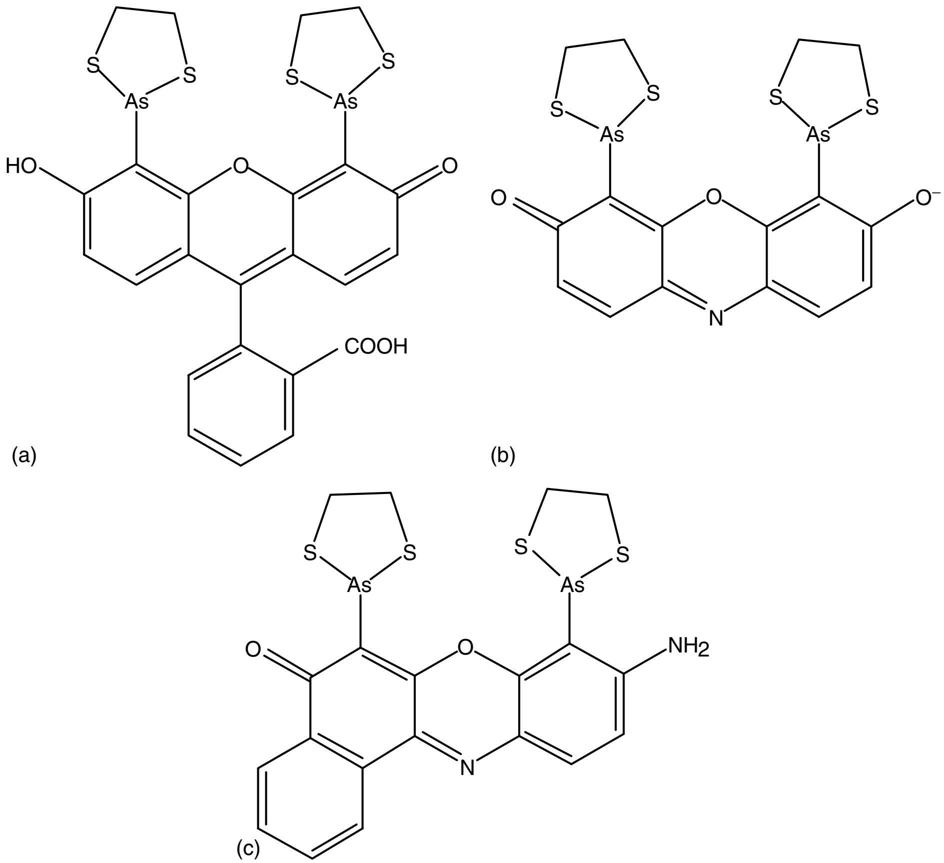


Figure 1 Structures of the bis(arsenical) fluorescent dyes that have been used to label tetracysteine peptide sequences in proteins expressed in live cells. (a) FlAsH-EDT$_2$. (b) ReAsH-EDT$_2$. (c) BArNile-EDT$_2$

changes in a calmodulin derivative. In their second paper (Nakanishi *et al*., 2004), this group further improved on the calcium responsiveness of their conjugates by incorporating a benzoic acid group.

One potential difficulty associated with in vivo labeling is the nonspecific attachment to other proteins, either due to cross-reaction with other cysteine-containing sequences or the natural occurrence of the target sequence. Both Griffin *et al*. (2000) and Stroffekova *et al*. (2001) identified nonspecific binding of FlAsH although they attributed it to different reasons (binding of the fluorophore to hydrophobic sites and reaction with other cysteine-containing proteins respectively).

2. Applications of in vivo labeling with bis(arsenicals)

In their original paper, Griffin *et al*. (1998) demonstrated the in vivo labeling of a tetracysteine-tagged enhanced cyan fluorescent protein (ECFP) via fluorescence resonance energy transfer (FRET) from the FlAsH moiety to the protein

fluorophore. Subsequent applications of this in vivo labeling method have been described and include the attachment of both FlAsH and ReAsH to tagged connexin43 (Gaietta *et al*., 2002; Sosinsky *et al*., 2003) to following trafficking of this protein across gap junctions. The availability of two different fluorophores that labeled cells at different time intervals allows the distinction between "old" and "young" protein and tracing of their individual fates (Gaietta *et al*., 2002). Moreover, the ReAsH moiety was able to bring about photopolymerization of 3,5-diaminobenzene so generating osmiophilic stains for subsequent electron microscopy so allowing localization of the labeled proteins at higher resolution than can be achieved by fluorescence microscopy. Many of the proteins that have been labeled in vivo have been recombinantly expressed in the host cell, however, this is not the only method by which a tetracysteine (TetCys) tag can be introduced; Andresen *et al*. (2004) used integrative epitope tagging to attach a TetCys tag (GSSGCCPGCC) to open reading frames (ORFs) in the chromosomes of the budding yeast *Saccharomyces cerevisiae*. To demonstrate this approach, these authors tagged β-tubulin and FlAsH-labeled it. In a systematic study of multiple TetCys tags as well as GFP (Green Fluorescent Protein), they demonstrated that the smaller tags (10-amino acids) had a less deleterious effect than the larger ones (20- or 30-amino acids) or than GFP on cell viability. After labeling, the modified tubulin was integrated into microtubules and could be readily imaged. In contrast to both Griffin *et al*. (2000) and Stroffekova *et al*. (2001), Andresen *et al*. (2004) observed very little nonspecific labeling in yeast cells.

In addition to its use for imaging, FlAsH-modified proteins have also been employed to selectively destroy the proteins to which they are attached by illumination of the cells at 488 nm. Marek and Davis (2002) demonstrated selective inactivation of FlAsH-modified synaptotagmin.

3. Alternative tag-selective labeling approaches

Several other examples of tag selective in vivo labeling chemistries have been described in the past 4 years, and while they differ in detail, they have the common aim of providing as unique as possible a site at which to attach the fluorophore (or other conjugating molecule).

Native chemical ligation exploits the unique reactivity of the thiol group of an N-terminal cysteine residue with thioesters and has been widely utilized in vitro for the attachment of many different small molecules site specifically to the N-terminus of target proteins (Yeo *et al*., 2004), as shown in Figure 2. While this technology has many uses, in the context of this review, it is the observation that the reaction shown in Figure 2 also occurs in vivo. The rarity of proteins with a natural N-terminal cysteine means that heterologous expression of a recombinant protein with an engineered N-terminal cysteine is selectively modified. Yeo *et al*. (2003) exploited exactly this approach to label glutathione-*S*-transferase (GST) having an engineered N-terminal cysteine with a variety of fluorophores in live *Escherichia coli* cells.

Strong, noncovalent interactions of suitable specificity can also form the basis of in vivo labeling methods. Miller *et al*. (2004) have taken advantage of the high

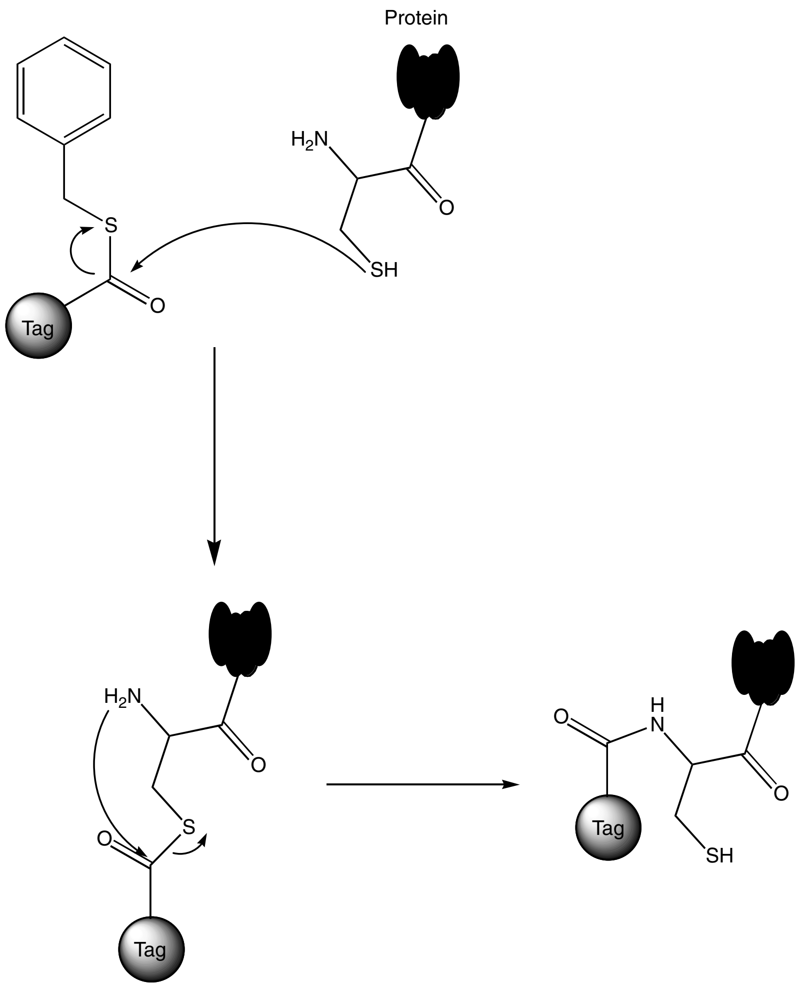


Figure 2 Native chemical ligation for the introduction of tags to proteins carrying an N-terminal cysteine residue

affinity (<nM) and slow dissociation rates (10^{-4} s^{-1}) of methotrexate derivatives from dihydrofolate reductase (DHFR) to target a fluorescent methotrexate to a DHFR fusion protein. To avoid binding to endogenous DHFR, the experiments were performed in $DHFR^{-}$ CHO (Chinese Hamster Ovary) cells. Fusing a myristoylation/palmitoylation sequence targeted the protein to the plasma membrane where the fluorescent dye was subsequently concentrated. To overcome the endogenous DHFR background, the authors propose to engineer DHFR to bind methotrexate derivatives not bound by the native protein. This particular method has close similarities to the well-established use of GFP fusion proteins. The smaller size of DHFR and the wider range of synthetic fluorophores may offer some advantages; however, overcoming the endogenous DHFR background will be essential to its further use.

An alternative approach is based on the specific covalent modification of fusion proteins that exploits the transfer of functional groups from alkylated guanine bases to the DNA repair enzyme O^6-alkylguanine-DNA alkyltransferase (hAGT). The reaction of hAGT with a labeled O^6 guanine derivative is shown in Figure 3. Methods to synthesize the labeled O^6 guanine derivatives were developed by Kindermann *et al*. (2004), and a mutant hAGT enzyme (W160) evolved to show a higher activity against O^6 benzylguanosine (Juillerat *et al*., 2003). Keppler *et al*., (2003, 2004a) demonstrated the transfer of a label to hAGT fused to CFP (cyan fluorescent protein), DHFR, GST, and a trimeric nuclear localization sequence from the SV40 large T antigen. Labeling of these fusion proteins was demonstrated in *E. coli*, *S. cerevisiae*, and CHO cells. Further, exemplification of this approach

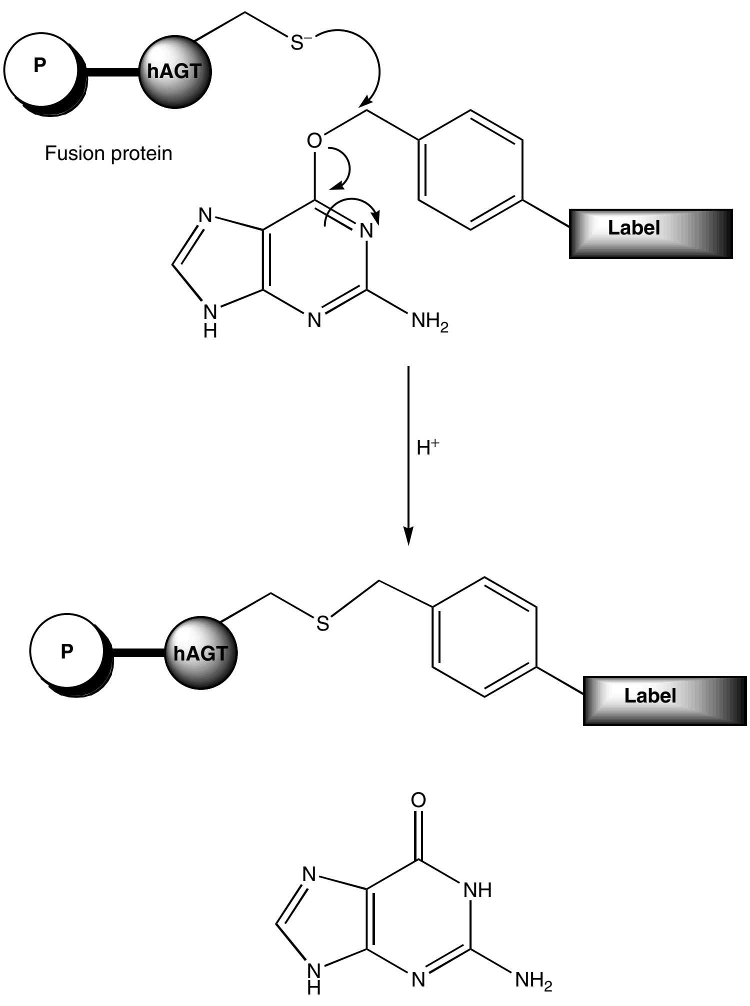


Figure 3 The mechanism of labeling of O^6-alkylguanine-DNA alkyltransferase fusion proteins with benzyl guanine derivatives

with a variety of animal cells was described by Keppler *et al*. (2004b), who fused different cell localization sequences to the hAGT and then labeled the expressed proteins in living cells.

4. Specific labeling of membrane proteins

Recently, Vogel and coworkers have described two approaches to live cell labeling of specific membrane proteins. Guignet *et al*. (2004) exploited the well-known ability of hexahisitdine tags to bind to nickel nitrilotriacetate (NTA) complexes. A nickel NTA complex modified with a fluorescent reporter group was found to complex with hexahisitdine-tagged membrane proteins (ionotropic and G-protein coupled receptors). The affinity of the metal chelate complexes were modest (μM), however, this was sufficient for imaging the proteins in cells.

Site-specific labeling of cell surface proteins has also been demonstrated by George *et al*. (2004) using an enzyme reaction (phosphopantetheine transferase (PPTase)). This enzyme transfers a modified CoA to a serine residue of the acyl carrier protein (ACP) fused to the membrane protein of interest. The fluorescent dye Cy3 labeled CoA was shown to be transferred to ACP fused to two membrane proteins (α-agglutinin receptor and neurokinin-1).

5. Conclusions

The past few years have seen a plethora of new approaches to the specific labeling of proteins in live cells. A combination of mutagenesis, heterologous expression, and fluorophore synthesis has generated new reagents for imaging the localization and trafficking of specific proteins that complement the now well-established GFP fusions.

Two directions for the development of this technology are signposted by recent work: the introduction of even less perturbing modifications and the use of these methods not just to follow protein localization but to use the proteins as indicators of ligand concentrations for both large and small molecules.

Further reading

Giriat I and Muir TW (2003) Protein semi-synthesis in living cells. *Journal of the American Chemical Society*, **125**(24), 7180–7181.

Marks KM, Braun PD and Nolan GP (2004) A general approach for chemical labeling and rapid, spatially controlled protein inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(27), 9982–9987.

References

Adams SR, Campbell RE, Gross LA, Martin BR, Walkup GK, Yao Y, Llopis J and Tsien RY (2002) New biarsenical Ligands and tetracysteine motifs for protein labeling in vitro and in vivo: synthesis and biological applications. *Journal of the American Chemical Society*, **124**(21), 6063–6076.

Andresen M, Schmitz-Salue R and Jakobs S (2004) Short tetracysteine tags to beta-tubulin demonstrate the significance of small labels for live cell imaging. *Molecular Biology of the Cell*, **15**(12), 5616–5622.

Gaietta G, Deerinck TJ, Adams SR, Bouwer J, Tour O, Laird DW, Sosinsky GE, Tsien RY and Ellisman MH (2002) Multicolor and electron microscopic imaging of connexin trafficking. *Science*, **296**(5567), 503–507.

George N, Pick H, Vogel H, Johnsson N and Johnsson K (2004) Specific labeling of cell surface proteins with chemically diverse compounds. *Journal of the American Chemical Society*, **126**(29), 8896–8897.

Griffin AB, Adams SR, Jones J and Tsien RY (2000) Fluorescent labeling of recombinant proteins in living cells with FlAsH. In *Methods in Enzymology*, Vol 327, Thorner J, Emr Scott D and Abelson JN (Eds.), Academic Press, pp. 565–578.

Griffin BA, Adams SR, Jones J and Tsien RY (1998) Specific Covalent Labelling of Recombinant Protein Molecules Inside Live Cells. *Science*, **281**, 269–272.

Guignet EG, Hovius R and Vogel H (2004) Reversible site-selective labeling of membrane proteins in live cells. *Nature Biotechnology*, **22**(4), 440–444.

Juillerat A, Gronemeyer T, Keppler A, Gendreizig S, Pick H, Vogel H and Johnsson K (2003) Directed evolution of O-6-alkylguanine-DNA alkyltransferase for efficient labeling of fusion proteins with small molecules in vivo. *Chemistry & Biology*, **10**(4), 313–317.

Keppler A, Gendreizig S, Gronemeyer T, Pick H, Vogel H and Johnsson K (2003) A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nature Biotechnology*, **21**(1), 86–89.

Keppler A, Kindermann M, Gendreizig S, Pick H, Vogel H and Johnsson K (2004a) Labeling of fusion proteins of O-6-alkylguanine-DNA alkyltransferase with small molecules in vivo and in vitro. *Methods*, **32**(4), 437–444.

Keppler A, Pick H, Arrivoli C, Vogel H and Johnsson K (2004b) Labeling of fusion proteins with synthetic fluorophores in live cells. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(27), 9955–9959.

Kindermann M, Sielaff I and Johnsson K (2004) Synthesis and characterization of bifunctional probes for the specific labeling of fusion proteins. *Bioorganic & Medicinal Chemistry Letters*, **14**(11), 2725–2728.

Marek KW and Davis GW (2002) Transgenically encoded protein photoinactivation (FIAsH-FALI): Acute inactivation of synaptotagmin I. *Neuron*, **36**(5), 805–813.

Miller LW, Sable J, Goelet P, Sheetz MR and Cornish VW (2004) Methotrexate conjugates: a molecular in vivo protein tag. *Angewandte Chemie (International Edition)*, **43**(13), 1672–1675.

Nakanishi J, Nakajima T, Sato M, Ozawa T, Tohda K and Umezawa Y (2001) Imaging of conformational changes of proteins with a new environment/sensitive fluorescent probe designed for site specific labeling of recombinant proteins in live cells. *Analytical Chemistry*, **73**(13), 2920–2928.

Nakanishi J, Maeda M and Umezawa Y (2004) A new protein conformation indicator based on biarsenical fluorescein with an extended benzoic acid moiety. *Analytical Sciences*, **20**(2), 273–278.

Sosinsky GE, Gaietta GM, Hand G, Deerinck TJ, Han A, Mackey M, Adams SR, Bouwer J, Tsien RY and Ellisman MH (2003) Tetracysteine genetic tags complexed with biarsenical ligands as a tool for investigating gap junction structure and dynamics. *Cell Communication and Adhesion*, **10**(4–6), 181–186.

Stroffekova K, Proenza C and Beam KG (2001) The protein-labeling reagent FLASH-EDT2 binds not only to CCXXCC motifs but also non-specifically to endogenous cysteine- rich proteins. *Pflugers Archiv-European Journal of Physiology*, **442**(6), 859–866.

Yeo DSY, Srinivasan R, Uttamchandani M, Chen GYJ, Zhu Q and Yao SQ (2003). Cell-permeable small molecule probes for site-specific labeling of proteins. *Chemical Communications*, (23), 2870–2871.

Yeo DSY, Srinivasan R, Chen GYJ and Yao SQ (2004) Expanded utility of the native chemical ligation reaction. *Chemistry-a European Journal*, **10**(19), 4664–4672.

Specialist Review

Using photoactivatable GFPs to study protein dynamics and function

Dmitry M. Chudakov and Konstantin A. Lukyanov
Institute of Bioorganic Chemistry, Moscow, Russia

1. Introduction

Green fluorescent protein (GFP) from hydroid jellyfish *Aequorea victoria* (Johnson *et al*., 1962; Prasher *et al*., 1992) and its mutants have transformed cell and molecular biology. Over the past few years, a number of GFP homologs of different colors including cyan, green, yellow, red fluorescent proteins, and nonfluorescent purple-blue chromoproteins have been identified and extensively characterized (Matz *et al*., 1999; Lukyanov *et al*., 2000). GFP-like proteins have been found not only in Cnidaria (in corals Anthozoa and jellyfishes Hydrozoa) but also in copepods (Arthropoda, Crustacea) (Shagin *et al*., 2004). The most remarkable feature of GFP-like proteins is their ability to form an intrinsic chromophore within the protein globule without the assistance of any other proteins or cofactors except oxygen (Heim *et al*., 1994). What makes the GFP family of proteins prominent is the fact that they represent the only genetically encoded fluorescent tags that can be widely used to label organisms, cells, organelles, and proteins, as well as to develop molecular sensors for different aspects of intracellular environment (Lippincott-Schwartz and Patterson, 2003; Verkhusha and Lukyanov, 2004).

GFP has turned out to be a very suitable tag for photobleaching experiments. It produces bright, stable fluorescence that does not fade at low levels of illumination. By contrast, with high illumination levels, the GFP fluorophore can be photobleached. Under conditions where the light intensity used for photobleaching does not noticeably perturb living cells (Patterson *et al*., 1997), a variety of photobleaching methods can be employed to monitor protein and organelle dynamics. This includes FRAP (fluorescence recovery after photobleaching) or FLIP (fluorescence loss in photobleaching), which can give information about protein diffusion rate, mobile fraction, and exchange rate between bleached and unbleached areas (Lippincott-Schwartz *et al*., 2003; *see also* Article 53, **Photobleaching (FRAP/FLIP) and dynamic imaging**, Volume 5). However, these techniques do not enable direct tracking of the chosen object's movement using its fluorescent signal as the bleached proteins are invisible.

In order to directly monitor the trafficking of a specific population of tagged proteins, other methods had to be developed. One method used photobleaching of an acceptor protein in a FRET (fluorescence resonance energy transfer) pair, which resulted in an increase in donor protein fluorescence (Marchant *et al*., 2001). An alternative FLAP (fluorescence localization after photobleaching) technique was also reported (Dunn *et al*., 2002), involving two fluorophores, one of which is bleached and the other is used as a fluorescent control. The FLAP signal was calculated as the difference between the intensity of the light emitted by the two fluorophores. For reliable direct tracking, however, it has been necessary to develop directly photoactivatable fluorescent proteins, that is, proteins in which light of definite wavelength and intensity induces structural changes leading to the appearance of novel and greatly enhanced fluorescent signal. Several important advances have recently been reported in the development of such photoactivatable fluorescent proteins, and these are discussed below.

2. GFP photoactivation

The GFP fluorescence excitation spectrum has two peaks - with maxima at 396 nm and 476 nm, which correspond to two close emission peaks with maxima at about 505 nm. Two excitation peaks, with a 6:1 ratio, correspond to protonated (neutral) and deprotonated (anionic) chromophore forms (Chattoraj *et al*., 1996; Niwa *et al*., 1996; Brejc *et al*., 1997). In 1996, it was reported that UV irradiation causes reproportioning of excitation peaks in favor of the 476-nm peak (Yokoe and Meyer, 1996). That was the first time that GFP had been shown to be capable of photoactivation. However, the contrast between the photoactivated and initial 476-nm-excited GFP fluorescence reached only threefold. This contrast was insufficient for most experiments, and the technique was not widely adopted.

A mechanism for UV-induced GFP photoconversion has been revealed more recently (van Thor *et al*., 2002) (Figure 1a, right panel). Mass-spectral and crystallographic studies of the photoconverted protein showed that decarboxylation of Glu222 occurs during this process. As a result, a proton transfer pathway responsible for the equilibrium between the protonated and deprotonated states of chromophore becomes disrupted. Thus, the chromophore in photoconverted GFP exists mainly in the anionic state, absorbing light at about 480 nm.

In 2002, a GFP mutant was reported (Patterson and Lippincott-Schwartz, 2002), capable of much more effective photoactivation. This mutant protein, named PA-GFP, contains a single T203H substitution that resulted in the prevalence of protonated chromophore state before photoactivation. In this state, the protein is characterized by a single absorption/excitation peak at 400 nm and green fluorescence at 517 nm (Figure 1a, Table 1). Thus, PA-GFP gives almost zero fluorescence when excited at 480–500 nm. Intense UV-violet light irradiation (about 400 nm) causes irreversible chromophore photoconversion to the anionic form, which possesses absorption/excitation maximum at 504 nm. The contrast in the 504 nm-excited fluorescence before and after PA-GFP activation reaches 100-fold, making it suitable for many applications.

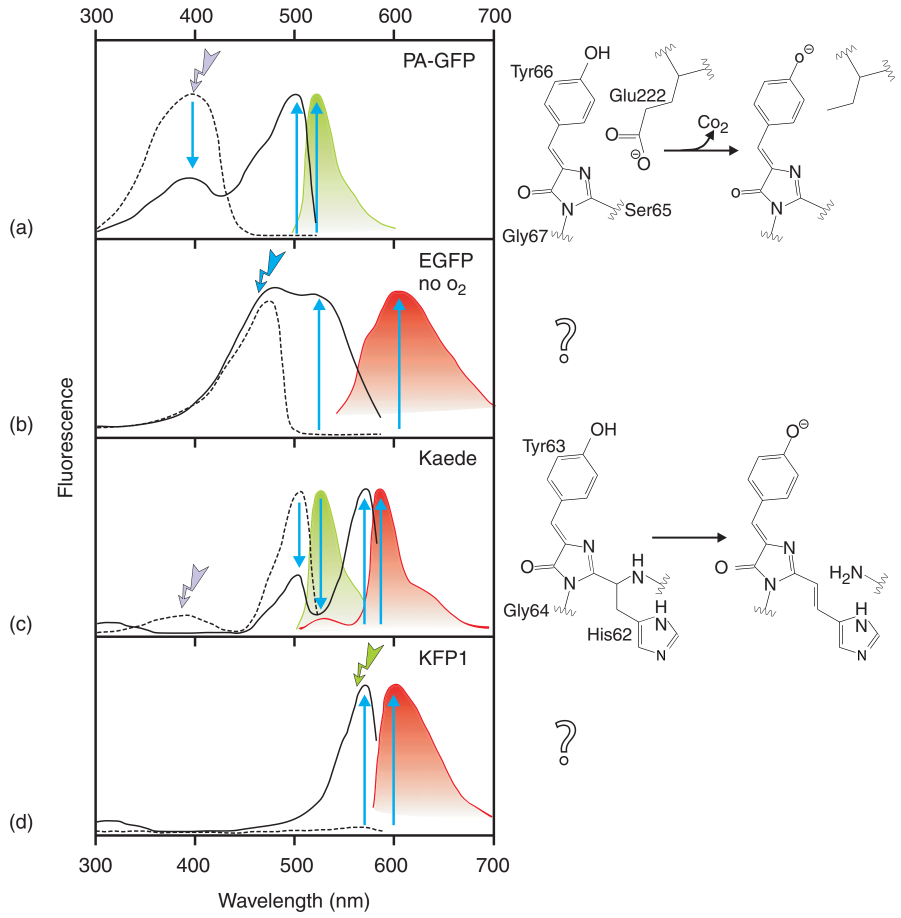


Figure 1 Schematic representation of spectral and structural changes underlying photoactivatable proteins conversion. Left: Excitation spectra before (dashed lines) and after (solid lines) photoactivation are shown black. Colored filled graphs show emission spectra. Straight blue arrows show spectral changes during photoconversion. Zigzag arrows show activating light. Right: Molecular mechanisms underlying photoactivation are shown for PA-GFP and Kaede

3. GFP photoactivation in anaerobic conditions

Evolution of this theme resembles a detective story. Soon after GFP photoactivation in aerobic conditions was published (Yokoe and Meyer, 1996), it was discovered that blue light (488 nm) GFP irradiation under anaerobic conditions made it red fluorescent (Elowitz *et al*., 1997; Sawin and Nurse, 1997). In the activated state, GFP possessed broad excitation-emission peaks at 525 and 600 nm, respectively (Figure 1b). This technology allowed red fluorescent signal activation and in vivo tracking of the movement of cellular proteins. The monomeric properties of GFP made it a useful tool, though restricted its application to anaerobic systems. This method was successfully applied (Elowitz *et al*., 1999) to measure a GFP diffusion coefficient in *Escherichia coli*.

Table 1 Photoactivatable fluorescent proteins

		PA-GFP	Kaede	KFP1[a]
Oligomeric state		Monomer	Tetramer	Tetramer
Activating light		UV-violet	UV-violet	Green
Quenching light		-	-	Blue, max at 450 nm
Excitation changes, nm		400–504	508–572	Growth at 590
Emission changes, nm		-	518–580	Growth at 600
Photoactivation reversibility		Irreversible	Irreversible	Reversible or irreversible
Fluorescence increase, fold		100	800	70 or 35
Ratio change, fold		~200	2000	-
Before activation	QY[b]	0.13	0.88	<0.001
	EC[c]	20 700	98 800	123 000
	pK_a	4.5	5.6	n.d.
	Brightness[d]	0.08	2.64	<0.004
After activation[a]	QY[b]	0.79	0.33	0.07
	EC[c]	17 400	60 400	59 000
	pK_a	n.d.	5.6	n.d.
	Brightness[d]	0.42	0.60	0.13

Protein advantages as a photoactivatable label are marked green, drawbacks - red.
[a]Contrast, quantum yield, and extinction coefficient after activation are given for the irreversibly kindled KFP1.
[b]QY, quantum yield.
[c]EC, extinction coefficient, $M^{-1}\,cm^{-1}$.
[d]As compared to brightness (extinction coefficient multiplied by quantum yield) of EGFP.

Sacchetti *et al*. (1999) reported that photoactivatable red signal observed in bacteria is produced by endogenous porphyrins, but not by the photoactivated GFP. Indeed, a substance they purified from *E. coli* definitely looked like porphyrins, according to the data given. Nevertheless, a marked difference in the shape of emission spectra between porphyrins and red photoactivated GFP gave rise to doubts.

In 2003, a novel article was published (Jakobs *et al*., 2003), describing a new successful application of GFP photoconversion to red form for investigating the continuity of mitochondrial matrix in living yeast. The authors also showed that fluorescence lifetime for the red photoconverted GFP chromophore is 2.11 ns, which was close to the usual green GFP fluorescence lifetime but not to that of porphyrins (about 15 ns).

Summing up, GFP photoconversion to the red fluorescent form under anaerobic conditions is a real phenomenon, which could be applied to object tracking within oxygen-free systems. Unfortunately, as GFP chromophore maturation necessarily requires oxygen, this technique was restricted to facultative anaerobes.

In contrast to GFP photoactivation under aerobic conditions, where the events underlying the photoconversion are generally clear (van Thor *et al*., 2002), the mechanism of GFP photoconversion to red fluorescent form remains unclear. This mechanism probably hides new important information about GFP chromophore functioning and also clues to the development of novel photoactivatable proteins.

4. Kaede

An intriguing fluorescent protein was cloned in 2002 from a stony coral *Trachyphyllia geoffroyi* (Ando *et al*., 2002). This protein was named "Kaede" ("Kaede" means a maple leaf in Japanese). In the dark, Kaede matures to a green fluorescent form. In response to UV-violet light irradiation, Kaede undergoes irreversible photoconversion to a red fluorescent form (Figure 1c, Table 1).

Soon after the Kaede discovery, the structural basis for its photoconversion was revealed (Mizuno *et al*., 2001) (Figure 1c, right panel). First, a usual GFP-like green-emitting chromophore is formed by His62-Tyr63-Gly64. Then, a red-emitting chromophore is formed under UV light irradiation as a result of protein backbone cleavage between amide N and αC of His62 and subsequent double bond formation between αC and βC of His62. Thus, the Kaede red chromophore includes the His62 ring in its conjugated π-system. The resulting increase in the red-to-green fluorescence ratio after photoactivation reaches more than 2000-fold, making it the most contrasting photoactivatable fluorescent protein currently developed.

5. Kindling fluorescent proteins (KFPs)

This group of photoactivatable proteins arose from a unique natural protein, which determines purple coloration of tentacles of sea anemone *Anemonia sulcata*. This protein was designated as asFP595 (Lukyanov *et al*., 2000), although it is variously also referred to asCP562 (Wiedenmann *et al*., 2000), asCP (Gurskaya *et al*., 2001), and asulCP (Labas *et al*., 2002). We will use the last designation.

Initially nonfluorescent asulCP becomes red fluorescent in response to irradiation with intense green light (kindling) and then relaxes back to a nonfluorescent state (Lukyanov *et al*., 2000). Blue light (action spectrum peaked at 450 nm) instantly reverts the activated protein to the nonfluorescent state (quenching). For the wild-type protein, both kindling and quenching are completely reversible processes.

Several mutants of asulCP with altered characteristics of kindling and quenching were generated (Chudakov *et al*., 2003a,b). Moreover, similar photoactivation behavior was transferred to two nonkindling, nonfluorescent coral chromoproteins, hcriCP from *Heteractis crispa* and cgigCP from *Condilactis gigantea* (Gurskaya *et al*., 2001), by point mutations of the chromophore environment (Chudakov *et al*., 2003a). All KFPs become red fluorescent after kindling (excitation peak at 580–590 nm, emission peak at 600–630 nm). Importantly, most mutants were capable of irreversible kindling in response to the more intense and/or enduring irradiation. The best mutant kindling variant of asulCP, named KFP1 (Chudakov *et al*., 2003b), demonstrated about a 70-fold reversible or 35-fold irreversible fluorescence increase in response to irradiation by intense green-yellow light (Figure 1d, Table 1).

Most probably, reversible photoactivation of KFPs can be explained by changes in protein conformation (chromophore cis-trans isomerization and/or rearrangement of chromophore and environment). However, the exact nature of both reversible and irreversible kindling effects remains unclear, waiting for further investigations.

6. Practical use of PA-GFP, Kaede, and KFP1

Among the GFP-like photoactivatable tags, only three proteins, PA-GFP, Kaede, and KFP1 are currently suitable for a wide range of practical applications (green-to-red GFP conversion requires anaerobic conditions and thus cannot be used in most biological models). Three "levels" of photolabeling could be recognized: labeling of whole cells, cell organelles, and proteins. Examples of potential applications of precise photolabeling are summarized in Table 2. In fact, photoactivatable proteins can find their use in any task aimed at monitoring the dynamics of objects in complex environments.

The photoactivation process could be divided onto three main steps: primary visualization of a photoactivatable tag, its precise photoactivation in a selected area, and tracing the resulting fluorescent signal. These steps differ significantly for PA-GFP, Kaede, and KFP1. Before activation, the green fluorescence of PA-GFP is visible under UV-violet light excitation (low-intensity light must be applied to PA-GFP to avoid premature photoconversion). Kaede initially represents a conventional GFP, thus it can be viewed when excited with blue light. Visualization of nonfluorescent KFP1 is possible due to reversible kindling by green light of medium intensity.

Irreversible photoactivation of a target object is achieved by strong irradiation with UV-violet (for PA-GFP and Kaede) or green (for KFP1) light. After photoconversion, PA-GFP acquires a new excitation peak at 504 nm. Its green fluorescence can be viewed in blue light against nonfluorescent background of nonactivated protein. Activated Kaede becomes red fluorescent and can be discriminated easily from green nonactivated protein. Irreversibly kindled KFP1 fluoresces red. It can be visualized after excitation with weak green light insufficient

Table 2 Potential applications of photoactivatable fluorescent proteins

Precise optical labeling of . . .	Possible areas of investigation
Cells	Cell tracking in individual development.
	Lymphocyte tracking.
	Cell tracking in cancer and metastasis.
	Tracking of free-living and pathogenic unicellular organisms.
Organelles	Visualization of individual cells of intricate shapes (e.g., neurons) in a complex network.
	Rate and direction of intracellular movement of organelles.
	Investigation of organelle continuity.
	Interchange of contents between organelles.
	Membrane fluidity.
	Intercellular exchange of organelles.
Proteins	Intracellular diffusion measurements.
	Rate and direction of movement of target proteins.
	Assembling–disassembling of cytoskeleton, nucleoli, etc.
	Protein passing through pathways of posttranslation modifications (Golgi, ER) and degradation (endosomes).
	Tracking viruses within a host.
	Intercellular exchange of proteins.

to activate the whole pool of KFP1. Background fluorescence of nonactivated KFP1 can be additionally decreased by quenching blue light.

None of the photoactivatable proteins reported are perfect, and their advantages and disadvantages are marked green and red in Table 1. PA-GFP is a monomeric protein, while Kaede and KFP1 are tetramers. Thus, only PA-GFP can be used safely in fusions with other proteins. The successful application of PA-GFP to protein tracking has been reported recently (Tulu *et al*., 2003). Tetrameric state of Kaede and KFP1 can impede proper folding and/or functioning of tagged proteins. So, the main area of Kaede and KFP1 applications lies in cell and organelle photolabeling. Considering the required wavelengths of activation light, one can see that KFP1 requires well-tolerated green light, while both PA-GFP and Kaede need much more phototoxic UV-violet light. Another advantage of KFP1 is that it "occupies" only one (red) fluorescence color allowing multicolor labeling together with other GFP-like proteins, for example, PA-GFP or any other GFP variant. Kaede demonstrates the highest contrast due to a simultaneous decrease in green and an increase in red fluorescence. However, Kaede engages two fluorescent colors, and thus it cannot be used simultaneously with GFP. Finally, red photoactivatable proteins, Kaede and KFP1, have the general advantages of red dyes, which suffer less interference from background cellular autofluorescence, and which penetrate deeper through animal tissues compared to shorter wavelength emission.

Development of novel monomeric photoactivatable proteins of different colors requiring activating light of different wavelengths remains one of the most urgent tasks. Red and far-red proteins that can be activated by low phototoxic orange-red or multiphoton infrared light are of special interest. A useful set of photoactivatable GFP-like proteins will extend the application area for noninvasive optical imaging methods to investigate living objects.

7. Photoactivatable fluorescent proteins in nature

Most photoactivatable proteins described in this review are natural or at least based on natural properties of wild-type proteins found in Anthozoa and Hydrozoa species. For example, a natural protein Kaede rapidly converts to red fluorescent form under UV-violet light irradiation of relatively low intensity. Its photoactivation capability was found accidentally – when the protein was left near the window! KFP was first cloned from the *Anemonia sulcata* – and it is possible that photoactivation of this protein plays a functional role in the anemone. PA-GFP activation is based on the decarboxylation of the Glu222 under UV-violet light irradiation, which probably occurs in nature in different GFP-like proteins as a way of posttranslational protein modification.

Many of the GFP-like proteins probably undergo some photoactivation in natural environment within the organisms of Cnidaria and Crustacea species under intense sunlight. Since the function(s) of GFP-like proteins remain(s) unclear, it is especially difficult to speculate the role of photoactivatable proteins in living organisms. However, it is possible to assume that active and self-dependent adaptation of these proteins to sunlight allows organisms to accommodate changing

light and temperature conditions without changes of gene expression. Anyway, it makes sense to continue to search for novel photoactivatable fluorescent proteins in nature.

Acknowledgments

This work was supported by European Commission FP-6 Integrated Project LSHG-CT-2003-503259, and Russian Academy of Sciences for the program "Molecular and Cellular Biology".

References

Ando R, Hama H, Yamamoto-Hino M, Mizuno H and Miyawaki A (2002) An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12651–12656.

Brejc K, Sixma TK, Kitts PA, Kain SR, Tsien RY, Ormo M and Remington SJ (1997) Structural basis for dual excitation and photoisomerization of the Aequorea victoria green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 2306–2311.

Chattoraj M, King BA, Bublitz GU and Boxer SG (1996) Ultra-fast excited state dynamics in green fluorescent protein: Multiple states and proton transfer. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 8362–8367.

Chudakov DM, Belousov VV, Zaraisky AG, Novoselov VV, Staroverov DB, Zorov DB, Lukyanov S and Lukyanov KA (2003b) Kindling fluorescent proteins for precise in vivo photolabeling. *Nature Biotechnology*, **21**, 191–194.

Chudakov DM, Feofanov AV, Mudrik NN, Lukyanov S and Lukyanov KA (2003a) Chromophore environment provides clue to "kindling fluorescent protein" riddle. *Journal of Biological Chemistry*, **278**, 7215–7219.

Dunn GA, Dobbie IM, Monypenny J, Holt MR and Zicha D (2002) Fluorescence localization after photobleaching (FLAP): A new method for studying protein dynamics in living cells. *Journal of Microscopy*, **205**, 109–112.

Elowitz MB, Surette MG, Wolf PE, Stock J and Leibler S (1997) Photoactivation turns green fluorescent protein red. *Current Biology*, **7**, 809–812.

Elowitz MB, Surette MG, Wolf PE, Stock JB and Leibler S (1999) Protein mobility in the cytoplasm of *Escherichia coli*. *Journal of Bacteriology*, **181**, 197–203.

Gurskaya NG, Fradkov AF, Terskikh A, Matz MV, Labas YA, Martynov VI, Yanushevich YG, Lukyanov KA and Lukyanov SA (2001) GFP-like chromoproteins as a source of far-red fluorescent proteins. *FEBS Letters*, **507**, 16–20.

Heim R, Prasher DC and Tsien RY (1994) Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 12501–12504.

Jakobs S, Schauss AC and Hell SW (2003) Photoconversion of matrix targeted GFP enables analysis of continuity and intermixing of the mitochondrial lumen. *FEBS Letters*, **554**, 194–200.

Johnson FH, Shimomura O, Saiga Y, Gershman LC, Reynolds G and Waters JR (1962) Quantum efficiency of Cybridina luminescences, with a note on that of Aequorea. *Journal of Cellular and Comparative Physiology*, **60**, 85–103.

Labas YA, Gurskaya NG, Yanushevich YG, Fradkov AF, Lukyanov KA, Lukyanov SA and Matz MV (2002) Diversity and evolution of the green fluorescent protein family. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4256–4261.

Lippincott-Schwartz J, Altan-Bonnet N and Patterson GH (2003) Photobleaching and photoactivation: Following protein dynamics in living cells. *Nature Cell Biology*, (Suppl) S7-S14.

Lippincott-Schwartz J and Patterson GH (2003) Development and use of fluorescent protein markers in living cells. *Science*, **300**, 87–91.

Lukyanov KA, Fradkov AF, Gurskaya NG, Matz MV, Labas YA, Savitsky AP, Markelov ML, Zaraisky AG, Zhao X, Fang Y, *et al*. (2000) Natural animal coloration can be determined by a non-fluorescent GFP homolog. *Journal of Biological Chemistry*, **275**, 25879–25882.

Marchant JS, Stutzmann GE, Leissring MA, LaFerla FM and Parker I (2001) Multiphoton-evoked color change of DsRed as an optical highlighter for cellular and subcellular labeling. *Nature Biotechnology*, **19**, 645–649.

Matz MV, Fradkov AF, Labas YA, Savitsky AP, Zaraisky AG, Markelov ML and Lukyanov SA (1999) Fluorescent proteins from nonbioluminescent *Anthozoa* species. *Nature Biotechnology*, **17**, 969–973.

Mizuno H, Sawano A, Eli P, Hama H and Miyawaki A (2001) Photo-induced peptide cleavage in the green-to-red conversion of a fluorescent protein. *Biochemistry*, **40**, 2502–2510.

Niwa H, Inouye S, Hirano T, Matsuno T, Kojima S, Kubota M, Ohashi M and Tsuji FI (1996) Chemical nature of the light emitter of the Aequorea green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13617–13622.

Patterson GH and Lippincott-Schwartz J (2002) A photoactivatable GFP for selective photolabeling of proteins and cells. *Science*, **297**, 1873–1877.

Patterson GH, Knobel SM, Sharif WD, Kain SR and Piston DW (1997) Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophysical Journal*, **73**, 2782–2790.

Prasher DC, Eckenrode VK, Ward WW, Prendergast FG and Cormier MJ (1992) Primary structure of the Aeqourea victoria green fluorescent protein. *Gene*, **111**, 229–233.

Sacchetti A, Cappetti V, Crescenzi C, Celli N, Rotilio D and Alberti S (1999) Red GFP and endogenous porphyrins. *Current Biology*, **9**, 391–393.

Sawin KE and Nurse P (1997) Photoactivation of green fluorescent protein. *Current Biology*, **7**, 606–607.

Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM, *et al*. (2004) GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Molecular Biology and Evolution*, **21**, 841–850.

Tulu US, Rusan NM and Wadsworth P (2003) Peripheral, non-centrosome-associated microtubules contribute to spindle formation in centrosome-containing cells. *Current Biology*, **13**, 1894–1899.

van Thor JJ, Gensch T, Hellingwerf KJ and Johnson LN (2002) Phototransformation of green fluorescent protein with UV and visible light leads to decarboxylation of glutamate 222. *Nature Structural Biology*, **9**, 37–41.

Verkhusha VV and Lukyanov KA (2004) The molecular properties and applications of Anthozoa fluorescent proteins and chromoproteins. *Nature Biotechnology*, **22**, 289–296.

Wiedenmann J, Elke C, Spindler KD and Funke W (2000) Cracks in the beta-can: Fluorescent proteins from Anemonia sulcata (Anthozoa, Actinaria). *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14091–14096.

Yokoe H and Meyer T (1996) Spatial dynamics of GFP-tagged proteins investigated by local fluorescence enhancement. *Nature Biotechnology*, **14**, 1252–1256.

Specialist Review

FRET-based reporters for intracellular enzyme activity

Moritoshi Sato and Yoshio Umezawa

The University of Tokyo, Tokyo, Japan

1. Introduction

Many breakthroughs in our understanding of signal transduction, especially in complex tissues such as those found in the central nervous system, have depended on decoding the spatial and temporal dynamics of intra- and extracellular signaling biomolecules. The most generally applicable and popular techniques with a high spatial and temporal resolution include the use of optical probes. Fluorescent reporters have been developed that report local concentrations of the signaling biomolecules or ions of interest by alternations in the amplitude or wavelength distribution of their excitation or emission spectra. For instance, synthetic fluorescent Ca^{2+} reporters including Fura-2, Indo-1, and Quin-2 contributed to reveal intracellular dynamics of Ca^{2+}, such as Ca^{2+} waves and Ca^{2+} oscillations, under a fluorescence microscope (Tsien, 1994). Also, synthetic fluorescent reporters were recently developed for other ions and small biomolecules such as Zn^{2+} and nitric oxide (Walkup and Imperiali, 1997; Kojima *et al*., 1998). Green fluorescent protein (GFP) was recently cloned from the jellyfish *Aequorea victoria* and used for genetically labeling the signaling proteins, instead of conventional synthetic dyes such as fluorescein and rhodamine (Tsien, 1998). This genetic labeling method of cellular proteins facilitated direct observation of the proteins not only in living cells but also in tissues and organs from transgenic animals expressing fusion proteins of GFP and the proteins of interest. Mutagenesis studies further generated different-colored GFP variants, such as blue fluorescent proterin (BFP), cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP) (Tsien, 1998). This GFP technology has accelerated the generation of fluorescent reporters for signaling molecules, because pairs of these different-colored GFP variants respectively act as a donor and acceptor for fluorescence resonance energy transfer (FRET) (Miyawaki and Tsien, 2000). So far, we and other groups have developed genetically encoded fluorescent reporters for visualizing cellular key signaling species, including Ca^{2+} (Miyawaki and Tsien, 2000), cyclic AMP (Zaccolo *et al*., 1999), cyclic GMP (Sato *et al*., 2000), small G-proteins (Mochizuki *et al*., 2001). These reporters revealed novel spatio-temporal dynamics of each signaling player in living cells. Further development of fluorescent reporters for pivotal signaling processes should be an

active area of interest. In this review, we discuss fluorescent reporters for protein phosphorylation–based signal transduction and lipid second messengers.

2. Protein phosphorylation

Protein phosphorylation by intracellular kinases plays one of the most pivotal roles in signaling pathways within cells (Hunter, 2000). To reveal the biological functions of these kinases, electrophoresis, immunocytochemistry, and in vitro kinase assay have been used. However, these conventional methods do not provide enough information about the spatial and temporal dynamics of the signal transduction based on protein phosphorylation and dephosphorylation in living cells. To overcome these limitations for investigating kinase signaling, we developed genetically encoded fluorescent reporters for visualizing protein phosphorylation in living cells (Sato *et al*., 2000) (Figure 1). Using these reporters, we visualized under a fluorescence microscope when, where, and how the protein kinases are activated in single living cells.

To construct the fluorescent reporters for protein phosphorylation, a substrate domain for a protein kinase of interest is fused with a phosphorylation recognition domain via a flexible linker sequence (Figure 1). The tandem fusion unit consisting

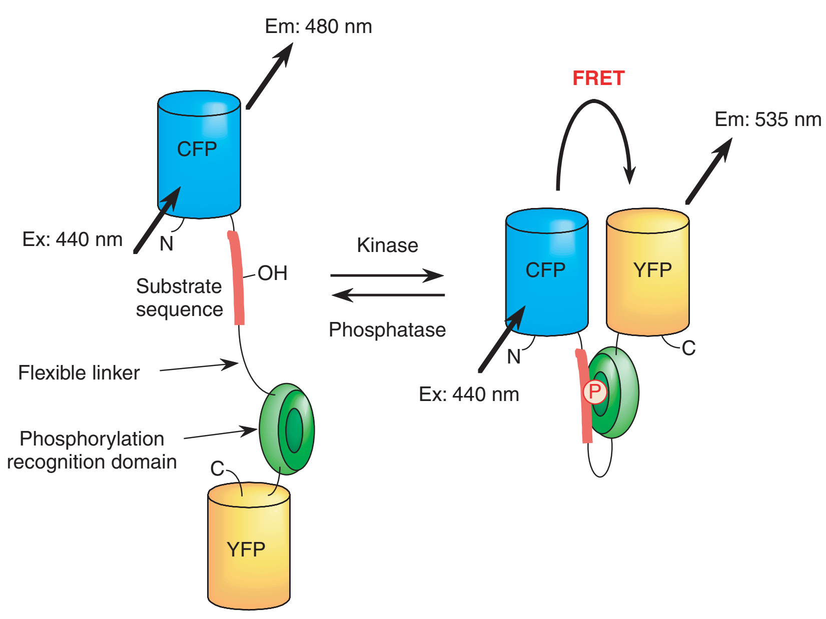


Figure 1 Principle of a fluorescent reporter for protein phosphorylation, which was named phocus. Upon phosphorylation of the substrate domain within phocus by the protein kinase, the adjacent phosphorylation recognition domain binds with the phosphorylated substrate domain, which changes the efficiency of FRET between the GFP mutants within phocus. By tethering a localization domain with phocus, the phocus can be localized in the specific intracellular locus of interest to visualize the local phosphorylation event there. Abbreviations: CFP: cyan fluorescent protein; YFP: yellow fluorescent protein; P in an open circle, the phosphorylated residue

of the substrate domain, linker sequence and phosphorylation recognition domain, is sandwiched with two different color fluorescent proteins, CFP and YFP, which serve as the donor and acceptor fluorophores for FRET (Figure 1). As a result of phosphorylation of the substrate domain and subsequent binding of the phosphorylated substrate domain with the adjacent phosphorylation recognition domain, FRET is induced between the two fluorescent units, which elicits the phosphorylation-dependent changes in fluorescence emission ratios of the donor and acceptor fluorophores (Figure 1). Upon activation of phosphatases, the phosphorylated substrate domain is dephosphorylated and the FRET signal is decreased (Figure 1). We named this reporter "phocus" (a fluorescent reporter for protein phosphorylation that can be custom-made). By using suitable substrate and phosphorylation recognition domains, we have developed a large number of phocuses for several key protein kinases including a receptor tyrosine kinase, the insulin receptor (Sato *et al*., 2002) and a serine/threonine protein kinase, Akt/PKB (Sasaki *et al*., 2003) (Table 1). In addition, substrates for protein kinases and phosphatases can be targeted to specific subcellular compartments including mitochondria, the Golgi apparatus, nucleus, and plasma membrane in living cells, which are thought to be critical for specific signal transduction in the respective intracellular loci. Thus, we further tailored our phocuses to analyze the phosphorylation events in such particular locations in single living cells (Table 1).

2.1. A phocus for imaging phosphorylation by insulin receptor

IRS-1 is one of the major substrates of insulin receptor. It contains a Peckstrin Homology (PH) domain and a phosphotyrosine binding (PTB) domain at its N-terminal end. The PH domain binds to phosphoinositides in the plasma membrane, and the PTB domain binds to the juxta-membrane domain of insulin receptor, which is tyrosine-phosphorylated almost immediately after insulin binding to the receptor. Thus, the concentration of IRS-1 is thought to be increased

Table 1 Detailed information on composition of each fluorescent indicator for protein phosphorylation and on each location to be directed in single living cells

Kinase	Indicator	Localization domain	Localization	Substrate domain	Phosphorylation recognition domain
Insulin receptor (tyrosine kinase)	Phocus-2pp	PH-PTB domain from IRS-1	Cytosol, nucleus, and membrane ruffles including insulin receptor	ETGTEEYMKMDLG	$p85\alpha_{330-429}$
	Aktus	–	Cytosol	RGRSRSAP	$14\text{-}3\text{-}3\eta_{82-235}$
Akt/PKB (serine/threonine kinase)	eNOS-Aktus	$eNOS_{1-35}$	Golgi	RGRSRSAP	$14\text{-}3\text{-}3\eta_{82-235}$
	Bad-Aktus	$Tom20_{1-33}$	Mitochondria	RGRSRSAP	$14\text{-}3\text{-}3\eta_{82-235}$

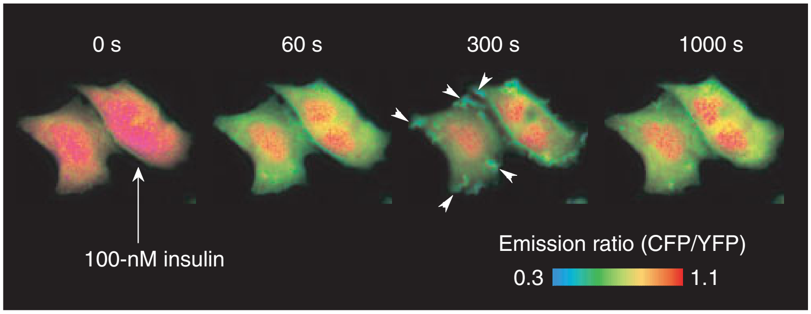


Figure 2 Fluorescence imaging with phocus-2pp upon insulin stimulation. Pseudocolor images of the CFP/YFP emission ratio are shown before (time 0 s) and at 60, 300, and 1000 s after the addition of 100-nM insulin at 25 °C, obtained from the CHO-IR cells expressing phocus-2pp. Insulin-induced accumulation of phocus-2pp at the membrane ruffles is indicated by white arrows in the image at 300 s

around the insulin receptor at the plasma membrane upon insulin stimulation. These PH and PTB domains were fused with phocus, named phocus-2pp, to locate the phocus around the insulin receptor like IRS-1 and to measure the local phosphorylation event there (Table 1). When phocus-2pp was expressed in CHO-IR cells, fluorescence was observed throughout the cells (Figure 2, time 0 s). Upon insulin stimulation, the CFP/YFP emission ratio expressed with pseudocolor was decreased in the cytosol owing to phosphorylation-induced FRET from CFP to YFP within phocus-2pp. 300 s after insulin stimulation, membrane ruffles, in which a large amount of phocus-2pp were accumulated, appeared around the plasma membrane and disappeared in 1000 s (Figure 2). In these membrane ruffles, phocus-2pp was found to colocalize with the insulin receptor. Interestingly, in these membrane ruffles, the extent of phocus-2pp phosphorylation was visualized to be ∼2-fold greater than that in the cytosol. This difference in phosphorylation levels between intracellular loci could be due to different balance of kinase and phosphatase activities between these intracellular loci. Phocus-2pp should contribute to uncovering the biological significance of such characteristic domains for tyrosine kinase signaling in membrane ruffles with high spatial and temporal resolution.

2.2. *Aktus for imaging phosphorylation by Akt/PKB*

Akt/PKB is a serine/threonine kinase that regulates a variety of cellular responses, such as cell proliferation, cell survival, and angiogenesis (Marte and Downward, 1997). To provide information on the spatial and temporal dynamics of the Akt activity in single living cells, we have developed a genetically encoded fluorescent reporter for Akt, named *Aktus* (Sasaki *et al*., 2003). Almost all Akt substrates are localized to subcellular regions. For example, eNOS (Fulton *et al*., 2001), which mediates a vasodilatory effect by nitric oxide production, is predominantly localized to the Golgi apparatus, while Bad (Chao and Korsmeyer, 1998), which is related to apoptosis promotion, is present in mitochondrial outer membranes. By

Table 2 Differential subcellular localization of activated Akt between cellular stimuli

	Cytosol[a]	Golgi[b]	Mitochondria[c]
Insulin	−	+	−
17β-estradiol	−	+	+

[a]Measured with Aktus.
[b]Measured with eNOS-Aktus.
[c]Measured with Bad-Aktus.

fusing the Aktus with the respectively subcellular localization domains within the eNOS and Bad, eNOS-Aktus and Bad-Aktus, which are respectively localized to the Golgi apparatus and mitochondrial outer membrane, were developed as shown in Table 1 and compared with the cytosolic diffusible reporter, Aktus (Sasaki *et al*., 2003). We have shown that in vascular endothelial cells, the Golgi-localized reporter, eNOS-Aktus, was phosphorylated upon stimulation with insulin and with 17β-estradiol, whereas the mitochondria-localized Bad-Aktus was phosphorylated in response to 17β-estradiol but not by insulin (Table 2). On the other hand, the diffusible reporter, Aktus, was not efficiently phosphorylated either upon insulin or 17β-estradiol stimulation (Table 2). These results suggest that activated Akt is localized to various subcellular compartments including the Golgi apparatus and/or mitochondria rather than diffusing throughout the cytosol, thereby efficiently phosphorylating its substrate proteins. Our observations with the mitochondria-localized reporter suggest that localization of activated Akt to mitochondria is triggered by 17β-estradiol but not by insulin. The present reporters and their applications are thus expected to contribute to the studies of a whole range of dynamics of activated Akt in living cells.

To summarize, we have developed a method for visualizing protein phosphorylation–based signal transduction in the living cells, which was herein exemplified for insulin receptor and Akt signaling pathways. The present method is also applicable for certain other protein kinases of interest by changing the substrate sequence, phosphorylation recognition domain, and localization domain. We suggest the use of substrate sequences selectively phosphorylated by kinases of interest to rule out the possibility that such tailor-made reporters are also phosphorylated by other kinases. Our method is advantageous not only for imaging kinase signaling in single live cells and tissues with high spatial and temporal resolution but also for multicell analysis aimed at high-throughput screening to identify pharmaceuticals that regulate kinase and phosphatase activities.

3. Lipid second messengers

Phosphatidylinositol-3,4,5-triphosphate (PIP_3), one of the lipid second messengers, regulates diverse cellular functions including cell proliferation and apoptosis, and any dysregulation in its production can lead to major diseases such as diabetes and cancer (Czech, 2000). This PIP_3 is generated at the cellular membrane by

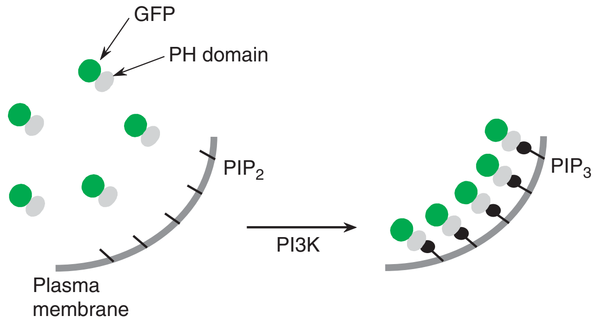


Figure 3 A conventional method for imaging PIP_3 accumulation at the plasma membrane. This reporter is a fusion protein of GFP and a PH domain that specifically recognizes PIP_3

phosphoinositide 3-kinases (PI3Ks), and recruits and activates its binding proteins at the cellular membranes, including Akt, PDK1, Btk, and GRP1 (Cantley, 2002). However, little is known about exactly when, where, and how PIP_3 is produced. Recently, fusion proteins of GFP and PIP_3-binding domains derived from Btk, GRP1, ARNO, or Akt have been reported as reporters for PIP_3 accumulation in the cellular membrane, in which the translocation of the fusion proteins from the cytosol to the membrane has been explained to reflect the PIP_3 accumulation (Varnai *et al*., 1999; Venkateswarlu *et al*., 1998a,b; Watton and Downward, 1999) (Figure 3). However, several factors such as membrane ruffles and changes in the cell shape, which are frequently observed during fluorescence imaging experiments, affect the fluorescence intensity change in the region of interest in a PIP_3-independent manner and cause serious artifacts. Moreover, it is difficult with these fluorescent fusion proteins to distinguish to which membranes the fusion proteins are translocated in the cell. To overcome these limitations, we have developed fluorescent reporters for PIP_3, based on FRET (Sato *et al*., 2003) (Figure 4). These novel PIP_3 reporters are composed of two different-colored mutants of GFP and a PIP_3-binding domain, and the PIP_3 level was observed by dual-emission ratio imaging, thereby allowed stable observation of PIP_3 without suffering from the artifacts described above. To construct the fluorescent reporters for PIP_3, a PH domain of GRP1, which selectively binds with PIP_3, is sandwiched between cyan (CFP) and yellow (YFP) variants of *Aequorea* GFP, via rigid α-helical linkers, consisting of repeated EAAAR sequences (Figure 4). Within one of the rigid linkers, a single diglycine motif is introduced as a hinge. This chimeric reporter protein is tethered with the membrane by fusing with a membrane localization sequence (MLS) via the rigid α-helical linker (Figure 4). When PIP_3 is produced at the membrane upon PI3K activation, the PH domain binds to the PIP_3, and a significant conformational change in the reporter protein quickly occurs around the flexible diglycine motif introduced in the rigid α-helical linker (Figure 4). This flip-flop-type conformational change of the reporter protein causes intramolecular FRET from CFP to YFP, which underlies the detection of PIP_3 dynamics at the cellular membrane (Figure 4). We named this reporter “fllip” (a fluorescent reporter for a lipid second messenger that can be tailor-made) (Sato *et al*., 2003).

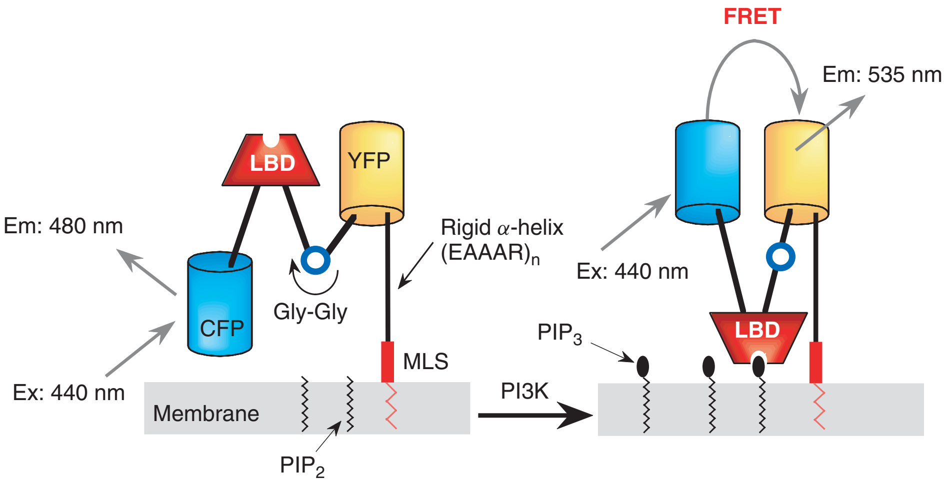


Figure 4 Principle of fllip for visualizing PIP_3. Upon binding of PIP_3 with the PH domain within fllip, a filp-flop-type conformational change of fllip takes place, which changes the efficiency of FRET from CFP to YFP. Abbreviations: LBD: lipid binding domain; MLS: membrane localization sequence

Fllip can be located to particular membrane surfaces of interest by connecting appropriate MLSs, such as lipidation sequences and transmembrane sequences. For example, using the CAAX box motif of N-Ras as the MLS yields a reporter for PIP_3 at the plasma membrane (fllip-pm). Using the mutated CAAX box sequence of N-Ras, in which the cysteine 181 was replaced with a serine, fllip was localized at the endomembranes to monitor the PIP_3 level there (fllip-em).

Using the fllip variants that exhibit each specific subcellular distribution, we revealed spatio-temporal regulations of the PIP_3 production in single living cells. We found that in response to ligand stimuli, PIP_3 was increased to a large extent at the endomembranes rather than at the plasma membrane (Figure 5). In addition, we revealed that this PIP_3 increase at the endomembranes was due to its *in situ* production at the endomembranes triggered by endocytosed receptor tyrosine kinases. The demonstration of PIP_3 production through receptor endocytosis addresses a long-lasting question about how its downstream signaling

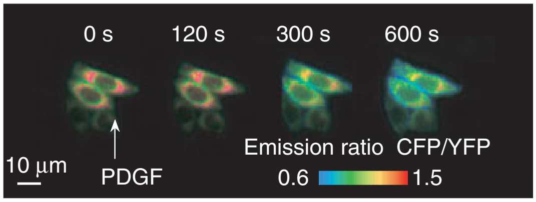


Figure 5 Response of fllip-em to PIP_3 at the endomembranes upon PDGF stimulation. Figure shows pseudocolor images of the CFP/YFP emission ratio before (time 0 s) and at 120, 300, 600 s after the addition of 50 ng mL^{-1} PDGF, obtained from the CHO-PDGFR cells expressing fllip-em. Blue shift in the CFP/YFP emission ratio indicates the production of PIP_3 at the endomembranes

pathways including Akt are activated at intracellular compartments remote from the plasma membrane, such as the Golgi and mitochondria.

From a methodological viewpoint, it should be mentioned that the present fllips, which are genetically encoded fluorescent reporters, have general applicability for other lipid second messengers as well. Fllips have two key sections, the PH domain and MLS, that can be tailor-made. Lipid second messengers other than PIP_3, such as diacylglycerol (Zhang *et al*., 1995) and phosphatidylinositol-3-phosphate (Misra *et al*., 2001), could be selectively detected by using appropriate binding domains for respective lipid messengers instead of the PH domain. Also, by connecting each specific MLS, fllips could be directed not only to the plasma membrane and endomembranes but also to other organelle membranes, such as the inner nuclear membrane and the outer membrane of mitochondria.

References

Cantley LC (2002) The phosphoinositide 3-kinase pathway. *Science*, **296**, 1655–1657.

Chao DT and Korsmeyer SJ (1998) Bcl-2 family: regulators of cell death. *Annual Review of Immunology*, **16**, 395–419.

Czech MP (2000) PIP_2 and PIP_3: complex roles at the cell surface. *Cell*, **100**, 603–606.

Fulton D, Gratton JP and Sessa WC (2001) Post-translational control of endothelial nitric oxide synthase: why isn't calcium/calmodulin enough? *The Journal of Pharmacology and Experimental Therapeutics*, **199**, 818–824.

Hunter T (2000) Signaling – 2000 and beyond. *Cell*, **100**, 113–127.

Kojima H, Nakatsubo N, Kikuchi K, Kawahara S, Kirino Y, Nagoshi H, Hirata Y and Nagano T (1998) Detection and imaging of nitric oxide with novel fluorescent indicators: diaminofluoresceins. *Analytical Chemistry*, **70**, 2446–2453.

Marte BM and Downward J (1997) PKB/Akt: connecting phosphoinositide 3-kinase to cell survival and beyond. *Trends in Biochemical Sciences*, **22**, 355–358.

Misra S, Miller GJ and Hurley JH (2001) Recognition of phosphatidylinositol 3-phosphate. *Cell*, **107**, 559–562.

Miyawaki A and Tsien RY (2000) Monitoring Protein Conformations and interactions by fluorescence resonance energy transfer between mutants of green fluorescent protein. *Methods in Enzymology*, **327**, 472–500.

Mochizuki N, Yamashita S, Kurokawa K, Ohba Y, Nagai T, Miyawaki A and Matsuda M (2001) Spatio-temporal images of growth-factor-induced activation of Ras and Rap1. *Nature*, **411**, 1065–1068.

Sasaki K, Sato M and Umezawa Y (2003) Fluorescent indicators for Akt/protein kinase B and dynamics of Akt activity visualized in living cells. *The Journal of Biological Chemistry*, **278**, 30945–30951.

Sato M, Hida N, Ozawa T and Umezawa Y (2000) Fluorescent indicators for cyclic GMP based on cyclic GMP-dependent protein kinase Iα and green fluorescent proteins. *Analytical Chemistry*, **72**, 5918–5924.

Sato M, Ozawa T, Inukai K, Asano T and Umezawa Y (2002) Fluorescent indicators for imaging protein phosphorylation in single living cells. *Nature Biotechnology*, **20**, 287–294.

Sato M, Ueda Y, Takagi T and Umezawa Y (2003) Production of $PtdInsP_3$ at endomembranes is triggered by receptor endocytosis. *Nature Cell Biology*, **5**, 1016–1022.

Tsien RY (1994) Fluorescence imaging creates a window on the cell. *Chemical and Engineering News*, **18**, 34–44.

Tsien RY (1998) The green fluorescent protein. *Annual Review of Biochemistry*, **67**, 509–544.

Varnai P, Rother KI and Balla T (1999) Phosphatidylinositol 3-kinase-dependent membrane association of the Bruton's tyrosine kinase pleckstrin homology domain visualized in single living cells. *The Journal of Biological Chemistry*, **274**, 10983–10989.

Venkateswarlu K, Gunn-Moore F, Tavaré JM and Cullen PJ (1998a) Nerve growth factor- and epidermal growth factor-stimulated translocation of the ADP ribosylation factor-exchange factor GRP1 to the plasma membrane of PC12 cells requires phosphatidylinositol 3-kinase and the GRP1 pleckstrin homology domain. *The Biochemical Journal*, **335**, 139–146.

Venkateswarlu K, Oatey PB, Tavaré JM and Cullen PJ (1998b) Insulin-dependent translocation of ARNO to the plasma membrane of adipocytes requires phosphatidylinositol 3-kinase. *Current Biology*, **8**, 463–466.

Walkup GK and Imperiali B (1997) Fluorescent chemosensors for divalent zinc based on zinc finger domains. Enhanced oxidative stability, metal binding affinity, and structual and functional characterization. *Journal of the American Chemical Society*, **119**, 3443–3450.

Watton J and Downward J (1999) Akt/PKB localisation and 3′ phosphoinositide generation at sites of epithelial cell-matrix and cell-cell interaction. *Current Biology*, **9**, 433–436.

Zaccolo M, De Giorgi F, Cho CY, Feng L, Knapp T, Negulescu PA, Taylor SS, Tsien RY and Pozzan T (1999) A genetically encoded fluorescent indicator for cyclic AMP in living cells. *Nature Cell Biology*, **2**, 25–29.

Zhang G, Kazanietz MG, Blumberg PM and Hurley JH (1995) Crystal structure of the Cys2 activator-binding domain of protein kinase C delta in complex with phorbol ester. *Cell*, **81**, 917–924.

Specialist Review

Quantitative EM techniques

John M. Lucocq
School of Life Sciences, University of Dundee, Dundee, UK

1. Introduction

This article examines strategies for quantification from sectioned biological material at the electron microscope (EM) level. Recently there has been increasing demand for rigorous quantification and for linking morphological results with physiological and biochemical data. To achieve this it has been essential to develop a body of methods that allows accurate, unbiased, and efficient estimation of thin section and 3D quantities. This article describes current approaches to quantitation of both structure and immunolabeling. Sectioning is a powerful tool for revealing the internal structure of cells and tissue. At the ultrastructural level it provides samples that are thin enough for clear display of internal cell components and presents the opportunity to sample 3D specimens appropriately. Here the focus is exclusively on quantification on thin sections of biological material.

Structural display is the sine qua non of good quantitative EM because one cannot quantify what cannot be seen. Therefore optimal specimen preparation and structure contrasting lie at the heart of all successful quantification regimes; however, these lie outside the scope of this article (see Griffiths, 1993; Liou *et al*., 1996). Methods for molecular contrasting at the EM are also now well established and these range from the less popular preembedding approaches, which use either particulate or enzyme-based methods producing electron dense enzyme reactions to the more widely used on-section labeling that utilizes particulate markers. Particulate markers have the inherent advantage of being countable and have been proved to yield a signal that can report on antigen concentration. Currently the most important particular marker system in widespread use is the colloidal gold (Lucocq, 1993a) but other methods such as quantum dots (Giepmans *et al*., 2005) are showing promise of improved labeling quality.

2. The basics of sampling

Using sections to provide readouts of structure or immunolabel introduces the problem of sampling (Lucocq, 1993b). All sections, whether they are immunolabeled or used for structural quantitation, are small samples ultimately derived from an animal/experiment, culture, organ, tissue, cell, or organelle. It is

important that the information contained in the sections used is representative of the "experimental universe" from which they are drawn, whether this is a set of animals or cell cultures. To ensure this representation is fair, sampling must be strictly nonselective and so simple uniform random sampling is the mainstay of all EM quantitative approaches. In fact this is a minimum requirement at all levels of the sampling hierarchy (animals/experiments, dishes, blocks, sections, EM grids, EM grid holes; see Figure 1). In heterogeneous biological systems simple random sampling may not always be the most efficient strategy and can be modified to improve the accuracy of results obtained from doing a particular amount of work. The best-known modification is *systematic random sampling*, in which a random start point ensures unbiasedness, and subsequent samples are spaced in a pattern at regular intervals through the organ, specimen, block, or section (see Figure 1). When biological samples are heterogeneous, systematic random sampling is usually more efficient, partly because it spreads-out "measurements" to sense local variations. By comparison, simple random samples may tend to cluster or become dispersed and thereby over- and underrepresent certain regions. Systematic random sampling is also simpler to set up than simple random sampling. As we shall see, depending on the intended readout it is almost always necessary to randomize the position of the section within the specimen of interest; for some parameters, it may also be important to randomize orientation as well.

3. Immunolabel quantitation

Because labeling systems are applied to sections that produce essentially 2D images of structure, this represents the simplest system to illustrate principles of quantification. After the introduction of the particulate markers, many studies used purely qualitative assessment of labeling. However, assessing small changes in labeling intensity or amounts of dispersed label is difficult by eye and comparisons between conditions, or between different compartments, demanded unbiased quantitative measures. Sectioning opens up cellular compartments and makes them available to the immunogold labeling reagents that are applied to the sections. The random placement of sections at the lower end of a random sampling-based hierarchy therefore ensures that the components of compartments/structures in the animal/cell line have similar probability of encountering the reagents. This is an important advantage over permeabilization techniques in which penetration of the reagents is at the mercy of both structural variables and permeabilization conditions that are difficult to control (Griffiths, 1993).

In an ideal world there would be a direct and proportionate relationship between particle labeling and protein antigens irrespective of their location in the cell, but in reality the relationship is influenced by the structural context in which the antigen is sitting and other factors such as modification of the antigen or the characteristics of the antibody probes. Thus, the number of particles labeling each protein (often termed the *labeling efficiency*) can vary from compartment to compartment or from structure to structure (Lucocq, 1994; Griffiths and Hoppeler, 1986). Important work from the Utrecht group (e.g., Posthuma *et al*., 1988) determined that variation is due to different penetration of labeling reagents and sought to eliminate these

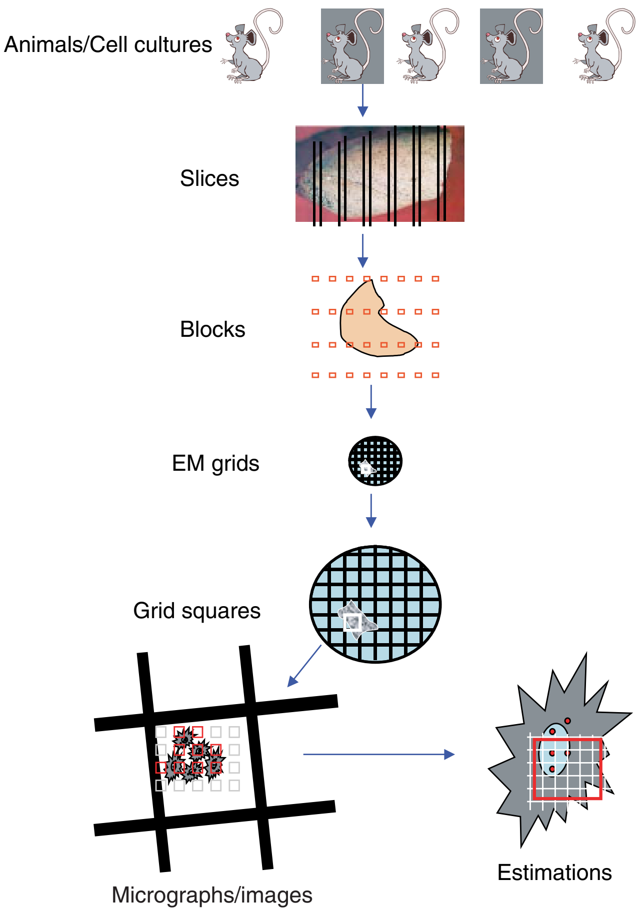


Figure 1 Sampling hierarchy. Quantitative estimations using electron microscopy are carried out on microscope viewing-screen images or micrographs of thin-sectioned material. These represent very small samples of the organ/tissue/cells from individual animals or cell cultures. The problem faced is how to sample slices, blocks, EM grids, EM grid holes, and micrographs, so that the estimations contain fair estimates of the parameters required. An absolute necessity is random selection of the items at each level of the sampling hierarchy. Increases in efficiency can be obtained from a systematic array of samples that is applied randomly. In the case shown here the organ has been systematically sampled first with slices and then as EM blocks. Sections from these blocks are sampled at a randomly selected EM grid hole and systematically spaced micrograph sections taken. Finally a randomly positioned geometric probe with systematic lattice structure is applied to each micrograph (see text for details and discussion of cases in which randomization of orientation is also a requirement). In general, studies have found that higher levels of the hierarchy are the major contributors to the overall variance and it is advisable to study a minimum of three and preferably five animals/cultures/experiments. Experience has shown that 10–20 micrographs and 100–200 events at the estimation stage may be sufficient to reduce the contribution from this level to a minimum. At the intermediate levels of the hierarachy, the number required will be determined by the heterogeneity of the specimen

differences by embedding fixed cells or tissues in either polyacrylamide or Lowicryl resins. As things stand, most workers accept the underlying variation in labeling efficiency between compartments. It is worth noting also that labeling efficiency over the same compartment take out may remain constant from experiment to experiment so that labeling intensity may correlate rather well with local changes in the concentration of antigen.

3.1. Estimation of labeling (antigen) distribution

Estimating the distribution of particulate labeling is of fundamental importance in quantitative EM. Mapping studies enable informed judgements about the antigen distribution and are a good way of identifying the compartments or structures that contain pools of antigen. It is important to realize that while the distribution of labeling over a number of compartments/structures may be relatively insensitive to overall changes in the intensity of the labeling signal, it may be biased by focal changes in labeling intensities in ways that are difficult to interpret (see discussion of labeling density below).

Strategies for estimation of labeling distribution depend on obtaining a systematic sample from a representative labelled section. As described above, random sampling (simple or systematic) of animals/experiments, dishes, and blocks with random positioning of the section within the blocks ensures fair representation of different elements in the sampling hierarchy (Figure 1; note that the issue of how section orientation affects labeling efficiency has not been addressed systematically and it may be pertinent to include measures that randomize orientation as well as position in labeling studies (see below)). When thin sections are mounted on EM grids they may be of variable quality or may be so extensive as to preclude sampling the whole section area. Under these conditions smaller regions of the section can be selected by random (or systematic random) selection of EM grid holes containing sections of interest. It is worth reiterating the adage that one cannot quantify what cannot be seen and so adequate contrasting of sections of the appropriate thickness to reveal the structure of interest is important.

The most efficient way of sampling a selected labeled section (or more limited area of the section) is to obtain a systematic random sample and this can be achieved in a number of ways. In one approach, the selected section area can be scanned in strips at the electron microscope (Figure 2a), at a magnification at which the particulate marker and the structures of interest can be identified. Initially the magnification selected should be the lowest that allows both structures/compartments and particulate marker to be clearly seen as this maximizes the sample size. The strips are defined by translocation of the section under the screen and 10–20 of them are systematically spaced (with a random start) and cover the whole of the selected area (see Lucocq *et al*., 2004). A second approach is to take micrographs at systematic random locations across the selected area of the section (Figure 2b). Again the magnification used should be the minimum that allows both particles and structure to be visualized. The number of compartments assessed would depend on the underlying biology and

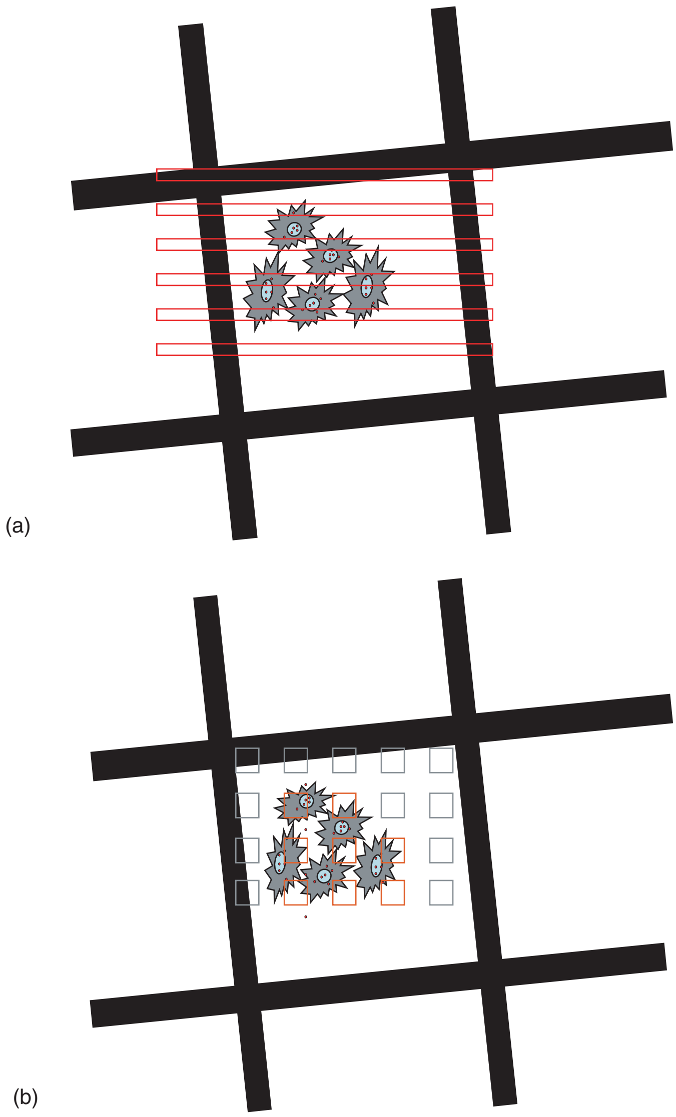


Figure 2 Labeling distribution. Systematic sampling of particulate immunolabeling can be achieved using all available sections on an EM grid. But usually an area of interest such as an EM grid hole (or holes) containing optimally contrasted section(s) is selected at random. In the scanning method (a) this area of interest is translocated under the viewing screen to create systematically spaced strips– initially at a magnification sufficient to visualize both the labeling and structures of interest. The uppermost corner of the grid hole can be used as a randomly placed marker to position the scans. In the micrograph method, (b) images (containing relevant cell compartments/labeling) are recorded at systematically spaced locations, again using the uppermost corner of the EM grid hole as a marker of random position

the goals of the experiment (see Mayhew *et al*., 2002; Griffiths *et al*., 2001). Particles are counted and assigned to compartments during the scanning procedure and approximately 200 particles per grid are enough to describe the distribution over 10–15 compartments (Lucocq *et al*., 2004). If a more precise value of the labeling fraction in one particular compartment is required, then 100–200 particles should be examined over that individual compartment (Lucocq *et al*., 2004). This level of sampling has also been found to be sufficient in many stereological studies. With these guidelines in mind it may be pertinent at some stage of a study to increase or decrease the total number of sampled particles. Increases in particle numbers counted may be achieved simply by increasing sample size (scan or micrographs), while decreases follow from increasing magnification used in the scanning approach or by decreasing the quadrat size on micrographs. It is generally advisable to apply unbiased counting rules to the scanning strips or to quadrats used for counting on micrographs, especially when the size of the gold particles becomes significant relative to the quadrat (>100th of the quadrat size; Lucocq, 1994).

Recently, powerful methods for comparison of the raw particle count distribution between groups have been introduced (Mayhew *et al*., 2004). This type of analysis is facilitated by the use of contingency table analysis with statistical degrees of freedom for chi-squared values being determined by the number of compartment and number of experimental groups of cells (see Mayhew *et al*., 2004 for discussion and details). The method enables identification of compartments/structures in which the major between-group differences reside.

3.2. Estimation of labeling density

The local intensity of particulate immunolabeling reflects the local concentration of a component antigen (Griffiths, 1993) and is therefore a biochemically relevant quantity for that compartment/structure. The labeling intensity can be found by relating label to the size of a structure/compartment, and in the classical approach, labeling density is related to the absolute size of profiles displayed on the ultrathin sections. Profile sizes can be estimated by interactions of geometrical probes such as points or lines as systematic test grids (see Figure 3a). Points are used to estimate organelle area and lines to estimate profile length (usually membranes) (see Griffiths, 1993; Lucocq, 1994). With the number of probe hits, the spacing of the probes in the grid and magnification, a simple formula converts the counts to areas or lengths from which the absolute density values can be estimated (see Figure 3). Notice that to ensure unbiasedness of the estimates, the grid of geometric probes is applied as a systematic series of points or lines. For unbiased estimations, there is a requirement for random placement of points (for area estimation) or random placement *and* orientation of lines (for profile length estimation); and a course for random sampling throughout the hierarchy (as yet the need for random orientation in 3D has not been systematically investigated; see discussion below for surface estimation). Traditionally when the grid is applied to micrographs/digital camera images, the density is reported per square micron or density per micron length of membrane, thereby allowing

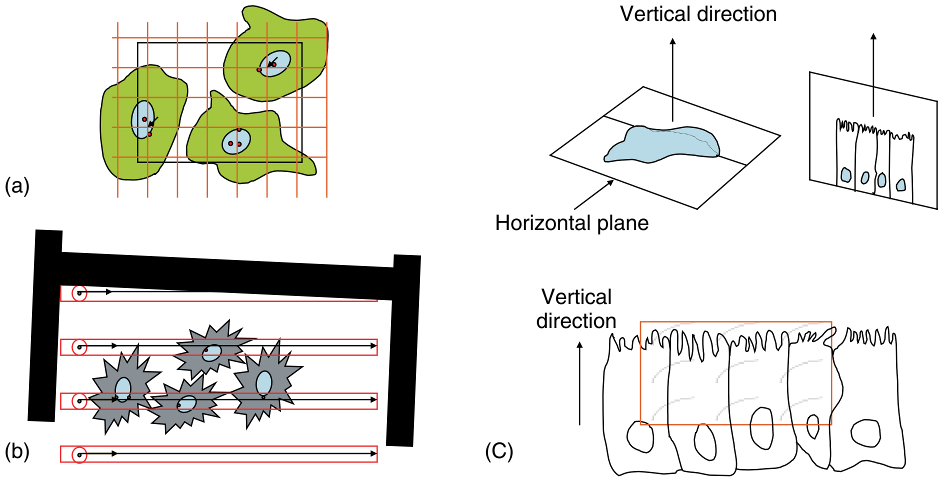


Figure 3 Labeling density, volume, and surface density. Labeling density can be expressed as labeling per unit profile area or per unit length of profile. As shown in (a), the area can be estimated using point hits over the structure. The points can be defined as the corners of lattice squares and the profile area, A, can be computed from the product of area associated with each point on the lattice (a/p) and the total number of point hits (ΣP) on the stucture. Here, there are two point hits (arrows) over the nucleoplasm and the number of labeling particles is two. Random placement of the lattice relative to the specimen is a requirement. The profile length of nuclear envelope (here represented by a single membrane trace) can be estimated from the number of intersections (I) of the envelope with lattice lines (in this case 9), which should be randomly oriented and positioned relative to the membrane profiles. Trace length is computed using $(4/\pi)Id$ where d is the real distance between the lines ($(2/\pi)Id$ is the formula if only one set of parallel lines are used). Assignment of gold particles to membranes (in this case 3) is decided on the basis of acceptance zones within which dispersion of label is considered to be possible – for example, two particle widths for 10-nm diameter particles (se Griffiths, 1989a). As described in the text, labeling may be related to point counts or intersections without conversion to absolute units. Relating labeling to intersection counts can also be achieved by an adaptation of the scanning approach in which gold particle counts are related to intersections of structures with a scanning line defined as a feature located in the EM viewing screen moves relative to the section (b).

The systematic lattice illustrated in (a) can also be used to estimate volume and surface density in 3D stereologically. For unbiased estimation of volume density, the lattice is randomly placed inside the specimen by ensuring random selection/placement at all levels in the hierarchy. The fraction of points that fall over the structure of interest relative to a reference space, reports on the volume density of the structure in the reference space.

Intersection of structures with lattice lines (I) enables estimation of the surface density using the formula $2I/L$ where L is the total line length applied to the specimen. Either the lines or the specimen must be randomly placed *and* oriented in space for unbiased estimations (see text for some details on generating isotropic lines). (c) Illustrates one method of generating isotropic uniform random lines in space. A recognizable horizontal plane (in this case the base of a columnar cell sheet) enables a vertical direction to be identified. A section along the vertical direction is taken in uniform random position and at random angle of rotation around the vertical axis. On this section plane, isotropic lines can be represented by sine weighted lines, for example, cycloid arcs usually applied as a systematic lattice (test system in red quadrat of figure). Intersections of linear profiles with these test lines allow surface density to be estimated according to the formula $2I/L$ (see Baddeley *et al*., 1987; Howard and Reed, 1997 for details)

interpretation and comparison with other studies. This type of density information can also be used in combination with section thickness measurements to estimate density of labeling over structures in 3D (see, e.g., Griffiths and Hoppeler, 1984).

More recently, for certain comparative purposes, particulate labeling has been related to point or intersection counts without "converting" to absolute area or

length units (see Figure 3). Density can be expressed as gold particles per point or intersection (Watt *et al*., 2002; 2004; Mayhew *et al*., 2002) and this index of labeling intensity has been used to test for preferential labeling over individual or groups of compartments. In one approach, the labeling densities are compared to those obtained if the existing gold labeling were randomly spread over all compartments to provide a relative labeling index (RLI; see Mayhew *et al*., 1996, 1999). The intensity data can be obtained on micrographs or from scanning in strips at the electron microscope using the viewing screen in which a feature acts to trace a scanning line for intersection counting (see Figure 3b). This type of approach cannot at present be applied to both organelle and membrane surfaces simultaneously.

4. Structure quantification

Analysis of structural parameters in 3D from sections requires specifically designed stereology that employs sampling with geometrical probes. The geometrical probes are generally applied to sections as systematic arrays of points, lines, planes, or brick-shaped volumes with strict rules for multistage random sampling animals, cells, blocks, sections, and micrographs. Simple formulae convert the counts into 3D quantities. As a rule of thumb the dimensions of the probe and the parameter to be estimated must add up to at least three, for example, points (0D) can only be used for volume estimation (3D) and lines (1D) but not points (0D) may be used to estimate surface (2D).

4.1. Volume fraction and surface density

Volume fraction and surface density are simply measures of the concentration of volume or surface inside the volume of some reference structure (see Figure 3a; for reasons of clarity and simplicity, length density estimation is not dealt with here (see Howard and Reed, 1998 for details). Volume fraction can be estimated using an array of points that are placed at a random location within a specimen. The fraction of points that fall on a structure relative to a reference space/structure is an estimate of the fraction occupied by the structure of interest. This is an extension from the principle outlined by Delesse (1847); see also Thompson, 1930), which states that the fractional area displayed by a structure reports on the fraction of the reference space occupied by that structure. The random/systematic random positioning of the points on the section again depends on correct sampling throughout the hierarchy and is employed down to the level of the section and the systematic array probes applied as a grid to the micrograph (Figure 3a). The orientation of the specimen can be arbitrary. Importantly, the readout is a ratio and is therefore sensitive to changes in volume of both the structure and the reference space. It is therefore advisable to have information on the size of the reference space (see below), or at least its stability. Once the reference space volume is known, then the absolute volume of the structure inside that space can be computed easily from the product of the volume fraction and the reference space volume. Volume fraction has been

one of the most used estimators in electron microscopy but is less often combined with reference space volume estimation.

Surface density is usually estimated in a volume. In cell biology, this might represent the packing density of nuclear membrane in the cytoplasm or mitochondrial cristae within mitochondria. Again caution is required in interpreting the readout ratio and knowledge of the reference space size enables absolute values for surface to be estimated from the product of surface density and reference volume. The estimates are most often made using systematic arrays of line probes that are randomly positioned and randomly oriented in 3D space, relative to the specimen (Figure 3b). A number of strategies have been developed to ensure randomness of orientation (isotropy) of probe or specimen. These include embedding the specimen in an isotropic sphere, which is rolled before embedment (the disector; Nyengaard and Gundersen, 1992); dicing tissue/cell pellets and allowing them to settle haphazardly (which is less rigorous; Stringer *et al*., 1982); and making two successive positioned oriented sections through a structure (the orientator; Mattfeldt *et al*., 1990). Finally, a powerful way to generate isotropic lines in space is to orient sections in a fixed direction vertical to an identifiable horizontal plane associated with the object, and then generate sine weighted isotropic lines on the sections (vertical section method; Baddeley *et al*., 1986; Michel and Cruz-Orive, 1988; see Figure 3c). This elegant solution allows orientation of sections along specific directions that display features of cell organization such as polarity and is well suited to surface estimation in cell monolayers sectioned in a direction perpendicular to the culture dish. Irrespective of how the isotropy is generated, the number of intersections of lines with the surface features of interest can be easily converted into a surface density value. The product of surface density and reference volume will then provide an estimate of the total surface (see Figure 3 legend and Griffiths *et al*., 1984, 1989a,b for examples).

For both volume and surface density estimations, errors may arise when structures are small compared to the section (e.g. 50–200 nm). Peripheral grazing sections result in lost caps that decrease the number of point or line encounters and underestimate surface or volume density. On the other hand as structures become small relative to the section, they overproject their profiles into the final image and lead to overestimation of size densities. Correction factors based on model shapes for the small structures are available but these should be used under carefully controlled conditions in which the models used correspond closely to the morphology of the structure in question (Weibel and Paumgartner, 1978; Weibel, 1979).

4.2. *Reference volume*

The volume of organs, tissue components, cells, or organelles, are of interest in their own right but they also represent important reference spaces in which stereological densities can be converted into absolute values.

One of the most powerful methods for volume estimation is that devised by the seventeenth century priest Bonaventura Cavalieri who was a disciple of Galileo

(Cavalieri, 1635). This approach is well suited to EM slices that are relatively thin compared to the object (a similar principle can also be applied using sections that are relatively thick Gual-Arnau and Cruz-Orive, 1998). A series of sections through the entire object are prepared at a regular spacing with the first slice lying at a uniform random start location within the section interval. The area of the object profiles as displayed on these sections is estimated by counting point hits of a lattice grid overlaid on the object profile (Gundersen and Jensen, 1987). A simple formula converts the point counts summed over all sections into the estimated object volume (Figure 4a). Importantly, 5–10 sections are generally enough to obtain precise estimates of objects irrespective of their orientation and shape. The precision of volume estimate obtained using the Cavalieri method can be assessed as the coefficient of error (Gundersen and Jensen, 1987), which decreases as the number of sections (sampling intensity) is increased. The Cavalieri method

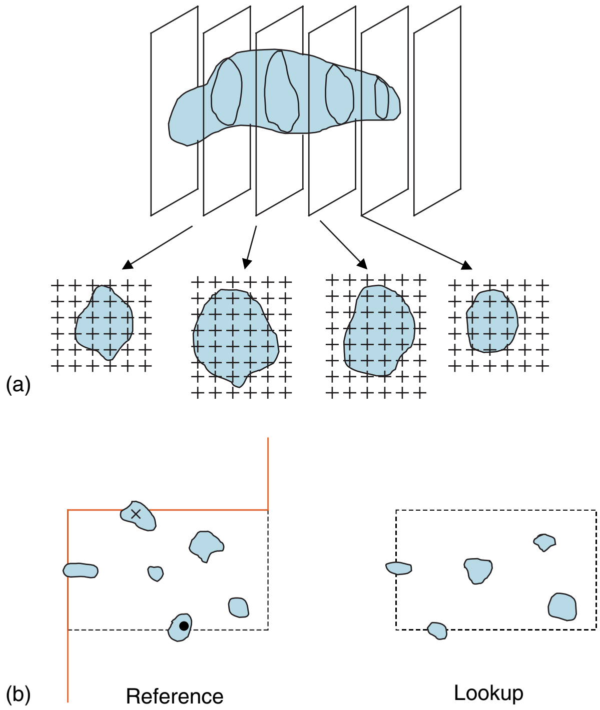


Figure 4 Volume and particle number. (a) Volume estimation using the Cavalieri sections method in EM. A systematic sample of equally spaced sections (usually 5–10) is placed at a random start location. The total profile area (A) is the product of the area associated with each point on the lattice (a/p) and the total number of point hits on the profiles (ΣP). The distance between the sections (k) is obtained from the product of the section thickness (t) and the number of sections in each interval (n_s). The volume estimate is computed from $A \times k$. (b) Principle of using a disector to count particles. In the reference section, particle profiles are selected if they are enclosed entirely in the quadrat or crossed by the dotted acceptance line. Profiles that encounter the continuous forbidden line are ignored. Selected particles that disappear in the lookup (·) section are counted. Although the particle (x) has no profile in the lookup it is not counted because its profile hits the forbidden line. See Gundersen (1988) for further details

is now widely used in light microscopy and radiology (Roberts *et al*., 2000) and in electron microscopy it is an excellent method for cell or organelle size estimation that can be combined with other estimators (Lucocq *et al*., 1989; McCullough and Lucocq, 2005). Other methods for estimating particle volume are outlined below.

4.3. Counting and sampling in 3D

The problem with sampling or counting items from 3D space with sections of cells or tissues is that the probability that a structure appears in a section is related to its size (actually its height in the direction of sectioning) and not its number. For many years quantitative biologists attempted to circumvent the problem by assuming model shapes, thereby enabling conversion of profile size distributions into particle numbers. Such model-based approaches (see Weibel, 1979 for discussion) can be inherently biased and the breakthrough in number estimation and sampling came in 1984 when Sterio reported using a two-section approach to identify and count particles in 3D space (Sterio, 1984; Mayhew and Gundersen, 1996). The essence of this approach is to sample profiles of particles displayed in a section using two-dimensional unbiased counting rules applied to a quadrat (Figure 4b) and to assess which of the selected particles are then *not* displayed in a second parallel section. This amounts to counting particle tops or edges and is particularly suited if the particles are entirely convex and therefore do not contain multiple edges. For vesicles or spheroidal or tubular structures such as nuclei or particulate mitochondria, this is not a problem. However for more complex organelles such as Golgi apparatus or interconnected mitochondria, the number of edges may be multiple and varied. The union of the sampling section (and its quadrat) with the second section (the lookup) is termed a *disector*. If disectors are placed at random or systematic random through the specimen, then the probability that a particle (nucleus, vesicle, etc.) is selected in this way is related only to its number. Once selected, this sampled structure can be further characterized. Implementation of disectors at the EM level demands accurate positioning of fields for sampling and lookup at appropriate magnifications. Important questions arise as to distance between sections in the disectors, relative to the particles in the population. Clearly the distance must be close enough to avoid missing any particles between the sections; but the method is based on comparing the profiles in sampling section and the lookup section – so it is advisable to keep the sections close enough to enable particles sectioned by both sections to be used as a reference. As a general guide disectors about one-third the height of the average particle are usually sufficient. Since the preparation of disectors for use in EM is fairly labor intensive, then each section in the disector can be used alternately as the reference and lookup to improve efficiency (see Smythe *et al*., 1989)

4.3.1. Uses of disectors

One straightforward readout from disectors is the particle density in the volume of any reference space that can be identified in the disector. The volume of this

reference space can be estimated using point counting applied to the structures displayed in the quadrat used for sampling. Interestingly the reciprocal of the particle density in particle volume turns out to be the mean number-weighted particle volume (Gundersen, 1986). The volume of particles can be also assessed using Cavalieri estimator although this requires further examination of a section stack through the entire sampled particle. At the EM level disectors have been used to sample and count coated endocytic structures (Smythe *et al*., 1989) and to characterize mitotic Golgi fragments (Lucocq *et al*., 1989).

An elegant application of the disector can be used to estimate the numerical ratio of a small structure to a larger one (the number of peroxisomes per cell nucleus, e.g., Gundersen, 1986). In a stack of sections one pair of sections is used to sample the larger structure and estimate its numerical density in a reference space. At a random location within this disector interval, a smaller disector, composed of two more closely spaced sections, is then used to sample the smaller structure and the density estimated. The ratio of the two densities reports the numerical ratio of the two structures and since the estimates are obtained using the same average section thickness, then no section thickness measurement is required and no determination of the total reference volume is required. Double disectors have been used in electron microscopy to estimate the number of coated pits per cell (Smythe *et al*., 1989), the number of Golgi fragments/vesicles per cell (Lucocq *et al*., 1989), the number of synapses (Jastrow *et al*., 1997; Mayhew 1996) and the number of gold particles per cell (Lucocq, 1992).

In the case of small structures such as vesicles there is a special case in which all structures sectioned by the ultrathin section do not have a profile in the next section. Thus the probability of finding a disappearing profile is 1. Under these conditions there may be no need for a lookup section to be used. This convenient adaptation of the disector must be used with caution and preliminary work is needed to determine the proportion of structures that are detected by one section but disappear in the next. This approach has been used to count Golgi vesicles in dividing HeLa cells (Lucocq *et al*., 1989).

4.4. Particle volume

As already discussed, the volume of structures sampled using disectors can be estimated either by using Cavalieri sections method or from the reciprocal of the particle density in the disector. However there are other methods for particle volume estimation. Unlike those just mentioned, they all require some element of isotropy. The selector (Cruz-Orive, 1987) samples particles with disectors and uses a systematic stack of sections through the particle to place sampling points within the particle volume. From these points line probes are used to estimate the volume of the particle. The distance between the sections need not be known but the orientation of the line probes must be isotropic. A refinement of the selector principle is the nucleator (Gundersen *et al*., 1988), which uses a central structure such as a nucleolus that can be sampled with constant probability as a source for the isotropic lines (Henrique *et al*., 2001). This tends to be more efficient

than the selector and uses fewer sections. Another powerful method is the rotator, which takes advantage of the Pappus theorem. This states that the product of the distance traveled by the center of gravity of a figure and its area can be used to estimate its total volume in rotation (Jensen and Gundersen, 1993). This gives the possibility of estimating the volume of an object when sections lie along a defined axis but are allowed to randomly assume any angle of rotation around this axis. This approach has been used at the EM level to estimate cell volume around the centriolar structures and the volume of the endoplasmic reticulum at different locations within mitotic HeLa cells (McCullough and Lucocq, 2005; Mironov and Mironov, 1998). Finally it is also worth mentioning the point sample intercept method (Gundersen and Jensen, 1985; Gundersen *et al*., 1988), which allows estimation of particle volume on single sections. Points are used as probes to probe particle volume and isotropic lines through these points can be used to estimate the particle volume, but this is weighted according to the volume and not number of particles.

5. 2D and 3D spatial distributions

Quantitation of total or aggregate quantities have been the main focus for EM quantitation and are often referred to as *first-order stereology*. However the statistical study of spatial distributions is of biological value in both 2D and in 3D and has been referred to as second-order stereology.

Statistical approaches for assessing the distribution of point processes in 2D have been presented by Diggle (1983) and Ripley (1981). The first-order property of such a point pattern is its intensity, expressed as number of items per unit area, whereas the second-order property is characterized, for example, by Ripley's K function, which is sensitive to the distribution of interparticle distances which can be compared to an expected number of neighbors within a distance of any point. Membrane lawns and freeze fracture techniques present 2D surfaces that can be labeled with particulate markers and the distribution of the labeling provides clues to clustering or receptors, lipids, and downstream signaling molecules. A recent application of these principles has been used to examine protein distribution at the plasma membrane (Prior *et al*., 2003). Development of 3D spatial analysis is in its infancy but the use of linear dipole probes has shown some promise in assessing the spatial distribution of volume features at the EM level (Mayhew, 1999; see also Reed and Howard, 1999). Molecular or subcellular organelle distributions have yet to be examined using this methodology.

Finally, this review covers only some of the basic principles and the reader is referred to more extensive basic texts (such as Howard and Reed, 1998) and to the International Society for Stereology (ISS) website (**http://www.stereologysociety.org/**) for further details. Advice from those experienced in this field can be invaluable to anyone embarking on a quantitative EM study.

References

Baddeley AJ, Gundersen HJ and Cruz-Orive LM (1986) Estimation of surface area from vertical sections. *Journal of Microscopy*, **142**, 259–276.

Cavalieri B (1635) *Geometria Indivisibilibus Continuorum. Typis Clemetis Feronij Bononi. Reprinted 1966 as Geometria degli Indivisibili*, Unione Tipografico-Editrice Torinese: Torino.

Cruz-Orive LM (1987) Particle number can be estimated using a disector of unknown thickness: the selector. *Journal of Microscopy*, **145**, 121–142.

Delesse MA (1847) Procédé mécanique pour déterminer la composition des roches. *Comptes Rendus de l'Academie des Sciences Paris*, **25**, 544–545.

Diggle PJ (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press: New York.

Giepmans BN, Deerinck TJ, Smarr BL, Jones YZ and Ellisman MH (2005) Correlated light and electron microscopic imaging of multiple endogenous proteins using quantum dots. *Nature Methods*, **2**, 743–749.

Griffiths G (1993) In *Fine Structure Immunocytochemistry*, Griffiths G (Ed.), Springer-Verlag: Berlin.

Griffiths G, Back R and Marsh M (1989a) A quantitative analysis of the endocytic pathway in baby hamster kidney cells. *The Journal of Cell Biology*, **109**, 2703–2720.

Griffiths G and Hoppeler H (1986) Quantitation in immunocytochemistry: correlation of immunogold labeling to absolute number of membrane antigens. *The Journal of Histochemistry and Cytochemistry*, **34**, 1389–1398.

Griffiths G, Fuller SD, Back R, Hollinshead M, Pfeiffer S and Simons K (1989b) The dynamic nature of the Golgi complex. *The Journal of Cell Biology*, **108**, 277–297.

Griffiths G, Lucocq JM and Mayhew TM (2001) Electron microscopy applications for quantitative cellular microbiology. *Cellular Microbiology*, **3**, 659–668.

Griffiths G, Warren G, Quinn P, Mathieu-Costello O and Hoppeler H (1984) Density of newly synthesized plasma membrane proteins in intracellular membranes. I. Stereological studies. *The Journal of Cell Biology*, **98**, 2133–2141.

Gual-Arnau X and Cruz-Orive L (1998) Variance prediction under systematic sampling with geometric probes. *Advances in Applied Probability*, **30**, 1–15.

Gundersen HJ (1986) Stereology of arbitrary particles. A review of unbiased number and size estimators and the presentation of some new ones, in memory of William R. Thompson. *Journal of Microscopy*, **143**, 3–45.

Gundersen HJ, Bagger P, Bendtsen TF, Evans SM, Korbo L, Marcussen N, Moller A, Nielsen K, Nyengaard JR, Pakkenberg B, *et al*. (1988) The new stereological tools: disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis. *APMIS*, **96**, 857–881.

Gundersen HJ and Jensen EB (1985) Stereological estimation of the volume-weighted mean volume of arbitrary particles observed on random sections. *Journal of Microscopy*, **138**, 127–142.

Gundersen HJG and Jensen EB (1987) The efficiency of stereological sampling in stereology and its prediction. *Journal of Microscopy*, **147**, 229–263.

Henrique RM, Rocha E, Reis A, Marcos R, Oliveira MH, Silva MW and Monteiro RA (2001) Age-related changes in rat cerebellar basket cells: a quantitative study using unbiased stereological methods. *Journal of Anatomy*, **198**, 727–736.

Howard CV and Reed MG (1998) *Unbiased Stereology. Three Dimensional Measurement in Microscopy*, BIOS: Oxford.

Jastrow H, Von Mach MA and Vollrath L (1997) Adaptation of the disector method to rare small organelles in TEM sections exemplified by counting synaptic bodies in the rat pineal gland. *Journal of Anatomy*, **191**, 399–405.

Jensen EB and Gundersen HJG (1993) The rotator. *Journal of Microscopy*, **170**, 35–44.

Liou W, Geuze HJ and Slot JW (1996) Improving structural integrity of cryosections for immunogold labeling. *Histochemistry and Cell Biology*, **106**, 41–58.

Lucocq J (1992) Quantitation of gold labeling and estimation of labeling efficiency with a stereological counting method. *The Journal of Histochemistry and Cytochemistry*, **40**, 1929–1936.

Lucocq JM (1993a). Particulate markers for immunoelectron microscopy. In *Fine structure immunocytochemistry*, Chap. 8, Griffiths G Springer-Verlag: Berlin. 1992. pp. 279–302.

Lucocq J (1993b) Unbiased 3-D quantitation of ultrastructure in cell biology. *Trends in Cell Biology*, **3**, 345–358.

Lucocq J (1994) Quantitation of gold labelling and antigens in immunolabelled ultrathin sections. *Journal of Anatomy*, **184**, 1–13.

Lucocq JM, Berger EG and Warren G (1989) Mitotic Golgi fragments in HeLa cells and their role in the reassembly pathway. *The Journal of Cell Biology*, **109**, 463–474.

Lucocq JM, Habermann A, Watt S, Backer JM, Mayhew TM and Griffiths G (2004) A rapid method for assessing the distribution of gold labeling on thin sections. *The Journal of Histochemistry and Cytochemistry*, **52**, 991–1000.

Mattfeldt T, Mall G, Gharehbaghi H and Moller P (1990) Estimation of surface area and length with the orientator. *Journal of Microscopy*, **159**, 301–317.

Mayhew TM (1996) How to count synapses unbiasedly and efficiently at the ultrastructural level: proposal for a standard sampling and counting protocol. *Neurocytology*, **25**, 793–804.

Mayhew TM (1999) Second-order stereology and ultrastructural examination of the spatial arrangements of tissue compartments within glomeruli of normal and diabetic kidneys. *Journal of Microscopy*, **195**, 87–95.

Mayhew T, Griffiths G, Habermann A, Lucocq J, Emre N and Webster P (2003) A simpler way of comparing the labelling densities of cellular compartments illustrated using data from VPARP and LAMP-1 immunogold labelling experiments. *Histochemistry and Cell Biology*, **119**, 333–341.

Mayhew TM, Griffiths G and Lucocq JM (2004) Applications of an efficient method for comparing immunogold labelling patterns in the same sets of compartments in different groups of cells. *Histochemistry and Cell Biology*, **122**, 171–177.

Mayhew TM and Gundersen HJ (1996) If you assume, you can make an ass out of u and me': a decade of the disector for stereological counting of particles in 3D space. *Journal of Anatomy*, **188**, 1–15.

Mayhew TM, Lucocq JM and Griffiths G (2002) Relative labelling index: a novel stereological approach to test for non-random immunogold labelling of organelles and membranes on transmission electron microscopy thin sections. *Journal of Microscopy*, **F205**, 153–164.

McCullough S and Lucocq J (2005) Endoplasmic reticulum positioning and partitioning in mitotic HeLa cells. *Journal of Anatomy*, **206**, 415–425.

Michel RP and Cruz-Orive LM (1988) Application of the Cavalieri principle and vertical sections method to lung: estimation of volume and pleural surface area. *Journal of Microscopy*, **150**, 117–136.

Mironov AA Jr and Mironov AA (1998) Estimation of subcellular organelle volume from ultrathin sections through centrioles with a discretized version of the vertical rotator. *Journal of Microscopy*, **192**, 29–36.

Nyengaard JR and Gundersen HJG (1992) The isector: a simple and direct method for generating isotropic, uniform random sections from small specimens. *Journal of Microscopy*, **165**, 427–431.

Posthuma G, Slot JW, Veenendaal T and Geuze HJ (1988) Immunogold determination of amylase concentrations in pancreatic subcellular compartments. *European Journal of Cell Biology*, **46**, 327–335.

Prior IA, Muncke C, Parton RG and Hancock JF (2003) Direct visualization of Ras proteins in spatially distinct cell surface microdomains. *The Journal of Cell Biology*, **160**, 165–170.

Reed MG and Howard CV (1999) Stereological estimation of covariance using linear dipole probes. *Journal of Microscopy*, **195**, 96–103.

Ripley BD (1981) *Spatial Statistic*, Wiley: New York.

Roberts N, Puddephat MJ and McNulty V (2000) The benefit of stereology for quantitative radiology. *The British Journal of Radiology*, **73**, 679–697.

Smythe E, Pypaert M, Lucocq J and Warren G (1989) Formation of coated vesicles from coated pits in broken A431 cells. *The Journal of Cell Biology*, **108**, 843–853.

Sterio DC (1984) The unbiased estimation of number and sizes of arbitrary particles using the disector. *Journal of Microscopy*, **134**, 127–136.

Stringer BMJ, Wynford-Thomas D and Williams ED (1982) Physical randomization of tissue architecture: an alternative to systematic sampling. *Journal of Microscopy*, **126**, 179–182.

Thompson E (1930) Quantitative microscopic analysis. *The Journal of Geology*, **38**, 193.

Watt SA, Kimber WA, Fleming IN, Leslie NR, Downes CP and Lucocq JM (2004) Detection of novel intracellular agonist responsive pools of phosphatidylinositol 3,4-bisphosphate using the TAPP1 pleckstrin homology domain in immunoelectron microscopy. *Biochemical Journal*, **377**, 653–663.

Watt SA, Kular G, Fleming IN, Downes CP and Lucocq JM (2002) Subcellular localization of phosphatidylinositol 4,5-bisphosphate using the pleckstrin homology domain of phospholipase C delta J. *Biochemical Journal*, **363**, 657–666.

Weibel ER (1979) *Stereological Methods Vol 1 Practical Methods for Biological Morphometry*, Academic Press: London.

Weibel ER and Paumgartner D (1978) Integrated stereological and biochemical studies on hepatocytic membranes. II. Correction of section thickness effect on volume and surface density estimates. *The Journal of Cell Biology*, **77**, 584–597.

Short Specialist Review

High content screening

D. Lansing Taylor and Kenneth A. Giuliano
Cellumen, Inc., Pittsburgh, PA, USA

1. Background

The field of "cellomics" emerged from the early success of genomics and proteomics and the term has been used interchangeably with "functional genomics" and more recently with "functional proteomics". Our definition of cellomics, as a field, is broader: "the study of the temporal and spatial activity of cells and cellular constituents that are responsible for cell functions".

High Content Screening (HCS) is a set of methods and tools that were originally created, developed, commercialized, and defined by Cellomics, Inc. in 1997 (Giuliano *et al*., 1997; Giuliano *et al*., 2003) as a platform technology for the emerging field of cellomics. In today's language, this would be called a "systems cell biology" platform. The original definition of HCS was "an automated method for analyzing arrays of cells that contain one or more fluorescent reporter molecules, where the fluorescent signals are converted into digital data and then the digital data are used to automatically make measurements of intensity and/or distribution of the fluorescent signals on or in the cells, where changes indicate a change in distribution, environment, or activity of the fluorescent reporter molecules". Variations in the original definition of HCS have evolved over the last few years as more scientists have used the technology in drug discovery and basic biomedical research. The most general definition offered to date has been "the use of biological probes in combination with imaging technologies to probe within individual cells and to screen for activity on a target at a subcellular level (Screening Review, 2004)".

HCS was conceived as a high-throughput, cell biology platform aimed at "industrializing" cell biology. The concept was analogous to the creation of automated DNA sequencing. The developers of the automated DNA sequencers that eventually allowed the human genome project to be successful in a reasonable amount of time and at a reasonable cost, did not develop all the technologies from scratch. They integrated previously developed methods of fluorescent dye tagging of nucleotides, running gels containing DNA fragments, reading the "ladders" in the gels and incorporating the sequence data into searchable databases. The major step from a technical and an intellectual property perspective was the integration and automation of the whole set of processes into one turnkey method. The output of the process of automated DNA sequencing was not a "ladder" read on the gel, but a DNA sequence that could be fed into a DNA search engine to find new genes.

The development of HCS followed the same path as automated DNA sequencing. Cellomics, Inc. successfully integrated and automated multiple, previously single technologies including fluorescence microscopy; imaging science; fluorescence-based reagents (Taylor, 1992; Taylor *et al*., 2001); cell assays; and searchable databases. The result, HCS, formed a platform technology that was positioned to industrialize the field of cellomics.

2. Present applications

To date, a wide range of cellular processes have been explored with HCS. The first applications were measuring the activation of transcription factors (e.g., NF-kB) by quantifying the translocation from the cytoplasm into the nucleus (Ding *et al*., 1998) and quantifying the internalization of receptors into vesicles (Conway *et al*., 1999). Subsequently, screens for a wide range of transcription factors and receptors have been performed. An increasing number of cellular processes ranging from apoptosis, cell migration, cell cycle, G-protein coupled receptor activation (Walker *et al*., 1999), microtubule stability (Giuliano, 2003), and cytotoxicity (Abraham *et al*., 2004) have been implemented. It is believed that any cellular constituent including "biosensor" proteins and organelles can be specifically labeled and quantitative measurements can form the basis of an HCS screen (Giuliano and Taylor, 1998; Meyer and Teruel, 2003).

An ability to take more than one measurement from a cell has allowed the so-called multiplexing. For example, while the function (e.g., translocation) of the target protein in question can be monitored using one fluorescence channel (e.g., using green fluorescent protein tags (*see* Article 50, **Using photoactivatable GFPs to study protein dynamics and function**, Volume 5), or staining with antibodies coupled to fluoresein (*see* Article 58, **Immunofluorescent labeling and fluorescent dyes**, Volume 5)), other parameters of the cell can be simultaneously measured in another fluorescence channel, such as nuclear volume by staining with Hoechst dyes. This increases the information content of the assay.

3. Future developments

HCS has been accepted as an important approach to drug discovery by the pharmaceutical industry and it is now being incorporated into large-scale screens in academic research. The first generation platforms will now evolve as the market demands more from the platform. There will be advances in the individual component technologies that make up HCS, as well as in the integrated solutions. The optical and mechanical systems will become more sophisticated with the inclusion of modules for: reading microarrays of cells on chips; spectral deconvolution in order to analyze greater than four fluorescent reporters in the same assay; large area detection to increase the throughput; and additional fluorescence spectroscopic measurements to broaden the molecular information gained. Imaging science will be harnessed to perform faster and more powerful pattern recognition analyses that

will increase the robustness of the measurements. Training sets of data will allow the end user to better define the positive and negative attributes that the systems will be programmed to detect and to measure. In addition, the increased speed will permit the measurement of all possible morphometric and fluorescence parameters that will help define the optimal parameters for making decisions.

More specific, as well as multiplexed fluorescence-based reagents will be created; especially for live-cell screens. Multiplexing reagents with distinct spectral distributions, fluorescence lifetimes, and fluorescence anisotropy will be used in combination with reagents that can "modulate" cell constituent functions such as RNA-based reagents that can "knock down" coding and even noncoding genes (*see* Article 60, **siRNA approaches in cell biology**, Volume 5), as well as "chemical switches" that can regulate specific gene expression. Individual, as well as multiple cellular pathways will be interrogated with these advanced reagents opening up more complicated assays involving multiplexed measurements in four dimensions (x,y,z space and time). The creation of even more data per assay/screen (many terabytes) with these new tools will drive the development and implementation of more advanced informatics and cellular bioinformatics. The process of rapidly flowing through the production of data, the extraction of information and finally, the creation of knowledge will become faster and more powerful.

HCS promises to have the same impact on the field of cellomics, as automated DNA sequencing had on the field of genomics.

References

Abraham VC, Taylor DL and Haskins JR (2004) High content screening applied to large-scale cell biology. *Trends in Biotechnology*, **22**, 15–22.

Conway BR, Minor LK, Xu JZ, Gunnet JW, DeBiasio R, D'Andrea MR, Rubin R, DeBiasio R, Giuliano K, DeBiasio L, *et al*. (1999) Quantitation of G-protein coupled receptor internalization using G-protein coupled-green fluorescent protein conjugates with the ArrayScan high content screening system. *Journal of Biomolecular Screening*, **4**, 75–86.

Ding GJF, Fischer PA, Boltz RC, Schmidt JA, Colaianne JJ, Gough A, Rubin RA and Miller DK (1998) Characterization and quantitation of NF-kB nuclear translocation induced by interleukin-1 and tumor necrosis factor-alpha. Development and use of a high capacity fluorescence cytometric system. *The Journal of Biological Chemistry*, **273**, 28897–28905.

Giuliano KA (2003) High Content profiling of drug-drug interactions: cellular targets involved in the modulation of microtubule drug action by the antifungal ketoconazole. *Journal of Biomolecular Screening*, **8**, 125–135.

Giuliano KA and Taylor DL (1998) Fluorescent-protein biosensors: new tools for drug discovery. *Trends in Biotechnology*, **16**, 135–140.

Giuliano KA, Haskins JR and Taylor DL (2003) Advances in high content screening for drug discovery. *Assays and Drug Development Technologies*, **1**, 565–577.

Giuliano KA, DeBiasio RL, Dunlay RT, Gough A, Volosky JM, Zock J, Pavlakis GN and Taylor DL (1997) High content screening: A new approach to easing the bottlenecks in the drug discovery process. *Journal of Biomolecular Screening*, **2**, 249–259.

Meyer T and Teruel MN (2003) Fluorescence imaging of signaling networks. *Trends in Cell Biology*, **13**, 101–106.

Screening Review (2004) Select Biosciences Ltd: **www.ScreeningReview.com**.

Taylor DL, Woo ES and Giuliano KA (2001) Real-time molecular and cellular analysis: the new frontier of drug discovery. *Current Opinion in Biotechnology*, **12**, 75–81.

Taylor DL, Nederlof M, Lanni F and Waggoner AS (1992) The new vision of light microscopy. *American Scientist*, **80**, 322–335.

Walker JK, Premont RT, Barak LS, Caron MG and Shetzline MA (1999) Properties of Secretin receptor internalization differ from those of the B (2) – adrenergic receptor. *The Journal of Biological Chemistry*, **274**, 31515–31523.

Short Specialist Review

Photobleaching (FRAP/FLIP) and dynamic imaging

George Banting
University of Bristol, Bristol, UK

1. Background

Fluorescence Recovery After Photobleaching (FRAP) and Fluorescence Loss In Photobleaching (FLIP) are related techniques, both of which are useful in the study of protein movement in live cells. Both techniques, particularly FRAP, have been used for many years to follow the diffusion of molecules (proteins and lipids) in the plane of the lipid bilayer of the plasma membrane (e.g., see Wey *et al*., 1981; Jacobson *et al*., 1976a,b; Axelrod *et al*., 1976; Edidin *et al*., 1976). Other related techniques are now also widely used (*see* Article 49, **Small molecule fluorescent probes for protein labeling and their application to cell-based imaging (including FlAsH etc.)**, Volume 5, Article 50, **Using photoactivatable GFPs to study protein dynamics and function**, Volume 5, Article 51, **FRET-based reporters for intracellular enzyme activity**, Volume 5, Article 55, **Imaging protein function using fluorescence lifetime imaging microscopy (FLIM)**, Volume 5 and Article 56, **Elucidating protein dynamics and function using fluorescence correlation spectroscopy (FCS)**, Volume 5, Article 57, **Quantum dots for multiparameter measurements**, Volume 5). Historically, specific cell-surface components were labeled with a fluorophore (e.g., fluorescein or tetramethylrhodamine) to enable them to be visualized during the imaging process. These early studies, pioneered by biophysicists, laid foundations upon which researchers from biophysicists to cell biologists now capitalize. The techniques of FRAP and FLIP became more accessible to cell biologists, particularly those with some molecular biological expertise, following the discovery of green fluorescent protein (GFP) and, more importantly, the isolation of the cDNA sequence that encodes it (Chalfie *et al*., 1994). GFP (including its spectral variants and related fluorescent proteins) is particularly attractive as a tool for studies in live cells since it is genetically encoded and emits visible light without any requirement for cofactors. Thus, over the past 10 years, many proteins have been tagged with GFP, expressed in cells, and then visualized by fluorescence microscopy. So, although the FRAP and FLIP techniques can, in theory, be applied to any fluorescently labeled molecule, this short review will assume that they are being used in the study of proteins that have been tagged with fluorescent proteins (FPs).

2. FRAP

Once a construct encoding a protein of interest tagged with GFP (or one of its spectral variants) has been generated and the recombinant protein has been successfully expressed in an appropriate cell (and shown to be correctly localized in that cell), it is possible to use FRAP and FLIP to study the movement of that protein. It is important to note that elevated expression of recombinant GFP-tagged proteins can lead to their mislocalization (e.g., see Girotti and Banting, 1996), so it is important that recombinant GFP-tagged proteins, as with any other epitope-tagged protein, are shown to be correctly localized before further studies are carried out. In the FRAP technique, a defined region of the cell (region of interest, ROI) being imaged is subjected to a short burst of high-intensity laser illumination. This treatment leads to photobleaching of the fluorophore (i.e., the FP) in the ROI. Recovery of fluorescence in the ROI can then be recorded by imaging the cell at regular intervals over a period of time (see Figure 1). This recovery of fluorescence is due to repopulation of the ROI by fluorescent molecules that have moved in from the surrounding area in exchange for photobleached molecules moving out. Thus, if we are considering a membrane protein, recovery of the fluorescent signal in the photobleached region can be considered to be an indicator of the diffusional mobility of that protein in the plane of the lipid bilayer. Indeed, this process can be followed quantitatively by time-lapse microscopy, with the fluorescent intensity of the ROI relative to that in the prebleach period (I/I_0) being plotted as a function of time (see Figure 2). This allows determination of the mobile and immobile fractions of the protein as well as the effective diffusion coefficient (D_{eff}) of the mobile fraction (Kubitscheck *et al*., 1994; Soumpasis, 1983; Axelrod *et al*., 1976), that is, how much of the protein is free to diffuse and how rapidly it does so if it can. Clearly, this technique is not only applicable to membrane proteins, but the

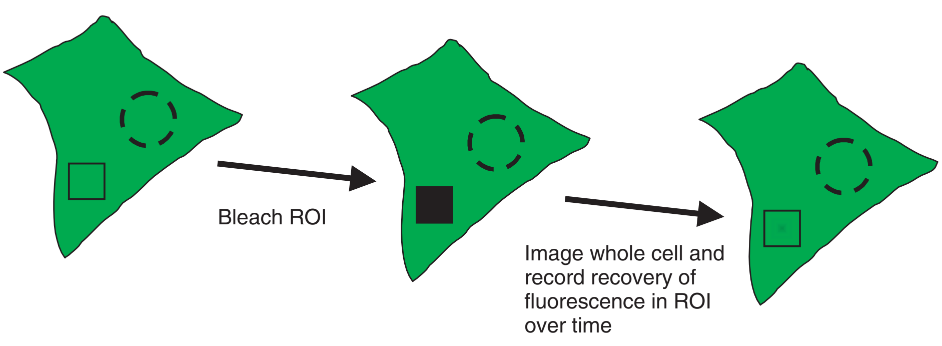


Figure 1 FRAP

- Image cell at low light intensity.
- Photobleach-defined area (ROI, boxed area above) using high light intensity.
- Image cell at low light intensity.
- Observe bleached area (black) adjacent to fluorescent area (green if using GFP-tagged protein).
- Observe repopulation of bleached area by fluorescently tagged protein.

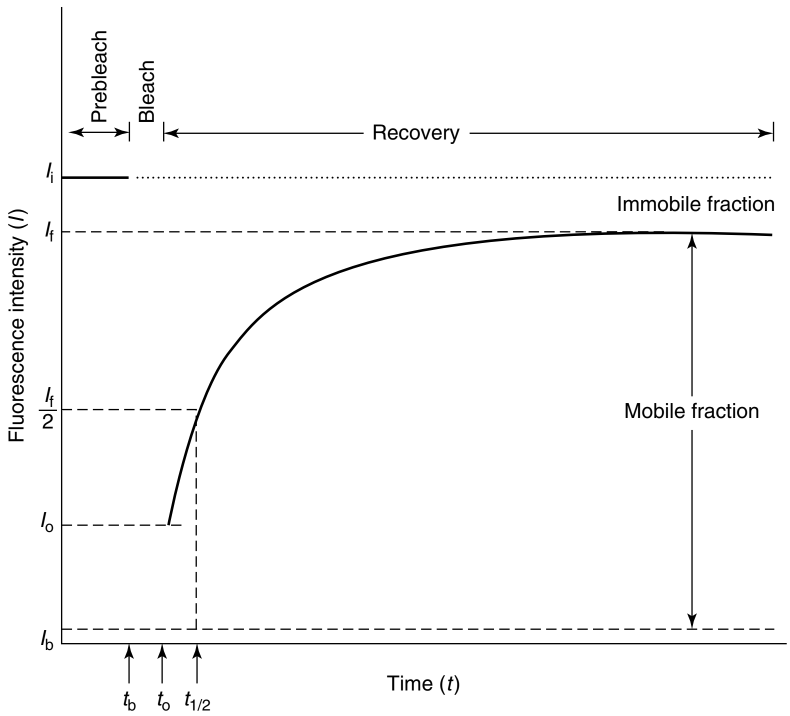


Figure 2 The region of interest (ROI) is photobleached between time t_b and t_0 and the fluorescence decreases from the initial fluorescence I_i to I_0. The fluorescence then recovers over time (due to replacement of bleached molecules by fluorescent ones) until recovery is complete (I_f). $t_{1/2}$ indicates the time taken for 50% recovery of the fluorescence signal. The intensity of the fluorescence signal in the ROI may well not return to the original value (I_i) following recovery (i.e., I_f does not equal I_i as shown above). This difference can be attributed to the immobile fraction of the fluorescently tagged protein

diffusion of soluble proteins in the cytosol can be assayed as can translocation of proteins (e.g., from cytosol to membrane or to nucleus) and protein exchange in macromolecular complexes (see Reits and Neefjes, 2001; Bastiaens and Pepperkok, 2000 for citations of examples).

3. FLIP

FLIP (see Figure 3) might be considered to be a corollary of FRAP. The difference between FRAP and FLIP is that whereas in FRAP the ROI is subjected to photobleaching only once, in FLIP it is repeatedly photobleached. The repeated bursts of photobleaching are interspersed with acquisition of images of the whole cell. This allows visualization of a gradual loss of fluorescence from the whole of the region that was originally fluorescent and occurs as fluorescent molecules move into the ROI, get photobleached, then move out again. Loss of fluorescence in different regions of the cell can be monitored over time, giving mobile and

immobile fractions in each of those regions as well as the D_{eff} for the mobile fraction. Clearly, this technique can also be used for both soluble and membrane proteins, and nicely complements FRAP.

4. Caveats

There are caveats that must be borne in mind when interpreting FRAP and FLIP data. For example, the use of FRAP to study protein diffusion in the lipid bilayer of the plasma membrane is relatively straightforward since the proteins being followed are present in a continuous two-dimensional membrane system. However, many intracellular membrane systems are highly dynamic and not necessarily continuous; therefore, fluorescence recovery in a ROI will be due to a combination of diffusion in the plane of the lipid bilayer, reconfiguration of the membrane system, and the degree of continuity/discontinuity of that membrane system. A further point that must be borne in mind when using FRAP (or FLIP) is that reversible photobleaching of GFP can occur (Swaminathan *et al*., 1997). As described elsewhere (Bastiaens and Pepperkok, 2000), a simple way to determine whether reversible photobleaching is occurring or not is to change the size of the ROI that is photobleached. The kinetics of recovery will be unaffected by the size of the ROI if recovery is due to reversible photobleaching. The effect of de novo protein synthesis on FRAP and FLIP results should also be considered, that is, is

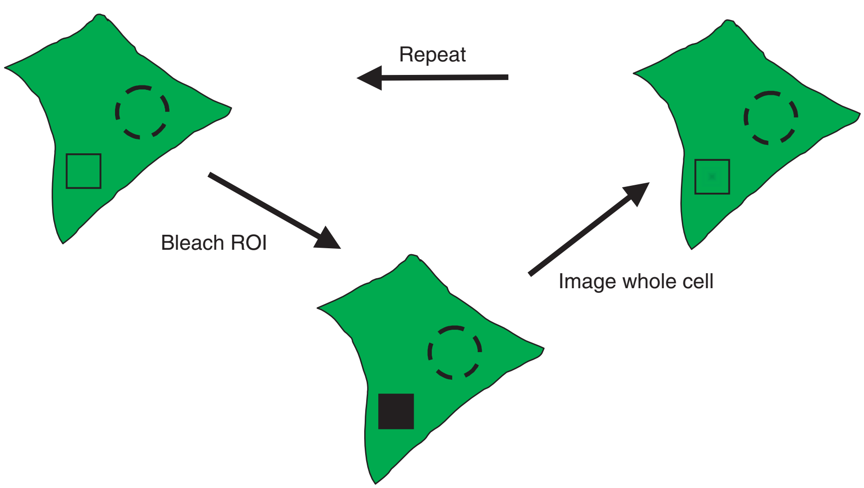


Figure 3 FLIP

- Image cell at low light intensity.
- Photobleach defined area (ROI, boxed area above) using high light intensity.
- Image cell at low light intensity.
- Repeat above two steps multiple times.
- Observe gradual loss of fluorescence throughout entire cell.

recovery of fluorescence in an ROI due to movement of an already synthesized protein from another location to the ROI or is it due to repopulation of the ROI by newly synthesized material? This question can be addressed by performing the FRAP and FLIP analysis on cells that are incubated in the presence of a protein synthesis inhibitor such as cycloheximide.

References

Axelrod D, Koppel DE, Schlessinger J, Elson E and Webb WW (1976) Mobility measured by fluorescence photobleaching recovery kinetics. *Biophysical Journal*, **16**, 1055–1069.

Bastiaens PIH and Pepperkok R (2000) Observing proteins in their natural habitat: the living cell. *Trends in Biochemical Sciences*, **25**, 631–637.

Chalfie M, Tu Y, Euskirchen G, Ward WW and Prasher DC (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802–805.

Edidin M, Zagyansky Y and Lardner TJ (1976) Measurement of membrane protein lateral diffusion in single cells. *Science*, **191**, 466–468.

Girotti M and Banting G (1996) TGN38-green fluorescent protein hybrid proteins expressed in stably transfected eukaryotic cells provide a tool for the real-time, in vivo study of membrane traffic pathways and suggest a possible role for ratTGN38. *Journal of Cell Science*, **109**, 2915–2926.

Jacobson K, Derzko Z, Wu ES, Hou Y and Poste G (1976a) Measurement of the lateral mobility of cell surface components in single, living cells by fluorescence recovery after photobleaching. *Journal of Supramolecular Structure*, **5**, 565–576.

Jacobson K, Wu E and Poste G (1976b) Measurement of the translational mobility of concanavalin A in glycerol-saline solutions and on the cell surface by fluorescence recovery after photobleaching. *Biochimica et Biophysica Acta*, **433**, 215–222.

Kubitscheck U, Wedekind P and Peters R (1994) Lateral diffusion measurement at high spatial resolution by scanning microphotolysis in a confocal microscope. *Biophysical Journal*, **67**, 948–956.

Reits EAJ and Neefjes JJ (2001) From fixed to FRAP: measuring protein mobility and activity in living cells. *Nature Cell Biology*, **3**, E145–E147.

Soumpasis DM (1983) Theoretical analysis of fluorescence photobleaching recovery experiments. *Biophysical Journal*, **41**, 95–97.

Swaminathan R, Hoang CP and Verkman AS (1997) Photobleaching recovery and anisotropy decay of green fluorescent protein GFP-S65T in solution and cells: cytoplasmic viscosity probed by green fluorescent protein translational and rotational diffusion. *Biophysical Journal*, **72**, 1900–1907.

Wey CL, Cone RA and Edidin MA (1981) Lateral diffusion of rhodopsin in photoreceptor cells measured by fluorescence photobleaching and recovery. *Biophysical Journal*, **33**, 225–232.

Short Specialist Review

Probing protein function in cells using CALI

Thomas J. Diefenbach and Daniel G. Jay
Tufts University School of Medicine, Boston, MA, USA

1. Introduction

Proteomics is advancing beyond the cataloging of the complement of expressed proteins in cells. Understanding protein function within the context of the cell is a key final step in proteomics. Traditionally, molecular genetics has been used to test protein function through knockout and overexpression of relevant genes. Chromophore-assisted laser inactivation (CALI) provides a direct means for acute inactivation of *in situ* protein function.

2. The process of CALI

Nonfunction-blocking antibodies (IgG, IgM, Fab fragments, single chain antibodies or scFvs) are covalently linked to a photosensitizer dye such as malachite green isothiocyanate (MG-ITC; Figure 1).

The labeled antibodies are loaded into cells using electroporation, trituration-loading, scrape-loading, and microinjection, or are simply applied to the surface of the cells or preparation to address surface proteins (Buchstaller and Jay, 2000). The preparation is then exposed to a specific wavelength of light. The light is absorbed by the dye and this energy raises the chromophore into an excited state. Upon relaxation of the excited dye, free radicals are generated and the cycle repeats (Figure 2). In the case of CALI utilizing malachite green, it is thought that energy is transferred to a water molecule, and a hydroxyl radical is formed with a lifetime of 1 ns and a half maximal radius of inactivation of 15 Å (Jay, 1988; Liao *et al*., 1994; Linden *et al*., 1992).

An illustration of the spatial specificity of the radius of action of CALI comes from a study of the three 25-kDa subunits of the T-cell receptor complex (Liao *et al*., 1995). CALI targets either of the subunits bound by the antibody, with a slight effect on the antibody binding to the neighboring subunit, and no other effects on other components of the receptor complex.

The hydroxyl radical can damage the protein in different ways through reactions with different amino acid side chains (arginine, histidine, methionine, phenylalanine, tyrosine, tryptophan) (Halliwell and Gutteridge, 1989). A fluorophore-based

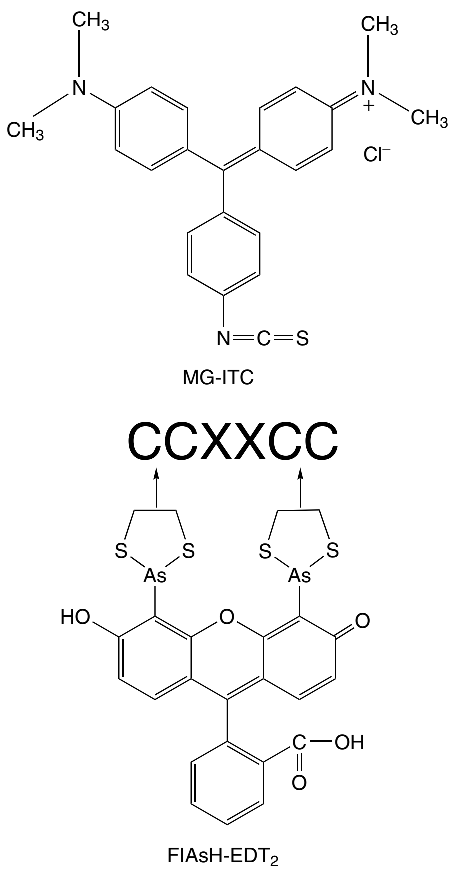


Figure 1 Atomic structures of two common CALI photoactivators, malachite green isothiocyante (MG-ITC), and fluorescein isothiocyanate (FITC), are covalently linked to free amines (lysine) on the antibodies or scFv fragments through reaction of the isothiocyanate (N=C=S) with amine to form a thiourea. Also shown are two membrane-permeable fluorophores, ReAsH-EDT$_2$ and FlAsH-EDT$_2$, each with two arsenics that bind a tetracysteine motif (X is any amino acid except cysteine) expressed as part of the recombinant target protein of interest

derivative of CALI utilizing fluorescein isothiocyanate (FITC; Figure 1) instead of malachite green has also been developed and is referred to as FALI (Fluorophore-assisted light inactivation; Beck *et al*., 2002). The reactive species for FALI is singlet oxygen, and experiments with free-radical quenchers suggest a larger half-maximal inactivation radius of ~40 Å (Beck *et al*., 2002). Singlet oxygen has a much greater lifetime in cells of 200 ns versus 1 ns for the hydroxyl radical. Although the singlet oxygen likely reacts with four amino acids (tryptophan, histidine, methionine, and cysteine), the reactions are complex, can proceed through several mechanisms yielding distinct decomposition products (Halliwell and Gutteridge, 1989), and are as much as 300-fold slower than the reaction of the hydroxyl radical with similar amino acids. The lifetimes of free radicals are limited by endogenous free-radical scavengers (ascorbate and glutathione for hydroxyl

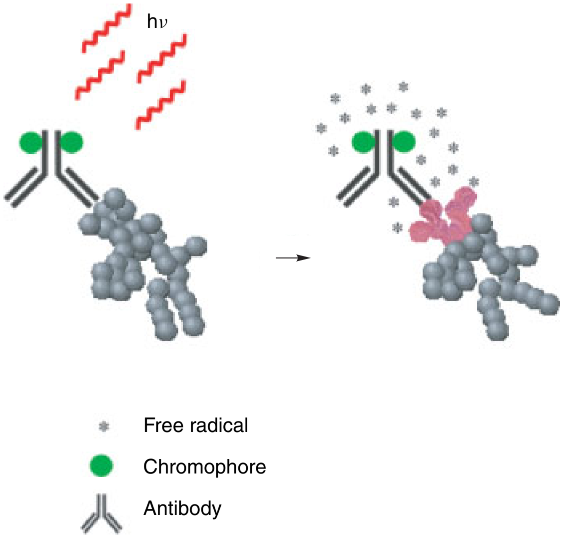


Figure 2 Schematic of the CALI process. The dye-labeled antibody binds the target protein and is irradiated with either laser or incandescent light (hν). Energy is transferred from the dye to water or free oxygen to form hydroxyl radical or singlet oxygen, respectively. These short-lived reactive species (*) cause local and irreversible damage to specific amino acids on the bound protein. This damage is spatially restricted, and may have many consequences, including perturbation of protein conformation, enzymatic activity, or protein–protein interactions

radicals, ergothioneine, and carotenoids for singlet oxygen; Chaudiere and Ferrari-Iliou, 1999). The cell's growth state may also affect its ability to deal with oxidative damage by free radicals, as proliferating or tumorigenic cells display an inherently greater capacity to deal with oxidative stress (for review see Das, 2002). These factors should be considered when assessing the applicability of different CALI methods.

CALI is performed on one of three scales: micro-CALI, macro-CALI, and FALI. Macro-CALI employs a large laser to produce a 2–5-mm diameter laser spot useful for irradiation of tissue samples, culture dishes, or microtiter wells. Micro-CALI uses a less powerful laser, the light of which is directed through the optics of an inverted microscope, yielding spot sizes of 5–100 μm in diameter. Micro-CALI can therefore be used to target protein populations in individual cells or within specific cellular compartments. FALI employs diffuse light, allowing many samples to be treated in parallel in multiwell plates.

Genetic compensation is not as much a concern with the acute loss of function afforded by CALI as it can be for RNAi and genetic loss of function. Furthermore, the RNAi approach cannot distinguish between different proteins arising through posttranslational modification, while CALI may distinguish between isoforms with isoform-specific antibodies. Unlike RNAi, CALI can be performed any time after antibody application or loading, and recovery of protein function will depend on turnover of the target protein. CALI is thus ideally suited to address protein function in the dynamic and rapidly changing cellular environment. Limitations of CALI

include: the requirement for a specific antibody, a need to load antibodies into cells for intracellular targets, and recovery will occur with *de novo* protein synthesis.

3. Developments in CALI technology

CALI technology has been aided by expanding the selection of photoactivatable dyes, or by changing the way in which those dyes bind to the target protein. FALI utilizes fluorescein, which is excited by higher energy, shorter wavelength light, permitting the use of conventional light sources, and the ability to illuminate entire multiwell plates for high-throughput, cell-based proteomic screens (Beck *et al*., 2002; Eustace *et al*., 2002). These screens have addressed cell surface targets by adding exogenous antibody to the cell preparations prior to FALI. One of the advantages of FALI as a screening tool is that it can take advantage of single chain antibody (scFv) libraries to bypass the traditional high-throughput bottlenecks in a "top–down" approach to screen protein function directly, as a complement to extensive mutational analyses. As part of such a proteomic approach, FALI identifies scFvs that are then used to identify their targets immunoprecipitation and mass spectrometry.

One molecular genetic approach to CALI employs the use of green fluorescent protein (GFP) as a genetically encoded photoactivator (Rajfur *et al*., 2002; *see also* Article 48, **Probing cellular function with bioluminescence imaging**, Volume 5 and Article 50, **Using photoactivatable GFPs to study protein dynamics and function**, Volume 5). GFP is relatively inefficient as a photoconverter likely on account of shielding of the fluorophore nested within its barrel structure, and may thus require longer illumination periods for full effect (Surrey *et al*., 1998). Two new methods for genetically targeted CALI have been developed, using genetically encoded protein tags to which a biarsenyl fluorophores selectively bind (*see* Article 49, **Small molecule fluorescent probes for protein labeling and their application to cell-based imaging (including FlAsH etc.)**, Volume 5). FlAsH-FALI (FlAsH, or fluorescein derivative with two As(III) substituents on the 4′ and 5′ positions) utilizes site-specific fluorescent labeling of recombinant proteins by genetically tagging the target protein with a tetracysteine motif. This motif is then recognized by a membrane-permeant, biarsenical fluorescein derivative (Figure 1). This labeling method was initially tested in cells by fusing the tetracysteine motif to the COOH-terminus of a mutant of cyan fluorescent protein (ECFP) (Griffin *et al*., 1998). FlAsH only fluoresces upon binding to the cysteine motif on the protein of interest (Adams *et al*., 2002), although nonspecific binding to endogenous tetracysteine motifs is a potential concern. FlAsH-FALI effects will be limited to expressed proteins, with endogenous proteins not affected except perhaps through collateral damage. Uniform penetration of the fluorescein derivative into tissue and residual background fluorescence through nonspecific binding are issues to be considered for FlAsH-FALI, depending on the system. FlAsH-FALI also requires the coaddition of 1,2-dithiols to outcompete FlAsH binding to endogenous cysteine pairs (Adams *et al*., 2002). Since FlAsH-FALI uses fluorescein, which requires higher energy, shorter wavelength light (typically 480–490 nm) that is

absorbed by cells, a membrane-permeant, red biarsenical fluorophore (ReAsH-EDT_2; Figure 1) was developed, which is excited by longer wavelength, 593-nm light (ReAsH-CALI; Tour *et al*., 2003). With ReAsH-CALI, maximal inactivation of connexin-43, a component of gap junctions, was achieved using a 150-W xenon arc lamp with only 10 to 30 s of irradiation (Tour *et al*., 2003). Combined, these approaches expand the utility of CALI and create a bridge for CALI as an adjunct to molecular genetic approaches.

4. Applications of CALI at the cellular level

CALI has been used to examine the roles of proteins that are members of larger protein families and that have not yet been engineered transgenically or lack specific pharmacological blockers. For example, the myosin superfamily of actin-based molecular motors comprises 18 classes. Genetic knockouts of individual myosins are problematic since a single cell possesses members from many myosin classes and often more than one isoform. CALI has been used to probe the function of distinct myosins: myosins 1c and Va (Wang *et al*., 1996), and myosins IIA and IIB (Diefenbach *et al*., 2002) in relation to the motility of the growing tips of nerve cell processes (for earlier review, see Buchstaller and Jay, 2000). CALI technology has also been applied to the family of neural cell adhesion molecules (NCAMs). Neurons from knockout mice that do not express the cell adhesion molecule, L1, fail to extend neurites on an L1 substrate (Dahme *et al*., 1997). In this case, the complete lack of the appropriate adhesion molecule precluded study of its role in cell shape change because the cells failed to initiate extension of their neuritic processes. Micro-CALI of L1 resulted in neurite retraction without affecting the protrusive activity of the neurite tips, while micro-CALI of NCAM-180 impaired protrusive activity at the tips of extending neurites without affecting neurite extension (Takei *et al*., 1999).

CALI has also been applied to examination of proteins implicated in cancer cell growth and migration, including the actin-associated protein ezrin (Lamb *et al*., 1997), and the protein tyrosine kinase, pp60 c-src (Hoffman-Kim *et al*., 2002). In pp60 c-src knockout mice, no major neurological deficits were found (Soriano *et al*., 1991). Yet, with acute inactivation of pp60 c-src, extension of neuronal processes in culture was dramatically increased, implicating this tyrosine kinase as a negative regulator of laminin-mediated neuronal growth. The molecular chaperone, heat shock protein 90α (hsp90α) was recently identified in a FALI-based, high-throughput proteomic screen, and found to facilitate, via an extracellular route, cancer cell migration through its interaction with matrix metalloproteases (Eustace *et al*., 2004). CALI also has potential in the field of prion protein research (Graner *et al*., 2000). These studies underscore the utility of CALI in dissecting the cellular roles of individual members of large protein families.

5. Applications of CALI in vivo

CALI can also address protein function at distinct developmental timepoints in vivo or during dynamic processes such as cell division and cell migration. CALI

of fasciclin I reduced bundling of pioneer axons in the developing grasshopper (Jay and Keshishian, 1990; Diamond *et al*., 1993), and CALI of patched protein altered cell fate in *Drosophila* (Schmucker *et al*., 1994). CALI of the repulsive guidance cue ephrin-A5 in chick optic tectum in vivo limits growth of axons of the retinotectal projection (Sakurai *et al*., 2002). In organ culture of retina and optic nerve, CALI of myelin-associated glycoprotein resulted in regeneration of retinal ganglion cells after an optic nerve crush (Wong *et al*., 2003). Thus, CALI technology has proven itself in many in vivo applications.

To conclude, CALI technology provides a direct means of assaying protein function and dynamics in cells and tissues with great temporal and spatial precision. It can serve as a complementary approach to knockout, RNAi, antisense, pharmacological or other approaches in that it provides a means of addressing the endpoint of a proteomic strategy – how the function of the protein directly impacts cellular function.

References

Adams SR, Campbell RE, Gross LA, Martin BR, Walkup GK, Yao Y, Llopis J and Tsien RY (2002) New biarsenical ligands and tetracysteine motifs for protein labeling in vitro and in vivo: synthesis and biological applications. *Journal of the American Chemical Society*, **124**, 6063–6076.

Beck S, Sakurai T, Eustace BK, Beste G, Schier R, Rudert F and Jay DG (2002) Fluorophore-assisted light inactivation: a high-throughput tool for direct target validation of proteins. *Proteomics*, **2**, 247–255.

Buchstaller A and Jay DG (2000) Micro-scale chromophore-assisted laser inactivation of nerve growth cone proteins. *Microscopy Research and Techniques*, **48**, 97–106.

Chaudiere J and Ferrari-Iliou R (1999) Intracellular antioxidants: from chemical to biochemical mechanisms. *Food and Chemical Toxicology*, **37**, 949–962.

Dahme M, Bartsch U, Martini R, Anliker B, Schachner M and Mantei N (1997) Disruption of the mouse L1 gene leads to malformations of the nervous system. *Nature Genetics*, **17** (3), 346–349.

Das UN (2002) A radical approach to cancer. *Medical Science Monitor*, **8** (4), RA79–RA92.

Diamond P, Mallavarapu A, Schnipper J, Booth J, Park L, O'Connor TP and Jay DG (1993) Fasciclin I and II have distinct roles in the development of grasshopper pioneer neurons. *Neuron*, **11**, 409–421.

Diefenbach TJ, Latham VM, Yimlamai D, Liu CA, Herman IM and Jay DG (2002) Myosin 1c and myosin IIB serve opposing roles in lamellipodial dynamics of the neuronal growth cone. *Journal of Cell Biology*, **158**, 1207–1217.

Eustace BK, Buchstaller A and Jay DG (2002) Adapting chromophore-assisted laser inactivation for high throughput functional proteomics. *Briefings in Functional Genomics and Proteomics*, **1** (3), 257–265.

Eustace BK, Sakurai T, Stewart JK, Yimlamai D, Unger C, Zehetmeier C, Lain B, Torella C, Henning SW, Beste G, *et al*. (2004) Functional proteomic screens reveal an essential extracellular role for hsp90 alpha in cancer cell invasiveness. *Nature Cell Biology* **6** (6), 507–514.

Graner E, Mercadante AF, Zanata SM, Martins VR, Jay DG and Brentani RR (2000) Lamini-induced PC-12 cell differentiation is inhibited following laser inactivation of cellular prion protein. *FEBS Letters*, **482** (3), 257–260.

Griffin BA, Adams SR, and Tsien RY (1998) Specific covalent labeling of recombinant protein molecules inside live cells. *Science* **281**(5374), 269–272.

Halliwell B and Gutteridge JMC (1989) *Free Radicals in Biology and Medicine*, Second Edition, Clarendon Press: Oxford.

Hoffman-Kim D, Kerner JA, Chen A, Xu A, Wang TF and Jay DG (2002) pp60(c-src) is a negative regulator of laminin-1-mediated neurite outgrowth in chick sensory neurons. *Molecular and Cellular Neuroscience*, **21** (1), 81–93.

Jay DG (1988) Selective destruction of protein function by chromophore-assisted laser inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 5454–5458.

Jay DG and Keshishian H (1990) Laser inactivation of fasciclin I disrupts axon adhesion of grasshopper pioneer neurons. *Nature*, **348** (6301), 548–550.

Lamb R, Ozanne BW, Roy C, McGarry L, Stipp C, Mangeat P and Jay DG (1997) Essential functions of ezrin in maintenance of cell shape and lamellipodial extension in normal and transformed fibroblasts. *Current Biology*, **7**, 682–688.

Liao JC, Roider J and Jay DG (1994) Chromophore-assisted laser inactivation of proteins is mediated by the photogeneration of free radicals. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2659–2663.

Liao JC, Berg L and Jay DG (1995) Chromophore-assisted laser inactivation of subunits of the T cell receptor in living cells is spatially restricted. *Photochemical Photobiology*, **62**, 923–929.

Linden KG, Liao JC and Jay DG (1992) Spatial specificity of chromophore-assisted laser inactivation of protein function. *Biophysical Journal*, **61**, 956–962.

Rajfur Z, Roy P, Otey C, Romer L and Jacobson K (2002) Dissecting the link between stress fibres and focal adhesions by CALI with EGFP fusion proteins. *Nature Cell Biology*, **4**, 286–293.

Sakurai T, Wong E, Drescher U, Tanaka H and Jay DG (2002) Ephrin-A5 restricts topographically specific arborization in the chick retinotectal projection *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America*, **99** (16), 10795–10800.

Schmucker D, Su A, Beermann A, Jackle H and Jay DG (1994) Laser inactivation of pached protein switches cell fate in the larval visual system of *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2664–2668.

Soriano P, Montgomery C, Gesake R and Bradley A (1991) Targeted disruption of the $^{c-src}$proto-oncogene leads to osteopetrosis in mice. *Cell*, **64**, 693–702.

Surrey T, Elowitz MB, Wolf PE, Yang F, Nedelec F, Shokat K and Leibler S (1998) Chromophore-assisted light inactivation and self-organization of microtubules and motors. *Proceedings of the National Academy of Sciences of the United States of America*, **95** (8), 4293–4298.

Takei K, Chan TA, Wang FS, Deng H, Rutishauser U and Jay DG (1999) The neural cell adhesion molecules L1 and NCAM-180 act in different steps of neurite outgrowth. *Journal of Neuroscience*, **19**, 9469–9479.

Tour O, Meijer RM, Zacharias DA, Adams SR and Tsien RY (2003) Genetically targeted chromophore-assisted light inactivation. *Nature Biotechnology*, **21**(12), 1505–1508.

Wang FS, Wolenski JS, Cheney RE, Mooseker MS, and Jay DG (1996) Function of myosin-V in filopodial extension of neuronal growth cones. *Science* **273**(5275), 660–663.

Wong EV, David S, Jacob MH and Jay DG (2003) Inactivation of myelin-associated glycoprotein enhances optic nerve regeneration. *Journal of Neuroscience*, **23**(8), 3112–3117.

Short Specialist Review

Imaging protein function using fluorescence lifetime imaging microscopy (FLIM)

Yan Gu and Daniel Zicha

Cancer Research UK London Research Institute, London, UK

1. Introduction to FLIM techniques

Fluorescence lifetime is the average time for molecules to spend in the excited state before returning to the ground state with either emitting a photon or transferring their energy nonradiatively to surrounding molecules. Lifetimes of most molecules are measured in nanoseconds and are influenced by local environment such as collisional quenching, pH variation, temperature, and Fluorescence Resonance Energy Transfer (FRET). Lifetime gives valuable insight into the binding of emitting molecules, and surrounding chemical and physical parameters.

Lifetime is usually measured by FLIM (fluorescence lifetime imaging microscopy), which can be broadly divided into frequency-domain approaches and time-domain approaches. Frequency-domain FLIM modulates the emission of fluorophores with sinusoidal excitation, and uses the modulation difference and phase delay between excitation and emission to calculate their lifetime (Lakowicz, 1999). Time-domain FLIM relies on fast-gated detection of individual photons emitted by fluorophores excited by a femtosecond laser pulse in order to "snap" their decay, from which their lifetime is processed. Time-domain approaches are implemented as either wide-field or point-scan FLIM. Wide-field FLIM uses a fast-gated optical intensifier and a CCD camera to acquire a time sequence of 2D decays with high-resolution time delay (Cole *et al*., 2000). Point-scan FLIM, on the other hand, relies on 2D scan to acquire spatial information. Recent improvements in the Time Correlated Single Photon Counting (TCSPC) technique have greatly facilitated FLIM measurements (Becker *et al*., 2003).

Other specialized FLIM techniques include anisotropy FLIM (rFLIM), spectral FLIM (sFLIM), and endoscopic FLIM (Becker *et al*., 2002; Hanley *et al*., 2002; Siegel *et al*., 2001). rFLIM uses perpendicularly polarized decay kinetics to explore the multiplicity of rotational correlated depolarization or anisotropy, reflecting heterogeneity in molecular populations and the size, shape, and internal motion of a complex (Clayton *et al*., 2002). sFLIM measures lifetime in a number of spectral bands and extends FLIM further into spectral dimension. In biological multicomponent systems, sFLIM provides better assessment to the lifetime

distribution of single fluorescent species, and increases contrast to distinguish multiple fluorophores or heterogeneity of local environment. Endoscopic FLIM combines FLIM with endoscopes, making possible in vivo lifetime measurements in tissue without staining (Tadrous *et al*., 2003).

Sensitivity is critical in lifetime measurements, especially when the molecular concentrations are low or interactions are weak. In many protein-interaction experiments, relative measurements such as lifetime ratio of different proteins or one protein over different time points are preferred to increase the sensitivity. A sensitive approach, referred as the *acceptor photobleaching FLIM*-FRET, was adopted from intensity FRET (Chen *et al*., 2003). Using this approach, the binding proteins can be solely identified by the lifetime difference of the donor before and after photobleaching. One of the important issues of this method is the oxygen-related radicals generated during photobleaching. Because they have very short lifetime, their presence would decrease the average lifetime of donors. Lampert *et al*. (1983) suggested removing the oxygen before the experiment to avoid these radicals and the related acceptor recovery after photobleaching.

Recent research has started to focus on live-cell FLIM approaches (Ng *et al*., 1999; Krishnan *et al*., 2003; Grecco *et al*., 2004; Errington *et al*., 2005; Treanor *et al*., 2005). FLIM on fixed cells provides a snapshot of protein interactions, whereas live-cell FLIM can reveal their dynamics. Live-cell FLIM can also avoid the artifacts caused by the fixation. The difficulties of live-cell FLIM, however, are associated with the fact that live cells are quite vulnerable to phototoxicity and photobleaching. Excitation strength must be controlled to acceptable levels. Therefore, FLIM measurements usually suffer from low signal-to-noise ratio and long acquisition time. A typical FLIM measurement takes several minutes, which makes it difficult to attain high temporal resolution. Faster FLIM approaches are often achieved with compromises in spatial resolution, spatial field of view, or detection accuracy. Increasing FLIM detection sensitivity seems to be the key issue to the live-cell FLIM. The influence of experimental conditions on the lifetime is another issue that has not been explored thoroughly. For instance, the local environment in fixed cell is stable, in contrast to a dynamic process in the living cell. Molecular diffusions in living cells provide rather different environment from that of fixed cells, and would in turn change the heterogeneity of lifetime.

2. FLIM data processing

Data processing is an important part of FLIM. As emission is a random process, each excited molecule has the same probability of emitting in a given time period. Statistically, the time-dependent fluorescence $I(t)$ reflects an exponential decay of the excited-state populations as (Lakowicz, 1999)

$$I(t) = \sum_{i=1}^{m} I_i(0)\mathrm{e}^{\frac{t}{\tau_i}} \tag{1}$$

where and $I_i(0)$ and τ_i are the emission (at t = 0) and the lifetime of ith fluorophore population, and m is the number of fluorophore populations in the system. In time-domain FLIM, original data are arranged in such a way that at each pixel, there

is a time sequence of intensities following the decay kinetics of molecules. The nonlinear least-squares fitting proposed by Grinvald and Steinberg (1974) is still widely used (Lakowicz, 1999). In a mono-exponential system, such a fit usually produces accurate estimation of the average lifetime. However, in multiexponential systems such as most biological systems, this algorithm can sometimes generate ambiguity, that is, more than one set of results will satisfy equation (1). Such a problem inherently originates from the fact that parameters $I_i(0)$ and τ_i are correlated, and one can either vary the lifetime to compensate the amplitude, or vice versa (Grinvald and Steinberg, 1974). In a typical protein–protein interaction system, at least two lifetimes (free and FRET-involved donors) need to be identified. Higher dimensions are required if spectral cross talk from acceptor and autofluorescence need to be considered. Fitting more than two components is very difficult and insensitive without many data points and auxiliary information (Cubeddu *et al*., 2002). In some cases, estimations between biexponential fit and triexponential fit to a true system are not quite distinguishable from Goodness of Fit (χ^2_R). Practical measures to avoid the ambiguity include (1) setting up lifetime boundaries using additional system knowledge, (2) obtaining images with high signal-to-noise ratio, (3) sampling as many data points as possible, and (4) acquiring additional information with auxiliary experiment to reduce the number of variables or set up additional restraints to the correlation between $I_i(0)$ and τ_i.

Various postprocessing algorithms have been developed to improve the signal-to-noise ratio and lifetime accuracy. Global analysis algorithm proposed by Verveer and Bastiaens (2003) assumes that neighboring pixels have similar lifetime, and different pixels in an image can be treated as independent measurements. They applied the algorithm in frequency-domain FLIM to study multiexponential systems with a single modulation frequency. Similar strategy is used in time-domain FLIM to increase the signal-to-noise ratio. Such an algorithm works well when images have good signal-to-noise ratio, and reducing spatial resolution is not an issue. Another option to improve the data quality without increasing data points is to optimize the sampling arrangement. Usually, data are sampled with equal time period. The fit process takes these points with equal weight. Apparently, variations of the decay signal influence the lifetime estimation differently. For instance, time points before 2τ have higher signal-to-noise ratio and therefore are more accurate than those after 3τ. It is helpful to arrange the sampling nonlinearly and emphasize more on the data with higher quality in the fit algorithm.

In this paper, we propose an approach using intensity-based spectral unmixing to provide auxiliary information to lifetime measurements. Spectral unmixing is a well-established image processing methodology, which is based on the approximation that the spectral signature of a test sample is the sum of a linear combination of abundance weighted spectra of constituent pure classes (Petrou and Foschi, 1999). Using the emission spectra of specified pure fluorochromes as a reference, the fluorescence of the corresponding fluorochrome within a complex can therefore be precisely determined from its composite spectrum. In a double-fluorophore system, the linear unmixing algorithm can be expressed as

$$F_m(x, y, \lambda) = C_1(x, y)F_1(\lambda) + C_2(x, y)F_2(\lambda) \qquad (2)$$

where $C_1(x, y)$ and $C_2(x, y)$ are unknown abundance factors of the two fluorochromes at pixel location (x, y), whose spectral responses are expressed as $F_1(\lambda)$ and $F_2(\lambda)$ respectively, and $F_m(x, y, \lambda)$ is the spectral response of a test sample. $F_1(\lambda)$ and $F_2(\lambda)$ are measured using single fluorochrome specimens (controls). To avoid interference from environment, controls and the test sample are made from the same specimen, and are imaged under exactly the same conditions whose influence is cancelled out during the unmixing process.

Assuming lifetime is detected within wavelength λ_1 and λ_2, the steady state emission ratio can be given as (Lakowicz, 1999)

$$\frac{I_1(0)\tau_1}{I_2(0)\tau_2} = \frac{C_1(x, y)\displaystyle\int_{\lambda_1}^{\lambda_2} F_1(\lambda)\,\mathrm{d}\lambda}{C_2(x, y)\displaystyle\int_{\lambda_1}^{\lambda_2} F_2(\lambda)\,\mathrm{d}\lambda} \tag{3}$$

Equation (3) provides an extra restraint to equation (1), which not only helps to solve the ambiguity problem but also improves the lifetime accuracy. If m fluorophores are in the complex, equation (3) consists of $m - 1$ ratios.

Nowadays, with the increasing demand of high spatial/temporal resolutions, more and more protein lifetime measurements are based on the combination of confocal microscopes, multiphoton excitation, and TCSPC (Becker & Hickl). Laser scanning confocal microscope LSM 510 META (Carl Zeiss Ltd.) allows multispectral images to be collected simultaneously without compromising resolution and efficiency. The 32 spectral detectors are aligned in such a way that they can cover any emission wavelength between 378.4 and 720.8 nm, with a maximum resolution of 11 nm. Outputs from a maximum of eight channels can provide continuous hyperspectral coverage of a sample's emission. Using spectral unmixing, we have successfully separated ECFP from EGFP and EYFP from EGFP. On the basis of LSM 510 META, we are exploring the possibility to associate FLIM with spectral-based measurements in a single cell mount, which provides multidimensional information (spatial, temporal, spectrum, and lifetime) for protein colocalization and interaction. Such combinations do not significantly increase the experimental load, as lifetime measurements and spectral emission can be accomplished in a single sample mounting. Since spectral measurements are much quicker than lifetime, and the additional steady state information helps to reduce data sampling in FLIM, it might speed up data acquisition. A macro program can be implemented to automatically conduct both measurements.

3. Protein functional interpretation using FLIM

Association with local environment makes lifetime a very powerful indicator to explore protein's neighborhood. However, it also makes the comparative measurement in a cell population very difficult. A fluorophore may have different lifetimes in different environments. For instance, ECFP has an average lifetime of 2.6 ns when fused with C/EBPα proteins expressed in GHFT1-5 cell nuclei (Chen

et al., 2003). This lifetime is shortened to 2.3 ns when transfected into HEK cells, and even becomes shorter (phase lifetime: 1.32 ns and modulation lifetime: 2.23 ns) when fused with nuclear bipartite nucleoplasmin in live Vero cells (Becker *et al*., 2004; Pepperkok *et al*., 1999). In FLIM-FRET, the variation of a donor's lifetime in a test sample from that of a donor-only control is an indicator of the energy transfer. Even if both samples are prepared under the same conditions and cells express at the same level, local heterogeneity between two independent cells will generate lifetime difference, regardless of the existence of FRET. Furthermore, since the resultant interaction distance is usually a statistical average, the quantified lifetime is often used only to deduce qualitative conclusions, such as reduced lifetime of donors which indicates the binding of proteins, while increased lifetime implies the opening of protein complexes (Ng *et al*., 2001).

According to literature, FLIM has been applied to a range of protein function studies, such as contrast enhancement of protein environment, and colocalization and interaction between proteins. Image contrast enhancement by lifetime can be used to distinguish molecules in different local environment, which may not be resolvable from their emission spectra. For example, collagens are abundant proteins in human tissue. Experimental data suggest that their autofluorescence is at least partly due to cross-linkage formed by the action of lysyl oxidase on lysine and hydroxylysine side chain polymer (Eyre, 1987). Such cross-linkages vary with tissue type and local environment such as glucose concentration, and increase with tissue age (Pongor *et al*., 1984). These variations alter the local lifetime and can provide contrast to distinguish collagens with different ages. As collagens emit naturally, FLIM has the potential to enhance the contrast of in vivo tissue imaging. Tadrous *et al*. (2003) used lifetime for histopathological assessment of breast cancer tissues.

Protein colocalization is another widely used lifetime application. Traditionally, colocalization refers to the pixel registration of different lifetime images. Owing to the diffraction limitation of optics, such a protein copresence is restricted to about 200 nm in resolution, which is quite coarse in protein's scale (several nanometers). The development of FLIM-FRET techniques greatly increases the resolution (Bacskai *et al*., 2003; Chen and Periasamy, 2004). When two different proteins or different parts of a protein are tagged with different fluorophores, FRET signal will report colocalizations or binding activities if the proximity of two fluorophores is less than 10 nm. Bacskai *et al*. (2003) applied double immunofluorescence to the brain tissue Tg2576 transgenic mice, which overexpress mutant human amyloid precursor protein, leading to the progressive formation of senile plaques similar to those found in Alzheimer's disease. FITC-10 d5 and rhodamine-labeled 3 d6 antibodies were used to recognize epitopes specifically bound to the N-terminal of amyloid-β (amino acids 3 to 6 and 1 to 5 respectively). The lifetime distributions revealed that they are colocalized on amyloid-β deposits within 10 nm, and two epitopes have different relative conformations. Apparently, protein colocalization by FRET only is limited within 10 nm. While pixel-based colocalization provides wider field of view and FRET-based colocalization gives high resolution, the combination of both techniques provides a high dynamical range and nanometer resolution solution to identify the formation of protein complexes.

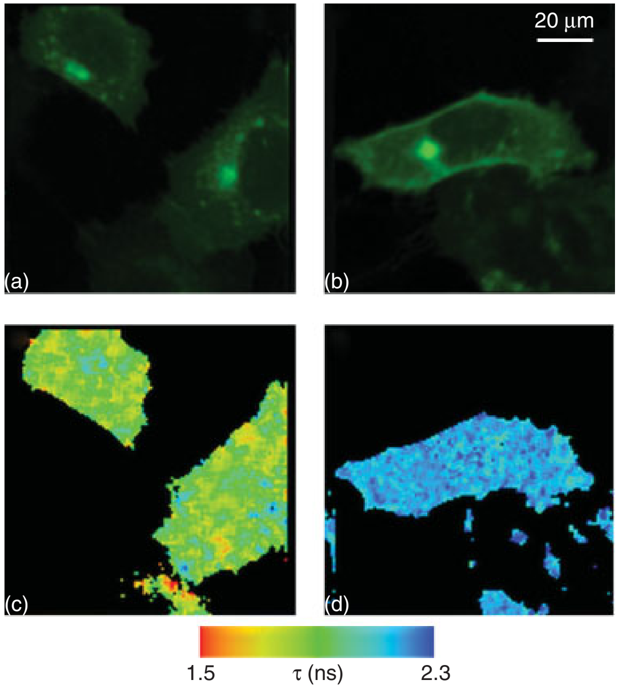


Figure 1 N-WASP/Cdc42 interaction in fixed NIH 3T3 fibroblasts. Images, acquired using a multiphoton TCSPC system, represent donor fluorescence intensity (a and b) and associated lifetime maps (c and d) for cells expressing EGFP-N-WASP & Myc-Cdc42, either in the presence (a and c) or absence (b and d) of Cy3-MycAntibody acceptor labeling. Lifetime images are presented in pseudocolor, defined by the associated scale

Typical interaction applications include quantitative ion concentration and calcium-binding studies, the pH fluctuations, and oxygen concentration change monitoring in cells (Bastiaens and Squire, 1999; Periasamy *et al*., 2001). In our studies of the mechanisms of cell motility, we found that inhibition of Cdc42 or N-WASP in NIH 3T3 fibroblasts led to 30% reduction of cell speed. This significant reduction can be explained by the fact that two molecules act together in a complex. In order to directly address this hypothesis, we used FLIM-FRET to visualize the interaction between Cdc42 and N-WASP (Figure 1), in collaboration with T. Ng, Randall Centre, King's College London, UK, and B. Vojnovic, Gray Cancer Institute, Northwood, UK (Monypenny, 2003). As shown in Figure 1, the shortening of lifetime of EGFP-N-WASP demonstrates the occurrence of FRET and therefore a direct indication of N-WASP and Cdc42 being in a very close proximity (Figure 1c in comparison to Figure 1d).

When protein–protein interaction lasts for a relatively long time period (compared with the measuring speed of FLIM), live-cell FLIM can be used to monitor its dynamics. Peter *et al*. (2004) applied live-cell FLIM to study the receptor-kinase interaction between the chemokine receptor (CXCR4) and protein kinase Cα (PKCα) in carcinoma cells. They found the CXCR4-PKCα complex forms concomitant with the endocytosis process, which would be difficult to detect in fixed

cells. With dynamic FLIM-FRET, Martin-Fernandez *et al*. (2004) observed a progressive decrease in FRET efficiency during the capsid disassembly for adenovirus serotype 5 (Ad5) in living CHO-CAR cells. On the basis of FRET efficiency, they concluded that the disassembly process has two sequential protein dissociation rates.

In conclusion, developments in confocal microscopy enable the spectral- and lifetime-based techniques to be combined in the lifetime measurement, providing more sensitive and accurate estimations. With the advance of technology, FLIM is being driven toward quantitative protein functional interpretations in live cells for both in vitro and in vivo applications.

References

Bacskai BJ, Skoch J, Hickey GA, Allen R and Hyman BT (2003) Fluorescence resonance energy transfer determinations using multiphoton fluorescence lifetime imaging microscopy to characterize amyloid-beta plaques. *Journal of Biomedical Optics*, **8**, 368–375.

Bastiaens P and Squire A (1999) Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. *Trends in Cell Biology*, **9**, 48–52.

Becker W, Bergmann A, Biskup C, Kelbauskas L, Zimmer T, Klöcker N and Benndorf K (2003) High resolution TCSPC lifetime imaging. *Proceedings of SPIE*, **4963**, 175–184.

Becker W, Bergmann A, Biskup C, Zimmer T, Klöcker N and Benndorf K (2002) Multiwavelength TCSOC lifetime imaging. *SPIE*, **4620**, 79–84.

Becker W, Bergmann A, Hink MA, König K, Benndorf K and Biskup C (2004) Fluorescence lifetime imaging by time-correlated single-photon counting. *Microscopy Research and Technique*, **63**, 58–66.

Chen Y, Mills JD and Periasamy A (2003) Protein localization in living cells and tissues using FRET and FLIM. *Differentiation*, **71**, 528–541.

Chen Y and Periasamy A (2004) Characterization of Two-photon Excitation Fluorescence Lifetime Imaging Microscopy for Protein Localization. *Microscopy Research and Technique*, **63**, 72–80.

Clayton AHA, Hanley QS, Arndt-Jovin DJ, Subramaniam V and Jovin TM (2002) Dynamic Fluorescence Anisotropy Imaging Microscopy in the Frequency Domain (rFLIM). *Biophysical Journal*, **83**, 1631–1649.

Cole MJ, Siegel J, Webb SED, Jones R, Dowling K and French PMW (2000) Whole-field optically sectioned fluorescence lifetime imaging. *Optics Letters*, **25**, 1361–1363.

Cubeddu R, Comelli D, D'Andrea C, Taroni P and Valentini G (2002) Time-resolved fluorescence imaging in biology and medicine. *Journal of Physics D: Applied Physics*, **35**, R61–R76.

Errington RJ, Ameer-beg SM, Vojnovic B, Patterson LH, Zloh M and Smith PJ (2005) Advanced microscopy solutions for monitoring the kinetics and dynamics of drug-DNA targeting in living cells. *Advanced Drug Delivery Reviews*, **57**, 153–167.

Eyre D (1987) Collagen cross-linking amino acids. *Methods in Enzymology*, **144**, 115–139.

Grecco HE, Lidke KA, Heintzmann R, Lidke DS, Spagnuolo C, Martinez OE, Jares-Erijman EA and Jovin TM (2004) Ensemble and Single Particle Photophysical Properties (Two-Photon Excitation, Anisotropy, FRET, Lifetime, Spectral Conversion) of Commercial Quantum Dots in Solution and in Live Cells. *Microscopy Research and Technique*, **65**, 169–179.

Grinvald A and Steinberg IZ (1974) On the Analysis of Fluorescence Decay Kinetics by the Method of Least-Squares. *Analytical Biochemistry*, **59**, 583–598.

Hanley QS, Arndt-Jovin DJ and Jovin TM (2002) Spectrally Resolved Fluorescence Lifetime Imaging Microscopy. *Applied Spectroscopy*, **56**, 155–166.

Krishnan RV, Saitoh H, Terada H, Centonze VE and Herman B (2003) Development of a multiphoton fluorescence lifetime imaging microscopy system using a streak camera. *The Review of Scientific Instruments*, **74**, 2714–2721.

Lakowicz JR (1999) *Principle of Fluorescence Spectroscopy*, Kluwer Academic / Plenum Publishers: New York.

Lampert RA, Chewter LA, Phillips D, O'Connor DV, Roberts AJ and Meech SR (1983) Standards for Nanosecond Fluorescence Decay Time Measurements. *Analytical Chemistry*, **55**, 68–73.

Martin-Fernandez M, Longshaw SV, Kirby I, Santis G, Tobin MJ, Clarke DT and Jones GR (2004) Adenovirus Type-5 Entry and Disassembly Followed in Living Cells by FRET, Fluorescence Anisotropy, and FLIM. *Biophysical Journal*, **87**, 1316–1327.

Monypenny JE (2003) The development of quantitative live cell imaging techniques and their applications in the study of inter-cellular communication and sarcoma cell motility. *PhD Thesis*, University College of London: London, 249–254.

Ng T, Parsons M, Hughes WE, Monypenny J, Zicha D, Gautreau A, Arpin M, Gschmeissner S, Verveer PJ, Bastiaens PIH, *et al*. (2001) Ezrin is a downstream effector of trafficking PKC-integrin complexes involved in the control of cell motility. *The EMBO Journal*, **20**, 2723–2741.

Ng T, Squire A, Hansra G, Bornancin F, Prevostel C, Hanby A, Harris W, Barnes D, Schmidt S, Mellor H, *et al*. (1999) Imaging Protein Kinase C_α Activation in Cells. *Science*, **283**, 2085–2089.

Pepperkok R, Squire A, Geley S and Bastiaens PIH (1999) Simultaneous detection of multiple green fluorescent proteins in live cells by fluorescence lifetime imaging microscopy. *Current Biology*, **9**, 269–272.

Periasamy A, Elangovan M, Wallrabe H, Barroso M, Demas JN, Brautigan DL and Day RN (2001) In *Methods in Cellular Imaging*, Periasamy A (Ed.), Oxford University Press: Hong Kong, pp. 295–308.

Peter M, Ameer-Beg SM, Hughes MKY, Keppler MD, Prag S, Marsh M, Vojnovic B and Ng T (2004) Multiphoton-FLIM Quantification of the EGFP-mRFP1 FRET Pair for Localization of Membrane Receptor-Kinase Interactions. *Biophysical Journal*, **88**, 1224–1237.

Petrou M and Foschi PG (1999) Confidence in Linear Spectral Unmixing of Single Pixels. *IEEE Transactions on Geoscience and Remote Sensing*, **37**, 624–626, Part 2.

Pongor S, Ulrich PC, Bencsath FA and Cerami A (1984) Aging of proteins: Isolation and identification of a fluorescent chromophore from the reaction of polypeptides with glucose. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 2684–2688.

Siegel J, Elson DS, Webb SED, Parsons-Karavassilis D, Leveque-Fort S, Cole MJ, Lever MJ, French PMW, Neil MAA, Juskaitis R, *et al*. (2001) Whole-field five-dimensional fluorescence microscopy combining lifetime and spectral resolution with optical sectioning. *Optics Letters*, **26**, 1338–1340.

Tadrous PJ, Siegel J, French PMW, Shousha S, Lalani EN and Stamp GWH (2003) Florescence lifetime imaging of unstained tissues: early results in human breast cancer. *Journal of Pathology*, **199**, 309–317.

Treanor B, Lanigan PMP, Suhling K, Schreiber T, Munro I, Neil MAA, Phillips D, Davis DM and French PMW (2005) Imaging fluorescence lifetime heterogeneity applied to GFP-tagged MHC protein at an immunological synapse. *Journal of Microscopy*, **217**, 36–43.

Verveer PJ and Bastiaens PIH (2003) Evaluation of global analysis algorithms for single frequency gluorescence lifetime imaging microscopy data. *Journal of Microscopy*, **209**, 1–7.

Elucidating protein dynamics and function using fluorescence correlation spectroscopy (FCS)

Petra Schwille
Dresden University of Technology, Dresden, Germany

Fluorescence correlation spectroscopy (FCS) was first described as a very sensitive relaxation technique by Magde *et al*. (1972) and during the following years has been applied by several groups to determine the concentration and mobility (in particular, diffusion) parameters as well as chemical kinetics. In contrast to classical fluorescence spectroscopy that measures the average signal, FCS resolves temporal variations in fluorescence, induced by statistical fluctuations of measurement parameters such as particle number and brightness. Consequently, all dynamic processes that change particle brightness, mobility, and concentration can be sensitively analyzed in any transparent medium.

The ability of FCS to measure local fluorophore concentrations and diffusion coefficients with minimal interference with the biological system has had important consequences for applying FCS to cell-based problems. However, it was not until the early 1990s when FCS was combined with a confocal microscopy setup that this promise was realized. The small volume resolution of FCS achieved using confocal imaging (down to the femtoliter range, which is of the order of the volume of an *E. coli* cell), has provided a real breakthrough in the FCS field. The problem of low signal-to-noise levels that had so far substantially limited its applications and frustrated many researchers, who were originally convinced by the conceptual beauty of the technique, was now overcome. Rigler and coworkers, who were the first to apply FCS in confocal setups, reported signal-to-noise levels up to 1000 and could easily resolve the transits of single fluorescent molecules through the illuminated focal measurement volume (Rigler *et al*., 1993; Eigen and Rigler, 1994).

It is fair to say, therefore, that the confocal FCS scheme introduced by Rigler *et al*. triggered a true renaissance of this technique. Biochemists, molecular and cell biologists then immediately recognized the potential of FCS for ultrasensitive analysis of biomolecular dynamics in homogeneous assays. Today's standard FCS applications range from the determination of concentrations and mobility parameters, rate constants of association and dissociation reactions, or enzymatic turnovers, to timescales of internal fluctuations due to structural or photophysical dynamics. In combination with specific dyes, internal fluctuation patterns can yield

environmental parameters such as pH, ionic strength, or oxygen levels. A review of solution-based applications is provided by Elson and Rigler (2001). In all these applications, the essential information can be drawn from the observation of single molecules entering and leaving a laser-illuminated confocal volume element by random diffusive motion, one by one, inducing characteristic fluctuations in the fluorescence signal. Molecular residence times in the observation volume gives direct access to mobility parameters. Fluctuation strength, on the other hand, is inversely related to the occupation number and, thus, the concentration. Internal dynamics can usually be resolved if the processes to be investigated (e.g., conformational changes) change the fluorescence properties of the fluorophore. The characteristic fluctuation pattern is then not only induced by the entering and leaving of the measurement volume but also by reversible on–off blinking of the fluorophore during the dynamic process to be investigated.

A key application of FCS in cell-based assays is in determination of protein–protein interactions, be it of receptor–ligand binding on cell membranes or other recognition processes. However, enzymatic turnovers in the living cell can also be monitored. Standard FCS is rather limited in following molecular interactions, because its key measurement parameters, the residence time of single molecules in the confocal spot or the local concentration, are rarely affected by specific reactions in a predictable and reliable way. Sufficient measurement selectivity can only be guaranteed if the interaction partners are labeled with distinct fluorescent probes, like in other fluorescence microscopy applications such as colocalization or FRET assays. However, in contrast to these imaging-based techniques, dual-color FCS does not depend on the spatial proximity of two labels that could or could not reflect true mutual protein binding. Rather, the so-called dual-color cross-correlation technique probes the concomitant *movement* of particles by quantifying the number of coincident fluctuation events in two independent measurement channels (see Figure 1). Noninteracting, single-labeled molecules induce only noncoordinated fluctuations in either one of the detection channels (here, green or red). Only true interactions between two differently labeled molecules make them codiffuse in a way that induces coincident fluorescence fluctuations in both measurement channels every time a complex molecule makes a transit (Schwille *et al*., 1997). Assuming the dual-color confocal setup is well calibrated, the amplitude of the so-called cross-correlation function is directly proportional to the relative number of bound species. This means that molecular interactions can be comfortably determined in situ, even in the complex cellular environment.

Despite the huge promise of dual-color cross-correlation as a proteomic technique, few intracellular applications have so far been reported, although commercial FCS systems are available. One potential reason for this surrounds the high sensitivity of the method, making it susceptible to several artifacts, including the confounding effects of nonspecific binding of the labeled molecules to cellular structures, or photobleaching of the probes. Furthermore, concentrations of fluorescent molecules above 100 nM seriously lower the signal-to-noise ratio and, thus, sensitivity.

A major technical challenge is guaranteeing the stable overlap of the green and the red excitation volumes in two-color applications. However, it has been found that two-photon excitation, a nonlinear excitation mode frequently employed in

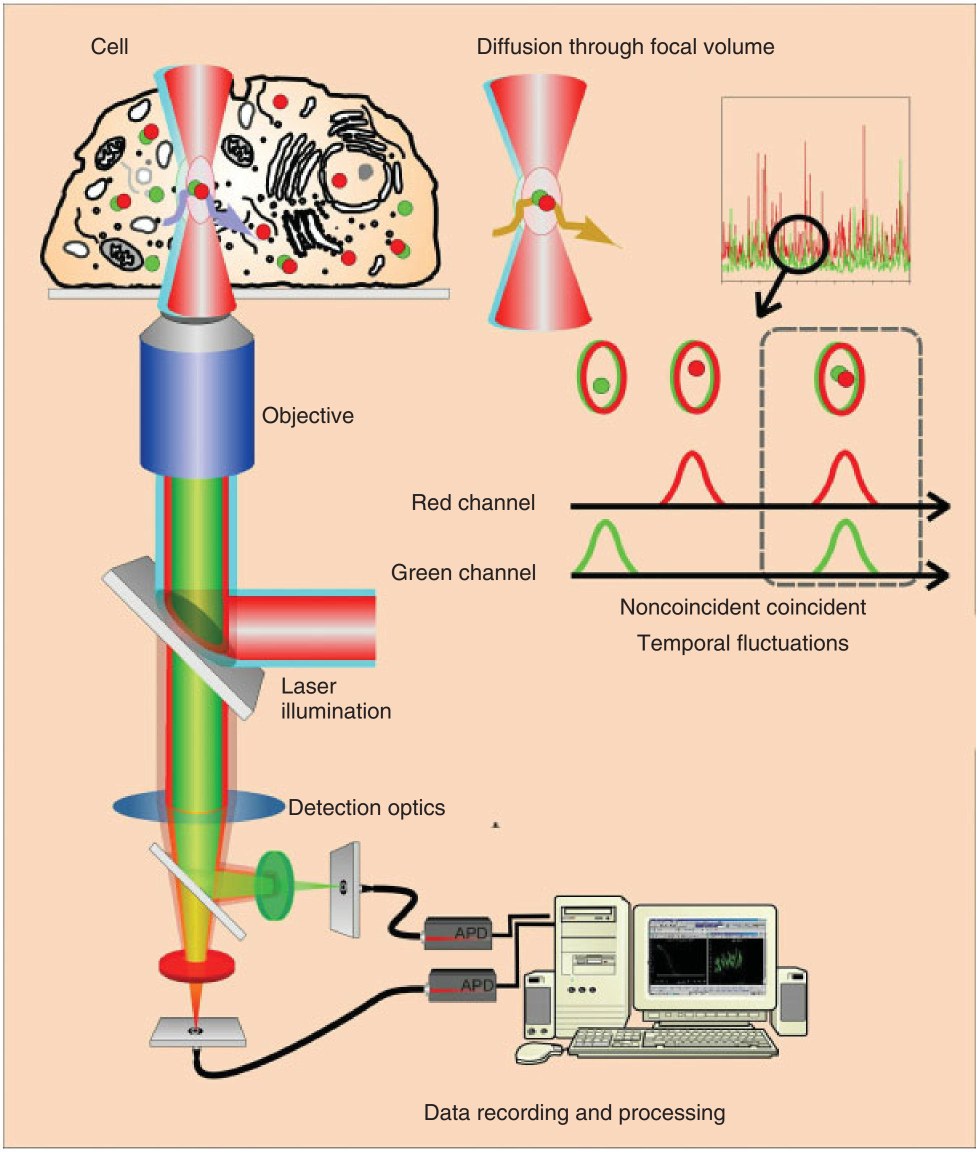


Figure 1 Principle of dual-color cross-correlation analysis for measurements in live cells: Concomitant movement of labels, which reflects molecular association of two different species, can be evidenced by coincident fluctuations in two different measurement channels. Hereby, protein–protein interaction can be probed in situ

fluorescence imaging, considerably improves cellular FCS applications. On the one hand, this is due to the dramatically reduced scattering and photobleaching in the spatially confined two-photon excitation spot, but also because of the broad absorption spectra for two-photon excitation of common probes (i.e., dyes with significantly different emission spectra can be simultaneously two-photon excited by the same infrared wavelength (Heinze *et al*., 2000)). This significantly

simplifies the optical setup, as only one excitation volume is now required. By this instrumental modification, two- or even three-color intracellular FCS applications are now possible, promising an enormous wealth of real-time studies on protein interaction patterns in situ (Kim *et al*., 2004).

A further and attractive application of dual-color cross-correlation has recently been proposed (Bacia *et al*., 2002). By following the diffusional movement of small endocytic vesicles with differentially labeled cargoes, various endocytic pathways can be distinguished and the preference of certain endocytosis mechanisms for different cargo species identified.

Combined with fluorescence resonance energy transfer (FRET), protein conformational dynamics can be addressable by FCS. This involves positioning the donor and acceptor dyes at opposite ends of a single protein. If reversible conformational changes occur on timescales that are fast compared to the observation time window, that is, the diffusion time of a molecule through the open focal volume, they can be dynamically analyzed by auto- and cross-correlation analysis (Margittai *et al*., 2003).

An interesting alternative to the standard confocal or two-photon excitation setup for sufficient background suppression in FCS applications, particularly on membranes, has been reported recently (Lieto *et al*., 2003). Total internal reflection (TIR) microscopy is a specific illumination feature that achieves extremely high axial resolution on the measurement volume to juxta-membrane regions. If laser light is incident upon a medium of lower refractive index under a critical angle, excitation will only occur in a thin surface layer of ca. 100 nm, the so-called evanescent field. In combination with a confocal pinhole for detection, unique resolution features can be obtained, and the separation of membrane-bound processes from fluorescent background in the solution above and below the membrane is facilitated. A combined TIR-FCS setup is particularly attractive to study the interaction of ligands with membrane-associated receptors, since it not only allows the user to distinguish free from bound ligand but also allows determination of binding and dissociation kinetics (i.e., a ligand can only be detected during its residence at the membrane where it is bound to the receptor).

Overall, it is fair to state that the era of cellular applications of FCS, as well as other single-molecule-based techniques, has just begun. A fascinating variety of applications are awaiting researchers in this special field of cellular biophysics. Presently, the bottleneck for these applications is not the optical technology but the availability of suitable fluorescence assays. Highly efficient site-specific fluorescent labeling of the molecule of interest, so minimizing the presence of unlabeled molecules and free dye, is often absolutely mandatory for quantitative studies. GFP technology with its broad palette of spectrally separable, genetically encoded probes certainly provides great promise for the development of multicolor assays. For example, GFP and its red homolog DsRed make a very good pairing for dual-color cross-correlation applications (Kohl *et al*., 2002). On the other hand, semiconductor nanoparticles such as "quantum dots" (*see* Article 57, **Quantum dots for multiparameter measurements**, Volume 5) may well provide a very attractive tool for FCS applications once their biocompatibility has been solved.

References

Bacia K, Majoul IV and Schwille P (2002) Probing the endocytic pathway in live cells using dual-color fluorescence cross-correlation analysis. *Biophysical Journal*, **83**, 1184–1193.

Eigen M and Rigler R (1994) Sorting single molecules, applications to diagnostics and evolutionary biotechnology. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 5740–5747.

Elson EL and Rigler R (Eds.) (2001) *Fluorescence Correlation Spectroscopy. Theory and Applications*, Springer: Berlin.

Heinze KG, Koltermann A and Schwille P (2000) Simultaneous two-photon excitation of distinct labels for dual-color fluorescence cross-correlation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 10377–10382.

Kim SA, Heinze KG, Waxham MN and Schwille P (2004) Intracellular calmodulin availability accessed with two-photon cross-correlation. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 105–110.

Kohl T, Heinze KG, Kuhlemann R, Koltermann A and Schwille P (2002) A protease assay for two-photon crosscorrelation and FRET analysis based solely on fluorescent proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12161–12166.

Lieto AM, Cush RC and Thompson NL (2003) Ligand-receptor kinetics measured by total internal reflection with fluorescence correlation spectroscopy. *Biophysical Journal*, **85**, 3294–3302.

Magde D, Elson EL and Webb WW (1972) Thermodynamic fluctuations in a reacting system – measurement by fluorescence correlation spectroscopy. *Physical Review Letters*, **29**, 705–708.

Margittai M, Widengren J, Schweinberger E, Schroder GF, Felekyan S, Haustein E, König M, Fasshauer D, Grubmüller H, Jahn R, *et al*. (2003) Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15516–15521.

Rigler R, Mets Ü, Widengren J and Kask P (1993) Fluorescence correlation spectroscopy with high count rates and low background, analysis of translational diffusion. *European Biophysics Journal*, **22**, 169–175.

Schwille P, Meyer-Almes F-J and Rigler R (1997) Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. *Biophysical Journal*, **72**, 1878–1886.

Short Specialist Review

Quantum dots for multiparameter measurements

Marcel P. Bruchez
Quantum Dot Corporation, Hayward, CA, USA

The optical properties of semiconductor nanocrystals are unique in fluorescence (Hotz, 2005; Bruchez *et al*., 1998; Chan and Nie, 1998). These particles are made from semiconductor materials at a size where the optical properties of the semiconductor are perturbed by quantum mechanical effects. This size range is known as the "quantum confinement" size regime. The emission spectra of high-quality quantum dots are very narrow (20–30-nm full width at half maximum), symmetric, and the emission color depends sensitively on the properties of the underlying particles, especially their size and composition. The absorbance of the particles is very large, typically 2–10 times any dye molecule at a single wavelength, and increases continuously at wavelengths shorter than the emission peak (see Figure 1). High-quality conjugated materials can be consistently prepared from blue to the near-infrared (450–900 nm), with quantum efficiencies of greater than 50%, resulting in the brightest fluorescently labeled biomolecules available. Because of the similar excitation properties for all colors of quantum dots, multiple colors can be efficiently excited simultaneously with a light source such as blue-violet filtered light or a 405- or 488-nm laser. Quantum dots are inorganic particles, and are well insulated from the environment, resulting in high photostability. This allows the materials to be tracked for many minutes to hours without any loss in fluorescence signal. The quantum dot materials provide a bright, sensitive family of easily distinguished fluorescent labels that can be imaged continuously for a variety of applications across biology.

Initial applications of these materials focused on labeling fixed antigens in cells and tissues with antibodies and biomolecules directly conjugated to quantum dots (Wu *et al*., 2003). Demonstrations of up to 5-color imaging with simple color filtering were easily performed with secondary detection using Qdot® conjugates to highly cross-adsorbed secondary antibodies (Figure 2). Such imaging with typical fluorophores cannot be accomplished without matrix unmixing of the fluorescent images owing to heavy overlap of the emission of the distinct dyes. Multicolor analysis of cells allows simultaneous imaging of morphological and functional parameters in a single cell, for instance, translocation and posttranslational modifications such as phosphorylation and glycosylation. In many cases, the state of a population of cells is just as interesting as the individual morphological

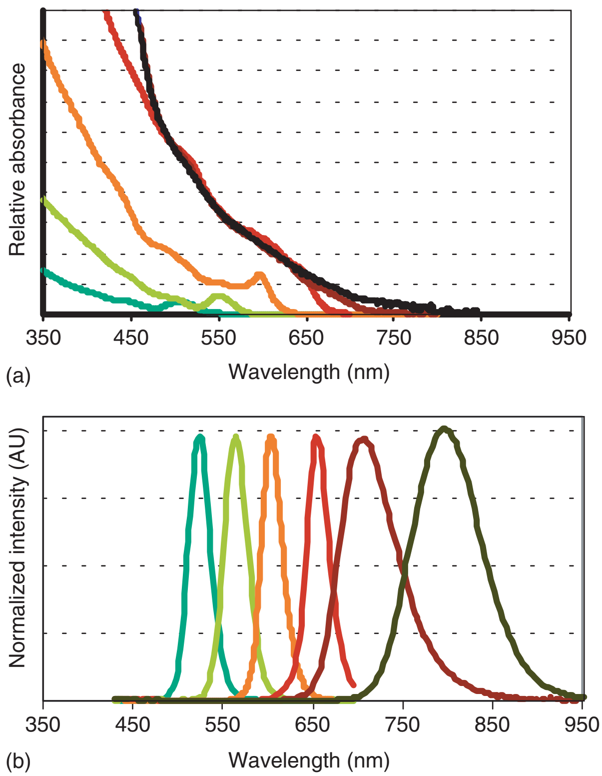


Figure 1 Absorbance (a) and emission (b) spectra of different color Qdot® conjugates

information. In these cases, quantum dots have been used for multiplexed western blotting, resulting in quantitative and sensitive multiprotein analysis in a single western blot. These applications are easily extended to investigate differences in protein expression and modification across a range of tissues and treated cells (Ornberg *et al*., 2005). Multiplexed quantitative analysis of protein modifications with quantum dots allows measurement of the activation of multiple pathways on fixed cells and cell lysates to determine signaling states. This returns more information than has been available from typical cellular analysis methods.

The photostability and brightness of the quantum dots provide investigators with new insights into receptor dynamics and ligand–receptor interactions in live-cell imaging (Dahan *et al*., 2003; Lidke *et al*., 2004). By labeling receptors with antibodies or with the native ligand, these groups discovered new dynamic properties of receptors at the cell surface that could not previously be observed. Dahan *et al*. demonstrated that single glycine receptors in the membrane of live neurons change their diffusion dynamics depending on the cellular site being sampled. At membrane sites away from the synapse, these receptors were highly mobile, though on diffusion into sites containing synaptic proteins they became pinned in place. By visualizing the diffusion of single receptors labeled with quantum dots for 20 min, these investigators were able to see these dynamics in

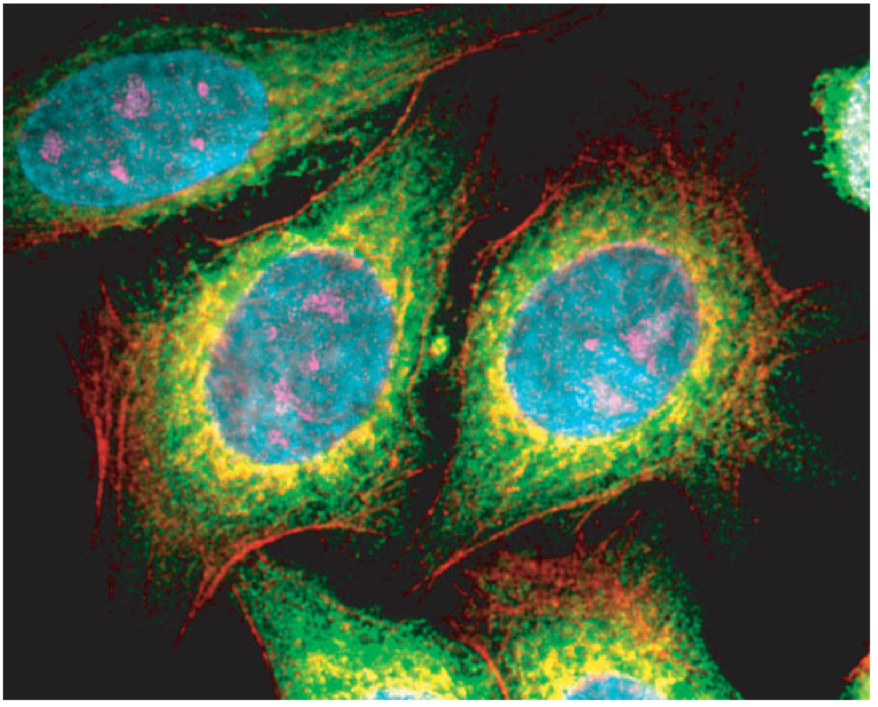


Figure 2 Multicolor immunolabeling with five colors of Qdot® conjugates. Fixed human epithelial cells were stained for Ki-67 (magenta, Qdot 605 anti-rabbit), nuclear antigens (cyan, Qdot 655 antihuman), mitochondria (orange, Qdot 525 anti-mouse), microtubules (green, Qdot 565 anti-rat), and actin filaments (red, Qdot 705 streptavidin). Individual gray scale images were captured with appropriate 20-nm bandpass emission filters, pseudocolored, and merged into a signal color image to show the relative cellular locations of the targets

real time with the knowledge that they were watching the behavior of a single molecule throughout the experiment. Previous studies had demonstrated that there were both highly mobile and immobile forms of this receptor, but they were unable to illustrate the key information that these two forms are the same receptor and that this is simply a context-dependent diffusion change. The ability to track biological events over physiologically relevant timescales is essential to the understanding of the roles of key biological components in signaling pathways, and quantum dots provide a nearly ideal tool for watching the interactions of extracellular components with the signal transduction machinery. Lidke *et al*. tracked quantum dot conjugated epidermal growth factor while interacting with a variety of visible fluorescent protein–tagged erb-B receptors in engineered cell lines. In doing this, they discovered new patterns of receptor dimerization and receptor–ligand trafficking in this important receptor tyrosine kinase family.

Recent applications of quantum dots in in vivo imaging have demonstrated the utility of quantum dots for noninvasive molecular analysis in living animals (Akerman *et al*., 2002; Ballou *et al*., 2004; Kim *et al*., 2004; Voura *et al*., 2004; Gao *et al*., 2004). These investigators demonstrated how quantum dots can be

designed to escape the reticuloendothelial system and can be used to track cells and to label vascular structures. In addition, these materials can target to cell surface antigens or vascular surface markers when conjugated to a targeting antibody or peptide. The strong absorbance and photostability of near-infrared-emitting quantum dots enable the detection of tissues and structures deep in animals, and suggest potential applications in diagnostics, surgery, and biological research. In addition, the ability of high-quality quantum dots to survive embedding, fixation, and typical tissue mounting procedures allows researchers to monitor their experiments noninvasively and subsequently perform detailed tissue analysis and molecular colocalization experiments.

The multiple colors provided by various sizes of quantum dots are currently being exploited in combination to create massively multiplexed systems of cells or beads for parallel analysis (Mattheakis *et al*., 2004; Han *et al*., 2001). Color combinations on beads can generate the spectral equivalent of a bar code, which, when coupled with distinct reporters, allows simultaneous analysis of many analytes in a sample (Figure 3). Initially, this capability was demonstrated with simple hybridization, and for multiplexed single nucleotide polymorphism analysis (Xu *et al*., 2003). This strategy has been developed into a system for multiplexed mRNA analysis and commercialized as the Mosaic™ Gene Expression Assay System by Quantum

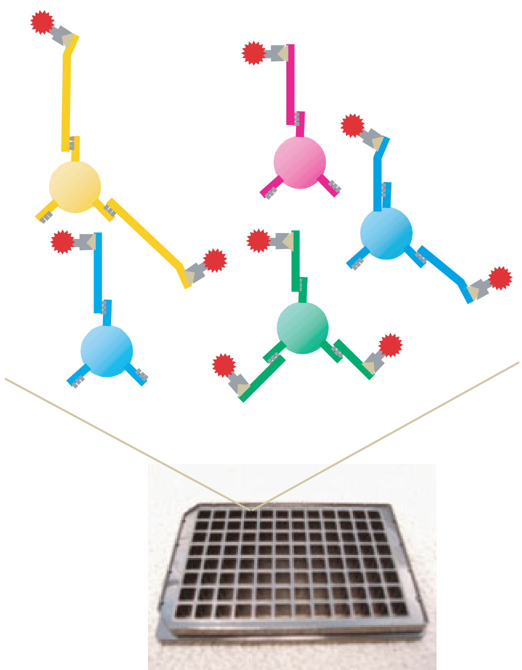


Figure 3 Schematic overview of a Mosaic™ gene expression assay. Beads with different spectral fingerprints are shown in different colors, and are attached to unique sequences. These are hybridized to the mRNA from a sample, and then quantitatively detected with a distinct color. This system is suitable for quantitative analysis of many genes in many samples in a multiwell plate format

Dot Corporation. The key advantage of this system is the large number of codes available. With room for plenty of distinct codes, a large number of internal controls and a high level of replication can be built into the system, allowing statistically valid, robust, and precise measurements to be performed. Building the appropriate quality control into each assay ensures that all of the assay steps are successful. This results in consistently meaningful data, with less time spent duplicating experiments. This level of gene expression analysis allows researchers to reproducibly and quantitatively compare hundreds of genes over thousands of samples.

Quantum dots are an enabling tool for modern fluorescence-based biology, and are poised to help researchers answer many questions raised by newly available proteomic and genomic information. The practice of following single markers has given way to the analysis of multiple cellular systems, along with upstream, downstream, and off-target effects, requiring more information per experiment. When multiplex capability, sensitivity, and stability are needed for looking at proteins or genes in live cells, fixed cells, tissues, animals, or other samples, quantum dot–based probes have pointed researchers to new questions and have given them new tools to get the answers.

Further reading

Derfus AM, Chan WCW and Bhatia SN (2004) Intracellular delivery of quantum dots for live cell labeling and organelle tracking. *Advanced Materials*, **16**(12), 961–966.

Lagerholm BC, Wang M, Ernst LA, Ly DH, Liu H, Bruchez MP and Waggoner AS (2004) Multicolor coding of cells with cationic peptide coated quantum dots. *Nano Letters*, **4**(10), 2019–2022.

References

Akerman ME, Chan WC, Laakkonen P, Bhatia SN and Ruoslahti E (2002) Nanocrystal targeting in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(20), 12617–12621.

Ballou B, Lagerholm BC, Ernst LA, Bruchez MP and Waggoner AS (2004) Noninvasive imaging of quantum dots in mice. *Bioconjugate Chemistry*, **15**(1), 79–86.

Bruchez M Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science*, **281**(5385), 2013–2016.

Chan WC and Nie S (1998) Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science*, **281**(5385), 2016–2018.

Dahan M, Levi S, Luccardini C, Rostaing P, Riveau B and Triller A (2003) Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking. *Science*, **302**(5644), 442–445.

Gao X, Cui Y, Levenson RM, Chung LW and Nie S (2004) In vivo cancer targeting and imaging with semiconductor quantum dots. *Nature Biotechnology*, **22**(8), 969–976.

Han M, Gao X, Su JZ and Nie S (2001) Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nature Biotechnology*, **19**(7), 631–635.

Hotz CZ (2005) Applications of quantum dots in biology: an overview. *NanoBiotechnology Protocols*, in press.

Kim S, Lim YT, Soltesz EG, De Grand AM, Lee J, Nakayama A, Parker JA, Mihaljevic T, Laurence RG, Dor DM, *et al*. (2004) Near-infrared fluorescent type II quantum dots for sentinel lymph node mapping. *Nature Biotechnology*, **22**(1), 93–97.

Lidke DS, Nagy P, Heintzmann R, Arndt-Jovin DJ, Post JN, Grecco HE, Jares-Erijman EA and Jovin TM (2004) Quantum dot ligands provide new insights into erbB/HER receptor-mediated signal transduction. *Nature Biotechnology*, **22**(2), 198–203.

Mattheakis LC, Dias JM, Choi YJ, Gong J, Bruchez MP, Liu J and Wang E (2004) Optical coding of mammalian cells using semiconductor quantum dots. *Analytical Biochemistry*, **327**(2), 200–208.

Ornberg RL, Liu H and Harper TF (2005) Western blot analysis with quantum dot fluorescence technology: a sensitive and quantitative method for multiplexed proteomics. *Nature Methods*, **2**(1), 79–81.

Voura EB, Jaiswal JK, Mattoussi H and Simon SM (2004) Tracking metastatic tumor cell extravasation with quantum dot nanocrystals and fluorescence emission-scanning microscopy. *Nature Medicine*, **10**(9), 993–998.

Wu X, Liu H, Liu J, Haley KN, Treadway JA, Larson JP, Ge N, Peale F and Bruchez MP (2003) Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots. *Nature Biotechnology*, **21**(1), 41–46.

Xu H, Sha MY, Wong EY, Uphoff J, Xu Y, Treadway JA, Truong A, O’Brien E, Asquith S, Stubbins M, *et al*. (2003) Multiplexed SNP genotyping using the Qbead system: a quantum dot-encoded microsphere-based assay. *Nucleic Acids Research*, **31**(8), e43.

Basic Techniques and Approaches

Immunofluorescent labeling and fluorescent dyes

Peter Watson, Krysten J. Palmer and David J. Stephens
University of Bristol, Bristol, UK

1. Introduction

Immunofluorescence microscopy is an important experimental tool for investigating the distribution of soluble and structural proteins within cells. This technique can give clues to the function of a poorly understood protein and enable direct comparisons of a protein's localization in various cell types and tissues at different stages of the cell cycle, during development or under stress or stimulation. Successful immunofluorescence requires fixation and permeabilization conditions that preserve the in vivo distribution of the protein(s) under examination. The quality of antibodies used is critical to ensure both specificity and high signal-to-noise on detection. Choice of direct versus indirect methods, amplification method if any, fluorescent dye conjugate, and detection method will all determine the success of immunofluorescence experiments, and careful consideration needs to be given to each. The complexity of protocols employed will also largely determine the applicability of the methods to high-throughput analyses. We also consider the application of technologies that, although not commonly employed at present, have great potential for the future; these include the use of recombinant antibodies and immunofluorescence in the context of live cell imaging. Here, we discuss the possible approaches and highlight some of the pros and cons of their use in the specific detection of molecules in cells and tissues.

2. Fixation protocols

The main aim of fixation is to freeze a sample in a state that is as close to "real life" as possible. In essence, this would render immunofluorescence impossible, as the cell membrane creates an effective barrier to this approach, and so certain compromises must be made between keeping cellular structures intact and having a working specimen in which to identify those structures. Soluble proteins have the unfortunate tendency to redistribute and/or become differentially extracted, while membrane proteins and cytoskeletal elements are subject to a lesser change.

Fixation and permeabilization methods must therefore be carefully selected and best suited to the protein(s) in question, considering size and charge as well as cell structure and type. Critical to the application of immunofluorescence to proteomic experiments is the reproducibility of the fixation technique used and indeed of the entire labeling protocol. One must ensure reproducibility across a large number of samples, particularly for comparative analyses of protein localization or readouts of cell function.

The two main processes of cellular fixation for subsequent immunofluorescence are chemical cross-linking (e.g., paraformaldehyde or glutaraldehyde) or protein precipitation by an organic solvent (e.g., methanol or acetone). The quality of cellular structures that remain following fixation can vary, depending on the method of fixation and the initial cell type used, and so this must be kept in mind. Excellent discussions of fixation methods can be found in Allan (1999) and Melan and Sluder (1992).

Aldehyde cross-linking forms methylene bridges between side groups. Formaldehyde (or paraformaldehyde) is most commonly used for immunofluorescence microscopy as it is readily available and is soluble in water. Glutaraldehyde was first used as a fixative for histological EM (electron microscopy) studies (Sabatini *et al*., 1963) and preserves internal structures well, presumably due to its ability to cross-link neighboring protein molecules via its two reactive groups. However, glutaraldehyde fixation may be problematic for immunofluorescence owing to an observed increase in autofluorescence of the sample and an inability of the antibody to reach the desired epitope because of the strong cross-linking effect. Other fixatives including ethylene glycol bis(succinimidylsuccinate) and dithiobis(succinimidylproprionate) are less commonly employed as they are poorly soluble in water. Both provide excellent fixation of microtubules and membranes and maintain antigenicity of many glutaraldehyde sensitive epitopes (Allan, 1999).

Permeabilization of chemically cross-linked cells is necessary for access to internal epitopes. Triton X-100 (0.1%) is regularly used as a nondenaturing detergent to permeabilize the cell membrane. Digitonin and saponin are also used. Analysis of a range of concentrations of these detergents is best to determine empirically the ideal conditions for a particular experiment. If the antibody only recognizes a denatured epitope, an incubation step with a denaturant (i.e., guanidine hydrochloride) can be included.

The most widely used alternative to aldehyde fixation is the use of organic solvents. Solvent fixation is very quick, and has the advantage of needing no permeabilization step or quenching of unreacted chemicals following fixation. Many intracellular structures are better preserved following one or other method, and similarly, antigenicity can be different following each method. Consideration must also be given to whether a particular fixation method might redistribute an antigen with a cell. Detailed protocols for most of these techniques can be found in Allan (1999).

3. Generating antibodies

Polyclonal, monoclonal, and recombinant antibodies all have their relative merits and disadvantages. An excellent discussion of antibody production methods

for proteomics can be found elsewhere (Bradbury *et al*., 2003a). Successful immunofluorescence is wholly dependent on the specificity of the primary antibody used. In all experiments, titration of each preparation of antibody is important in order to determine the optimal concentration for fluorescence microscopy. Polyclonal antibodies often provide enhanced signals compared to monoclonal equivalents in immunofluorescence, and affinity purification can ensure high specificity. Polyclonal antibodies are also cheaper to produce and can be readily made in sufficient quantity for most purposes. However, unlike with monoclonal antibodies, the continued supply necessary for long-term projects is not guaranteed. Although an antibody may work for immunoprecipitation or immunoblotting, it does not automatically follow that it will be suitable for immunofluorescence. Optimally, a number of antibodies to different epitopes on the same protein are compared in order to identify the most suitable for immunofluorescence.

4. Recombinant antibodies

The generation of recombinant antibodies using phage display technology is now relatively simple and can be readily scaled up using high-throughput technology (Hust and Dubel, 2004). The process involves the selection and purification of recombinant antibody fragments (single chain fragment V, scFv). A library of antibody fragments presenting bacteriophage undergoes several rounds of affinity selection (panning) against the protein of choice (Figure 1(a) shows an example of panning for antibodies directed against the small GTPase, Rab6). Once a bacteriophage with a positive antibody fragment has been identified, the antibody fragment can be expressed and purified, effectively in unlimited quantity. The simplicity and speed of this approach also makes it amenable to the production of conformation-specific antibodies directed against a single protein (Nizak *et al*., 2003), for example, one that preferentially recognizes the GTP-bound form of Rab6 and has a low affinity for the GDP-bound form (Nizak *et al*., 2003). These antibodies are suitable for immunofluorescence (Figure 1b and d show specific localization of Rab6GTP to the Golgi, consistent with its known function). Immunofluorescence need not be restricted to fixed cells and combining recombinant antibody and green fluorescent protein (GFP) technology permits the use of these reagents in live cell imaging experiments (Figure 1c) (Nizak *et al*., 2003). Thus, antigens can be observed in situ without the need to fix and permeabilize cells and eliminating the need for secondary detection.

There is a greater chance of finding antibodies to a selected epitope, and an increased chance of selecting a high affinity epitope, as the size of the initial phage library increases; this underlies the key limitation of the technology. Higher affinity antibodies can be created by mutation of the initial isolated phage followed by further rounds of selection. The potential of this approach, particularly for large-scale proteomic approaches (e.g., its use as a fluorescent biosensor) is huge, and will doubtless undergo considerable development in future years.

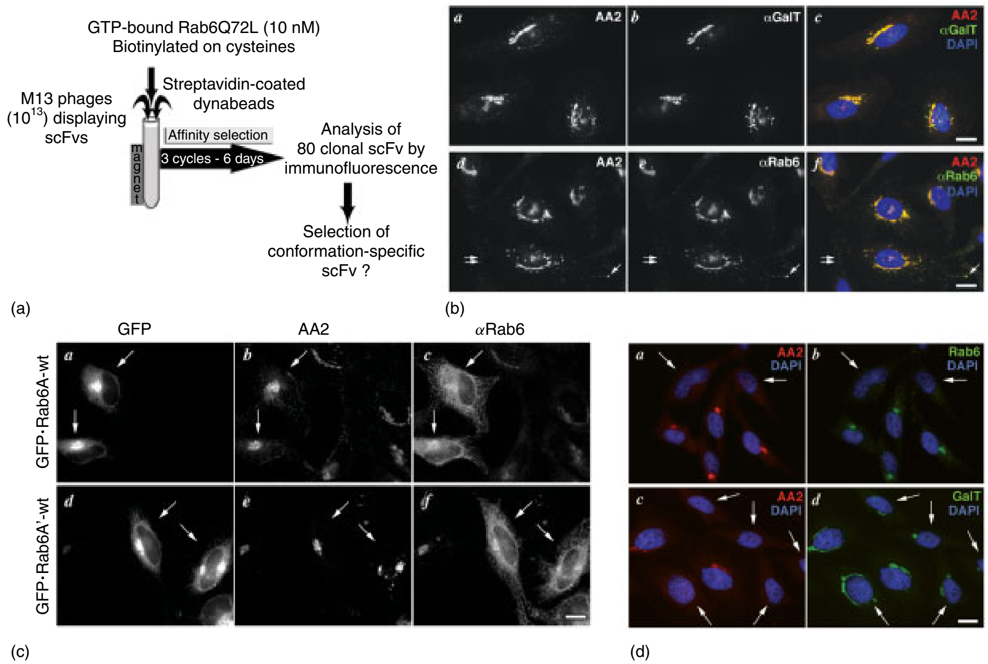


Figure 1 Recombinant antibodies for immunofluorescence. (a) The production of specific recombinant antibodies is achieved by sequential panning of a phage library against a subcellular fraction or protein of interest (here, recombinant Rab6-GTP). (b) This allows localization of endogenous proteins in living, unfixed cells. Here, antibody AA2 shows specific localization of endogenous Rab6-GTP to the Golgi as shown by colocalization with galactosyltransferase (GalT). (c) These antibodies can be produced such that they recognize only specific conformations of a target protein, here, the GTP-bound conformation of the small GTPase Rab6. Differential localization of total versus GTP-bound pools of Rab6 using a GFP fusion of AA2. (d) Specificity is confirmed by showing a loss off immunofluorescent labeling after depletion of Rab6A using RNA interference (Reprinted with permission from Nizak *et al*. (2003). Copyright (2003) AAAS)

5. Secondary antibodies

Most frequently, immunofluorescence is performed indirectly; that is, the primary antibody used is detected by a separate fluorescently labeled secondary antibody. There are three main considerations for the secondary antibody: the species that it was raised in, the level of cross adsorption against other species it has gone through during the purification process, and the fluorophore that it is linked to. Cross adsorption against other species reduces the possibility of spurious recognition of other antibody preparations used for colabeling, and should result in lower background binding to your chosen cell type. Careful choices need to be made depending on host species for the generation of these antibodies. For example, use of primary antibody raised in the goat would preclude the use of secondary antibodies raised in the same species. Fortunately, major manufacturers provide suitable solutions to this through the availability of antibodies raised in the goat and the donkey.

6. Choice of fluorescent dye

Direct immunofluorescence can be done by conjugating primary antibodies (whole molecules or Fab' fragments) directly to fluorophores. Many manufacturers provide kits to achieve this through succinimidyl esters, isothiocyanates, sulfonyl chlorides, tetrafluorophenyl esters, or thiol-reactive probes. Tetrafluorophenyl esters produce stable carboxamide bonds and are less susceptible to hydrolysis than succinimidyl esters. Isothiocyanate conjugates deteriorate over time. Similarly, these reactive groups are used to conjugate dyes to secondary antibodies for indirect immunofluorescence.

For multiple labeling experiments, the fluorophore used depends upon limiting the crossover between emission spectra, and the visualization methods (i.e., the excitation and emission filters) that can be used. New approaches to spectral separation through hardware and software developments (Zimmermann *et al*., 2002) are now enabling the simultaneous analysis of more and more probes in a single cell or tissue. These techniques are applicable to both confocal and wide-field imaging.

Previously, commonly used dyes such as fluorescein and rhodamine suffered from low quantum yield (brightness) and rapid photobleaching. Newer dyes such as the Alexa Fluor dyes from Molecular Probes or cyanine dyes (CyDye Fluors) from Amersham Biosciences are more photostable than previously used dyes such as fluorescein isothiocyanate (FITC). This allows repeated viewing, time-lapse imaging of living cells, and the production of three-dimensional (3D) data sets with reduced photobleaching. Conjugation at high molecular ratios of the fluorophore to the protein can lead to fluorescence quenching, presumably owing to dye–dye interactions (Randolph and Waggoner). It is possible to attain a higher dye-protein ratio with the Alexa Fluor dyes before quenching becomes noticeable (Berlier *et al*., 2003). These dyes are also available with a diverse array of spectral characteristics, permitting the use of two, three, four, or more dyes simultaneously.

It is now possible to combine the simplicity and flexibility of fluorescent imaging by light microscopy with the ultrastructural resolution provided by electron microscopy (EM). This can be done using a combination of GFP and immunogold labeling (Mironov *et al*., 2000) or through the use of probes such as Fluoronanogold that contain both fluorescent labels for light microscopy and a Nanogold cluster for EM (Takizawa and Robinson, 2000). The use of the Nanogold label avoids the problem of fluorescent quenching seen with colloidal gold conjugates. (For further information, see **http://www.nanoprobes.com**.) The labor-intensive nature of EM means its use in proteomic experiments is likely to be limited.

7. Quantum dots

Quantum dots (Qdots) consist of a semiconductor nanocrystal core, a semiconductor shell, and a polymer coat (Seydel, 2003) and are fluorescent as a result of the quantum confinement effect. The polymer coat can be conjugated to biomolecules, including streptavidin and antibodies, allowing binding specificity while retaining the optical qualities of the crystal (Chan and Nie, 1998). Qdots are stable and

nontoxic inside the cytosolic compartment (Dubertret *et al*., 2002), and can be used for immunofluorescent labeling of cells and tissues (Wu *et al*., 2003; Jaiswal *et al*., 2003). A key problem with quantum dots is their bioincompatibility (high nonspecific binding to the sample) and physical size. Qdots are discussed in more detail elsewhere in this volume (*see* Article 57, **Quantum dots for multiparameter measurements**, Volume 5).

8. Image acquisition and high-throughput technologies

Imaging of fluorescent samples can be done in a number of ways including scanning confocal microscopy and wide-field imaging (see Stephens and Allan, (2003) for a review) as well as flow cytometry. Imaging of multiple fluorophores, time-lapse imaging and multidimensional imaging (3D and 4D), deconvolution, and image reconstruction can all add significantly to specific experiments and these topics are dealt with elsewhere in this section. Immunofluorescence labeling and detection are highly amenable to automation scaled to 96- or 384-well formats (Bradbury *et al*., 2003b). Robotic handling can be applied to the fixation and labeling steps, and automated object recognition algorithms (Liebel *et al*., 2003; Danckaert *et al*., 2002) now permit the detection of subcellular localization across a population of cells. A recent example is the assembly of a system for rapid automated screening of cDNAs (Liebel *et al*., 2003), which has been applied to high content screening of processes such as intracellular traffic using immunofluorescence to define protein function.

9. Summary

The application of immunofluorescence toward diagnostics and studies of protein function depend largely on reliable, reproducible methodologies. The development of antibody chips will facilitate much wider analyses of protein function on a proteomic scale, and these can also be used for immunofluorescence localization (Wang, 2004). The possibilities of recombinant antibodies, live cell imaging, high throughput, and high content screening mean that immunofluorescence of both fixed and living cells will continue to be a valuable tool to define protein function.

References

Allan V (1999) *Protein Localization by Fluorescence Microscopy*, Oxford University Press: Oxford.

Berlier JE, Rothe A, Buller G, Bradford J, Gray DR, Filanoski BJ, Telford WG, Yue S, Liu J, Cheung CY, *et al*. (2003) Quantitative comparison of long-wavelength Alexa Fluor dyes to Cy dyes: fluorescence of the dyes and their bioconjugates. *Journal of Histochemistry and Cytochemistry*, **51**, 1699–1712.

Bradbury A, Velappan N, Verzillo V, Ovecka M, Chasteen L, Sblattero D, Marzari R, Lou J, Siegel R and Pavlik P (2003a) Antibodies in proteomics I: generating antibodies. *Trends in Biotechnology*, **21**, 275–281.

Bradbury A, Velappan N, Verzillo V, Ovecka M, Chasteen L, Sblattero D, Marzari R, Lou J, Siegel R and Pavlik P (2003b) Antibodies in proteomics II: screening, high-throughput characterization and downstream applications. *Trends in Biotechnology*, **21**, 312–317.

Chan WC and Nie S (1998) Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science*, **281**, 2016–2018.

Danckaert A, Gonzalez-Couto E, Bollondi L, Thompson N and Hayes B (2002) Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, **3**, 66–73.

Dubertret B, Skourides P, Norris DJ, Noireaux V, Brivanlou AH and Libchaber A (2002) *In vivo* imaging of quantum dots encapsulated in phospholipid micelles. *Science*, **298**, 1759–1762.

Hust M and Dubel S (2004) Mating antibody phage display with proteomics. *Trends in Biotechnology*, **22**, 8–14.

Jaiswal JK, Mattoussi H, Mauro JM and Simon SM (2003) Long-term multiple color imaging of live cells using quantum dot bioconjugates. *Nature Biotechnology*, **21**, 47–51.

Liebel U, Starkuviene V, Erfle H, Simpson JC, Poustka A, Wiemann S and Pepperkok R (2003) A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Letters*, **554**, 394–398.

Melan MA and Sluder G (1992) Redistribution and differential extraction of soluble proteins in permeabilized cultured cells. Implications for immunofluorescence microscopy. *Journal of Cell Science*, **101**(Pt 4), 731–743.

Mironov AA, Polishchuk RS and Luini A (2000) Visualizing membrane traffic *in vivo* by combined video fluorescence and 3D electron microscopy. *Trends in Cell Biology*, **10**, 349–353.

Nizak C, Monier S, del Nery E, Moutel S, Goud B and Perez F (2003) Recombinant antibodies to the small GTPase Rab6 as conformation sensors. *Science*, **300**, 984–987.

Sabatini DD, Bensch K and Barrnett RJ (1963) Cytochemistry and electron microscopy. The preservation of cellular ultrastructure and enzymatic activity by aldehyde fixation. *Journal of Cell Biology*, **17**, 19–58.

Seydel C (2003) Quantum dots get wet. *Science*, **300**, 80–81.

Stephens DJ and Allan VJ (2003) Light microscopy techniques for live cell imaging. *Science*, **300**, 82–86.

Takizawa T and Robinson JM (2000) FluoroNanogold is a bifunctional immunoprobe for correlative fluorescence and electron microscopy. *Journal of Histochemistry and Cytochemistry*, **48**, 481–486.

Wang Y (2004) Immunostaining with dissociable antibody microarrays. *Proteomics*, **4**, 20–26.

Wu X, Liu H, Liu J, Haley KN, Treadway JA, Larson JP, Ge N, Peale F and Bruchez MP (2003) Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots. *Nature Biotechnology*, **21**, 41–46.

Zimmermann T, Rietdorf J, Girod A, Georget V and Pepperkok R (2002) Spectral imaging and linear un-mixing enables improved FRET efficiency with a novel GFP2-YFP FRET pair. *FEBS Letters*, **531**, 245–249.

Quantitative image analysis and the Open Microscopy Environment

Jason R. Swedlow, Chris Allan, Andrea Falconi and Jean-Marie Burel
University of Dundee, Dundee, UK

1. Introduction

Microscopy has traditionally been a technique devoted to understanding the structure and dynamics of the cell. The combination of genetically encoded fluorescence based on fluorescent proteins (Tsien, 1998; *see also* Article 50, **Using photoactivatable GFPs to study protein dynamics and function**, Volume 5) and the availability of easy-to-use digital imaging systems has revolutionized applications for microscopy in modern biology. A vast array of imaging modes is already available, with more on the way (*see* Article 48, **Probing cellular function with bioluminescence imaging**, Volume 5, Article 51, **FRET-based reporters for intracellular enzyme activity**, Volume 5, Article 54, **Probing protein function in cells using CALI**, Volume 5, Article 55, **Imaging protein function using fluorescence lifetime imaging microscopy (FLIM)**, Volume 5, Article 56, **Elucidating protein dynamics and function using fluorescence correlation spectroscopy (FCS)**, Volume 5 and Article 57, **Quantum dots for multiparameter measurements**, Volume 5). Most critically, microscopy is not just visualization, or "seeing", but now also quantification, often at the molecular level (Gerlich *et al*., 2001; Swedlow, 2003). Quantitative imaging can be used to characterize the signal and noise characteristics of different microscopes (Murray, 1998; Swedlow *et al*., 2002; Zucker and Price, 2001), and it also allows image-based assays, where the localization and movement of macromolecules can be measured in fixed and living cells (Lippincott-Schwartz *et al*., 2001; *see also* Article 53, **Photobleaching (FRAP/FLIP) and dynamic imaging**, Volume 5 and Article 56, **Elucidating protein dynamics and function using fluorescence correlation spectroscopy (FCS)**, Volume 5). Taken to the extreme, image-based assays can be expanded to assay the effects of libraries of small molecule inhibitors (Yarrow *et al*., 2003; Straight *et al*., 2003), or libraries of genes or even whole genomes (Simpson *et al*., 2000; Kiger *et al*., 2003; *see also* Article 52, **High content screening**, Volume 5).

A major problem in these assays is the systematic analysis of the large amounts of data that are produced. There are two parts to this problem. First, data must be stored in a secure and accessible fashion, with redundant backups. Second, while there is much emphasis on developing image acquisition systems and analysis algorithms for cell-based assays, systematic analysis requires an informatics infrastructure that maintains relations between binary image data, experimental and image metadata, and any processing and analytic results. In this article, we discuss our efforts toward both of these goals.

2. Secure and available storage of image data

Historically, high-density fixed storage devices have not been affordable by most academic laboratories, so most image data storage has been stored on whatever high-density removable media devices were available. For instance, various optical disc formats were available during the late 1980s and these provided a convenient method for storing data. Unfortunately, the formats used for these devices became obsolete within a few years, and archived data effectively was lost on unreadable media. More recently, compact disks (CDs) and digital versatile disks (DVDs) have been used; however, these are not permanent media and usually are not indexed with the files they contain. The result is usually a mass of unindexed archived data, spread across a large number of CDs or DVDs. When students leave a lab, these collections of disks and all their precious data are placed in a box and stored, often in the principal investigator's office, to gather dust as the media steadily degrades.

Recently, large-scale data storage and archiving systems have become available and it is now possible to purchase multiterabyte data storage systems at reasonable cost. However, just like any other data storage medium, these systems have fixed capacity and, therefore, will eventually fill to capacity. Scalability is a crucial consideration to ensure that management burden and costs do not also scale each time the storage system is expanded.

To solve this problem, we have employed a hierarchical storage system, consisting of separate pools of storage media with different price/performance ratios that ensure that the most recently acquired and accessed data is immediately accessible. An automatic archiving system checks the on-line filesystem every five minutes for all files that have not been accessed after a certain date and moves everything except the first 0.5 MB of the file off to a tape archive. Migrated files are still visible in a user's directory, and are automatically returned from the archive if accessed. Files that are either forgotten or never used are then automatically archived on more cost-effective media. While this strategy involves a time penalty for infrequently accessed files (data transfer only requires a few seconds for >500 MB, but finding and loading tapes can take up to 90 s), the advantage is that on-line filesystem space is reserved for actively used files. Moreover, if systems are bought strategically, the different storage media pools can be expanded where it is necessary to expand the overall capacity of the storage system. On-line disk arrays currently cost \$3000–12 000/Tbyte, while tape storage or lower performance disks cost \$100–500/Tbyte. While any efficient storage system will include a facility for a "permanent" archive to off-line storage, the hierarchical approach that we have implemented allows data

to be kept available for analysis and further experiments at a price that is affordable by most laboratories. A diagram of our system is shown in Figure 1.

It should be noted that the infrastructure to use such a solution requires a significant up-front investment for servers, fiber channel switches, and so on, and significant expertise for configuration. *The most important investment we have made is in the skilled personnel to run these facilities!!!*

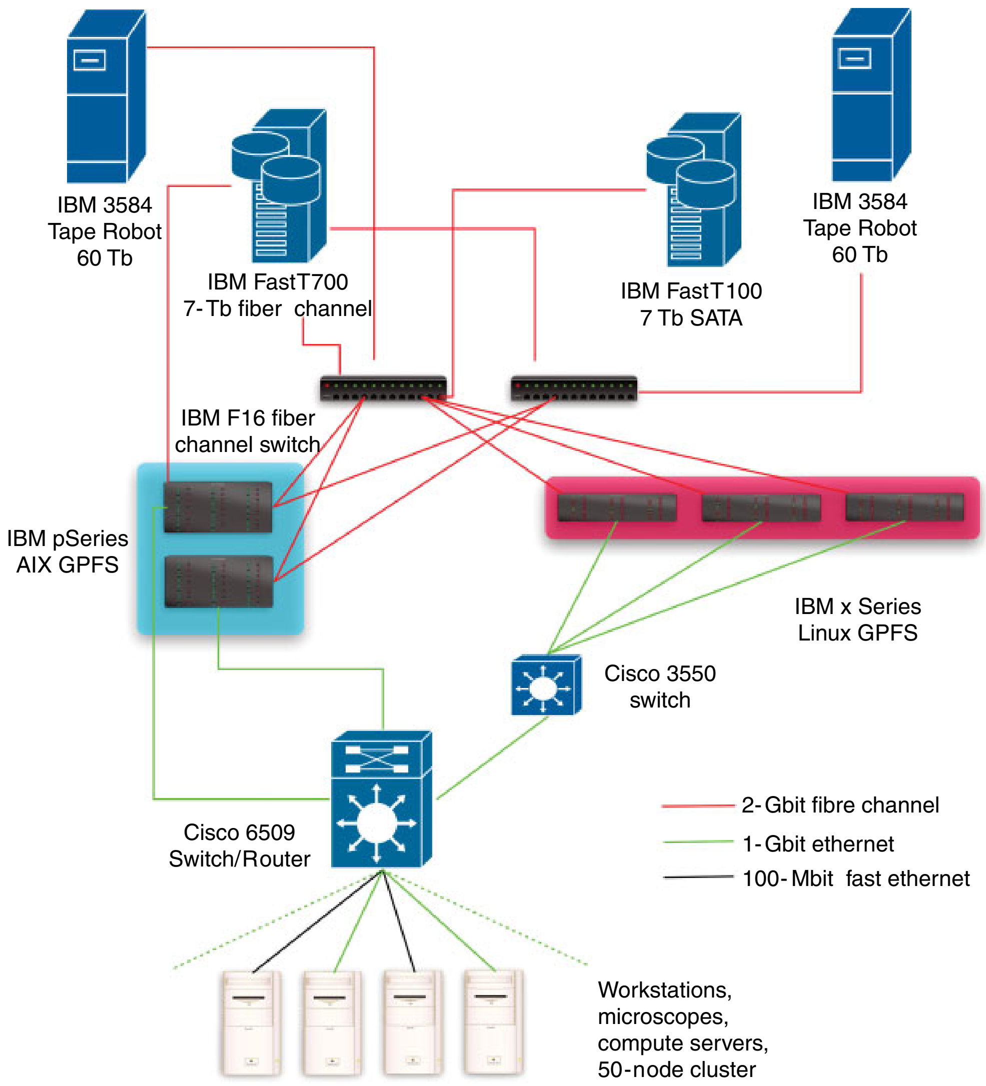


Figure 1 Hierarchical data storage system. The general layout and infrastructure of the Light Microscopy Facility data storage system at the Wellcome Trust Biocentre, University of Dundee. The specifics of network connections and redundancy and hardware configuration are not included for clarity. The hierarchical file management is run by Tivoli Storage Manager. Most of our storage equipment is provided by IBM, but similar solutions are available from other vendors. The types of network connections between the different hardware devices is indicated

3. Toward image informatics – the Open Microscopy Environment

Biological image data is "content-rich": the distribution of biological molecules are recorded, usually within the context of spatial location at a specific time, and often relative to the distribution of at least a few other molecules. Methods that measure the environment of a fluorophore, its binding affinity, molecular interactions, and chemical neighborhood only add to this complexity. Unlike genomic data, where the output of a DNA sequencing system ("... GAGCTC ... ") directly identifies a specific sequence, the mapping of image pixel values to a "result" requires much more contextual knowledge – the contrast method of imaging (e.g., fluorescence, DIC, etc.), the mode (e.g., fluorescence lifetime, resonance energy transfer, etc.), and large numbers of descriptors about the structure of the data (e.g., time lapse, illumination, objective lens) (Andrews *et al*., 2002). Given the pace of methodology and application development in biological imaging, it has been impossible to develop a single data format standard that would support the needs of the whole field. More fundamentally, an "image" is only a small part of the data relevant to an imaging experiment. Knowledge of the experiment (cell type, mutant, staining method, etc.) as well as the results of any analyses (rates, intensity changes, etc.) all represent critical measurements that must be linked to all other data elements. Without these links, loss of one data element (e.g., the cell line used for a specific image, the time lapse interval) can invalidate all of the data related to it.

To solve this problem, we are building an open-source image data, metadata, and analysis management system known as the Open Microscopy Environment (OME; **http://openmicroscopy.org**) (Swedlow *et al*., 2003). OME is specifically designed for the storage and analysis of digital microscope image data and metadata. The major focus of OME is not on creating novel analysis algorithms, but instead on development of a structure that ultimately allows any application to read and use any data associated with or generated from digital imaging microscopes. The critical role for metadata and analytical results in biological imaging is discussed in more detail elsewhere (**http://openmicroscopy.org/applications**). OME is designed for installation to support the work in a single lab or imaging facility, although there are larger use cases suggested by others to support the centralized provision of large numbers of images to the research community (Schuldt, 2004).

The heart of the OME system is the OME Data Model, which provides the framework for all the data relationships managed by OME. This model is instantiated in two ways. The OME XML File casts the OME Data Model into a self-describing entity using XML, the standard methodology for providing access to data. The Data Model and the XML file are both described at **http://openmicroscopy.org/api/xml/** and are the subject of a future publication. In addition, the Data Model is the foundation of a data warehousing, management, visualization, and tracking system, known simply as "OME", the basic layout of which is shown in Figure 2 (see Swedlow *et al*., 2003 for more info). A key concept in our implementation is the possibility for local extensibility. A number of basic data elements will be shared between all data (User Name, Image Dimensions, etc.), but local customization will always be required. The use of XML intrinsically supports this extensibility such that the Data Model, the OME XML File, and an instantiation of OME software

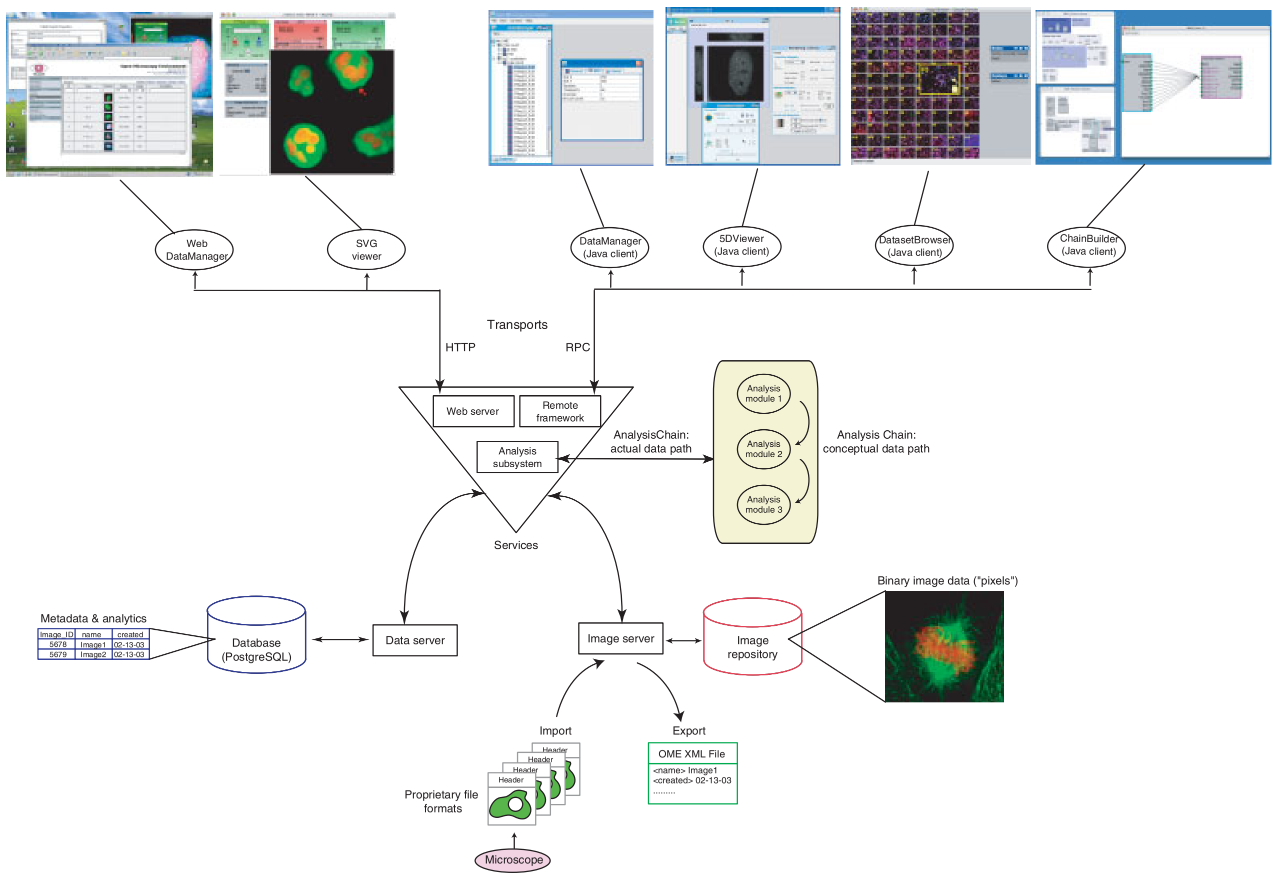


Figure 2 General layout of OME. A diagram showing the different parts of the OME server and interface systems. This figure is not a software architecture document and serves only to define the general functionality of the system. Imported data from a microscope system is stored within a database and an image repository (see text). The OMEIS and OMEDS servers manage all access to the image repository and database, respectively. Data is accessed for visualization and analysis via a services layer that supports HTTP and XML-RPC. Currently, OME2.2 is distributed with support for web browser access and a stand-alone Java-based image visualization and analysis tool ("Shoola") and a DatasetBrowser for viewing large groups of images. Screenshots of these tools are shown at the top of the figure

can all be updated to adapt to specific needs. The details of this are described at **http://openmicroscopy.org/api/xml/**.

In our current system, OME2.2 (**http://cvs.openmicroscopy.org.uk**), image data is imported from a user's microscope using OME tools that read specified commercial file formats. OME considers an "image" as a five-dimensional data structure (space, time, and channel) (Andrews *et al*., 2002). All relevant metadata is recovered from the user or the incoming data file and entered into the database using the OME Data Server (OMEDS). The actual pixel values form a large amount of data and are stored separately in a specified, protected filesystem, the OME Repository, and accessed through the OME Image Server (OMEIS). OMEDS and OMEIS can exist on the same or different servers as necessary. During import, a large number of statistical measurements (e.g., minimum, maximum, and mean pixel values) are performed and stored in the database.

All access to the data within OME is made through a transport layer that supports a variety of access methods. Data visualization and management using a web browser are supported using standard HTTP protocols. A more flexible interface is provided using "Shoola" and OME-JAVA, application programming interfaces (APIs) that support access from Java remote client programs. More general access to the OME Services is provided over XML-RPC. To take advantage of these systems, we have also built a number of Java-based client applications that provide cross-platform access to OME for visualization, data management, and analysis. Examples of the web browser and Java interfaces are available at **http://openmicroscopy.org/getting-started/**.

A major focus for OME is the integration of image processing and analysis into a single data model. This affords a significant convenience in that all data pertinent to an experiment is maintained in a coherent or linked manner. More profoundly, this approach generates a transactional record of all processing and analytic events that occur on an image. In practice, OME captures all of the input and output parameters for an analysis, as well as the specific results – a new image, rates, sizes, and so on. Each analysis event is uniquely identified, so that the consequences of each analysis, and each parameter change for a given analysis, can be accessed.

4. Image informatics and OME – the future

The OME project was initiated simply on the basis of a need for large-scale data management that no one else had filled. In the past two years, the OME project has successfully transformed from a prototype system – OME1.0 – to a fully integrated multisite open-source software development project. There are currently eight full-time developers working on OME at the University of Dundee (Dundee, UK), National Institutes of Health (Baltimore, USA), Massachusetts Institute of Technology (Cambridge, USA), Deutsches Krebsforschungszentrum (Heidelberg, Germany), and the University of Wisconsin, Madison (Madison, USA) (see **http://openmicroscopy.org/about/**). All work is coordinated through a CVS server running at the University of Dundee (**http://cvs.openmicroscopy.org.uk**). The project has met a number of milestones over the last two years, released in two

major releases, OME2.0 (released June 2003) and OME2.2 (released June 2004), with our final release in this series, OME2.4, scheduled for March 2005. These releases will finish our initial efforts to define what an image data management tool can do. For the last six months, OME2.2 has been tested with real use cases covering large-scale image assays for cell biology and high-throughput screening. These tests have shown the founding concepts and implementation in the OME Data Model to be sound, but have revealed a number of limitations to the system. For example, the current version of OME only supports very limited abilities to delete data, once it exists within OME. This has been deliberate, as the model and implementation errs on the side of caution on this issue. Nonetheless, there are real cases where data can be archived and then deleted from a database, and this functionality must be supported. More generally, our interface and server structure has been designed as an initial prototype. With our current testing experience, there are now clear avenues for substantially improving the performance of the system, and the project will begin implementing these changes in 2005.

In the future, OME can serve as a data management service for single labs and as a portal to a Grid-enabled data management network. This technology promises to share data, processing algorithms, and processing power across many different sites. While the tools for these ideas are still developing, OME's provision of a managed data system at a lab or imaging facility will provide the basis for a larger network of image data sharing and analysis.

Acknowledgments

We gratefully acknowledge helpful discussions with our academic and commercial partners (Swedlow, 2004) and members of the Swedlow laboratory. Research in the authors' laboratory is supported by a grant from the Wellcome Trust (Ref 068046 to J. R. S.). J. R. S. is a Wellcome Trust Senior Research Fellow.

References

Andrews PD, Harper IS and Swedlow JR (2002) To 5D and beyond: Quantitative fluorescence microscopy in the postgenomic era. *Traffic*, **3**, 29–36.

Gerlich D, Beaudouin J, Gebhard M, Ellenberg J and Eils R (2001) Four-dimensional imaging and quantitative reconstruction to analyse complex spatiotemporal processes in live cells. *Nature Cell Biology*, **3**, 852–855.

Kiger A, Baum B, Jones S, Jones M, Coulson A, Echeverri C and Perrimon N (2003) A functional genomic analysis of cell morphology using RNA interference. *Journal of Biology*, **2**, 27.

Lippincott-Schwartz J, Snapp E and Kenworthy A (2001) Studying protein dynamics in living cells. *Nature Reviews. Molecular Cell Biology*, **2**, 444–456.

Murray JM (1998) Evaluating the performance of fluorescence microscopes. *Journal of Microscopy*, **191**, 128–134.

Schuldt A (2004) Images to reveal all? *Nature Cell Biology*, **6**, 909.

Simpson JC, Wellenreuther R, Poustka A, Pepperkok R and Wiemann S (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Reports*, **1**, 287–292.

Straight AF, Cheung A, Limouze J, Chen I, Westwood NJ, Sellers JR and Mitchison TJ (2003) Dissecting temporal and spatial control of cytokinesis with a myosin II Inhibitor. *Science*, **299**, 1743–1747.

Swedlow JR (2003) Quantitative fluorescence microscopy and image deconvolution. *Methods in Cell Biology*, **72**, 349–367.

Swedlow JR (2004) **http://www.openmicroscopy.org/about/partners.html**.

Swedlow JR, Goldberg I, Brauner E and Sorger PK (2003) Informatics and quantitative analysis in biological imaging. *Science*, **300**, 100–102.

Swedlow JR, Hu K, Andrews PD, Roos DS and Murray JM (2002) Measuring tubulin content in *Toxoplasma gondii*: A comparison of laser-scanning confocal and wide-field fluorescence microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2014–2019.

Tsien RY (1998) The green fluorescent protein. *Annual Review of Biochemistry*, **67**, 509–544.

Yarrow JC, Feng Y, Perlman ZE, Kirchhausen T and Mitchison TJ (2003) Phenotypic screening of small molecule libraries by high throughput cell imaging. *Combinatorial Chemistry and High Throughput Screening*, **6**, 279–286.

Zucker RM and Price O (2001) Evaluation of confocal microscopy system performance. *Cytometry*, **44**, 273–294.

siRNA approaches in cell biology

Alexandra Gampel and Harry Mellor
University of Bristol, Bristol, UK

1. Introduction

Small interfering RNAs (siRNAs) allow for the targeted silencing of gene expression in cultured cells and whole organisms. The simplicity of the technique allows it to be applied to the analysis of whole genomes and, when coupled with recent advances in microarray screening and proteomic techniques, siRNAs provide an incredibly powerful tool for postgenomic research. It seems clear that siRNAs will play a key role in biological research in the coming decade. Here, we review the basis of the technique and examine some of the key uses to which it is being applied.

2. Mechanisms of action

Scientists have known that double-stranded RNA is a powerful mediator of targeted gene suppression for some time (Fire *et al*., 1998). This technique has been exploited by the *Caenorhabditis elegans* and *Drosophila* research communities to allow routine silencing of individual genes, as well as genome-wide screening protocols. One major limitation of this technique, however, was its lack of application to mammalian cells. In mammalian cells, double-stranded RNA triggers a number of innate defense mechanisms designed to counter viral infection (Samuel, 2001). So, while double-stranded RNA is effective in gene silencing in mammalian cells, the toxicity of these others responses makes this procedure unfeasible in practical terms. The breakthrough discovery was made recently by Tuschl and colleagues, who showed that small, 21–25 nucleotide double-stranded RNAs were capable of mediating gene silencing in mammalian cells without invoking the toxic antiviral responses (Elbashir *et al*., 2001a). The basis for this finding has become clearer with the elucidation of the mechanisms of siRNA action (Figure 1; reviewed in (Hannon, 2002)). In multicellular eukaryotes, double-stranded RNA is targeted by the enzymes of the Dicer family of class III ribonucleases, which process RNA into short, ∼22 nucleotide double-stranded RNA oligonucleotides. These short RNA duplexes are then bound by the RISC (RNA-induced silencing) complex, which discards one strand and uses the other to seek a complementary target mRNA. Having found its prey, the nuclease activity of RISC

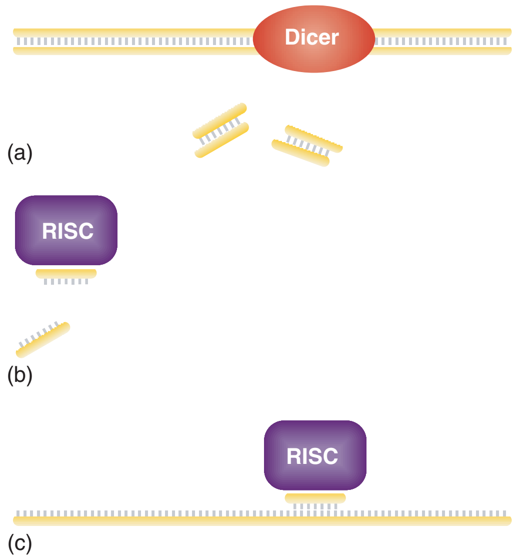


Figure 1 A model for siRNA action. (a) dsRNA is targeted by class III ribonucleases of the Dicer family, which bind as a dimer and fragment the template into siRNAs – short, double-stranded oligonucleotides of approximately 21 nt in length, with 2 nt overhangs at the 3′-ends. (b) The siRNAs are bound by the RISC complex, whose intrinsic helicase activity separates the strands and discards one, keeping the other as a guide RNA. (c) The RISC complex binds to target mRNA with high specificity through Watson–Crick base pairing with the guide RNA. The endonuclease activity of the RISC complex then degrades the target. Synthetic siRNAs are used directly by the RISC complex, sidestepping Dicer, and also avoiding activation of the antiviral response, which requires longer dsRNA

then degrades the mRNA, leading to functional silencing of gene expression. Short synthetic siRNAs are effective substrates for the RISC complex, but are too short to activate effectors of the toxic antiviral response such as the PKR kinase. By feeding straight to the RISC complex, siRNAs allow for effective gene silencing in mammalian cells with little or no toxicity.

3. Design and delivery

One major advantage of siRNA mediated gene silencing is how few constraints there are in the design of successful reagents; a marked contrast to antisense technologies. In our experience, approximately 30% of siRNAs will mediate significant silencing when chosen solely on the basis of annealing temperature. Ongoing research into the mechanisms of siRNA action has already identified criteria that determine the specificity (Elbashir *et al*., 2001b) and efficiency (Schwarz *et al*., 2003) of siRNA action, and it seems likely that the design of effective siRNAs will soon become a routine operation. Several siRNA supply companies now offer extremely useful on-line siRNA design utilities, some of which also include automatic screening of the nonredundant UniGene database to filter out nonspecific sequences (**http://www.dharmacon.com**).

A second key advantage of siRNAs is their high target specificity. Although it is becoming clear that siRNAs have a range of off-target effects on signaling pathways and gene expression (Sledz *et al*., 2003; Persengiev *et al*., 2004), microarray analysis of gene expression suggests that siRNAs are highly specific tools if their design and delivery are optimized (Semizarov *et al*., 2003). Most potential problems can be addressed by the inclusion of two important experimental controls. First, experiments should always include controls without siRNA, and with targetless siRNAs – this allows for any general effects of siRNAs to be excluded. Second, all observations require confirmation through the use of at least two siRNAs directed to different sites on the target mRNA – this dramatically reduces the possibility that observations are due to effects outside of the intended target. As long as these simple precautions are followed, siRNAs appear to be a robust and highly specific experimental tool.

The third important benefit of siRNAs is their ease of delivery. The small size of siRNAs and the potential for RISC to degrade multiple target mRNA molecules each contribute to this. Because they act in the cytosol, siRNAs have a direct advantage over antisense oligonucleotides in that they only have to cross the plasma membrane to reach their site of action. A number of lipid-based transfer protocols have been developed specifically for siRNA delivery and these have been used successfully on a wide range of primary and cultured cell lines. In our experience, a rough rule of thumb is that cell types that can be transfected with plasmids to at least 25% efficiency can be transfected with siRNAs to 100% efficiency. siRNAs are surprisingly stable in cell culture, with a half-life of around 2–5 days. This lifetime can be increased further through chemical modification of the oligonucleotides (**http://www.dharmacon.com**). Nevertheless, there are still situations where longer-term availability of siRNA is required. Agami and coworkers devised a vector-based system where siRNAs are encoded as small hairpin sequences (shRNA), transcribed from an RNA polymerase III promoter, allowing for long-term gene silencing in cells (Brummelkamp *et al*., 2002a). Equivalent retroviral systems allow for high efficiency of delivery and stable integration into the host genome (Brummelkamp *et al*., 2002b). These varied and efficient delivery systems have important implications for the potential clinical use of siRNA technology; however, for most research scientists, the most exciting recent advance has been the demonstration of stable germline transmission of virally encoded shRNA in a mouse line (Carmell *et al*., 2003). Compared to conventional techniques for targeting of mouse genes through homologous recombination, this is a rapid and straightforward route to achieve gene silencing in whole animals. Perhaps most significantly, it is an enabling technology that makes whole animal studies of gene function feasible for a much larger section of the scientific community.

4. Strategies and screens

Scientists are still coming to terms with the new experimental freedoms that siRNA allows. Perhaps the most significant of these is the feasibility of genome-wide screens in organisms that have not previously been accessible to these techniques.

Schultz and coworkers have recently constructed an siRNA library representing 8000 human genes in a novel plasmid expression system that transcribes the two RNA strands from opposed promoters. Having transfected this arrayed library into cultured mammalian cells, the authors performed an assay for TNFα-induced activation of the transcription factor NF-κB and identified several novel members of this signaling pathway (Zheng *et al*., 2004). In an alternative approach, Mousses and colleagues have perfected microarrays of oligonucleotide siRNAs in a transfection matrix that can be overlaid with monolayers of cultured mammalian cells (Mousses *et al*., 2003). It seems that it will soon be possible to carry out genome-wide screens in cultured mammalian cells for any process that can be quantified on a cellular basis. An example of this approach in a genetically tractable organism has recently been published by Perrimon and coworkers. They used RNA interference with dsRNAs to target 994 *Drosophila* genes and scored the morphology of cells grown in culture, allowing them to identify 160 genes that contribute to cell shape (Kiger *et al*., 2003). Although this work involved *Drosophila* cells and long dsRNAs, it could just as easily have been applied on a larger scale to cultured mammalian cells using siRNAs. The recent advances in high-throughput cell imaging and semi-automated image analysis (Price *et al*., 2002) suggest that such scaling is a practical proposition.

Genome-wide screening experiments fall outside of the day-to-day business of most laboratories; however, siRNA allows for other equally powerful approaches to the analysis of gene function that are easily accessible by most researchers. One generalized situation that deserves highlighting is the analysis of branching pathways. Transcription factors often initiate complex profiles of gene expression; however, DNA microarrays now allow relatively straightforward analysis of expression patterns across whole genomes. By combining siRNA silencing of specific transcription factors with gene profiling, it should be possible to map the gene targets of mammalian transcription factors, and so begin to unravel the complexities of transcriptional regulation – one of the key goals of postgenomic research. Certainly, similar approaches in *C. elegans* suggest that this approach is feasible (Murphy *et al*., 2003). A comparable problem is that of mapping the targets of protein kinases. The human genome contains over 500 protein kinases and approximately 30% of cellular proteins are phosphorylated. Recent advances in the identification of phosphoproteins by solid-phase capture techniques and tandem mass spectrometry has made possible the analysis of complex mixtures of phosphoproteins such as cell lysates (Kalume *et al*., 2003). By combining siRNA silencing of protein kinases with differential analysis of target protein phosphorylation it would seem that we can begin to unravel the tangled networks of protein phosphorylation that constitute so much of cellular signaling.

5. Conclusions

siRNAs combine high specificity with ease of use to make them powerful tools in the assignment of gene function. Most of us are still inwardly digesting the experimental freedoms afforded by these new reagents, and we are in a period of ingenious and imaginative development of siRNA applications. Perhaps the

most exciting aspect of these emerging technologies is that they are widely accessible, allowing the community as a whole to participate in the coming decade of postgenomic research.

Further reading

x. (x) **http://www.dharmacon.com**.

References

Brummelkamp TR, Bernards R and Agami R (2002a) A system for stable expression of short interfering RNAs in mammalian cells. *Science*, **296**, 550–553.

Brummelkamp TR, Bernards R and Agami R (2002b) Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell*, **2**, 243–247.

Carmell MA, Zhang L, Conklin DS, Hannon GJ and Rosenquist TA (2003) Germline transmission of RNAi in mice. *Nature Structural Biology*, **10**, 91–92.

Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K and Tuschl T (2001a) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.

Elbashir SM, Martinez J, Patkaniowska A, Lendeckel W and Tuschl T (2001b) Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *The EMBO Journal*, **20**, 6877–6888.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

Hannon GJ (2002) RNA interference. *Nature*, **418**, 244–251.

Kalume DE, Molina H and Pandey A (2003) Tackling the phosphoproteome: tools and strategies. *Current Opinion in Chemical Biology*, **7**, 64–69.

Kiger A, Baum B, Jones S, Jones M, Coulson A, Echeverri C and Perrimon N (2003) A functional genomic analysis of cell morphology using RNA interference. *Journal of Biology (Online)*, **2**, 27.

Mousses S, Caplen NJ, Cornelison R, Weaver D, Basik M, Hautaniemi S, Elkahloun AG, Lotufo RA, Choudary A and Dougherty ER (2003) RNAi microarray analysis in cultured mammalian cells. *Genome Research* **13**, 2341–2347.

Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H and Kenyon C (2003) Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. *Nature*, **424**, 277–283.

Persengiev SP, Zhu X and Green MR (2004) Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA*, **10**, 12–18.

Price JH, Goodacre A, Hahn K, Hodgson L, Hunter EA, Krajewski S, Murphy RF, Rabinovich A, Reed JC and Heynen S (2002) Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *Journal of Cellular Biochemistry. Supplement*, **39**, 194–210.

Samuel CE (2001) Antiviral actions of interferons. *Clinical Microbiology Reviews*, **14**, 778–809, table of contents.

Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N and Zamore PD (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.

Semizarov D, Frost L, Sarthy A, Kroeger P, Halbert DN and Fesik SW (2003) Specificity of short interfering RNA determined through gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6347–6352.

Sledz CA, Holko M, de Veer MJ, Silverman RH and Williams BR (2003) Activation of the interferon system by short-interfering RNAs. *Nature Cell Biology* **5**, 834–839.

Zheng L, Liu J, Batalov S, Zhou D, Orth A, Ding S and Schultz PG (2004) An approach to genomewide screens of expressed small interfering RNAs in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 135–140.

Introductory Review

Posttranslational modification of proteins

Martin R. Larsen and Ole N. Jensen
University of Southern Denmark, Odense, Denmark

1. Introduction

Basic biochemistry courses teach that proteins consist of linear chains of amino acid residues, which encode their three-dimensional structure and biological function. Less emphasis is put on the fate and regulation of proteins once they have been synthesized and released from ribosomes. Exceptions include glycogeneogenic enzymes, for example, glycogen phosphorylase, which is regulated by phosphorylation/dephosphorylation as a function of the energy status of the cell, and blood clotting factors such as fibrinogen, which is activated by proteolytic cleavage as a response to injury.

All cells and tissues rely on proper localization and regulation of proteins in order to maintain integrity and biological function and to be able to respond to external stimulation: How do proteins know where to go in the cell and when to perform their duties? For example, how do proteins help propagate rapid signals from the eye lens to the brain and in turn respond by activation of certain muscles in our arms or legs to perform a desired action? The answer to these questions is that proteins are transiently or permanently modified by other proteins/enzymes in a process referred to as posttranslational modification. Posttranslational modifications may encompass covalent attachment (or removal) of chemical moieties, for example, addition of phosphate groups (*see* Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6) or carbohydrate groups (*see* Article 62, **Glycosylation**, Volume 6, Article 64, **Structure/function of *N*-glycans**, Volume 6, and Article 69, **Glycosylation in bacteria: that, what, how, why, now what?**, Volume 6), which provides the cue for the protein to change its conformation and become active or to target the protein to a particular cellular compartment where it may fulfill its function. Yet other types of posttranslational modifications such as ubiquitination destine proteins for degradation. Thus, at any given time, a given protein species may exist in several different forms each of which may have different localization, activity, and function. A list of common protein modifications and their function is outlined in Table 1.

Mass spectrometry (MS), a technique widely used for molecular mass determination of peptides and proteins, is ideally suited to the study of posttranslational

Table 1 Some commonly encountered posttranslational modifications of proteins

PTM type	ΔMass (Da)	Stability	Function and notes
Phosphorylation			Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signaling
– pTyr	+80	+++	
– pSer, pThr		+/++	
Acetylation	+42	+++	Protein stability, protection of N-terminus. Regulation of protein–DNA interactions (Histones)
Methylation	+14	+++	Gene regulation (Histones)
Acylation, fatty acid modification			Cellular localization and targeting signals, membrane tethering, mediator of protein–protein interactions
– farnesyl	+204	+++	
– myristoyl	+210	+++	
– palmitoyl	+238	+/++	
– etc.			
Glycosylation			
– N-linked	>1000	+/++	Excreted proteins, cell–cell recognition/signaling
– O-linked	162	+/++	*O*-GlcNAc, reversible, regulatory functions
GPI-anchor	>1000	++	Glycosylphosphatidylinositol (GPI)-anchor, membrane tethering of enzymes and receptors
Disulfide bond formation	−2	++	Intra- and intermolecular cross-link, protein stability
Deamidation	+1	+++	Probably a regulator of protein-ligand and protein–protein interactions, also a common chemical artifact
Pyroglutamic acid	−17	+++	Protein stability, blocked N-terminus
Ubiquitination	>1000	+/++	Destruction signal
Nitration of Tyr	+45	+/++	Oxidative damage during inflammation

For a comprehensive list of protein modifications and the corresponding Δmass values, visit DeltaMass at **http://www.abrf.org/index.cfm/dm.home**.

modifications. The basic premise is that the addition of chemical moieties to proteins will lead to an increase in protein molecular mass, which can be measured by MS. In reality, proteins are usually digested with proteolytic enzymes to generate peptides, which are more amendable to MS analysis. In this case, the challenge is to retrieve, identify, and characterize the peptides that carry the modified amino acid residues (Figure 1). Electrospray ionization (ESI) MS (Fenn *et al*., 1989) and matrix-assisted laser desorption/ionization (MALDI) MS (Karas and Hillenkamp, 1988) are used in such studies to determine posttranslational modifications, including tandem mass spectrometry (MS/MS) for sequencing of the modified peptides (Mann and Jensen, 2003; Jensen, 2004).

The aim of this introductory review is to describe protein phosphorylation (*see* Article 63, **Protein phosphorylation analysis by mass spectrometry**,

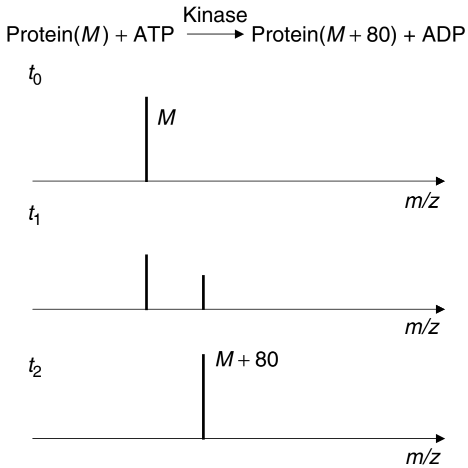


Figure 1 Monitoring posttranslational modification of protein by mass spectrometry. A polypeptide of molecular weight M is incubated with ATP and a protein kinase and aliquots of the sample are analyzed by mass spectrometry at three timepoints. Modification of the protein is evident by observation of an increase in protein molecular weight of 80 Da, corresponding to addition of one phosphate group

Volume 6 and Article 73, **Protein phosphorylation analysis – a primer**, Volume 6) and protein processing (*see* Article 67, **Posttranslational cleavage of cell-surface receptors**, Volume 6). We will outline techniques to investigate these posttranslational modifications by mass spectrometry and complementary biochemical techniques.

2. Protein phosphorylation

Protein phosphorylation is the covalent addition of a phosphate group to one or more amino acid residues within a protein. Each phosphate group adds 80 Da to the molecular mass of the protein or the proteolytically derived peptides and is the basis for MS-based detection techniques. The phosphorylation reaction is catalyzed by protein kinases, which usually utilize ATP or GTP as the phosphate donor. In eukaryote cells, serine, threonine, or tyrosine residues are the substrate sites located within kinase recognition sequence motifs in proteins. For example, protein kinase A recognizes the consensus sequence (R/K)(R/K)x(S/T), that is, a dibasic site followed by any amino acid residue followed by the acceptor serine or threonine residue, such as the motif RRQSLA in human Engrailed-2 protein (Hjerrild *et al*., 2004). Thus, computational sequence analysis can be used to predict putative phosphorylation sites in proteins by a range of well-characterized protein kinases (Blom *et al*., 2004; Hjerrild *et al*., 2004). The human genome contains on the order of 500 genes that encode protein kinases, illustrating the importance of this class of enzyme (Manning *et al*., 2002).

Receptor tyrosine kinases (RTKs) include the insulin receptor, epidermal growth factor receptor (EGFr) and the platelet-derived growth factor receptor (PDGFr),

which play important roles in cellular development and differentiation. Whereas the occurrence of phosphotyrosine is less abundant than phosphoserine and phosphothreonine in mammalian cells, the signaling pathways propagated by RTKs and tyrosine phosphorylation have been relatively well described during the last decade because good reagents are available for their analysis. These include specific inhibitors of RTKs and their downstream targets and anti-phosphotyrosine antibodies for western blotting and immunoprecipitation of phosphotyrosine proteins. Several successful proteomics approaches for studying cell signaling are based on anti-phosphotyrosine antibodies for the recovery of phosphoprotein signaling complexes for subsequent identification by MS (Blagoev *et al*., 2003; Salomon *et al*., 2003; Pandey *et al*., 2000).

Phosphoproteins are substrates for endogenous phosphoprotein phosphatases that remove phosphate groups. When purifying phosphoproteins from cells and tissues it is therefore important to include phosphatase inhibitors, for example, Calyculin A. Most phosphatases recognize specific phosphorylated residues and consensus sequence sites within proteins. However, some phosphatases are nonspecific and will efficiently remove a majority of phosphate on proteins and peptides, for example, alkaline phosphatase and calf intestinal phosphatase.

Purified or recombinant protein kinases and phosphoprotein phosphatases are very useful in MS-based in vitro protein phosphorylation assays, as the addition or removal of phosphate groups leads to a mass increase or decrease of 80 Da (HPO3) or multiples thereof, which is easily monitored by MS (Stensballe *et al*., 2001; Larsen *et al*., 2001b; Qin and Chait, 1997).

Phosphorylation changes the physiochemical properties of proteins as the addition of a phosphate group to an amino acid leads to a lowering of the pK_a value. Phosphorylated amino acids may reduce the cleavage efficiency of proteases, leading to a low yield of phosphopeptide for MS analysis. Furthermore, the lower pK_a value of phosphopeptides relative to unmodified peptides often results in differential detection of phosphopeptide, particularly in MALDI MS. Several methods for enhancing phosphopeptide signals have been reported, including the use of matrix additives in MALDI MS (Asara and Allison, 1999) and immobilized metal affinity chromatography (IMAC) (Neville *et al*., 1997; Figeys *et al*., 1998; Cao and Stults, 1999; Posewitz and Tempst, 1999; Stensballe *et al*., 2001), graphite columns (Larsen *et al*., 2004; Chin and Papac, 1999; Vacratsis *et al*., 2002) and antibodies (Sickmann and Meyer, 2001) for enrichment of phosphopeptides prior to MS analysis. Such methods improve the probability of MS-based detection of phosphorylated peptides in complex mixtures.

3. Phosphopeptide analysis by tandem mass spectrometry

Tandem mass spectrometry is utilized for phosphopeptide detection and sequencing in several different ways. The most commonly used approach is the product ion-scanning mode (*see* Article 18, **Techniques for ion dissociation (fragmentation) and scanning in MS/MS**, Volume 5) in which a phosphopeptide candidate is sequenced by MS/MS to determine which amino acid residue is phosphorylated. This is a very efficient approach, which relies on “picking” the right ion for

sequencing. In those cases in which no phosphopeptide candidates are readily called, then the MS/MS precursor ion-scanning method may reveal the phosphopeptides via selective detection of diagnostic fragment ions, such as *m/z* 216 for phosphotyrosine (Carr *et al*., 1996; Neubauer and Mann, 1999; Steen *et al*., 2001). MS/MS neutral loss scanning (*see* Article 18, **Techniques for ion dissociation (fragmentation) and scanning in MS/MS**, Volume 5) is yet another method that relies on the gas-phase elimination of phosphoric acid (H3PO4, $m = 98$) from phosphoserine- and phosphothreonine-containing peptides (Bateman *et al*., 2002; Schroeder *et al*., 2004).

Whereas most phosphoprotein determinations have been performed by ESI MS/MS, the advent of sensitive and efficient MALDI MS/MS instruments has revived the interest in using this approach for phosphoprotein analysis. Several improvements have been reported, including optimized sample preparation techniques (Kjellstrom and Jensen, 2004; Larsen *et al*., 2004; Stensballe and Jensen, 2004) and use of various types of MALDI tandem mass spectrometers (Bennett *et al*., 2002; Medzihradszky *et al*., 2000; Krutchinsky *et al*., 2001).

Owing to the importance of protein phosphorylation in many biological processes, the systematic determination of phosphoproteins combined with the assignment and quantitation of phosphorylation sites, known as phosphoproteomics, is receiving much attention. Methods for enrichment of phosphoproteins from cell lysates are used in combination with multidimensional chromatography (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5) and mass spectrometry in large-scale studies aimed at elucidating molecular mechanisms of signal transduction in cellular development, differentiation, and defense. Antibodies specific for phosphoproteins, for example, anti-phosphotyrosine antibodies and anti-PKA substrate antibodies have proven very useful in this endeavor (Pandey *et al*., 2000; Gronborg *et al*., 2002; Salomon *et al*., 2003). Similarly, methods for enrichment of phosphopeptides by IMAC have been extensively used in large-scale phosphoproteomics (Ficarro *et al*., 2003; Brill *et al*., 2004; Nuhse *et al*., 2003). The final stage of phosphoproteomics analysis is tandem mass spectrometry followed by data analysis and interpretation. Quantitative methods based on stable isotope labeling (*see* Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5) enable relative quantitation of protein expression and also determination of changes in protein phosphorylation (Mann *et al*., 2002).

4. Protein processing and truncation

Proteolytic processing is a commonly observed posttranslational modification. Signal peptides, which are used to target and localize proteins to different compartments in the cell, are removed during the maturation of the protein. The proteolytic processing of the polypeptide chain is very specific, meaning that the cleavage is taking place at a given amino acid motif. A combination of unspecific and specific processing can be observed during or after changes in the environment, for example, due to cell aging (Ling *et al*., 2003) or osmotic stress (Larsen *et al*.,

2001a). Identifying the exact sites of processing of proteins in the cell is often required in order to elucidate the biological significance of the processing. Several tools for prediction of signal peptides and secretion signals of proteins are available (**www.cbs.dtu.dk**).

The first step in the characterization of proteolytically processed proteins is the efficient detection of the corresponding protein products. Detection of proteolytic processing in proteomics is often performed using two-dimensional gel electrophoresis (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5 and Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5), since processed proteins can be efficiently detected on the basis of migration changes brought about by their reduced size. Depending on the specificity of the processing of the protein, it will migrate into a single spot (specific processing) or give rise to a smear on the gel (unspecific processing) on the 2D gel. Non-gel-based proteomics methods, which rely on the detection of individual tryptic peptides rather than measurement of the intact, processed polypeptides, are not well suited for the study of proteolytically processed proteins because of the difficulty in capturing and defining the N- and C-terminal peptides.

Once isolated, the proteolytically processed protein can be characterized using mass spectrometry in order to locate the exact site for processing, that is, the site of proteolytic cleavage. The most frequently used methods for localizing the exact site is by differential peptide mass mapping using proteolytic enzymes with different cleavage specificities, for example, trypsin and endoproteinase Asp-N (Larsen *et al*., 2001a). Successful analysis relies on the detection of the terminal peptide(s) of the processed protein. Peptides that cannot be immediately mapped to the protein sequence or differ from the peptide mass map of the mature protein are sequenced by tandem mass spectrometry, as they may be derived from the processed termini of the polypeptide. Selective enrichment of the C-terminal peptide by anhydrotrypsin chromatography (ATC), an affinity chromatography method that selectively adsorbs peptides with Arg, Lys, or AECys (*S*-aminoethyl cysteine) residues at the C-terminus under weak acidic conditions, has been demonstrated (e.g., Kumazaki *et al*., 1987).

In a recent study of stress-induced changes in protein expression in *Saccharomyces cerevisiae*, a total of 10 protein spots on a silver-stained 2D gel were identified to represent C- and/or N-terminal processed forms of enolase 2 (Larsen *et al*., 2001a). The C-terminal peptides were identified by mass spectrometry analysis of the peptides derived by in-gel digestion of the proteins in a buffer containing $H_2{}^{16}O$ and $H_2{}^{18}O$ in the proportion 1:1. This results in incorporation of $^{16}O/^{18}O$ (1:1) in all newly generated carboxy termini except the original C-terminal peptide, resulting in a split signal separated by 2 Da for "internal" peptides. Thus, the original C-terminal peptide will exhibit a single peptide ion signal, whereas all other peptides produce a split signal in the mass spectrum.

Presently, the detection of N-terminal peptides is complicated by the reactivity of the N-terminal amino group, which is very similar to the reactivity of the lysine amino group, resulting in no efficient chemistry to tag or label the N-terminal peptide.

5. Conclusions

The elucidation of posttranslational modification of proteins is a prerequisite for understanding the molecular details of living organisms, from bacteria to plants and mammals. The ability to detect and monitor posttranslational modifications of proteins by mass spectrometry and other analytical methods is already providing exciting insights into key regulatory events in cells and organisms, in health and in disease.

References

Asara JM and Allison J (1999) Enhanced detection of phosphopeptides in matrix-assisted laser desorption/ionization mass spectrometry using ammonium salts. *Journal of the American Society for Mass Spectrometry*, **10**, 35–44.

Bateman RH, Carruthers R, Hoyes JB, Jones C, Langridge JI, Millar A and Vissers JPC (2002) A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *Journal of the American Society for Mass Spectrometry*, **13**, 792–803.

Bennett KL, Stensballe A, Podtelejnikov AV, Moniatte M and Jensen ON (2002) Phosphopeptide detection and sequencing by matrix-assisted laser desorption/ionization quadrupole time-of-flight tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, **37**, 179–190.

Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ and Mann M (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nature Biotechnology*, **21**, 315–318.

Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S and Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Brill LM, Salomon AR, Ficarro SB, Mukherji M, Stettler-Gill M and Peters EC (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Analytical Chemistry*, **76**, 2763–2772.

Cao P and Stults JT (1999) Phosphopeptide analysis by on-line immobilized metal-ion affinity chromatography-capillary electrophoresis-electrospray ionization mass spectrometry. *Journal of Chromatography A*, **853**, 225–235.

Carr SA, Huddleston MJ and Annan RS (1996) Selective detection and sequencing of phosphopeptides at the femtomole level by mass spectrometry. *Analytical Biochemistry*, **239**, 180–192.

Chin ET and Papac DI (1999) The use of a porous graphitic carbon column for desalting hydrophilic peptides prior to matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Analytical Biochemistry*, **273**, 179–185.

Fenn JB, Mann M, Meng CK, Wong SF and Whitehouse CM (1989) Electrospray Ionization for mass-spectrometry of large biomolecules. *Science*, **246**, 64–71.

Ficarro S, Chertihin O, Westbrook VA, White F, Jayes F, Kalab P, Marto JA, Shabanowitz J, Herr JC, Hunt DF, *et al*. (2003) Phosphoproteome analysis of capacitated human sperm - Evidence of tyrosine phosphorylation of a kinase-anchoring protein 3 and valosin-containing protein/p97 during capacitation. *The Journal of Biological Chemistry*, **278**, 11579–11589.

Figeys D, Gygi SP, Zhang Y, Watts J, Gu M and Aebersold R (1998) Electrophoresis combined with novel mass spectrometry techniques: Powerful tools for the analysis of proteins and proteomes. *Electrophoresis*, **19**, 1811–1818.

Gronborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON and Pandey A (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies - Identification of a novel protein, Frigg, as a protein kinase a substrate. *Molecular & Cellular Proteomics*, **1**, 517–527.

Hjerrild M, Stensballe A, Rasmussen TE, Kofoed CB, Blom N, Sicheritz-Ponten T, Larsen MR, Brunak S, Jensen ON and Gammeltoft S (2004) Identification of phosphorylation sites in protein kinase a substrates using artificial neural networks and mass spectrometry. *Journal of Proteome Research*, **3**, 426–433.

Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, **8**, 33–41.

Karas M and Hillenkamp F (1988) Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons. *Analytical Chemistry*, **60**, 2299–2301.

Kjellstrom S and Jensen ON (2004) Phosphoric acid as a matrix additive for MALDI MS analysis of phosphopeptides and phosphoproteins. *Analytical Chemistry*, **76**(17), 5109–5117.

Krutchinsky AN, Kalkum M and Chait BT (2001) Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Analytical Chemistry*, **73**, 5066–5077.

Kumazaki T, Terasawa K and Ishii S (1987) Affinity-chromatography on immobilized anhydrotrypsin - general utility for selective isolation of C-terminal peptides from protease digests of proteins. *Journal of Biochemistry (Tokyo)*, **102**, 1539–1546.

Larsen MR, Graham ME, Robinson PJ and Roepstorff P (2004) Improved detection of hydrophilic phosphopeptides using graphite powder microcolumns and mass spectrometry - evidence for in vivo doubly phosphorylated dynamin I and dynamin III. *Molecular & Cellular Proteomics*, **3**, 456–465.

Larsen MR, Larsen PM, Fey SJ and Roepstorff P (2001a) Characterization of differently processed forms of enolase 2 from *Saccharomyces cerevisiae* by two-dimensional gel electrophoresis and mass spectrometry. *Electrophoresis*, **22**, 566–575.

Larsen MR, Sorensen GL, Fey SJ, Larsen PM and Roepstorff P (2001b) Phospho-proteomics: evaluation of the use of enzymatic de-phosphorylation and differential mass spectrometric peptide mass mapping for site specific phosphorylation assignment in proteins separated by gel electrophoresis. *Proteomics*, **1**, 223–238.

Ling Y, Morgan K and Kalsheker N (2003) Amyloid precursor protein (APP) and the biology of proteolytic processing: relevance to Alzheimer's disease. *The International Journal of Biochemistry & Cell Biology*, **35**, 1505–1535.

Mann M and Jensen ON (2003) Proteomic analysis of post-translational modifications. *Nature Biotechnology*, **21**, 255–261.

Mann M, Ong SE, Gronborg M, Steen H, Jensen ON and Pandey A (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends in Biotechnology*, **20**, 261–268.

Manning G, Whyte DB, Martinez R, Hunter T and Sudarsanam S (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML and Burlingame AL (2000) The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Analytical Chemistry*, **72**, 552–558.

Neubauer G and Mann M (1999) Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: potentials and limitations. *Analytical Chemistry*, **71**, 235–242.

Neville DCA, Rozanas CR, Price EM, Gruis DB, Verkman AS and Townsend RR (1997) Evidence for phosphorylation of serine 753 in CFTR using a novel metal-ion affinity resin and matrix-assisted laser desorption mass spectrometry. *Protein Science*, **6**, 2436–2445.

Nuhse TS, Stensballe A, Jensen ON and Peck SC (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Molecular & Cellular Proteomics*, **2**, 1234–1243.

Pandey A, Podtelejnikov AV, Blagoev B, Bustelo XR, Mann M and Lodish HF (2000) Analysis of receptor signaling pathways by mass spectrometry: Identification of Vav-2 as a substrate

of the epidermal and platelet-derived growth factor receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 179–184.

Posewitz MC and Tempst P (1999) Immobilized gallium(III) affinity chromatography of phosphopeptides. *Analytical Chemistry*, **71**, 2883–2892.

Qin J and Chait BT (1997) Identification and characterization of posttranslational modifications of proteins by MALDI ion trap mass spectrometry. *Analytical Chemistry*, **69**, 4002–4009.

Salomon AR, Ficarro SB, Brill LM, Brinker A, Phung QT, Ericson C, Sauer K, Brock A, Horn DM and Schultz PG (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 443–448.

Schroeder MJ, Shabanowitz J, Schwartz JC, Hunt DF and Coon JJ (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Analytical Chemistry*, **76**, 3590–3598.

Sickmann A and Meyer HE (2001) Phosphoamino acid analysis. *Proteomics*, **1**, 200–206.

Steen H, Kuster B, Fernandez M, Pandey A and Mann M (2001) Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Analytical Chemistry*, **73**, 1440–1448.

Stensballe A, Andersen S and Jensen ON (2001) Characterization of phosphoproteins from electrophoretic gels by nanoscale Fe(III) affinity chromatography with off-line mass spectrometry analysis. *Proteomics*, **1**, 207–222.

Stensballe A and Jensen ON (2004) Phosphoric acid enhances the performance of Fe(III) affinity chromatography and MALDI MS/MS for recovery, detection and sequencing of phosphopeptides. *Rapid Communications in Mass Spectrometry: RCM*, **18**(15), 1721–1730.

Vacratsis PO, Phinney BS, Gage DA and Gallo KA (2002) Identification of in vivo phosphorylation sites of MLK3 by mass spectrometry and phosphopeptide mapping. *Biochemistry*, **41**, 5613–5624.

Introductory Review

Glycosylation

Richard D. Cummings
University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

1. Definition of glycosylation

Glycosylation is the enzymatic process of adding sugars in glycosidic linkages to other molecules to generate glycoconjugates or complex carbohydrates, which are defined as molecules that have covalently attached sugars. The process of glycosylation generates all the classes of glycoconjugates, which include glycoproteins, glycosylphosphoinositol (GPI)-anchored glycoproteins, glycolipids, glycosaminoglycans and proteoglycans, peptidoglycans, lipopolysaccharides, free oligosaccharides and polysaccharides, and modified drugs and toxins (Hakomori, 2000; Lis and Sharon, 1993; Power and Jennings, 2003; Rademacher *et al*., 1988; Raetz and Whitfield, 2002; Reiter, 2002; Samuel and Reeves, 2003; Scheible and Pauly, 2004; Spiro, 2002; Upreti *et al*., 2003; Varki, 1993; Wilson, 2002; Zachara and Hart, 2004; Zitzmann *et al*., 2000). Even nucleic acids are a kind of glycoconjugate, since they contain monosaccharide residues, although they are not synthesized by the typical process of glycosylation. Glycoconjugates are found in abundance in higher animals and plants, but most organisms, including all bacteria and most viruses, contain one or more glycoconjugates and thus directly or indirectly carry out glycosylation. Examples of some glycoconjugates in mammalian cells produced by the complex processes of glycosylation are shown in Figure 1. There is also a nonenzymatic mechanism termed *glycation* where sugars chemically react with an amino acid side chain in proteins to generate a covalent attachment of a monosaccharide, but this process is distinguishable from glycosylation (Kikuchi *et al*., 2003).

2. Monosaccharides in glycoconjugates

Mammalian glycoconjugates are built from 10 basic building blocks or sugar precursors (Figure 1), which include the 6-carbon sugars termed hexoses (Glucose – Glc, Galactose – Gal, Mannose – Man, and Fucose – Fuc), the 6-carbon hexosamines (*N*-acetylglucosamine – GlcNAc and *N*-acetylgalactosamine – GalNAc), the 9-carbon Sialic acids (Sia, including *N*-acetylneuraminic acid – NeuAc and *N*-glycolylneuraminic acid – NeuGc), the 6-carbon hexuronic acids (Glucuronic acid – GlcA and Iduronic acid – IdA), and the 5-carbon

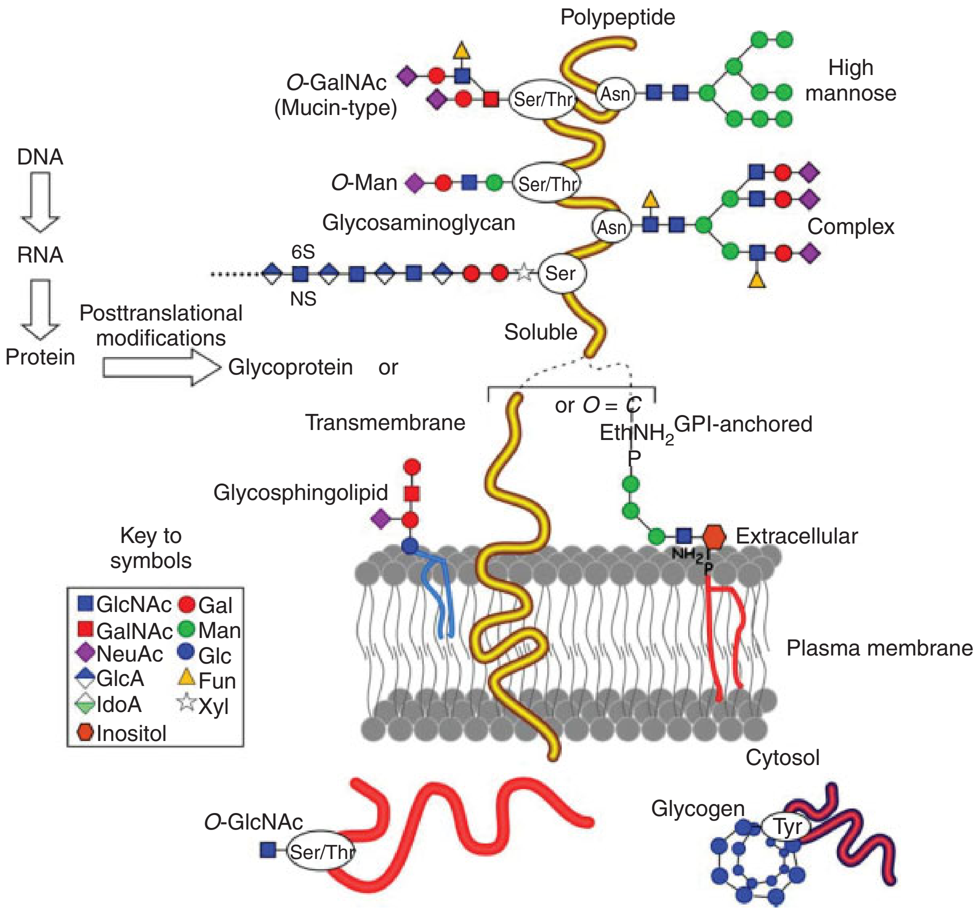


Figure 1 Glycosylation generates the greatest diversity of macromolecules and the largest class of posttranslational modifications of proteins. Examples of composite glycoproteins are shown in the secretions, plasma membrane, and cytoplasm of a mammalian cell, along with glycosphingolipid in the membrane. Monosaccharides are represented by the symbols identified in the key

pentose Xylose (Xyl). Other monosaccharides found in nucleic acids are the 5-carbon pentoses (ribose – Rib and deoxyribose – deoxyRib). While these are major monosaccharides in mammals, other life forms may contain these and/or many different monosaccharides. For example, sialic acids are found in all mammals, but not in all animals and plants. Plant glycoconjugates often contain sugars not found in animals, such as the 5-carbon sugar arabinose (Ara), and the 6-carbon sugars rhamnose (Rha) and fructose (Fruc). Bacteria express many unusual sugars not commonly found in mammalian glycoconjugates, such as 7-carbon sugars termed heptoses, an 8-carbon sugar 2-keto-3-deoxyoctonoic acid (KDO), and a 9-carbon sugar *N*-acetylmuramic acid (MurNAc). Altogether, well over a hundred different monosaccharides have been identified in various glycoconjugates made by different organisms.

3. Structures of glycoconjugates

The monosaccharides in glycosylated molecules are usually linked together within oligosaccharides, which have from 2 to approximately 20 sugar residues, or polysaccharides, which contain more than 20 sugar residues in a chain (Taylor and Drickamer, 2003; Varki *et al*., 1999) (Figure 1) (*see* Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6). These oligosaccharides and polysaccharides are generally now referred to as "glycans". Glycan structures are directional. The "reducing end" has the reducing terminus (i.e., the chemical reducing activity of an aldehyde or ketone) when the glycan is free. Most sugars in glycoconjugates are aldoses, that is, sugars with an aldehyde group at carbon 1, such as Glc, but some glycoconjugates also contain sugars that are ketoses, that is, sugars with a keto group at carbon 2, such as Fruc. Glycans are usually linked through their reducing end to another molecule. The opposite end of the glycan is termed the "nonreducing end". In regard to glycoproteins, there is enormous diversity in the ways they can be glycosylated. Some glycoproteins contain a single sugar residue or a single glycan linked to an amino acid. Some glycoproteins, such as mucins, contain thousands of glycans each linked to specific amino acids within the mucin. Over two dozen linkages of sugars to proteins have been identified so far. The linkage of one sugar to another or to a non-carbohydrate component (termed an aglycone) is typically an O-glycosidic linkage (or bond) characterized by −C−O−C−that is formed by the double condensation of an aldehyde with alcohols. The O-glycosidic linkage is known as an acetal. In typical glycans, the monosaccharide residues occur as 6-member rings known as a pyranose, but some glycans have monosaccharides that occur as 5-member rings known as a furanose. The linkage between each sugar is through an anomeric carbon that is optically active and in either alpha or beta anomeric linkage to the other component. The carbon at the reducing end of a sugar residue is the anomeric carbon (usually carbon 1 for aldoses and carbon 2 for ketoses). Each monosaccharide occurs in nature as either an optically active D or L enantiomer (an enantiomer is a molecule that is a mirror image isomer of another molecule). The presence of enantiomers is not random because one is usually preferred over the other in biology. For example, all glucose residues in animals are D enantiomers (D-Glc), and all fucose residues appear as L enantiomers (L-Fuc). Enzymes that recognize a particular enantiomer (e.g., D-Glc or D-Gal) usually do not recognize the L-form of that sugar. Consider as an example the disaccharide sequence D-Gal-beta-1-4-D-Glc (abbreviated Galβ1→4Glc), which can be read as showing that D-Gal is β-linked from its carbon 1 through the hydroxyl group at carbon 4 of D-Glc. The D-Glc residue is at the reducing end and the Gal residue is at the nonreducing end. This is the structure of lactose, which is the common sugar found in milk. A conjugate in nature containing this sugar is the glycolipid lactosylceramide, which has the structure Galβ1→4Glcβ1→ceramide and is one of the simplest glycolipids found in animals. Linkages of sugars to each other can produce di-, tri-, tetra-, penta-, hexasaccharides, and so forth. But unlike other building blocks for other biological polymers, such as nucleic acids and proteins, monosaccharides within glycans may be branched. This branching is a key feature of many glycans and allows a tremendous diversity of structures to be generated from a relatively small number

of building blocks. Thus, many glycans may have the same size and composition, but because of differences in the way they are branched, or the anomeric linkage between them, glycoconjugates of the same monosaccharide composition can vary enormously in structure. It has been calculated that six monosaccharides can be theoretically linked in different ways to make as many as 1×10^{12} different hexasaccharides (Laine, 1994). By contrast, six amino acids can be theoretically linked in different ways to make only 6.4×10^{7} different hexapeptides. Sugars can be linked to amino acids in proteins in multiple ways, but the most common are O-glycosidic linkages –C–O–C–and *N*-glycosidic linkages –C–N–involving amide linkages of sugar to an amino acid. In animal cell glycoproteins, the *N*-glycans (also termed Asn-linked oligosaccharides) are attached to the amino acids asparagine (Asn) as *N*-glycosides such as *R*-GlcNAcβ1→Asn (Figure 1). *O*-glycans are usually linked as *O*-glycosides to the hydroxylated amino acids serine (Ser) and threonine (Thr), hydrolysine, hydroxyproline, and tyrosine (Tyr) (*see* Article 65, **Structure/function of *O*-glycans**, Volume 6). There are also many less common forms of attaching glycans to proteins, such as C-mannosylation of tryptophan, which involves a direct carbon–carbon (–C–C–) bond between the sugar and an amino acid bond rather than a glycosidic bond. Linkages of sugars to lipids are often through a glycosidic bond to a lipid alcohol group, as seen above for lactosylceramide or through a phosphoester linkage. Phosphoester linkages between sugars are also found within some polysaccharides, such as the protozoan parasite-derived lipophosphoglycans and in bacterial teichoic acids, which are polymers containing phosphate groups joining glycerol or ribitol (a reduced form of ribose).

4. Occurrence of glycoconjugates

Glycoconjugates are found on the surfaces of cells as either intrinsic or extrinsic membrane components, but they also occur within cells (cytoplasmic localization and in organelles), and in cell secretions. Glycoconjugates are the basis of the human ABO blood group antigens, and many other less-well-publicized blood group antigens, such as the Lewis antigens and P antigens (Hakomori, 1999) that occur on cell surface and secreted glycoconjugates. It has been estimated that in mammals, a majority of the gene-encoded proteins are glycosylated (Apweiler *et al*., 1999). Glycosylation is considered to be the most common, and complex, type of "posttranslational modification" of proteins. While the number of human genes is now estimated to be between 20 000–25 000, the number of differently glycosylated proteins (glycoforms) may be in the many millions. This is because different cell types often glycosylate the same protein differently, and can differentially glycosylate amino acids within a single glycoprotein. Differences in glycan structures on a single glycoprotein species are sometimes referred to as *microheterogeneity*. The field of study that attempts to define all the structures of all the glycoconjugates is termed glycomics (Drickamer and Taylor, 2002; Sutton-Smith *et al*., 2002). It has analogy to genomics and proteomics, which are the sequences of all genes and definition of all expressed proteins, respectively, in an organism (*see* Article 74, **Glycoproteomics**, Volume 6). Bacteria are covered by a glycoconjugate coat, which is composed of peptidoglycan (a polysaccharide

containing covalently attached amino acids), and lipopolysaccharides, a composite of sugars linked to membrane lipids. Some bacteria also make glycoproteins where sugars are covalently linked amino acids (Schaffer and Messner, 2004), as found in animal glycoproteins, although the glycan structures are different. Each type of bacteria generates different glycan structures; in pathogenic bacteria that infect animals, these glycans are often antigenic. Many antibiotics, such as penicillin and bacitracin, interfere with glycoconjugate assembly in bacteria (*see* Article 69, **Glycosylation in bacteria: that, what, how, why, now what?**, Volume 6). In general, all cells appear to contain abundant amounts of glycosylated molecules on their surfaces and in their secretions.

5. Sites of glycosylation in cells

All of the sugars that a cell needs for glycosylation are synthesized directly within cells from de novo metabolic synthesis or interconversions, but cells can also transport many sugars into their cytoplasm from the outside. Animals, fungi, bacteria, and possibly carnivorous plants can acquire monosaccharides by uptake of monosaccharides released through the degradation of glycoconjugates in food. Glycosylation in animal cells occurs in many different cellular compartments such as the cytosol, the secretory organelles, which include the endoplasmic reticulum and Golgi apparatus, and the plasma membrane (Young, 2004). Glycosylation in bacteria occurs primarily in the outer membrane. Viral glycoproteins are normally synthesized by enzymes produced by the host cell that the virus invades.

6. Process of glycosylation

Higher animals and plants have hundreds of individual genes encoding the enzymes that are involved directly in glycosylation to catalyze the transfer of sugar, for example, transferases (Davies and Henrissat, 2002). There are also hundreds of other genes involved in synthesis and turnover of glycosylation precursors and intermediates, such as the monosaccharides and lipids. There appear to be at least five major classes of enzymes that catalyze glycosylation reactions (Figure 2). The most common class comprises the glycosyltransferases that transfer sugar to an acceptor molecule using sugar nucleotide donors; in most cases sugar addition occurs to the nonreducing end of the glycan, but in some special cases it may occur to the reducing end. Examples of sugar nucleotide donors are uridine-diphospho-glucose (UDP-Glc) and UDP-GlcNAc. Some less common sugar nucleotide donors are guanosine-diphospho-fucose (GDP-Fuc) and cytosine-monophospho-sialic acid (CMP-Sia). Other types of transferases transfer sugar from a lipid-phospho-glycan donor, such as dolichol-phospho-mannose and dolichol-pyrophospho-oligosaccharide to acceptor molecules. The addition of sugar to Asn residues in animal cell glycoproteins occurs by the en bloc addition of a preformed glycan from a dolichol-pyrophospho-oligosaccharide during protein translation (cotranslational modification) in the rough endoplasmic reticulum, coupled by the release of dolichol-pyrophosphate (Yan and Lennarz, 1999) (Figure 2). Such

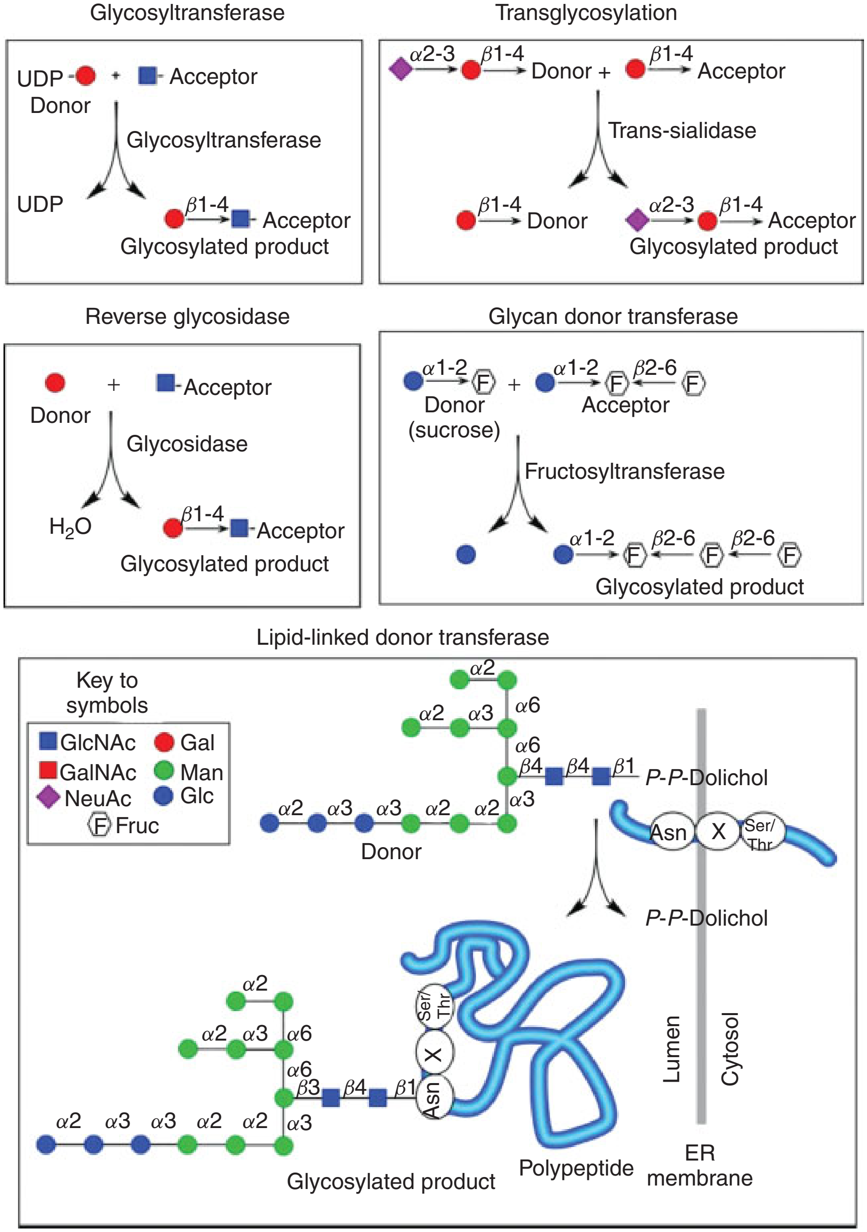


Figure 2 There are several types of glycosylation reactions. Shown are five major pathways in which acceptor molecules are enzymatically glycosylated. Donors for the reactions are shown on the left side within each box and the acceptor and glycosylated product are shown on the right side. Monosaccharides are represented by the symbols identified in the key

glycosylation can be blocked in cells treated with the antibiotic tunicamycin, which blocks the formation of the dolichol-pyrophospho-oligosaccharide donor. Bacteria lipopolysaccharides are also synthesized by the en bloc transfer of preformed lipid-phospho-glycans intermediates to growing polysaccharide chains. The recycling of the lipid-phosphate in this bacterial process is inhibited by the antibiotic bacitracin. Within the Golgi apparatus of eukaryotic cells, other sugars are enzymatically removed and added to glycoproteins in a stepwise fashion through glycan "trimming" and "processing" reactions. Glycoproteins acquire their "mature" structures upon completion of this complex trafficking (Kornfeld and Kornfeld, 1985; *see also* Article 64, **Structure/function of *N*-glycans**, Volume 6). Some organisms use transglycosidases that catalyze sugar addition through a transglycosylation reaction whereby a terminal sugar is enzymatically removed from one glycan and donated to another. Such reactions are used by pathogenic protozoans to coat themselves with sialic acid from host tissues (Pereira-Chioccola and Schenkman, 1999). Interestingly, sugar addition can also occur by the reverse action of glycosidases, which normally remove sugars from a glycoconjugate (Bertozzi and Kiessling, 2001). Finally, in a pathway possibly unique to plants, a disaccharide can also be a donor in the glycosylation of a larger polysaccharide, as is seen for sucrose (a disaccharide of glucose linked to fructose). The enzyme fructosyl transferase transfers Fruc from sucrose to a growing Fruc-containing polysaccharide (fructan), with the release of free Glc (Cairns, 2003).

7. Sizes of glycoconjugates

Glycoconjugates come in all sizes and are among the largest and the smallest polymers in nature (Rademacher *et al*., 1988; Varki *et al*., 1999). Glycogen, the major storage polysaccharide of animals, which is composed of repeating and branched α-linked Glc residues, is actually a glycoprotein, where a glucose residue is linked to the amino acid Tyr to form the glycoprotein glycogenin (Lomako *et al*., 2004). The attached glycogen chain may have a molecular size over 10 million Da. Plants might have a related protein to which starch is linked. Starch is also a long branched polymer of α-linked Glc residues. The alpha linkages between Glc residues in glycogen and starch help them to structurally fold in a compact form and conserve space. Plants also synthesize probably one of the largest glycosylated molecules, which is cellulose, a linear polymer of β-linked Glc residues, which may have a size of many millions of daltons. The beta linkage of the Glc residues in cellulose causes it to be more linear and unable to fold like glycogen or starch. Crustaceans generate chitin, which is a linear polysaccharide of GlcNAc residues in beta linkage to each other and may have a size greater than 10 million Da. Some glycosaminoglycans in animals can also have enormous sizes. Hyaluronic acid, a polysaccharide of alternating GlcNAc and GlcA residues, has a size greater than 5 million Da. Animal heparin is an anticoagulant polysaccharide composed of alternating sulfated GlcNAc and GlcA or IdA residues with a size up to 40 000 Da. In animals, glycoproteins come in all sizes; some are extremely large and some are very small. For example, a human mucin termed MUC4 that is present in mucous secretions is a membrane-associated glycoproteins over 2 µm

in length, with a polypeptide backbone of over 900 000 Da (>8000 amino acids) that is heavily glycosylated on multiple threonine and serine residues (Lamblin *et al*., 2001). The actual molecular weight of such a mucin approaches several million daltons (*see* Article 70, ***O*-glycan processing**, Volume 6). One of the smallest glycoproteins in humans is CD59, which is the principal cellular inhibitor of the C5b-9 membrane attack complex (MAC) of human complement (Smith, 2004). CD59 is a GPI-anchored membrane glycoprotein with only 77 amino acids in the mature protein (*see* Article 66, **GPI anchors**, Volume 6 and Article 77, **Glycosylphosphatidylinositol anchors – a structural perspective**, Volume 6). It has a glycan in amide linkage to an asparagine residue at position 18 and the C-terminal amino acid is linked to a GPI-anchor in the membrane. The glycan attached to that asparagine residue can account for 30–50% of the total mass of the glycoprotein. Glycolipids also vary greatly in size but they are almost always smaller than glycoproteins (Hakomori, 2000). Typical glycolipids in animal cells range in size from 1000 to 4000 Da. One of the largest polymeric complexes containing sugar in nature is the peptidoglycan on bacteria that is cross-linked, as seen, for example, in mycobacteria (Crick *et al*., 2001), to provide a virtual single molecular net covering that promotes pathogenesis and protects bacteria from osmotic shock.

8. Functions of glycosylation

Glycosylation has many important biological functions. Glycans function as structural components (e.g., glycosaminoglycans and proteoglycans, cellulose, chitin, glycolipids), energy sources (e.g., sucrose, glycogen, starch), and recognition components for carbohydrate-binding proteins (CBPs). Many different glycans are known to be bound by CBPs (also termed lectins) (Sharon and Lis, 2004; Taylor and Drickamer, 2003). As discussed below, defects in glycosylation often result in disease and death. Glycosylation of proteins in the cytosol and nucleus by the addition of a single monosaccharide GlcNAc to Ser or Thr amino acids in proteins can regulate gene expression, protein interactions, and protein turnover (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6). During biosynthesis of glycoproteins in the rough endoplasmic reticulum of animal cells, the addition of glycans to Asn amino acid residues is part of a process of protein folding termed quality control (QC) (Helenius and Aebi, 2004; Parodi, 1999). This QC process is complex and involves a large number of enzymes that modify glycans and proteins that recognize the glycans on newly synthesized glycoproteins. Defects in the QC process regarding glycoproteins are often detrimental to cell survival. Glycoproteins on the surfaces of cells participate in cell–cell adhesion and recognition through binding by CBPs, as is seen for sperm–egg binding essential to fertilization, and in white blood cell (lymphocyte) adhesion and extravasation or movement into tissues and lymph nodes (Lowe, 2002; Lowe and Marth, 2003; Sharon and Lis, 2004). Surface glycoconjugates are involved in a myriad of cell signaling activities, since virtually all proteins on the surfaces of animal cells are glycosylated. Glycoconjugates within the extracellular matrix are essential to tissue integrity and to cell adhesion within the matrix. Many hormones are glycoproteins,

such as erythropoietin, which stimulates red blood cell production. Glycosylation of toxins (xenobiotics) is a part of the mechanism for clearance and detoxification. A typical toxin glycosylation is the addition of GlcA from UDP-GlcA donors in the liver, resulting in enhanced toxin solubility and inhibition of its toxin function. Even common hormones, such as testosterone and estrogen, are glycosylated by the addition of GlcA. Glycoconjugates are also essential nutrient components for cells. Thus, it is understandable that defects in glycosylation are associated with many diseases and disorders.

9. Disorders associated with glycosylation

A large number of human disorders and diseases are caused by defects in glycosylation. From a historical perspective, it has been known for nearly a century that some people have genetic defects resulting in an inability to degrade certain types of glycoconjugates (the process termed catabolism). Such defects are recognized as lysosomal storage diseases (Neufeld, 1991). Many dozens of such disorders are known. Some of the more noteworthy ones are Tay-Sachs disease, caused by a deficiency of β-hexosaminidase resulting in accumulation of an undegraded ganglioside (a glycolipid) GM2 in the brain, Gaucher disease, caused by a deficiency in β-glucocerebrosidase resulting in accumulation of several undegraded glycolipids, and Hurler syndrome, caused by a loss of α-L-iduronidase resulting in the accumulation of undegraded glycosaminoglycans. But in recent years, scientists have shown that many other human genetic disorders result from defects in the ability to glycosylate proteins or generally synthesize correctly glycosylated molecules. Many of these disorders are grouped as Congenital Disorders of Glycosylation (CDGs) (Jaeken, 2004). Because glycoconjugates are important in animal biology and development, these genetic diseases are rather uncommon. Some CDG patients are unable to synthesize sugar nucleotides, the precursors for glycosylation, while others inefficiently add glycans to newly synthesized proteins. But there are many other known defects in glycosylation outside the CDG categories, such as I-cell disease (Kornfeld and Sly, 1985), congenital muscular dystrophy (Endo and Toda, 2003), and defects in glycosaminoglycan biosynthesis (Zak *et al*., 2002). In I-cell disease, patients lack the enzyme that adds GlcNAc-phosphate residues to Man residues on lysosomal hydrolases. Normally, the presence of phosphate on Man residues is required for the correct localization of lysosomal hydrolases in their eventual destination within the lysosome. However, for I-cell patients, lysosomal enzymes are not correctly targeted to lysosomes and the fibroblasts in a patient's body inefficiently degrade a wide variety of glycoconjugates and proteins. In some forms of congenital muscular dystrophy, patients lack an enzyme that adds GlcNAc residues to O-linked mannose residues on glycoproteins, resulting in muscle weakness (*see* Article 71, ***O*-Mannosylation**, Volume 6). Mouse genetics has also shown the requirement for correct glycosylation in animal development (Lowe and Marth, 2003). Mice engineered to lack some glycosyltransferases die as embryos or fail to develop properly postpartum. Tumor cells also appear to express altered

glycoconjugates due to changes in glycosyltransferase and glycoprotein expression (Alper, 2003; Hakomori, 2001; Varki and Varki, 2001). It is now clear that glycosylation of macromolecules is a common and essential biological process.

10. Conclusions

In the past decade, there has been an explosion of interest in glycosylation and the biological functions of glycoconjugates. Each cell type within an animal appears to generate a unique repertoire of glycoconjugate structures, determined in large part by the regulation of the glycosylation machinery. Glycans function as structural components, energy sources, and they also function as recognition components for a large array of carbohydrate-binding proteins. A large number of genes in higher animals regulate glycosylation, and genetic defects in these pathways result in developmental abnormalities and embryonic death.

Acknowledgments

The author thanks Dr. Ziad Kawar for critical comments on the manuscript and gratefully acknowledges the support of the NIH (AI 48075 and CH/HD54832-01).

References

Alper J (2003) Glycobiology. Turning sweet on cancer. *Science*, **301**, 159–160.

Apweiler R, Hermjakob H and Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta*, **1473**, 4–8.

Bertozzi CR and Kiessling LL (2001) Chemical glycobiology. *Science*, **291**, 2357–2364.

Cairns AJ (2003) Fructan biosynthesis in transgenic plants. *Journal of Experimental Botany*, **54**, 549–567.

Crick DC, Mahapatra S and Brennan PJ (2001) Biosynthesis of the arabinogalactan-peptidoglycan complex of mycobacterium tuberculosis. *Glycobiology*, **11**, 107R–118R.

Davies GJ and Henrissat B (2002) Structural enzymology of carbohydrate-active enzymes: implications for the post-genomic era. *Biochemical Society Transactions*, **30**, 291–297.

Drickamer K and Taylor ME (2002) Glycan arrays for functional glycomics. *Genome Biology*, **3**, REVIEWS1034.

Endo T and Toda T (2003) Glycosylation in congenital muscular dystrophies. *Biological & Pharmaceutical Bulletin*, **26**, 1641–1647.

Hakomori S (1999) Antigen structure and genetic basis of histo-blood groups A, B and O: their changes associated with human cancer. *Biochimica et Biophysica Acta*, **1473**, 247–266.

Hakomori S (2000) Traveling for the glycosphingolipid path. *Glycoconjugate Journal*, **17**, 627–647.

Hakomori S (2001) Tumor-associated carbohydrate antigens defining tumor malignancy: basis for development of anti-cancer vaccines. *Advances in Experimental Medicine and Biology*, **491**, 369–402.

Helenius A and Aebi M (2004) Roles of N-linked glycans in the endoplasmic reticulum. *Annual Review of Biochemistry*, **73**, 1019–1049.

Jaeken J (2004) Congenital disorders of glycosylation (CDG): update and new developments. *Journal of Inherited Metabolic Disease*, **27**, 423–426.

Kikuchi S, Shinpo K, Takeuchi M, Yamagishi S, Makita Z, Sasaki N and Tashiro K (2003) Glycation–a sweet tempter for neuronal death. *Brain Research. Brain Research Reviews*, **41**, 306–323.

Kornfeld R and Kornfeld S (1985) Assembly of asparagine-linked oligosaccharides. *Annual Review of Biochemistry*, **54**, 631–664.

Kornfeld S and Sly WS (1985) Lysosomal storage defects. *Hospital Practice (Off Ed)*, **20**, 71–75, 78–82.

Laine RA (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10(12) structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology*, **4**, 759–767.

Lamblin G, Degroote S, Perini JM, Delmotte P, Scharfman A, Davril M, Lo-Guidice JM, Houdret N, Dumur V, Klein A, *et al*. (2001) Human airway mucin glycosylation: a combinatory of carbohydrate determinants which vary in cystic fibrosis. *Glycoconjugate Journal*, **18**, 661–684.

Lis H and Sharon N (1993) Protein glycosylation. Structural and functional aspects. *European Journal of Biochemistry*, **218**, 1–27.

Lomako J, Lomako WM and Whelan WJ (2004) Glycogenin: the primer for mammalian and yeast glycogen synthesis. *Biochimica et Biophysica Acta*, **1673**, 45–55.

Lowe JB (2002) Glycosylation in the control of selectin counter-receptor structure and function. *Immunological Reviews*, **186**, 19–36.

Lowe JB and Marth JD (2003) A genetic approach to Mammalian glycan function. *Annual Review of Biochemistry*, **72**, 643–691.

Neufeld EF (1991) Lysosomal storage diseases. *Annual Review of Biochemistry*, **60**, 257–280.

Parodi AJ (1999) Reglucosylation of glycoproteins and quality control of glycoprotein folding in the endoplasmic reticulum of yeast cells. *Biochimica et Biophysica Acta*, **1426**, 287–295.

Pereira-Chioccola VL and Schenkman S (1999) Biological role of Trypanosoma cruzi trans-sialidase. *Biochemical Society Transactions*, **27**, 516–518.

Power PM and Jennings MP (2003) The genetics of glycosylation in Gram-negative bacteria. *FEMS Microbiology Letters*, **218**, 211–222.

Rademacher TW, Parekh RB and Dwek RA (1988) Glycobiology. *Annual Review of Biochemistry*, **57**, 785–838.

Raetz CR and Whitfield C (2002) Lipopolysaccharide endotoxins. *Annual Review of Biochemistry*, **71**, 635–700.

Reiter WD (2002) Biosynthesis and properties of the plant cell wall. *Current Opinion in Plant Biology*, **5**, 536–542.

Samuel G and Reeves P (2003) Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydrate Research*, **338**, 2503–2519.

Schaffer C and Messner P (2004) Surface-layer glycoproteins: an example for the diversity of bacterial glycosylation with promising impacts on nanobiotechnology. *Glycobiology*, **14**, 31R–42R.

Scheible WR and Pauly M (2004) Glycosyltransferases and cell wall biosynthesis: novel players and insights. *Current Opinion in Plant Biology*, **7**, 285–295.

Sharon N and Lis H (2004) History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology*, **14**, 53R–62R.

Smith LJ (2004) Paroxysmal nocturnal hemoglobinuria. *Clinical Laboratory Science*, **17**, 172–177.

Spiro RG (2002) Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, **12**, 43R–56R.

Sutton-Smith M, Morris HR, Grewal PK, Hewitt JE, Bittner RE, Goldin E, Schiffmann R and Dell A (2002) MS screening strategies: investigating the glycomes of knockout and myodystrophic mice and leukodystrophic human brains. *Biochemical Society Symposium*, **69**, 105–115.

Taylor ME and Drickamer K (2003) *Introduction to Glycobiology*, Oxford University Press: Oxford.

Upreti RK, Kumar M and Shankar V (2003) Bacterial glycoproteins: functions, biosynthesis and applications. *Proteomics*, **3**, 363–379.
Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, **3**, 97–130.
Varki A, Cummings RD, Esko J, Freeze H, Hart G and Marth J (1999) *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, New York.
Varki A and Varki NM (2001) P-selectin, carcinoma metastasis and heparin: novel mechanistic connections with therapeutic implications. *Brazilian Journal of Medical and Biological Research*, **34**, 711–717.
Wilson IB (2002) Glycosylation of proteins in plants and invertebrates. *Current Opinion in Plant Biology*, **12**, 569–577.
Yan Q and Lennarz WJ (1999) Oligosaccharyltransferase: a complex multisubunit enzyme of the endoplasmic reticulum. *Biochemical and Biophysical Research Communications*, **266**, 684–689.
Young WW Jr (2004) Organization of Golgi glycosyltransferases in membranes: complexity via complexes. *The Journal of Membrane Biology*, **198**, 1–13.
Zachara NE and Hart GW (2004) O-GlcNAc a sensor of cellular state: the role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress. *Biochimica et Biophysica Acta*, **1673**, 13–28.
Zak BM, Crawford BE and Esko JD (2002) Hereditary multiple exostoses and heparan sulfate polymerization. *Biochimica et Biophysica Acta*, **1573**, 346–355.
Zitzmann N, Mehlert A, Carrouee S, Rudd PM and Ferguson MA (2000) Protein structure controls the processing of the N-linked oligosaccharides and glycosylphosphatidylinositol glycans of variant surface glycoproteins expressed in bloodstream form Trypanosoma brucei. *Glycobiology*, **10**, 243–249.

Specialist Review

Protein phosphorylation analysis by mass spectrometry

Hanno Steen

Children's Hospital and Harvard Medical School, Boston, MA, USA

Judith A. Jebanathirajah

Harvard Medical School, Boston, MA, USA

1. Introduction

One of the most common types of posttranslational protein modifications is protein phosphorylation. In prokaryotes many different amino acid residues including histidine, aspartic acid, glutamic acid, lysine, arginine, and cysteine, residues can be found in their phosphorylated forms, whereas the main targets for phosphorylation in eukaryotes are serine, threonine, and tyrosine residues (Yan *et al*., 1998). It is estimated that about one-third of all proteins in mammalian cells are phosphorylated at any given time and that about 5% of a vertebrate genome encodes for protein kinases and protein phosphatases (Hunter, 1998), underscoring the importance of this type of protein modification. The presence of this large variety of protein kinases and phosphatases permits the use of reversible protein phosphorylation in a vast number of different, highly regulated pathways and functions. The fast reversibility of protein phosphorylation renders it versatile for cell regulation, such as metabolism, cell division, cell differentiation, cell proliferation, and signal transduction, which require "off"- and "on"-states and fast switching between these two states for fully functional regulatory mechanisms.

The exact localization of phosphorylation sites is of pivotal interest when trying to understand detailed mechanisms of cellular regulatory processes. This calls for fast and sensitive methods for the detection and localization of protein phosphorylation. The most sensitive methods for the detection of phosphorylation utilize radioactive phosphorus isotopes such ^{32}P or ^{33}P. In these experiments, radioactive phosphorus isotopes are incorporated in vivo or in vitro into proteins of interest, which are subsequently digested, and phosphopeptide mapping is performed either using two-dimensional thin layer chromatography or liquid chromatography. Autoradiography and/or densitometry can then be used to localize and/or quantitate the labeled spots or fractions (Hunter and Sefton, 1991). Edman-degradation-based approaches can also be used, whereby the release of radioactivity

during one particular cycle from the immobilized peptide can be correlated with the site of phosphorylation (Campbell and Morrice, 2002). However, the in vivo incorporation of the radioactive isotope is not possible in the case of tissue samples and is very inefficient in the case of cell culture due to the presence of endogenous unlabeled adenosinetriphosphate (ATP) such that large amounts of radioactively labeled ATP are required to achieve a degree of in vivo phosphorylation that is sufficient for sensitive detection. Thus, there is a need for other, nonradioactive methods, especially for the analysis of in vivo protein phosphorylation. One commonly followed strategy utilizes generic anti-phosphoserine, -threonine, or -tyrosine antibodies either for profiling and the subsequent comparison of different (cell) states (Godovac-Zimmermann *et al*., 1999). Antibodies can also be used for detailed analyses of phosphorylation sites. This is performed by using site-specific antibodies in contrast to antibodies against global phosphoamino acid residues used for the former type of experiment described. However, the latter approach requires the identification and localization of the phosphorylation site by other means such as mutational scanning or by utilizing radioactivity or nonradioactive methods. The most versatile nonradioactive method is mass spectrometry, and this technology provides unrivaled sensitivity and speed when compared with traditional biochemical methods. This review summarizes the most commonly used mass spectrometry–based approaches for phosphorylation analysis and highlights new developments in the field of phosphoproteomics as of end of 2004.

2. What are the challenges in analyzing protein phosphorylation?

Currently, protein identification by mass spectrometry is a standard technique that can be automated to a large degree by companies and core facilities around the world (*see* Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5). This is in contrast to protein phosphorylation analysis (or protein modification analysis in general) which is still far from being routine and/or automated. The main difference between these procedures is that for identification purposes any set of peptides or even one peptide is sufficient for an unambiguous protein identification, whereas for protein phosphorylation analysis, the modified peptide has to be found and analyzed such that for the "complete" protein characterization, close to 100 % sequence coverage is required – and this is not normally achieved because of certain intrinsic limitations: (1) mass spectrometry is most sensitive in a particular *m/z* range such that peptides that are on either extreme ends of the *m/z* scale are inefficiently detected; (2) the cleavage efficiency varies for different proteins and different domains within the proteins, such that some expected peptides are not generated; (3) peptides have different ionization and detection responses, and thus signal intensity can easily vary by orders of magnitude for peptides derived from the same protein (*see* Article 16, **Improvement of sequence coverage in peptide mass fingerprinting**, Volume 5). The latter point is far from being fully understood; however, some preliminary studies showed that the peptide length, number of lysine and glutamic acid residues as well as the propensity to

form α-helices seem to have an effect on the ionization and detection efficiencies of the different peptides (Schirle *et al*., 2004). These general reasons for not observing all peptides are further complicated by other phosphospecific reasons for phosphopeptides being missed during mass spectrometric analysis. These include substoichiometric degrees of phosphorylation and heterogeneous phosphorylation, both of which lead to lower amounts of phosphopeptides of a specific *m/z* as compared to unmodified peptides from the same protein of interest. Furthermore, phosphopeptides tend to have lower ionization/detection efficiencies when analyzed by MALDI mass spectrometry especially when 4-hydroxy-α-cyanocinnamic acid is the matrix used – one of the reasons for this might be that fragmentation of the labile phosphomoiety is induced by in-source decay during the ionization process (*see* Article 11, **Nano-MALDI and Nano-ESI MS**, Volume 5 and Article 14, **Sample preparation for MALDI and electrospray**, Volume 5).

Serine and threonine-phosphorylated peptides cause additional problems during mass spectrometric analyses because of their fragmentation behavior when using collisional activation conditions normally used to obtain sequence-revealing fragment ions. The product-ion spectra of these species are normally dominated by fragment ions resulting from the neutral loss of the phosphomoiety as phosphoric acid (H_3PO_4 corresponding to 98 Da; note that the difference in the mass spectrum is given by n*98/z, n being the number of phosphomoieties on the peptide of interest and z being the charge state of the precursor of interest). Since this loss of phosphoric acid is energetically favored, other sequence/phosphorylation site-revealing fragment ions are often absent (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5). This is especially relevant in ion trap instruments where a resonant frequency is used for the excitation (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) collisionally induced dissociation (CID) once the predominant fragment ions that lost the phosphoric acid group are formed they are off-resonance with the applied excitation frequency and are no longer excited/fragmented (see Figure 1). This is in contrast to collisional activation in high-pressure multipoles (non-ion trap); here the fragment ion that underwent the neutral loss usually fragments further by additional collisions while traversing the collision cell, such that useful sequence information is often obtained.

Since this neutral-loss ion is normally the first fragment ion formed, the originally phosphorylated residues are converted into dehydroalanine (69 Da) and dehydroaminobutyric acid (83 Da) residues. These residues are indicative of the phosphorylation site and have to be considered during interpretation of the product-ion spectra in order to localize the site of phosphorylation; unfortunately, these residues are often not considered by some of the current protein-identification programs and as such many phosphopeptides may not be identified.

This problem of a predominant loss of the phosphomoiety can be circumvented by using a novel (peptide) fragmentation process called electron capture dissociation (ECD) (Zubarev *et al*., 1998). This process is based on the capture of thermal electrons that induce radical chain reactions resulting in the predominant peptide cleavage at the N–αC bond without fragmenting the most labile side chain modifications including phosphorylation (see e.g., Stensballe *et al*., 2000) such that the site of modification can be easily and unambiguously identified even if a

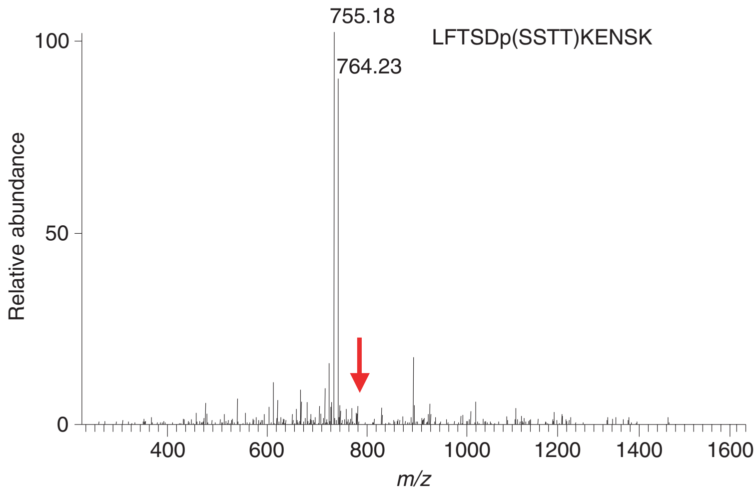


Figure 1 Product ion spectrum of a precursor at *m/z* 813.16 (red arrow) acquired on a quadrupole ion trap mass spectrometer. The predominant neutral loss of 98 Da (H_3PO_4) and 116 Da ($H_3PO_4 + H_2O$) is clearly visible. However, other sequence-revealing fragment ions are only of minor abundance such that it is only possible to identify the peptide but not the exact site of phosphorylation (shown as p(SSTT))

highly phosphorylated species containing numerous serine and threonine residues is fragmented (see e.g., Figure 2).

The disadvantages of this method include (1) the requirement of expensive FT-ICR instruments (*see* Article 5, **FT-ICR**, Volume 5), (2) low efficiency of electron capture and hence low fragmentation yield, and (3) the fact that ECD fragmentation in an LC-compatible timescale is still a challenge. However, preliminary data for

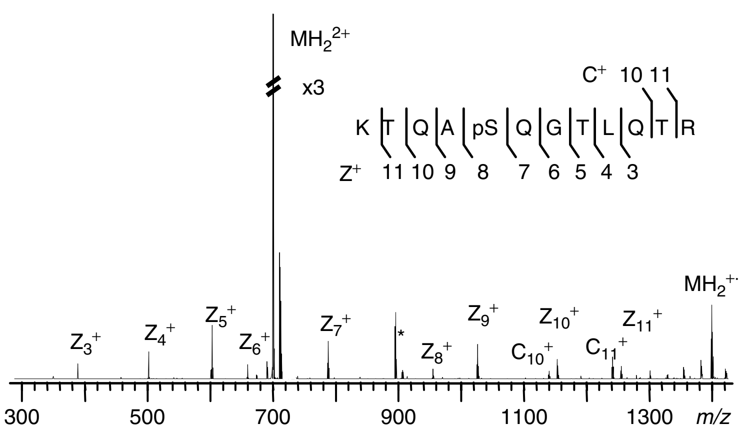


Figure 2 An ECD product ion spectrum of the phosphorylated peptide KTQA(pS)QGTLQTR. In contrast to collision induced activation, no loss of H_3PO_4 (−98 Da) is observed owing to the nature of ECD and instead of B- and Y-type fragment ions, only C- and Z-type fragment ions are generated (Figure courtesy of Dr. Bogdan Budnik, then University of Southern Denmark)

ECD fragmentation in quadrupole ion traps have recently been presented (Silivra *et al*., 2004). In addition, another novel fragmentation approach, Electron Transfer Dissociation (ETD) has been introduced in linear ion traps. ETD utilizes gas-phase reactions between protonated peptides and a basic radical anions, transferring electrons upon collision, thereby inducing ECD-like peptide fragmentations with all its advantages (and also its disadvantage of low fragmentation yields) in a cheaper type of mass spectrometer. This technique shows promise for obtaining intact phosphopeptide fragmentation data in LC/MS compatible timescales (Syka *et al*., 2004).

3. Given the obstacles, how is protein phosphorylation analyzed by mass spectrometry?

3.1. Brute force

The most common means of phosphorylation analysis by mass spectrometry is by using "Brute Force" methods. This approach entails the extensive analysis of the sample of interest by LC/MS prior to exhaustive data mining using database search programs such as MASCOT, Sequest, Sonar, or ProID where mass differences of 80 Da (and multiples thereof) due to (multiple) phosphorylation are considered in database searches. Although such a strategy is valid as a first-pass analysis, the disadvantages are obvious; since only "observed" peptides (conditional on the parameter used for the data dependent MS^2 experiments) are fragmented, the vast majority of the fragmented peptides are unmodified peptides such that the majority of information collected is redundant as the identity of the protein of interest is already known after the first unmodified peptide from any given protein is successfully fragmented and searched against a database. This means that the analysis of phosphopeptides is only a side product of such a brute force strategy and the shortcomings of such an analysis are that (1) phosphopeptides of low abundance are often missed, especially during the analysis of more complex samples and (2) the analysis often only provides insufficient fragmentation information to unambiguously localize the phosphorylation because of poor phosphospecific fragmentation (see above or below). A further complication might be caused by sulfation, another protein modification that gives rise to the same mass increase as phosphorylation but shows a different fragmentation behavior. Although the mass shift of the precursor ion caused by a modification is taken into account by the currently used database search programs, the differing fragmentation patterns are often not considered such that sulfation can easily be mistaken for phosphorylation. It was thought that sulfation only occurred on tyrosine residues of extracellular proteins; however, a recent study provides evidence that serine or threonine residues of intracellular mammalian proteins can also be sulfated (Medzihradszky *et al*., 2004).

In order to avoid all the shortcomings associated with the brute force technique listed above, different approaches are necessary for more efficient phosphorylation analysis. The most commonly used approaches utilize of the selective/specific mass spectrometric detection of phosphorylated species based on (1) their characteristic

fragmentation behavior upon collisionally induced dissociation and/or (2) the selective enrichment of phosphorylated species using affinity purification strategies. Often, combinations of these methods are utilized. Other approaches including chemical derivatization and phosphatase treatment have been described as well.

3.2. Selective detection

As detailed previously, phosphorylated peptides exhibit a characteristic behavior on collisionally induced activation in tandem mass spectrometers: (1) In the negative ion mode (detection of anions, i.e., deprotonated species), PO_3^- is observed as characteristic fragment ion at *m/z* −79. It is highly specific for any type of phosphorylated species, that is, in the case of phosphopeptides and proteins this ion at *m/z* −79 is observed irrespective of whether serine, threonine, or tyrosine is phosphorylated. (2) In the positive ion mode (detection of cations, i.e., protonated species), phosphopeptides tend to lose the phosphomoiety as a neutral phosphoric acid molecule giving a "signature" ion pair in the MS/MS spectrum spaced by a *m/z* of 98/*z* (*z* being the charge state of the precursor) for each phosphomoiety contained in the peptide of interest. A caveat of this signature fragmentation is that this loss is only observed for phosphoserine- and phosphothreonine-containing peptides. Phosphotyrosine-containing peptides give rise to a characteristic fragment ion at *m/z* 216.043, which can be utilized when high resolution, high accuracy tandem mass spectrometers such as quadrupole TOF instruments are employed (*see* Article 10, **Hybrid MS**, Volume 5; Steen *et al*., 2001; Steen *et al*., 2002).

3.3. Precursor ion scanning

Precursor ion scanning allows for the selective detection of only those parent ion species that give rise to the phosphospecific fragment ion upon peptide dissociation. Although precursor ion scanning has a superb specificity and is extremely sensitive such that species of interest hidden in the chemical noise can easily be detected, limitations have been described for precursor of *m/z* −79 and precursor of *m/z* 216.043 experiments (see Figure 3).

In the former case, the deprotonated peptide giving rise to the specific PO_3^- ion is preferentially formed in alkaline solution and requires that the instrument be run in negative ion mode. This prevents the immediate sequencing of the peptide as deprotonated ions normally do not give rise to sequence-revealing fragment ions. The latter phosphotyrosine-specific experiments require a high resolution, high accuracy quadrupole TOF instrument, which does not allow precursor ion scans in a LC-compatible timescale as the second mass analyzer, the TOF tube, cannot be operated in a mass filter mode. Instead, distinct product-ion spectra have to be acquired for each precursor mass prior to computing the intensity of the characteristic reporter ion versus the *m/z* value of the precursor. Use of such an approach requires 30 s or more and this time frame is incompatible with LC/MS experiments that are currently standard.

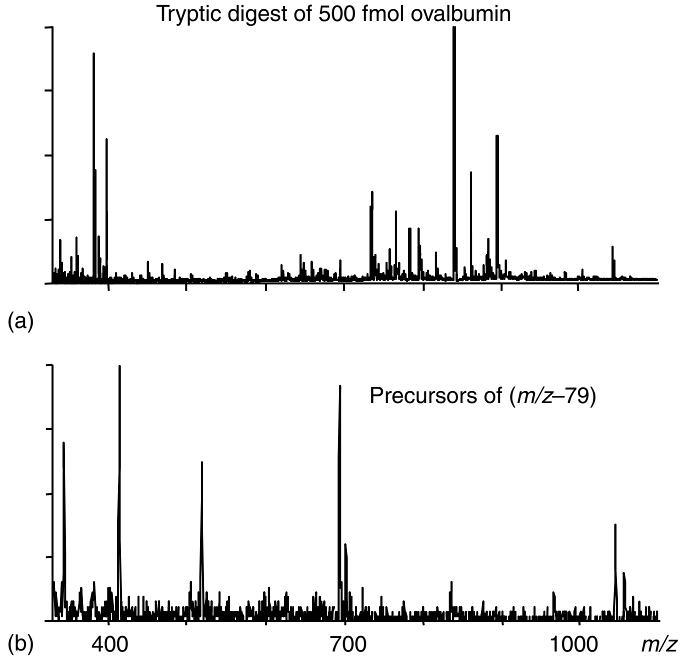


Figure 3 (a) Mass spectrum in negative ion mode of 500 fmol of a tryptic ovalbumin digest. Numerous peptide ion signals are apparent, however, it is not obvious which signals correspond to phosphorylated species. (b) Precursor ion experiment of ovalbumin that selectively detects the phosphorylated species giving rise to the characteristic PO_3^- fragment ion at *m/z* −79

Approaches to solve both of these limitations have recently been introduced:

1. Since deprotonated, anionic species can also be generated from strongly acidic solutions (e.g., LC effluents), if a negative potential is applied to the electrospray ionization (ESI) needle tip during a normal LC/MS experiment, a precursor ion scan in the negative ion mode can still be performed with an experimental strategy described as follows: a precursor of *m/z* −79 experiment in negative ion mode is performed on the deprotonated peptide ions eluting from the column. An ion signal in the precursor ion experiment triggers switching of the polarity of the instrument to the positive ion mode and a fragment ion spectrum of the protonated species, whose deprotonated complement gave rise to the signal in the negative ion mode precursor ion experiment, is acquired (Le Blanc *et al*., 2003). While performing this experiment, an *m/z* difference of 2 has to be considered to account for mass difference between the deprotonated and protonated species. Furthermore, one has to account for the fact that the observed degree of deprotonation does not necessarily reflect the degree of protonation, for example, a doubly deprotonated species does not necessarily give rise to a doubly protonated species such that attempts to fragment the phosphopeptides of interest can fail and/or additional scans have to be performed.
2. Niggeweg *et al*. recently introduced a way to circumvent the problem of precursor ion experiments on quadrupole TOF instruments. Fragment ion spectra

for each precursor mass were not acquired, instead the first quadrupole was operated in rf-only mode, that is, all precursors eluting off the column at any given time were fragmented simultaneously. The assignment of the precursor ion giving rise to the fragment ion species of interest at a given time is accomplished by correlating the LC elution profiles of the reporter ion(s) with the different peptide species eluting at that given moment. Possessing information about the exact elution time and exact precursor *m/z* value allows directed sequencing experiments to be performed in a consecutive LC/MS experiment of the same sample (Niggeweg *et al*., 2004).

3.4. Constant neutral loss scanning

The second fragmentation-specific detection method is the constant neutral loss scanning experiment that utilizes the characteristic neutral loss of phosphoric acid (H_3PO_4) from the phosphorylated species. This neutral loss reaction can be of significant abundance in triple quadrupole instruments or quadrupole TOF instruments, and is normally the predominant fragmentation in quadrupolar ion trap mass spectrometers. Although this loss occurs under acidic conditions in positive ion mode (i.e., it is easily compatible with LC/MS experiments), certain problems are caused by the fact that neutral losses are also charge-dependent, that is, a neutral loss of 98 Da translates into a *m/z* difference of 49 or 32.6 for doubly and triply charged species, respectively. Another problem with this method is a result of the low-resolution instruments (ion traps and triple quadrupole instruments) normally used for this type of experiment, where a loss of *m/z* 32 (methanesulfenic acid CH_3SOH) that occurs from a doubly charged peptide containing an oxidized methionine residue can be mistaken for a loss of a phosphoric acid moiety from a triply charged species. This is a particular problem when fast, low-resolution/low-accuracy scans are performed.

An interesting combination of the neutral loss and the brute force approach can be easily realized in quadrupole ion traps (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) during a standard LC/MS run by triggering an MS^3 experiment when a peptide selected for fragmentation during a normal data-dependent acquisition displays a characteristic neutral loss of *m/z* 49 or 32.6 (see Figure 4). This MS^3 experiment of the "dephosphorylated" species (dephosphorylated in this context refers to the loss of H_3PO_4 and not HPO_3, as observed when proteins are treated with phosphatases) provides sequence information such that the position of the newly formed dehydroalanine and/or dehydroaminobutyric acid residues and hence the localization of the former site of phosphorylation can be determined. Although such an approach still requires that the species of interest is selected for fragmentation (many phosphorylated species are found in substoichiometric amounts making this an unlikely event), it allows one to selectively spend valuable measuring time on species of interest, thereby increasing the chances of an unambiguous localization of the phosphorylation sites (Beausoleil *et al*., 2004).

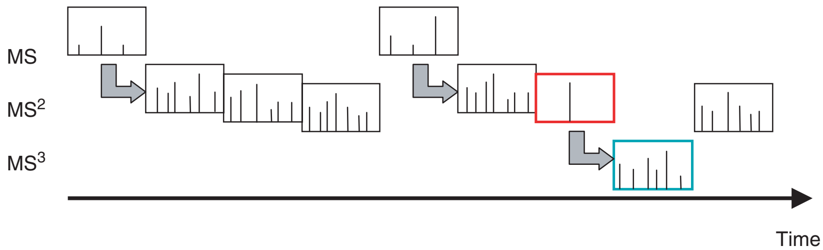


Figure 4 Schematic workflow of the data-dependent pseudoneutral loss/MS^3 experiments. The three most intense peptide ion signal are chosen for subsequent peptide sequencing by MS^2 after a survey MS. If one of the product-ion spectra shows a characteristic loss of phosphoric acid as major fragment ion (see red MS^2 spectrum), this particular fragment ion is further sequenced in an MS^3 experiment, often enabling the unambiguous localization of the phosphorylation site. Subsequently, the instrument returns to its normal data-dependent MS-MS^{2}cycle

3.5. Selective enrichment

Several different strategies for the selective enrichment of phosphopeptides and phosphoproteins have been devised. These are either based on antibodies or on affinity chromatographic methods utilizing the special physicochemical properties of phosphomonoesters, that is, phosphopeptides and phosphoproteins. It should be noted that the two approaches are somewhat complimentary because antibodies work much better for proteins, whereas affinity chromatographic approaches are more applicable to phosphopeptides.

3.6. Phosphospecific antibodies

Several commercial antibodies against generic motifs containing phosphoserine, phosphothreonine, and phosphotyrosine residues are available. The anti-phosphotyrosine-specific antibodies have been used in numerous studies for the selective enrichment of tyrosine-phosphorylated proteins, subsequent protein identification, and phosphorylation mapping, and hence have demonstrated their applicability (see Figure 5 and e.g., Godovac-Zimmermann *et al.*, 1999; Steen *et al.*, 2002).

Nevertheless, it should be noted that the different commercial monoclonal anti-phosphotyrosine antibodies such as 4G10 (Upstate) or RC20 (Transduction Laboratory) do show different specificities to an extent that a mixture of different anti-phosphotyrosine antibodies should be used in order to enrich for as many different tyrosine-phosphorylated species as possible. The performance of the commonly used anti-phosphotyrosine antibodies for enrichment/affinity purification is superior to the performance of the commercially available anti-phosphoserine and -threonine antibodies that do not show good specificity or yields of enrichment when used for immunoprecipitation (Gronborg *et al.*, 2002). However, Cell Signaling Technology has since shown that antibodies raised against peptide

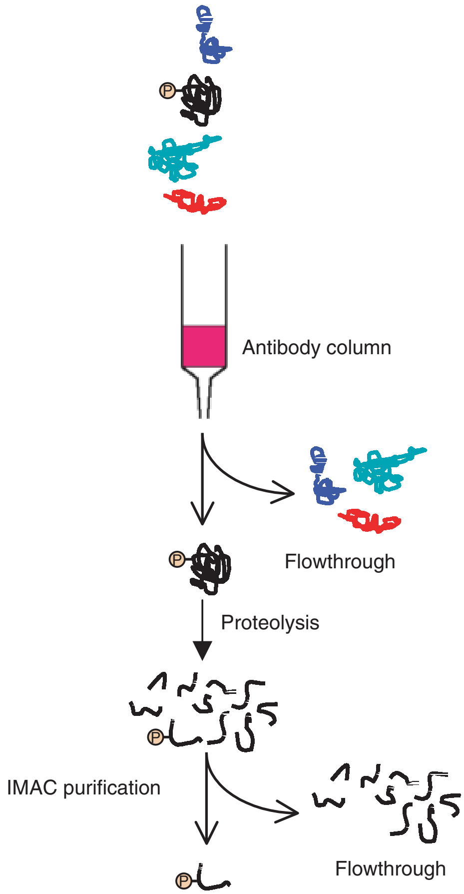


Figure 5 A general strategy using antibodies and immobilized metal affinity chromatography (IMAC) to enrich for phosphoproteins and subsequently for phosphopeptides

libraries composed of specific phosphorylated consensus motifs can be very useful for affinity purification of proteins containing the motif (Gronborg *et al*., 2002; Zhang *et al*., 2002a).

Antibodies are normally used for the enrichment of phosphorylated proteins although a recent study showed that antibodies can also be applied to phosphopeptides (Rush *et al*., 2005).

3.7. *Affinity chromatography*

It has been known for almost 20 years that phosphorylated species have a high affinity to Fe^{3+}-ions and that immobilizing this transition metal ion using different chelators provides a means to selectively enrich for phosphoproteins and phosphopeptides by Immobilized Metal Affinity Chromatography (IMAC; see Figure 6).

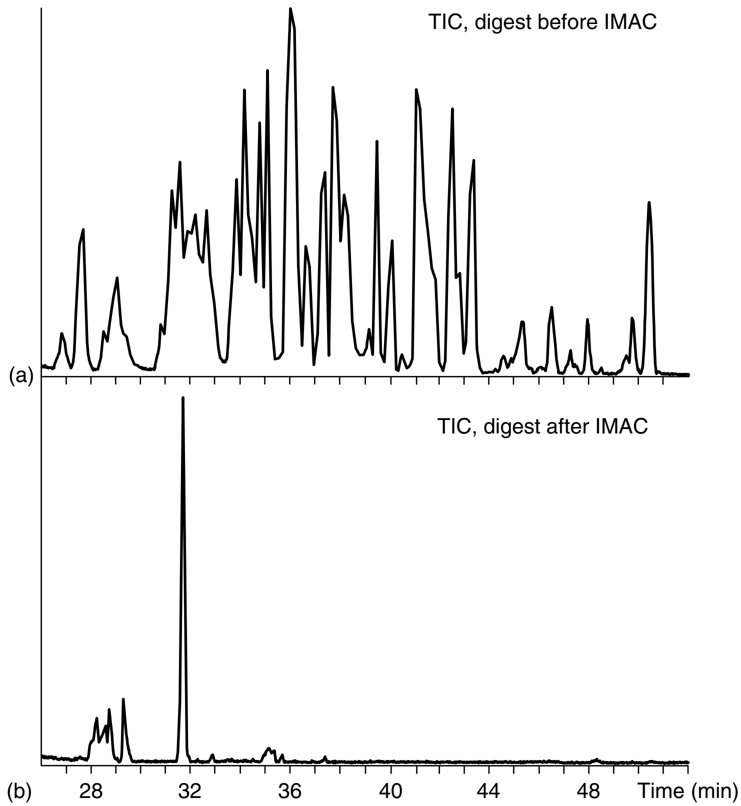


Figure 6 (a) The total ion chromatogram (TIC) of a tryptic digest of a phosphoprotein without immobilized metal affinity purification (IMAC). (b) The TIC of the same digest after IMAC enrichment of the phosphopeptides

Different metal cations and chelators have been tested and the most useful proven cations and chelators for proteomic purposes appear to be Fe^{3+}, Ga^{3+}, and ZrO^{2+}, and immobilized iminodiacetic acid (IDA) (e.g., POROS MC from Applied Biosystems) and nitrilotriacetic acid (NTA) (e.g., Silica-NTA from QIAGEN). These conclusions are based on studies testing the binding and the elution efficiencies of phosphorylated species. The question as to which combination is "the best" is debatable and is probably dependent on the particular sample and the operator's preference and experience (Nuhse *et al*., 2003; Posewitz and Tempst, 1999). Additional issues of concern with respect to IMAC for the enrichment of phosphopeptides are specificity (false-negatives) and selectivity (false-positives), that is, loss of some phosphorylated species and nonspecific binding of unphosphorylated peptides. The former issue is crucially dependent on the sample preparation and bed volume; that is, buffers with high salt concentrations or phosphate-based buffers severely affect the binding of phosphorylated species in general and underdimensioned bed volumes are biased toward multiply phosphorylated species (due to replacement-chromatography type processes). The selectivity of IMAC is complicated by amino acids within peptides with acidic

(Glu and Asp) and electron donor functionalities (e.g., His, Cys), which can show affinities to immobilized cations based on ionic and polar interactions. To increase the selectivity, Ficarro *et al*. devised a change in the established IMAC enrichment protocol for large-scale phosphoproteome studies by masking the acidic side chains of glutamic and aspartic acid residues by esterification (Ficarro *et al*., 2002). The use of this procedure is still debated (Brill *et al*., 2004), and there might be a sample-dependent trade-off between losses due to additional derivatization steps, increased complexity due to incomplete derivatization and side reactions, and lower selectivity.

Other strategies for the enrichment of subsets of phosphopeptides are based on the strong anionic character of the phosphomoiety at pH values > 2. The overall charge state of phosphorylated tryptic peptide is reduced to 1+ or less at a pH of 3 provided no additional basic residues besides the C-terminal lysine and arginine of tryptic peptides are present in the peptides. Thus, cation exchange beads can be used to enrich for phosphopeptides by collecting the flowthrough or the early/low salt fractions (Beausoleil *et al*., 2004). This is complemented by anion exchange strategies, by selecting salt concentrations at which predominantly phosphorylated species, especially multiply phosphorylated species, bind to anion exchange material and unphosphorylated peptides do not bind to the anion exchange material (Nuhse *et al*., 2003).

Although ion exchange strategies do not show both the specificity and selectivity of IMAC purification, they are still extremely useful for the enrichment of a subset of phosphopeptides and/or providing a method to reduce the complexity of the sample due to their robustness and superior capacity.

3.8. Derivatization of the phosphomoiety

The only enzymatic reaction on protein- or peptide-linked phosphomoieties that has been described (to our knowledge) is the complete removal of the phosphate group with phosphatases, which regenerates the original protein/peptide such that no information pertaining the site of phosphorylation is obtained. Thus, chemical means have to be exploited for the derivatization of the phosphomoiety leading to selective detection or enrichment of the original phosphopeptides. So far two different chemical strategies have been described:

1. Zhou *et al*. (2001) described a selective derivatization of the phosphomoiety utilizing 1-Ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC) chemistry subsequent to a stepwise protection of the different functional groups within the peptides. This approach led to the formation of thiol-containing phosphoamidates, allowing the selective immobilization and purification of the phosphorylated species. Although conceptually interesting, this approach proved to be difficult because of the six different reaction steps and numerous reversed-phase purifications in between these reactions, compromising the throughput and sensitivity of this approach.
2. The other widely tried and tested chemical derivatization strategy applicable to serine and threonine phosphorylation, but not to tyrosine phosphorylation, is

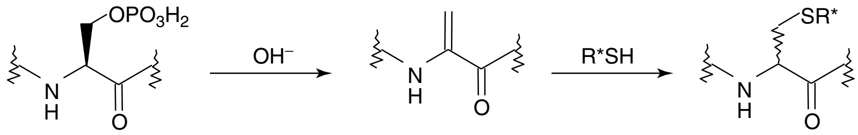


Figure 7 A reaction scheme showing the derivatization of phosphoserine residues using the β-elimination/Michael addition strategy. Please note that the former L-serine residue is racemized such that diastereomeric peptide species are often separated by LC

based on β-elimination of the phosphomoiety followed by a Michael addition of a strong nucleophile, such as a thiol to the in situ formed dehydroalanine and/or dehydro-2-aminobutyric acid (Meyer *et al*., 1986) (see Figure 7). Numerous thiols have been used for this derivatization strategy, enabling the specific detection of phosphopeptides based on identification of characteristic peak pairs (Molloy and Andrews, 2001); by precursor ion scanning in positive ion mode (Steen and Mann, 2002); using specific proteolytic cleavages (Knight *et al*., 2003; Rusnak *et al*., 2002); by allowing the selective immobilization utilizing solid-phase chemistry approaches (McLachlin and Chait, 2003; Qian *et al*., 2003); and using avidin-biotin interactions (Oda *et al*., 2001). An additional advantage of this β-elimination/Michael addition approach is that isotopically labeled compounds can be incorporated, which allow for the relative quantitation of protein phosphorylation (see below). However, this approach has limitations: (1) the reaction conditions leading to partial hydrolysis (Steen and Mann, 2002), (2) side reactions with unmodified serine and threonine residues (Li *et al*., 2003; McLachlin and Chait, 2003), (3) side reactions with other serine and threonine modifications such as O-glycosylation (Wells *et al*., 2002), or (4) side reactions with modified and incompletely reduced cysteine residues (Steen and Mann, 2002). This indicates that the application of this chemical derivatization strategy to phosphopeptide mapping of substoichiometric phosphopeptides ($<5\%$) may be problematic.

3.9. *Other strategies for the phosphorylation analysis*

One commonly used approach for the identification of phosphopeptides within mixtures utilizes alkaline phosphatase for the complete dephosphorylation of the sample. Acquiring spectra – usually MALDI-TOF mass spectra – before and after the phosphatase treatment allows the unambiguous identification of phosphopeptides based on the mass shift of 80 Da or multiples thereof (Larsen *et al*., 2001; Yip and Hutchens, 1992). However, since the enzymatic dephosphorylation regenerates the original unmodified peptide structure, it is not possible to determine the site of original phosphorylation using the phosphatase-treated sample. For the detailed phosphorylation site determination, another fraction of the sample has to be reanalyzed, but since the masses of the phosphopeptides are known sensitive and time efficient, directed experiments can be performed.

Another alternative for qualitative and quantitative (see below) analysis of protein phosphorylation was introduced by Lehmann and coworkers (Wind *et al*.,

2001). They coupled liquid chromatography to inductively coupled plasma (ICP) mass spectrometry. In ICPMS, the sample is atomized/ionized at 8000 K using an inductively coupled plasma, which allows for the selective detection of phosphorylated species eluting from the reversed-phase column based on the presence of a signal for $^{31}P^{+}$, an element that is not present in unmodified proteins/peptides. Such an experimental setup enables the identification of a peptide as being phosphorylated species, and accurate quantitation of the phosphorylated species can be made. However, as the peptides are completely atomized, it is not possible to identify the phosphopeptide and/or to determine the site of phosphorylation, that is, further "traditional" LC/MS experiments have to be performed. This can be done in a more directed way, since the retention times of the phosphorylated species are known.

3.9.1. Quantitation of protein phosphorylation

Since protein phosphorylation is highly dynamic, it becomes increasingly apparent that it is no longer sufficient to merely identify the site of phosphorylation. Information pertaining to the extent of phosphorylation, and/or the comparison between the phosphorylation states of different samples must also be acquired. Although quantitative proteomics is in its infancy, several approaches, mainly based on stable isotope labeling, have been introduced (reviewed in Lill, 2003; Steen and Pandey, 2002), and it is obvious that the quantitation of protein phosphorylation combines the difficulties of quantitative proteomics with the problems encountered in modification-specific proteomics. In addition, quantitation of phosphorylation has two aspects: (1) the relative quantitation to determine the increase or decrease of phosphorylation in different samples and (2) absolute quantitation to determine the degree of phosphorylation.

Apart from the methods that aim to reduce complexity of the samples by labeling particular amino acid residues, for example, ICAT (Gygi *et al*., 1999; *see also* Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5), all other described methods are, in principle, applicable to the relative quantitation of phosphorylation (or modifications in general), especially if combined with the different specific enrichment or detection strategies described above. Examples for application of metabolic isotopic labeling strategies or peptide derivatization approaches subsequent to proteolysis of samples for the quantitation of protein phosphorylation can be found in Ibarrola *et al*. (2003) and Bonenfant *et al*. (2003).

In addition, several phosphospecific mass spectrometry–based relative quantitation methods have been introduced. One of them is based on β-elimination/Michael addition utilizing pairs of isotopologues for the labeling of the different samples. Proofs of concept have been provided by Weckwerth *et al*. (2000) who used normal or deuterium-labeled ethanthiol or by Goshe *et al*. (2001) who combined β-elimination/Michael addition with the ICAT strategy by using isotopologues of thiol-terminated biotin derivatives – a strategy coined PhIAT for Phosphoprotein Isotope-Coded Affinity Tag. Both of these approaches suffer from the same limitations of β-elimination/Michael addition as described above and thus should be applied and the results interpreted with caution.

Another very sensitive and precise approach that can be used for the quantitation of proteins and other general modifications makes use of synthetic, isotopically labeled peptides as internal standards, which can be spiked into the samples in defined quantities (Ayad *et al*., 2003; Gerber *et al*., 2003). Using this method, it is sufficient to ensure that equal amounts of the synthetic phosphopeptide are added to the different samples for relative quantitation, whereas for absolute quantitation the synthetic peptide must be quantitated by, for example, amino acid analysis such that defined amounts of the internal standard are added to the sample(s). This method shows very good accuracies and detection limits. If a particular phosphorylation site is being monitored in a large number of samples, the method is very cheap; however, as the method entails peptide synthesis, purification, and quantitation, it is expensive for single-use experiments.

Other methods for the absolute quantitation of the degree of phosphorylation entail splitting the sample and labeling the two fractions with different isotopologues prior to dephosphorylation of one of the fractions. Subsequently, the two fractions are combined and analyzed by mass spectrometry; the ratio of the two dephosphorylated isotopologues species enable the calculation of the degree of phosphorylation (Hegeman *et al*., 2004; Zhang *et al*., 2002b).

In addition to the different quantitation approaches that are based on the comparison of the signal intensities of isotopologues, there is an increasing notion that the current LC/MS setup with autosampler is sufficiently quantitative and reproducible such that semi quantitative information with respect to the degree of phosphorylation can be derived simply by monitoring the peak intensity during an LC/MS experiment. Using this setup provides a general, cheap, and fast way to quantitate protein phosphorylation. In order to normalize for run-to-run variations and for different amounts of starting material in the different samples (a point that is often ignored in many methods and studies), the signal intensity of one or more "robust" peptides from the numerous unmodified peptides identified from the phosphorylated protein of interest can be used as an internal standard (Ruse *et al*., 2002; Steen *et al*., 2005). This allows for the relative quantitation of protein phosphorylation.

Assuming that the decrease (increase) of the ion signal intensity of the phosphorylated species is reflected by the increase (decrease) of the ion signal intensity of the unphosphorylated complement, it is possible to calculate the ratio of the ionization/detection responses of both peptide species. This in turn allows for the calculation of the degree of phosphorylation by simply comparing the signal intensities of the phosphorylated and the unphosphorylated peptide without assuming equal ionization/detection responses of the two species. At least two different measurements are necessary for a singly phosphorylated species, whereas three different states are necessary to calculate the relative ionization/detection responses for doubly phosphorylated peptides and their singly phosphorylated and unphosphorylated complements, and so on. Once the relative ionization/detection responses have been determined, they can be used for the particular peptide signals in following experiments – presuming the experimental and instrumental parameters did not change – such that the extent of phosphorylation can easily be calculated even in a single sample. If only one sample is available, phosphatase treatment for complete or partial dephosphorylation can be used to generate samples

representing different degrees of phosphorylation, which can then be used to calculate the extent of phosphorylation (Steen *et al*., 2005).

4. Conclusion and outlook

The large number of different mass spectrometric methods available for the analysis of protein phosphorylation underscores the notion that no single method provides a comprehensive phosphorylation map of a protein. As such, numerous methods have to be applied and correspondingly larger amounts of protein are necessary for phosphorylation mapping in comparison to simple protein-identification studies (Chen *et al*., 2002; Vihinen and Saarinen, 2000). Global phosphoproteomic experiments pose different problems to single protein phosphorylation mapping; in such cases, the goal is to identify as many phosphorylation sites as possible. This generates a huge amount of valuable information such that missing some phosphorylation sites is irrelevant. Impressive progress has been made in this field starting with the yeast phosphoproteome study by Ficarro *et al*. (2002) who identified several hundreds of phosphorylation sites. Since this study, several other studies have been published, which clearly show that large phosphoproteomics experiments can be performed (Beausoleil *et al*., 2004; Nuhse *et al*., 2003). However, to avoid terming these large-scale phosphoproteomic studies as "interesting stamp collections", large-scale quantitative phosphorylation studies should be carried out as well. These types of experiments would allow the comparison of phosphorylation states between, for example, different cell states, thus providing a more biological insight into signaling processes. Unfortunately, most of the quantitative approaches only allow for the comparison of two states/samples. However, these types of studies will become a reality with the novel labeling reagent such as the iTRAQ from Applied Biosystems that facilitates the comparison of up to four cell states/samples and with the implementation of software that allows the robust quantitation of particular peptide ion signals without any labeling. This limitation should be solved and more quantitative phosphoproteomic studies can be expected in the near future.

Acknowledgments

The authors would like to thank their former and present colleagues at University of Southern Denmark, Harvard Medical School and MDS-SCIEX. In addition, we would like to apologize to all those authors whose work in the field of mass spectrometric analysis of protein phosphorylation could not be cited due to space constraints.

References

Ayad NG, Rankin S, Murakami M, Jebanathirajah J, Gygi S and Kirschner MW (2003) Tome-1, a trigger of mitotic entry, is degraded during G1 via the APC. *Cell*, **113**, 101–113.

Beausoleil SA, Jedrychowski M, Schwartz D, Elias J, Villen J, Li J, Cohn MA, Cantley LC and Gygi SP (2004) Identification.

Bonenfant D, Schmelzle T, Jacinto E, Crespo JL, Mini T, Hall MN and Jenoe P (2003) Quantitation of changes in protein phosphorylation: a simple method based on stable isotope labeling and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 880–885.

Brill LM, Salomon AR, Ficarro SB, Mukherji M, Stettler-Gill M and Peters EC (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Analytical Chemistry*, **76**, 2763–2772.

Campbell D and Morrice N (2002) Identification of protein phosphorylation sites by a combination of mass spectrometry and solid phase Edman sequencing. *Journal of Biomolecular Techniques*, **13**, 119–130.

Chen SL, Huddleston MJ, Shou W, Deshaies RJ, Annan RS and Carr SA (2002) Mass spectrometry-based methods for phosphorylation site mapping of hyperphosphorylated proteins applied to Net1, a regulator of exit from Mitosis in yeast. *Molecular & Cellular Proteomics: MCP*, **1**, 186–196.

Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF and White FM (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnology*, **20**, 301–305.

Gerber SA, Rush J, Stemman O, Kirschner MW and Gygi SP (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6940–6945.

Godovac-Zimmermann J, Soskic V, Poznanovic S and Brianza F (1999) Functional proteomics of signal transduction by membrane receptors. *Electrophoresis*, **20**, 952–961.

Goshe MB, Conrads TP, Panisko EA, Angell NH, Veenstra TD and Smith RD (2001) Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Analytical Chemistry*, **73**, 2578–2586.

Gronborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON and Pandey A (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Molecular & Cellular Proteomics*, **1**, 517–527.

Gygi S, Rist B, Gerber S, Turecek F, Gelb M and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **19**, 994–999.

Hegeman AD, Harms AC, Sussman MR, Bunner AE and Harper JF (2004) An isotope labeling strategy for quantifying the degree of phosphorylation at multiple sites in proteins. *Journal of the American Society for Mass Spectrometry*, **15**, 647–653.

Hunter T (1998) The Croonian lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Philosophical Transactions of the Royal Society of London. Series B*, **353**, 583–605.

Hunter T and Sefton BM (1991) Protein Phosphorylation, Part A and Part B. *Methods in Enzymology*, 200 and 201.

Ibarrola N, Kalume DE, Gronborg M, Iwahori A and Pandey A (2003) A proteomic approach for quantitation of phosphorylation using stable isotope labeling in cell culture. *Analytical Chemistry*, **75**, 6043–6049.

Knight ZA, Schilling B, Row RH, Kenski DM, Gibson BW and Shokat KM (2003) Phosphospecific proteolysis for mapping sites of protein phosphorylation. *Nature Biotechnology*, **21**, 1047–1054.

Larsen MR, Sorensen GL, Fey SJ, Larsen PM and Roepstorff P (2001) Phospho-proteomics: evaluation of the use of enzymatic de- phosphorylation and differential mass spectrometric peptide mass mapping for site specific phosphorylation assignment in proteins separated by gel electrophoresis. *Proteomics*, **1**, 223–238.

Le Blanc JC, Hager JW, Ilisiu AM, Hunter C, Zhong F and Chu I (2003) Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications. *Proteomics*, **3**, 859–869.

Li W, Backlund PS, Boykins RA, Wang G and Chen HC (2003) Susceptibility of the hydroxyl groups in serine and threonine to beta-elimination/Michael addition under commonly used moderately high-temperature conditions. *Analytical Biochemistry*, **323**, 94–102.

Lill J (2003) Proteomic tools for quantitation by mass spectrometry. *Mass Spectrometry Reviews*, **22**, 182–194.

McLachlin DT and Chait BT (2003) Improved beta-elimination-based affinity purification strategy for enrichment of phosphopeptides. *Analytical Chemistry*, **75**, 6826–6836.

Medzihradszky KF, Chalkley RC and Burlingame AL (2004) 80 Da modification of protein serine and threonine residues: not always phosphate! *Paper presented at 52nd ASMS Conference on Mass Spectrometry, Nashville, 23–27 May 2004*.

Meyer HE, Hoffmann-Posorske E, Korte H and Heilmeyer LM Jr (1986) Sequence analysis of phosphoserine-containing peptides. Modification for picomolar sensitivity. *FEBS Letters*, **204**, 61–66.

Molloy MP and Andrews PC (2001) Phosphopeptide derivatization signatures to identify serine and threonine phosphorylated peptides by mass spectrometry. *Analytical Chemistry*, **73**, 5387–5394.

Niggeweg R, Gentzel M and Wilm M (2004) A multiplexed, precise, fast and sensitive precursor ion scan like analysis mode for LC quadrupole time of flight based instruments. *Presented at 52nd ASMS Conference on Mass Spectrometry, Nashville, 23–27 May 2004*.

Nuhse TS, Stensballe A, Jensen ON and Peck SC (2003) Large-scale analysis of in Vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Molecular & Cellular Proteomics*, **2**, 1234–1243.

Oda Y, Nagasu T and Chait BT (2001) Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nature Biotechnology*, **19**, 379–382.

Posewitz MC and Tempst P (1999) Immobilized gallium(III) affinity chromatography of phosphopeptides. *Analytical Chemistry*, **71**, 2883–2892.

Qian WJ, Goshe MB, Camp DG II, Yu LR, Tang K and Smith RD (2003) Phosphoprotein isotope-coded solid-phase tag approach for enrichment and quantitative analysis of phosphopeptides from complex mixtures. *Analytical Chemistry*, **75**, 5441–5450.

Ruse CI, Willard B, Jin JP, Haas T, Kinter M and Bond M (2002) Quantitative dynamics of site-specific protein phosphorylation determined using liquid chromatography electrospray ionization mass spectrometry. *Analytical Chemistry*, **74**, 1658–1664.

Rush J, Moritz A, Lee KA, Quo A, Goss VL, Spek EJ, Zhang H, Zha XM, Polakiewicz RD and Comb MJ (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature Biotechnology*, **23**, 94–101.

Rusnak F, Zhou J and Hathaway GM (2002) Identification of phosphorylated and glycosylated sites in peptides by chemically targeted proteolysis. *Journal of Biomolecular Techniques* **13**, 228–237.

Schirle M, Mallick P, Gagneur J, Schmitt R, Aebersold R and Kuster B (2004) Proteome-wide prediction and verification of idiotypic peptides for protein identification. *Presented at 52nd ASMS Conference on Mass Spectrometry, Nashville, 23–27 May 2004*.

Silivra OA, Ivonin IA, Kjeldsen F and Zubarev RA (2004) Ion-electron reactions in a quadrupole ion trap: realization, first results and prospects. *Presented at 52nd ASMS Conference on Mass Spectrometry, Nashville, 23–27 May 2004*.

Steen H, Jebanathirajah JA, Springer M and Kirschner MW (2005) Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, In press.

Steen H, Kuster B, Fernandez M, Pandey A and Mann M (2001) Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Analytical Chemistry*, **73**, 1440–1448.

Steen H and Mann M (2002) A new derivatization strategy for the analysis of phosphopeptides by precursor ion scanning in positive ion mode. *Journal of the American Society for Mass Spectrometry*, **13**, 996–1003.

Steen H and Pandey A (2002) Proteomics goes quantitative: measuring protein abundance. *Trends in Biotechnology*, **20**, 361.

Steen H, Pandey A, Andersen JS and Mann M (2002) Analysis of tyrosine phosphorylation sites in signaling molecules by a phosphotyrosine-specific immonium ion scanning method. *Science's STKE: Signal Transduction Knowledge Environment*, **2002**, PL16.

Stensballe A, Jensen ON, Olsen JV, Haselmann KF and Zubarev RA (2000) Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Communications in Mass Spectrometry*, **14**, 1793–1800.

Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J and Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9528–9533.

Vihinen H and Saarinen J (2000) Phosphorylation site analysis of semliki forest virus nonstructural protein 3. *The Journal of Biological Chemistry*, **275**, 27775–27783.

Weckwerth W, Willmitzer L and Fiehn O (2000) Comparative quantification and identification of phosphoproteins using stable isotope labeling and liquid chromatography/mass spectrometry. *Rapid Communications MS*, **14**, 1677–1681.

Wells L, Vosseller K, Cole RN, Cronshaw JM, Matunis MJ and Hart GW (2002) Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Molecular & Cellular Proteomics: MCP*, **1**, 791–804.

Wind M, Edler M, Jakubowski N, Linscheid M, Wesch H and Lehmann WD (2001) Analysis of protein phosphorylation by capillary liquid chromatography coupled to element mass spectrometry with 31P detection and to electrospray mass spectrometry. *Analytical Chemistry*, **73**, 29–35.

Yan J, Packer N, Gooley A and Williams K (1998) Protein phosphorylation: technologies for the identification of phosphoamino acids. *Journal of Chromatography. A*, **808**, 23–41.

Yip T-T and Hutchens TW (1992) Mapping and sequence-specific identification of phosphopeptides in unfractionated protein digest mixtures by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *FEBS Letters*, **308**, 149–153.

Zhang H, Zha X, Tan Y, Hornbeck PV, Mastrangelo AJ, Alessi DR, Polakiewicz RD and Comb MJ (2002a) Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *The Journal of Biological Chemistry*, **277**, 39379–39387.

Zhang X, Jin QK, Carr SA and Annan RS (2002b) N-Terminal peptide labeling strategy for incorporation of isotopic tags: a method for the determination of site-specific absolute phosphorylation stoichiometry. *Rapid Communications in Mass Spectrometry: RCM*, **16**, 2325–2332.

Zhou H, Watts JD and Aebersold R (2001) A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnology*, **19**, 375–378.

Zubarev RA, Kelleher NL and McLafferty FW (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society*, **120**, 3265–3266.

Specialist Review

Structure/function of *N*-glycans

Miriam V. Dwek
University of Westminster, London, UK

Susan Brooks
Oxford Brookes University, Oxford, UK

1. Background – the importance and function of N-linked glycosylation

There is an increasing interest in the posttranslational modification of proteins, and among these modifications, glycosylation appears to be one of the commonest and most functionally significant. Glycosylation is an extremely common event and analysis of gene sequences reveals that more than 50% of all proteins have one or more glycosylation site (Apweiler *et al*., 1999; *see also* Article 65, **Structure/function of *O*-glycans**, Volume 6, Article 66, **GPI anchors**, Volume 6, and Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6). Despite this, only a very small proportion of glycosylated proteins have been fully characterized as the analysis of glycans remains a technically challenging and specialist field.

2. Glycosylation of proteins – a brief overview

The oligosaccharides or glycans synthesized by human cells are composed of seven monosaccharide units: glucose (Glc), galactose (Gal), mannose (Man), fucose (Fuc), *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GalNAc), and sialic or neuraminic acids (SA). SA is a negatively charged molecule and its presence often alters the chemical properties of the glycan substantially. The individual monosaccharides link together in α- or β-linkages between different carbon atoms of their ring structures (e.g., $\alpha 1 \rightarrow 3$, $\beta 1 \rightarrow 4$, $\alpha 2 \rightarrow 3$ etc.), giving rise to a bewildering heterogeneity of potential glycan structures. Although all the theoretical permutations do not exist in vivo, glycans still remain an extraordinarily structurally diverse group of compounds and their biochemical analysis is technically challenging.

3. What are N-linked glycoproteins?

Glycoproteins are proteins containing one or more covalently linked carbohydrate group(*see* Article 62, **Glycosylation**, Volume 6). In N-linked glycoproteins, the

oligosaccharide chain is attached by its first monosaccharide residue, GlcNAc, in a β-N-glycosidic bond to the nitrogen of the amino group of asparagine (Asn) on the polypeptide chain.

4. The enzymes of glycosylation – glycosidases and glycosyltransferases

Unlike proteins and nucleic acids, oligosaccharides are not directly the products of genes; rather, they are built by the coordinated activity of glycosidases (enzymes that cleave glycans) and glycosyltransferases (enzymes responsible for the attachment of monosaccharide units). Over 250 different enzymes are known to be involved in mammalian glycan synthesis, and the activity of 30 or more enzymes can be involved in synthesis of a single N-linked oligosaccharide chain.

Several enzymes are critical in the early stages of N-linked glycan synthesis. The first important enzyme is the oligosaccharyltransferase (OST), a multimeric enzyme complex, which is responsible for attaching a preformed oligosaccharide precursor from a dolichol lipid intermediate to the amide nitrogen of the Asn residue of the nascent polypeptide. Mannosidases and glucosidases then trim the large oligosaccharide precursor. Once the N-linked oligosaccharide has been attached to the protein and trimmed, it is extended by the sequential addition of monosaccharides. These monosaccharides are added via a transglycosylation reaction, catalyzed by a specific glycosyltransferase: here, a monosaccharide is transferred from a high-energy nucleotide donor, commonly uridine 5′-diphosphatase (UDP), guanosine 5′-diphosphatase, (GDP) or cytidine 5′-diphosphatase (CDP), onto the oligosaccharide acceptor. The high-energy nucleotide donors are synthesized in the cytosol and transported across the membrane of the endomembrane system by antiporters that ensure the reciprocal return of spent nucleotides to the cytosol. This step is essential for successful glycosylation as the presence of spent nucleotides is potently inhibitory to transglycosylation reactions.

Until recently, it was believed that a single glycosyltransferase was responsible for catalyzing the transglycosylation reaction between a given donor and acceptor to yield a single, specific linkage – often referred to as the “one enzyme – one linkage rule”. It is now known that although usual, this is not universally the case. Sometimes, more than one enzyme, usually belonging to the same enzyme family, can catalyze identical linkages, and also, sometimes, a single enzyme may be capable of catalyzing more than one reaction.

5. Location and retention of the enzymes of glycosylation in the secretory pathway

As the enzymes of glycosylation act sequentially in trimming and building oligosaccharides, so, generally, they are located along the secretory pathway in the same sequence in which they act to modify oligosaccharides. For example, the α-mannosidase I involved in early trimming of N-linked glycans is located in the *cis*-Golgi; the *N*-acetylglucosaminyltransferase I involved in early elongation is located

in *medial*-Golgi and β1,4 galactosyltransferase and α2,6 sialyltransferase involved in terminal glycosylation are located in the *trans*-Golgi network. The situation is, however, slightly more complex than this, with some glycosyltransferases found throughout the Golgi stacks, and others being located in different parts of the secretory pathway in different cell types, or in the same cell types under different physiological conditions, at different stages in development or in disease. Such enzyme subcompartmentalization may control the types and range of oligosaccharides synthesized by a cell. The regulation of glycan synthesis, which is not well understood, is the subject of much conjecture. Several theories have been put forward to explain the positioning of enzymes of glycosylation within the secretory pathway and their retention there, as opposed to their being delivered to the plasma membrane and/or other parts of the cell along with other proteins passing through the secretory pathway (Colley, 1997). As there is little sequence homology between glycosidases and glycosyltransferases from different families, a common Golgi retention signal, analogous to KDEL/HDEL endoplasmic reticulum localization sequence or akin to the mannose-6-phosphate system responsible for lysosome targeting, is unlikely. However, they do share a common domain structure. They are all type II transmembrane proteins consisting of a cytoplasmic tail, a signal anchor transmembrane domain, a stem region, and a large luminal catalytic domain. Their positioning and retention in the Golgi apparatus may be related to the conformation of their shared domain structure.

Two major theories are currently proposed to explain Golgi retention and positioning. The first suggests that differing concentrations of cholesterol in the lipid bilayer in different parts of the endomembrane system result in membranes with different characteristics of thickness and stiffness. It has been proposed that glycosyltransferases locate to different microdomains that have characteristics appropriate to their particular transmembrane segment and that their short transmembrane region allows retention within the Golgi apparatus because they are not physically able to incorporate into cholesterol-rich transport vesicles destined for the plasma membrane. An alternative "kin recognition" model proposes that glycosyltransferases that are involved in specific parts of the glycobiosynthetic pathway aggregate together through interactions between their stem regions into insoluble homo- or hetero-oligomers too large to enter transport vesicles. Further interaction between cytosolic components and the enzyme complex with Golgi apparatus membrane proteins may additionally anchor these complexes in place (Nilsson *et al*., 1993). There have also been suggestions that the pH gradient from the ER (endoplasmic reticulum) through the Golgi apparatus may play a role in correct localization of the glycosyltransferase enzymes.

6. Regulation of glycosyltransferase expression and activity – what glycans are synthesized?

The factors that determine the glycosylation repertoire of a cell are multiple, tissue specific, and highly regulated during differentiation and sometimes cellular proliferation, and are not well understood. Many glycosidases and glycosyltransferases

are constitutively expressed at low levels in all mammalian somatic cells, presumably to provide the everyday machinery of glycosylation for the biosynthesis of cellular glycoconjugates. Cells may also upregulate the expression of specific glycosyltransferases in response to changes in conditions, to supply a need for higher levels of expression of glycosylated molecules.

Glycosyltransferase and glycosidase expression appears to be regulated predominantly at the level of transcription but differential processing of mRNA, alterations in mRNA stability and differences in efficiency of translation may also influence final enzyme activity. The expression of closely related isoenzymes with different levels of activity or substrate preference appears to be one mechanism by which differential glycosylation is achieved.

Other factors that influence the glycosylation profile of the cell are the availability of sugar donors, competition for glycoprotein or sugar donor between different competing glycosyltransferase enzymes, competition for glycosyltransferase enzymes between different potential glycoproteins, the conformational structure of the protein, and the rate at which the glycoprotein passes through the secretory pathway – which influences contact time of transferase with substrate. Enzymes that have different pH requirements may also function optimally in different parts of the secretory pathway as the internal pH decreases from ER to *trans*-Golgi network.

7. N-linked protein glycosylation

N-linked protein glycosylation is much better understood than O-linked glycosylation because of the range of chemical inhibitors of various parts of the N-linked biosynthetic pathway which have facilitated its careful study (Elbein, 1987). N-linked glycosylation begins with the synthesis in the cytoplasm of a branched oligosaccharide ($Man_5GlcNAc_2$) linked to a dolichol molecule embedded in the membrane of the ER. The $Man_5GlcNAc_2$ structure is synthesized by the action of, sequentially, first, an *N*-acetylglucosaminylphosphotransferase enzyme, which catalyzes the reaction Dol-P + UDP-GlcNAc → GlcNAc-P-Dol + UMP; then, an *N*-acetylglucosaminyltransferase catalyzes the addition of a GlcNAc (from the donor UDP-GlcNAc) to form GlcNAc(β1 → 4)GlcNAc-P-Dol; this is called the *chitobiose core*. The chitobiose core is then acted upon by mannosyltransferase I and a mannose residue (from the donor GDP-Man) is added to yield Man(β1 → 4)GlcNAc(β1 → 4)GlcNAc-P-Dol. Then, mannosyltransferases II to IV come into play, adding, sequentially, three mannose residues to give the branched dolichol-linked oligosaccharide precursor $Man_5GlcNAc_2$ structure, illustrated in Figure 1(a).

This cumbersome hydrophillic structure, is flipped across the hydrophobic lipid bilayer of the plasma membrane of the ER, from the cytoplasmic face of the membrane and onto the lumenal face, in a process that remains incompletely understood (Helenius *et al*., 2002) but includes the Rft1 flippase protein (Helenius *et al*., 2002). The $Man_5GlcNAc_2$ structure is then extended by the addition of four mannose and three glucose units to form an oligosaccharide intermediate $Glc_3Man_9GlcNAc_2$ (Figure 1b). This oligosaccharide intermediate is then attached to an Asn residue (within a consensus sequence Asn-Xaa-Ser/Thr, where Xaa is any amino acid except proline) of the nascent polypeptide chain. Attachment

of the oligosaccharide intermediate to the Asn is catalyzed by the enzyme oligosaccharyltransferase (OST). N-linked glycosylation is thus a cotranslational event – it commences during the synthesis of the polypeptide.

The oligosaccharide intermediate is then sequentially trimmed by the action of first α-glucosidase I, which cleaves the terminal Glc residue, then α-glucosidase II, which cleaves the remaining two terminal Glc residues, then α-mannosidase, which removes a terminal Man residue. This yields the structure $Man_8GlcNAc_2$, illustrated in Figure 1(c). This trimming is important for a number of reasons. First, α-glucosidase I acts while the oligosaccharide intermediate is still attached to the OST and is necessary for its release. Secondly, trimming of the Glc residues is an important control mechanism in protein folding (Parodi, 2000). Folding glycoproteins are recognized by two ER resident lectins, membrane-bound calnexin, and

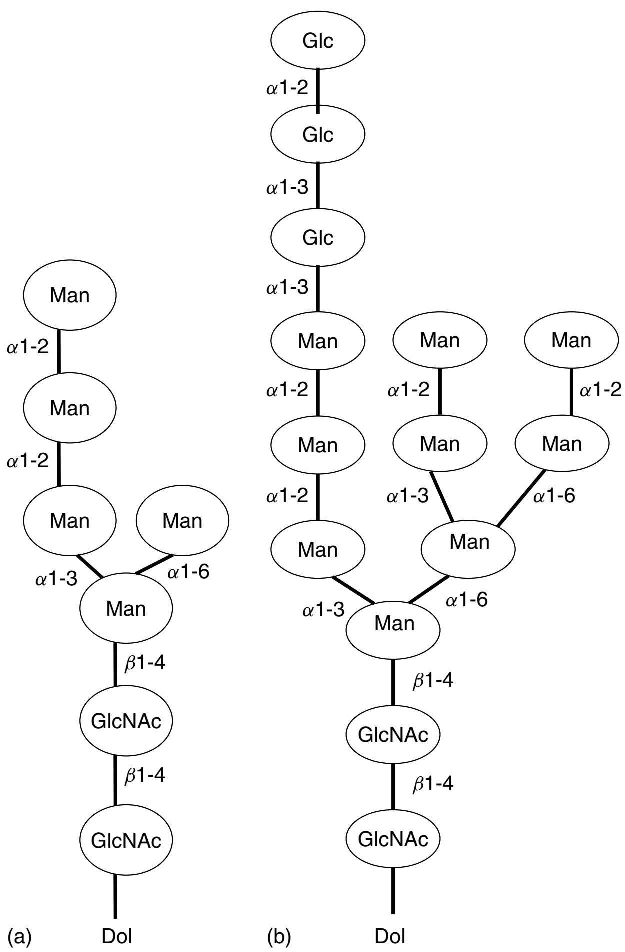


Figure 1 (a) Man5GlcNAc2 oligosaccharide precursor. (b) Glc3Man9GlcNAc2 oligosaccharyl intermediate. (c) $Man_8GlcNAc_2$. (d) $Man_5GlcNAc_2$. (e) A high mannose–type N-linked glycan. (f) A complex-type N-linked glycan

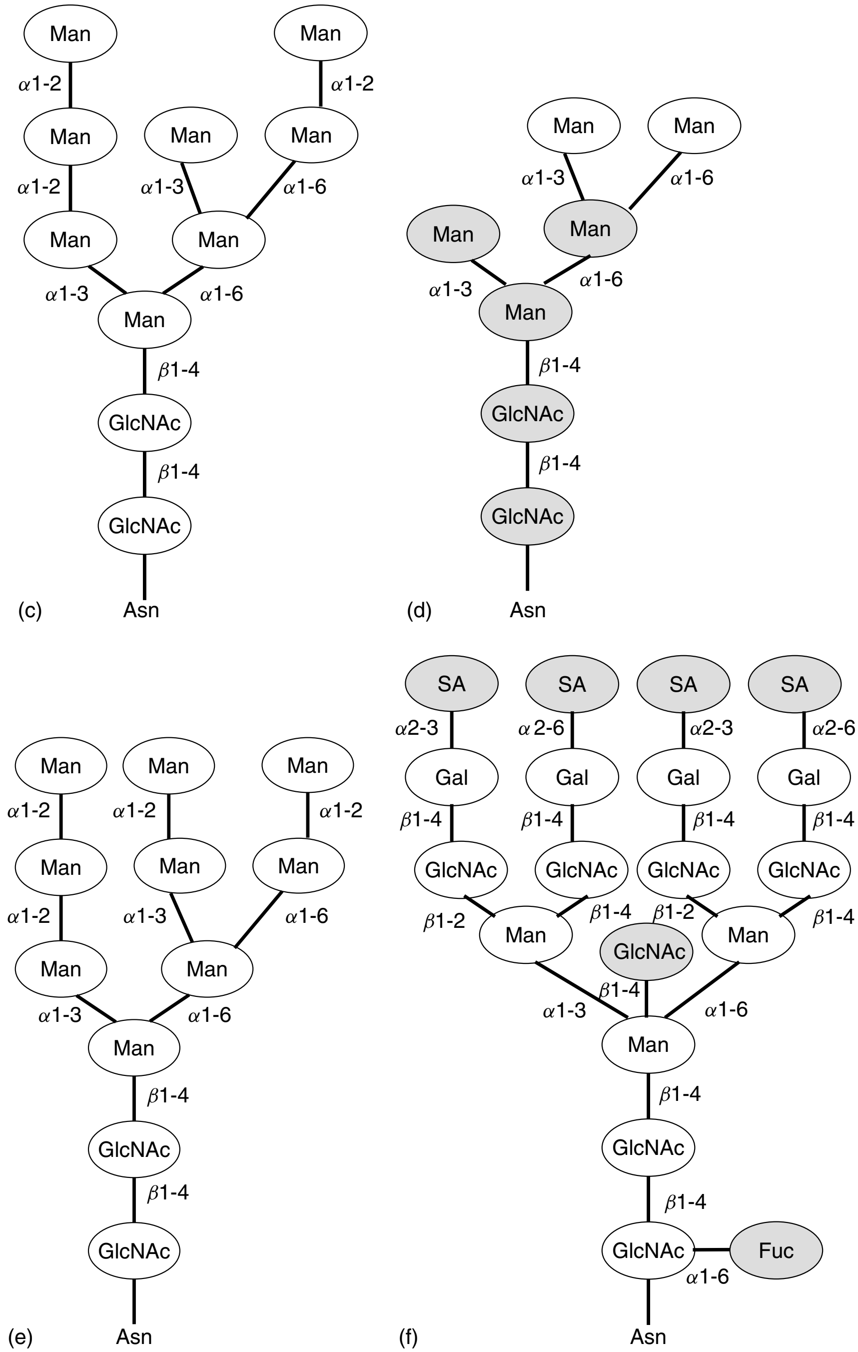


Figure 1 (*continued*)

its soluble homolog calreticulin. These lectins recognize the $\alpha 1 \rightarrow 3$ linked Glc of the partially trimmed $Glc_1Man_9GlcNAc_2$ intermediate, which is created from $Glc_3Man_9GlcNAc_2$ by the action of glucosidases I and II. Once they have bound the lectins, disulfide bonds are formed between cysteine residues on the polypeptide chain of the glycoprotein. If the protein is now correctly folded, the further action of glucosidase II results in the removal of the final $1 \rightarrow 3$ linked Glc yielding $Man_9GlcNAc_2$ and the calnexin and calreticulin are released from the glycoprotein. This signals the exit of the correctly folded protein from the ER and its transport to the Golgi apparatus. If, however, it is not properly folded, the physical features of the misfolded protein, for example, the presence of hydrophobic patches, trigger the action of UDP:Glc: glycoprotein glucosyltransferase (GT), which reglucosylates the oligosaccharide intermediate. The glycoprotein is then once again recognized by the lectins and returns through the cycle of reglucosylation and folding until correct folding is achieved. GT thus acts as a sensor of glycoprotein misconformation. Proteins that are terminally misfolded are eventually transported from the ER into proteosomes in the cytosol and degraded.

Another important signaling oligosaccharide is that which targets glycoproteins to the lysosome for degradation. This structure is mannose-6-phosphate and it is synthesized by the modification of the $Man_8GlcNAc_2$ intermediate structure by the addition of phosphorylated GlcNAc (catalyzed by the action of GlcNAc-1-phosphotransferase) to the subterminal Man; the GlcNAc is then cleaved, leaving mannose-6-phosphate. This structure is recognized by a mannose-6-phosphate receptor on the lysosome.

If the protein is successfully folded, it is transported to the Golgi apparatus and glycosylation continues. Further Man residues are removed to yield a $Man_5GlcNAc_2$ structure, illustrated in Figure 1(d), which may then either be trimmed further and extended by the addition of other monosaccharides or extended without further trimming. A heterogeneous range of branching N-linked glycans are thus possible; in general, the function of this microheterogeneity in glycan structure is poorly understood.

Nevertheless, all N-linked glycans share a common trimannosyl core, shaded in Figure 1(d), and are classified as (1) high mannose, (2) complex, or (3) hybrid type. High mannose type have between five and nine Man units attached to the trimannosyl core, illustrated in Figure 1(e); complex type have no further Man units attached to the trimannosyl core, but have branches containing chains of the disaccharide GlcNAc($\beta 1 \rightarrow 4$)Gal (termed polylactosamine chains). An example is given in Figure 1(f). Hybrid type, as the name suggests, contain a mixture of structural features of both high mannose and complex types. The hybrid and complex type N-linked glycans are significant because their synthesis appears to be a late evolutionary event that is associated with animals possessing a blood circulatory system. This is of relevance to the biotechnology industry as it is the cells of simpler organisms, such as bacteria, yeasts, plants, and insects that are commonly used for the production of recombinant proteins. The N-linked glycans that these cells synthesize differ significantly from those of human cells, and, if they are destined for human use, may result in them having unpredictable biological activities, serum half-life or immunogenicity.

The antennae of N-linked glycans may also be further elaborated by diverse glycan structures. A bisecting GlcNAc can be attached to the trimannosyl core and fucose residues can be attached to various antennae structures. Antennae are often terminated by $\alpha 2 \rightarrow 3$ or $\alpha 2 \rightarrow 6$ linked charged sialic acid residues. These features are illustrated, shaded, in the complex-type N-glycan in Figure 1(f).

8. N-linked glycosylation – functional aspects

Our understanding of the function of N-linked glycans on proteins has been helped considerably by the availability of inhibitors that serve to prevent the normal biosynthesis of N-linked structures. Defining the function of N-linked glycans on proteins has often been difficult because of technical aspects related to their evaluation and this topic is considered in more detail later. Glycosylation influences numerous aspects of protein biochemistry; for example, correct protein folding, susceptibility to proteases, serum half-life and functional activity. Many of these characteristics are of particular significance with regard to recombinant proteins destined for potential human therapeutic use, and therefore a better understanding of glycosylation and its control is of great relevance to the burgeoning biotechnology industry (Brooks, 2004). The glycosylation of proteins is pivotal to many processes involving cell–cell interactions, for example, fertilization, embryogenesis and development, immune defense, inflammation, and in disease processes, including cancer metastasis (Dennis *et al*., 1999; Orntoft and Vestergaard 1999) and glycosylation is of relevance in diverse aspects of biology and medicine. Our understanding of the function of N-linked glycans on proteins has been greatly assisted by the availability of specific inhibitors that prevent the normal biosynthesis of N-linked structures; another crucial technology has been site-directed mutagenesis to evaluate the effect of preventing glycan occupancy at particular N-linked glycosylation sites. Using these, and other technical approaches, it has been possible to elucidate the function of N-linked glycosylation. Many examples exist in which N-linked glycosylation is known to play a crucial role in protein function and some examples of these are given below.

8.1. Protein structure

The importance of N-linked glycans in ensuring correct protein folding was mentioned briefly earlier. This aspect of N-linked glycosylation has received much attention and the crucial role of N-linked glycans in ensuring the correct physiological structure of proteins has been shown in a number of different biological systems. In the case of the immunoglobulin G (IgG), for example, the constant domains in the hinge region of IgG contain N-linked glycans and the $\alpha 1 \rightarrow 3$ branch of the trimannosyl core on one N-linked glycan interacts with the trimannosyl core on the other glycan. This interaction changes the conformation of the N-linked glycan structures to allow interaction with the binding pocket on the constant domain surface of the IgG. This is thought to be an important structural feature of N-linked glycans on IgG in later enabling monocyte binding.

8.2. Protein clearance

One of the best-described functions for N-linked glycosylation of proteins is in modulating the clearance of serum proteins from the circulation. The role of N-linked glycans in process was first discovered in the 1960s when the copper-transporting protein ceruloplasmin was found to have a much reduced serum half-life concurrent with an absence of sialic acid residues on its N-linked glycans. Since then, it has become recognized that a carbohydrate-binding protein (lectin) in the liver, the hepatic asialoglycoprotein receptor recognizes asialo- (sialic acid free) serum proteins. When asialylated proteins bind to the receptor, they are removed from the circulatory system and degraded in the lysosomes. This important receptor is responsible for the clearance of, amongst other proteins, erythropoietin, whose serum half-life is dramatically reduced when non-sialylated N-linked glycans are attached (Ashwell and Harford, 1982). Importantly, other lectin-receptors have been reported that bind to the monosaccharides found in N-linked glycans and further functions relating to the role of N-linked glycosylation in protein clearance are likely to be described.

8.3. Enzyme function

There are numerous accounts showing a role for N-linked glycans in enzyme function; these include altering the kinetics of enzyme-substrate interactions, maintaining appropriate tertiary structure of the enzyme, or alternatively, the glycans may be important for other aspects of enzyme function such as the binding of divalent cations. In the case of tissue plasminogen activator (tPA), a serine protease that converts plasminogen to plasmin and induces fibrinolysis, the N-linked glycans appear to have a multitude of important functions. For example, glycan occupation at Asn 184 has been shown to allow the formation of a double-chain form of tPA, which results in an increase in the size of the lysine binding site in fibrin thereby reducing fibrinolysis; additionally, glycosylation has been shown to modulate the kinetics of enzyme–substrate interaction of both tPA and its substrate plasminogen (which is itself variably glycosylated).

8.4. Hormone function

There is now considerable evidence, mostly from studies of the pituitary hormones that show N-linked glycans are important molecules in signaling events. The N-linked glycans attached to lutenizing hormone (LH), thyroid-stimulating hormone (TSH) and chorionic gonadotrophin (CG) has been shown to have a variety of functions. The first is to ensure the correct folding of the two (α and β) subunits on LH, FSH (follicle stimulating hormone), and CG, each of which contains one or two occupied N-linked glycosylation sites. The N-linked glycans, of LH rather than containing terminal Gal capped with SA contain terminal GalNAc variably capped with sulfate groups. The presence of the terminal GalNAc with/without sulfate is important for serum clearance of the hormones by the Kupffer cells of the liver and

acts as an important feedback signal to ensure that more LH is produced by the pituitary gland and subsequent induction of ovulation via the LH/CG receptor in the ovary. Notably, the enzymes involved in the attachment of GalNAc and sulfate to the LH fluctuate according to the levels of estrogen in the circulation (Bielinska and Boime, 1995).

In conclusion, the diversity of N-linked glycans and their role in a wide range of protein functions means that there is considerable interest in their structural analysis. The well-defined biosynthetic pathway for N-linked glycans has aided our study of this type of glycosylation. It is now widely accepted that N-linked glycosylation is often important for a range of physiological functions including to ensure correct protein folding, clearance of proteins from the circulation, and as signaling molecules.

Further reading

Brooks SA, Dwek MV and Schumacher U (2002) *Functional and Molecular Glycobiology*, Publ. Bios Scientific Publishers: Oxford.

Hart C, Schulenberg B, Steinberg TH, Leung WY and Patton WF (2003) Detection of glycoproteins in polyacrylamide gels and on electroblots using Pro-Q Emerald 488 dye, a fluorescent periodate Schiff-base stain. *Electrophoresis*, **23**(4), 588–598.

Helenius J and Aebi M (2002) Transmembrane movement of dolichol linked carbohydrates during N-glycoprotein biosynthesis in the endoplasmic reticulum. *Seminars in Cell & Developmental Biology*, **13**, 171–178.

Lowe JB and Marth JD (2003) A genetic approach to mammalian glycan function. *Annual Review of Biochemistry*, **72**, 643–691.

Munro S (1998) Review: localisation of proteins in the Golgi apparatus. *Trends in Cell Biology*, **8**, 1–15.

Sears P and Wong C-H (1998) Enzyme action in glycoprotein synthesis. *Cell Molecular Life Sciences*, **54**, 223–252.

References

Apweiler R, Hermjakob H and Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta*, **1473**, 4–8.

Ashwell G and Harford J (1982) Carbohydrate-specific receptors of the liver. *Annual Reviews in Biochemistry*, **51**, 531–534.

Bielinska M and Boime I (1995) The glycoprotein hormone family: structure and function of the carbohydrate chains. In *Glycoproteins*, Montreuil J, Schachter H and Vliegenthart JEG (Eds.), Elsevier Science: Amsterdam, pp. 565–587.

Brooks SA (2004) Appropriate glycosylation of recombinant proteins for human use: implications of choice of expression system. *Molecular Biotechnology*, **28**(3), 241–255.

Colley KJ (1997) Mini review: Golgi localization of glycosyltransferases; more questions than answers. *Glycobiology*, **7**, 1–13.

Dennis JW, Granovsky M and Warren CE (1999) Glysoprotein glycosylation and cancer progression. *Biochimica et Biophysica Acta*, **1473**, 21–34.

Elbein AD (1987) Inhibitors of the biosynthesis and processing of N-linked oligosaccharide chains. *Annual Review of Biochemistry*, **56**, 497–534.

Helenius J, Ng DT, Marolda CL, Walter P, Valvano MA and Aebi M (2002) Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. *Nature*, **415** (6870), 447–450.

Nilsson T, Slusarewicz P, Hoe MH and Warren G (1993) Kin recognition. A model for the retention of Golgi enzymes. *FEBS Letters*, **330**, 1–4.

Orntoft T and Vestergaard EM (1999) Clinical aspects of altered glycosylation of glycoproteins in cancer. *Electrophoresis*, **20**, 362–371.

Parodi AJ (2000) Protein glucosylation and its role in protein folding. *Annual Review of Biochemistry*, **69**, 69–93.

Specialist Review

Structure/function of *O*-glycans

Anthony P. Corfield

University of Bristol, Bristol, UK

1. Introduction

Glycosylation is a major posttranslational modification affecting well over 50% of all proteins (Van den Steen *et al*., 1998; Varki *et al*., 1999). Recent developments in glycobiology have seen significant improvements in oligosaccharide sequencing, the establishment of databases of oligosaccharide structures, and clear demonstration of a range of biological functions associated with glycans. The *O*-glycans represent one of the three main groups of glycosylation modifications, the others being *N*-glycans and glycosylphosphatidyl inositol anchors (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6 and Article 66, **GPI anchors**, Volume 6) (Varki *et al*., 1999; Brockhausen *et al*., 2001). *O*-glycans fall into several groups: (1) the oligosaccharides having GalNAc linked to serine or threonine, often termed *mucin type glycosylation*, which is the focus of this article; (2) *N*-acetylglucosamine linked to serine or threonine (*O*-GlcNAc) on cytosolic and nuclear proteins (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6), (3) Fucose and Glucose O-linked to serine or threonine in the epidermal growth factor domains in a variety of proteins; (4) O-mannosylation detected in mammalian systems and identified in dystroglycan and brain proteoglycans, together with the oligomannosyl glycans found in yeasts (*see* Article 71, ***O*-Mannosylation**, Volume 6) and finally, (5) proteoglycans, a large family of molecules with distinct structural and functional differences to the other groups listed here. This short review outlines the structural characteristics of the large number of oligosaccharides linked to serine and threonine through GalNAc, and groups the molecular and biological functions associated with the presence of these *O*-glycans. More detailed consideration of the processing of the *O*-glycan chains and their role in protein processing is covered elsewhere in this volume (*see* Article 70, ***O*-glycan processing**, Volume 6).

2. The structure of O-linked glycans

2.1. O-GalNAc oligosaccharide structure

Oligosaccharides in the Ser(Thr)-GalNAc class vary in size from a single GalNAc through disaccharides to chains with up to 18 or more sugar residues. The

larger oligosaccharides comprise three distinct regions: (1) a core unit linking the oligosaccharide chain to serine and threonine residues in the protein; (2) a backbone, found in extended glycans, but not always present; and (3) a peripheral nonreducing terminus, which is largely responsible for the considerable diversity found in the *O*-glycans (Schachter and Brockhausen, 1992).

Addition of a single GalNAc residue occurs in a small number of proteins under normal conditions. This has been termed *the Tn antigen* and has been used as a cancer marker as it is also expressed in many tumors.

Eight core classes have been identified and are shown in Table 1. Core 1 and Core 2 are the most common structures found, with lower occurrence of Core 3 and 4, while the remaining structures are relatively rare. The type of core has a direct influence on the final structure of the extended oligosaccharides and is also determined by the biosynthetic pathways outlined briefly below (see Figure 1 for an example of biosynthetic pathways for Core 1–4 structures (Brockhausen *et al*., 2001)). This has implications for disease-related changes where the core class is modified resulting in abnormal, cancer-related glycosylation. Branched chains are generated from Core 2 and Core 4, and these are also subject to predictable extension through the biosynthetic pathways. In larger extended oligosaccharide chains, a backbone made up of Galβ1-3/4GlcNAc repeat units may be found, and these include the i/I antigens that may remain as nonreducing terminal structures (Table 1 (Schachter and Brockhausen, 1992; Brockhausen *et al*., 2001)). The huge variation of *O*-glycan oligosaccharide structures arises from the decoration of the core, and backbone units by a variety of peripheral units. These modifications include fucosylation, sialylation, sulfation, acetylation, and methylation. Many of these peripheral units are structures with biological function such as the blood group and sialylated/fucosylated/sulfated antigens. Some examples from the large databases of *O*-glycan structures are given in Table 1.

2.2. *Identification of O-glycan sites*

The identification of those serine and threonine residues in peptide sequences that are selected for O-glycosylation does not follow the simple tripeptide sequon model for N-linked oligosaccharides. Potentially any serine or threonine may be a site, and in contrast to *N*-glycan sites, often appear as clusters. Peptide domains with high serine, threonine proline content have been identified in many proteins but no general consensus sequence has been identified. However, a series of general rules have been identified and an algorithm has been developed to predict O-glycosylation sites (Gupta *et al*., 1999; Thanka Christlet and Veluraja, 2001). The identification and characterization of the transferases that add the first GalNAc residues to the peptide (ppGaNTases) has shown that the selection of O-glycosylation sites is regulated at the enzyme level ((Ten Hagen *et al*., 2003) and see below) and explains, in part, the difficulty in predicting sites from primary sequence alone. A number of features can be identified that appear to govern the siting of *O*-glycans. First, a primary sequence preference can be identified, which varies for serine and threonine, reflecting the greater level of threonine substitution. Positioning of proline in the peptide adjacent to O-glycosylation sites has been

Table 1 Features of *O*-glycan structure. Structures are color-coded to show core (black), backbone, and peripheral regions

Nomenclature	Structure	Common name
Initial glycosylation with GalNAc		
GalNAc protein	GalNAcα-*O*-Ser/Thr	Tn antigen
Core classes		
Core 1	GalNAcα-*O*-Ser/Thr Galβ1-3	Thomsen Friedenreich (TF or T) antigen
Core 2	GlcNAcβ1-6 GalNAcα-*O*-Ser/Thr Galβ1-3	
Core 3	GalNAcα-*O*-Ser/Thr GlcNAcβ1-3	
Core 4	GlcNAcβ1-6 GalNAcα-*O*-Ser/Thr GlcNAcβ1-3	
Core 5	GalNAcα-*O*-Ser/Thr GalNAcα1-3	
Core 6	GalNAcα-*O*-Ser/Thr GlcNAcβ1-6	
Core 7	GalNAcα-*O*-Ser/Thr GalNAcα1-6	
Core 8	GalNAcα-*O*-Ser/Thr Galα1-3	
Backbone (elongation) units		
Type 1	Galβ1-3GlcNAc	
Type 2	Galβ1-4GlcNAc	*N*-acetyllactosamine
Poly *N*-acetyl lactosamine type 2	(Galβ1-4GlcNAcβ1-3-)$_n$	i antigen
Branched *N*-acetyl lactosamine type 2	Galβ1-4GlcNAcβ1-6 Galβ1 Galβ1-4GlcNAcβ1-3	I antigen
Peripheral nonreducing termini		
Blood group H type 1 (Galβ1-3GlcNAc)	Fucα1-2Galβ1-3GlcNAcβ1-	Blood group H antigen
Blood group A type 2 (Galβ1-4GlcNAc)	GalNAcα1-3Galβ1-4GlcNAcβ1- |α1-2 Fuc	Blood group A antigen
Blood group B type 1 (Galβ1-3GlcNAc)	Gal α1-3Galβ1-3GlcNAcβ1- |α1-2 Fuc	Blood group B antigen
Sialyl Lewis[a]	Neu5Acα2-3Galβ1-3GlcNAcβ1- |α1-4 Fuc	Sialyl Lewis[a] antigen
Sialyl Lewis[x]	Neu5Acα2-3Galβ1-4GlcNAcβ1- |α1-3 Fuc	Sialyl Lewis[x] antigen
Sulfo Lewis[a]	SO^{3-}3 Galβ1-3GlcNAcβ1- |α1-4 Fuc-	Sulfo Lewis[a] antigen
Sialyl-Tn	Neu5Acα2-6GalNAc-alpha-*O*-Ser/Thr	Sialyl-Tn antigen
Sialyl-T-antigen	Neu5Acα2-3Galβ1-3GalNAcα*O*-Ser/Thr Neu5Ac |α2-6 Galβ1-3GalNAcα*O*-Ser/Thr	Sialyl T or sialyl Core 1 antigen

Ser: serine, Thr: threonine, Neu5Ac: sialic acid, GalNAc: *N*-acetylgalactosamine, GlcNAc: *N*-acetylglucosamine, Fuc: Fucose.
α, β, and numbers indicate the glycosidic linkage between neighboring monosaccharides.

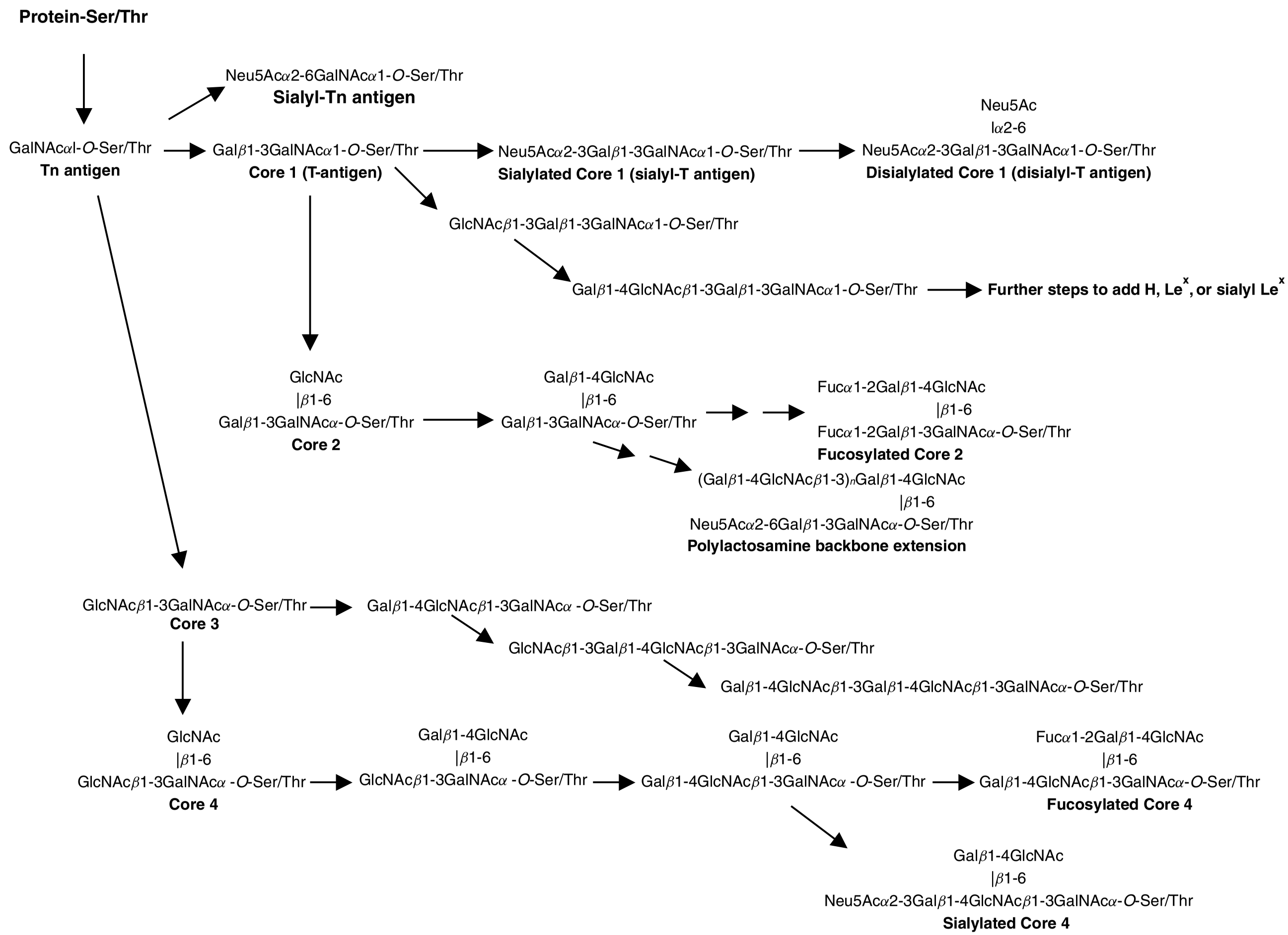


Figure 1 Pathways of Cores 1–4 and extended oligosaccharide biosynthesis. Pathways for the formation of Cores 1–4 and selected examples for their extension. Single arrows indicate the action of one glycosyltransferase, double arrows indicate the action of several enzymes. Ser: serine, Thr: threonine, Neu5Ac: sialic acid, GalNAc: *N*-acetylgalactosamine, GlcNAc: *N*-acetylglucosamine, Fuc: Fucose. α, β, and numbers indicate the glycosidic linkage between neighboring monosaccharides

identified in windows. The windows predict positions for Ser/Thr/Pro, but also exclude certain amino acids including Cys, Trp, Met, and Asp (Gupta *et al*., 1999; Thanka Christlet and Veluraja, 2001).

Second, only those serine and threonine residues that are exposed will be glycosylated. This implies a role for the conformation of the protein, and is confirmed with the identification of *O*-glycans on β-turns and extended structures where proline residues play a role. Furthermore, hydrophobic domains have no *O*-glycans, and bulky amino acids do not feature close to O-glycosylation sites.

Finally, the patterns of O-glycosylation are tissue specific, a function of the initial action of the ppGaNTases and the remaining glycosyltransferases. This has been well demonstrated for the secreted mucin genes where the tandem repeat ser/thr/pro domains contain the same amino acid sequence but where variable glycosylation is found on a tissue specific basis. MUC5AC is expressed in the stomach, respiratory tract, and the conjunctiva, each of which show different glycosylation patterns (Corfield *et al*., 2001). Accordingly, a primary amino acid sequence for *O*-glycan location will not be obvious for this protein.

2.3. Biosynthesis of O-linked glycans

Formation of the *O*-glycans follows a series of biosynthetic pathways that are located in regions of the endoplasmatic reticulum and Golgi apparatus (Van den Steen *et al*., 1998; Brockhausen *et al*., 2001). The primary sequence of oligosaccharides is therefore controlled by substrate specificity of the glycosyltransferases. Initiation of biosynthesis, by addition of GalNAc to serine and threonine residues in acceptor proteins, is mediated by a family of UDP-N-acetylgalactosamine: polypeptide *N*-acetylgalactosaminyltransferases (ppGaNTases) (Ten Hagen *et al*., 2003). These transferases have substrate specificities depending on the location of the serine or threonine residues in the peptide and whether they are already substituted by a GalNAc residue. The enzymes act in a hierarchical manner implying the need for coordinated enzyme action to achieve full O-glycosylation of potential sites. As these enzymes are also expressed in a tissue specific manner, a differential glycosylation pattern will be expected in an organ/tissue dependent manner. The substitution of GalNAc is thus a crucial guiding step in the positioning, level of occupancy, and oligosaccharide extension in defined proteins.

The biosynthesis of the core, backbone, and peripheral regions follows the action of the ppGaNTases to generate the vast library of *O*-glycan structures known. These pathways consist of families of glycosyltransferases, each with a substrate specificity, which defines the characteristic *O*-glycan structures. These enzymes have been reviewed in detail (e.g., (Schachter and Brockhausen, 1992; Brockhausen *et al*., 2001)). To illustrate the action of these enzymes, a pathways map for some of the common core classes and extended oligosaccharide structures is show in Figure 1.

In addition, further modification of some sugars occurs generating new epitopes. The most important of these events are sulfation of Gal and GalNAc (Brockhausen, 2003) and O-acetylation of sialic acids (Corfield *et al*., 1999).

3. The functions of O-linked glycosylation

The functions of *O*-glycans are expected to be wide, ranging in view of the variety of O-glycosylated proteins described. However, only in recent years, clear structure–function relationships have been demonstrated. This section draws attention to the major areas where these interactions have been demonstrated.

3.1. Protein structure and stability

The substitution of an *O*-glycan onto a peptide or protein backbone will have significant influence on the overall structural arrangement of the completely folded molecule. A number of studies have investigated this phenomenon and have shown that many of the properties exhibited are not exclusive but are often shared within one molecule.

3.1.1. Protein conformation, tertiary structure

The mucins and mucin-like molecules with serine threonine proline-rich domains have been used as models to identify changes in protein tertiary structure resulting from O-glycosylation. The addition of the initial GalNAc residue and subsequent elongation of the oligosaccharides results in an extended conformation often referred to as the "bottle brush" structure. A series of investigations on the characteristic tandem repeat sequences found in the MUC genes, in particular MUC1, have demonstrated the modifications to peptide conformation as a result of glycosylation (Otvos and Cudic, 2003). These studies have been correlated with the biosynthetic introduction of the initial GalNAc residues and the specificity of the family of ppGaNAc transferases (Kirnarsky *et al*., 2003). The elegant work of Gerken with porcine submaxillary mucin has shown that the hydroxyamino acid spacing may contribute to the regulation of glycan length, thereby, providing a mechanism for maintaining an optimally expanded mucin conformation (Gerken *et al*., 2002).

Further examples of proteins with "mucin-like" or O-glycosylated "stalk" domains show conformational arrangements that have direct importance in molecular organization, orienting functional/binding domains away from the cell surface. A number of examples exist, including nerve growth factor receptor, CD2, LFA-3, CD8$\alpha\beta$ coreceptor, and calcitonin.

3.1.2. Protein quaternary structure and molecular association

The extended conformations resulting from O-glycosylation offer increased opportunities for molecular interactions at a quaternary level leading to molecular networks and gel structures (McMaster *et al*., 1999). These have been demonstrated typically for the epithelial secreted mucins that have physiological function in adherent gels at mucosal surfaces. In contrast, the prevention of molecular association or aggregation and, therefore, regulation of molecular function can be achieved

through O-glycosylation. This has been demonstrated for lactase phlorizin, secretory IgA, and CD45 dimerization (Van den Steen *et al*., 1998; Xu and Weiss, 2002).

3.1.3. Protein stability protease and heat resistance

The stability of proteins through O-glycosylation is also reflected in protective roles against proteolytic degradation and thermal disruption. These protective properties are lost or absent if the oligosaccharides are removed or are not present. Decay accelerating factor (CD55) is protected in part against proteolysis by O-glycosylation and this regulates the half-life for its biological activity (Van den Steen *et al*., 1998).

The highly O-glycosylated core of mucus glycoproteins is resistant to proteolytic degradation, and this property has been used to prepare mucin glycopeptides. However, protection of protease sensitive domains may also be achieved by site specific O-glycosylation. Thermal stability of protein has been shown in highly O-glycosylated examples such as the mucins and in glucoamylase. Improved thermostability could be introduced into glucoamlyase using mutants designed to increase O-glycosylation in the already highly O-glycosylated belt region (Liu *et al*., 2000).

3.2. Recognition phenomena

O-glycans play roles in a variety of processes that come under the general heading of recognition. These interactions are often complex and may also involve *N*-glycans (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6) present in the same molecules. Such functions continue to accumulate and are briefly reviewed here.

3.2.1. Cell proliferation and growth

Processes associated with cell growth have been found to be closely linked with O-glycosylation. Three examples illustrate the kind of involvement found. (1) The development of muscle tissues is regulated differentially through the polysialic acid containing and other O-linked glycans found in a muscle-specific domain on neural cell adhesion molecule (NCAM). These domains mediate myoblast fusion and the normal development of muscle (Suzuki *et al*., 2003). (2) The hormone-mediated proliferation of breast cancer cells has been shown to be dependent on the O-glycosylation of human sex hormone-binding globulin (SHBG) for binding to a membrane located receptor. The elimination of *O*-glycans on SHBG prevents the inhibition of estradiol-induced proliferation by this hormone (Raineri *et al*., 2002). (3) Induction of apoptosis by agents that cause endoplasmic reticulum stress lead to O-glycosylation of both β-catenin and the E-cadherin cytoplasmic domain. The cadherin/catenin adhesion complexes are regulated by recycling from the plasma membrane and by proteolysis during apoptosis. O-glycosylation of

newly synthesized E-cadherin blocks cell surface transport, resulting in reduced intercellular adhesion (Zhu *et al*., 2001).

3.2.2. Glycoprotein clearance

The clearance of glycoproteins from the circulation is mediated through several different receptors expressed in hepatic, renal, endothelial, and lymphatic tissues. The recognition and binding of the glycoproteins depends on their glycosylation patterns. The hepatic asialo-glycoprotein receptor binds asialo-*N*-glycans in serum glycoproteins but also *O*-glycans with terminal GalNAc. Related receptors have been identified for the elimination of the glycoprotein hormones and insulin-like growth factor binding proteins (Marinaro *et al*., 2000).

3.2.3. Glycoprotein trafficking

The targeting of proteins within the cell has assumed great importance for many aspects of glycosylation and a number of examples have been described supporting a role for *O*-glycans in trafficking pathways (Huet *et al*., 2003). In polarized cells, specific sorting mechanisms are found in the trans-Golgi network. Apical targeting occurs via apical lipid raft carriers and the targeted glycoproteins have glycans mediating their association with rafts and ensuring transport to the correct location (Huet *et al*., 2003). The details of this system remain controversial.

Cellular targeting to the plasma membrane, regulated by *O*-glycan expression has been shown for an increasing number of molecules (Huet *et al*., 2003) including ceramidase, sucrase/isomaltase, CD44, aminopeptidase N, dipeptidyl peptidase IV, human neurotrophin receptor, and C1qRP/CD93. Exit from the endoplasmic reticulum is also determined in this way for cytochrome b5 and human delta opioid receptor G-protein.

3.2.4. Immunological recognition

The immune process relies on the recognition of foreign proteins to initiate a response that leads to protection. The detection of these non–self structures depends on antibodies and T-cells. Glycosylation has a major influence on both of these immunological systems and is also represented in the form of well-known antigenic structures such as the ABH blood groups (Van den Steen *et al*., 1998; Rudd *et al*., 2001).

The human ABH blood groups (see Table 1) are glycan structures carried on both O-and N-linked oligosaccharides in many glycoproteins. Individuals produce antibodies to the non–self blood groups while remaining tolerant to their own. *O*-glycans also contribute to the epitopes in glycophorin A, responsible for the MN and Miltenberger blood groups.

The antibodies IgA1 and IgD have serine/threonine/proline-rich protein domains in the hinge region between the Fab and Fc regions (Novak *et al*., 2000). These peptide sequences are O-glycosylated with short oligosaccharides based on the

Core 1 Galβ1-3GalNAc disaccharide and are largely sialylated. The function of these domains is related to the conformation and stability of the proteins and thus, represents a specific example of the functions already noted above. In the case of IgA1, resistance to proteolytic attack is an important biological requirement for the secretory form of the antibody, which is the major antibody secreted at mucosal surfaces. The conformation of the glycosylated hinge region confers protection to common proteases (Novak *et al*., 2000). Both IgA1 and IgD have been shown to bind to T-lymphocytes through a cell surface lectin that recognizes the *O*-glycans (Van den Steen *et al*., 1998).

Selection of a peptide by antigen presenting cells takes place through the MHC-II-invariant chain Ii system, leading to the appearance of the MHC-II complex at the cell surface. Subsequently, the peptide is recognized by the T-cell receptor on $CD4^+$ T-lymphocytes. Ii contains *O*-glycans that play a role in the stabilization of the intracellular MHC-II-Ii complex (Rudd *et al*., 2001). As noted earlier, thymocyte CD8$\alpha\beta$ coreceptor stalk O-glycosylation affects MHC-I ligand binding (Moody *et al*., 2001). Core 1 sialic acid in addition to CD8β on mature thymocytes decreases CD8$\alpha\beta$-MHCI avidity by altering CD8$\alpha\beta$ domain–domain association and/or orientation.

The binding of peptides selected for presentation to the MHC complex is affected considerably by glycosylation. Glycosylated peptides are processed and presented, but the presence of glycans can also prevent binding or give rise to molecular mimicry. This has been demonstrated in a number of cases in which the binding to the MHC complex groove occurs with O-glycosylated peptides, which mimic certain amino acid residues in nonglycosylated structures (Van den Steen *et al*., 1998; Rudd *et al*., 2001). Peptides substituted with *O*-GlcNAc (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6) have been found to play a role both in the binding to the MHC complex and in the recognition of MHC complex bound epitopes by the T-cell receptor.

A further role for *O*-glycans in immune cells has been found with the demonstration of an *O*-glycan marker on the surface of effector CD8 T-cells. This *O*-glycan was upregulated on virus-specific CD8 T-cells during the effector phase of the primary cytotoxic T lymphocyte response, showed a strong correlation with the acquisition of effector function, and was downregulated on memory CD8 T-cells (Harrington *et al*., 2000).

The mechanisms responsible for autoimmunity also include molecular mimicry, and examples have been described for O-glycosylated glycoproteins (Van den Steen *et al*., 1998). This applies to the truncated Core 1 disaccharide-related structures arising from glycosidase action in infectious diseases such as pneumonia, glycosylation of hydroxy lysine residues in collagen type II leading to arthritis, and antibodies against GlcNAc in Streptococcus group A specific carbohydrate, that are able to cross-react with a cytokeratin peptide.

3.2.5. Selectin ligands

Considerable interest has been focused on the O-glycosylation of selectin ligands as this area has provided excellent examples of structure–function relationships related to the regulation of leukocyte circulation. The different selectins (P, L, and

E) recognize various protein ligands with serine–threonine-rich domains carrying *O*-glycans (Varki, 1997).

PSGL-1 is the major P-selectin ligand. High-affinity interactions with P-selectin require *O*-glycans with a branched Core-2 structure containing the sialyl Lewis^x antigen (see Table 1 for these structures) or other fucosylated glycans. In addition, the coexpression of sulfated tyrosine residues near the critical *O*-glycan structures is necessary (Leppanen *et al*., 2002).

L-selectin mediates tethering and rolling of lymphocytes on high endothelial venules in secondary lymphoid organs binding with specific ligands. The major ligands identified are CD34, GlyCAM, MadCAM-1, podocalyxin, and human endomucin. Recognition by L-selectin depends on peripheral sulfo, sialyl Lewis^x units with Gal-6- or GlcNAc-6-sulfate carried on *O*-glycans. The sulfotransferases generating these sulfated *O*-glycans on L-selectin ligands have been identified and support the structural studies for L-selectin binding requirements (Bruehl *et al*., 2000).

E-selectin, expressed on endothelial cells, mediates adhesion of leukocytes and tumor cells to endothelium. E-selectin ligands have proved more difficult to identify but include PSGL-1, MUC1, and several cancer derived mucin molecules. The ligands also show a requirement for sialyl Lewis^x or sialyl Lewis^a units carried on *O*-glycans, but do not appear to need sulfate for recognition.

There is currently great interest in the regulation of selectin binding through the expression of soluble sialyl Lewis^x substituted glycoproteins during inflammation and disease-related events. In addition, a role for selectin ligands in metastasis has been described focused on cellular mucins carrying sialyl Lewis structures (Varki and Varki, 2001).

3.2.6. Multivalent and multimeric binding and signal transduction

The binding of ligands to receptors can be substantially amplified through interactions with molecules having a repeat structure. Such multivalent ligands have been described for many receptors. The nature of *O*-glycan substitution in serine–threonine-rich protein domains lends itself to the formation of multivalent glycosylated ligands. Increased strength of binding can also be achieved through receptors with multiple units. The combination of mutivalent ligand and multimeric receptor binding can greatly enhance the sensitivity of such systems. The selectins are good examples of this phenomenon. The binding of CD22 to PSGL-1 through sialyl Lewis^x is enhanced through multiple substitution on the ligand and generates the stronger interactions necessary to enable rolling and arrest of circulating leukocytes in the circulation (Van den Steen *et al*., 1998).

3.2.7. Fertilization

Glycosylation is a vital part of recognition processes involved in fertilization. A role for *O*-glycans has been demonstrated in the structure of the oocyte zona pellucida or mucus barrier where differences in structure are species specific. Considerable evidence supports a role for glycans in the fertilization of the oocyte by spermatozoa. This binding is thought to require the interaction of *O*-glycans

linked to a specific protein, zona pellicuda protein 3 (ZP3) on the oocyte, with egg-binding proteins coating the sperm plasma membrane. Sperm cell binding is dependent on a specific ZP3 domain carrying *O*-glycans with nonreducing terminal α-Gal or β-GlcNAc residues (Dell *et al*., 2003). Spermadhesins have been identified as the binding partners in these interactions and may themselves be regulated through glycosylation, which blocks the normal binding. Similar functional molecules have been identified in different species and which also depend on O-glycosylation for their activity. Binding of the sperm to ZP3 triggers release of glycosidase activity, which destroys the *O*-glycan receptors on other sperm cells and prevents polyspermia.

3.3. Activity of signaling molecules

The direct action of glycosylation on the activity of signaling molecules is largely limited to examples where *N*-glycosylation (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6) or *O*-GlcNAc (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6) glycans have a role. A few examples have been described for cytokine and hormone dependency on *O*-GalNAc-based glycans but these interactions remain unclear and are linked with the sophisticated biosynthesis of the initial glycosyltransfer of GalNAc as noted above. The biological activity of interleukin-5 is decreased if IL-5 is O-glycosylated (Kodoma *et al*., 1993), while the presence of an *O*-glycan in granulocyte colony-stimulating factor increases its activity (Nissen *et al*., 1994). Both of these molecules also share stabilizing properties donated by the addition of the *O*-glycans.

The activity of thyroid stimulating hormone and the thyrotropic action of human chorionic gonadotrophin are reduced by desialylation or loss of their *O*-glycans. These few examples indicate variable roles for O-glycosylation and underline the need to examine the complete range of potential functions of *O*-glycans in any individual glycoprotein.

References

Brockhausen I (2003) Sulphotransferases acting on mucin-type oligosaccharides. *Biochemical Society Transactions*, **31**, 318–325.

Brockhausen I, Yang J, Lehotay M, Ogata S and Itzkowitz S (2001) Pathways of mucin O-glycosylation in normal and malignant rat colonic epithelial cells reveal a mechanism for cancer-associated Sialyl-Tn antigen expression. *Biological Chemistry*, **382**, 219–232.

Bruehl RE, Bertozzi CR and Rosen SD (2000) Minimal sulfated carbohydrates for recognition by L-selectin and the MECA-79 antibody. *Journal of Biological Chemistry*, **275**, 32642–32648.

Corfield AP, Carroll D, Myerscough N and Probert CS (2001) Mucins in the gastrointestinal tract in health and disease. *Frontiers in Bioscience*, **6**, D1321–D1357.

Corfield AP, Myerscough N, Warren BF, Durdey P, Paraskeva C and Schauer R (1999) Reduction of sialic acid O-acetylation in human colonic mucins in the adenoma-carcinoma sequence. *Glycoconjugate Journal*, **16**, 307–317.

Dell A, Chalabi S, Easton RL, Haslam SM, Sutton-Smith M, Patankar MS, Lattanzio F, Panico M, Morris HR and Clark GF (2003) Murine and human zona pellucida 3 derived from mouse eggs express identical O-glycans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15631–15636.

Gerken TA, Zhang J, Levine J and Elhammer A (2002) Mucin core O-glycosylation is modulated by neighboring residue glycosylation status. Kinetic modeling of the site-specific glycosylation of the apo-porcine submaxillary mucin tandem repeat by UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases T1 and T2. *Journal of Biological Chemistry*, **277**, 49850–49862.

Gupta R, Birch H, Rapacki K, Brunak S and Hansen JE (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Research*, **27**, 370–372.

Harrington LE, Galvan M, Baum LG, Altman JD and Ahmed R (2000) Differentiating between memory and effector CD8 T cells by altered expression of cell surface O-glycans. *Journal of Experimental Medicine*, **191**, 1241–1246.

Huet G, Gouyer V, Delacour D, Richet C, Zanetta JP, Delannoy P and Degand P (2003) Involvement of glycosylation in the intracellular trafficking of glycoproteins in polarized epithelial cells. *Biochimie*, **85**, 323–330.

Kirnarsky L, Suryanarayanan G, Prakash O, Paulsen H, Clausen H, Hanisch FG, Hollingsworth MA and Sherman S (2003) Conformational studies on the MUC1 tandem repeat glycopeptides: implication for the enzymatic O-glycosylation of the mucin protein core. *Glycobiology*, **13**, 929–939.

Kodoma S, Tsujimoto M, Tsuruoka N, Sugo T, Endo T and Kobata A (1993) Role of sugar chains in the in vitro activity of recombinant human interleukin 5. *European Journal of Biochemistry*, **211**, 903–908.

Leppanen A, Penttila L, Renkonen O, McEver RP and Cummings RD (2002) Glycosulfopeptides with O-Glycans Containing Sialylated and Polyfucosylated Polylactosamine Bind with Low Affinity to P-selectin. *Journal of Biological Chemistry*, **277**, 39749–39759.

Liu HL, Doleyres Y, Coutinho PM, Ford C and Reilly PJ (2000) Replacement and deletion mutations in the catalytic domain and belt region of Aspergillus awamori glucoamylase to enhance thermostability. *Protein Engineering*, **13**, 655–659.

Marinaro JA, Neumann GM, Russo VC, Leeding KS and Bach LA (2000) O-glycosylation of insulin-like growth factor (IGF) binding protein-6 maintains high IGF-II binding affinity by decreasing binding to glycosaminoglycans and susceptibility to proteolysis. *European Journal of Biochemistry*, **267**, 5378–5386.

McMaster TJ, Berry M, Corfield AP and Miles MJ (1999) Atomic force microscopy of the submolecular architecture of hydrated ocular mucins. *Biophysical Journal*, **77**, 533–541.

Moody AM, Chui D, Reche PA, Priatel JJ, Marth JD and Reinherz EL (2001) Developmentally regulated glycosylation of the CD8alphabeta coreceptor stalk modulates ligand binding. *Cell*, **107**, 501–512.

Nissen C, Dalle Carbonare V and Moser Y (1994) In vitro comparison of the biological potency of glycosylated versus non-glycosylated rG-CSF. *Drug Investigation*, **7**, 346–352.

Novak J, Tomana M, Kilian M, Coward L, Kulhavy R, Barnes S and Mestecky J (2000) Heterogeneity of O-glycosylation in the hinge region of human IgA1. *Molecular Immunology*, **37**, 1047–1056.

Otvos L Jr and Cudic M (2003) Conformation of glycopeptides. *Mini Reviews in Medicinal Chemistry*, **3**, 703–711.

Raineri M, Catalano MG, Hammond GL, Avvakumov GV, Frairia R and Fortunati N (2002) O-Glycosylation of human sex hormone-binding globulin is essential for inhibition of estradiol-induced MCF-7 breast cancer cell proliferation. *Molecular and Cellular Endocrinology*, **189**, 135–143.

Rudd PM, Elliott T, Cresswell P, Wilson IA and Dwek RA (2001) Glycosylation and the immune system. *Science*, **291**, 2370–2376.

Schachter H and Brockhausen I (1992) The biosynthesis of serine (threonine)-N-acetylgalactosamine-linked carbohydrate moieties. In *Glycoconjugates*, Allen HJ and Kisailus EC (Eds.), Marcel Dekker: New York, Basel, Hong Kong, pp. 263–332.

Suzuki M, Angata K, Nakayama J and Fukuda M (2003) Polysialic Acid and mucin type *O*-glycans on the neural cell adhesion molecule differentially regulate myoblast fusion. *Journal of Biological Chemistry*, **278**, 49459–49468.

Ten Hagen KG, Fritz TA and Tabak LA (2003) All in the family: the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases. *Glycobiology*, **13**, 1R–16R.

Thanka Christlet TH and Veluraja K (2001) Database analysis of *O*-glycosylation sites in proteins. *Biophysical Journal*, **80**, 952–960.

Van den Steen P, Rudd PM, Dwek RA and Opdenakker G (1998) Concepts and principles of *O*-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology*, **33**, 151–208.

Varki A (1997) Sialic acids as ligands in recognition phenomena (review). *FASEB Journal*, **11**, 248–255.

Varki A, Cummings RD, Esko J, Freeze H, Hart G and Marth JD (1999) *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.

Varki A and Varki NM (2001) P-selectin, carcinoma metastasis and heparin: novel mechanistic connections with therapeutic implications. *Brazilian Journal of Medical and Biological Research*, **34**, 711–717.

Xu Z and Weiss A (2002) Negative regulation of CD45 by differential homodimerization of the alternatively spliced isoforms. *Nature Immunology*, **3**, 764–771.

Zhu W, Leber B and Andrews DW (2001) Cytoplasmic *O*-glycosylation prevents cell surface transport of E-cadherin during apoptosis. *EMBO Journal*, **20**, 5999–6007.

Specialist Review

GPI anchors

Nigel M. Hooper
University of Leeds, Leeds, UK

1. Introduction

Although the majority of integral membrane proteins are anchored by transmembrane hydrophobic polypeptides, numerous proteins are now known to be anchored in the membrane by a covalently attached glycosyl-phosphatidylinositol (GPI) moiety. This complex glycan structure is attached to the C-terminal amino acid of the mature protein and serves to stably anchor the protein in the outer leaflet of the lipid bilayer (Figure 1). As such, a GPI anchor can be considered as an alternative to a C-terminal hydrophobic transmembrane polypeptide domain, as in type 1 integral membrane proteins. GPI-anchored proteins are present in organisms at most stages of eukaryotic evolution, including protozoa, yeast, slime molds, plants, invertebrates, and vertebrates (McConville and Ferguson, 1993). A diverse range of GPI-anchored proteins have been described, including protozoal surface coat proteins, receptors, adhesion molecules, ectoenzymes, and differentiation antigens. Possibly, the most infamous GPI-anchored protein is the prion protein, the causative agent of the transmissible spongiform encephalopathies such as Creutzfeldt–Jakob disease in humans and mad cow disease (bovine spongiform encephalopathy).

In addition to being found attached to proteins, nonprotein-linked or free GPI structures are also present in cells and appear to be functionally important components in their own right (McConville and Menon, 2000). Particularly in certain protozoan parasites, free GPIs are the major class of glycolipids and are essential for parasite growth and infectivity. For example, GPIs from *Plasmodium falciparum* have the properties of the malarial toxin, and synthetic GPI has been reported as a candidate antitoxic vaccine in a model of malaria (Schofield *et al*., 2002).

2. GPI anchor structure

Structural determination of the GPI anchors on proteins from protozoa, yeast, plants, and mammals has revealed the common core structure: ethanolamine-PO_4-6Manα1-2Manα1-6Manα1-4GlcN$H_2\alpha$1-6*myo*-inositol-1-PO_4-lipid (Figure 1). The ethanolamine is amide-bonded to the α-carboxyl group of the C-terminal amino acid of the mature protein. The conserved core structure may possess a variety

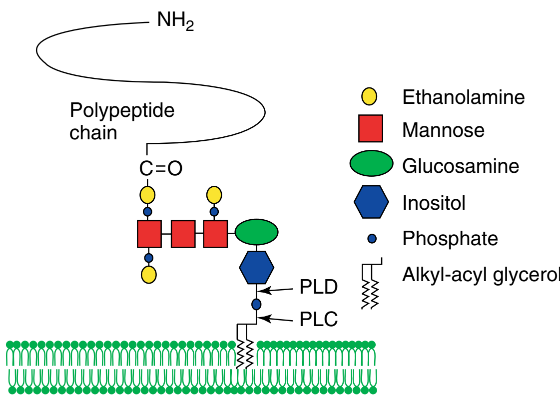


Figure 1 Schematic representation of a GPI anchor. The polypeptide chain is covalently linked through its C-terminus to the core GPI structure ethanolamine-phosphate-(mannose)$_3$-glucosamine-phosphatidylinositol. The lipid, alkyl-acyl-glycerol in the example shown, interacts with the outer leaflet of the lipid bilayer. Additional ethanolamine phosphate groups are shown on the first and third mannose residues. The sites of cleavage by phospholipase D (PLD) and phospholipase C (PLC) are shown on either side of the phosphate within the phosphatidylinositol of the anchor by phospholipase D (PLD) and phospholipase C (PLC) are shown. (Reprinted from *Immunology Today*, **20**, Horejsi V, Drbal K, Cebecauer M, Cerny J, Brdicka T, Angelisova P and Stockinger H, GPI-microdomains: a role in signalling via immunoreceptors, 356–361, (1999), with permission from Elsevier)

of side-chain modifications (additional phosphoethanolamines and sugars e.g., *N*-acetyl-galactosamine, galactose, mannose) that appear to be protein, tissue, or species specific (Brewis *et al*., 1995). The lipid moieties range from ceramide in most yeast and slime mold GPI-anchored proteins to diacylglycerol in protozoa and (predominantly) 1-alkyl-2-acylglycerol in mammalian proteins (*see* Article 77, **Glycosylphosphatidylinositol anchors – a structural perspective**, Volume 6). In some cases, the inositol ring contains an additional lipid modification in the form of an ester-linked palmitic acid (Ferguson, 1992) that renders the anchor resistant to the action of phospholipase C (PLC) (*see* Section 4.1).

3. GPI anchor synthesis and addition

The preformed GPI anchor core is built up by sequential addition of the individual sugars and phosphoethanolamine to a phosphatidylinositol (PI) molecule on the cytoplasmic side of the rough endoplasmic reticulum (ER) (Ferguson, 1999; McConville and Menon, 2000). The process is initiated by the transfer of *N*-acetylglucosamine (GlcNAc) from UDP-GlcNAc to PI, followed by de-N-acetylation of the resulting GlcNAc-PI to give Glc-PI. The Glc-PI is then mannosylated with three mannose residues, each being derived from dolichol-phosphomannose, which in turn is synthesized from GDP-mannose and dolichol phosphate. In mammalian cells and yeast, the Glc-PI must first be acylated on the inositol residue before mannose addition can occur. In other organisms, inositol acylation may occur at a later stage of the biosynthetic pathway. The mannosylated GPIs are then modified with phosphoethanolamine, which is derived

from phosphatidylethanolamine. Most of the enzymes involved in each of these steps have been identified and, to varying extents, characterized (see McConville and Menon, 2000). The completed precursor is then flipped across the membrane into the lumen of the rough ER prior to addition by the transamidase enzyme to a protein with a C-terminal GPI anchor addition signal (see below).

To date, all GPI-anchored proteins are present either on the outer leaflet of the plasma membrane or on an intracellular membrane with the same disposition (i.e., in the lumen of the ER or secretory vesicles). Thus, all GPI-anchored proteins possess two signal sequences in their nascent polypeptide chain (Figure 2a). The first is a hydrophobic cleavable N-terminal signal sequence that is recognized by the signal recognition particle and that cotranslationally directs the protein through the Sec61 translocon complex into the lumen of the rough ER (*see* Article 37, **Signal peptides and protein localization prediction**, Volume 7). The second is another hydrophobic peptide of some 15 to 25 amino acid residues that lies at the very C-terminus of the nascent chain. A hydropathy plot of the nascent (cDNA predicted) amino acid sequence of GPI-anchored proteins clearly reveals these two strongly hydrophobic regions at either end of the polypeptide (Figure 2b). The C-terminal hydrophobic sequence is preceded by a consensus sequence (ω, $\omega + 1$, and $\omega + 2$) for GPI anchor addition (Figure 2a). Cleavage of the polypeptide chain occurs on the C-terminal side of the ω residue, with concomitant addition of the preassembled GPI anchor (see above) to the newly exposed α-carboxyl group of the ω residue by the transamidase enzyme (Figure 3). This transamidase is a

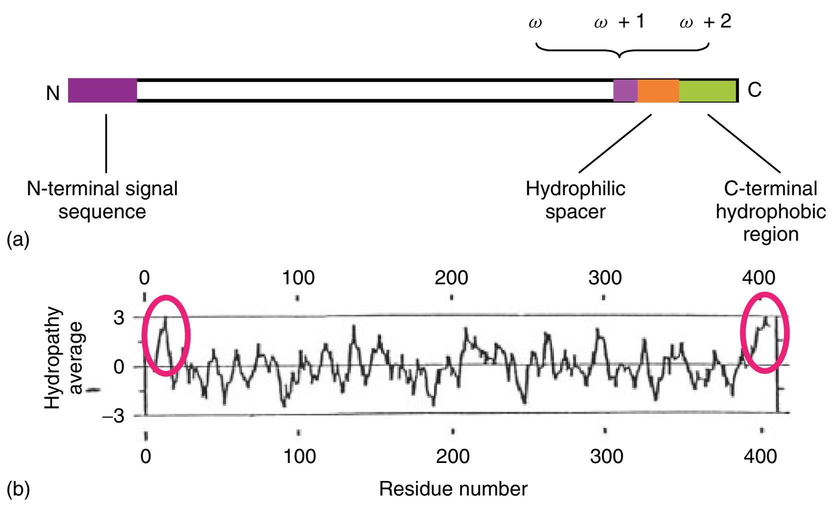


Figure 2 Schematic representation of a typical nascent GPI-anchored protein; (a) the cDNA-derived amino acid sequence of a typical GPI-anchored protein contains an N-terminal signal peptide and a C-terminal signal for directing addition of the GPI anchor. This C-terminal signal consists of a hydrophobic region, a hydrophilic spacer, and three residues ω, $\omega + 1$, and $\omega + 2$. The ω residue is the site of cleavage of the polypeptide chain and of addition of the preformed GPI anchor by the transamidase enzyme (see text for details); (b) hydropathy plot of the nascent polypeptide chain of the GPI-anchored porcine membrane dipeptidase (Rached *et al*., 1990). The strongly hydrophobic regions at the extreme N-terminus and C-terminus are circled

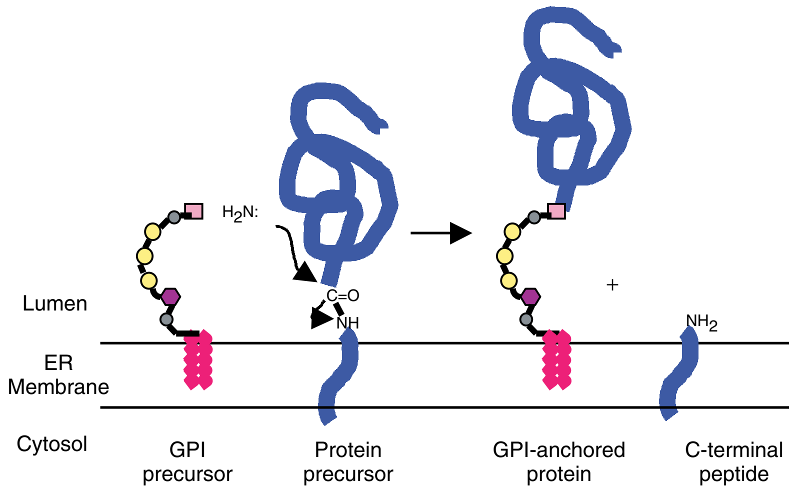


Figure 3 Addition of a GPI anchor to a protein. The C-terminal hydrophobic peptide transiently tethers the nascent polypeptide chain in the membrane of the rough ER following its translation and translocation into the lumen. The GPI anchor is built up by sequential addition of sugars and phosphoethanolamine to phosphatidylinositol (see text for details). The transamidase complex then cleaves the polypeptide chain between the ω and $\omega + 1$ residues (see Figure 2a) and concomitantly adds on the preformed GPI anchor. The released C-terminal peptide is degraded

complex of several proteins, including Gpi8p, which contains the active site of the enzyme, Gaa1p, Gpi16p (PIG-T), and Gpi17p (PIG-S) (Fraering *et al*., 2001). Once added to the protein, addition of other sugars to the core GPI structure may occur as the protein is trafficked through the ER and Golgi.

Analysis of native GPI-anchored protein sequences and extensive site-directed mutagenesis of these residues in the GPI-anchored proteins alkaline phosphatase and decay-accelerating factor (Udenfriend and Kodukula, 1995) has shown that the ω residue is restricted to amino acids with small side chains (Ala, Asn, Asp, Cys, Gly, or Ser), whereas $\omega + 1$ can be any residue except for Pro and Trp and $\omega + 2$ is usually Gly or Ala, or occasionally Ser or Thr. The potential for a protein to be GPI anchored, as well as the most probable site of GPI anchor attachment in the protein, can now be predicted in a manner analogous to the prediction of the site of signal peptidase cleavage at the N-terminus of a protein, using web-based algorithms such as the "big-Π predictor" (available at **http://mendel.imp.univie.ac.at/gpi/gpi_prediction.html** or via **http://ca.expasy.org/tools/**).

4. Identification of a GPI anchor

Although the potential for a protein to be GPI anchored can be predicted from analysis of its cDNA-derived amino acid sequence (see above), the presence of such an anchor structure can only be verified experimentally. The identification of a GPI anchor on a protein does not have to rely upon direct structural determination of the anchor using biophysical techniques such as mass spectroscopy, which require relatively large amounts of highly purified protein and access to such equipment.

Techniques exist whereby the presence of such a structure can be determined even if the protein is present in low abundance or is not readily purified (see Hooper and McIlhinney, 1999; Hooper, 2001).

4.1. Release by bacterial phosphatidylinositol-specific phospholipase C

The susceptibility of a protein to be released from the membrane by bacterial phosphatidylinositol-specific phospholipase C (PI-PLC) is probably the simplest and most commonly used criterion to demonstrate the presence of a GPI anchor. PI-PLCs from *Bacillus thuringiensis* and *B. cereus* are available commercially. Membranes or cells containing the protein of interest are incubated with bacterial PI-PLC and the loss of the protein from the membrane and/or its appearance in the supernatant is monitored following centrifugation (Brewis *et al*., 1994). Alternatively, following incubation with PI-PLC, the membrane sample can be subjected to temperature-induced phase separation in Triton X-114 to separate the intact amphipathic protein from the cleaved hydrophilic form that lacks the fatty acids (Bordier, 1981; Hooper and Bashir, 1991). The lack of release of a protein by PI-PLC does not necessarily rule out the presence of a GPI anchor. Several GPI-anchored proteins are resistant to the action of PI-PLC due to additional acylation of the inositol ring (Ferguson, 1992). However, removal of this acyl group by alkaline hydroxylamine treatment then renders the protein susceptible to cleavage by PI-PLC (Hooper, 2001).

4.2. Detection of the cross-reacting determinant

Polyclonal antisera raised in rabbits against the PLC-cleaved soluble form of one GPI-anchored protein often cross-react with other unrelated GPI-anchored proteins. The site of cross-reactivity is the cross-reacting determinant (CRD), which is cryptic in the membrane-bound amphipathic form of GPI-anchored proteins and is only exposed after cleavage of the GPI anchor with PLC. The major epitope involved in this recognition is the inositol 1,2-cyclic monophosphate that is generated on PLC cleavage of the GPI anchor (Zamze *et al*., 1988; Hooper *et al*., 1991). Thus, the recognition of a protein by an antiCRD antiserum following PLC cleavage is virtually unequivocal evidence for the presence of a GPI anchor.

4.3. Metabolic labeling

If the protein of interest can be expressed in a suitable cell line, then metabolic labeling with components of the GPI anchor can be used to detect the presence of a GPI anchor. The most widely used substrate for labeling GPI anchors is [^{3}H] or [^{14}C] ethanolamine, although other components such as inositol or fatty acids can be used (Hooper and McIlhinney, 1999; Hooper, 2001). Following an appropriate labeling period, the cells are harvested and the protein of interest

immunoprecipitated following either PI-PLC release or detergent solubilization of the membranes. The presence of the radiolabel in the protein can then be assessed by SDS-PAGE followed by fluorography. When using a radiolabelled fatty acid, treatment with PI-PLC or nitrous acid (Hooper, 2001) should be performed to confirm that the fatty acid is in a GPI anchor and not some other lipid modification.

4.4. Detergent insolubility

GPI-anchored proteins, particularly those in mammalian cells, are relatively resistant to solubilization by certain detergents, such as Triton X-100 at 4°C (Hooper and Turner, 1988). In contrast, they are readily solubilized by detergents such as *n*-octyl-β-D-glucopyranoside. This detergent insolubility is an intrinsic property of the GPI anchor due to its clustering in glycosphingolipid- and cholesterol-rich membrane microdomains, so-called lipid rafts (see Section 5.2) (Brown and Rose, 1992). Indeed, the most commonly used method for the biochemical isolation of lipid rafts exploits this insolubility in detergents such as Triton X-100 (Hooper, 2001). Although a few transmembrane polypeptide-anchored proteins are present in lipid rafts, the majority are excluded from such domains and are effectively solubilized from the membrane with Triton X-100 at 4°C. Thus, the relative insolubility in certain detergents can be used to predict the presence of a GPI anchor on a protein.

5. Functions and properties of GPI anchors

The basic function of a GPI anchor is to stably attach the protein to the membrane; however, such a structure can convey additional properties on the protein that may modulate its function.

5.1. Cleavage by phospholipases

In addition to the bacterial PI-PLCs that are routinely used to experimentally demonstrate the presence of a GPI anchor on a protein (see Section 4.1), endogenous phospholipases exist that can cleave the GPI anchor in situ. Such cleavage may be a mechanism for generating a soluble form of the protein whose properties are either the same as or subtly different from that of the membrane-bound form and/or be a means of downregulating the surface expression of the protein. A GPI-specific phospholipase D (PLD) is present in blood plasma that can cleave GPI anchors in vitro, although its role in vivo remains unclear (Low and Huang, 1991). Although several mammalian GPI-anchored proteins have been shown to exist in a soluble form and possess the CRD indicative of cleavage by a PLC (Movahedi and Hooper, 1997; Park *et al*., 2002), to date, this enzyme has not been identified or isolated.

5.2. Association with lipid rafts

Within the plasma membrane of mammalian cells are domains enriched in sphingomyelin, glycosphingolipids, and cholesterol that have been termed lipid

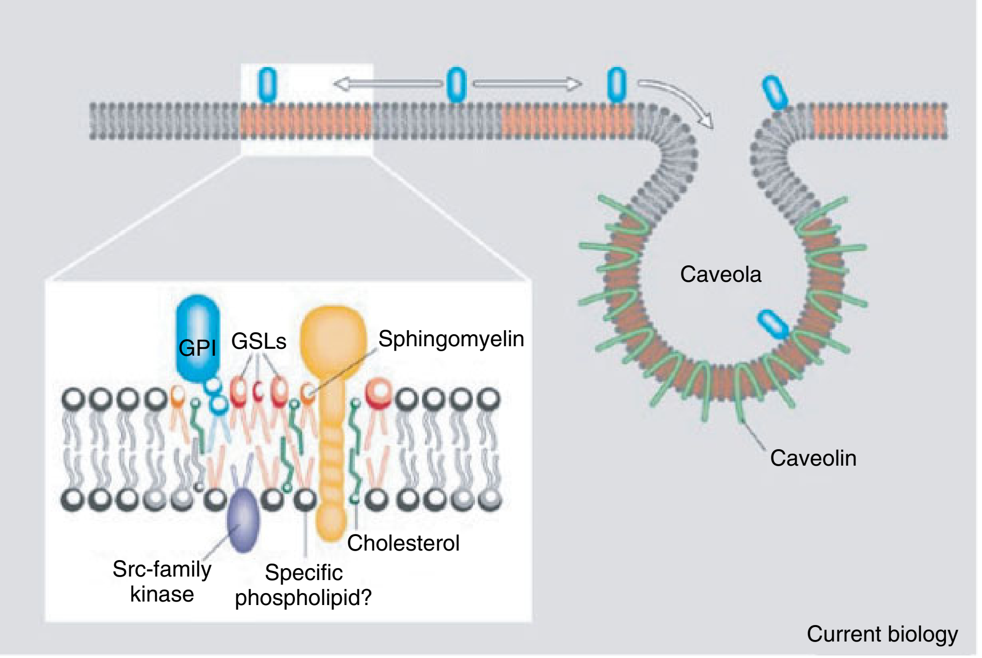


Figure 4 A model for the organization of lipid rafts and caveolae in the membrane. Lipid rafts (red) segregate from the other regions (gray) of the bilayer. The lipid bilayer in rafts is asymmetric, with sphingomyelin (orange) and glycosphingolipids (GSLs; red) enriched in the outer leaflet. Cholesterol (dark green) is present in both leaflets. GPI-anchored proteins (blue) are attached to the outer leaflet of the bilayer, and acylated proteins, such as Src-family nonreceptor tyrosine kinases (purple), are integrated into the cytoplasmic leaflet. Some transmembrane polypeptide-anchored proteins (yellow) are also associated with rafts. Caveolae, with their coat protein caveolin (light green) are subsets of rafts possibly involved in endocytosis. Individual proteins and lipids may move both within and between different rafts. (Reprinted from *Current Biology*, **8**, Hooper NM, Do glycolipid microdomains really exist? R114–R116, (1988), with permission from Elsevier)

rafts (Figure 4) (Simons and Ikonen, 1997; Simons and Toomre, 2000). Owing to intrinsic hydrophobic interactions between the saturated acyl chains of the GPI anchor and the surrounding lipids, as well as interactions between the glycan portion of the anchor and the head groups of the glycosphingolipids, GPI-anchored proteins cluster within these lipid rafts. Recent data show that the two GPI-anchored proteins Thy-1 and the prion protein occupy different, albeit closely adjacent, domains in the plasma membrane (Madore *et al*., 1999) and it is becoming evident that rafts in cells appear to be heterogeneous both in terms of their protein and their lipid content (Pike, 2004).

5.3. *Intracellular targeting signal*

The presence of a GPI anchor often, but not always, causes the intracellular targeting of the protein to the apical surface of polarized epithelial cells (such as intestinal or kidney cells) and to the axonal surface of neurons (Ikonen, 2001).

The association of GPI-anchored proteins with lipid rafts as the proteins and lipids are sorted in the Golgi apparatus may underlie this polarized targeting.

5.4. Transmembrane signaling

Although GPI anchors have no cytoplasmic domain, they can often affect transmembrane signaling. Cross-linking of GPI-anchored proteins with antibodies can mediate an intracellular response, such as a rise in intracellular Ca^{2+}, tyrosine phosphorylation, proliferation, and cytokine induction (Horejsi *et al*., 1999; Ilangumaran *et al*., 2000). The localization of the GPI-anchored proteins in lipid rafts appears to be critical for this function and almost certainly is related to the fact that the cytoplasmic leaflet of lipid rafts is enriched in numerous signaling proteins, such as nonreceptor tyrosine kinases (Figure 4). However, the molecular mechanism by which the signal is passed across the membrane has still to be elucidated.

5.5. Endocytosis via caveolae

GPI-anchored proteins are generally excluded from the classical clathrin-mediated endocytic pathway. Instead, some GPI-anchored proteins, for example the folate receptor, are endocytosed through another structure termed *caveolae* (Nichols, 2003). Caveolae are flask-shaped invaginations of the plasma membrane that contain the coat protein caveolin (Figure 4). Caveolae are also enriched in cholesterol and glycosphingolipids and are insoluble in certain detergents at low temperature and probably represent a subpopulation of the detergent-insoluble lipid rafts.

5.6. Lateral mobility

GPI-anchored proteins lack a cytoplasmic domain and therefore cannot interact directly with the cytoskeleton inside the cell, leading to the assumption that such proteins would be more laterally mobile in the plane of the bilayer. However, the clustering of GPI-anchored proteins in lipid rafts would restrict their lateral movement. Experimentally, some GPI-anchored proteins appear to move more rapidly in the plane of the bilayer (Ishihara *et al*., 1987), while others display a restricted pattern of movement (Zhang *et al*., 1991), indicating that the lateral mobility of any one GPI-anchored protein probably depends on its particular membrane environment.

5.7. Intercellular transfer

Several GPI-anchored proteins have been shown to be transferred from the surface of one cell to another, a phenomenon referred to as *cell painting* (Kooyman *et al*., 1995; Anderson *et al*., 1996). The lack of a cytoplasmic domain certainly is the key to this process and it is known that purified GPI-anchored proteins can spontaneously insert into lipid bilayers. The physiological significance of this process is not clear, but it may enable a particular cell type that does not normally synthesize a particular protein or a cell with limited or no biosynthetic capacity

(e.g., erythrocytes and spermatozoa) to express the GPI-anchored protein on its cell surface under particular circumstances.

Acknowledgments

Work in the author's laboratory on GPI anchors has been supported by the Wellcome Trust, the Medical Research Council, and the Biotechnology and Biological Sciences Research Council of Great Britain.

Further reading

Hooper NM (1998) Do glycolipid microdomains really exist? *Current Biology*, **8**, R114–R116.

References

Anderson SM, Yu G, Giattina M and Miller JL (1996) Intercellular transfer of a glycosylphosphatidylinositol (GPI)-linked protein: release and uptake of CD4-GPI from recombinant adeno-associated virus-transduced HeLa cells. *Proceedings of the National Academy of Sciences USA*, **93**, 5894–5898.

Bordier C (1981) Phase separation of integral membrane proteins in Triton X-114 solution. *Journal of Biological Chemistry*, **256**, 1604–1607.

Brewis IA, Ferguson MAJ, Mehlert A, Turner AJ and Hooper NM (1995) Structures of the glycosyl-phosphatidylinositol anchors of porcine and human membrane dipeptidase. Interspecies comparison of the glycan core structures and further structural studies on the porcine anchor. *Journal of Biological Chemistry*, **270**, 22946–22956.

Brewis IA, Turner AJ and Hooper NM (1994) Activation of the glycosyl-phosphatidylinositol-anchored membrane dipeptidase upon release from porcine kidney membranes by phospholipase C. *Biochemical Journal*, **303**, 633–638.

Brown DA and Rose JK (1992) Sorting of GPI-anchored proteins to glycolipid-enriched membrane subdomains during transport to the apical cell surface. *Cell*, **68**, 533–544.

Ferguson MAJ (1992) Site of palmitoylation of a phospholipase C-resistant glycosyl-phosphatidylinositol membrane anchor. *Biochemical Journal*, **284**, 297–300.

Ferguson MAJ (1999) The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. *Journal of Cell Science*, **112**, 2799–2809.

Fraering P, Imhof I, Meyer U, Strub JM, van Dorsselaer A, Vionnet C and Conzelmann A (2001) The GPI transamidase complex of *Saccharomyces cerevisiae* contains Gaa1p, Gpi8p, and Gpi16p. *Molecular Biology of the Cell*, **12**, 3295–3306.

Hooper NM (2001) Determination of glycosyl-phosphatidylinositol membrane protein anchorage. *Proteomics*, **1**, 748–755.

Hooper NM and Bashir A (1991) Glycosyl-phosphatidylinositol-anchored membrane proteins can be distinguished from transmembrane polypeptide-anchored proteins by differential solubilization and temperature-induced phase separation in Triton X-114. *Biochemical Journal*, **280**, 745–751.

Hooper NM, Broomfield SJ and Turner AJ (1991) Characterization of antibodies to the glycosyl-phosphatidylinositol membrane anchors of mammalian proteins. *Biochemical Journal*, **273**, 301–306.

Hooper NM and McIlhinney RA (1999) Lipid modification of proteins. In *Post-Translational Processing: A practical Approach*, Higgins SJ and Hames BD (Eds.), Oxford University Press: Oxford, pp. 175–203.

Hooper NM and Turner AJ (1988) Ectoenzymes of the kidney microvillar membrane. Differential solubilization by detergents can predict a glycosyl-phosphatidylinositol membrane anchor. *Biochemical Journal*, **250**, 865–869.

Horejsi V, Drbal K, Cebecauer M, Cerny J, Brdicka T, Angelisova P and Stockinger H (1999) GPI-microdomains: a role in signalling via immunoreceptors. *Immunology Today*, **20**, 356–361.

Ikonen E (2001) Roles of lipid rafts in membrane transport. *Current Opinion in Cell Biology*, **13**, 470–477.

Ilangumaran S, He HT and Hoessli DC (2000) Microdomains in lymphocyte signalling: beyond GPI-anchored proteins. *Immunology Today*, **21**, 2–7.

Ishihara A, Hou Y and Jacobson K (1987) The Thy-1 antigen exhibits rapid lateral diffusion in the plasma membrane of rodent lymphoid cells and fibroblasts. *Proceedings of the National Academy of Sciences USA*, **84**, 1290–1293.

Kooyman DL, Byrne GW, McClellan S, Nielsen D, Tone M, Waldmann H, Coffman TM, McCurry KR, Platt JL and Logan JS (1995) In vivo transfer of GPI-linked complement restriction factors from erythrocytes to the endothelium. *Science*, **269**, 89–92.

Low MG and Huang K-S (1991) Factors affecting the ability of glycosylphosphatidylinositol-specific phospholipase D to degrade the membrane anchors of cell surface proteins. *Biochemical Journal*, **279**, 483–493.

Madore N, Smith KL, Graham CH, Jen A, Brady K, Hall S and Morris R (1999) Functionally different GPI proteins are organised in different domains on the neuronal surface. *EMBO Journal*, **19**, 6917–6926.

McConville MJ and Ferguson MAJ (1993) The structure, biosynthesis and function of glycosylated phosphatidylinositols in the parasitic protozoa and higher eukaryotes. *Biochemical Journal*, **294**, 305–324.

McConville MJ and Menon AK (2000) Recent developments in the cell biology and biochemistry of glycosylphosphatidylinositol lipids. *Molecular Membrane Biology*, **17**, 1–16.

Movahedi S and Hooper NM (1997) Insulin stimulates the release of the GPI-anchored membrane dipeptidase from 3T3-L1 adipocytes through the action of a phospholipase C. *Biochemical Journal*, **326**, 531–537.

Nichols B (2003) Caveosomes and endocytosis of lipid rafts. *Journal of Cell Science*, **116**, 4707–4714.

Park SW, Choi K, Lee HB, Park SK, Turner AJ, Hooper NM and Park HS (2002) Glycosyl-Phosphatidylinositol (GPI)-anchored renal dipeptidase is released by a phospholipase C in vivo. *Kidney and Blood Pressure Research*, **25**, 7–12.

Pike LJ (2004) Lipid rafts: heterogeneity on the high seas. *Biochemical Journal*, **378**, 281–292.

Rached E, Hooper NM, James P, Semenza G, Turner AJ and Mantei N (1990) cDNA cloning and expression in *Xenopus laevis* oocytes of pig renal dipeptidase, a glycosyl-phosphatidylinositol-anchored ectoenzyme. *Biochemical Journal*, **271**, 755–760.

Schofield L, Hewitt MC, Evans K, Siomos MA and Seeberger PH (2002) Synthetic GPI as a candidate anti-toxic vaccine in a model of malaria. *Nature*, **418**, 785–789.

Simons K and Ikonen E (1997) Functional rafts in cell membranes. *Nature*, **387**, 569–572.

Simons K and Toomre D (2000) Lipid rafts and signal transduction. *Nature Reviews Molecular Cell Biology*, **1**, 31–39.

Udenfriend S and Kodukula K (1995) How glycosyl-phosphatidylinositol-anchored membrane proteins are made. *Annual Review of Biochemistry*, **64**, 563–591.

Zamze SE, Ferguson MAJ, Collins R, Dwek RA and Rademacher TW (1988) Characterisation of the cross-reacting determinant (CRD) of the glycosyl-phosphatidylinositol membrane anchor of Trypanosoma brucei variant surface glycoprotein. *European Journal of Biochemistry*, **176**, 527–534.

Zhang F, Crise B, Su B, Hou Y, Rose JK, Bothwell A and Jacobson K (1991) Lateral diffusion of membrane-spanning and glycosylphosphatidylinositol-linked proteins: toward establishing rules governing the lateral mobility of membrane proteins. *Journal of Cell Biology*, **115**, 75–84.

Specialist Review

Glycosylation and hepatacellular carcinoma

Anand Mehta
Drexel University College of Medicine, Doylestown, PA, US

1. Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is a cancer arising from the liver and is referred to as primary liver cancer. HCC is the fifth most common cancer in the world and is the third leading cancer killer worldwide (Block *et al*., 2003). Indeed, liver cancer kills almost all patients who have it within a year. The World Health Organization (WHO) estimated that in 2005 there were 560 000 new cases of liver cancer worldwide, and a similar number of patients died as a result of this disease (**http://www.who.int/en**). The low survival rates have been attributed to the late diagnosis and poor therapeutic options (Di Bisceglie *et al*., 1998).

The frequency of liver cancer in southeast and east Asia and sub-Saharan Africa is greater than 20 cases per 100 000 people. In contrast, the frequency of liver cancer in North America and western Europe is much lower; less than 5 per 100 000 people (Parkin, 2006). However, recent data show that the frequency and mortality of liver cancer in the United States is rising. This increase is primarily due to the chronic infection of hepatitis C virus (HCV), an infection of the liver that leads to cirrhosis and HCC (Marrero, 2006).

HCC is characterized by its propensity for vascular invasion, and microscopic venous or macroscopic portal vein invasion are recorded as being the most significant risk factors for recurrence. Indeed, both intrahepatic metastasis and/or multicentric occurrence contribute to the recurrence in the liver remnant as does an initial large tumor size (especially >5 cm). Even after liver transplantation, often viewed as a cure for HCC, intrahepatic tumor recurrence occurs and is especially frequent in tumors >3 cm (Brechot *et al*., 2000). Although the cause of the tumor (viral infection, alcohol etc.) does not appear to be a significant risk factor for recurrence, there are reports of lower rates of recurrence in hepatitis B virus (HBV)-infected individuals compared to HCV patients (Ichikawa *et al*., 1997; Kumada *et al*., 1997).

2. Risk factor for HCC: hepatitis

Infection with HBV and/or HCV is the major etiology of HCC (Benvegnu *et al*., 1994; Brechot, 1996; Hoofnagle, 1999). Both HBV and HCV can

cause chronic infections of the liver and most chronically infected individuals remain asymptomatic for many years and the clinical disease is not apparent until decades later. Nearly 25–40% of all chronic carriers of HBV or HCV eventually develop untreatable liver cancer (Costello, 1999). Indeed, HBV infection is associated with over 80% of all HCC cases worldwide and can be as high as 96% in regions where HBV is endemic (Beasley, 1988). More than 350 million people worldwide are chronically infected with HBV, including 1.25 million in the United States (Hoofnagle, 1997). HCV is estimated to chronically infect 170 million people worldwide, with 4 million in the United States. With 140 000–320 000 new cases of HBV/HCV reported in the United States each year, the *at-risk* population for HCC has been consistently rising.

The progression of liver disease in asymptomatic chronic carriers of HBV and HCV is monitored by serum liver function tests (LFTs) and ultrasound imaging for the detection of small masses in the liver (Marrero, 2006). Many of the constituents of the LFT panel, for example, transaminase levels, vary throughout the course of chronic hepatitis and are of limited use in the early detection of HCC. Ultrasound detection requires at least a 3-cm tumor mass to be present, and often occurs at a stage at which the prognosis is very poor (Brechot, 1987; Hoofnagle, 1997). Since early surgical and chemotherapeutic intervention is the best hope for patient survival (Brechot, 1987; Di Bisceglie *et al*., 1998), early detection of HCC is necessary to identify the need for intervention. Thus, with a highly defined risk population, the potential for a significant impact on the disease is present.

2.1. Protein glycosylation

The attachment of oligosaccharides to proteins, referred to as glycosylation, is a common co- and posttranslational modification that profoundly affects the biological function of proteins (Varki, 1993). There are a multitude of different types of glycosylation; however, the focus of this article is centered on common glycosylation types associated with proteins that pass through the secretory pathway. The biosynthesis of the oligosaccharide part of the glycoprotein is not genetically encoded, in contrast to the amino acid backbone of the protein. Therefore, information about these biologically important glycan structures cannot be directly gained by a genetic approach. In addition, although the same enzymatic glycosylation machinery is available to all proteins that enter the secretory pathway in a given cell, most glycoproteins emerge with a characteristic (conserved) but potentially heterogeneous glycosylation pattern. Therefore, each protein's glycosylation "profile" is a necessary parameter required for complete protein evaluation in any proteomic study.

The majority of proteins that pass through the secretory pathway are glycoproteins (Monteuil *et al*., 1995) and, consequently, the majority of cell surface proteins are glycosylated (Gahmberg and Tolvanen, 1996). Glycans can be attached to proteins via a variety of linkages and mechanisms. The most common types of glycosylation for cell surface and secreted molecules are asparagine (N-) and

oxygen (O-) linked glycosylation. N-glycosylation is defined by the addition of a preassembled oligosaccharide by the oligosaccharyltransferase (OST) to the side chain nitrogen of an asparagine residue located within a primary sequence signal motif (Asn-X-Ser/Thr, where X can be any amino acid except proline). The presence of this sequon is required for N-glycosylation, but not all sequons are occupied. The transfer of the tetradecasaccharide occurs cotranslationally in the rough endoplasmic reticulum (RER). The oligosaccharide can be involved in RER protein folding (Ou *et al*., 1993; Hammond *et al*., 1994) before undergoing remodeling prior to delivery to the target destination. The extent of remodeling, carried out by a series of competing enzymatic reactions throughout the ER and Golgi apparatus, gives rise to the different classes of N-glycans: oligomannose (least remodeling), hybrid (intermediate), and complex (full remodeling) (Kornfeld and Kornfeld, 1985). Although all N-glycans retain a common pentasaccharide core structure, diversity is potentially huge, for example, the cell surface molecule CD59 contains over 100 different oligosaccharides attached to its single N-glycosylation site (Rudd *et al*., 1997). Mucin-like O-glycosylation is initiated by the transfer of a single monosaccharide (GalNAc) and elongation leads to potentially more complex structures due to the absence of a single conserved core structure. There is a family of transferases which catalyze the initial GalNAc transfer (Clausen and Bennett, 1996; Bennett *et al*., 1998), each member with different specificity (Wandall *et al*., 1997; Hassan *et al*., 2000), consequently, no single O-glycosylation sequon emerges, although certain rules emerge following the evaluation of the enzyme family members.

3. Ubiquity of glycosylation and changes in cancer and development

An investigation of the SWISS-PROT database has shown that almost two thirds of all proteins contain the N-glycosylation sequon, NXS/T, and these may be glycoproteins (Apweiler *et al*., 1999). Assessing the probability of the occupancy of the sequon, it was concluded that half of all proteins are glycoproteins. The number of potential O-glycosylation sites cannot be accurately assessed; therefore, the true number of glycoproteins could be even greater than is currently believed.

Presently, less than 10% of proteins present in the databases are recorded as glycoproteins and even fewer ($<$0.01%) have been well characterized. These figures reflect the fact that accurate and reliable detection, quantification, and characterization of oligosaccharides has been technically very difficult and time consuming. This is unfortunate since alteration in the oligosaccharides associated with glycoproteins, in particular, at the cell surface, is one of the many molecular changes that accompany malignant transformations (Hakomori, 1996; Kobata, 1998). For example, it is well established that in HCC there is an elevation in the amount of fucosylated (Fc) oligosaccharides associated with the secreted α-feto protein (AFP) and is due to, in part, an abnormal secretion system (Miyoshi *et al*., 1999), and other secreted proteins have been shown to be aberrantly glycosylated

(Leray *et al*., 1989; Yamashita *et al*., 1989; Aoyagi, 1995). Cell surface proteins have yet to be evaluated in this system. In addition, there are many examples of abnormal O-glycosylation in cancers of the gastrointestinal tract, pancreas, ovaries, breast, and lung (Brockhausen, 1999; Brockhausen, 2000).

4. Evidence of molecular changes associated with hepatocellular carcinoma

The literature reports many biomolecular changes that occur during the development of HCC. These include (but are not limited to) increased serum activin-A (Pirisi *et al*., 2000), decreased serum retinol levels (Newsome *et al*., 2000), expression of cellular ets-1 (Baumann *et al*., 2000); repression of the cell surface marker CD81 (Inoue *et al*., 2001); and numerous changes that can be identified by comparative genomic hybridization (Zondervan *et al*., 2000) or, in hepatoma cell lines, by complementary DNA (cDNA) microarray analysis (Kawai *et al*., 2001; Patil *et al*., 2005). However, by far the most useful change associated with HCC is the increase in serum AFP (Tatarinov, 1963). AFP is an oncofetal glycoprotein that is (hypothesized to be) produced by transformed liver cells after the development of HCC. However, AFP can be produced under many circumstances, including other liver diseases (Alpert *et al*., 1968; Ruoslahti *et al*., 1974; Di Bisceglie and Hoofnagle, 1989), and is not a definitive marker for the development of HCC. Analysis of the regulatory mechanisms of increased AFP synthesis in hepatic injury and in malignant transformations (Sawadaishi *et al*., 1988; Nakabayashi *et al*., 1989) has been unsuccessful in distinguishing elevation of AFP between HCC and chronic liver disease. The detection of AFP is currently one of the methods by which the development of HCC is detected, although the poor sensitivity of this marker has led people to question its true usefulness in the detection of HCC (Bruix and Sherman, 2005).

5. Change in glycosylation associated with HCC

The literature indicates that changes in glycosylation occur during the development of HCC. The most notable change is an increase in the level of core α-1,6-linked fucosylation of AFP (Breborowicz *et al*., 1981; Miyazaki *et al*., 1981). In HCC and in testicular cancer, the glycosylation of AFP shifts from a simple biantennary glycan to an α-1,6-linked core fucosylated biantennary glycan (Figure 1). These changes have been observed by both direct glycan sequencing of AFP and by increased reactivity of AFP with a variety of lectins that preferentially bind to fucose-containing glycan (Johnson *et al*., 2000). The species of AFP that reacts preferentially with the lectin *lens culinaris* (LCH) is referred to as AFP-L3. Several reports have clearly indicated that AFP-L3 is a more specific marker of HCC than the total AFP protein levels (Taketa *et al*., 1985; Taketa *et al*., 1990; Shiraki *et al*.,

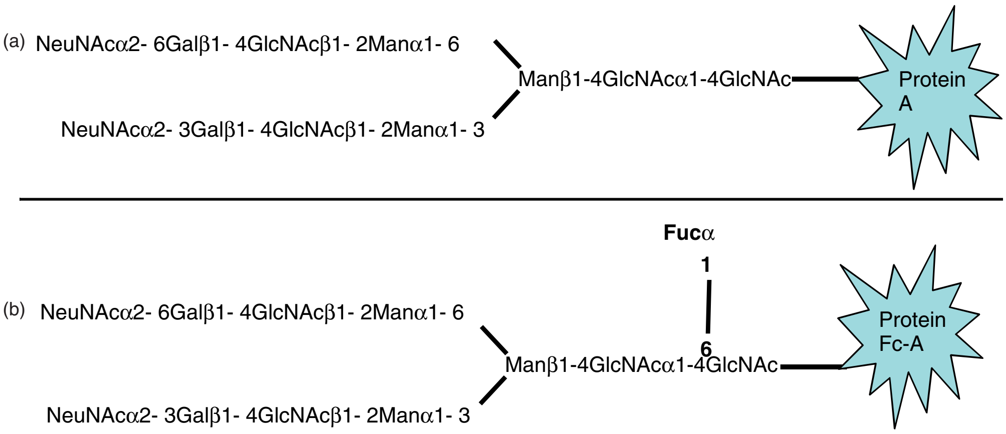


Figure 1 Example of glycoforms. A protein with the exact same amino acid sequence but with a different sugar chain is referred to as a glycoform. Normally, liver cells produce glycoproteins that contain the sugar structure seen in panel A. This is referred to as a biantennary glycan (A2G2). In HCC, the liver cells start to attach a fucose residue to the glycan chain, creating a fucosylated glycoprotein (glycoform, referred to with the Fc prefix). For picture: *N*-acetylglucosamine (GlcNAc); mannose (Man); galactose; (Gal); sialic acid (NeuNAc); fucose (Fuc)

1995). Indeed, AFP-L3 gained approval from the US Food and Drug administration (FDA) in 2005 as the only diagnostic assay for HCC.

In addition to the increases in fucosylation observed on AFP, other changes in N-linked glycosylation have also been observed. These changes include the addition of bisecting *N*-acetylglucosamine (GlcNAc) residues along with increased α-2,6-linked sialation (Morelle *et al*., 2006). These changes have been observed on a more global scale, rather than through the examination of a single glycoprotein.

Although the molecular mechanism of increased fucosylation in HCC is not clear (Noda *et al*., 1998), (Hutchinson *et al*., 1991; Miyoshi *et al*., 1999), it is known that the increase is not restricted to AFP (Yamashita *et al*., 1989; Naitoh *et al*., 1999). Results from several groups have indicated that other liver-derived glycoproteins such as α-1 acid glycoprotein, and α-1 antitrypsin also become fucosylated with the development of HCC, and a recent study has proposed that these glycoforms may be valuable biomarkers of HCC (Naitoh *et al*., 1999). However, a comprehensive comparative analysis of all the fucosylated glycoproteins in HCC patients has not been performed. This type of study has been limited because of the absence of a suitable technology to allow the examination of large pools of unknown proteins. With the advent of sensitive glycan analysis and proteomic technologies, the ability to comprehensively identify all the fucosylated proteins in patients with HCC and to identify these proteins for the development of diagnostic markers is now a possibility.

Recent work has identified global glycan changes that can be observed that correlate with the presence of HCC. This has been performed either on glycan analysis of whole serum (Callewaert *et al*., 2004) or serum that has been depleted of immunoglobulin (Block *et al*., 2005; Comunale *et al*., 2006). In both circumstances, an increase in core fucosylation was observed with the development of HCC. An example of the increase in fucosylation is shown in Figure 2,

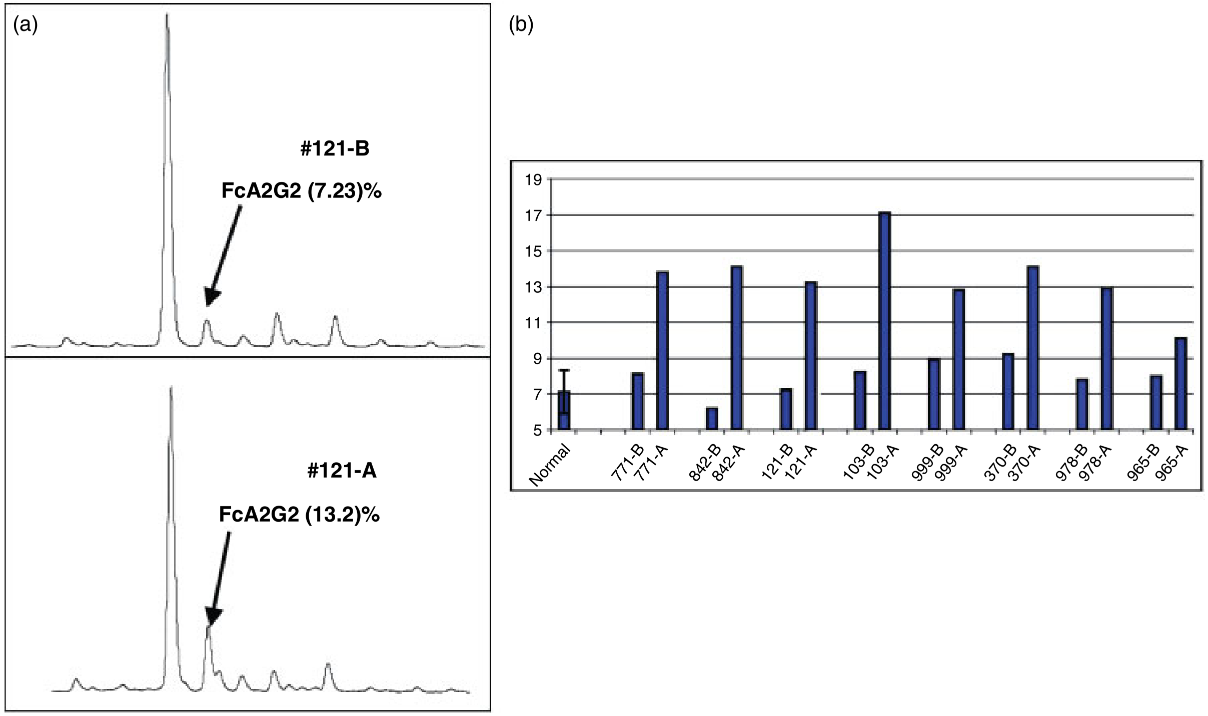


Figure 2 The level of the FcA2G2 glycan in people as a function of time. (a) The level of the FcA2G2 structure in a patient either before (B) or after (A) the diagnosis of cancer as determined by normal phase HPLC analysis (Guile *et al*., 1996). The level of the FcA2G2 structure increases from 7.23% of the total glycan pool to over 13% of the total glycan pool after the diagnosis of cancer. (b) Levels of the FcA2G2 structure in eight individuals either before or after the diagnosis of cancer. For graph, the Y-axis is the percentage of FcA2G2 structure in each individual as a function of total released glycan. The X-axis is the sample number. From Comunale *et al*., 2006

where serum from patients was analyzed by normal phase high performance liquid chromatography (HPLC) (Guile *et al*., 1996; Rudd *et al*., 1999). This small sample set consisted of 16 samples from eight patients at two time points: before the development of cancer or after the diagnosis of cancer. In this case, all patients classified as "before" were clinically diagnosed with cirrhosis (Comunale *et al*., 2006). Samples classified as "after" were clinically diagnosed with HCC by ultrasound and biopsy. As shown in Figure 2, patients before the diagnosis of HCC have much lower levels of the core fucosylated biantennary glycan (FcA2G2) as compared to after the diagnosis of HCC. For example, patient 121 had 7.23% of the FcA2G2 glycan before the diagnosis of HCC and 13.2% after the diagnosis of HCC. As shown in Figure 2(b), this upward trend was true of all patients examined. Consistent with our published results, these results highlight the fact that increased levels of α-1,6-linked core fucosylation are associated with the development of HCC. What was more impressive was that increased levels of core fucosylation were observed in patients that were AFP negative, and we had not been able to make this observation using the conventional screening methods. That is, as shown in figure 2(b), patients 965, 370, 842, 999, and 978 were all AFP negative, yet had increased levels of the FcA2G2 glycan (Comunale *et al*., 2006).

To identify those proteins that had increased fucosylation, the fucosylated glycoproteins associated with sera from either pooled normal or pooled HCC-positive individuals were extracted using fucose-specific lectins and the proteome analyzed by either two-dimensional gel electrophoresis (2DE) or by a simple LC MS/MS–based methodology designed to identify fucosylated peptides (described in Comunale *et al*., 2006). Using these methodologies, it was determined that glycoproteins such as fucosylated α-1-acid glycoprotein, Fc-ceruloplasmin, Fc-α-2-macroglobulin, Fc-hemopexin, Fc-Apo-D, Fc-HBsAg, and Fc-Kininogen increased in patients with HCC, while the levels of Fc-haptoglobin decreased in the these patients (all listed in Figure 3) (Comunale *et al*., 2006). Many of these proteins are acute phase proteins that are secreted by the liver. Although the analysis of a panel of fucosylated proteins as a marker of liver disease has been postulated before (Turner, 1992; Turner, 1995), a comprehensive analysis of the total fucosylated proteome as a function of the disease has not been performed. However, with the current protein chip–based systems, the analysis of a large number of fucosylated proteins can now be used to determine if a fucosylated "biosignature" exists that can be utilized to diagnose the disease.

6. Conclusion

Changes in N-linked glycosylation have long been associated with the development of the disease (Kim and Varki, 1997; Dennis *et al*., 1999; Taylor-Papadimitriou *et al*., 1999; Kellokumpu *et al*., 2002; Peracaula *et al*., 2003). In the case of liver cancer, it has been clearly shown that an increase in core fucosylation is associated with the development of the disease, in both animal models and in people (Breborowicz *et al*., 1981; Miyazaki *et al*., 1981; Taketa *et al*., 1985; Taketa *et al*., 1990; Kelleher *et al*., 1992; Ajdukiewicz *et al*., 1993; Shiraki *et al*., 1995; Block *et al*., 2005; Comunale *et al*., 2006). This finding has led to the only diagnostic

Identified Fucosylated Protein	
Fc-GP-73	Dr
Fc-Hemopexin	
Fc-HBsAg	
Fc-AFP	
Fc-Alpha-acid glycoprotein	
Fc-Alpha-1-antichymotrypsin	
Fc-Alpha-1-antitrypsin	
Fc-Serotransferrin	
Fc-Ceruloplasmin	
Fc-Alpha-2-macroglobulin	
Fc-Alpha-2-HS-glycoprotein (Fetuin A)	
Fc-Haptoglobin	
Fc-Fibrinogen gamma chain precursor	
Fc-Complement factor B	
Fc-IgG	
Fc-IgA	
Fc-APO-D	
Fc-Ig M	
Fc-Kininogen	
Fc-Histidine rch glycoprotein	
Fc-complement C1s component	
Fc-alpha 1 B glycoprotein	
Fc-B-2-glycoprotein (apo H)	

Figure 3 The proteins identified as being altered in patients with HCC. Figure is modified from Comunale *et al*., 2006

marker of HCC approved by the US Food and Drug Administration. However, for HCC, this may represent just the tip of the iceberg. As many more glycoproteins with altered glycans are identified, a marker that can identify early liver cancer with 100% sensitivity and 100% specificity may soon be at hand.

References

Ajdukiewicz AB, Kelleher PC, Krawitt EL, Walters CJ, Mason PB, Koff RS and Belanger L (1993) Alpha-fetoprotein glycosylation is abnormal in some hepatocellular carcinoma, including white patients with a normal alpha-fetoprotein concentration. *Cancer Letters*, **74**(1-2), 43–50.

Alpert ME, Uriel J and de Nechaud B (1968) alpha fetogloblin in the diagnosis of human hepatoma. *The New England Journal of Medicine*, **278**, 984–986.

Aoyagi Y (1995) Carbohydrate-based measurements on alpha-fetoprotein in the early diagnosis of hepatocellular carcinoma. *Glycoconjugate Journal*, **12**(3), 194–199.

Apweiler R, Hermjakob H and Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta*, **1473**(1), 4–8.

Baumann NA, Vidugiriene J, Machamer CE and Menon AK (2000) Cell surface display and intracellular trafficking of free glycosylphosphatidylinositols in mammalian cells. *The Journal of Biological Chemistry*, **275**(10), 7378–7389.

Beasley RP (1988) Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer*, **61**(10), 1942–1956.

Bennett EP, Weghuis DO, Merkx G, van Kessel AG, Eiberg H and Clausen H (1998) Genomic organization and chromosomal localization of three members of the UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferase family. *Glycobiology*, **8**(6), 547–555.

Benvegnu L, Fattovich G, Noventa F, Tremolada F, Chemello L, Cecchetto A and Alberti A (1994) Concurrent hepatitis B and C virus infection and risk of hepatocellular carcinoma in cirrhosis. *Cancer*, **74**(9), 2442–2448.

Block TM, Comunale MA, Lowman M, Steel LF, Romano PR, Fimmel C, Tennant BC, London WT, Evans AA, Blumberg BS, *et al* (2005) Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(3), 779–784.

Block TM, Mehta AS, Fimmel CJ and Jordan R (2003) Molecular viral oncology of hepatocellular carcinoma. *Oncogene*, **22**(33), 5093–5107.

Breborowicz J, Mackiewicz A and Breborowicz D (1981) Microheterogeneity of alpha-fetoprotein in patient serum as demonstrated by lectin affino-electrophoresis. *Scandinavian Journal of Immunology*, **14**(1), 15–20.

Brechot C (1987) Hepatitis B virus (HBV) and hepatocellular carcinoma. HBV DNA status and its implications. *Journal of Hepatology*, **4**(2), 269–279.

Brechot C (1996) Hepatitis B and C viruses and primary liver cancer. *Bailliere's Clinical Gastroenterology*, **10**(2), 335–373.

Brechot C, Gozuacik D, Murakami Y and Paterlini-Brechot P (2000) Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). *Seminars in Cancer Biology*, **10**(3), 211–231.

Brockhausen I (1999) Pathways of O-glycan biosynthesis in cancer cells. *Biochimica et Biophysica Acta*, **1473**(1), 67–95.

Brockhausen I (2000) O-linked chain glycosyltransferases. *Methods in Molecular Biology*, **125**, 273–293.

Bruix J and Sherman M (2005) Management of hepatocellular carcinoma. *Hepatology*, **42**(5), 1208–1236.

Callewaert N, Van Vlierberghe H, Van Hecke A, Laroy W, Delanghe J and Contreras R (2004) Noninvasive diagnosis of liver cirrhosis using DNA sequencer based total serum protein glycomics. *Nature Medicine*, **10**(4), 429–434.

Clausen H and Bennett EP (1996) A family of UDP-GalNAc: polypeptide N-acetylgalactosaminyl-transferases control the initiation of mucin-type O-linked glycosylation. *Glycobiology*, **6**(6), 635–646.

Comunale MA, Lowman M, Long RE, Krakover J, Philip R, Seeholzer S, Evans AA, Hann HWL, Block TM and Mehta AS (2006) Proteomic analysis of serum associated fucosylated glycoproteins in the development of primary hepatocellular carcinoma. *Journal of Proteome Research*, **6**(5), 308–315.

Costello CE (1999) Bioanalytic applications of mass spectrometry. *Current Opinion in Biotechnology*, **10**(1), 22–28.

Dennis JW, Granovsky M and Warren CE (1999) Glycoprotein glycosylation and cancer progression. *Biochimica et Biophysica Acta*, **1473**(1), 21–34.

Di Bisceglie AM, Carithers RL Jr and Gores GJ (1998) Hepatocellular carcinoma. *Hepatology*, **28**(4), 1161–1165.

Di Bisceglie AM and Hoofnagle JH (1989) Elevations in serum alpha-fetoprotein levels in patients with chronic hepatitis B. *Cancer*, **64**(10), 2117–2120.

Gahmberg CG and Tolvanen M (1996) Why mammalian cell surface proteins are glycoproteins [published erratum appears in (1996) *Trends in Biochemical Sciences* Dec;21(12):491]. *Trends in Biochemical Sciences*, **21**(8), 308–311.

Guile GR, Rudd PM, Wing DR, Prime SB and Dwek RA (1996) A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. *Analytical Biochemistry*, **240**(2), 210–226.

Hakomori S (1996) Tumor malignancy defined by aberrant glycosylation and sphingo(glyco)lipid metabolism. *Cancer Research*, **56**(23), 5309–5318.
Hammond C, Braakman I and Helenius A (1994) Role of N-linked oligosaccharide recognition, glucose trimming, and calnexin in glycoprotein folding and quality control. *Proceedings of the National Academy of Sciences of the United States of America*, **91**(3), 913–917.
Hassan H, Reis CA, Bennett EP, Mirgorodskaya E, Roepstorff P, Hollingsworth MA, Burchell J, Taylor-Papadimitriou J and Clausen H (2000) The lectin domain of UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase-T4 directs its glyco-peptide specificities. *The Journal of Biological Chemistry*, **275**(49): 38197–38205.
Hoofnagle JH (1997) Hepatitis C: the clinical spectrum of disease. *Hepatology*, **26**(3 Suppl 1): 15S–20S.
Hoofnagle JH (1999) Management of hepatitis C: current and future perspectives. *Journal of Hepatology*, **31**(Suppl 1), 264–268.
Hutchinson WL, Du MQ, Johnson PJ and Williams R (1991) Fucosyltransferases: differential plasma and tissue alterations in hepatocellular carcinoma and cirrhosis. *Hepatology*, **13**(4), 683–688.
Ichikawa N, Fujimoto J, Okamoto E, Yamanaka N and Nishigami T (1997) Cellular DNA content and histopathological analysis in hepatocellular carcinoma with multiple nodules. *Surgery Today*, **27**(6), 483–490.
Inoue G, Horiike N and Onji M (2001) The CD81 expression in liver in hepatocellular carcinoma. *International Journal of Molecular Medicine*, **7**(1), 67–71.
Johnson PJ, Poon TC, Hjelm NM, Ho CS, Blake C and Ho SK (2000) Structures of disease-specific serum alpha-fetoprotein isoforms. *British Journal of Cancer*, **83**(10), 1330–1337.
Kawai HF, Kaneko S, Honda M, Shirota Y and Kobayashi K (2001) Alpha-fetoprotein-producing hepatoma cell lines share common expression profiles of genes in various categories demonstrated by cDNA microarray analysis. *Hepatology*, **33**(3), 676–691.
Kelleher PC, Walters CJ, Myhre BD, Tennant BC, Gerin JL and Cote PJ (1992) Altered glycosylation of alpha-fetoprotein in hepadnavirus-induced hepatocellular carcinoma of the woodchuck. *Cancer Letters*, **63**(2), 93–99.
Kellokumpu S, Sormunen R and Kellokumpu I (2002) Abnormal glycosylation and altered Golgi structure in colorectal cancer: dependence on intra-Golgi pH. *FEBS Letters*, **516**(1–3), 217–224.
Kim YJ and Varki A (1997) Perspectives on the significance of altered glycosylation of glycoproteins in cancer. *Glycoconjugate Journal*, **14**(5), 569–576.
Kobata A (1998) A retrospective and prospective view of glycopathology. *Glycoconjugate Journal*, **15**(4), 323–331.
Kornfeld R and Kornfeld S (1985) Assembly of asparagine-linked oligosaccharides. *Annual Review of Biochemistry*, **54**, 631–664.
Kumada T, Nakano S, Takeda I, Sugiyama K, Osada T, Kiriyama S, Sone Y, Toyoda H, Shimada S, Takahashi M, *et al* (1997) Patterns of recurrence after initial treatment in patients with small hepatocellular carcinoma. *Hepatology*, **25**(1), 87–92.
Leray G, Deugnier Y, Jouanolle AM, Lehry D, Bretagne JF, Campion JP, Brissot P and LeTreut A (1989) Biochemical aspects of alpha-L-fucosidase in hepatocellular carcinoma. *Hepatology*, **9**, 249–252.
Marrero JA (2006) Hepatocellular carcinoma. *Current Opinion in Gastroenterology*, **22**(3), 248–253.
Miyazaki J, Endo Y and Oda T (1981) Lectin affinities of alpha-fetoprotein in liver cirrhosis, hepatocellular carcinoma and metastatic liver tumor. *Acta Hepatologica Jpn*, **22**, 1559–1568.
Miyoshi E, Noda K, Yamaguchi Y, Inoue S, Ikeda Y, Wang W, Ko JH, Uozumi N, Li W and Taniguchi N (1999) The alpha1-6-fucosyltransferase gene and its biological significance. *Biochimica et Biophysica Acta*, **1473**(1), 9–20.
Monteuil J, Vliegenthart JFG and Schachter H (1995) *Glycoproteins*, Elsevier: Oxford.
Morelle W, Flahaut C, Michalski JC, Louvet A, Mathurin P and Klein A (2006) Mass spectrometric approach for screening modifications of total serum N-glycome in human diseases: application to cirrhosis. *Glycobiology*, **16**(4), 281–293.

Naitoh A, Aoyagi Y and Asakura H (1999) Highly enhanced fucosylation of serum glycoproteins in patients with hepatocellular carcinoma. *Journal of Gastroenterology and Hepatology*, **14**(5), 436–445.

Nakabayashi H, Watanabe K, Saito A, Otsuru A, Sawadaishi K and Tamaoki T (1989) Transcriptional regulation of alpha-fetoprotein expression by dexamethasone in human hepatoma cells. *The Journal of Biological Chemistry*, **264**(1), 266–271.

Newsome PN, Beldon I, Moussa Y, Delahooke TE, Poulopoulos G, Hayes PC and Plevris JN (2000) Low serum retinol levels are associated with hepatocellular carcinoma in patients with chronic liver disease. *Alimentary Pharmacology and Therapeutics*, **14**(10), 1295–1301.

Noda K, Miyoshi E, Uozumi N, Gao CX, Suzuki K, Hayashi N, Hori M and Taniguchi N (1998) High expression of alpha-1-6 fucosyltransferase during rat hepatocarcinogenesis. *International Journal of Cancer*, **75**(3), 444–450.

Ou WJ, Cameron PH, Thomas DY and Bergeron JJ (1993) Association of folding intermediates of glycoproteins with calnexin during protein maturation. *Nature*, **364**(6440), 771–776.

Parkin DM (2006) The global health burden of infection-associated cancers in the year 2002. *International Journal of Cancer*, **118**(12): 3030–3044.

Patil MA, Chua MS, Pan KH, Lin R, Lih CJ, Cheung ST, Ho C, Li R, Fan ST, Cohen SN, *et al* (2005) An integrated data analysis approach to characterize genes highly expressed in hepatocellular carcinoma. *Oncogene*, **24**(23), 3737–3747.

Peracaula R, Tabares G, Royle L, Harvey DJ, Dwek RA, Rudd PM and de Llorens R (2003) Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins. *Glycobiology*, **13**(6), 457–470.

Pirisi M, Fabris C, Luisi S, Santuz M, Toniutto P, Vitulli D, Federico E, Del Forno M, Mattiuzzo M, Branca B, *et al* (2000) Evaluation of circulating activin-A as a serum marker of hepatocellular carcinoma. *Cancer Detection and Prevention*, **24**(2), 150–155.

Rudd PM, Mattu TS, Zitzmann N, Mehta A, Colominas C, Hart E, Opdenakker G and Dwek RA (1999) Glycoproteins: rapid sequencing technology for N-linked and GPI anchor glycans. *Biotechnology and Genetic Engineering Reviews*, **16**, 1–21.

Rudd PM, Morgan BP, Wormald MR, Harvey DJ, van den Berg CW, Davis SJ, Ferguson MA and Dwek RA (1997) The glycosylation of the complement regulatory protein, human erythrocyte CD59. *The Journal of Biological Chemistry*, **272**(11), 7229–7244.

Ruoslahti E, Salaspuro M, Pihko H, Andersson L and Seppala M (1974) Serum alpha-fetoprotein: diagnostic significance in liver disease. *British Medical Journal*, **2**(918), 527–529.

Sawadaishi K, Morinaga T and Tamaoki T (1988) Interaction of a hepatoma-specific nuclear factor with transcription-regulatory sequences of the human alpha-fetoprotein and albumin genes. *Molecular and Cellular Biology*, **8**(12), 5179–5187.

Shiraki K, Takase K, Tameda Y, Hamada M, Kosaka Y and Nakano T (1995) A clinical study of lectin-reactive alpha-fetoprotein as an early indicator of hepatocellular carcinoma in the follow-up of cirrhotic patients. *Hepatology*, **22**(3), 802–807.

Taketa K, Ichikawa E, Taga H and Hirai H (1985) Antibody-affinity blotting, a sensitive technique for the detection of alpha-fetoprotein separated by lectin affinity electrophoresis in agarose gels. *Electrophoresis*, **6**, 492–497.

Taketa K, Sekiya C, Namiki M, Akamatsu K, Ohta Y, Endo Y and Kosaka K (1990) Lectin-reactive profiles of alpha-fetoprotein characterizing hepatocellular carcinoma and related conditions. *Gastroenterology*, **99**(2), 508–518.

Tatarinov YS (1963) *The discovery of fetal globulin in patients with primary cancer of the liver*, First International Biochemistry Congress of the USSR, Moscow Academy of Sciences USSR: Moscow.

Taylor-Papadimitriou J, Burchell J, Miles DW and Dalziel M (1999) MUC1 and cancer. *Biochimica et Biophysica Acta*, **1455**(2-3), 301–313.

Turner GA (1992) N-glycosylation of serum proteins in disease and its investigation using lectins. *Clinica Chimica Acta*, **208**(3), 149–171.

Turner GA (1995) Haptoglobin. A potential reporter molecule for glycosylation changes in disease. *Advances in Experimental Medicine and Biology*, **376**, 231–238.

Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, **3**(2), 97–130.

Wandall HH, Hassan H, Mirgorodskaya E, Kristensen AK, Roepstorff P, Bennett EP, Nielsen PA, Hollingsworth MA, Burchell J, Taylor-Papadimitriou J, *et al* (1997) Substrate specificities of three members of the human UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase family, GalNAc-T1, -T2, and -T3. *The Journal of Biological Chemistry*, **272**(38), 23503–23514.

Yamashita K, Koide N, Endo T, Iwaki Y and Kobata A (1989) Altered glycosylation of serum transferrin of patients with hepatocellular carcinoma. *The Journal of Biological Chemistry*, **264**(5), 2415–2423.

Zondervan PE, Wink J, Alers JC, JN IJ, Schalm SW, de Man RA and van Dekken H (2000) Molecular cytogenetic evaluation of virus-associated and non-viral hepatocellular carcinoma: analysis of 26 carcinomas and 12 concurrent dysplasias. *The Journal of Pathology*, **192**(2), 207–215.

Specialist Review

Protein glycosylation and renal cancer

Roisean E. Ferguson, Naveen S. Vasudev, Peter J. Selby and Rosamonde E. Banks
Cancer Research UK Clinical Centre, University of Leeds, Leeds, UK

1. Introduction

Cancer is the second most common cause of death in the Western world and was responsible for over 6.2 million deaths in 2000 worldwide (International Agency for Research on Cancer, 2003). It is a disease that poses many challenges to the clinician, including early detection, monitoring of disease progression, predicting patient prognosis, optimization of treatment selection, and monitoring of therapeutic response.

One of the critical factors in the survival from cancer is detection at an early stage, as, with few notable exceptions, survival rates for advanced cancers have remained relatively static over the last 20 years (National Cancer Institute, Surveillance Epidemiology, and End Results program **http://seer.cancer.gov/**). The impact on survival by combining early detection with current therapies would be significant as outcomes generally improve when treatment is applied to a disease that is confined to the tissue of origin. The identification of diagnostic markers for earlier diagnosis coupled with the identification of novel therapeutic targets is central to the future successful treatment of cancer. To this end, an ongoing challenge in cancer research is to identify disease-associated differences between normal and cancer cells so that they might be harnessed for diagnosis and targeted for more selective treatments.

It is well established that cancer development proceeds via a succession of genetic alterations, each providing an advantage that ensures their survival, generally resulting in an insensitivity to anti-growth signals, evasion of apoptosis, sustained angiogenesis, and metastasis (Hanahan and Weinberg, 2000). Ultimately, these genetic mutations are manifest at a protein level, resulting in changes in expression, post-translational modification, activity, and localization, all of which affect cellular function. It is important to characterize the functional consequences of such genetic alterations in terms of protein structure and function to address what changes are involved in the transformation of a normal cell to a malignant one. The identification and understanding of these changes are fundamental to the identification of novel markers and therapeutic targets.

There is abundant evidence to show that malignant transformation is associated with a dramatic alteration of cellular glycosylation (Dennis *et al*., 1999a; Orntoft and Vestergaard, 1999). Since glycosylated proteins (glycoproteins) coat all eukaryotic cells, such changes have effects on carbohydrate recognition mechanisms implicated in cell adhesion and motility as well as intracellular signaling events, enabling tumor development (Kim and Varki, 1997; Dennis *et al*., 1999a). Recent technological advances in the field of glycobiology have shed some light on the functional significance of glycosylation changes, and there is now much research effort to elucidate disease-specific glycans to allow specific targeting and destruction of cancer cells or to employ these glycans for diagnostics (Shriver *et al*., 2004; Dube and Bertozzi, 2005; Fuster and Esko, 2005).

2. Glycans in cancer development

The nature of glycosylation and its role in biological processes is covered elsewhere within this section. There is abundant evidence to suggest that cell surface glycosylation is altered in cancer. Earlier observations noted that the glycoproteins in transformed cells were larger than those found in healthy cells (Meezan *et al*., 1969), and this has been demonstrated many times since. This can reflect altered cellular metabolism, but changes in glycosylation have also been directly associated with tumor development and progression. Examples include the role of glycans in cell–cell recognition, cancer cell metastasis, and homing (Dennis *et al*., 1999a; Orntoft and Vestergaard, 1999; Gorelik *et al*., 2001; Kannagi *et al*., 2004).

A number of common changes in glycosylation occur. Epithelial tumors often overproduce mucin glycoproteins such as MUC1 and CA125 (MUC16), and a number of tumor marker assays currently in use, including CA 125 and CA 19-9, are based on epitopes that are carried on mucins (Hollingsworth and Swanson, 2004). Changes in expression levels of glycosyltransferases in cancer cells lead to the frequent expression of a number of terminal glycans such as sLe^x, sLe^y, sLe^a sialyl-Tn, GloboH, and polysialic acid (for a review, see Hakomori, 2000). A common feature of malignant cells is an increase in size and branching of *N*-glycans, attributed to the increased expression of *N*-acetylglucosaminyltransferase (GnT-V (Dennis *et al*., 1987)). The association between β1-6 branching and tumor progression (Kim and Varki, 1997; Dennis *et al*., 1999a; Dennis *et al*., 1999b; Orntoft and Vestergaard, 1999) explains the observations of the increased size of tumor membrane glycoproteins (Meezan *et al*., 1969). The increased branching produces increased sites for sialic acids, which, in conjunction with upregulation of the sialyltransferases, leads to a global increase in sialylation. Formation of less complex structures in O-linked mucin-type carbohydrate chains is another common change (Kim and Varki, 1997; Orntoft and Vestergaard, 1999). MUC1 is a good example of the latter where it is both upregulated and aberrantly glycosylated in $>90\%$ of breast cancers (Taylor-Papadimitriou *et al*., 2002). Truncation of O-linked glycans leads to the appearance of novel glycan epitopes such as the sialyl-Tn and Thomsen–Friedenreich antigens while simultaneously making the peptide of the protein more exposed. Some cancers exhibit an increased expression of gangliosides such as GM1, GM2, GD2, and GD3 (Hakomori, 2000). In addition,

high levels of hyaluronan, which is a major component of the extracellular matrix, are associated with malignant progression in many cancers including breast and colorectal cancers (Fjeldstad and Kolset, 2005). There are numerous other examples where the carbohydrate phenotype has been an indicator of the malignant potential. In particular, a correlation between the expression of sLe^x and increased branching of *N*-linked glycans and metastatic potential has been demonstrated in a number of studies (Demetriou *et al*., 1995; Granovsky *et al*., 2000). Specific examples of alterations associated with renal cancer are detailed later.

3. Studying the cancer glycoproteome: a potentially rich source of biomarkers and therapeutic targets

Because cancer cell glycans differ from those found in normal healthy cells, it follows that it should be possible to target them specifically on the basis of their altered glycosylation or use the cancer-specific glycans for diagnosis (Dwek and Brooks, 2004). In some cases, glycoproteins produced by the cancer cells may be secreted or released into the blood stream or urine, forming the basis of a diagnostic test for the cancer.

Unlike proteomics and genomics, the field of glycoproteomics has been hindered by the sheer complexity of the macromolecules involved. Progress toward delineating the molecular basis of the glycan function, in particular, has been extremely challenging as it has many roles that depend not only on the protein to which it is attached but also the precise site of attachment, thereby making it necessary to document the structure and function of the glycoprotein as a whole. Glycan synthesis is nontemplate driven, hence it has greater structural diversity than both proteins and nucleic acids and needs to be considered as a heterogeneous mixture of different chemical structures when isolated from cells or tissues. The difficulty in synthesising complex carbohydrates to decipher structure–function correlates has also been a significant challenge.

Several approaches have been undertaken to characterize cancer-associated glycosylation changes. Mass spectrometry, in combination with various chromatographic methodologies, is one of the most powerful and versatile techniques for the structural analysis of oligosaccharides (Tabares *et al*., 2006). The use of monoclonal antibodies against specific glycan structures (Sell, 1990), as well as studying the binding pattern of lectins to the glycoproteins of healthy compared to cancerous tissue sections (Streets *et al*., 1996) or lectin-based blotting of normal and tumor tissue glycoproteins that have been resolved by 1D- or 2D-PAGE, has also been tried (Osborne and Brooks, 2006). Much has been gleaned on the function of particular cellular glycans by perturbation of their synthesis through the use of glycosylation inhibitors or glycosidases, as well as alteration of the glycosyltransferase expression (Bertozzi and Kiessling, 2001; Moremen, 2002; Patsos *et al*., 2005). Genetic knockouts have provided invaluable insight into the importance of particular glycans in tumor progression (Granovsky *et al*., 2000; Lowe and Marth, 2003), and there is now great impetus to decipher disease-specific glycans so that they are targeted to specifically destroy cancer cells that display them or their presence is used for diagnostic means.

3.1. Biomarkers

Currently, a number of circulating tumor-derived glycoproteins such as CA125, prostate-specific antigen (PSA) and carcinoembryonic antigen (CEA) function as tumor markers; however, they suffer from a lack of specificity. Their expression is not unique to cancer cells as they are often raised in a variety of non-malignant conditions and their levels are not always elevated with cancer (Ugrinska *et al*., 2002).

More recently, detailed analysis of the glycan structures of PSA and CA125 proteins has revealed cancer-specific differences, and may ultimately enable further discrimination between cancer and benign disease (Kui-Wong *et al*., 2003; Tabares *et al*., 2006). The glycan changes *per se* that indicate malignancy are already being employed in diagnosis; the commonly used CA19-9 test for pancreatic cancer measures serum levels of sLe^a and currently represents the most specific test for this cancer (Magnani, 2004). Tumor-specific glycosylation may also be analyzed via lectin staining of histological samples; in fact, much of our initial insight into the changes in glycosylation of cancer cells emerged from the differential binding of lectins between healthy and cancerous tissue (Ching *et al*., 1988; Blonski *et al*., 2005). First discovered in the 1880s, lectins are carbohydrate-binding proteins of nonimmunological origin. They have specific but overlapping affinities for particular glycan structures, thus their binding to normal and tumor glycoproteins provides information on the nature of the changes being observed (for a detailed review, see Sharon and Lis, 2004). One lectin, in particular, *Helix pomatia* agglutinin (HPA), has been extensively used in cancer research as an indicator of poor prognosis in breast, lung, colorectal, gastric, and prostate cancer (Mitchell and Schumacher, 1999). More recently, aberrant glycosylation in cancer cells has been employed for molecular imaging to monitor breast cancer in murine tumor models using peptide conjugates that bind to the underglycosylated form of MUC1 (Moore *et al*., 2004).

3.2. Glycotherapeutics

Although it is evident that glycosylation is altered with cancer, dissecting the role of specific glycans has been extremely challenging. This is because cellular heterogeneity is very common during the evolution of neoplasias, but, in addition, the micro heterogeneity of glycosylation presents further problems in correlating the presence of a specific glycans to the biology of a given tumor. Despite this, there are numerous examples where the altered glycosylation has been targeted for therapeutic purposes, some of which are outlined below.

Interactions between circulating tumor cells and remote vascular endothelium are key early events in cancer metastasis. The interaction of selectins on endothelial cells with the carbohydrate determinants, sLe^x and sLe^a, which are commonly overexpressed on cancer cells constitutes an important step in this process (Kannagi *et al*., 2004). Numerous studies have demonstrated that the expression of these antigens is inversely correlated with patient survival (Nakagoe *et al*., 2002; Saito *et al*., 2003). Modification of tumor cell surface antigens with GlcNAc

analogues has been used to reduce tumor cell adhesion (Woynarowska *et al*., 1996) and small molecule mimetics of sLe^x or sLe^a selectin ligands (Dey and Witczak, 2003; Fuster *et al*., 2003) as well as disaccharide decoys that interfere with selectin–carbohydrate interaction have been considered for the treatment of metastatic diseases. Therapies based on the blocking of tumor-associated glycan formation using inhibitors of glycoprocessing enzymes are another possibility. GnT-V is a key enzyme in the formation of branching of *N*-linked oligosaccharides, which is strongly linked to tumor invasion and metastasis of colon and breast cancers (Fernandes *et al*., 1991) and its expression is an independent predictor of superficial bladder cancer recurrence (Takahashi *et al*., 2006); however, currently no such inhibitors are undergoing therapeutic evaluation. Of the carbohydrate-processing inhibitors presently available, the alkaloid swainsonine, a Golgi α-mannosidase II inhibitor, was the first to have been selected for clinical testing based on its anticancer activity. It also exhibited low toxicity in a number of Phase I cancer trials (Fuster and Esko, 2005). Since increased sialylation is associated with metastatic potential, inhibition of the sialyltransferases has also been considered as a therapeutic strategy (Drinnan *et al*., 2003) and inhibition of ST8aII by synthetic sialic acid precursors has been demonstrated *in vitro,* but, as yet, has not been replicated *in vivo* (Horstkorte *et al*., 2004). Several glycan-based vaccines against tumor glycans are presently undergoing clinical trials, with variable success (for a comprehensive review, see Dube and Bertozzi, 2005; Ouerfelli *et al*., 2005), and multiantigenic vaccines are currently undergoing preclinical evaluation (Ragupathi and Livingston, 2002). The use of lectins for targeted drug delivery has recently been demonstrated *in vivo* (Robinson *et al*., 2004).

4. Renal cancer: a challenging malignancy

Renal cancer represents approximately 3% of adult malignancies with a male-to-female ratio of 2:1, causing 102 000 deaths worldwide (Parkin *et al*., 2005). However, the incidence of this cancer is increasing with a 126% increase in incidence since the 1950s in the United States of America (Pantuck *et al*., 2001) and a reported 18% increase in incidence rate in the last 10 years in females in the United Kingdom, which is one of the largest rises in the incidence rate of any cancer in women (Cancer Research UK Cancer Stats, 2005).

Approximately 70% of renal tumors are histologically classified as conventional or clear cell renal cell carcinomas (RCC), thought to arise from the proximal tubule (Pantuck *et al*., 2001). Tumor stage and grade are the most useful prognostic indicators with a 5-year survival of 16–32% for those with metastatic disease (Pantuck *et al*., 2001). Surgical resection of the primary tumor remains the main treatment option for organ-confined disease; however, owing to the relatively asymptomatic nature of the disease, over 50% of patients present with a locally advanced stage of metastatic disease. Their prognosis remains poor since renal cancer is unresponsive to radiotherapy and also represents one of the most chemoresistant of tumors as none of the commonly used cytotoxic agents provide appreciable activity (Motzer and Russo, 2000). More promising response rates have been reported with biological therapies such as interferon and interleukin-2;

however, these are associated with overall response rates of less than 20% (Zhou and Rubin, 2001).

There is a clear need to identify biomarkers for detecting renal cancer while still resectable, as well a need for novel therapeutic targets. There are currently no tumor markers in routine clinical use for detection, staging, or monitoring of renal cancer evolution and treatment. The identification of diagnostic and prognostic markers is thus an intense area of investigation and the National Cancer Institute (NCI) has identified the development of biomarkers and therapeutic strategies for renal cancer as a priority area for research (National Cancer Institute, 2002).

5. Glycosylation and renal cancer: the quest for biomarkers and therapeutic targets

While in recent years much progress has been made in characterizing and understanding the changes in glycosylation that occur in cancer, much of the effort has been focused around the more common cancer types. By comparison, therefore, relatively little is known regarding glycosylation and renal cancer. Although the studies are somewhat limited, it is clear that renal cancer demonstrates an altered pattern of glycosylation compared to the normal kidney.

The majority of the studies examining glycosylation changes in renal cancer are based around descriptive studies of lectin-staining patterns in the normal kidney compared to that in renal cancer tissues. Lectin histochemistry is a useful tool to detect subtle alterations in glycosylation of cancer tissues, and the lectin-staining properties within individual tissues may have relevance to the neoplastic process, as alterations in cell morphology alone may not indicate the underlying malignant potential and patient prognosis. While some detailed studies have been made, comprehensive studies are few and the numbers of specimens evaluated are often very small (in fact, in some cases only one patient has been examined). In the normal adult kidney, particular sugar residues are strictly confined to various parts of the nephron. For example, the binding of the lectin *Lotus tetragonolobus* agglutinin (LTA), specific for α-1-fucose residues, is limited to the proximal convoluted tubule. Others, such as Soybean agglutinin (SBA), Dolichos biflorus (DBA), and Peanut agglutinin (PNA), that bind to α- or β-linked *N*-acetylgalactosamine (and to a lesser extent, galactose residues), *N*-acetyl-α-D-galactosamine, and β-D-galactose, respectively, are restricted to the distal convoluted tubules and collecting ducts (Holthofer *et al*., 1981).

In the malignant kidney, the pattern of lectin staining is consistently reported to be altered; however, there are numerous discrepancies in the changes that are described, hence little consensus can be drawn. This is likely to reflect technical differences in methodologies and also the inherent heterogeneity that is associated with renal cancers. Studies into renal cancer have variably used fluorochrome-labeled lectins, biotinylated lectins, and horseradish peroxidase–coupled lectins to label frozen or paraffin-embedded tissues as well as single cell suspensions. Such variation in methodology can seriously affect the pattern of lectin staining observed and therefore the correlations that are made (Brooks *et al*., 1996). It is apparent

that further complexity is presented with renal cancer, as there is considerable interpatient heterogeneity reported within studies (Raedler *et al*., 1982; Ulrich *et al*., 1985).

Binding of LTA in RCCs has been variably reported as absent, to approximately 60% positivity and as high as 90% in well-differentiated tumors (Holthofer *et al*., 1983; Hanioka *et al*., 1990; Ulrich *et al*., 1985). Lectins that are normally confined to distal convoluted tubule, such as SBA, show positivity rates of between 37.5 and 95% (Iizumi *et al*., 1986; Hanioka *et al*., 1990) and DBA between 0 and 25% (Holthofer *et al*., 1983; Ulrich *et al*., 1985).

Ongoing work within our own laboratory has sought to identify the renal cancer-associated glycoproteins/glycoforms, where the aim is that they may form the basis of novel biomarkers or therapeutic targets. Using a combination of lectin histochemistry on matched normal kidney and cancer tissue specimens, as well as lectin-based blotting of tissue lysates separated by one-dimensional and two-dimensional polyacrylamide gel electrophoresis (PAGE), the binding pattern of a number of lectins has been examined, several of which exhibited alterations in binding between matched tumor and normal tissues. Examples of both tissue lectin histochemistry and lectin blots of tissue glycoproteins are shown in Figure 1.

(a) SNA binding is present in the glomeruli of the normal kidney, but is largely absent from normal tubular epithelial cells. In contrast, strong membranous staining of tumor cell membranes is apparent.

(b) Lectin blot of normal and tumor tissue proteins resolved by 2D PAGE. A number of glycoproteins are present in the tumor only.

These studies help in establishing the aberrant nature of glycan expression in RCCs, but, by themselves, reveal little about specific changes or the role of glycans in renal cancer. Ulrich and colleagues reported a correlation between tumor grade and LTA positivity (with high-grade tumors showing universal absence of binding),

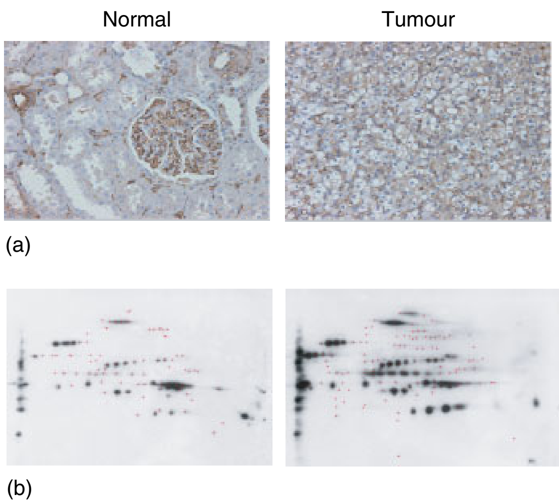


Figure 1 Lectin histochemistry and lectin blotting of matched normal kidney and tumor tissues with *Sambucus nigra* (SNA)

suggesting that well-differentiated low-grade tumors express the carbohydrate components of the tubular cells from which they are supposed to originate with a higher incidence than their more malignant counterparts (Ulrich *et al*., 1985). Like many other tumors, the expression of sLe^x has been correlated with both histological parameters and prognosis in patients with RCC. Immunohistochemistry has shown that tumors with higher levels of expression are associated with higher grade, vascular infiltration, and early relapse (Tozawa *et al*., 2005). One of the molecular mechanisms underlying this overexpression appears to be linked to the switching from oxidative to elevated anaerobic glycolysis (Warburg effect) that tumors often undergo in response to hypoxia. The resultant accumulation of hypoxia inducible transcription factors (HIF) (often also constitutively activated in RCC via loss of the von Hippel Lindau tumor suppressor gene) is thought to lead to the enhancement of glucose transporters and glycosyltransferases related to the sLe^x-determinant expression (Koike *et al*., 2004; Kannagi, 2004).

Galectins are a major class of glycoproteins that bind to ß-galactoside sugars and are involved in cell–cell interactions. Galectins-1 and -3 have been implicated in the progression of various cancers (for a review, see van den Brûle *et al*., 2004). Galectin-3 expression in RCC has been reported as being associated with tumor indolence, being expressed predominantly in low-grade conventional RCCs, indolent chromophobe RCCs, and benign oncocytomas (Young *et al*., 2001). In another study, RCC aggressiveness was related to a decrease in galectin-1 binding sites (and to a lesser extent, galectin-1 expression) and a trend toward decreased galectin-3 binding sites (Francois *et al*., 1999).

Our current understanding of the mechanisms that underlie these changes in surface glycosylation in RCC, and indeed in other cancers, is extremely poor. The most well-characterized enzymes involved in glycan assembly are the *N*-acetylglucosaminyltransferases (GnT), which play a critical role in the antennary number, linkage, and bisecting structure of glycans. They are commonly observed to be overexpressed in many tumors, and, in particular, in GnT-V (Taniguchi *et al*., 2001). In contrast to this, the level of GnT-V activity in RCC has been reported as being almost identical to the normal kidney, and the level of activity of GnT-III and -IV to be considerably reduced (Zhu *et al*., 1997). This finding is further supported by the observation that gamma-glutamyltransferase (GGT) is differentially glycosylated in RCC in comparison to the normal kidney. RCC GGT contained comparatively fewer complex-type bisecting GlcNAc chains, consistent with a reduced expression of GnT-III in RCC (Yoshida *et al*., 1995).

6. Summary and future challenges

Although there has been enormous progress in understanding the mechanisms of cancer development, paralleled advances in cancer treatment have not been forthcoming (Weir *et al*., 2003). Cancer is a highly complex disease, there are many different cancer types, and within each cancer there are different subtypes, which are often associated with different clinical behaviors and outcomes. Earlier detection, as well as the identification of novel therapeutic targets, is essential for the future successful treatment of this disease. This is particularly true for renal cancer, where

early surgical resection of the primary tumor is essential for successful treatment; however, because it is relatively asymptomatic, it is too often diagnosed too late. A major goal of cancer research therefore is to identify diagnostic and prognostic biomarkers that are associated with malignancy from patient tissue, serum, or urine. Protein glycosylation changes dramatically with cancer and glycans are engaged in many aspects of tumor progression; therefore, they represent opportunities for diagnostics and therapeutics.

Changes in glycosylation are apparent in cancer; however, no one change seems to distinctly differentiate normal and malignant tissue, rather each malignancy is characterized by a set of changes in glycan expression. Hence, combinations of changes, a "glycan signature" of the disease, may be most informative. So far in the clinic, altered glycosylation has been harnessed mainly for biomarkers due to the secretion of tumor-derived glycoproteins into the bloodstream or into other bodily fluids, and it is important that this continues and further discoveries are exploited. The current effort is now directed at understanding the role of glycosylation in cancer, and the possibilities of manipulating glycan function to minimize processes such as metastasis and increasing immune response toward cancer cells are currently being explored. Glycoproteomics is a relatively new field and important breakthroughs in glycobiology fueled by rapid technology advances have resulted in a renaissance in the past few years. It is an overwhelming task to elucidate precise mechanistic functions for each type of aberrant glycosylation seen with cancer. However, the potential gain for diagnosis and therapeutics justifies perseverance in this difficult and complex field of research. To advance this burgeoning field, several international collaborative efforts, such as the Consortium for Functional Glycomics, an international initiative funded by the National Institute of General Medical Sciences, have been established (Raman *et al*., 2006) and the progress described so far is likely only to be the very tip of the iceberg.

References

Bertozzi CR and Kiessling LL (2001) Chemical glycobiology. *Science*, **291**, 2357–2364.

Blonski K, Schumacher U, Burkholder I, Edler L, Nikbakth H, Boeters I, Peters A, Kugler C, Horny HP, Langer M, *et al* (2005) Binding of recombinant mistletoe lectin (aviscumine) to resected human adenocarcinoma of the lung. *Anticancer Research*, **25**, 3303–3307.

Brooks SA, Lymboura M, Schumacher U and Leathem AJ (1996) Histochemistry to detect Helix pomatia lectin binding in breast cancer: methodology makes a difference. *Journal of Histochemistry and Cytochemistry*, **44**, 519–524.

van den Brûle F, Califice S and Castronovo V (2004) Expression of galectins in cancer: a critical review. *Glycoconjugate Journal*, **19**, 537–542.

Cancer Research UK Stats Sheets 2001, (2005).

Ching CK, Black R, Helliwell T, Savage A, Barr H and Rhodes JM (1988) Use of lectin histochemistry in pancreatic cancer. *Journal of Clinical Pathology*, **41**, 324–328.

Demetriou M, Nabi IR, Coppolino M, Dedhar S and Dennis JW (1995) Reduced contact-inhibition and substratum adhesion in epithelial cells expressing GlcNAc-transferase V. *The Journal of Cell Biology*, **130**, 383–392.

Dennis JW, Granovsky M and Warren CE (1999a) Glycoprotein glycosylation and cancer progression. *Biochimica et Biophysica Acta*, **1473**, 21–34.

Dennis JW, Granovsky M and Warren CE (1999b) Protein glycosylation in development and disease. *Bioessays*, **21**, 412–421.

Dennis JW, Laferte S, Waghorne C, Breitman ML and Kerbel RS (1987) Beta 1-6 branching of Asn-linked oligosaccharides is directly associated with metastasis. *Science*, **236**, 582–585.
Dey PM and Witczak ZJ (2003) Functionalized S-thio-di- and S-oligosaccharide precursors as templates for novel SLex/a mimetic antimetastatic agents. *Mini Reviews in Medicinal Chemistry*, **3**, 271–280.
Drinnan NB, Halliday J and Ramsdale T (2003) Inhibitors of sialyltransferases: potential roles in tumor growth and metastasis. *Mini Reviews in Medicinal Chemistry*, **3**, 501–517.
Dube DH and Bertozzi CR (2005) Glycans in cancer and inflammation–potential for therapeutics and diagnostics. *Nature Reviews. Drug Discovery*, **4**, 477–488.
Dwek MV and Brooks SA (2004) Harnessing changes in cellular glycosylation in new cancer treatment strategies. *Current Cancer Drug Targets*, **4**, 425–442.
Fernandes B, Sagman U, Auger M, Demetrio M and Dennis JW (1991) Beta 1-6 branched oligosaccharides as a marker of tumor progression in human breast and colon neoplasia. *Cancer Research*, **51**, 718–723.
Fjeldstad K and Kolset SO (2005) Decreasing the metastatic potential in cancers–targeting the heparan sulfate proteoglycans. *Current Drug Targets*, **6**, 665–682.
Francois C, van Velthoven R, De Lathouwer O, Moreno C, Peltier A, Kaltner H, Salmon I, Gabius HJ, Danguy A, Docaestecker C, *et al* (1999) Galectin-1 and galectin-3 binding pattern expression in renal cell carcinomas. *American Journal of Clinical Pathology*, **112**, 194–203.
Fuster MM, Brown JR, Wang L and Esko JD (2003) A disaccharide precursor of sialyl Lewis X inhibits metastatic potential of tumor cells. *Cancer Research*, **63**, 2775–2781.
Fuster MM and Esko JD (2005) The sweet and sour of cancer: glycans as novel therapeutic targets. *Nature Reviews. Cancer*, **5**, 526–542.
Gorelik E, Galili U and Raz A (2001) On the role of cell surface carbohydrates and their binding proteins (lectins) in tumor metastasis. *Cancer Metastasis Reviews*, **20**, 245–277.
Granovsky M, Fata J, Pawling J, Muller WJ, Khokha R and Dennis JW (2000) Suppression of tumor growth and metastasis in Mgat5-deficient mice. *Nature Medicine*, **6**, 306–312.
Hakomori S (2000) Traveling for the glycosphingolipid path. *Glycoconjugate Journal*, **17**, 627–647.
Hanahan D and Weinberg RA (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
Hanioka K, Imai Y, Watanabe M and Ito H (1990) Lectin histochemical studies on renal tumors. *The Kobe Journal of Medical Sciences*, **36**, 1–21.
Hollingsworth MA and Swanson BJ (2004) Mucins in cancer: protection and control of the cell surface. *Nature Reviews. Cancer*, **4**, 45–60.
Holthofer H, Miettinen A, Paasivuo R, Lehto VP, Linder E, Alfthan O and Virtanen I (1983) Cellular origin and differentiation of renal carcinomas. A fluorescence microscopic study with kidney-specific antibodies, antiintermediate filament antibodies, and lectins. *Laboratory Investigation*, **49**, 317–326.
Holthofer H, Virtanen I, Pettersson E, Tornroth T, Alfthan O, Linder E and Miettinen A (1981) Lectins as fluorescence microscopic markers for saccharides in the human kidney. *Laboratory Investigation*, **45**, 391–399.
Horstkorte R, Muhlenhoff M, Reutter W, Nohring S, Zimmermann-Kordmann M and Gerardy-Schahn R (2004) Selective inhibition of polysialyltransferase ST8SiaII by unnatural sialic acids. *Experimental Cell Research*, **298**, 268–274.
Iizumi T, Yazaki T, Kanoh S, Koiso K, Koyama A and Tojo S (1986) Fluorescence study of renal cell carcinoma with antibodies to renal tubular antigens, intermediate filaments, and lectins. *Urologia Internationalis*, **41**, 57–61.
International Agency for Research on Cancer (2003) *World Cancer Report*.
Kannagi R (2004) Molecular mechanism for cancer-associated induction of sialyl Lewis X and sialyl Lewis A expression-The Warburg effect revisited. *Glycoconjugate Journal*, **20**, 353–364.
Kannagi R, Izawa M, Koike T, Miyazaki K and Kimura N (2004) Carbohydrate-mediated cell adhesion in cancer metastasis and angiogenesis. *Cancer Science*, **95**, 377–384.
Kim YJ and Varki A (1997) Perspectives on the significance of altered glycosylation of glycoproteins in cancer. *Glycoconjugate Journal*, **14**, 569–576.

Koike T, Kimura N, Miyazaki K, Yabuta T, Kumamoto K, Takenoshita S, Chen J, Kobayashi M, Hosokawa M, Taniguchi A, *et al* (2004) Hypoxia induces adhesion molecules on cancer cells: A missing link between Warburg effect and induction of selectin-ligand carbohydrates. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 8132–8137.

Kui-Wong N, Easton RL, Panico M, Sutton-Smith M, Morrison JC, Lattanzio FA, Morris HR, Clark GF, Dell A and Patankar MS (2003) Characterization of the oligosaccharides associated with the human ovarian tumor marker CA125. *Journal of Biological Chemistry*, **278**, 28619–28634.

Lowe JB and Marth JD (2003) A genetic approach to Mammalian glycan function. *Annual Review of Biochemistry*, **72**, 643–691.

Magnani JL (2004) The discovery, biology, and drug development of sialyl Lea and sialyl Lex. *Archives of Biochemistry and Biophysics*, **426**, 122–131.

Meezan E, Wu HC, Black PH and Robbins PW (1969) Comparative studies on the carbohydrate-containing membrane components of normal and virus-transformed mouse fibroblasts. II. Separation of glycoproteins and glycopeptides by sephadex chromatography. *Biochemistry*, **8**, 2518–2524.

Mitchell BS and Schumacher U (1999) The use of the lectin Helix pomatia agglutinin (HPA) as a prognostic indicator and as a tool in cancer research. *Histology and Histopathology*, **14**, 217–226.

Moore A, Medarova Z, Potthast A and Dai G (2004) In vivo targeting of underglycosylated MUC-1 tumor antigen using a multimodal imaging probe. *Cancer Research*, **64**, 1821–1827.

Moremen KW (2002) Golgi alpha-mannosidase II deficiency in vertebrate systems: implications for asparagine-linked oligosaccharide processing in mammals. *Biochimica et Biophysica Acta*, **19**, 225–235.

Motzer RJ and Russo P (2000) Systemic therapy for renal cell carcinoma. *The Journal of Urology*, **163**, 408–417.

Nakagoe T, Fukushima K, Tanaka K, Sawai T, Tsuji T, Jibiki M, Nanashima A, Yamaguchi H, Yasutake T, Ayabe H, *et al* (2002) Evaluation of sialyl Lewis(a), sialyl Lewis(x), and sialyl Tn antigens expression levels as predictors of recurrence after curative surgery in node-negative colorectal cancer patients. *Journal of Experimental and Clinical Cancer Research*, **21**, 107–113.

National Cancer Institute (2002) *Report of the kidney/bladder cancers progress review group*.

Orntoft TF and Vestergaard EM (1999) Clinical aspects of altered glycosylation of glycoproteins in cancer. *Electrophoresis*, **20**, 362–371.

Osborne C and Brooks SA (2006) SDS-PAGE and Western blotting to detect proteins and glycoproteins of interest in breast cancer research. *Methods in Molecular Medicine*, **120**, 217–229.

Ouerfelli O, Warren JD, Wilson RM and Danishefsky SJ (2005) Synthetic carbohydrate-based antitumor vaccines: challenges and opportunities. *Expert Review of Vaccines*, **4**, 677–685.

Pantuck AJ, Zisman A and Belldegrun AS (2001) The changing natural history of renal cell carcinoma. *The Journal of Urology*, **166**, 1611–1623.

Parkin DM, Bray F, Ferlay J and Pisani P (2005) Global cancer statistics, 2002. *CA: A Cancer Journal for Clinicians*, **55**, 74–108.

Patsos G, Hebbe-Viton V, San MR, Paraskeva C, Gallagher T and Corfield A (2005) Action of a library of O-glycosylation inhibitors on the growth of human colorectal cancer cells in culture. *Biochemical Society Transactions*, **33**, 721–723.

Raedler A, Boehle A, Otto U and Raedler E (1982) Differences of glycoconjugates exposed on hypernephroma and normal kidney cells. *The Journal of Urology*, **128**, 1109–1113.

Ragupathi G and Livingston P (2002) The case for polyvalent cancer vaccines that induce antibodies. *Expert Review of Vaccines*, **1**, 193–206.

Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S and Sasisekharan R (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82R–90R, [EPUB ahead of print].

Robinson MA, Charlton ST, Garnier P, Wang XT, Davis SS, Perkins AC, Frier M, Duncan R, Savage TJ, Wyatt DA, *et al* (2004) LEAPT: lectin-directed enzyme-activated prodrug

therapy. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14527–14532.
Saito K, Fujii Y, Kawakami S, Hayashi T, Arisawa C, Koga F, Kageyama Y and Kihara K (2003) Increased expression of sialyl-Lewis A correlates with poor survival in upper urinary tract urothelial cancer patients. *Anticancer Research*, **23**, 3441–3446.
Sell S (1990) Cancer-associated carbohydrates identified by monoclonal antibodies. *Human Pathology*, **21**, 1003–1019.
Sharon N and Lis H (2004) History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology*, **14**, 53R–62R.
Shriver Z, Raguram S and Sasisekharan R (2004) Glycomics: a pathway to a class of new and improved therapeutics. *Nature Reviews. Drug Discovery*, **3**, 863–873.
Streets AJ, Brooks SA, Dwek MV and Leathem AJ (1996) Identification, purification and analysis of a 55 kDa lectin binding glycoprotein present in breast cancer tissue. *Clinical Chimica Acta*, **54**, 47–61.
Tabares G, Radcliffe CM, Barrabes S, Ramirez M, Aleixandre RN, Hoesel W, Dwek RA, Rudd PM, Peracaula R and de Llorens R (2006) Different glycan structures in prostate-specific antigen from prostate cancer sera in relation to seminal plasma PSA. *Glycobiology*, **16**, 132–145.
Takahashi T, Hagisawa S, Yoshikawa K, Tezuka F, Kaku M and Ohyama C (2006) Predictive value of N-acetylglucosaminyltransferase-V for superficial bladder cancer recurrence. *The Journal of Urology*, **175**, 90–93.
Taniguchi N, Ihara S, Saito T, Miyoshi E, Ikeda Y and Honke K (2001) Implication of GnT-V in cancer metastasis: a glycomic approach for identification of a target protein and its unique function as an angiogenic cofactor. *Glycoconjugate Journal*, **18**, 859–865.
Taylor-Papadimitriou J, Burchell JM, Plunkett T, Graham R, Correa I, Miles D and Smith M (2002) MUC1 and the immunobiology of cancer. *Journal of Mammary Gland Biology and Neoplasia*, **7**, 209–221.
Tozawa K, Okamoto T, Kawai N, Hashimoto Y, Hayashi Y and Kohri K (2005) Positive correlation between sialyl Lewis X expression and pathologic findings in renal cell carcinoma. *Kidney International*, **67**, 1391–1396.
Ugrinska A, Bombardieri E, Stokkel MP, Crippa F and Pauwels EK (2002) Circulating tumor markers and nuclear medicine imaging modalities: breast, prostate and ovarian cancer. *Quarterly Journal of Nuclear Medicine and Molecular Imaging*, **46**, 88–104.
Ulrich W, Horvat R and Krisch K (1985) Lectin histochemistry of kidney tumors and its pathomorphological relevance. *Histopathology*, **9**, 1037–1050.
Weir HK, Thun MJ, Hankey BF, Ries LA, Howe HL, Wingo PA, Jemal A, Ward E, Anderson RN and Edwards BK (2003) Annual report to the nation on the status of cancer, 1975–2000, featuring the uses of surveillance data for cancer prevention and control. *Journal of the National Cancer Institute*, **95**, 1276–1299.
Woynarowska B, Dimitroff CJ, Sharma M, Matta KL and Bernacki RJ (1996) Inhibition of human HT-29 colon carcinoma cell adhesion by a 4-fluoro-glucosamine analogue. *Glycoconjugate Journal*, **13**, 663–674.
Yoshida K, Sumi S, Honda M, Hosoya Y, Yano M, Arai K and Ueda Y (1995) Serial lectin affinity chromatography demonstrates altered asparagine-linked sugar chain structures of gamma-glutamyltransferase in human renal cell carcinoma. *Journal of Chromatography. B, Biomedical Applications*, **672**, 45–51.
Young AN, Amin MB, Moreno CS, Lim SD, Cohen C, Petros JA, Marshall FF and Neish AS (2001) Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers. *American Journal of Pathology*, **158**, 1639–1651.
Zhou M and Rubin MA (2001) Molecular markers for renal cell carcinoma: impact on diagnosis and treatment. *Seminars in Urologic Oncology*, **19**, 80–87.
Zhu TY, Chen HL, Gu JX, Zhang YF, Zhang YK and Zhang RA (1997) Changes in N-acetylglucosaminyltransferase III, IV and V in renal cell carcinoma. *Journal of Cancer Research and Clinical Oncology*, **123**, 296–299.

Short Specialist Review

Posttranslational cleavage of cell-surface receptors

Martin Stacey, John Davies, Siamon Gordon and Hsi-Hsien Lin
Sir William Dunn School of Pathology, Oxford, UK

1. Introduction

In eukaryotic cells, nascent polypeptides generated by ribosomes rarely represent the true form of mature proteins. Instead, many newly synthesized proteins undergo posttranslational modifications that may mediate either subtle or profound effects on the function of the protein. Once translated, peptides may be altered by the covalent modification of either side chains or terminal amino acids through a variety of enzymatic processes. These include phosphorylation, deamination, hydroxylation, ubiquitination, prenylation, glycosylation, acylation, and methylation. Peptide bond formation may even occur between the side groups of different peptides via an isopeptide bond. Isopeptide formations are mediated by transglutaminases that catalyze the cross-linking of different proteins via ε-(γ-glutamyl)lysine bonds. Proteolytic cleavage differs from the above mechanisms of posttranslational modification in that it is irreversible. Site-specific proteolysis plays an important role in a diverse range of biological processes such as the proenzyme and hormone activation, determination of cell fate, release of cell-associated growth factors, protein trafficking, tissue remodeling, and apoptosis. In all, posttranslational modifications, whether they involve covalent attachment or deletion, or proteolytic processing, serve to modify the function of nascent proteins. Protein function may be regulated through the activation/inhibition of enzymatic or signaling activities, via the structural stabilization against mechanical stress or protease degradation or by altering trafficking/targeting. In this review, we focus on the proteolytic cleavage of three disparate families of cell-surface receptors. Notch receptors, protease-activated receptors (PARs), and G-protein receptor proteolytic site (GPS) containing receptors all exemplify how posttranslational modification by proteolysis can greatly affect protein function.

2. The proteolytic cleavage of Notch

The Notch receptor is crucial in the development of virtually all metazoans, regulating the temporal and spatial cell-fate decisions of multipotent precursor cells. Notch

is a multiple-domain, single-span membrane protein that exemplifies how posttranslational proteolytic cleavage can play a central role in the maturation, activation, and regulation of protein function. Notch is processed at five distinct stages (Figure 1). The first proteolytic processing of the precursor protein occurs in the endoplasmic recticum (ER) by a signal peptidase, which removes the signal peptide from the N-terminus. The second occurs in the Golgi by a furin-like convertase (Logeat *et al*., 1998) (a Ca^{2+}-dependent serine endoprotease, belonging to the subtilisin/kexin-like family), which cleaves at an RXR/KR dibasic recognition site about 70 amino acids N-terminal to the membrane region. Noncovalent reassociation after processing, results in the expression of a heterodimer at the cell surface, which is essential for the subsequent signaling via CSL (CBF1, Suppressor of Hairless, Lag-1) family of transcription factors (Jarriault *et al*., 1998). Recently, in human Notch, low levels of unprocessed cell-surface molecules have been shown to signal by a CSL-independent mechanism, demonstrating that proteolytic cleavage can have a profound effect by directing the signaling pathway chosen by a cell-surface receptor (Bush *et al*., 2001). The third proteolytic event of Notch occurs upon binding to its ligands, Delta, Serrate, and Jagged. Evidence suggests that the endocytosis of the ligand–receptor complex by the ligand-bearing cell results in a conformational change in the Notch protein. This leads to removal of the ectodomain close to the membrane via an ADAM/TACE protease (a disintegrin and metalloprotease/TNF-α converting enzyme), thus preventing further interactions with ligand-bearing cells (Brou *et al*., 2000). Once the ectodomain has been cleaved, the mechanism of Notch signal transduction differs from that of other proteins relying on phosphorylation or secondary messenger generation in that intramembrane cleavage of Notch is performed by a proteolytic complex termed γ-secretase (Kimberly and Wolfe, 2003). The complex comprises Presenilin, an eight span–membrane protein thought to be an aspartyl protease, Nicastrin a single-span membrane, and other putative proteins. The aspartyl protease activity of Presenilin releases the intracellular domain of Notch, which is then translocated into the nucleus where it interacts with CSL proteins. This either displaces repressing histone deacetylases or recruits transcriptional activators, leading to the activation or suppression of target genes. Finally, Notch signaling is downregulated when the intracellular domain is phosphorylated by a Sel-10 containing complex, resulting in subsequent ubiquitination and targeting for degradation by the proteasome (Hubbard *et al*., 1997; Oberg *et al*., 2001). Similar mechanisms involving regulated intramembrane proteolysis have now been proposed for a number of other receptors including the amyloid precursor protein, the adhesion molecules CD44 and E-Cadherin, a tyrosine kinase receptor ErbB4, LRP, a member of the LDL receptor family, and neurotrophin (Okamoto *et al*., 2001; Kanning *et al*., 2003; May *et al*., 2002; Ni *et al*., 2001; Roncarati *et al*., 2002). These data show that intramembrane cleavage is not unique to Notch but instead is a more widespread signaling phenomenon.

3. Protease-activated receptors

The protease-activated receptors (PAR1–4) are G-protein coupled receptors (GPCRs) that have been implicated in a range of physiological roles including

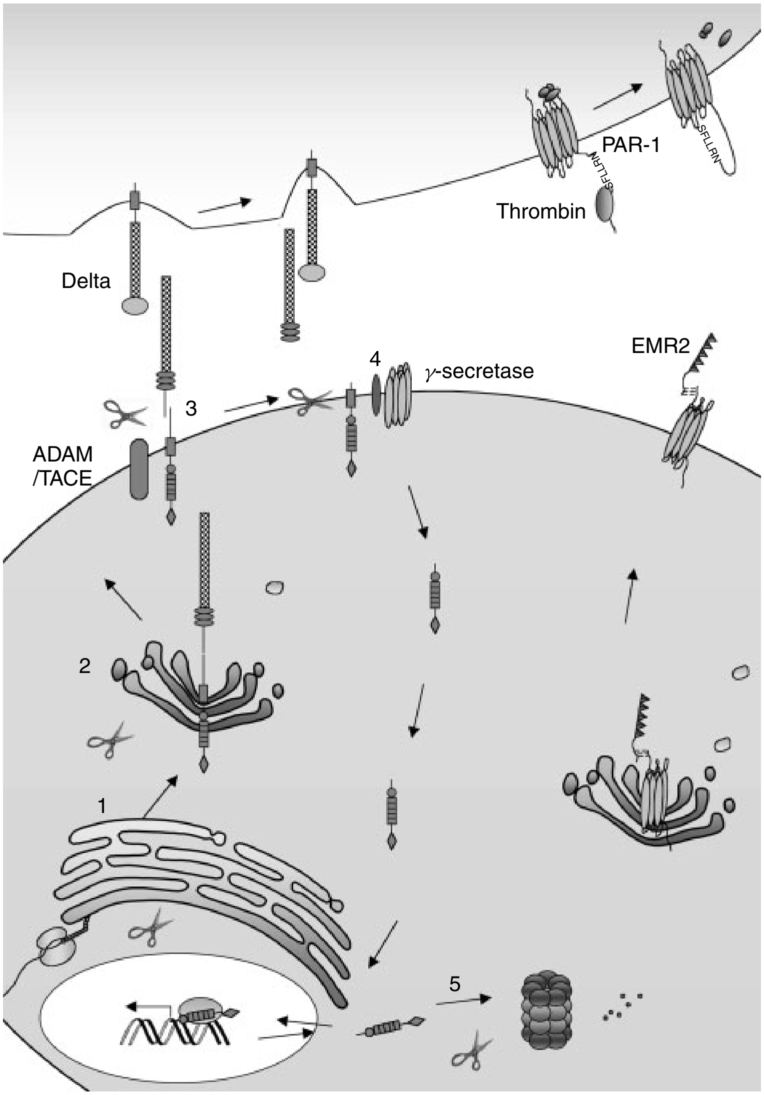


Figure 1 Notch, PAR, and GPS-protein processing. Notch-1: 1. Within the ER, the first of five proteolytic events occurs when the signal peptide is removed by a signal peptidase. 2. Further processing is mediated by a furin-like convertase in the golgi. After cleavage Notch reassociates and is expressed on the cell surface as a heterodimer. 3. The interaction of Delta or other ligands with Notch initiates endocytosis, which in turn induces conformational changes in Notch allowing shedding by an ADAM/TACE protease. 4. A γ-secretase protease complex then cleaves Notch within its membrane domain releasing its intracellular domain that is translocated into the nucleus where it interacts with CSL proteins. 5. Finally, Notch signaling is downregulated when the intracellular domain is phosphorylated by a Sel-10 containing complex, resulting in subsequent ubiquitination and targeting for degradation by the proteasome. PAR-1: After translocation and processing of its signal peptide in the ER, PAR-1 is expressed as a multispan receptor on the cell surface. Cleavage by the serine protease, Thrombin, reveals a previously masked tethered ligand that subsequently interacts with the second extracellular loop, activating the PAR receptor. GPS receptors: EMR2 a leukocyte-restricted receptor is shown as a paradigm for GPS receptors. The protein is initially targeted to the ER where its signal peptide is removed and an autocatalytic processing event occurs. The receptor then reassociates and is transported to the cell surface where it interacts with its ligands as a functional heterodimer. Scissors denote stages at which proteolytic cleavage occurs

platelet activation, inflammation, immune cell function, and embryonic development. However, unlike normal GPCRs, which bind their ligands in a reversible fashion, the PAR receptors are irreversibly activated by specific proteases. After the proteolytic cleavage of the N-terminus, a previously masked peptide sequence acts as a tethered ligand for the receptor causing subsequent signaling.

The mechanism of PAR activation, which was initially demonstrated for PAR-1 while studying platelet activation, now appears to represent a general paradigm for the other family members (Figure 1) (Vu *et al*., 1991). During the search for the human thrombin receptor, it was found that thrombin, a serine protease binds to the ectodomain of PAR-1 and cleaves the N-terminus between Arg41 and Ser42. This causes the unmasking of the peptide sequence SFLLRN that was previously sterically hindered by the proceeding N-terminal residues, thereby allowing it to act as a ligand for PAR-1. Synthetic peptides corresponding to the new N-terminus also act as agonists but are independent of thrombin cleavage. These peptides have proved invaluable in dissecting the crucial residues involved in receptor ligation and have demonstrated that protonation of the new N-terminal serine and the presence of phenylalanine, arginine, and asparagine residues are essential for the interaction with a conserved region within the second extracellular loop of the receptor (Scarborough *et al*., 1992). The binding of synthetic agonists or the N-terminus of postcleaved PAR-1 is thought to cause a conformational change allowing the receptor to interact with heterotrimeric G-proteins. The resulting activation of phospholipase Cβ and production of IP_3 and DAG in turn leads to increases in cytosolic Ca^{2+} levels. This is followed by integrin activation, shape change, release of dense and α-granules, and platelet–platelet aggregation, all of which are associated with platelet activation and the formation of thrombi.

As receptor cleavage effectively results in permanent ligand engagement and receptor activation, it follows that the extent of the cellular response is determined by the termination of PAR signaling. Although the exact mechanisms of desensitization and downregulation of PARs are not well characterized, it is presumed they are similar to other GPCRs. Phosphorylation by GPCR kinases and subsequent interactions with β-arrestins, which have been shown to play an important role in the desensitization of many GPCRs have also been shown to desensitize PAR-1. Receptor internalization and degradation via clathrin-coated vesicles, AP_2 interactions and lysosomal sorting also appear to be crucial in the downregulation of PAR-1 activity.

Although the general mechanism of PAR-1 activation is similar to that of PAR-2, -3, and -4, significant differences have been demonstrated. For example, PAR-2 is the only member of this family that is not cleaved by thrombin. Instead, it is cleaved by a number of trypsin-like serine proteases such as trypsin, neutrophil proteinase 3, factor VIIa/Xa, mast-cell tryptase and membrane-tethered serine protease-1 (Cottrell *et al*., 2003). In addition, PAR-3 is not activated by soluble synthetic peptides. These observations suggest that cleavage-mediated signaling in vivo is likely to be much more complex than initially envisaged, involving many different proteases and signaling mechanisms.

4. GPS-containing receptors

In recent years, a novel proteolytic motif known as the GPCR proteolytic site (GPS) has been identified in over 40 human cell-surface receptors (SMART, **http://smart.embl-heidelberg.de**). The majority of the GPS-containing receptors show homology to class B GPCRs within their transmembrane moieties. However, they also possess large N-terminal extracellular domains containing multiple adhesion motifs, coupled to a GPS-containing region (Fredriksson *et al*., 2003). This has led to the notion that these receptors might couple extracellular adhesion to intracellular G-protein signaling. The receptors, classified as LNB-TM7 receptors include the largest known cell-surface receptor VLG-1 (an impressive 6307 amino acids) (McMillan *et al*., 2002), the receptor for the black widow spider toxin, latrophilin (Lelianova *et al*., 1997), and the leukocyte restricted EGF-TM7 receptors (Stacey *et al*., 2000). Many of these LNB-TM7 receptors have now been demonstrated to undergo early proteolytic processing during protein maturation (Figure 1). Evidence now suggests that the cleavage occurs in the ER and requires an intact GPS region of ∼200 amino acids. The exact structure of this GPS region has not yet been elucidated; however, four conserved cysteines and invariant glycine and tryptophan residues appear to be crucial for cleavage to occur at a conserved membrane proximal site (HX↓S/T). Curiously, the extracellular subunit generated by cleavage tightly reassociates to as few as eight amino acids of the membrane subunit in a noncovalent fashion. It is then expressed on the cell surface as a heterodimer. The fact that the GPS domain is conserved in both vertebrates and invertebrates suggests that this proteolytic event must have an important functional role in receptor activity and is likely to be mediated by a common mechanism. Much work has been performed in the search for the protease(s) that cleave the GPS site but with no success. However, recent work by Lin *et al*. (2004) using purified protein of a GPS receptor, in a cell-free system, has clearly established that the cleavage mechanism is in fact an autocatalytic process occurring via an intramolecular cleavage event. As the GPS motif is highly conserved, it is likely that the autocatalytic event will be a paradigm for the other receptors. So what is the functional significance of the autocatalytic proteolysis? A clue to the potential function has come from studies of the multiple membrane-spanning protein responsible for autosomal-dominant polycystic kidney disease, PKD-1/polycystin-1. Polycystin-1 has been found to mediate Ca^{2+} influx by its interaction with the membrane protein polycystin-2. However, although Polycystin-1 GPS-mutants found in PKD patients have been shown to express normal levels of surface receptor, they do not undergo receptor cleavage or respond in functional assays (Qian *et al*., 2002), indicating that cleavage is essential for receptor signaling. As yet, the functional role of most of the other GPS-containing proteins is not known. However, it is likely that posttranslational cleavage events will play a crucial role in their biological activities.

Although Notch, PAR, and GPS receptor families only represent a small subset of eukaryotic genes, they demonstrate how proteolysis can have profound effects on

the function of proteins. With the renewed focus on proteins in the postgenomic era, there is no doubt that many more protein-cleavage events will soon be discovered, revealing many more intriguing biological mechanisms.

References

Brou C, Logeat F, Gupta N, Bessia C, LeBail O, Doedens JR, Cumano A, Roux P, Black RA and Israel A (2000) A novel proteolytic cleavage involved in Notch signaling: the role of the disintegrin-metalloprotease TACE. *Molecular Cell*, **5**, 207–216.

Bush G, diSibio G, Miyamoto A, Denault JB, Leduc R and Weinmaster G (2001) Ligand-induced signaling in the absence of furin processing of Notch-1. *Developmental Biology*, **229**, 494–502.

Cottrell GS, Amadesi S, Schmidlin F and Bunnett N (2003) Protease-activated receptor 2: activation, signalling and function. *Biochemical Society Transactions*, **31**, 1191–1197.

Fredriksson R, Gloriam DE, Hoglund PJ, Lagerstrom MC and Schioth HB (2003) There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini. *Biochemical and Biophysical Research Communications*, **301**, 725–734.

Hubbard EJ, Wu G, Kitajewski J and Greenwald I (1997) *sel-10* a negative regulator of *lin-12* activity in *Caenorhabditis elegans*, encodes a member of the CDC4 family of proteins. *Genes & Development*, **11**, 3182–3193.

Jarriault S, Le Bail O, Hirsinger E, Pourquie O, Logeat F, Strong CF, Brou C, Seidah NG and Isra lA (1998) Delta-1 activation of Notch-1 signaling results in HES-1 transactivation. *Molecular Cell Biology*, **18**, 7423–7431.

Kanning KC, Hudson M, Amieux PS, Wiley JC, Bothwell M and Schecterson LC (2003) Proteolytic processing of the p75 neurotrophin receptor and two homologs generates C-terminal fragments with signaling capability. *The Journal of Neuroscience*, **23**, 5425–5436.

Kimberly WT and Wolfe MS (2003) Identity and function of gamma-secretase. *Journal of Neuroscience Research*, **74**, 353–360.

Lelianova VG, Davletov BA, Sterling A, Rahman MA, Grishin EV, Totty NF and Ushkaryov YA (1997) Alpha-latrotoxin receptor, latrophilin, is a novel member of theRT secretin family of G protein-coupled receptors. *The Journal of Biological Chemistry*, **272**, 21504–21508.

Lin HH, Chang GW, Davies JQ, Stacey M, Harris J and Gordon S (2004) Autocatalytic cleavage of the EMR2 receptor occurs at a conserved G protein-coupled receptor proteolytic site motif. *The Journal of Biological, Chemistry* **279**(30), 31823–31832.

Logeat F, Bessia C, Brou C, LeBail O, Jarriault S, Seidah NG and Israel A (1998) The Notch1 receptor is cleaved constitutively by a furin-like convertase. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8108–8112.

May P, Reddy YK and Herz J (2002) Proteolytic processing of low density lipoprotein receptor-related protein mediates regulated release of its intracellular domain. *The Journal of Biological Chemistry*, **277**, 18736–18743.

McMillan DR, Kayes-Wandover KM, Richardson JA and White PC (2002) Very large G protein-coupled receptor-1, the largest known cell surface protein, is highly expressed in the developing central nervous system. *The Journal of biological chemistry*, **277**, 785–792.

Ni CY, Murphy MP, Golde TE and Carpenter G (2001) Gamma -Secretase cleavage and nuclear localization of ErbB-4 receptor tyrosine kinase. *Science*, **294**, 2179–2181.

Oberg C, Li J, Pauley A, Wolf E, Gurney M and Lendahl U (2001) The Notch intracellular domain is ubiquitinated and negatively regulated by the mammalian Sel-10 homolog. *The Journal of biological chemistry*, **276**, 35847–35853.

Okamoto I, Kawano Y, Murakami D, Sasayama T, Araki N, Miki T, Wong AJ and Saya H (2001) Proteolytic release of CD44 intracellular domain and its role in the CD44 signaling pathway. *The Journal of cell biology*, **155**, 755–762.

Qian F, Boletta A, Bhunia AK, Xu H, Liu L, Ahrabi AK, Watnick TJ, Zhou F and Germino GG (2002) Cleavage of polycystin-1 requires the receptor for egg jelly domain and is disrupted by human autosomal-dominant polycystic kidney disease 1-associated mutations. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16981–16986.

Roncarati R, Sestan N, Scheinfeld MH, Berechid BE, Lopez PA, Meucci O, McGlade JC, Rakic P and D'Adamio L (2002) The gamma-secretase-generated intracellular domain of beta-amyloid precursor protein binds numb and inhibits Notch signaling. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7102–7107.

Scarborough RM, Naughton MA, Teng W, Hung DT, Rose J, Vu TK, Wheaton VI, Turck CW and Coughlin SR (1992) Tethered ligand agonist peptides. Structural requirements for thrombin receptor activation reveal mechanism of proteolytic unmasking of agonist function. *The Journal of Biological Chemistry*, **267**, 13146–13149.

Stacey M, Lin HH, Gordon S and McKnight AJ (2000) LNB-TM7, a group of seven-transmembrane proteins related to family-B G-protein-coupled receptors. *Trends in Biochemical Sciences*, **25**, 284–289.

Vu TK, Hung DT, Wheaton VI and Coughlin SR (1991) Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation. *Cell*, **64**, 1057–1068.

Short Specialist Review

S-nitrosylation and thiolation

Yunfei Huang and Solomon H. Snyder
Johns Hopkins University, Baltimore, MD, USA

1. Introduction

The redox status of cysteine residues in proteins is essential to maintain the structure and function of numerous enzymes, receptors, and transcriptional factors. It has long been believed that the free sulfhydryl group (−SH) of a cysteine residue reacts easily with reactive oxygen/nitrogen species, resulting in irreversible oxidation, therefore serving as a marker of oxidative stress inside cells. However, recent evidence suggests that reversible posttranslational modifications of reactive cysteines, such as S-nitrosylation and thiolation, play a greater role in cell signaling. S-nitrosylation, in particular, regulates a multitude of physiological processes.

2. S-nitrosylation

Three terms, nitrosation, nitrosylation, and nitration, are frequently used to describe NO-mediated protein modifications. The first two sometimes employed interchangeably. Nitrosation designates the addition of a nitroso group (NO diatomic group), while nitrosylation refers to the addition of a nitrosyl group (NO radical). Thus, covalent attachment of a nitroso group (NO diatomic group) to another chemical group is called nitrosation and the direct incorporation of a nitrosyl group (NO radical) to a metal by a coordinating bond is called nitrosylation. S-nitrosylation, equivalent to S-nitrosation in some literatures, specifically refers to the formation of an S−NO bond between the NO moiety and a sulfur atom in a protein cysteine residue. Nitration is an irreversible modification of the covalent attachment of triatomic group ($-NO_2$) to proteins by peroxynitrite.

S-nitrosylation is a naturally occurring physiological modification that mediates nitric oxide (NO) signaling. Endogenous *S*-nitrosothiols (SNO) can be detected in many organs under physiological conditions (Bryan *et al*., 2004; Jaffrey *et al*., 2001). The production of SNO is directly linked to endogenous nitric oxide synthase (NOS) activities. S-nitrosylation modulates biological activities of cellular key regulators via altering enzymatic activities, and protein−protein or protein−DNA interactions (Gu *et al*., 2002; Matsushita *et al*., 2003; Kim *et al*., 2002). Aberrant S-nitrosylation is associated with selected disease states (Foster *et al*., 2003). In contrast to reversible modifications driven by enzymes, such

as protein phosphorylation, S-nitrosylation may involve fewer constrains, as it involves a simple chemical reaction between thiol groups and a free diffusible radical NO or its derivatives. Several reaction processes have been proposed in the biosynthesis of endogenous SNO (Table 1) (Foster *et al*., 2003).

Signal transduction pathways typically possess temporal and spatial specificity. Evidence for such specificity in S-nitrosylation, although still fragmentary, is beginning to accumulate. For instance, of more than 50 cysteines in ryanodine receptor, only one is nitrosylated. Jaffrey *et al*. (2001) identified a subset of proteins in brain that is S-nitrosylated. There may be several possible scenarios for such specificity. First, free cysteines in the lipophilic pocket of proteins and in the lipid membrane are uniquely susceptible to nitrosylation. NO appears to react with O_2 to form more reactive NO-derived species including NO_2 and N_2O_3 that S-nitrosylate cysteine thiols in target proteins. NO and O_2 prefer hydrophobic environments, in which the rate of formation of NO-derived species is about 300 times faster than in a liquid phase. Second, NO selectively nitrosylates cysteines within an acid–base motif. These charged amino acids in the consensus motif stabilize the reactive thiolated anion, facilitating nucleophilic attack. Third, subcellular localization and calcium-dependent activation of NOS contribute to the selectivity of S-nitrosylation. NO is a free diffusible gas that cannot be stored so that specificity of NO signaling is largely determined by the intracellular compartmentalization of NOS (Bredt, 2003). Thus, the NMDA receptor NR2a and a brain-enriched small G-protein Dexras1, recruited to the postsynaptic protein complex with nNOS by postsynaptic density protein 95 (PSD95) and CAPON respectively, are S-nitrosylated as they are brought into proximity with NO sources (Lane *et al*., 2001; Jaffrey *et al*., 1998; Fang *et al*., 2000).

S-nitrosylation is dynamically regulated in vivo. Steady state levels of S-nitrosylation are determined by the rates of nitrosylation and denitrosylation. S-nitrosylation requires NOS activity, mediated either by eNOS/nNOS activated by Ca^{2+}, or by iNOS induced by cytokines. S-nitrosylation is diminished when NOS activities are depleted genetically or chemically (Bryan *et al*., 2004; Jaffrey *et al*., 2001), suggesting that boosted NO production is required for S-nitrosylation under physiological conditions. Denitrosylation is a process that removes NO from protein thiols. Acute inhibition of NOS activity leads to rapid turnover of *S*-nitrosyl groups,

Table 1 Mechanisms of S-nitrosylation

Cofactors	Chemical reactions	Examples
O_2	$2NO + O_2 \rightarrow 2NO_2$	Ryanodine receptor
Lipid membrane or protein	$NO_2 + NO \rightarrow N_2O_3 \rightarrow {}^{+}NO \ldots NO_2^-$	
Hydrophobic pockets	${}^{+}NO\ NO_2^- + RSH \rightarrow RSNO + NO_2$	
Motif (K/R/H/D/E-C-D/E)	$RSH + NO \rightarrow RSNO$	Ras
Metals (iron, copper)	$NO + M^{n+} \rightarrow M{-}NO$	Hemoglobin β-chain
	$M{-}NO + RSH \rightarrow RSNO + H^+ + M^{n-1}$	
RSNO	$RSNO + R'SH \rightarrow RSH + R'SNO$	Anion exchange protein 1
Superoxide when $[NO] >> [O_2^-]$	$NO + O_2^- \rightarrow ONOO^- \rightarrow 2ONO{-}NO$	*S*-nitrosothiols

RSNO: Protein *S*-nitrosothiols; M: Transient metals.

suggesting that denitrosylation is constitutively active (Bryan *et al*., 2004). Multiple pathways have been found to modulate denitrosylation. Protein S-nitrosylation can be reversed by transnitrosation, a process by which an NO equivalent is transferred from one thiol to other thiols (Foster *et al*., 2003). It can also be decomposed by endogenous reductants such as ascorbate, or by metals such as copper (Foster *et al*., 2003). But the view of denitrosylation as a simple chemical reaction is probably incomplete. Recent studies have found that several enzymes regulate the catabolism of protein SNO. The *S*-nitrosoglutathione reductase (GSNOR), previously named as glutathione-dependent formaldehyde dehydrogenase (GDFDH) or alcohol dehydrogenase III (ADH III), selectively degrades *S*-nitrosoglutathione (GSNO) (Liu *et al*., 2004). Deficiency of GSNOR leads to accumulation of S-nitrosylated proteins and GSNO, suggesting that GSNO, an endogenous NO carrier, mediates the turnover of protein SNO driven by GSNOR (Figure 1). Additionally, cellular SNO is also maintained by the thioredoxin/thioredoxin reductase system (Haendeler *et al*., 2002). Interestingly, Cu/Zn SOD, a superoxide dismutase that catabolizes O^{2-}, also releases NO from GSNO (Johnson *et al*., 2001). As GSNO participates in the metabolism of protein SNO, it should come as no surprise that ceruloplasmin and γ-glutamyl transpeptidase (γ-GT) (Foster *et al*., 2003) that are known to regulate the production and breakdown of GSNO might also regulate the metabolism of protein SNO.

Selective catabolism of protein SNO is of biologically importance either to sustain or switch off NO signaling. Compartmentalization of protein SNO and cellular reductive powers might contribute to this selectivity. S–NO bonds are labile in reductive environment. Therefore, protein SNO in subcellular organelles, including vesicles and mitochondria, in which protein SNO could be shielded from

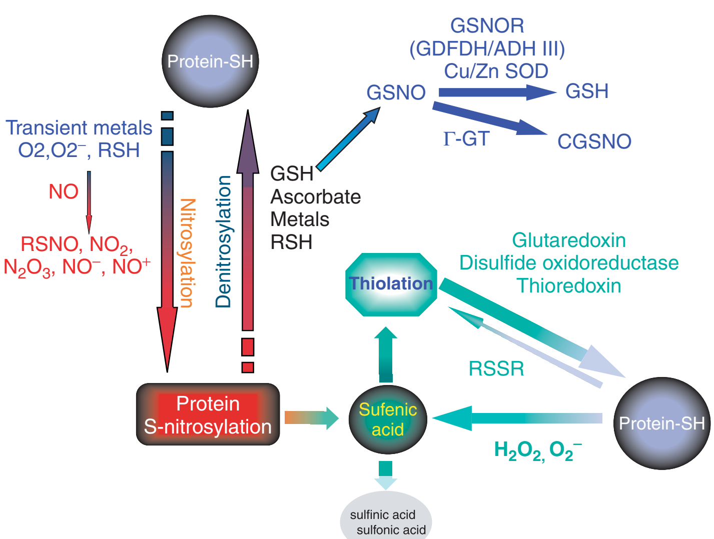


Figure 1 A schematic diagram illustrating the protein thiol modifications mediated by direct chemical or enzyme-assisted reactions in cells

such cellular reductive agents as ascorbate and free reductive thiols in cytosol (1 mM glutathione or GSH), seem to be relatively resistant to denitrosylation. This is exemplified by the fact that translocation of the S-nitrosylated caspase 3 from mitochondria to the reductive cytosol results in its denitrosylation and activation during apoptosis (Mannick *et al*., 2001).

It has been over a decade since NO was discovered as an endogenous signaling molecule generated by NOS. NO signaling was first thought to derive solely from activation of soluble guanylyl cyclase (sGC) and involves binding of NO to the heme moiety in sGC to produce cGMP that can be readily detected. Studies on S-nitrosylation have been limited because of the lack of a more reliable and sensitive detection method. Nevertheless, numerous methods have been developed to detect the S–NO bond. Gas-phase chemiluminescence methods, by using various strategies to cleave the S–NO bond (including photolytic cleavage, Cu^{+}/cysteine and I_2/I^{-}), are used to detect total SNO and individual S-nitrosylated proteins in combination with immunoprecipitation (Bryan *et al*., 2004; Liu *et al*., 2004). Other methods, such as [^{15}N] NMR and mass addition with electrospray ionization, are also used. Unfortunately, these techniques require specialized equipment that is not commonly available (Lane *et al*., 2001). A polyclonal antibody against SNO detects protein S-nitrosylation by immunohistochemistry, western blot, and immunoprecipitation (Matsushita *et al*., 2003). Conventional colorimetric methods including the Saville assay and measurement of the absorbance at 340 nm are not sensitive enough to detect SNO in vivo. Recently, a biotin-switch technique has been developed to identify S-nitrosylated proteins in vivo (Jaffrey *et al*., 2001). This approach provides a convenient tool in general laboratory settings and permits identification in vivo of S-nitrosylated proteins proteomically when coupled to mass spectrometry. The S-nitrosylated peptides can be identified by a modification of this procedure, providing nitrosopeptide mapping analogous to phosphopeptide mapping (Jaffrey *et al*., 2002).

3. Thiolation

Thiolation is another cysteine-based, oxidative, but reversible, posttranslational modification involving formation of mixed disulfides between protein thiols and low molecular weight (LMW) thiols such as GSH. This modification may serve as an antioxidant activity, preventing proteins from oxidative damage such as formation of disulfides among proteins resulting in protein aggregation, and irreversible oxidation of the thiol groups to a higher oxidation state (Thomas *et al*., 1995). Thiolation may be a physiologic cellular response to oxidative stress, mediating redox signaling by reversibly altering activities of cellular proteins (Murata *et al*., 2003). Thiolation is a nonenzymatic chemical reaction between reduced LMW thiols and reactive thiyl radicals or sulfenic acid intermediates that are derived from partial oxidation of protein thiols by a number of cellular oxidants, such as nitric oxide and its derivatives, superoxide anion, reactive lipids, and hydrogen peroxide (Table 2). It also can occur through thiol-disulfide exchange between protein thiols and oxidized glutathione and cysteine (GSSG and cystrine). Thiolation is selective with only some of cysteines in proteins being thiolated (Mallis *et al*., 2001).

Table 2 Mechanisms of thiolation

Mechanisms	Chemical reactions
Thiol-disulfide exchange	Protein-SH + RSSR → protein-S-SR + RS^-
Mediated by thiyl radical or sulfenic acid	Protein SH + ROS/RNS → protein-S*
	Protein-S* + RSH → protein-S-S-R
	Protein SH + ROS/RNS → protein-SOH
	Protein-SOH + RSH → protein-S-S-R
Irreversible oxidation	Protein SH + ROS/RNS → protein-SOH → protein-SO_2H/SO_3H

ROS: Reactive oxygen species; RNS: Reactive nitrogen species; Protein-S*: Protein thiyl radical; RSH: Low molecular weight thiols.

Moreover, it is also protein specific. For example, among GAPDH isozymes, only Tdh3, but not Tdh2, is thiolated when yeast is exposed to H_2O_2, though both isoforms can be thiolated by H_2O_2 in vitro (Grant *et al*., 1999). Thiolation does not apparently correlate well with cellular redox state (Shenton *et al*., 2002). Mechanisms underlying selective thiolation remain unclear. It is generally thought that the targeted cysteines must be exposed to solvent (liquid phase), providing access to LMW thiols in the cytosol. Features of the local microenvironment, such as a high pK_a that enhances ionization of the sulfhydryl groups, promote thiolation. GSH and GSSG are bulky and negatively charged molecules so that electrostatic interaction and steric interference are determinants of thiolation.

Protein thiolation is readily reversible when the oxidative stress is alleviated. Dethiolation is more likely to be an enzymatic process because the rate of direct dethiolation by GSH appears very slow. Indeed, several dithiol family enzymes, including thioredoxin, glutaredoxin, and protein disulfide isomerase, markedly facilitate dethiolation (Thomas *et al*., 1995) (Figure 1). Glutaredoxin specifically turns over thiolation but not other disulfide bonds. Recently, a novel monothiol-glutaredoxin, glutaredoxin 5, has been found to selectively dethiolate Tdh3 isoenzyme of GAPDH in yeast (Shenton *et al*., 2002). Substrate-specific dethiolation has recently been implicated in cell apoptotic signaling (Murata *et al*., 2003).

It is much more straightforward to detect thiolation than S-nitrosylation due to the relatively stable mixed disulfide bonds. In addition, thiolation can be labeled with [^{35}S].

4. Interaction between S-nitrosylation and thiolation

Ionized protein thiols are susceptible both to S-nitrosylation and thiolation. Indeed, many proteins can be nitrosylated or thiolated on the same reactive cysteine, for example, Cys^{181} in H-ras (Mallis *et al*., 2001) and Cys^{199} in oxyR (Kim *et al*., 2002). The S–NO bond is labile and easily oxidized to sulfenic acid that readily reacts with GSH, forming mixed disulfides (Figure 1); thus, protein S-nitrosylation promotes thiolation. Thiolation in vitro requires a higher concentration of GSNO

than does S-nitrosylation (Kim *et al*., 2002). The difference in effectiveness of these two modifications is likely to be determined by the distinctive nucleophilic attack properties of thiolated anions with either the nitrogen or with the sulfur. Electrostatic or steric interference for the bulky and charged GSH restrict access to reactive protein thiols more than small diffusible NO-derived species. Accordingly, S-nitrosylation should have a lower threshold than thiolation in response to NO signaling. Thus, protein S-nitrosylation occurs under normal physiological conditions, while thiolation is most likely to prevail with prolonged nitrosative stress.

References

Bredt DS (2003) Nitric oxide signaling specificity–the heart of the problem. *Journal of Cell Science*, **116**, 9–15.

Bryan NS, Rassaf T, Maloney RE, Rodriguez CM, Saijo F, Rodriguez JR and Feelisch M (2004) Cellular targets and mechanisms of nitros(yl)ation: an insight into their nature and kinetics in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4308–4313.

Fang M, Jaffrey SR, Sawa A, Ye K, Luo X and Snyder SH (2000) Dexras1: a G protein specifically coupled to neuronal nitric oxide synthase via CAPON. *Neuron*, **28**, 183–193.

Foster MW, McMahon TJ and Stamler JS (2003) S-nitrosylation in health and disease. *Trends in Molecular Medicine*, **9**, 160–168.

Grant CM, Quinn KA and Dawes IW (1999) Differential protein S-thiolation of glyceraldehyde-3-phosphate dehydrogenase isoenzymes influences sensitivity to oxidative stress. *Molecular and Cellular Biology*, **19**, 2650–2656.

Gu Z, Kaul M, Yan B, Kridel SJ, Cui J, Strongin A, Smith JW, Liddington RC and Lipton SA (2002) S-nitrosylation of matrix metalloproteinases: signaling pathway to neuronal cell death. *Science*, **297**, 1186–1190.

Haendeler J, Hoffmann J, Tischler V, Berk BC, Zeiher AM and Dimmeler S (2002) Redox regulatory and anti-apoptotic functions of thioredoxin depend on S-nitrosylation at cysteine 69. *Nature Cell Biology*, **4**, 743–749.

Jaffrey SR, Erdjument-Bromage H, Ferris CD, Tempst P and Snyder SH (2001) Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. *Nature Cell Biology*, **3**, 193–197.

Jaffrey SR, Fang M and Snyder SH (2002) Nitrosopeptide mapping: a novel methodology reveals s-nitrosylation of dexras1 on a single cysteine residue. *Chemistry & Biology*, **9**, 1329–1335.

Jaffrey SR, Snowman AM, Eliasson MJ, Cohen NA and Snyder SH (1998) CAPON: a protein associated with neuronal nitric oxide synthase that regulates its interactions with PSD95. *Neuron*, **20**, 115–124.

Johnson MA, Macdonald TL, Mannick JB, Conaway MR and Gaston B (2001) Accelerated s-nitrosothiol breakdown by amyotrophic lateral sclerosis mutant copper, zinc-superoxide dismutase. *The Journal of Biological Chemistry*, **276**, 39872–39878.

Kim SO, Merchant K, Nudelman R, Beyer WF Jr, Keng T, DeAngelo J, Hausladen A and Stamler JS (2002) OxyR: a molecular code for redox-related signaling. *Cell*, **109**, 383–396.

Lane P, Hao G and Gross SS (2001) S-nitrosylation is emerging as a specific and fundamental posttranslational protein modification: head-to-head comparison with O-phosphorylation. *Science's STKE: Signal Transduction Knowledge Environment*, **2001**, RE1.

Liu L, Yan Y, Zeng M, Zhang J, Hanes MA, Ahearn G, McMahon TJ, Dickfeld T, Marshall HE, Que LG, *et al*. (2004) Essential roles of S-nitrosothiols in vascular homeostasis and endotoxic shock. *Cell*, **116**, 617–628.

Mallis RJ, Buss JE and Thomas JA (2001) Oxidative modification of H-ras: S-thiolation and S-nitrosylation of reactive cysteines. *The Biochemical Journal*, **355**, 145–153.

Mannick JB, Schonhoff C, Papeta N, Ghafourifar P, Szibor M, Fang K and Gaston B (2001) S-Nitrosylation of mitochondrial caspases. *The Journal of Cell Biology*, **154**, 1111–1116.

Matsushita K, Morrell CN, Cambien B, Yang SX, Yamakuchi M, Bao C, Hara MR, Quick RA, Cao W, O'Rourke B, *et al*. (2003) Nitric oxide regulates exocytosis by S-nitrosylation of N-ethylmaleimide-sensitive factor. *Cell*, **115**, 139–150.

Murata H, Ihara Y, Nakamura H, Yodoi J, Sumikawa K and Kondo T (2003) Glutaredoxin exerts an antiapoptotic effect by regulating the redox state of Akt. *The Journal of Biological Chemistry*, **278**, 50226–50233.

Shenton D, Perrone G, Quinn KA, Dawes IW and Grant CM (2002) Regulation of protein S-thiolation by glutaredoxin 5 in the yeast *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry*, **277**, 16853–16859.

Thomas JA, Poland B and Honzatko R (1995) Protein sulfhydryls and their role in the antioxidant function of protein S-thiolation. *Archives of Biochemistry and Biophysics*, **319**, 1–9.

Short Specialist Review

Glycosylation in bacteria: that, what, how, why, now what?

Ben J. Appelmelk, Christina M. J. E. Vandenbroucke-Grauls and Wilbert Bitter
Vrije Universiteit Medical Center, Amsterdam, The Netherlands

1. That

The difference between pro- and eukaryotes suddenly appeared smaller when in 1974 Mescher and Strominger demonstrated that prokaryotes *also* could glycosylate their polypeptides. Initial work was on *Archeae* (like *Halobacterium halobium*) but later studies demonstrated that less exotic *Eubacteria* behaved similarly, and the S-layer glycoproteins were for years paradigmatic. More recently, it has become clear that also many common pathogens such as *Campylobacter jejuni*, *Neisseria meningitidis*, *Mycobacterium tuberculosis*, and *Pseudomonas aeruginosa* glycosylate polypeptides, often modifying outer membrane proteins (OMPs) or well-recognized virulence factors such as pili or flagella (Messner and Schaffner, 2003; Power and Reeves, 2003; Schmidt *et al.*, 2003; Szymanski *et al.*, 2003). Thus, the ability to express glycoproteins is widely distributed over the prokaryote kingdom.

2. What

As in eukaryotes, prokaryote glycans can occur O-linked to serine/threonine/tyrosine or N-linked through asparagine (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6, Article 65, **Structure/function of *O*-glycans**, Volume 6, and Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6). The eukaryotic consensus motif for N-glycosylation (Ser/Thr-X-Asn) has also been found in prokaryotes. While in eukaryotes, the (N-linked) glycan chains can be grouped into complexes, high mannose and hybrid, diversity in prokaryotes precludes such grouping. Their glycan substitutions vary between simple monosaccharides (Man, GlcNAc, GalNac) and complex oligosaccharides ($n > 15$) that include residues not encountered in man (like pseudaminic acid, an analog of sialic acid); these glycans may even consist of repeating units. Finally, to come full circle, high mannose glycans (8/9 mannoses), similar to those found in eukaryotes are expressed by the intracellular parasite *Chlamydia trachomatis*. Thus, prokaryotes protein glycans show a bewildering diversity and are sometimes completely eukaryotic-like.

3. How

The machinery required for glycosylating bacterial proteins has been elucidated recently in various species, and is best exemplified by *C. jejuni,* in which separate pathways for O- and N-glycosylation have been identified (Szymanski *et al*., 2003). These pathways are also physically separated on the chromosome, each organized in a glycosylation island. The O-glycosylation island (locus) consists of about 50 genes (between Cj1293 and Cj1337 in the genome strain 11168) and is adjacent to the flagellin structural genes *flaA* and *flaB*. Surprisingly, for this size of locus, the glycan chains of flagellin consist of a monosaccharide only (pseudaminic acid and derivatives). However, the degree of glycosylation is unsurpassed for a bacterial protein: 19 Ser/Thr residues carry this monosaccharide (=10% of the MW of native flagellin). For several genes in this locus, a likely role in the biosynthesis of pseudaminic acid could be assigned. For instance, Cj1293 codes for a functional UDP-GlcNAc dehydratase, Cj1313 and Cj1321 are acetyltransferases, and Cj1331 (called *ptmB*) is a CMP-NANA synthetase homolog, hence is likely to be involved in synthesis of the pseudaminic acid-CMP, the hypothesized glycosyl donor of this residue. Strikingly however, the glycosyltransferase itself has not been identified yet.

The N-glycosylation locus (also called *pgl*, for protein glycosylation) is much smaller (Cj1119c-1130c) but puts on a much more complex chain, namely, the heptasaccharide GalNAc-α1,4-GalNAc-α1,4-[Glcβ1,3]GalNAc-α1,4-GalNAc-α1,4-GalNAc-α1,3-bacillosamine (bacillosamine=2,4-diacetamido-2,4,6-trideoxyglucose). This machinery is also called the *general glycosylation pathway*, as it modifies >30 polypeptides that were identified after affinity chromatography with the GalNAc-specific lectin soybean agglutinin and SDS-PAGE. Mechanistic insight on how these glycans are synthesized was obtained though the identification of the Cj1126c (also called *pglB*) as an oligosaccharidyltransferase, highly homologous to its eukaryotic (yeast) counterpart STT3. STT3 transfers completed lipid-bound oligosaccharides to the Asn, as present in the consensus motif Asn-X-Ser/Thr. On the basis of analogy with eukaryotic *N*-glycan synthesis, the following scenario for *C. jejuni N*-glycan synthesis seems plausible (Figure 1): *pglC* (Cj1124c) first attaches an acetylhexosamine (starting with a UDP-acetylhexosamine donor) to a membrane-inserted polyprenyl lipid carrier; the subsequent actions of the *pglF* and other gene products of this locus, which contains at least five glycosyltransferases, form the complete heptasaccharide. Then, in analogy with eukaryotes, the lipid-bound heptasaccharide flip-flops from cytoplasmic to the periplasmic side of the membrane (by Cj1130c, WlaB, an ABC-transporter). The final step takes place when PglB transfers the heptasaccharide to Asn.

While in *C. jejuni* separate pathways exist for protein- and lipid glycosylation, in *P. aeruginosa* these routes intersect and the same trisaccharide is present on pilin and in LPS. Transfer of the trisaccharide from a lipid carrier toward pilin by a rather nonspecific oligosaccharidyltransferase (*pilO*) seems likely; hence, this route has similarities to the N-glycosylation system of *C. jejuni*.

Thus, the prokaryotic glycosylation machinery has been elucidated in quite substantial detail, in particular through whole-genome sequencing but not always functional data (enzyme activities) are available. Clearly, the N-glycosylation

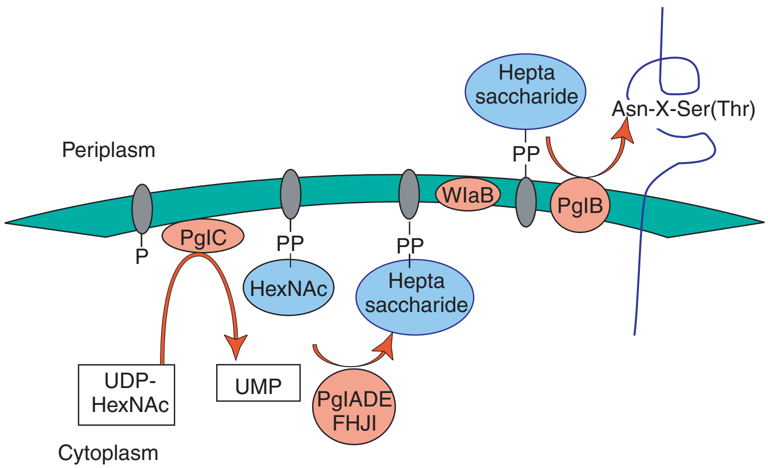


Figure 1 Model for the biosynthesis of N-linked glycoproteins in *Campylobacter jejuni* (adapted from Szymanski CM, Logan SM, Linton D and Wren B (2003) *Campylobacter jejuni* – a tale of two glycosylation systems. *Trends in Microbiology*, **11**, 233–238). Nucleotide-activated sugars are assembled on a lipid carrier, which results in the formation of a specific heptasaccharide. Subsequently, the glycan moiety is flipped across the inner membrane by WlaB (Cj1130c) and transferred to the appropriate asparagine residue by PglB. P: lipid-carrier phosphate; PP: lipid-carrier pyrophosphate; UDP: uridine diphosphate; UMP: uridine monophosphate; HexNAc: acetylhexosamine

pathway, formerly thought to be present in eukaryotes only, is present in Prokaryotes.

4. Why

Recent data suggest that bacterial protein glycosylation has, apart from a stabilizing or structural role, an important role in the adaptation of prokaryotes to their environment, like host–pathogen interaction. Two examples are given below.

4.1. The role of pilin glycosylation in gonococcal pathogenesis

Neisseria gonorrhoeae is the causative agent of gonorrhoea, a disease that can cause major complications upon be dissemination. PilE, the major structural pilin protein, is O-glycosylated with Galα1.3GlcNAc, with *pgtA* coding for the galactosyltransferase (Banerjee *et al*., 2002). In *pgtA*, a poly G-tract can be present, a signature for phase variation through DNA slippage (*see* Article 62, **The neisserial genomes: what they reveal about the diversity and behavior of these species**, Volume 4) (slipped-strand mispairing). When, for instance, a G14 tract is present in the parent bacterial cell, an intact GalT is formed and a disaccharide glycan is present. Upon cell division, owing to DNA slippage, phase-variant daughter cells arise with a G15 tract, which causes downstream frame-shifting and a

concomitant lack of a functional GalT and of galactosylated pili. Strikingly, strains that cause noncomplicated gonorrhea (i.e., confined to genitourinary epithelium) lack G-tracts (pili subunits are always modified), while disseminating strains display phase variation. A plausible mechanism is that the gonococci are recognized by an epithelial lectin and that for invasiveness prior detachment through phase variation is essential. Alternatively, the disaccharide might be the target of bactericidal serum antibodies, and hence phase variation would serve immune evasion.

4.2. *The role of Pseudomonas syringae flagellin glycosylation in host specificity*

Pseudomonas syringae is a plant pathogen of tobacco, soybean, tomato and others. Within *P. syringae*, pathovars (pv.) exist with strict host specificity (Takeuchi *et al*., 2003): while *P. syringae* pv. tabaci can infect its host the tobacco plant (compatibility), it is not compatible with nonhost plants such as soybean; vice versa, *P. syringae* pv. glycinea is compatible with soybean but not with tobacco. When a pathovar encounters a noncompatible host, a host defense (hypersensitivity, HR) reaction occurs, involving host cell apoptosis. Glycosylated pili are responsible for this HR reaction. In *P. syringae* pv. glycinea, a small glycosylation island (three orfs of which two are glycosyltransferases) was detected between two flagellin structural genes *fliC* and *fliL*. Disruption of the glycosyltransferases led to the production of non-/less-glycosylated pili and compatibility with tobacco, that is, this bacterium did no longer elicit the HR (Figure 2) reaction and could cause symptoms in this normally nonhost. Conversely, the mutants did not infect their original soybean host any more. These data prove that specific glycosylation and recognition is responsible for host specificity.

5. Now what?

By necessity, the study of prokaryote glycosylation started as “stamp collection”, that is, collecting structures of sometimes complex prokaryote glycan residues or chains. Now, the field is rapidly moving ahead into understanding the functional significance of glycosylation. The driving force is the explosion of -omics, including genomics, proteomics, and glycomics (Drickamer and Dell, 2002; *see also* Article 74, **Glycoproteomics**, Volume 6 and Article 75, **Mass spectrometry**, Volume 6). Protein spots present in 2D gels can be picked and, owing to the revolution in mass spectrometry (high throughput, enhanced sensitivity), be identified almost momentarily, compared to whole genomes, and assigned a function through increasingly sophisticated bioinformatics. In parallel, the field benefits from developments in carbohydrate synthesis (oligosaccharide libraries and glyco-chips). From the picture that is now emerging, it becomes clear that as in eukaryotes, prokaryote glycans can serve as information carriers, fine-tuning the interplay between prokaryotes and their environment, including human and nonhuman hosts.

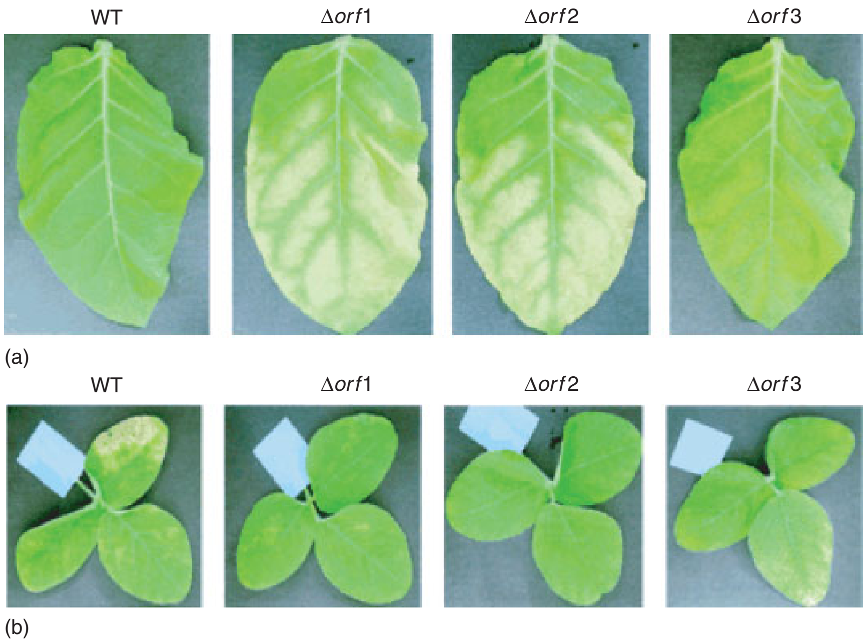


Figure 2 The role of pilin glycosylation on compatibility of *Pseudomonas syringae* pv. glycinea with tobacco (a) and soy (b) (Reproduced from Takeuchi K, Taguchi F, Inagaki Y, Toyoda K, Shraishi T and Ichose Y (2003) Flagellin glycosylation in *Pseudomonas syringae* pv. glycinea and its role in host specificity. *Journal of Bacteriology*, **185**, 6658–6665 by permission of American Society for Microbiology). Wild-type *Pseudomonas syringae* pv. glycinea is not compatible with (i.e., cannot infect) its nonhost tobacco but is so with its host soy, where it causes symptoms (whitish discoloration). After inactivation of glycosyltransferases required for pilin glycosylation (Δ*orf1* or Δ*orf2*), mutants are compatible with tobacco but now compatibility with the original host soy is lost

References

Banerjee A, Wang R, Supernavage SL, Ghosh SK, Parker J, Ganesh NF, Wang PG, Gulati S and Rice PA (2002) Implications of phase variation of a gene (*pgtA*) encoding a pilin galactosyltransferase in gonococcal pathogenesis. *The Journal of Experimental Medicine*, **196**, 147–162.

Drickamer K and Dell A (2002) *Glycogenomics: The Impact of Genomics and Informatics on Glycobiology*, Portland Press: London.

Messner P and Schaffner C (2003) Prokaryotic glycoproteins. In *Progress in the Chemistry of Organic Natural Products*, Vol. 85, Herz W, Falk H and Kirby GW (Ed.), Springer-Verlag: Vienna, pp. 51–124.

Power PM and Reeves MP (2003) The genetics of glycosylation in Gram-negative bacteria. *FEMS Microbiology Letters*, **218**, 211–222.

Schmidt MA, Riley LW and Benz I (2003) Sweet new world: glycoproteins in bacterial pathogens. *Trends in Microbiology*, **11**, 554–561.

Szymanski CM, Logan SM, Linton D and Wren B (2003) *Campylobacter jejuni*–a tale of two glycosylation systems. *Trends in Microbiology*, **11**, 233–238.

Takeuchi K, Taguchi F, Inagaki Y, Toyoda K, Shraishi T and Ichose Y (2003) Flagellin glycosylation in *Pseudomonas syringae* pv. glycinea and its role in host specificity. *Journal of Bacteriology*, **185**, 6658–6665.

Short Specialist Review

O-glycan processing

Anthony P. Corfield
University of Bristol, Bristol, UK

1. Introduction

Proteins may carry glycan chains linked by O- (*see* Article 65, **Structure/function of *O*-glycans**, Volume 6) or N-glycosidic bonds (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6) to the protein backbone. The biosynthesis and processing for each of these two types of glycan is quite different, reflecting discrete regulation of structure and biological function. The synthesis and processing of *O*- and *N*-glycans occurs in the endoplasmic reticulum and the Golgi through the controlled action of a series of glycosyltransferase and glycosidase reactions. The *N*-glycan chains are assembled through lipid (dolichol) intermediates and further processed by a series of specific glycosidases (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6). This is not the case for the O-glycans that do not rely on lipid intermediates.

The formation and processing of *N*- and *O*-glycans is regulated at the level of gene expression, mRNA, and enzyme protein activity. Additional control exists through substrate and cofactor concentrations at the subcellular site of synthesis. These processes are integrated to yield the huge variety of glycan structures utilized by different species, tissue types, cell types, and observed in different states of development and differentiation.

This short review focuses on the formation and processing of the Ser (Thr)-GalNAc class of *O*-glycans and those biological processing events for specific proteins that depend on the occurrence and manipulation of their O-linked oligosaccharides. Further examples of *O*-glycan processing events can be found for *N*-acetylglucosamine linked to serine or threonine (*O*-GlcNAc) on cytosolic and nuclear proteins (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6) and fucose and glucose O-linked to serine or threonine in the epidermal growth factor domains in a variety of proteins.

2. Processing of *O*-glycan chains

The assembly of the Ser (Thr)-GalNAc class of *O*-glycans is achieved by the action of a series of biosynthetic pathways that are located in regions of the endoplasmic reticulum and Golgi apparatus (*see* Article 65, **Structure/function of**

***O*-glycans**, Volume 6). These pathways involve largely glycosyltransferases, but also include *O*-glycan specific sulfotransferases and *O*-acetyltransferases, leading to the formation of defined oligosaccharide sequences, which have been confirmed by structural analysis (Brockhausen, 1999). The pathways lead to predictable structures with core, backbone, and peripheral units that are found linked to specific proteins. The subcellular organization of the enzymes, in particular the glycosyltransferases, into functional units acting on individual proteins is believed to deliver the specific glycosylation patterns found for each protein (Brockhausen, 1999). The examination of the biological function of individual proteins and the status of their glycans has revealed that processing of the oligosaccharide chains frequently plays a role (Van den Steen *et al*., 1998). The processing of *O*-glycans to generate biologically active glycoproteins may arise in several ways. Firstly, the attachment of *O*-glycans due to the sequential action of glycosyltransferases by the known pathways may deliver completed structures through the action of all enzymes in the pathways. However, in some cases, "incomplete" or truncated structures are formed, which have biological function. These glycans arise because of the action of a reduced number of glycosyltransferases in the biosynthetic units. The addition of backbone and/or peripheral monosaccharides does not occur and shortened or "incomplete" oligosaccharides result. This is the case for IgA hinge region *O*-glycans (Novak *et al*., 2000). Alternatively, the processing of completed *O*-glycan chains may be made by the action of catabolic enzymes including glycosidases, sulfatases, or esterases. Thus, *O*-glycans can be processed while attached to proteins in a manner analogous to the normal catabolic pathways responsible for the degradation and recycling of glycoproteins. These events usually involve removal of terminal residues such as sialic acids, fucose, or sulfate but can also include the total removal of *O*-glycan chains through the combined action of several exoglycosidases such as sialidases, fucosidase, and galactosidases together with endoglycosidases and *O*-glycanases responsible for the internal cleavage of the oligosaccharide chains or their removal for the peptide backbone respectively. The literature describing such events suggests that a variety of glycoproteins are associated with *O*-glycan processing phenomena and some examples are given below.

3. The role of O-linked glycosylation in protein expression and processing

Many O-glycosylated glycoproteins that are known to be biologically processed rely on their *O*-glycans to achieve these events. This posttranslational modification is an integral step enabling the function of the mature glycoprotein in vivo. Examples have been reported in several contexts and are grouped below to represent a selection from the current literature. Some of these aspects can also be found summarized in Article 65, **Structure/function of *O*-glycans**, Volume 6. In addition, glycoproteins with both *O*- and *N*-glycans are common and differential roles for the glycan chains relating to processing and other biological functions have been reported.

3.1. Protein synthesis and expression

The expression of proteins can be modulated at different stages by the presence of glycan chains. The synthesis of mature glycophorin A has been found to be dependent on the presence of *O*-glycans when expressed in glycosylation-deficient CHO cells. In this case, the *N*-glycans normally present in glycophorin A are not necessary for expression (Remaley *et al*., 1991). Further investigation showed that expression of glycophorin A without *O*-glycans could be achieved provided specific *N*-glycan structures were present. Thus, independent expression of *O*- and *N*-glycans can enable glycophorin A expression.

3.2. Proteolytic processing

A number of proteins depend on the addition of *O*-glycans at specific sites in order to prevent proteolytic cleavage destroying biological activity or preventing stable residence of the protein at its required subcellular location.

Insulin-like growth factor-2 (IGF-2) is expressed in most embryonic tissues and is required for normal development during gestation but is only expressed in adults during tumorigenesis. Two forms occur in serum – a big IGF-2 and a normal form. Proteolytic cleavage occurs to convert big to normal form, and this is mediated at Thr75 by the presence of an *O*-glycan. Big IGF-2 has no *O*-glycan and is not clipped (Duguay *et al*., 1998).

The expression of LDL receptors at the cell surface relies on the presence of *O*-glycan chains, which prevent proteolytic cleavage of the extracellular domains of these proteins (Kozarsky *et al*., 1988). A further example is the transferrin receptor (TfR). TfR is normally transported to the cell surface and then cycles between the cell surface and the endosomes. A soluble form, sTfR, is formed by proteolytic cleavage in the endosomes. The protease responsible is sensitive to the presence of an *O*-glycan at Thr104. Desialylation also abolishes the protection of the *O*-glycans against proteolytic activity. (Rutledge and Enns, 1996).

Human meprinβ, a member of the astacin family of zinc metalloendopeptidases, shows similar properties of proteolytic release from the cell surface. The O-linked carbohydrate side chains are clustered around a 13 amino acid sequence that contains the main cleavage site for proteolytic processing of the subunit. Prevention of O-glycosylation by specific inhibitors leads to enhanced proteolytic processing and the consequence is an increased release of the enzyme (Leuenberger *et al*., 2003).

Further examples of proteolytic processing are found for the blood clotting factors X and Xa.

The activation of the glycoproteins X and Xa occurs through proteolytic cleavage, by factor IXa and its cofactor VIIIa in the intrinsic cascade and by factor VIIIa and tissue factor in the extrinsic cascade, ultimately leading to thrombin formation. The activation of factor X has been reported to depend on the presence of sialic acid in *O*-glycans as asialo forms show reduced activation. This remains controversial, as other reports have not observed this relationship. However, the removal of the

O-glycans from factor X results in a significant reduction in the k_{cat} for the activation reaction (Inoue and Morita, 1993).

An example where O-glycosylation promotes proteolytic action in processing is found for CD44 in cells of melanocytic lineage. The proteolytic cleavage and shedding is dependent on the presence of partial or complete O-glycosylation of four serine-glycine motifs localized in the membrane-proximal CD44 ectodomain. Mutation of these serine residues, as well as extensive metabolic O-deglycosylation, block spontaneous CD44 shedding (Gasbarri *et al*., 2003).

3.3. Subcellular localization of proteins

The transport of proteins to the Golgi apparatus from the endoplasmic reticulum is closely linked with O-glycosylation. The biosynthesis, maturation, and intracellular transport of galactosyltransferase, an established trans-Golgi enzyme, are disrupted by brefeldin A that induces a microtubule-dependent backflow of Golgi components to the endoplasmic reticulum. The targeting of the galactosyltransferase for the Golgi apparatus is dependent on O-glycosylation (Bosshart *et al*., 1991). Brefeldin A has also been used to demonstrate O-glycosylation of Ribophorin I, a type I transmembrane glycoprotein specific to the rough endoplasmic reticulum. Brefeldin A treatment leads to O-glycosylation of Ribophorin I by glycosyltransferases that are redistributed from the Golgi apparatus to the endoplasmic reticulum (Ivessa *et al*., 1992).

Human neurotrophin receptor (p75 NTR) contains a cluster of O-glycans in the ectodomain located close to the membrane. The transport and sorting of p75 NTR in Caco-2 cells requires these O-glycans. After initial synthesis, the receptor is targeted to the basolateral membrane and subsequently migrates to the apical membrane. Localization to the apical surface is controlled jointly by the O-glycans together with a membrane anchor domain (Breuza *et al*., 1999).

3.4. Developmental expression

The expression of *O*-glycans during developmental processes associated with specific biological activity has been reported. Two examples illustrate this phenomenon.

The CD8$\alpha\beta$ coreceptor possesses an O-glycosylated polypeptide stalk connecting the glycoprotein to the thymocyte surface. Immature CD4(+) CD8(+) double-positive thymocytes bind major histocompatibility complex class I (MHCI) tetramers more avidly than mature CD8 single-positive thymocytes. This differential binding is regulated by a developmentally expressed sialyltransferase (ST3Gal-I) leading to O-glycan modification. The appearance of core 1 sialic acid linked to CD8β on mature thymocytes decreases CD8$\alpha\beta$ -MHCI avidity and thus modulates the binding of dimeric CD8 to MHCI (Moody *et al*., 2001).

The brush border proteins pro-sucrase-isomaltase (pro-SI) and dipeptidyl peptidase IV (DPPIV) are sorted for apical location during the differentiation of the crypt cells. These glycoproteins have both *N*- and *O*-glycans. Full O-glycosylation

is mediated by the N-glycans and the subsequent polarized sorting of these proteins to the apical membrane requires intact *O*-glycans. This example demonstrates interrelated but distinct roles for *N*- and *O*-glycans and their processing in brush border enzyme expression during intestinal cell differentiation. (Naim *et al*., 1999).

4. Immune cell expression and response

A regulatory function for T-lymphocyte Core 2 β-1,6-*N*-acetylglucosaminyltransferase, (C2GnT) in the peripheral immune system has been identified.

CD43 is the major cell-surface glycoprotein and an activation antigen expressed on both CD4 and CD8 single-positive T lymphocytes and carries core 2 O-glycan structures. Downregulation of the activation antigen occurs in thymic positive selection implying that the C2GnT modulated expression of CD43 glycoforms is involved in thymic selection events. (Ellies *et al*., 1996).

Additional support for this type of role for the Core 2 β-1,6-*N*-acetylglucosaminyltransferase comes from overexpression of the enzyme in transgenic mice. Isolation of T lymphocytes from these mice showed a reduced immune response due to impaired cell–cell interaction (Tsuboi and Fukuda, 1997).

References

Bosshart H, Straehl P, Berger B and Berger EG (1991) Brefeldin A induces endoplasmic reticulum-associated O-glycosylation of galactosyltransferase. *Journal of Cellular Physiology*, **147**, 149–156.

Breuza L, Monlauzeur L, Arsanto JP and Le Bivic A (1999) Identification of signals and mechanisms of sorting of plasma membrane proteins in intestinal epithelial cells. *Journal de la Societe de Biologie*, **193**, 131–134.

Brockhausen I (1999) Pathways of O-glycan biosynthesis in cancer cells. *Biochimica et Biophysica Acta*, **1473**, 67–95.

Duguay SJ, Jin Y, Stein J, Duguay AN, Gardner P and Steiner DF (1998) Post-translational processing of the insulin-like growth factor-2 precursor. Analysis of O-glycosylation and endoproteolysis. *The Journal of Biological Chemistry*, **273**, 18443–18451.

Ellies LG, Tao W, Fellinger W, Teh HS and Ziltener HJ (1996) The CD43 peripheral T cell activation antigen is downregulated in thymic positive selection. *Blood*, **88**, 1725–1732.

Gasbarri A, Del Prete F, Girnita L, Martegani MP, Natali PG and Bartolazzi A (2003) CD44s adhesive function spontaneous and PMA-inducible CD44 cleavage are regulated at post-translational level in cells of melanocytic lineage. *Melanoma Research*, **13**, 325–337.

Inoue K and Morita T (1993) Identification of O-linked oligosaccharide chains in the activation peptides of blood coagulation factor X. The role of the carbohydrate moieties in the activation of factor X. *European Journal of Biochemistry*, **218**, 153–163.

Ivessa NE, De Lemos-Chiarandini C, Tsao YS, Takatsuki A, Adesnik M, Sabatini DD and Kreibich G (1992) O-glycosylation of intact and truncated ribophorins in brefeldin A-treated cells: newly synthesized intact ribophorins are only transiently accessible to the relocated glycosyltransferases. *The Journal of Cell Biology*, **117**, 949–958.

Kozarsky K, Kingsley D and Krieger M (1988) Use of a mutant cell line to study the kinetics and function of O-linked glycosylation of low density lipoprotein receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 4335–4339.

Leuenberger B, Hahn D, Pischitzis A, Hansen MK and Sterchi EE (2003) Human meprin beta: O-linked glycans in the intervening region of the type I membrane protein protect the C-terminal

region from proteolytic cleavage and diminish its secretion. *The Biochemical Journal*, **369**, 659–665.

Moody AM, Chui D, Reche PA, Priatel JJ, Marth JD and Reinherz EL (2001) Developmentally regulated glycosylation of the CD8$\alpha\beta$ co-receptor stalks modulates ligand binding. *Cell*, **107**, 501–512.

Naim HY, Joberty G, Alfalah M and Jacob R (1999) Temporal association of the N- and O-linked glycosylation events and their implication in the polarized sorting of intestinal brush border sucrase-isomaltase, aminopeptidase N, and dipeptidyl peptidase IV. *The Journal of Biological Chemistry*, **274**, 17961–17967.

Novak J, Tomana M, Kilian M, Coward L, Kulhavy R, Barnes S and Mestecky J (2000) Heterogeneity of *O*-glycosylation in the hinge region of human IgA1. *Molecular Immunology*, **37**, 1047–1056.

Remaley AT, Ugorski M, Wu N, Litzky L, Burger SR, Moore JS, Fukuda M and Spitalnik SL (1991) Expression of human glycophorin A in wild type and glycosylation- deficient Chinese hamster ovary cells. Role of N- and O-linked glycosylation in cell surface expression. *The Journal of Biological Chemistry*, **266**, 24176–24183.

Rutledge EA and Enns CA (1996) Cleavage of the transferrin receptor Is influenced by the composition of the O-linked carbohydrate at position 104. *Journal of Cellular Physiology*, **168**, 284–293.

Tsuboi S and Fukuda M (1997) Branched O-linked oligosaccharides ectopically expressed in transgenic mice reduce primary T-cell immune responses. *The EMBO Journal*, **16**, 6364–6373.

Van den Steen P, Rudd PM, Dwek RA and Opdenakker G (1998) Concepts and principles of O-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology*, **33**, 151–208.

Short Specialist Review

O-Mannosylation

Tamao Endo
Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan

1. Introduction

Over 60% of the proteins produced by the human body are thought to contain sugar chains (*see* Article 62, **Glycosylation**, Volume 6). Recent data have revealed the importance of sugar chains as biosignals for multicellular organisms, including cell–cell communication, intracellular signaling, protein folding, and targeting of proteins within cells. There is growing evidence that sugar chains have a variety of roles in cellular differentiation and development, as well as in disease processes.

The major sugar chains of glycoproteins can be classified into two groups according to their sugar-peptide linkage regions. Those that are linked to asparagine (Asn) residues of proteins are termed *N*-glycans (*see* Article 64, **Structure/function of *N*-glycans**, Volume 6), while those that are linked to serine (Ser) or threonine (Thr) residues are called *O*-glycans (*see* Article 65, **Structure/function of *O*-glycans**, Volume 6). In *N*-glycans, the reducing terminal *N*-acetylglucosamine (GlcNAc) is linked to the amide group of Asn via an aspartylglycosylamine linkage. In *O*-glycans, the reducing terminal *N*-acetylgalactosamine (GalNAc) is attached to the hydroxyl group of Ser and Thr residues. In addition to the abundant *O*-GalNAc forms, several unique types of protein *O*-glycosylation have been found, such as *O*-linked fucose, glucose, GlcNAc (*see* Article 72, **O-linked *N*-acetylglucosamine (*O*-GlcNAc)**, Volume 6), and mannose, which have been shown to mediate diverse physiological functions. For example, *O*-mannosylation has recently been shown to be important in muscle and brain development.

2. Structure of *O*-mannosyl glycan

O-Mannosylation is known as a yeast-type modification, and *O*-mannosylated glycoproteins are abundant in the yeast cell wall (Strahl-Bolsinger *et al*., 1999). In unicellular eukaryotic organisms, all *O*-mannosyl glycan structures elucidated so far are neutral linear glycans consisting of 1 to 7 mannose residues (Figure 1). *O*-Mannosylation of proteins has been shown to be vital in yeast, and its absence may affect cell wall structure and rigidity. Additionally, a deficiency in protein *O*-mannosylation in the fungal pathogen *Candida albicans* leads to defects in multiple cellular functions, including expression of virulence. In addition to fungi and yeast,

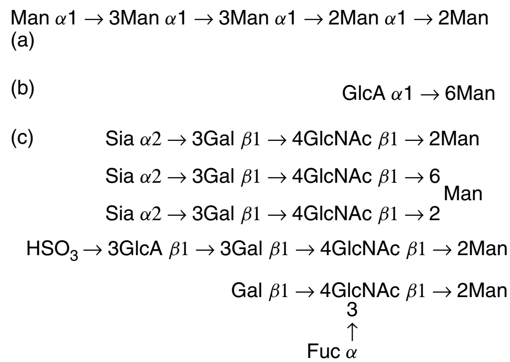


Figure 1 *O*-Mannosyl glycans found in (a) yeast, (b) clam worm, and (c) mammals. Man: mannose; GlcA: glucuronic acid; Sia: sialic acid; Gal: galactose; GlcNAc: *N*-acetylglucosamine; Fuc: fucose

clam worm has an *O*-mannosyl glycan (a glucuronylα1-6mannosyl disaccharide) in skin collagen (Spiro and Bhoyroo, 1980).

Mammalian *O*-mannosylation is an unusual type of protein modification that was first identified in chondroitin sulfate proteoglycans of brain, and is present in a limited number of glycoproteins of brain, nerve, and skeletal muscle (Endo, 1999). In brief, one of the best known *O*-mannosyl-modified glycoproteins is α-dystroglycan, which is a central component of the dystrophin-glycoprotein complex isolated from skeletal muscle membranes. We previously found that the glycans of α-dystroglycan include *O*-mannosyl oligosaccharides and that a sialyl *O*-mannosyl glycan, Siaα2-3Galβ1-4GlcNAcβ1-2Man, is a laminin-binding ligand of α-dystroglycan (Chiba *et al*., 1997). Interestingly, we found the same *O*-mannosyl glycan in rabbit skeletal muscle α-dystroglycan. After our reports of the sialylated *O*-mannosyl glycan, an HNK-1 epitope (sulfoglucuronyl lactosamine) carrying *O*-mannosyl glycan (HSO_3-3GlcAβ1-3Galβ1-4GlcNAcβ1-2Man) was detected in total brain glycopeptides. It is noteworthy that these oligosaccharides have not only 2-substituted mannose but also 2,6-disubstituted mannose. Further, dystroglycan from sheep brain has a Galβ1-4(Fucα1-3)GlcNAcβ1-2Man structure and mouse J1/tenascin, which is involved in neuron-astrocyte adhesion, contains the *O*-mannosyl glycans. Therefore, it is likely that a series of *O*-mannosyl glycans, with heterogeneity of mannose-branching and peripheral structures, is present in mammals. Further studies are needed to clarify the distribution of such *O*-mannosyl glycans in various tissues.

3. Biosynthesis of *O*-mannosyl glycan

Identification and characterization of the enzymes involved in the biosynthesis of mammalian type *O*-mannosyl glycans will help to elucidate the function and regulation of expression of these glycans.

A key difference among yeast, clam worm, and mammalian-type *O*-mannosyl glycans is that those in mammals have the GlcNAc β1-2Man linkage (Figure 1).

This linkage is assumed to be catalyzed by a glycosyltransferase, UDP-GlcNAc: protein *O*-mannose β1,2-*N*-acetylglucosaminyltransferase (POMGnT1). POMGnT1 catalyzes the transfer of GlcNAc from UDP-GlcNAc to *O*-mannosyl glycoproteins, and human *POMGnT1* gene was cloned (Yoshida *et al*., 2001). The nucleotide sequence indicated that human POMGnT1 is a 660 amino acid protein and is a type II membrane protein. This topology was similar to the topologies of other Golgi glycosyltransferases. As described already, mammalian *O*-mannosyl glycan has 2,6-disubstituted mannose. Very recently, a gene for this 6-branching enzyme (GnT-IX) has been cloned (Inamori *et al*., 2004).

In yeast, the family of protein *O*-mannosyltransferases catalyzes the transfer of a mannosyl residue from dolichyl phosphate mannose (Dol-P-Man) to Ser/Thr residues of certain proteins (Strahl-Bolsinger *et al*., 1999). However, attempts to detect protein *O*-mannosyltransferase activity and to characterize the enzyme(s) responsible for the biosynthesis of *O*-mannosyl glycans in vertebrates have not been successful. Two human homologs, *POMT1* and *POMT2*, were found but their gene products did not show any protein *O*-mannosyltransferase activity (Jurado *et al*., 1999; Willer *et al*., 2002). POMT1 and POMT2 share almost identical hydropathy profiles that predict both to be integral membrane proteins with multiple transmembrane domains. Recently, we developed a new method to detect the enzymatic activity of protein *O*-mannosyltransferase in mammalian cells and tissues. Using this new method, we demonstrated that human POMT1 and POMT2 have protein *O*-mannosyltransferase activity, but only when they are coexpressed (Manya *et al*., 2004). This suggests that POMT1 and POMT2 form a hetero-complex to express enzymatic activity similar to the complex in yeast (Girrbach and Strahl, 2003). *POMT1* and *POMT2* are expressed in all human tissues, but *POMT1* is highly expressed in fetal brain, testis, and skeletal muscle, and *POMT2* is predominantly expressed in testis (Jurado *et al*., 1999; Willer *et al*., 2002). *O*-Mannosylation seems to be uncommon in mammals, and only a few *O*-mannosylated proteins have been identified. It will be of interest to determine the regulatory mechanisms for protein *O*-mannosylation in each tissue.

Enzymes for galactosylation, sialylation, fucosylation, glucuronylation, or sulfation of *O*-mannosyl glycans are not identified yet.

4. Defects of *O*-mannosylation and congenital muscular dystrophies

Muscular dystrophies are genetic diseases that cause progressive muscle weakness and wasting (Burton and Davies, 2002). Recent data suggest that aberrant *O*-mannosylation of α-dystroglycan is the primary cause of some forms of congenital muscular dystrophy and neuronal migration disorder.

Muscle-eye-brain disease (MEB: OMIM 253280) is an autosomal recessive disorder characterized by congenital muscular dystrophy, ocular abnormalities, and brain malformation (type II lissencephaly). MEB has been observed mainly in Finland. After we screened the entire coding region and the exon/intron flanking sequences of the *POMGnT1* gene for mutations in patients with MEB, we identified

13 independent disease-causing mutations in these patients (Taniguchi *et al*., 2003; Yoshida *et al*., 2001). We have not detected these 13 substitutions in any of the 300 normal chromosomes, indicating that they are pathogenic and that the *POMGnT1* gene is responsible for MEB. To confirm that the mutations observed in patients with MEB are responsible for the defects in the synthesis of *O*-mannosyl glycan, we expressed all of the mutant proteins and found that none of them had enzymatic activity (Manya *et al*., 2003; Yoshida *et al*., 2001). Together, these findings indicate that MEB is inherited as a loss-of-function of the *POMGnT1* gene. If POMGnT1 does not work, no peripheral structure can be formed on *O*-mannosyl glycans. Because these structures are involved in adhesive processes, a defect of *O*-mannosyl glycan may severely affect cell migration and cell adhesion. Additionally, a selective deficiency of α-dystroglycan in MEB patients was found. This finding suggests that α-dystroglycan is a potential target of POMGnT1 and that hypoglycosylation of α-dystroglycan may be a pathomechanism of MEB. MEB muscle and brain phenotypes can be explained by abnormal *O*-mannosylation.

Walker–Warburg syndrome (WWS: OMIM 236670) is another form of congenital muscular dystrophy that is characterized by severe brain malformation (type II lissencephaly) and eye involvement. Patients with WWS are severely affected from birth and usually die within their first year. WWS has a worldwide distribution. Recently, 20% of WWS patients (6 of 30 unrelated WWS cases) has been found to have mutations in *POMT1* (Beltrán-Valero de Bernabé *et al*., 2002), but none of the 30 cases studied had mutations in another homolog, *POMT2*. This suggests that other as yet unidentified genes are responsible for this syndrome.

In WWS patients, as in MEB patients, a highly glycosylated α-dystroglycan was selectively deficient in skeletal muscle. WWS and MEB are clinically similar disorders, but WWS is a more severe syndrome than MEB. The difference of severity between the two diseases may be explained as follows. If POMGnT1, which is responsible for the formation of the GlcNAcβ1-2Man linkage of *O*-mannosyl glycans, is nonfunctional, only *O*-mannose residues may be present on α-dystroglycan in MEB. On the other hand, *POMT1* mutations cause complete loss of *O*-mannosyl glycans in WWS. It is possible that the attachment of a single mannose residue on α-dystroglycan in MEB is responsible for the difference in the clinical severity of WWS and MEB.

Interestingly, the *rt* mutation in *Drosophila*, which causes defects of myogenesis, was found to be due to a mutation in a homolog of *POMT1* (Martin-Blanco and Garcia-Bellido, 1996; Willer *et al*., 2002). Although the *rt* gene product is not known to initiate the biosynthesis of *O*-mannosyl glycans, *O*-mannosylation is an evolutionarily conserved protein modification (Endo, 1999), and may be essential for muscle development in both vertebrates and invertebrates.

After reporting that MEB and WWS are caused by defects of *O*-mannosylation, some muscular dystrophies have been suggested to be caused by abnormal glycosylation of α-dystroglycan, for example, Fukuyama-type congenital muscular dystrophy (FCMD: OMIM 253800), congenital muscular dystrophy type 1C (MDC1C: OMIM 606612), congenital muscular dystrophy type 1D (MDC1D), and the myodystrophy (*myd*) mouse. However, although highly glycosylated α-dystroglycan was found to be selectively deficient in the skeletal muscle of these patients and the gene products were thought to be putative glycosyltransferases, it is still unclear

due to the defects of *O*-mannosylation. Identification of these defects may provide new clues to the glycopathomechanism of muscular dystrophy.

5. Conclusions

O-Mannosylation is an uncommon protein modification in mammals, but it is important in muscle and brain development. Since a few *O*-mannosylated proteins have been identified, further proteomic studies are needed to clarify the distribution of *O*-mannosyl glycans in various tissues and to examine their changes during development. Future studies may also reveal that presently uncharacterized forms of muscular dystrophy are caused by defects in other glycosyltransferases. A major challenge will be to integrate the forthcoming structural, cell biological, and genetic information to understand how α-dystroglycan *O*-mannosylation contributes to muscular dystrophy and neuronal migration disorder.

Acknowledgments

The author gratefully acknowledges the financial support of Research Grants for Nervous and Mental Disorders (14B-4) and Research on Psychiatric and Neurological Diseases and Mental Health from the Ministry of Health, Labour and Welfare of Japan, and of a Grant-in-Aid for Scientific Research on Priority Area (14082209) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, throughout this project.

References

Beltrán-Valero de Bernabé D, Currier S, Steinbrecher A, Celli J, van Beusekom E, van Der Zwaag B, Kayserili H, Merlini L, Chitayat D, Dobyns WB, *et al*. (2002) Mutations in the *O*-mannosyltransferase Gene *POMT1* give rise to the severe neuronal migration disorder Walker-Warburg syndrome. *American Journal of Human Genetics*, **71**, 1033–1043.

Burton EA and Davies KE (2002) Muscular dystrophy–reason for optimism? *Cell*, **108**, 5–8.

Chiba A, Matsumura K, Yamada H, Inazu T, Shimizu T, Kusunoki S, Kanazawa I, Kobata A and Endo T (1997) Structures of sialylated *O*-linked oligosaccharides of bovine peripheral nerve α-dystroglycan. The role of a novel *O*-mannosyl-type oligosaccharide in the binding of α-dystroglycan with laminin. *Journal of Biological Chemistry*, **272**, 2156–2162.

Endo T (1999) *O*-Mannosyl glycans in mammals. *Biochimica et Biophysica Acta*, **1473**, 237–246.

Girrbach V and Strahl S (2003) Members of the evolutionarily conserved PMT family of protein *O*-mannosyltransferases form distinct protein complexes among themselves. *Journal of Biological Chemistry*, **278**, 12554–12562.

Inamori KI, Endo T, Gu J, Matsuo I, Ito Y, Fujii S, Iwasaki H, Narimatsu H, Miyoshi E, Honke K, *et al*. (2004) *N*-Acetylglucosaminyltransferase IX acts on the GlcNAcβ1,2-Manα1-Ser/Thr moiety, forming a 2,6-branched structure in brain *O*-mannosyl glycan. *Journal of Biological Chemistry*, **279**, 2337–2340.

Jurado LA, Coloma A and Cruces J (1999) Identification of a human homolog of the *Drosophila rotated abdomen* gene (*POMT1*) encoding a putative protein *O*-mannosyl-transferase, and assignment to human chromosome 9q34.1. *Genomics*, **58**, 171–180.

Manya H, Chiba A, Yoshida A, Wang X, Chiba Y, Jigami Y, Margolis RU and Endo T (2004) Demonstration of mammalian protein *O*-mannosyltransferase activity: Coexpression

of POMT1 and POMT2 required for enzymatic activity. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 500–505.

Manya H, Sakai K, Kobayashi K, Taniguchi K, Kawakita M, Toda T and Endo T (2003) Loss-of-function of an *N*-acetylglucosaminyltransferase, POMGnT1, in muscle-eye-brain disease. *Biochemical and Biophysical Research Communications*, **306**, 93–97.

Martin-Blanco E and Garcia-Bellido A (1996) Mutations in the *rotated abdomen* locus affect muscle development and reveal an intrinsic asymmetry in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 6048–6052.

Spiro RG and Bhoyroo VD (1980) Studies on the carbohydrate of collagens. Characterization of a glucuronic acid-mannose disaccharide unit from *Nereis* cuticle collagen. *Journal of Biological Chemistry*, **255**, 5347–5354.

Strahl-Bolsinger S, Gentzsch M and Tanner W (1999) Protein *O*-mannosylation. *Biochimica et Biophysica Acta*, **1426**, 297–307.

Taniguchi K, Kobayashi K, Saito K, Yamanouchi H, Ohnuma A, Hayashi YK, Manya H, Jin DK, Lee M, Parano E, *et al*. (2003) Worldwide distribution and broader clinical spectrum of muscle-eye-brain disease. *Human Molecular Genetics*, **12**, 527–534.

Willer T, Amselgruber W, Deutzmann R and Strahl S (2002) Characterization of POMT2, a novel member of the *PMT* protein *O*-mannosyltransferase family specifically localized to the acrosome of mammalian spermatids. *Glycobiology*, **12**, 771–783.

Yoshida A, Kobayashi K, Manya H, Taniguchi K, Kano H, Mizuno M, Inazu T, Mitsuhashi H, Takahashi S, Takeuchi M, *et al*. (2001) Muscular dystrophy and neuronal migration disorder caused by mutations in a glycosyltransferase, POMGnT1. *Developmental Cell*, **1**, 717–724.

O-linked *N*-acetylglucosamine (*O*-GlcNAc)

Stephen A. Whelan and Gerald W. Hart
The Johns Hopkins University School of Medicine, Baltimore, MD, USA

1. Introduction

O-linked *β*-*N*-acetylglucosamine (*O*-GlcNAc) was discovered in 1984 in murine lymphocytes (Torres and Hart, 1984). Virtually all of the *O*-GlcNAc modification is located on nucleocytoplasmic proteins (Holt and Hart, 1986), thus disproving the dogma that glycosylated proteins were only secreted or associated with biological membranes such as ER, Golgi, and other subcellular organelles. *O*-GlcNAc was shown to be both abundant and dynamic, occurring exclusively on serine/threonine residues. *O*-GlcNAc, like phosphorylation, has been detected on functionally diverse types of proteins including, kinases and signal transduction molecules, phosphatases, cytoskeletal proteins, tumor suppressors, hormone receptors, nuclear pore proteins, oncogenes, transcription factors, transcriptional and translational machinery, and viral proteins (Hart, 1997; Whelan and Hart, 2003).

O-GlcNAc plays an analogous role to phosphorylation in cellular regulation with rapid cycling in response to cellular signaling. *O*-GlcNAc occurs at sites that are also modified by protein kinases. *O*-GlcNAc is often reciprocal with phosphorylation at or near adjacent sites on proteins such as, c-Myc, estrogen receptor-*β*, and RNA Pol II (for review, see Hart, 1997). Serine and threonine residues may exist in at least three states, glycosylated, phosphorylated, or unmodified. Multiple *O*-GlcNAc and phosphorylation sites dramatically increase the complexity of protein regulation. The cell's proteome is further complicated with combinations of other abundant posttranslational modifications, as well as acetylation, methylation, and ubiquination. The functions of *O*-GlcNAc modification remain elusive due to the technical difficulties associated with the study of this saccharide. Recently, a number of tools and proteomic techniques have allowed initial deciphering of the complexity of the role of *O*-GlcNAc in the proteome (Roquemore *et al*., 1994; Whelan and Hart, 2003).

2. Enzymatic addition of *O*-GlcNAc

O-linked *β*-*N*-acetylglucosamine transferase (OGT) is the enzyme responsible for the addition of *O*-GlcNAc to nucleocytoplasmic proteins (Haltiwanger *et al*., 1990).

However, unlike the many kinases, there is currently only one known eukaryotic gene that encodes the catalytic subunit of OGT. In addition, knockout studies of the *OGT* gene showed that OGT resides on the X chromosome (Xq13.1) and is necessary for embryonic stem cell viability (Shafi *et al*., 2000). OGT is also evolutionarily conserved from *Caenorhabditis elegans* to man, with rat and human sequences being virtually identical (>99%), further indicating the importance of this enzyme's function in metazoans (for review, see Iyer and Hart, 2003).

Although there are multiple splice variants of OGT, the main native enzyme is composed of three identical 110-kDa peptides (Kreppel *et al*., 1997; Lubas *et al*., 1997). There is also a 78-kDa OGT subunit that is expressed in some tissues, which may be a result of proteolysis or derived via an alternate start site from OGT cDNA. A 103-kDa mitochondrial isoform has also been described (Hanover *et al*., 2003). The TPR domain of OGT is involved in protein–protein interactions as well as trimerization and stability, allowing OGT to interact with a large variety of proteins, including the characterized mSin3A (Yang *et al*., 2002) and OIP106/GRIF-1 (for review, see Iyer and Hart, 2003). OGT itself is modified by *O*-GlcNAc and by tyrosine phosphorylation, suggesting it may be regulated by tyrosine kinase–mediated signal transduction cascades (Kreppel *et al*., 1997).

3. Enzymatic removal of *O*-GlcNAc from proteins

Nucleocytoplasmic β-D-*N*-acetylglucosaminidase (*O*-GlcNAcase) hydrolyzes *O*-GlcNAc (Dong and Hart, 1994). *O*-GlcNAcase is also highly conserved throughout evolution from *C. elegans* to man. *O*-GlcNAcase is a member of the GCN5-related acetyltransferases family (GNAT), with the C-terminal half having the *O*-GlcNAcase activity and the N-terminal half containing putatively weak hyaluronidase activity (Schultz and Pils, 2002). *O*-GlcNAcase is ubiquitously expressed in two forms, a 130-kDa form that is primarily found in the cytosolic fraction and a nuclear localized inactive 75-kDa splice variant.

4. *O*-GlcNAc is a nutrient sensor

The hexosamine biosynthetic pathway (HBP) is responsible for the conversion of 2–5% of the glucose that enters the cell into uridine 5′-diphosphate (UDP)-GlcNAc, the activated sugar donor for OGT. OGT is highly responsive to intracellular UDP-GlcNAc concentrations having varying K_m values for the addition of *O*-GlcNAc to proteins (Kreppel and Hart, 1999). Within the cell, *O*-GlcNAc levels correlate with the level of UDP-GlcNAc. OGT and *O*-GlcNAc appear to be nutrient sensors, since UDP-GlcNAc synthesis requires and/or responds to glucose, amino acid, fatty acid, and nucleotide metabolism (for review, see Wells *et al*., 2003).

5. Physiological function

O-GlcNAc dysregulation has been implicated in disease pathologies such as Alzheimer's disease, cancer, and Type II Diabetes (for review, see Whelan and Hart,

2003). The etiology of Alzheimer's disease involves the formation of neurofibrillary tangles due to the hyperphosphorylation of the microtubule associated protein, tau. Under normal conditions, tau is modified by *O*-GlcNAc. It is hypothesized that *O*-GlcNAc modification protects against hyperphosphorylation and the etiology of Alzheimer's disease.

Primary normal breast tissue compared to breast cancer tissue was shown to have decreased *O*-GlcNAc levels and increased *O*-GlcNAcase activity. More specifically, the proto-oncogene c-Myc mutational hot spot (Thr58) for Burkitt's lymphoma is also the major *O*-GlcNAc modification site. Induction of insulin resistance, the hallmark of type II diabetes, by glucose, glucosamine, or by PUGNAc inhibition of *O*-GlcNAcase, results in hyper *O*-GlcNAc modification of many proteins, including downstream insulin signaling proteins IRS-1, IRS-2, PI3K, and β-catenin. The ongoing development of proteomic strategies has helped to define functional roles for *O*-GlcNAc on a variety of proteins and disease states (Whelan and Hart, 2003).

O-GlcNAc modification is known to have functional roles in protein–protein interactions and degradation. For example, *O*-GlcNAc-modified Sp1 appears to inhibit its binding to TATA-binding protein-associated factor II 110 and decreases Sp1 transcriptional activity on certain promoters (Yang *et al*., 2001). The Rpt2 ATPase in the mammalian proteasome 19S cap is modified by *O*-GlcNAc and as the modification increases, the proteasome function in the degradation of substrates decreases (Zhang *et al*., 2003).

6. *O*-GlcNAc proteomics

The enormous dynamic range of protein expression and multiple posttranslational modifications on one protein combinatorially has made the study of cellular responses regulation very complex. As with phosphorylation, there are a number of proteomic approaches that have been developed to facilitate the characterization of *O*-GlcNAc (Greis and Hart, 1998; Gygi, 1999; Wells *et al*., 2002). The development of *O*-GlcNAc-specific antibodies has significantly increased the efficiency of identifying *O*-GlcNAc proteins. Before the use of these antibodies, galactosyltransferase labeling of *O*-GlcNAc with UDP[^{3}H]Gal was the primary method and still remains the standard for the detection and characterization of *O*-GlcNAc (Roquemore *et al*., 1994).

Affinity chromatography (i.e., CTD 110.6 or sWGA-sepharose) is often used to enrich fractions in order to reduce complexity for mapping sites. The use of β-elimination and Michael addition of an affinity tag to map phosphorylation sites has been adapted to map *O*-GlcNAc sites (Wells *et al*., 2002). The β-elimination and Michael addition of dithiothreitol (BEMAD), not only stabilizes the labile O-linkage during collision induced dissociation by LC-MS/MS mass spectrometry but also allows for further enrichment of deuterated DTT-modified peptides by thiol-Sepharose. Both phosphorylation and *O*-GlcNAc-modified sites may be simultaneously quantified with selective use of enzymes, in conjunction with differential labeling of DTT and deuterated DTT. Continued development of sensitive proteomic techniques is necessary to elucidate the biological significance

of the site-specific, dynamic, and global changes of *O*-GlcNAc, together with other posttranslational modifications in the proteome.

Acknowledgments

Original research is funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under Contract No. N01-HV-28180 and by NIH DK61671 to G.W.H. Under a licensing agreement between Covance Research Products and The Johns Hopkins University, G.W.H. receives a share of royalties received by the university on sales of the CTD 110.6 antibody. The terms of this arrangement are managed by The Johns Hopkins University in accordance with its conflict-of-interest policies.

References

Dong DL and Hart GW (1994) Purification and characterization of an O-GlcNAc selective N-acetyl-beta-D-glucosaminidase from rat spleen cytosol. *The Journal of Biological Chemistry*, **269**, 19321–19330.

Greis KD and Hart GW (1998) Analytical methods for the study of O-GlcNAc glycoproteins and glycopeptides. *Methods in Molecular Biology*, **76**, 19–33.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **10**, 994–999.

Haltiwanger RS, Holt GD and Hart GW (1990) Enzymatic addition of O-GlcNAc to nuclear and cytoplasmic proteins. Identification of a uridine diphospho-N-acetylglucosamine:peptide beta-N-acetylglucosaminyltransferase. *The Journal of Biological Chemistry*, **265**, 2563–2568.

Hanover JA, Yu S, Lubas WB, Shin SH, Ragano-Caracciola M, Kochran J and Love DC (2003) Mitochondrial and nucleocytoplasmic isoforms of O-linked GlcNAc transferase encoded by a single mammalian gene. *Archives of Biochemistry and Biophysics*, **409**, 287–297.

Hart GW (1997) Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annual Review of Biochemistry*, **66**, 315–335.

Holt GD and Hart GW (1986) The subcellular distribution of terminal N-acetylglucosamine moieties. Localization of a novel protein-saccharide linkage, O-linked GlcNAc. *The Journal of Biological Chemistry*, **261**, 8049–8057.

Iyer SP and Hart GW (2003) Roles of the tetratricopeptide repeat domain in O-GlcNAc transferase targeting and protein substrate specificity. *The Journal of Biological Chemistry*, **278**, 24608–24616.

Kreppel LK, Blomberg MA and Hart GW (1997) Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *The Journal of Biological Chemistry*, **272**, 9308–9315.

Kreppel LK and Hart GW (1999) Regulation of a cytosolic and nuclear O-GlcNAc transferase. Role of the tetratricopeptide repeats. *The Journal of Biological Chemistry*, **274**, 32015–32022.

Lubas WA, Frank DW, Krause M and Hanover JA (1997) O-Linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *The Journal of Biological Chemistry*, **272**, 9316–9324.

Roquemore EP, Chou TY and Hart GW (1994) Detection of O-linked N-acetylglucosamine (O-GlcNAc) on cytoplasmic and nuclear proteins. *Methods in Enzymology*, **230**, 443–460.

Schultz J and Pils B (2002) Prediction of structure and functional residues for O-GlcNAcase, a divergent homologue of acetyltransferases. *FEBS Letters*, **529**, 179–182.

Shafi R, Iyer SP, Ellies LG, O'Donnell N, Marek KW, Chui D, Hart GW and Marth JD (2000) The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic

stem cell viability and mouse ontogeny. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5735–5739.

Torres CR and Hart GW (1984) Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *The Journal of Biological Chemistry*, **259**, 3308–3317.

Wells L, Vosseller K, Cole RN, Cronshaw JM, Matunis MJ and Hart GW (2002) Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Molecular & Cellular Proteomics*, **1**, 791–804.

Wells L, Vosseller K and Hart GW (2003) A role for N-acetylglucosamine as a nutrient sensor and mediator of insulin resistance. *Cellular and Molecular Life Sciences*, **60**, 222–228.

Whelan SA and Hart GW (2003) Proteomic approaches to analyze the dynamic relationships between nucleocytoplasmic protein glycosylation and phosphorylation. *Circulation Research*, **93**, 1047–1058.

Yang X, Su K, Roos MD, Chang Q, Paterson AJ and Kudlow JE (2001) O-linkage of N-acetylglucosamine to Sp1 activation domain inhibits its transcriptional capability. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6611–6616.

Yang X, Zhang F and Kudlow JE (2002) Recruitment of O-GlcNAc transferase to promoters by corepressor mSin3A: coupling protein O-GlcNAcylation to transcriptional repression. *Cell*, **110**, 69–80.

Zhang F, Su K, Yang X, Bowe DB, Paterson AJ and Kudlow JE (2003) O-GlcNAc modification is an endogenous inhibitor of the proteasome. *Cell*, **115**, 715–725.

Short Specialist Review

Posttranslational modifications to plants – glycosylation

Yoram Tekoah
Ben-Gurion University of the Negev, Beer-Sheva, Israel

1. Introduction

Like other eukaryotic organisms, plants have glycoproteins with *N*- and *O*-glycans (see Article 64, **Structure/function of *N*-glycans**, Volume 6 and Article 65, **Structure/function of *O*-glycans**, Volume 6). *N*-Glycans are attached to asparagine residues and O-linked glycans are attached to hydroxyproline, serine, or threonine residues. Only a few plant proteins have been described with both types of glycans. Animal and plant cells have a similar capability to assemble protein subunits but the systems are not identical as far as posttranslational modifications (PTMs) are concerned (Houdebine, 2002). The complex *N*-glycans of plants differ markedly in structure from those of animals. They more closely resemble those of insects and molluscs (Figure 1). Plant *N*-glycans can be divided into two main types: high mannose type and complex type. The high mannose types are derived from the original $Man_9GlcNAc_2$ structure common to plant and animal glycans. The general formula $Man_{5-9}GlcNAc_2$ represents glycans that are not fully processed. The complex type glycans may contain an α-(1,3)-fucose (Fuc) linked to the proximal GlcNAc residue and/or a β-(1,2)-xylose (Xyl) residues attached to the α-linked mannose (Man) of the glycan core. This glycan type is named *paucimannosidic* – type *N*-glycan (Lerouge *et al*., 1998). In plants, these additional N-linked glycans are considered antigenic (Faye *et al*., 1993) and/or allergenic (van Ree *et al*., 2000). In various cases additional modifications such as additional mannose or GlcNAc sugars are found. Plant glycans are not found to contain sialic acid residues that are present in mammalian glycans.

2. N-linked glycans

The biosynthesis of N-linked high-mannose glycans in plants is identical to that in yeast and mammals and involves the en bloc transfer of the $Glc_3Man_9GlcNAc_2$ oligosaccharide from dolichylpyrophosphate to specific asparagine residues constitutive of the N-glycosylation sequences, Asn-X-Ser/Thr (where X is any amino acid

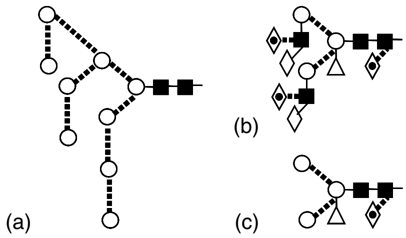


Figure 1 Diagrammatic representation of structures of various *N*-linked glycans found in plant glycoproteins. (a) High mannose type, (b) complex type, and (c) paucimannosidic type glycan. The symbols of the glycan structures are: mannose: clear circle; GlcNAc: dark square; galactose: diamond; fucose: diamond with a dot; xylose: triangle; |: 1–2 linkage; /: 1–3 linkage; –: 1–4 linkage; \: 1–6 linkage; solid line: β-linkage; dotted line: α-linkage

except Pro) in the nascent polypeptide chains. Once transferred onto the nascent protein and while the glycoprotein is transported along the secretary pathway, the oligosaccharide N-linked to Asn (*N*-glycan) undergoes several maturation steps involving the removal and the addition of sugar residues in the endoplasmic reticulum (ER) and the Golgi complex. The removal of glucose residues is achieved by glucosidase I and II in the lumen of the ER. This is then followed by transport of the protein to the Golgi. In the late Golgi apparatus, plant and mammalian *N*-glycan maturations are found to differ. In the Golgi, additional modification of the high-mannose glycans that are accessible to the various modifying enzymes that exist in the Golgi, is achieved to produce complex glycans. These modifications resemble the process found in mammalian cells. The process is initiated by the removal of four mannose residues by mannosidase I. It then continues by the attachment of a peripheral GlcNAc by GlcNAc transferase I (GlcNAcT-I) to the α-1–3-linked mannose attached to the central β-1–4-linked mannose. This is followed by the removal of two more mannose residues by mannosidase II, and attachment of a second peripheral GlcNAc residue. Only after the addition of at least one GlcNAc a β-1,2-xylose residue is transferred by a β-1,2-xylosyltransferase (β-1,2-XylT) to the β-linked mannose of the core. As an independent event, an α-1,3-fucose residue is added to the proximal GlcNAc of the chitobiose core by α-1,3-fucosyltransferase (α1,3-FucT) (Lerouge *et al*., 1998). These events occur in the medial and trans Golgi cistenae respectively (Fitchette-Laine *et al*., 1994). It is important to note that mammalian glycoproteins have not been reported to contain α-1,3-fucose or β-1,2-xylose residues. However, similar constituents and epitopes are reported on the glycoproteins of lower animals (e.g., molluscs, insects, and spiders) (Staudacher *et al*., 1999). In some cases, plant glycoproteins are found to be immunogenic when injected into mammals. It is likely that the α-1,3-fucose and β-1,2-xylose residues account for this immunogenicity, and the potential for inducing the allergenic characteristic of plants in mammals (Tekoah *et al*., 2004), although it is sometimes thought that the fucose alone is the main problematic epitope (Staudacher *et al*., 1999). In other cases, no glycan specific antibodies were produced in mice that were injected with plant produced antibodies (Chargelegue *et al*., 2000). After the addition of the second GlcNAc residue and the fucose and xylose substitutions, the glycan can be

either transported to the vacuole or can be further modified. The transport of such glycoproteins to the vacuole invariably results in the removal of the peripheral GlcNAc residues, because vacuoles contain *N*-acetylglucosaminidases (Lerouge *et al*., 1998). Similarly, high-mannose glycans that escaped the action of mannosidase I in the Golgi may be further trimmed since the vacuole contains also α-mannosidase. Thus, we can find that complex glycans on vacuolar proteins are rather small and have the structure (Xyl)Man_3GlcNAc(Fuc)GlcNAc. This structure, called *paucimannose*, can be found also in insects. In the case of hybrid type glycans one can encounter structures such as Man_4Xyl(Fuc)GlcNAc or Man_5Xyl(Fuc)GlcNAc. If the protein is not transferred to the vacuole but rather secreted, then the glycoproteins retain their peripheral GlcNAc residues because the cell wall has little or no *N*-acetylglucosaminidase activities. At this stage, the complex type *N*-glycans can be further processed by the addition of a terminal fucose residue by an α-1,4-fucosyltransferase (FucT) and a galactose residue by a β-1,3-galactosyltransferase (GalT) (Lerouge *et al*., 1998). This leads to a structure, Gal-β-1,3 (Fuc α-1,4) GlcNAc, known as *Lewis*^{a} (Le^a). This Le^a antigen that has been detected only in some plant proteins, is also found on cell surface glycoconjugates in mammalian systems and is usually involved in cell–cell recognition (Chen *et al*., 2005).

To date, peptide-N^4-(*N*-acetyl-β-glucosaminyl) asparagine amidase A (PNGase A) from almond emulsin is the only *N*-glycanase commercially available that is able to release N-linked glycans that contain α-1,3-core-fucose (Tretter *et al*., 1991). Since it is thought that α-1,3-core-fucose is an epitope involved in pollen and food allergy in addition to the β-1,2-xylose (Tretter *et al*., 1993; Wilson *et al*., 1998), it is most important to use PNGase A rather than peptide *N*-glycanase F (PNGase F) from *Flavobacterium meningosepticum* in order to ensure that all plant glycans have been released when studying glycoproteins from plants. An alternative is to release the glycans chemically with a procedure using hydrazine (Patel *et al*., 1993, see also Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6).

Glycans affect the physicochemical properties of a protein, including resistance to thermal denaturation, protection from proteolytic degradation, and solubility. It can also alter essential biological functions, such as immunogenicity, clearance rate, specific activity, and ligand–receptor interaction. The function of *N*-glycans on plant glycoproteins is to stabilize them against proteolytic degradation. Although a wide variety of glycans exist in plants, the function of complex glycans (as opposed to high-mannose glycans) is not understood. Studies showed that an *Arabidopsis* mutant (*cgl*) that lacks GlcNAc transferase I (GlcNAcT-I) and is unable to convert high-mannose glycans to complex glycans has no visible phenotype when grown under standard laboratory conditions (von Schaewen *et al*., 1993) unlike the case found in mammals (Ioffe and Stanley, 1994).

3. O-linked glycans

O-linked glycosylation can be found in amino acids that contain a hydroxy group (i.e., Ser, Thr, hydroxyproline (Hyp), and hydroxylysine). *O*-Glycans found in mammalian systems are usually linked to Ser or Thr residues of the polypeptide

backbone. These glycan structures can range from simple monosugars (Fuc, Glc, or GalNAc) to several sugars such as mucinlike oligosaccharides attached to core GalNAc linked to Ser/Thr residues. While N-linked glycans involve a single-step cotranslational addition of the N-linked oligosaccharide precursor on Asn residues followed by additional modifications in the ER and Golgi apparatus, O-glycosylation of Ser/Thr residues is a posttranslational multiple-step procedure in mammals. Like *N*-glycan maturation, *O*-glycans are transferred by glycosyltransferases from nucleotide-sugar donors onto *O*-glycans in the Golgi apparatus.

Very little information is yet available on mammalian-like *O*-glycans in plants (Faye *et al*., 2005). Some studies of O-linked glycans attached to Ser or Thr residues of plant glycoproteins have been reported. An O-glycosylated Ser residue with a single galactose was found in the cell wall glycoprotein extensin (Cho and Chrispeels, 1976). A similar monogalactosylation of Ser residues was found in proteins such as sweet potato sporamin (Matsuoka *et al*., 1995). Various proteins from rice have been shown to have mammalian mucin-type glycosylation (Mitsui *et al*., 1990; Kishimoto *et al*., 1999). In a study of glycosylation of therapeutic proteins expressed in plants, up to six proline/hydroxyproline conversions and variable amounts of arabinosylation (Pro/Hyp + Ara) were found in the hinge region of maize-expressed *h*IgA1 heavy chain (HC) (Karnoup *et al*., 2005). In some cases plants have been found to contain an additional O-linked *N*-acetylglucosamine (O-GlcNAc) similar to that found in mammalian proteins and is known as a key regulator of nuclear and cytoplasmatic protein activity (Love and Hanover, 2005).

4. Transgenic plants and glycosylation

Lately, it has been shown that plants can be used as a production system for recombinant mammalian proteins (Faye *et al*., 2005). The majority of therapeutic proteins require PTMs such as glycosylation to be effective. Nevertheless, if plants are to be used to produce mammalian proteins which are glycoproteins, they will usually be found to carry plantlike glycosylation that is usually highly immunogenic and potentially allergenic. In this case, it is most important to understand the structure of the glycans attached to the proteins to evaluate the potential use of these proteins (Tekoah *et al*., 2004). One of the attractive uses of transgenic plants is for the production of recombinant pharmaceuticals such as antibodies or lysosomal enzymes that are needed in large quantities (molecular pharming). Production in plants has various advantages such as the lowering of production costs of the therapeutic proteins compared to alternative methods. Projects are already under way to produce a number of mammalian proteins in plants, with clinical trials set to begin in the near future. Since glycosylation differs in plants and mammals, much work has been done recently and is still needed to overcome this drawback (Chen *et al*., 2005).

Several approaches can be used to avoid these potential problems. If mammalian proteins are produced in GlcNAcT-I-deficient mutant plants similar to those described by von Schaewen (von Schaewen *et al*., 1993) then all the glycans will be of the Man_5 $GlcNAc_2$ type, and such glycans will not be immunogenic. However, proteins containing mannose-terminating glycans will be rapidly cleared by the

mannose receptor of macrophage cells in mammalian systems. A second approach would be to eliminate (knockout) certain glycosidases and glycosyltransferases that can be problematic and add other specific mammalian glycosyltransferases. For instance, plant cells transformed with mammalian GlcNAcT-I, a mammalian Golgi enzyme, have been shown to function correctly in a plant cellular environment, and thus it should be possible to obtain plants that make *N*-glycans without xylose and with α-1,6-fucose on the proximal GlcNAc (Gomez and Chrispeels, 1994). The pathway for the synthesis of sialic acids does not exist in plants and therefore the production of glycoproteins that contain sialic acid residues would be more difficult. This would need to include enzymes to make this sugar, to attach it to cytidine monoshosphate (CMP). The protein to transport CMP-Sia across the Golgi cisternal membrane would have to be incorporated into the transgenic plants as well as a sialyltransferase. A third approach is to produce the mammalian proteins with carboxy-terminal signals such as amino acid sequences Lys/His-Asp-Glu-Leu (KDEL or HDEL) that cause the proteins to be retained in the ER. Proteins that have been modified in this way have been shown to accumulate to quite a large extent in the ER of leaves. Such proteins typically have high-mannose glycans similar to those of other ER residents (Ko *et al*., 2003; Tekoah *et al*., 2004). Plant systems have several advantages for the production of therapeutic proteins compared to animal systems, in addition to low cost of production. The easy scale-up, and the lack of animal pathogens, is not less important. Future studies will enhance our knowledge regarding the possibilities of "humanizing" plant glycosylation on therapeutic proteins.

5. Glycosylation in algae and production of therapeutic proteins

Very little is known about glycosylation in algae. The diversity of the various species of algae makes it an interesting alternative for production of therapeutic proteins. Furthermore, glycosylation can be similar to that of higher plants, such as is thought to be the case in the green algae *Chlamydomonas reinhardtii*. In other cases, totally different glycosylation patterns may exist, such as in the red algae *Porphyridium* sp. (Tekoah, unpublished data). Various studies have shown the potential of the production of such proteins in *C. reinhardtii* (Mayfield *et al*., 2003; Mayfield and Franklin, 2005). The benefits of utilizing algae for the production of therapeutic proteins are the same as in plants. In addition, the high growth rate, and the contained and controlled environment of growth and diversity of algae, can enhance these benefits. All this will not be feasible without the proper future study of glycosylation in algae that will help in the understanding of the glycosylation pathways, both in plants and algae, and help set potential goals for "humanization" of glycosylation.

6. Summary

The study of glycosylation in plants and algae has still to be fully elucidated. Plant genetic engineering can lead to the economical production of therapeutic

proteins, which are safe and a viable alternative to current methods of production in animal systems. Findings indicate the benefits that may arise from modifications of glycosylation of these proteins, using plant genetic engineering. But the importance of studying the glycosylation in therapeutic proteins to monitor the modifications of the glycans is important for ruling out potential harmful PTMs of glycans (Tekoah *et al*., 2004).

References

Chargelegue D, Vine ND, van Dolleweerd CJ, Drake PMW and Ma JK-C (2000) A murine monoclonal antibody produced in transgenic plants with plant-specific glycans is not immunogenic in mice. *Transgenic Research*, **9**(3), 187–194.

Chen M, Liu X, Wang Z, Song J, Qi Q and Wang PG (2005) Modification of plant N-glycans processing: the future of producing therapeutic protein by transgenic plants. *Medicinal Research Reviews*, **25**(3), 343–360.

Cho Y-P and Chrispeels MJ (1976) Serine-O-galactosyl linkages in glycopeptides from carrot cell walls. *Phytochemistry*, **15**(1), 165–169.

Faye L, Boulaflous A, Benchabane M, Gomord V and Michaud D (2005) Protein modifications in the plant secretory pathway: current status and practical implications in molecular pharming. *Vaccine*, **23**(15), 1770–1778.

Faye L, Gomord V, Fitchette-Laine AC and Chrispeels MJ (1993) Affinity purification of antibodies specific for Asn-linked glycans containing α 1-3 fucose or β 1-2xylose. *Analytical Biochemistry*, **109**, 104–108.

Fitchette-Laine A-C, Gomord V, Chekkafi A and Faye L (1994) Distribution of xylosylation and fucosylation in the plant Golgi apparatus. *The Plant Journal*, **5**(5), 673–682.

Gomez L and Chrispeels M (1994) Complementation of an arabidopsis thaliana mutant that lacks complex asparagine-linked glycans with the human cDNA encoding N-acetylglucosaminyl-transferase I. *The Proceedings of the National Academy of Sciences*, **91**(5), 1829–1833.

Houdebine LM (2002) Antibody manufacture in transgenic animals and comparisons with other systems. *Current Opinion in Biotechnology*, **13**(6), 625–629.

Ioffe E and Stanley P (1994) Mice lacking N-acetylglucosaminyltransferase I activity die at mid- gestation, revealing an essential role for complex or hybrid N-linked carbohydrates. *The Proceedings of the National Academy of Sciences*, **91**(2), 728–732.

Karnoup AS, Turkelson V and Anderson WHK (2005) O-Linked glycosylation in maize-expressed human IgA1. *Glycobiology*, **15**(10), 965–981.

Kishimoto T, Watanabe M, Mitsui T and Hori H (1999) Glutelin basic subunits have a mammalian mucin-type o-linked disaccharide side chain. *Archives of Biochemistry and Biophysics*, **370**(2), 271–277.

Ko K, Tekoah Y, Rudd PM, Harvey DJ, Dwek RA, Spitsin S, Hanlon CA, Rupprecht C, Dietzschold B, Golovkin M, *et al*. (2003) Function and glycosylation of plant-derived antiviral monoclonal antibody. *The Proceedings of the National Academy of Sciences*, **100**(13), 8013–8018.

Lerouge P, Cabanes-Macheteau M, Rayon C, Fischette-Lainé A-C, Gomord V and Fayé L (1998) N-Glycoprotein biosynthesis in plants: recent developments and future trends. *Plant Molecular Biology*, **38**(1-2), 31–48.

Love DC and Hanover JA (2005) The hexosamine signaling pathway: deciphering the "O-GlcNAc Code" 10.1126/stke.3122005re13. *Science's STKE [electronic resource]: Signal Transduction Knowledge Environment*, **2005**(312), 13.

Matsuoka K, Watanabe N and Nakamura K (1995) O-glycosylation of a precursor to a sweet potato vacuolar protein, sporamin, expressed in tobacco cells. *The Plant Journal*, **8**(6), 877–889.

Mayfield SP and Franklin SE (2005) Expression of human antibodies in eukaryotic micro-algae. *Vaccine*, **23**(15), 1828–1832.

Mayfield SP, Franklin SE and Lerner RA (2003) Expression and assembly of a fully active antibody in algae. *The Proceedings of the National Academy of Sciences*, **100**(2), 438–442.

Mitsui T, Kimura S and Igaue I (1990) Isolation and characterization of golgi membranes from suspension-cultured cells of rice (Oryza-Sativa-L). *Plant and Cell Physiology*, **31**(1), 15–25.

Patel T, Bruce J, Merry A, Bigge C, Wormald M, Jaques A and Parekh R (1993) Use of hydrazine to release in intact and unreduced form both N- and O-linked oligosaccharides from glycoproteins. *Biochemistry*, **32**(2), 679–693.

Staudacher E, Altmann F, Wilson IBH and Marz L (1999) Fucose in N-glycans: from plant to man. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1473**(1), 216–236.

Tekoah Y, Ko K, Koprowski H, Harvey DJ, Wormald MR, Dwek RA and Rudd PM (2004) Controlled glycosylation of therapeutic antibodies in plants. *Archives of Biochemistry and Biophysics*, **426**(2), 266–278.

Tretter V, Altmann F, Kubelka V, Marz L and Becker WM (1993) Fucose alpha 1,3-linked to the core region of glycoprotein N-glycans creates an important epitope for IgE from honeybee venom allergic individuals. *International Archives of Allergy and Immunology*, **102**(3), 259–266.

Tretter V, Altmann F and Marz L (1991) Peptide-N4-(N-acetyl-beta-glucosaminyl)asparagine amidase F cannot release glycans with fucose attached α 1-3 to the asparagine-linked N-acetylglucosamine residue. *European Journal of Biochemistry*, **199**(3), 647–652.

van Ree R, Cabanes-Macheteau M, Akkerdaas J, Milazzo JP, Loutelier-Bourhis C, Rayon C, Villalba M, Koppelman S, Aalberse R, Rodriguez R, *et al*. (2000) Beta(1,2)-xylose and alpha(1,3)-fucose residues have a strong contribution in IgE binding to plant glycoallergens. *The Journal of Biological Chemistry*, **275**(15), 11451–11458.

von Schaewen A, Sturm A, O'Neill J and Chrispeels MJ (1993) Isolation of a mutant arabidopsis plant that lacks N-acetyl glucosaminyl transferase I and is unable to synthesize golgi-modified complex N-linked glycans. *Plant Physiology*, **102**(4), 1109–1118.

Wilson IB, Harthill JE, Mullin NP, Ashford DA and Altmann F (1998) Core α1,3-fucose is a key part of the epitope recognized by antibodies reacting against plant N-linked oligosaccharides and is present in a wide variety of plant extracts. *Glycobiology*, **8**(7), 651–661.

Protein phosphorylation analysis – a primer

Judith A. Jebanathirajah
Harvard Medical School, Boston, MA, USA
MDS Sciex, Concord, ON, Canada

Hanno Steen
Children's Hospital and Harvard Medical School, Boston, MA, USA

1. Introduction

Phosphorylation is one of the most prevalent and important intracellular protein modifications. It is estimated that at any given time approximately 30% of all proteins are phosphorylated (Hunter, 1998). The fast reversibility of protein phosphorylation makes it very attractive for regulating cellular processes as diverse as metabolism, signaling, and cell proliferation. Numerous amino acid residues can be phosphorylated including aspartic acid, cysteine, and histidine. However, major research efforts so far have been focused on serine, threonine, and tyrosine phosphorylation because these amino acid residues are the main target for phosphorylation in eukaryotes. In addition, they are chemically more stable and thus considerably easier to analyze (Yan *et al*., 1998). Phosphorylation analysis is composed of several steps: (1) the detection of a known phosphorylation site, (2) residue-resolved localization of (novel) phosphorylation sites, (3) monitoring changes in phosphorylation (relative quantitation), and (4) determining the degree of phosphorylation (absolute quantitation of phosphosylation stoichiometry).

Common strategies for the analysis of protein phosphorylation utilize radioactive phosphorus isotopes, phosphospecific or phosphorylation site specific antibodies, and/or mass spectrometry (MS). Each method has its advantages and disadvantages for phosphorylation analysis, which will be summarized in the following section of this article.

2. Radioactive labeling

A well-established method for the analysis of protein phosphorylation is the incorporation of the radioactive phosphorus isotopes ^{32}P or ^{33}P. In order to detect a particular phosphorylation site, this method is combined with phosphopeptide

mapping by two-dimensional thin layer chromatography (2D-TLC) or liquid chromatography (LC) followed by radioactivity measurements to identify the spots/fractions of interest (Yan *et al*., 1998). However, this approach requires that a detailed analysis of the different phosphopeptides is carried out in order to unambiguously correlate a spot/fraction with a particular phosphorylation site. Since the detection of radioactivity is quantitative this method is suitable for relative quantitation, but not necessarily for absolute quantitation as the initial amount of protein is not always known.

To utilize radioactive labeling for residue-resolved localization of phosphorylation sites, Edman degradation–based approaches have proven to be useful. After isolating and purifying the phosphopeptides by, for example, 2D-TLC or LC, the identity of the peptide is determined by MS. This is followed by Edman degradation where the peptide is sequentially hydrolyzed and the release of radioactivity during one particular cycle from the immobilized peptide can be correlated with the primary sequence as determined by MS prior to the Edman degradation. The sequence of events described above identifies the site of phosphorylation (Campbell and Morrice, 2002).

True "de novo" sequencing of phosphopeptides by Edman degradation is problematic as the Phenylthiohydantoin (PTH) derivative of phosphotyrosine is barely soluble (Aebersold *et al*., 1991) and phosphoesters of threonine and serine undergo β-elimination under the conditions used for Edman sequencing. This side reaction gives rise to numerous undefined products such that phosphoserine or phosphothreonine residues are often only assigned with ambiguity (Campbell *et al*., 1986; Dedner *et al*., 1988).

Radioactive labeling is very sensitive and well suited for in vitro phosphorylation assays. However, in vivo incorporation of the radioactive isotopes is not possible in the case of tissue samples or is very inefficient in the case of cell culture due to the presence of endogenous unlabeled adenosine triphosphate (ATP). Hence, to achieve a degree of in vivo phosphorylation that is sufficient for sensitive detection, very large amounts of radioactively labeled ATP are required such that other pleiotropic effects cannot be excluded.

3. Phosphospecific antibodies

Three different types of antibodies are used for protein phosphorylation analysis:

1. *Anti-phosphoamino acid antibodies:* Anti-phosphotyrosine antibodies were invaluable for the elucidation of phosphotyrosine-based mitogen and cytokine signaling. They can be used for profiling the global tyrosine phosphorylation state of a particular sample after 1D or 2D gel analysis (*see* Article 22, **Two-dimensional gel electrophoresis**, Volume 5; Figure 1b), and for the enrichment of tyrosine-phosphorylated proteins by immunoprecipitation (Pandey *et al*., 2000). Although antibodies are mainly used for proteins, a recent study showed that these antibodies are also applicable to the enrichment of tyrosine-phosphorylated peptides (Rush *et al*., 2005). The development of good anti-phosphoserine and anti-phosphothreonine antibodies has been less successful so far (Gronborg *et al*., 2002).

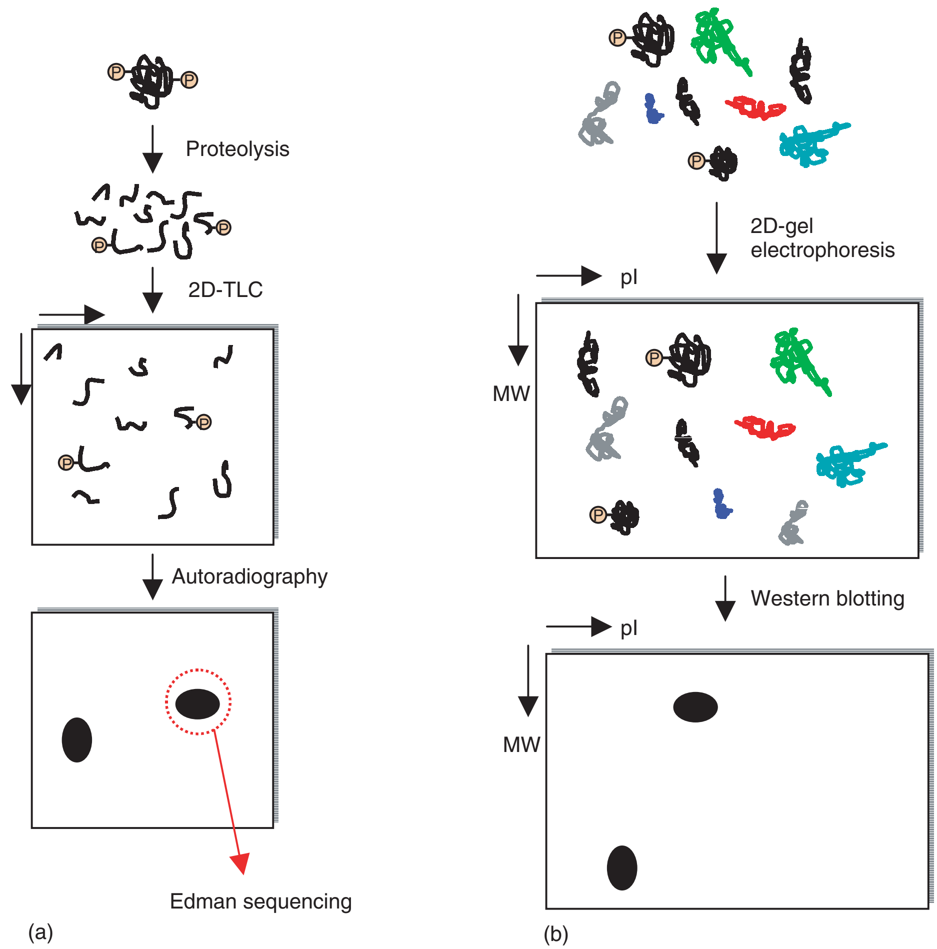


Figure 1 (a) A schematic portraying the radioactivity-based analysis of phosphorylated peptides involving proteolysis, two-dimensional thin layer chromatography (2D-TLC), and autoradiography to visualize the radioactively labeled peptides prior to Edman sequencing to determine the amino acid sequence. (b) A schematic for the identification of, for example, tyrosine-phosphorylated peptides using western blotting with anti-phosphotyrosine antibodies subsequent to two-dimensional gel electrophoresis (2D-GE). 2D-GE separates proteins in the first dimension on the basis of the isoelectric point (pI) and in the second dimension, on the basis of molecular weight

2. *Antibodies against phosphorylation consensus motifs:* These antibodies are raised against oriented peptide libraries mimicking particular phosphorylation consensus motifs (available from e.g., Cell Signaling Technology). They are a suitable compromise for serine and threonine phosphorylation motifs and can be used in a fairly broad context as well as for specific questions (Zhang *et al*., 2002a).
3. *Phosphorylation site–specific antibodies:* As antibodies can be the most sensitive, specific, and fastest way to detect a particular protein phosphorylation in complex protein mixtures, such as whole cell lysates, the use of these types of antibodies is becoming more widespread. However, localization of

the phosphorylation site by other means such as MS, Edman degradation with radioactive labeling is a requirement because the unbiased localization of novel phosphorylation sites using site specific antibodies is not possible.

Another advantage of antibodies is the semiquantitative nature of the Western blotting such that relative (semi-)quantitation is possible. However, absolute quantitation is impossible if one only uses phosphospecific antibodies. Other limitations of the use of antibodies stem from the fact that it is not always possible to raise antibodies against all phosphorylation sites and that specificities can be questionable in cases where multiple phosphorylation sites are found in close proximity.

4. Mass spectrometry

Mass spectrometry is very commonly used for the analysis of protein phosphorylation especially due to its improved sensitivity and speed as compared to many traditional biochemical methods such as Edman degradation. Another major advantage of MS is its versatility – that is, it can be used for any kind of phosphorylation analysis. If the identity of a protein and its phosphorylation sites are known, then directed mass spectrometric experiments can be performed in order to determine whether a particular site is phosphorylated or not as the mass-to-charge ratio of the expected proteolytic phosphopeptide can be calculated.

1. Relative quantitation of protein phosphorylation can be performed using stable isotope labeling approaches using metabolic, enzymatic, or chemical means to incorporate the label (Ibarrola *et al*., 2003; Oda *et al*., 1999). Alternatively, stable isotope-free methods can be used to monitor changes in the degree of phosphorylation after normalizing the raw data to account for run-to-run variations and variations in the starting amount of material (Ruse *et al*., 2002; Steen *et al*., 2005).
2. Absolute quantitation of the phosphorylation stoichiometry can be achieved by using either isotopically labeled internal standards (Stemmann *et al*., 2001), isotope labeling in combination with phosphatase treatment (Zhang *et al*., 2002b), or a completely stable isotope-free methods (Steen *et al*., 2005).

The true power of MS is its ability to sequence peptides in a second to subsecond timescale and to unambiguously localize phosphorylation sites down to the residue level in an unbiased manner (Figure 3a and b; Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5). Different strategies have been devised to identify and analyze the phosphorylated peptides within a complex mixture such as proteolytic protein digests.

The most successful strategies for mass spectrometric analysis of protein phosphorylation is to utilize either selective enrichment of phosphorylated species utilizing, for example, generic antibodies (see above) or immobilized metal affinity chromatography (IMAC; see below) approaches (Figure 2) (Posewitz and Tempst, 1999) and/or utilize the characteristic fragmentations of phosphopeptides upon

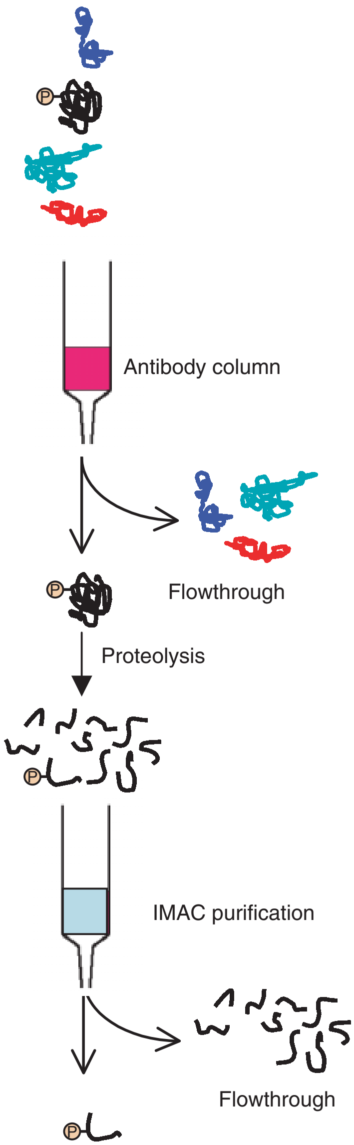


Figure 2 A general strategy using antibodies and immobilized metal affinity chromatography (IMAC) to enrich for phosphoproteins and subsequently phosphopeptides. Each of the enrichment steps can also be used separately

collisionally induced dissociation (CID) for so-called (constant) neutral loss and/or precursor ion experiments (Covey *et al*., 1991; Huddleston *et al*., 1993; *see also* Article 18, **Techniques for ion dissociation (fragmentation) and scanning in MS/MS**, Volume 5).

In general, it is always advisable to reduce the complexity of the sample as much as possible by enriching the protein(s) of interest or even the phosphorylated form(s)

of the protein(s) of interest using, for example, antibodies. However, it should be noted that quantitation of the phosphorylation stoichiometry is not possible if only the phosphorylated species (protein or peptide) is enriched. Following the enrichment of the protein(s) of interest, the sample is digested generating a complex mixture of peptides some of them being modified, that is, phosphorylated, but the vast majority being unmodified. The analytical challenge is to maximize the time spent on the analysis of the phosphopeptides and to minimize the time spent on the unmodified peptides. However, a bias against substoichiometric species of low abundance has to be expected when simple data dependent acquisition (DDA) routines are used for the acquisition of data since the peptides are selected for sequencing simply on the basis of the signal intensities of the peptide in the survey scan. This is less of a problem for samples of fairly low complexity, for example, single protein digests, as long as the number of peptides eluting off the LC column does not exceed the number of peptides that can be analyzed with these DDA routines. Thus, although a simple DDA experiment with a standard LC setup is valid as a first screen/first pass experiment, one must exercise caution. If phosphopeptides of very low abundance are expected or the sample is highly complex, such that at any given time during an LC/MS experiment more peptides elute than can be analyzed using DDA routines, numerous phosphopeptides might not be sequenced. Then, either enrichment protocols aimed at enriching the phosphopeptides or selective detection methods are advisable to efficiently utilize the valuable measuring time for analysis of the species of interest.

4.1. Selective detection

The phosphopeptide-specific mass spectrometric detection is based on characteristic fragmentations of phosphopeptides when fragmenting them in tandem mass spectrometers. For instance, in the negative ion mode (which is generally not suitable for peptide sequencing), a highly characteristic fragment ion at *m/z* −79 corresponding to PO_3^- is observable. In contrast, in the positive ion mode, a characteristic loss of H_3PO_4 (98 Da) is often prevalent for phosphoserine- and threonine-containing peptides, but is not commonly observed for tyrosine-phosphorylated peptides. The novel linear ion trap instruments (QTRAP or LTQ) are especially useful in utilizing these phosphopeptide-specific features on an LC-compatible timescale (Beausoleil *et al*., 2004; Le Blanc *et al*., 2003).

4.2. Selective enrichment

The most commonly used enrichment method is based on IMAC, which selectively binds phosphopeptides. Commercially available and in-house-made IMAC materials are commonly used and have all shown promising results. However, because of the mode of binding, it is expected that a certain number of acidic peptides, that is, aspartic acid and glutamic acid–rich peptides will be enriched as well. This problem of nonspecificity is more prevalent with highly complex samples (e.g., whole cell lysates) such that esterification of the carboxyl groups is recommended in those cases (Brill *et al*., 2004; Ficarro *et al*., 2002).

Other enrichment methods that are easily implemented are based on strong cation exchange (SCX) (Beausoleil *et al*., 2004) or strong anion exchange materials (SAX) (Nuhse *et al*., 2003). Ion exchange chromatography approaches utilize the low/high affinity of phosphopeptides to the respective ion exchange material. Both studies show impressive results using highly complex peptide mixtures as starting material. However, the binding/nonbinding of both the SCX and SAX approaches rely on the net charge of the peptide, that is, the amino acid composition. Thus, they are biased against certain subsets of peptides, for example, histidine-containing phosphopeptides in the case of SCX-based strategies (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5).

If matrix assisted laser desorption/ionization (MALDI) is employed for the phosphopeptide analysis, then "hot" (dissociation-inducing) matrices such as α-cyano-4-hydroxycinnamic acid should be avoided as they lead to the loss of the phospho moiety from serine and threonine-phosphorylated peptides, reducing the signal intensity of species of interest, that is, the phosphopeptides (*see* Article 14, **Sample preparation for MALDI and electrospray**, Volume 5). Thus, the use of "cold" matrices such as dihydroxybenzoic acid are recommended. Also, the addition of nitric or phosphoric acid has shown to increase the ionization/detection efficiencies of phosphopeptides (Stensballe and Jensen, 2004). If a MALDI mass spectrometer without MS/MS capability is used for the MALDI analysis, phosphopeptides can be selectively identified in peptide mixtures by performing the MALDI MS analysis before and after phosphatase treatment of the sample. The phosphorylated species decreases in weight by 80 Da or multiples thereof upon enzymatic dephosphorylation (Figure 3c). However, for residue-resolved localization of the phosphorylation sites, additional MS/MS experiments on the species of interest are required.

5. Other methods

Instead of using generic antibodies against phosphoamino acid to profile the phosphorylation state of samples after 1D or 2D gel analysis, it is also possible to use a phosphospecific fluorescence dye called Pro-Q Diamond (Molecular Probes) (*see* Article 27, **Detecting protein posttranslational modifications using small molecule probes and multiwavelength imaging devices**, Volume 5). This reagent provides detection limits in the subpicomole range and is compatible with subsequent in-gel digestion and MS analysis protocols. Hence, more detailed analysis of the gel spots of interest is possible. Furthermore, the binding of the fluorescence probe can be used for quantitation up to 3 orders of magnitude (Steinberg *et al*., 2003).

6. Conclusion

The methods described above are only a small selection of a vast array of different strategies used for the analysis of protein phosphorylation. The aim of this tutorial is to give a brief synopsis of the most broadly applicable and promising detection

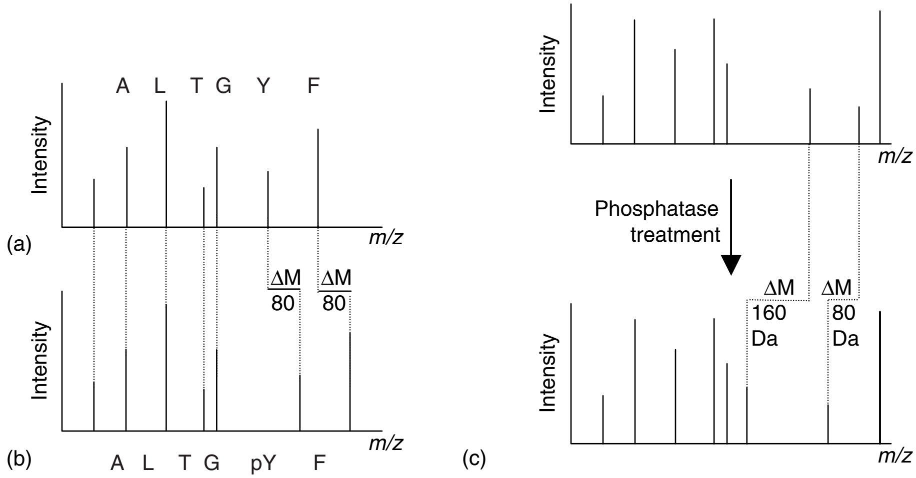


Figure 3 A schematic portraying how to localize phosphorylation sites using mass spectrometric phosphopeptide sequencing. Sequence-revealing peptide fragment ions derived from the phosphopeptide (b), that carry the phospho moiety and thereby localize the site of phosphorylation, are 80 Da heavier than the corresponding fragment ions in the product-ion spectrum of the unmodified cognate (a). (c) Identification of phosphopeptides within a peptide mixture based on the mass shift induced by phosphatase treatment of the peptides. The phosphopeptide masses decrease by 80 Da or multiples thereof depending on the number of phospho moieties enzymatically removed (here 2 and 1 phospho moieties, respectively)

and analytical methods, as well as their strengths and limitations. Experimental details should be explored in the references given. All of the methods have their particular advantages, such that the choice of method should be based on the question asked. An additional implication is that none of the methods is likely to provide a complete picture of the protein phosphorylation of a particular protein. For more comprehensive analysis, several methods should be combined as demonstrated by Vihinen and Saarinen (2000), and Chen *et al*. (2002) who combined five and more different strategies.

Acknowledgments

We would like to thank our former and current colleagues as well as Nick Morrice (MRC Phosphorylation Unit, University of Dundee), John Rush (Cell Signaling Technology) and the organizers and participants of the NSF Plant Protein Phosphorylation Workshop for fruitful discussions.

References

Aebersold R, Watts J, Morrison H and Bures E (1991) Determination of the site of tyrosine phosphorylation at the low picomole level by automated solid-phase sequence analysis. *Analytical Biochemistry*, **199**, 51–60.

Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC and Gygi SP (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins.

Proceedings of the National Academy of Sciences of the United States of America, **101**, 12130–12135.

Brill LM, Salomon AR, Ficarro SB, Mukherji M, Stettler-Gill M and Peters EC (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Analytical Chemistry*, **76**, 2763–2772.

Campbell D, Hardie D and Vulliet P (1986) Identification of four phosphorylation sites in the N-terminal region of tyrosine hydroxylase. *The Journal of Biological Chemistry*, **261**, 10489–10492.

Campbell D and Morrice N (2002) Identification of protein phosphorylation sites by a combination of mass spectrometry and solid phase Edman sequencing. *Journal of Biomolecular Techniques*, **13**, 119–130.

Chen SL, Huddleston MJ, Shou W, Deshaies RJ, Annan RS and Carr SA (2002) Mass spectrometry-based methods for phosphorylation site mapping of hyperphosphorylated proteins applied to Net1, a regulator of exit from mitosis in yeast. *Molecular & Cellular Proteomics*, **1**, 186–196.

Covey T, Shushan B, Bonner R, Schröder W and Hucho F (1991) LC/MS and LC/MS/MS screening for the sites of post-translational modifications in proteins. In *Methods in Protein Sequence Analysis*, Jörnvall H, Höög J-O and Gustavsson A-M (Eds.), Birkhäuser Verlag: Basel, pp. 249–256.

Dedner N, Meyer HE, Ashton C and Wildner GF (1988) N-terminal sequence analysis of the 8 kDa protein in *Chlamydomonas reinhardii*. *FEBS Letters*, **236**, 77–82.

Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF and White FM (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnology*, **20**, 301–305.

Gronborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON and Pandey A (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, frigg, as a protein kinase a substrate. *Molecular & Cellular Proteomics*, **1**, 517–527.

Huddleston MJ, Annan RS, Bean MF and Carr SA (1993) Selective detection of phosphopeptides in complex mixtures by electrospray liquid chromatography / mass spectrometry. *Journal of the American Society for Mass Spectrometry*, **4**, 710–717.

Hunter T (1998) The Croonian lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Philosophical Transactions of the Royal Society of London, Series B*, **353**, 583–605.

Ibarrola N Kalume DE Iwahori M, Gronborg A and Pandey A (2003). A proteomic approach for quantitation of phosphorylation using stable isotope labeling in cell culture. *Analytical Chemistry* **75**, 6043–6049.

Le Blanc JC, Hager JW, Ilisiu AM, Hunter C, Zhong F and Chu I (2003) Unique scanning capabilities of a new hybrid linear ion trap mass spectrometer (Q TRAP) used for high sensitivity proteomics applications. *Proteomics*, **3**, 859–869.

Nuhse TS, Stensballe A, Jensen ON and Peck SC (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Molecular & Cellular Proteomics*, **2**, 1234–1243.

Oda Y, Huang K, Cross FR, Cowburn D and Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6591–6596.

Pandey A, Podtelejnikov AV, Blagoev B, Bustelo XR, Mann M and Lodish HF (2000) Analysis of receptor signaling pathways by mass spectrometry: identification of vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 179–184.

Posewitz MC and Tempst P (1999) Immobilized gallium(III) affinity chromatography of phosphopeptides. *Analytical Chemistry*, **71**, 2883–2892.

Ruse CI, Willard B, Jin JP, Haas T, Kinter M and Bond M (2002) Quantitative dynamics of site-specific protein phosphorylation determined using liquid chromatography electrospray ionization mass spectrometry. *Analytical Chemistry*, **74**, 1658–1664.

Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, Zhang H, Zha XM, Polakiewicz RD and Comb MJ (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature Biotechnology*, **23**, 94–101.

Steen H, Jebanathirajah JA, Springer M and Kirschner MW (2005). Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by mass spectrometry. *Proceedings of the National Academy of Sciences of the United State of America*, In press.

Steinberg TH, Agnew BJ, Gee KR, Leung WY, Goodman T, Schulenberg B, Hendrickson J, Beechem JM, Haugland RP and Patton WF (2003) Global quantitative phosphoprotein analysis using multiplexed proteomics technology. *Proteomics*, **3**, 1128–1144.

Stemmann O, Zou H, Gerber SA, Gygi SP and Kirschner MW (2001) Dual inhibition of sister chromatid separation at metaphase. *Cell*, **107**, 715–726.

Stensballe A and Jensen ON (2004) Phosphoric acid enhances the performance of Fe(III) affinity chromatography and matrix-assisted laser desorption/ionization tandem mass spectrometry for recovery, detection and sequencing of phosphopeptides. *Rapid Communications in Mass Spectrometry*, **18**, 1721–1730.

Vihinen H and Saarinen J (2000) Phosphorylation site analysis of semliki forest virus nonstructural protein 3. *The Journal of Biological Chemistry*, **275**, 27775–27783.

Yan J, Packer N, Gooley A and Williams K (1998) Protein phosphorylation: technologies for the identification of phosphoamino acids. *Journal of Chromatography, A*, **808**, 23–41.

Zhang H, Zha X, Tan Y, Hornbeck PV, Mastrangelo AJ, Alessi DR, Polakiewicz RD and Comb MJ (2002a) Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *The Journal of Biological Chemistry*, **277**, 39379–39387.

Zhang X, Jin QK, Carr SA and Annan RS (2002b) N-terminal peptide labeling strategy for incorporation of isotopic tags: a method for the determination of site-specific absolute phosphorylation stoichiometry. *Rapid Communications in Mass Spectrometry*, **16**, 2325–2332.

Basic Techniques and Approaches

Glycoproteomics

Miriam V. Dwek
University of Westminster, London, UK

Philippa Mills
University College London, London, UK

1. Background

It is now widely accepted that correct protein glycosylation is important for the normal function of many cellular and secreted proteins. Glycosylation has been shown to be necessary for many diverse aspects of protein function, for example, to ensure correct protein folding, to protect proteins from proteolytic attack, as recognition molecules for protein clearance from the serum, and as adhesion molecules. In parallel to the role of glycosylation in normal protein function, the number of examples in which changes occur to the glycosylation of proteins in disease expands. This "basic techniques" article considers some of the issues that would be faced by researchers interested in studying glycosylation of proteins as an adjunct to studying the proteome of a tissue or cell extract.

2. When might glycan analysis be indicated?

The inherent complexity of glycans, compared to that of a linear peptide chain, has meant that analysis of protein glycosylation has been largely neglected in proteomics. Nevertheless, there are circumstances in which glycan analysis might provide useful information as part of a proteomics study.

The standard approach for the identification of (glyco)proteins after separation of the proteome by 1D- or 2D-PAGE is generally by MALDI-MS (matrix-assisted laser desorption ionization mass spectrometry, *see* Article 2, **Sample preparation for proteomics**, Volume 5) fingerprint analysis. The separated proteins are digested, in-gel, with a protease such as trypsin, analyzed on a mass spectrometer (*see* Article 3, **Tandem mass spectrometry database searching**, Volume 5), and the resulting masses are compared with those generated by the theoretical digestion of protein sequences in databases (SWISSPROT and NCBI). Using this approach, a proportion of peptide masses may not be identified as originating from the proteolytic digestion of the protein. These difficulties in protein identification are often explained by the occurrence of posttranslational modifications such

as glycosylation and phosphorylation, which are not predicted by the majority of databases whose sequences, in the main, have been determined from the characterization of cDNAs. These "modified" peptides are often ignored when the sole interest is to identify the protein. Computational tools, such as FindMod, PeptIdent, and Peptide-Mass (**http://www.expasy.ch/tools/**) can calculate the difference between the observed and predicted masses giving a useful insight into the types of possible modifications that may have occurred to the protein. The complexity of protein glycosylation, however, has precluded its inclusion in the majority of database search engines with many only considering the addition of single O-GlcNAc residues. In the case of N-linked glycans, the consensus sequence Asn-Xaa-Ser/Thr (where Xaa is any amino acid other than Pro) allows putative identification of N-linked glycosylation sites. No similar sequence has been determined for O-linked glycosylation and the assumption that is generally made is that any Ser/Thr residue is potentially occupied by O-linked glycans, although an increase in Pro has been observed close to Thr O-link glycosylation sites. To address this matter, three different nine-residue peptides were used to evaluate the ability of recombinant GalNAc transferases to produce glycosylated peptides. From these data, it has been suggested that extended beta-like substrate conformations were favored for successful enzyme–substrate binding, suggesting that these regions of peptides are of some importance when trying to predict occupied O-linked glycosylation sites (Kirnarsky *et al*., 1998). Another approach to identifying likely O-link glycosylation sites has emerged; this utilizes a neural network algorithm and incorporates into an O-glycbase database, which provides free access and information regarding the type of glycan, the glycosylation site, species from which the glycoprotein was derived and so on. The database can be found at **http://www.cbs.dtu.dk/databases/OGLYCBASE/Oglyc.base.html**

3. Staining 2D-PAGE gels to identify glycosylated proteins

A simple starting point to determine whether or not a protein is glycosylated is to use the periodic acid-Schiff (PAS) stain. In this procedure, the gel is incubated in the presence of periodic acid (this oxidizes hydroxyl groups; and converts hydroxyl and amino groups to aldehydes) the aldehydes are then reacted with Schiff reagent (fuschin–sulfuric acid) to give a pink-red colored product. The PAS stain is sensitive to approximately 0.1 μg of protein. Combining the periodic acid reaction with fluorescent dye technology enables the detection of glycoproteins at nanogram sensitivity (Hart, 2003). One of the advantages of using this approach for detection of glycosylated proteins is that it allows the gel to be stained first for glycoprotein detection and later with a "total" protein stain such as Coomassie brilliant blue, silver nitrate, or Sypro ruby. An "overlay" can be produced to allow the identification of glycosylated proteins that may be subjected to further analysis.

4. On-membrane analysis of protein glycosylation

Proteins separated by 2D-PAGE can be transferred to inert membranes, usually made of either nitrocellulose or polyvinylidene difluoride, by a process known

as *Western blotting* (Towbin, 1979). The glycosylation of the proteins can then be examined by detecting the monosaccharides that are present by using carbohydrate-binding proteins, for example, lectins, or antibodies. The sensitivity of detection of glycan residues may be enhanced to nanomol concentrations by the use of biotin conjugated lectins and/or antibodies using streptavidin-horse radish peroxidase, to produce either a colored substrate or a light reaction. Useful data can be obtained with this approach to glycan analysis, particularly when a variety of lectins are used in series.

5. Release of N-linked glycans

Most of the methods for detailed analysis of N-linked glycans involve the removal of the glycan moiety from the protein backbone first, although there have been recent reports of analysis of intact glycoproteins after separation by 2D-PAGE (Emmett, 2003; *see also* Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6). For a glycan analysis strategy to be successful, the protein spot of interest would normally be clearly visible on a Coomassie blue stained gel (i.e., there would be microgram quantities of protein available); full structural analysis is not currently achievable using protein spots detected by silver nitrate or fluorescent dyes, although glycan "fingerprinting" may still be achievable depending on the extent to which the protein is glycosylated.

The most frequently used approach for glycan release is endoglycosidase digestion and several enzymes are commercially available (Harvey, 2001). The most widely used is PNGase-F (peptide *N*-glycosidase F). Enzymatic treatment is favored over chemical methods, such as hydrazinolysis or β-elimination, for the release of N-linked sugars because the glycans are less extensively modified and the protein can also be recovered intact (Harvey, 2001). Various strategies have been developed for in situ enzymatic release of N-linked glycans after separation of the individual glycoforms (Küster *et al*., 2001). Protocols involve the release of the glycan moiety from the protein either directly from excised gel spots or after electroblotting of the glycoproteins onto polyvinylidene difluoride membranes.

6. Release of O-linked glycans

While the techniques for the analysis of N-linked glycans released from glycoproteins are becoming routine, the same is not true for the analysis of O-linked glycans. Unfortunately, there is no enzyme that efficiently releases all O-linked oligosaccharides from the polypeptide backbone; instead, chemical methods such as hydrazinolysis and, more usually, nonreductive β-elimination (Harvey, 2001) or trifluoromethanesulfonic acid (Edge, 2003) are employed (*see* Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6).

7. Glycan analysis

As mentioned above, glycan analysis is usually only attempted for glycoproteins that are available in microgram quantities. Once released, various analytical

techniques are available for the characterization of carbohydrate structures. These include chromatography, electrophoresis, and mass spectrometry (MS). In order to improve sensitivity of detection, prior derivatization is required for the majority of these analytical procedures (Lamari *et al*., 2003; Harvey, 1999).

MS has now reached a level of sensitivity in which the study of glycoproteins and the attached oligosaccharide structures can be undertaken on proteins separated by gel electrophoresis. While the advanced "soft ionization" techniques of fast-atom bombardment (FAB), electrospray ionization (ESI), and MALDI permit the analysis of intact polymer units (oligosaccharide, polypeptide, or glycopeptide), full structural characterization of a glycoprotein or glycopeptide also requires determination of branching, linkages, and configurations and identification of same-mass sugar isomers of glycans (i.e., microheterogeneity). Using a combination of 2D-PAGE with MALDI-TOF it is possible to determine the macroheterogeneity, that is, the variation in different N-linked glycans that occupy a single glycosylation site (Mills *et al*., 2001). This technology coupled with site-directed mutagenesis enables studies to be performed to address the topic of glycan function at particular glycosylation sites. The ionization technique of MALDI-MS is arguably the most sensitive of the "soft ionization" techniques at least for simple mass analysis and is ideally suited to carbohydrate analysis, because, unlike many other forms of MS, spectra can be obtained from unmodified glycans (Harvey, 2001). MALDI ionization yields very few fragment ions and therefore is the preeminent technique for screening molecular ions, permitting simple and rapid mass analysis of glycans. Since N-linked glycans are synthesized by a well-defined pathway, an accurate mass measurement will generally permit prediction of the structure. To obtain high-quality mass spectra of the released oligosaccharides, it is essential that all compounds other than the analyte, such as buffers and salts, are removed (Harvey, 2001).

Microheterogeneity, that is, subtle differences in glycan branching or bonding (anomericity) will need to be investigated by combining several techniques, including exoglycosidase digestion (Harvey, 2001), analysis of hydrolysates by gas-chromatography electron impact (EI)-MS or FAB-MS, and tandem mass spectrometry. A recent review discusses the available mass spectrometric techniques for the ionization and fragmentation of oligosaccharides and computer-based approaches for interpretation of oligosaccharide product-ion mass spectra (Zaia, 2004). There are a limited number of Web-based Bioinformatic analysis tools that enable the interpretation of sugar mass spectrometric data. They include GlycoMod (**www.expasy.ch/tools/glycomod**) and GlycoSuiteDB (**http://tmat.proteomesystems.com/glycosuite/**).

8. Summary

N-linked glycan analysis can be undertaken as an adjunct to 2D-PAGE and mass spectrometry in proteomics strategies with relative ease but more detailed analysis of structures, particularly O-linked glycans, often requires a variety of different analytical tools.

Further reading

Schulz BL, Packer NH and Karlsson NG (2002) Small-scale analysis of O-linked oligosaccharides from glycoproteins and mucins separated by gel electrophoresis. *Analytical Chemistry*, **74**, 6088–6097.

References

Edge AS (2003) Deglycosylation of glycoproteins with trifluoromethane sulphonic acid: Elution of molecular structure and function. *Biochemical Journal*, **376**, 339–350.

Emmett MR (2003) Determination of post-translational modifications of proteins by high-sensitivity, high-resolution Fourier transform ion cyclotron resonance mass spectrometry. *Journal of Chromatography A*, **1013**, 203–213.

Hart C, Schulenberg B, Steinberg TH, Leung WY and Patton WF (2003) Detection of glycoproteins in polyacrylamide gels on electroblots using Pro-Q Emerald 488 dye, a fluorescent periodate Schiff-base stain. *Electrophoresis*, **23**(4), 588–598.

Harvey DJ (1999) Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrometry Review*, **18**(6), 349–450.

Harvey DJ (2001) Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics*, **1**, 311–328.

Kirnarsky L, Nomoto M, Ikematsu Y, Hassan H, Bennett EP, Cerny RL, Clausen H, Hollingsworth MA and Sherman S (1998) Structural analysis of peptide substrates for mucin-type O-glycosylation. *Biochemistry*, **37**, 12811–12817.

Küster B, Krogh TN, Mørtz E and Harvey DJ (2001) Glycosylation analysis of gel-separated proteins. *Proteomics*, **1**, 350–361.

Lamari FN, Kuhn R and Karamanos NK (2003) Derivitization of carbohydrates for chromatographic, electrophoretic and mass spectrometric structure analysis. *Journal of Chromatography B*, **793**, 15–36.

Mills K, Mills PB, Clayton PT, Johnson AW, Whitehouse DB and Winchester BG (2001) Identification of α1-antitrypsin variants in plasma using proteomic technology. *Clinical Chemistry*, **47**, 2012–2002.

Towbin H, Staehelin T and Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences United States of America*, **76**, 4350–4354.

Zaia J (2004) Mass spectrometry of oligosaccharides. *Mass Spectrometry Reviews*, **23**, 161–227.

Basic Techniques and Approaches

Mass spectrometry

David J. Harvey
University of Oxford, Oxford, UK

1. Introduction

Analysis by mass spectrometry involves the production of gas-phase ions from the analyte and the measurement of their mass-to-charge ratio (*m/z*) by their movement in electric or magnetic fields. Although the technique has its origins in the late nineteenth century, it was not until the late 1980s that satisfactory methods had been developed for producing ions from the large biopolymers such as proteins, glycoproteins, and nucleic acids. In order to exploit these ionization techniques, there has been a move away from traditional magnetic sector instruments to more modern methods for separation and measuring ions such as time-of-flight (TOF) instrumentation.

2. Ionization techniques

2.1. *Electron impact (EI) and chemical ionization (CI)*

Early techniques, suitable only for volatile organic molecules with masses up to about 1000 Da, involved bombarding the molecules, in the vapor phase, with beams of electrons (electron-impact (EI) ionization) to produce "odd-electron" ions of the type $M^{+\cdot}$, where the charge originates from the removal of an electron. Ions of this type are unstable and dissociate into fragments to produce the characteristic mass spectrum. The technique, coupled with gas chromatography and specific derivatization reactions, is used for the analysis of monosaccharides in experiments designed to identify the constituent sugars (Merkle and Poppe, 1994) of protein-linked carbohydrates and for determinations of linkage position (Hellerqvist, 1990). More stable "even-electron" molecular ions of the type $[M + H]^+$ can be produced by chemical ionization (CI), where the electron beam first ionizes a reagent gas such as methane, *iso*-butane, or ammonia to produce species such as CH_5^+ or NH_4^+, which then transfer a proton to the analyte with little excess energy.

2.2. *Fast-atom bombardment (FAB)*

One of the most important developments for the analysis of larger organic molecules occurred in the early 1980s with the introduction of fast-atom bombardment (FAB)

ionization (Barber *et al*., 1981). In this technique, the sample, often derivatized to a hydrophobic form, is dissolved in a liquid with a low vapor pressure, such as glycerol, and bombarded with a stream of atoms generated from an inert gas, or with ions such as Cs^+. The result is a beam of even-electron ions from a target that does not require heat, both conditions that are conducive to the production of stable molecular ions. Molecules with masses of up to about 10 000 Da can be ionized. The most suitable matrix for glycan ionization appears to be a mixture of glycerol and thioglycerol. The technique was used extensively in the 1980s for analysis of carbohydrates (Dell, 1987), particularly N-linked glycans, as it produced both abundant molecular ions and fragments. However, low sensitivity, caused by a high background from matrix ions, together with a requirement for derivatization, usually by permethylation, has resulted in this technique now being mainly of historical interest.

The real breakthrough in mass spectrometry for proteomics occurred in the late 1980s with the development of two ionization techniques, matrix-assisted laser desorption/ionization (MALDI) (Karas and Hillenkamp, 1988) and electrospray ionization (ESI) (Fenn, 1993), techniques that are both capable of ionizing molecules with masses in excess of 1 000 000 Da.

2.3. Electrospray ionization

The electrospray ion source works at atmospheric pressure and consists of a stainless steel or metalized silica capillary held at a high potential, typically about 3 kV, and placed a few millimeters from a counterelectrode (Figure 1). A solution of the sample flows through the capillary and charged droplets emerge into the ion source where they meet a counterflow of warm nitrogen gas, sometimes known as a curtain gas, to vaporize the solvent. As the droplet diameter decreases, a point is reached where electrostatic repulsion causes it to disintegrate into smaller droplets. The process continues until only ionized sample molecules remain. These are sampled by a sampling cone placed off-axis or orthogonal to the solvent flow and introduced into the mass spectrometer. Typical solvents are mixtures of water and either methanol or acetonitrile; pure water or nonpolar organic solvents are less effective. Traces of acetic or formic acid can be added to the sample solution to promote protonation. Other additives such as metal salts can be used to

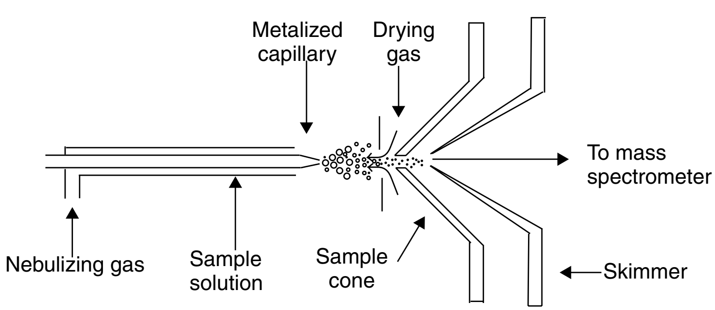


Figure 1 Electrospray ion source

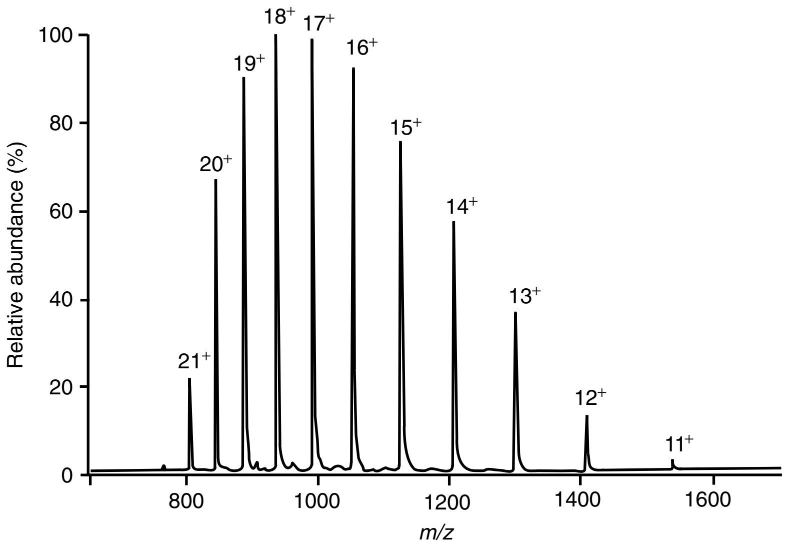


Figure 2 Electrospray spectrum of a protein

produce $[M + Metal]^+$ ions or to promote negative ion formation. Efficiency can be improved by use of a coaxial nebulizing flow of nitrogen around the spray, a technique known as ion spray.

Unlike most other ionization methods, electrospray ionization produces multiple-charged ions of the type $[M + nH]^{n+}$ or $[M - nH]^{n-}$, typically one charge for each kilodalton in mass for an average protein, thus allowing most proteins to produce an array of ions with *m/z* values within the range of most mass spectrometers (Figure 2). As many pairs of peaks are available for molecular weight calculations, an extremely accurate measurement of the mass can be obtained, usually better than 0.01%.

Carbohydrates are amenable to electrospray ionization and can be induced to produce a variety of ions, depending on the conditions. Neutral carbohydrates produce $[M + H]^+$, $[M + Na]^+$, and various doubly charged ions depending on whether the solvent contains an excess of protons or sodium ions. However, the larger glycans are not efficiently ionized. Stronger signals can be obtained in the negative ion mode, particularly from acidic glycans that contain sialic acid or sulfate, and more highly charged ions can form if several acid groups are present. Strong signals from neutral glycans can be obtained in the presence of a proton scavenger such as Cl^- or, particularly, NO_3^-.

2.4. *Matrix-assisted laser desorption/ionization (MALDI)*

In this technique, the sample is cocrystallized with a UV absorbing matrix and ionized with a laser, usually operating in the UV range (Figure 3). Ions are extracted with a high potential (in the order of 20 kV) and accelerated into the

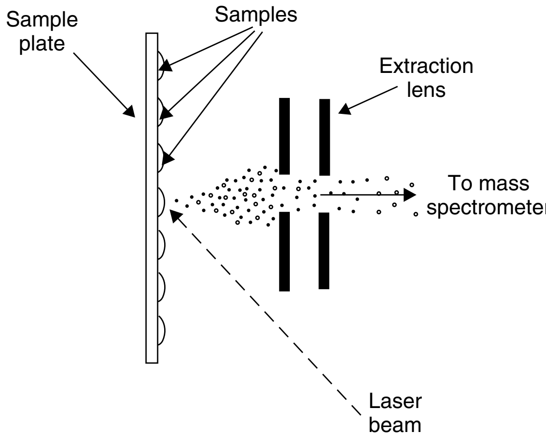


Figure 3 MALDI ion source

mass spectrometer. The purpose of the matrix is to dilute the sample and absorb the laser energy for transmission to the sample, but the exact mechanism by which this occurs is, as yet, unclear. Unlike electrospray ionization, the main products from proteins and glycoproteins are singly charged ions such as $[M+H]^+$, $[M-H]^-$, or $[M+Na]^+$ (Figure 4), but carbohydrates give almost exclusively $[M+Na]^+$ ions.

Various matrices are available; some of the most popular are listed in Table 1.

The laser energy is often sufficient to cause fragmentation, either within the ion source region (in-source decay (ISD)) or after the ion source (postsource decay (PSD)). A variation of MALDI is surface-enhanced laser desorption/ionization (SELDI) (Merchant and Weinberger, 2000) in which the laser target is a surface designed for adsorption of specific analytes such as peptides. MALDI is ideal for the ionization of carbohydrates (Harvey, 1999); however, as with electrospray, the sensitivity is not as high as that for peptides because the latter compounds are easily protonated. Nevertheless, carbohydrates can be derivatized with groups that either contain a charged function or that can readily be protonated or deprotonated. In order to achieve high sensitivity, it is essential to use clean samples from which contaminants such as salts, buffers, and antioxidants have been removed. Suitable methods are adsorption onto membranes (Börnsen, *et al*., 1995) or the use of various resins, often packed into small pipette tips (*see* Article 14, **Sample preparation for MALDI and electrospray**, Volume 5).

2.5. *Mass spectrometers*

There are many techniques for analyzing ions depending on their mass and method of production. Early instruments used magnetic fields to deflect ions according to their *m/z* values and, later, instruments with combined electrostatic and magnetic fields were developed in order to provide high resolution and accurate mass measurements for the determination of ionic composition. These analyzers required a continuous beam of ions, but highest masses were limited by magnet size to about

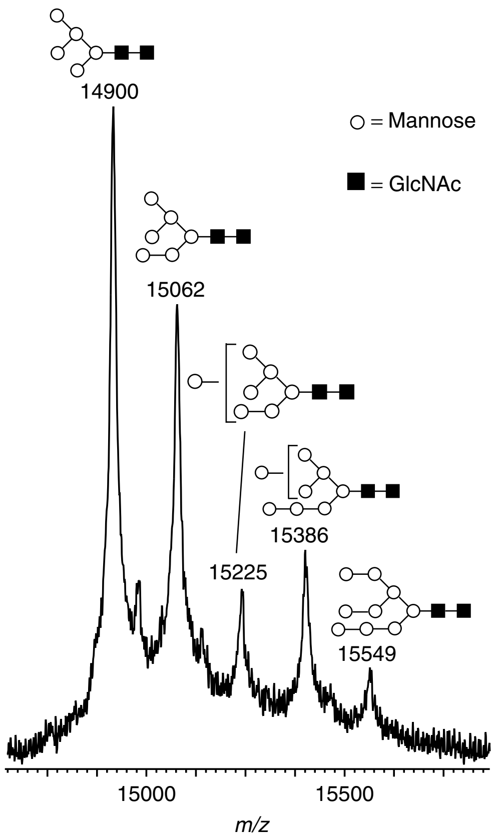
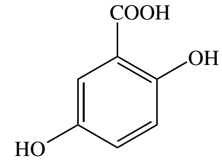
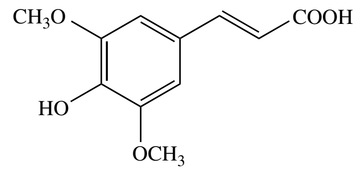
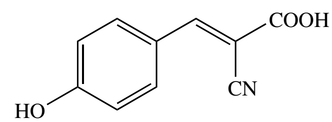
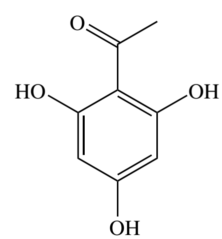
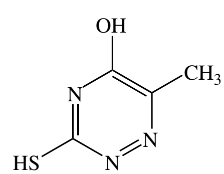
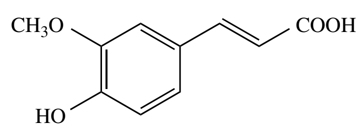
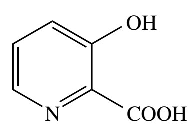


Figure 4 Typical MALDI mass spectrum of a glycoprotein

100 000. They were used extensively for EI, CI, and FAB studies. Quadrupole analyzers were a development of the 1970s. These instruments separated ions by a combination of direct and oscillating electric fields applied to four concentric rods (Figure 5). Again, continuous beams were required and, although relatively cheap, lightweight, and convenient to control with computers, upper mass limits and resolution were limited to a few thousands (*see* Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5).

Time-of-flight (TOF) instruments (*see* Article 7, **Time-of-flight mass spectrometry**, Volume 5), (second analyzer in Figure 6), on the other hand, have no theoretical upper mass limit and are sensitive because all ions can potentially reach the detector; with magnetic sector and quadrupole instruments that scan the mass range, most ions are discarded. Modern instruments produce resolutions approaching those provided by double-focusing magnetic sector spectrometers. TOF instruments require a pulsed ion beam, such as that from a MALDI ion source, and measure *m/z* values by timing ions transmitted through a flight tube typically about 1 m in length. High-resolution measurements and focusing of PSD ions is achieved by reflecting the ions with an ion mirror or reflectron. Two other

Table 1 Common matrices for MALDI mass spectrometry

Matrix	Molecular weight	Analyte	Structure
2,5-Dihydroxybenzoic acid (DHB)	154.1	General, particularly sugars	
Sinapinic acid (3,5-dimethoxy-4-hydroxycinnamic acid)	224.2	Proteins	
α-Cyano-4-hydroxycinnamic acid (HCCA)	189.2	Peptides, lipids	
2,4,6-trihydroxyacetophenone (THAP)	167.2	Sugars. Particularly those containing sialic acid, oligonucleotides	
6-Aza-2-thiothymine (ATT)	143.2	Gangliosides	
Ferulic acid	194.2	Glycoproteins	
3-Hydroxypicolinic acid (HPA)	139.1	Oligonucleotides	

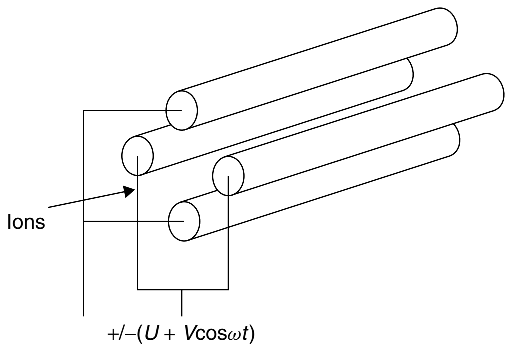


Figure 5 Quadrupole mass analyzer

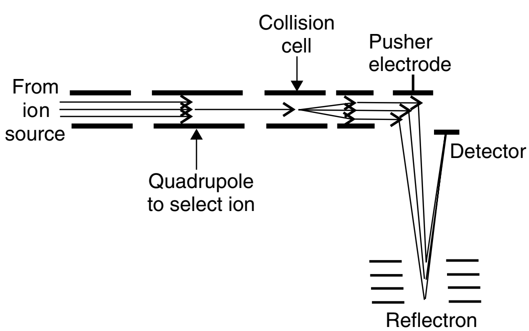


Figure 6 Q-Tof mass spectrometer

analyzers of interest are ion traps (*see* Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5) and Fourier transform ion-cyclotron resonance (FT-ICR) instruments (*see* Article 5, **FT-ICR**, Volume 5), both of which can trap ions and perform successive stages of fragmentation. The latter instruments are capable of very high resolution.

Analyzers can be connected in series in order to obtain fragment ions from a selected parent (*see* Article 10, **Hybrid MS**, Volume 5). The parent ion is selected from a mixture by using the first analyzer in a static condition, ions are fragmented in a relatively high-pressure region of the instrument, and the fragments are analyzed by a second analyzer. One of the most popular instruments of this type for proteomics is the Q-Tof tandem instrument that contains a quadrupole analyzer for initial ion selection, a collision cell, usually filled with argon to fragment the ions by collision, and a TOF analyzer to mass-measure the products at high sensitivity (Figure 6). Ionization usually employs an electrospray ion source although MALDI-Q-Tof instruments have recently been developed (Krutchinsky, *et al*., 1998). Ion trap and triple quadrupole instruments are also commonly found, the latter producing results similar to that of the Q-Tof instrument but at lower sensitivity because the third quadrupole is a scanning analyzer.

2.6. Uses of mass spectrometry for carbohydrate analysis

Both ESI and MALDI are able to produce spectra from glycoproteins but the resolution is usually not sufficient to resolve glycoforms for any but the smaller molecules, and then only those with single glycosylation sites (Figure 4). Glycans are, thus, usually removed either chemically (e.g., hydrazine) or enzymatically with endoglycosidases such as protein-*N*-glycosidase F (PNGase F) prior to structural analysis (*see* Article 76, **Analysis of N- and O-linked glycans of glycoproteins**, Volume 6). Site occupancy, however, can often be determined by cleaving the glycoprotein with a protease, such as trypsin, in order to localize a specific glycosylation site to a small glycopeptide. Mass spectrometric analysis of such glycopeptides can reveal the glycoforms at that site, and the carbohydrate compositions can be determined from their masses obtained by subtraction of the peptide mass, if the sequence is known. N-link site occupancy can be determined by cleaving the carbohydrate with PNGase F and measuring the mass of the residual peptide because this amidase leaves aspartic acid, which is one mass unit heavier, in place of the original asparagine.

Carbohydrates attached to glycoproteins contain only a few mass-different monosaccharides (Table 2). Consequently, it is relatively easy to deduce the

Table 2 Residue masses of common monosaccharides

Oligosaccharide	Residue formula	Residue mass
Deoxypentose	$C_5H_8O_3$	116.047 116.117
Pentose	$C_5H_8O_4$	132.042 132.116
Deoxyhexose	$C_6H_{10}O_4$	146.058 146.143
Hexose	$C_6H_{10}O_5$	162.053 162.142
Hexosamine	$C_6H_{11}NO_4$	161.069 161.158
HexNAc	$C_8H_{13}NO_5$	203.079 203.195
Hexuronic acid	$C_6H_8O_6$	176.032 176.126
N-Acetyl-neuraminic acid	$C_{11}H_{17}NO_8$	291.095 291.258
N-glycoyl-neuraminic acid	$C_{11}H_{17}NO_9$	307.090 307.257

Top figure = monoisotopic mass (based on C = 12.000000, H = 1.007825, N = 14.003074, O = 15.994915); Lower figure = average mass (based on C = 12.011, H = 1.00794, N = 14.0067, O = 15.9994).
Masses of intact glycans can be obtained by addition of the residue masses plus the figure for the terminal group (H_2O, 18.011 (monoisotopic) and 18.0153 (average)).

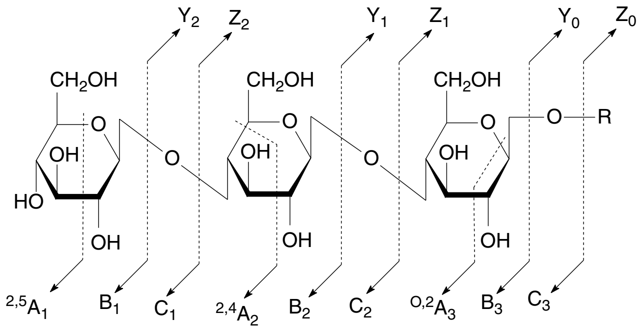


Figure 7 Nomenclature for describing fragment ions from carbohydrates

composition of these carbohydrates in terms of their isobaric monosaccharide composition by means of a single mass measurement. Mass spectrometry is not, however, able to distinguish isomers by this means; in order to achieve this, some chromatographic method of isomer separation, such as on-line high-performance liquid chromatography (HPLC) must be employed. Fine structural details are traditionally determined with the aid of exoglycosidases, with the products monitored either by HPLC after fluorescent labeling, or by MALDI MS.

With the advent of instruments such as the Q-Tof tandem mass spectrometer, fragmentation is assuming greater importance. Carbohydrates give mainly two types of fragment ions: glycosidic fragments where cleavage occurs between the sugar rings and which provides information on sequence and branching, and cross-ring fragments that provide information on linkage. Fragment ion nomenclature is as proposed by Domon and Costello (1988), and is outlined in Figure 7. $[M + H]^+$ ions give almost exclusively glycosidic fragment ions but $[M + Na]^+$ ions produce more complicated spectra with both glycosidic and cross-ring fragment ions. It has recently become evident that fragmentation of $[M - H]^-$ or $[M + X]^-$ ions, where X is a proton scavenger such as NO_3^-, produces much more informative spectra than fragmentation of positive ions. For N-linked glycans, for example, they produce fragments that reveal specific structural details such as individual antenna composition and the presence of a bisecting GlcNAc residue.

With the recent development of more sensitive instrumentation, better means of producing ions, and efficient coupling of contrasting techniques such as HPLC and Q-Tof mass spectrometry, the analysis of carbohydrates will increasingly depend on these technologies in the coming years.

Further reading

Ashcroft AE (1997) *Ionization Methods in Organic Mass Spectrometry*, Royal Society of Chemistry: Cambridge.

Chapman JR (1996) *Protein and Peptide Analysis by Mass Spectrometry, Methods in Molecular Biology*, Humana: Totowa.

Cole RB (1997) *Electrospray Ionization Mass Spectrometry, Fundamentals, Instrumentation and Applications*, Wiley: New York.

Yates JR III (1998) Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry*, **33**, 1–19.

References

Barber M, Bordoli RS, Sedgwick RD and Tyler AN (1981) Fast atom bombardment of solids (FAB): a new ion source for mass spectrometry. *Journal of the Chemical Society, Chemical Communications*, **7**, 325–327.

Börnsen KO, Mohr MD and Widmer HM (1995) Ion exchange and purification of carbohydrates on a Nafion(r) membrane as a new sample pretreatment for matrix-assisted laser desorption-ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **9**, 1031–1034.

Dell A (1987) FAB Mass spectrometry of carbohydrates. *Advances in Carbohydrate Chemistry and Biochemistry*, **45**, 19–72.

Domon B and Costello CE (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal*, **5**, 397–409.

Fenn JB (1993) Ion formation from charged droplets: roles of geometry, energy and time. *Journal of the American Society for Mass Spectrometry*, **4**, 524–535.

Harvey DJ (1999) Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrometry Reviews*, **18**, 349–451.

Hellerqvist CG (1990) Linkage analysis using Lindberg method. *Methods in Enzymology*, **193**, 554–573.

Karas M and Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Analytical Chemistry*, **60**, 2299–2301.

Krutchinsky AN, Loboda AV, Spicer VL, Dworschak R, Ens W and Standing KG (1998) Orthogonal injection of matrix-assisted laser desorption/ionization ions into a time-of-flight spectrometer through a collisional damping interface. *Rapid Communications in Mass Spectrometry*, **12**, 508–518.

Merchant M and Weinberger SR (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*, **21**, 1164–1177.

Merkle RK and Poppe I (1994) Carbohydrate composition analysis of glycoconjugates by gas-liquid chromatography/mass spectrometry. *Methods in Enzymology*, **230**, 1–15.

Basic Techniques and Approaches

Analysis of N- and O-linked glycans of glycoproteins

Tony H. Merry, Louise Royle, Catherine M. Radcliffe, Raymond A. Dwek and Pauline M. Rudd
Glycobiology Institute, University of Oxford, Oxford, UK

Sviatlana A. Astrautsova
Medical University of Grodno, Grodno, Belarus

1. Introduction

Sequence analysis has shown that nearly 70% of proteins carry the N-glycosylation sequon and that a further 10% can become O-glycosylated. The majority of proteins are therefore potentially glycosylated (Apweiler *et al*., 1999), and the analysis of the associated glycans is essential in their complete structural characterization. Unlike nucleic acids or proteins, both of which have linear sequences, most of the glycans attached to glycoproteins are branched structures with various position and linkage isomers. High resolution–separation techniques, which are often used concurrently, are required to characterize these diverse structures. Glycoproteins generally consist of a mixture of glycoforms in which a range of different sugars are present at each glycosylation site. In addition, some glycoproteins contain glycosylation sites that are variably occupied. Few glycoproteins have been fully characterized, although some have undergone extensive analysis (Takeuchi *et al*., 1988; Parekh *et al*., 1989; Rudd *et al*., 1997). In practice, it is possible to use the techniques described here to identify the majority of the glycans. Increasingly, techniques are becoming available for the analysis of the glycosylation of intact glycopeptides (Harvey, 2001; Iwase *et al*., 2001; Harvey, 2003). Monitoring the glycosylation of proteins and disease-associated alterations or determining the effect of alterations in the activity of glycan processing enzymes on glycosylation profiles is now a realistic proposition. Changes in glycosylation have been described in cancer (Litynska *et al*., 2000; Hakomori, 2001; Peracaula *et al*., 2003a,b; Callewaert *et al*., 2004; Ohyama *et al*., 2004), congenital enzyme deficiencies (Butler *et al*., 2003), autoimmune diseases (Parekh *et al*., 1985), and in various tissues of knockout mice in which genes for specific enzymes have been deleted (Furukawa *et al*., 2001; Sutton-Smith *et al*., 2002; Kui Wong *et al*., 2003). Glycosylation changes

may provide disease markers (Parekh *et al*., 1985) and give insights into disease pathogenesis (Malhotra *et al*., 1995).

2. Techniques

2.1. *Glycan release*

Figure 1 shows an overall strategy for glycan analysis. The most widely used technique for releasing N-linked glycans involves peptide-*N*-glycanase F (PNGaseF). This enzyme (EC 3.5.1.52) is an asparagine amidase, which cleaves the linkage between the core GlcNAc and the asparagine residue on the protein, releasing all classes of N-linked glycans (Kuhn *et al*., 1994) with the exception of those containing a fucose α 1–3 linked to the core GlcNAc (Altmann *et al*., 1995), which occur frequently in plant glycans. Glycans containing this α 1–3 linked fucose can however be released by the enzyme peptide-*N*-glycanase A used in conjunction with trypsin (Navazio *et al*., 2002; Ko *et al*., 2003; Tekoah *et al*., 2004). Alternatively, endoglycosidases that cleave N-linked glycans between the two core GlcNAcs can be used. Suitable enzymes include endo-H, which specifically cleaves oligomannose and hybrid N-links (Robbins *et al*., 1981), or endo-$F_{1,2\&3}$ (EC 3.2.1.96), which cleave oligomannose and hybrid, oligomannose and biantennary, bi- and triantennary structures respectively (Waddling *et al*., 2000). Enzymatic release has been demonstrated to release sugars efficiently, particularly from denatured proteins, and may be used to release N-linked glycans from individual glycoproteins separated by SDS-PAGE (Küster *et al*., 1997). Figure 2 shows the NP-HPLC profiles of N-linked glycans released by PNGaseF from the heavy chain, J chain, and secretory component (SC) of secretory IgA.

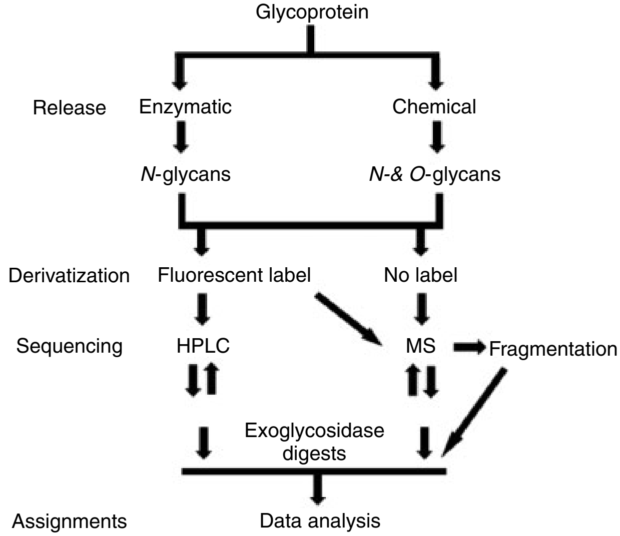


Figure 1 Diagram of a glycan analysis scheme using HPLC and mass spectrometry

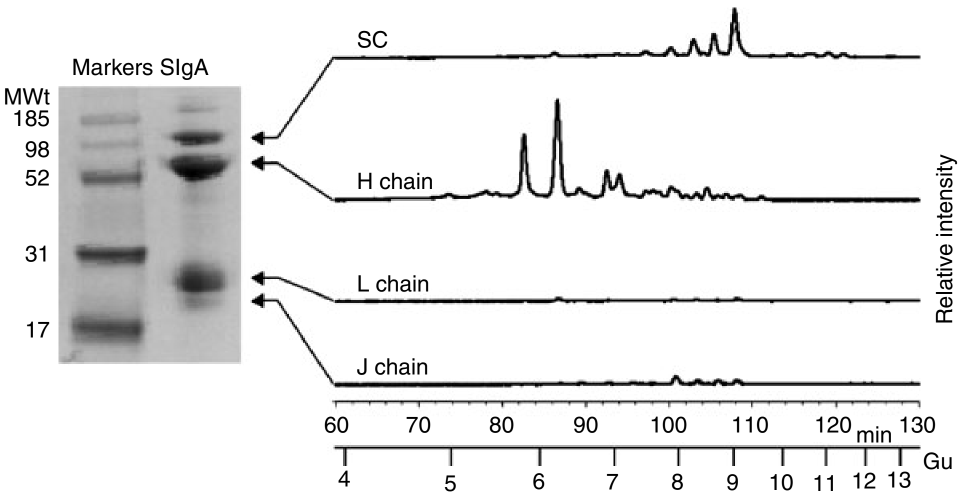


Figure 2 Secretory IgA (from human colostrum) was separated into secretory component (SC), heavy chain (H chain), light chain (L chain), and joining chain (J chain), on 10% BisTris SDS-PAGE after reduction and alkylation. NP-HPLC profiles of the 2-AB labeled *N*-glycans released by PNGaseF from gel bands show that each peptide contains a different population of *N*-glycans (Reproduced by permission of The American Society for Biochemistry & Molecular Biology)

Both N- and O-linked glycans can be released by chemical means, for example, by base-catalyzed β-elimination (Carlson and Blackwell, 1968; Aminoff *et al*., 1980; Strecker *et al*., 1995) or hydrazinolysis (Fukuda *et al*., 1976; Strecker *et al*., 1981; Yamashita *et al*., 1984; Bendiak and Cumming, 1985; Patel *et al*., 1993). For the release of O-linked glycans, chemical release is the only effective method as no generic *O*-glycanases are currently known. The glycans released by β-elimination under the strong reducing conditions needed to prevent degradation (Aminoff *et al*., 1980), are converted to the alditol form, which is difficult to derivatize. This may not be a problem for analysis by mass spectrometry (MS) but it does preclude the use of high sensitivity–fluorescent detection. To preserve intact O- and N-linked glycans and a nonreducing terminal for derivatization, hydrazinolysis is preferred (Patel *et al*., 1993; Merry *et al*., 2002).

2.2. *Derivatization of glycans*

Released glycans are most commonly derivatized with a fluorophore to enable detection and quantitation in the sub-picomole range. Commonly used fluorophores include 2-aminopyridine (2-AP), 2-aminobenzamide (2-AB), 2-aminoanthranilic acid (2-AA), 2-aminoacridone (AMAC), and 8-aminonaphthalene-1,3,6-trisulfonic acid (ANTS) (see Anumula, 2000 and references therein). The choice of fluorescent label depends on the separation technology used. For example, a charged label is specifically required for capillary electrophoresis. Although underivatized glycans can be detected by mass spectrometry, derivatization may aid their ionization (Harvey, 1999; Harvey, 2000).

2.3. *Glycan sequencing*

The most commonly used high-resolution techniques for separating glycans for sequence analysis are high performance liquid chromatography (HPLC), capillary electrophoresis, and mass spectrometry.

Several different types of HPLC columns can be used to separate glycans: weak anion exchange (WAX), which separates glycans by charge (Tran *et al*., 2000; Royle *et al*., 2002) most commonly derived from sialic acid residues or sulfated sugars; reverse-phase (RP), which separates oligosaccharides on the basis of hydrophobic interactions (Kuraya and Hase, 1996; Guile *et al*., 1998; Royle *et al*., 2002); and normal-phase (NP), which separates glycans on amide columns (Takahashi *et al*., 1993; Guile *et al*., 1996) where retention time mainly depends on the first approximation of the size of the glycan (Guile *et al*., 1996; Guile *et al*., 1998; Royle *et al*., 2002). Normal-phase HPLC, which is used most widely, resolves neutral and charged glycans in a single run. By standardizing the retention time of each peak to that of an internal standard of hydrolyzed dextran as described by Guile *et al*. (1996), reproducible glucose unit (GU) values can be obtained for individual glycans. The incremental GU values are measures of the affinity of each glycan for the column matrix and are related to the hydrophilicity of each monosaccharide residue in the chain. The affinity also depends on how much of the hydrophilic surface is exposed to the column matrix. It is this feature that gives the fine specificity of the columns, allowing arm-specific isomers such as the biantennary mono galactosylated glycans, where the galactose may be on either the 3 arm (GU 6.48) or the 6 arm (GU 6.29), to be resolved. Preliminary structures are assigned from a database of GU values from known structures (e.g., the Oxford Glycobiology 2-AB Glycan Database) (Guile *et al*., 1996; Royle *et al*., 2002) and are confirmed by exoglycosidase digestions.

Glycans with similar compositions often coelute on NP-HPLC. To identify individual structures in complex mixtures more than one separation method may be required, for example, fraction collection from WAX-HPLC followed by normal-phase HPLC of the fractionated glycans (Guile *et al*., 1998; Royle *et al*., 2003). Figure 3 shows the profiles from mono- and disialylated glycans which coelute on NP-HPLC after separation by WAX-HPLC.

To determine the exact monosaccharide sequence and linkages, treatment with specific exoglycosidases, either individually or in arrays, followed by profiling of the digestion products by NP-HPLC, is often used (Hase *et al*., 1987; Tomiya *et al*., 1988; Anumula and Dhume, 1998). Changes in the positions of the peaks relative to the dextran standard (GU values) are related to both the type and the number of monosaccharides removed (Guile *et al*., 1996). Combinations of exoglycosidases can be used where they have sufficiently similar pH optima and buffer requirements and conditions may need to be optimized for complete removal of glycans from specific glycoproteins (Edge *et al*., 1992; Prime and Merry, 1998). In such cases, it may be necessary to check the specificity under the incubation conditions used, rather than those given by the supplier, as low levels of contaminating enzymes may be present. Figure 4 shows the NP-HPLC profiles of the *N*-glycans from secretory component of IgA following a series of exoglycosidase array digestions with the

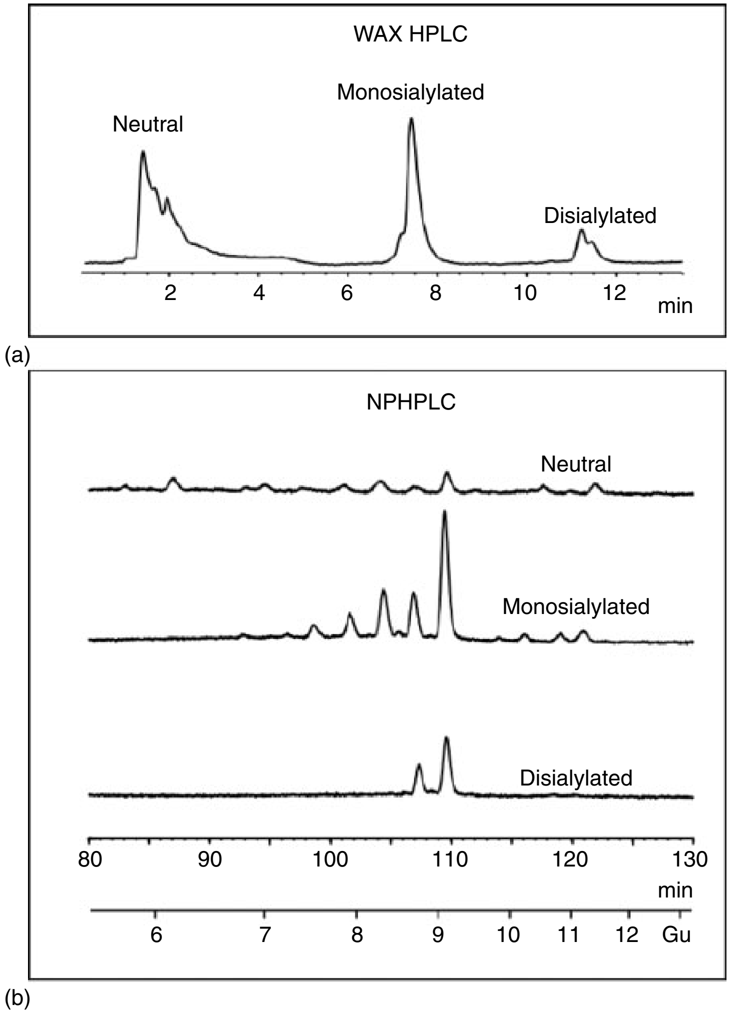


Figure 3 (a) shows the *N*-glycans from secretory component of human colostrum secretory IgA separated by weak anion exchange (WAX) HPLC. The differently charged fractions were collected and run on NP-HPLC (b). The data show that the two peaks that contain disialylated glycans elute with the same GU values as two monosialylated glycan structures (Reproduced by permission of The American Society for Biochemistry & Molecular Biology)

major structures annotated. In the top (undigested glycan pool) profile, both the two largest peaks contain a mixture of a di- and a monosialylated structures. The data from the WAX separation in Figure 3 was used to assign these structures.

The analysis of O-linked glycans is complicated by the presence of at least eight different core structures (*see* Article 65, **Structure/function of *O*-glycans**, Volume 6), compared with the trimannosyl-chitobiose core common to all N-linked glycans (Kornfeld and Kornfeld, 1985). This makes data interpretation more difficult than for the N-linked glycans.

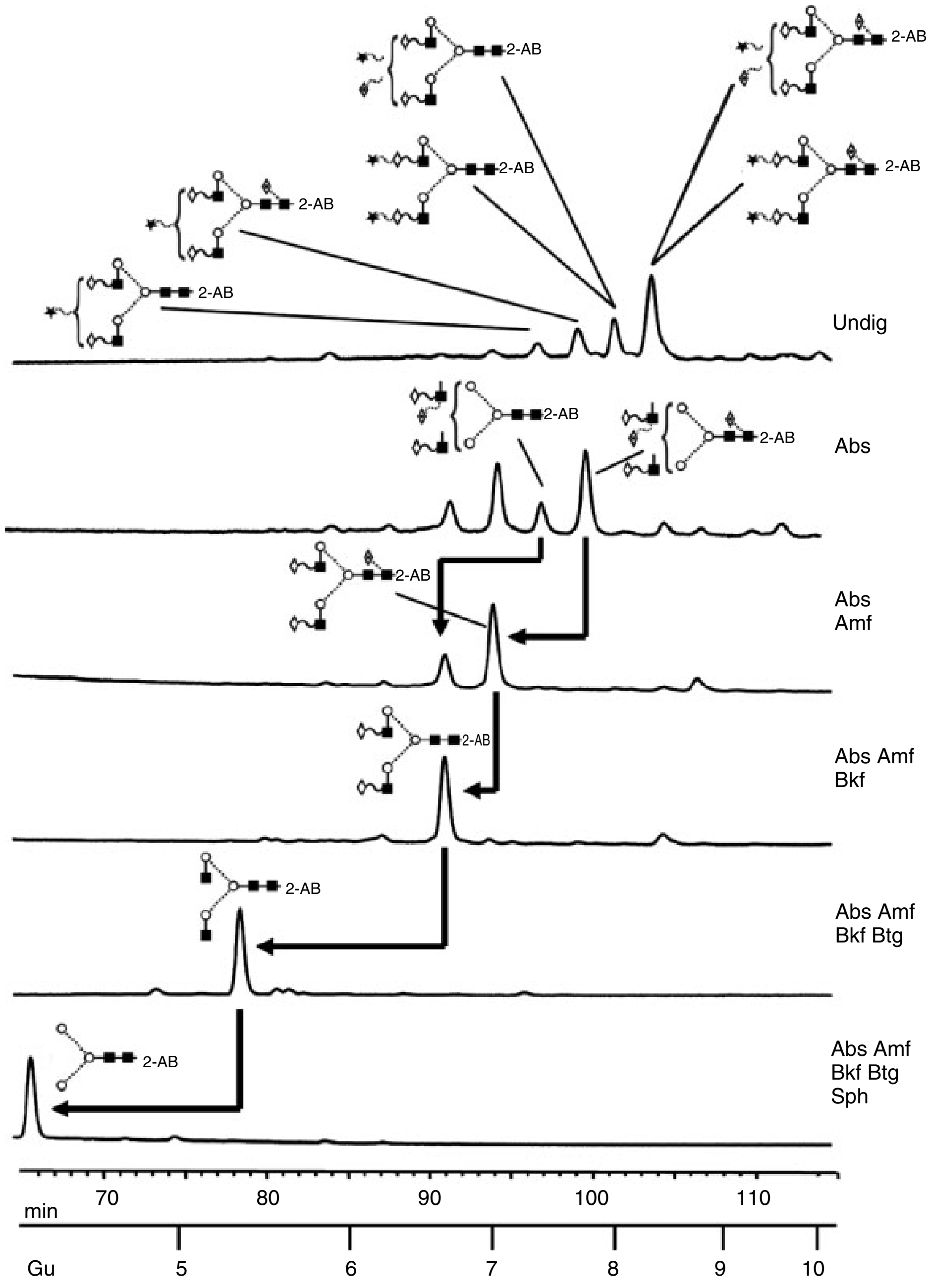


Figure 4 NP-HPLC profiles of 2-AB labeled *N*-glycans from secretory component of human colostrum secretory IgA. Aliquots of the total glycan pool were incubated with arrays of exoglycosidase enzymes then analyzed by NP-HPLC (undig = undigested). Exoglycosidase abbreviations: *Arthrobacter ureafaciens* sialidase (Abs); Bovine kidney α-fucosidase (Bkf); almond meal α-fucosidase (Amf); bovine testes β-galactosidase (Btg); *Streptococcus pneumoniae* β-*N*-acetylhexosaminidase (Sph). The major peaks are annotated using the scheme shown in Figure 6; arrows indicate movement of peaks following digestion

Recent advances in mass spectrometry instrumentation now allow profiling of a large number of different glycans (*see* Article 75, **Mass spectrometry**, Volume 6) but fragmentation or derivatization is required to give detailed information. Two main types of MS are currently used, namely, matrix assisted laser desorption ionization (MALDI) or electrospray (ESI-MS).

A combination of HPLC separation, mass detection, and fragmentation in LC-MS/MS with exoglycosidase digestion data can be used to sequence complicated N- and O-linked glycans (Mattu *et al*., 2000; Peracaula *et al*., 2003a,b; Royle *et al*., 2003). Figure 5 shows an example of MS fragmentation data for an *O*-glycan. If the glycoprotein comes from a source where the biosynthetic pathway is known, and where characteristic glycoproteins have already been sequenced, glycan sequences can be predicted from NP-HPLC data and confirmed with a limited number of enzymatic digestions.

Capillary electrophoresis (Camilleri *et al*., 1995) and high performance anion exchange chromatography (HPAEC) with pulsed amperometric detection (PAD) (Goodarzi and Turner, 1998) can also be used to analyze glycans. These techniques require careful standardization and optimization but can be used routinely for high-throughput screening (Goodarzi and Turner, 1998; Taverna *et al*., 1998). DNA sequencing instruments have also been modified for glycan profiling, particularly for the analysis of de-sialylated glycans (Callewaert *et al*., 2004).

2.4. *Presentation of data*

The established IUPAC nomenclature (1982) gives a complete and unambiguous description of each glycan but it is difficult to visualize the structures from the α-numeric code of the nomenclature, especially when comparing similar glycans. It is therefore common practice to use symbols that are more readily recognized by the human eye to depict the glycan structure, but there is no universal system in place, with different types of representation widely used. A format incorporating features from several existing schemes has been proposed (see Figure 6) that enables complex structural information to be displayed in simple cartoon format. Geometric symbols are used to represent individual monosaccharides, the shading of which has been chosen to represent derivatives of the native monosaccharide structure. For example, an open symbol means the sugar is not substituted, the filled symbol indicates that the monosaccharide contains an *N*-acetyl group, and a dot indicates a deoxy sugar. Hence, Gal is represented by an open diamond, GalNAc by a filled diamond, and deoxy-galactose (fucose) by an open diamond with a dot inside. The linkage positions are represented by the angle of the line linking adjacent monosaccharides (as first proposed by Kamerling and Vliegenthart, 1992) and most sugars are linked via the C1 reducing end (except sialic acids, which are linked via C2); this information is not represented in these diagrams. It is the position to which the reducing end of the sugar to the left attaches to the sugar on the right, which is represented by the bond angle. A continuous line indicates a β-linkage, a dotted line an α-linkage. Where the exact monosaccharide is unknown (e.g., from mass spec data), an appropriately shaded symbol is used; for example, an open hexagon for a hexose and a filled hexagon for a HexNAc. Unknown linkages are

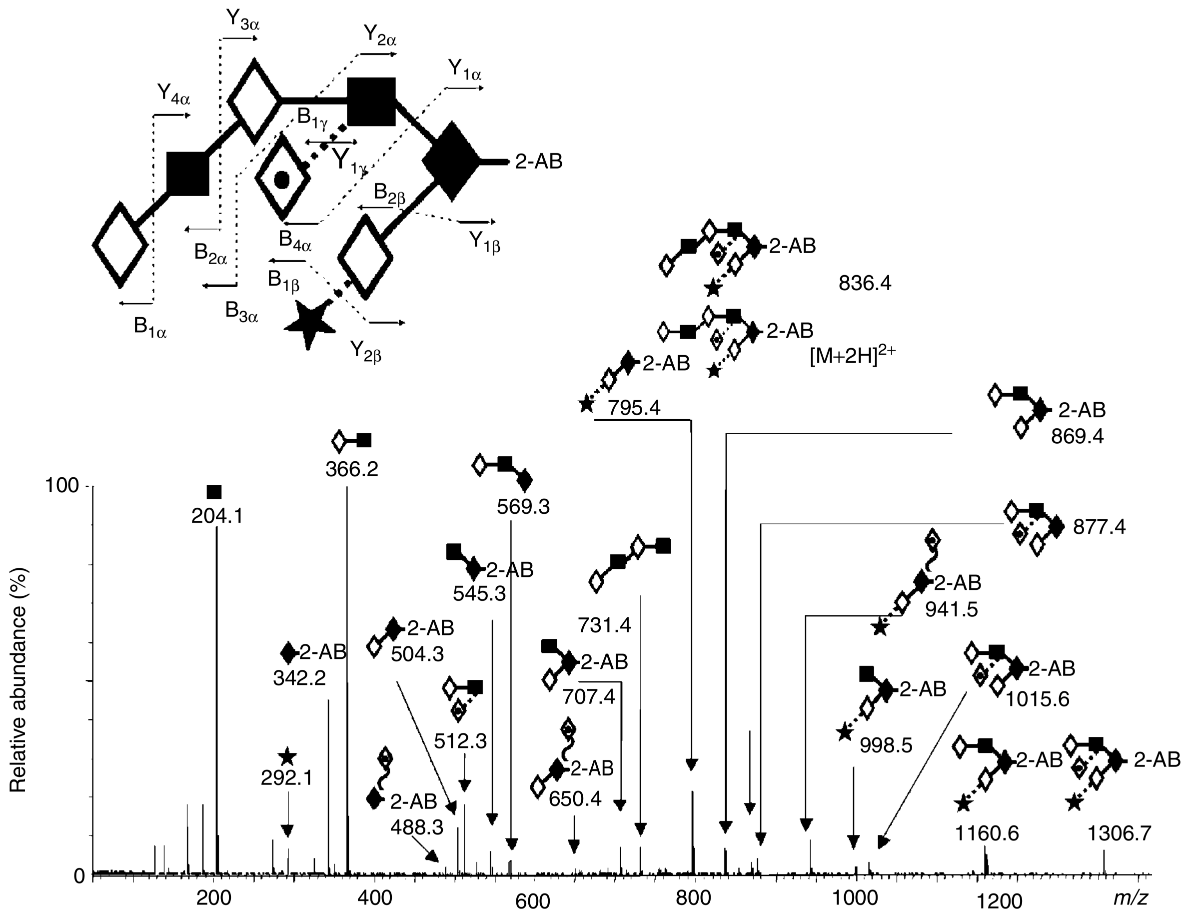


Figure 5 MS fragmentation of an *O*-glycan from secretory IgA by electrospray ionization following online NP-HPLC. The fragmentation scheme of the $[M+2H]^{2+}$ ion with *m/z* 836.4 is based on that of Domon and Costello (Domon and Costello, 1988) (Reprinted from Analytical Chemistry, **304**, Royle L. *et al*., pp 70–90, Copyright 2002, with permission from Elsevier)

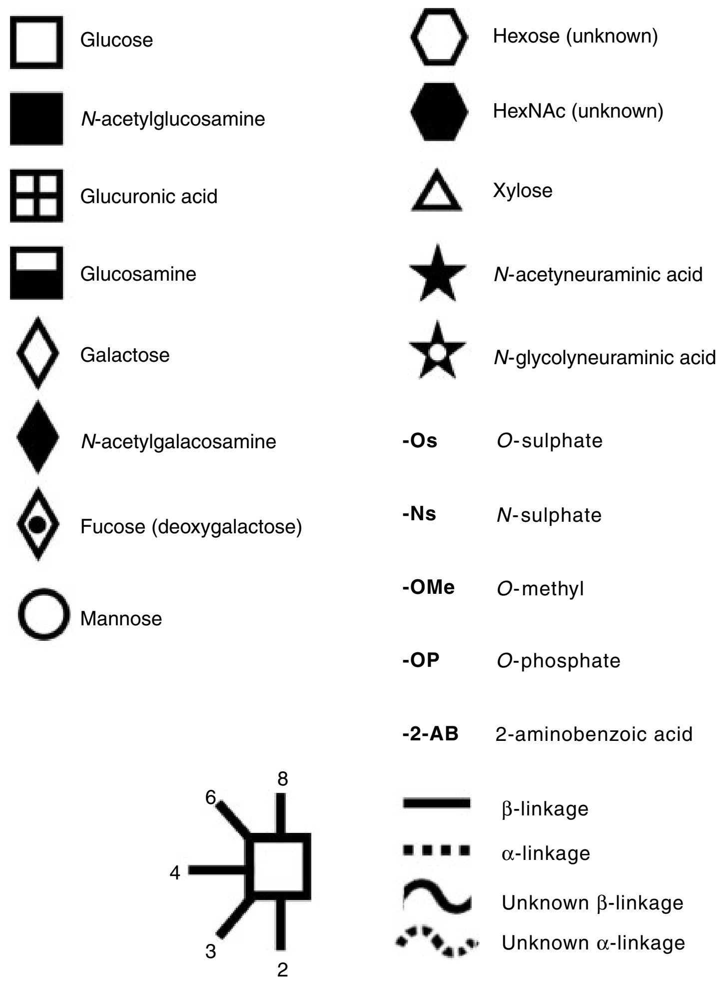


Figure 6 A scheme for the diagrammatic representation of glycans

represented by a wavy line. The normal D or L configuration or ring size (pyranose or furanose) that the monosaccharide adopts in glycans is assumed; for example, an open square represents D-glucopyranose. For fucose, the L configuration is taken as the default. Substituents such as sulfation or phosphorylation can also be shown by use of appropriate letters. This simple visual representation of both sugars and linkages clearly conveys the structural information needed when working with complex glycan structures.

References

Altmann F, Schweiszer S and Weber C (1995) Kinetic comparison of peptide: N-glycosidases F and A reveals several differences in substrate specificity. *Glycoconjugate Journal*, **12**(1), 84–93.

Aminoff D, Gathmann WD, McLean CM and Yadomae T (1980) Quantitation of oligosaccharides released by the beta-elimination reaction. *Analytical Biochemistry*, **101**(1), 44–53.
Anumula KR (2000) High-sensitivity and high-resolution methods for glycoprotein analysis. *Analytical Biochemistry*, **283**(1), 17–26.
Anumula KR and Dhume ST (1998) High resolution and high sensitivity methods for oligosaccharide mapping and characterization by normal phase high performance liquid chromatography following derivatization with highly fluorescent anthranilic acid. *Glycobiology*, **8**(7), 685–694.
Apweiler R, Hermjakob H and Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica and Biophysica Acta*, **1473**(1), 4–8.
Bendiak B and Cumming DA (1985) Hydrazinolysis-N-reacetylation of glycopeptides and glycoproteins. Model studies using 2-acetamido-1-N-(L-aspart-4-oyl)-2-deoxy-beta-D-glucopyranosy lamine. *Carbohydrate Research*, **144**(1), 1–12.
Butler M, Quelhas D, Critchley AJ, Carchon H, Hebestreit HF, Hibbert RG Vilarinho L, Teles E, Matthijs G, Schollen E, *et al*. (2003) Detailed glycan analysis of serum glycoproteins of patients with congenital disorders of glycosylation indicates the specific defective glycan processing step and provides an insight into pathogenesis. *Glycobiology*, **13**(9), 601–622.
Callewaert N, Van Vlierberghe H, Van Hecke A, Laroy W, Delanghe J and Contreras R (2004) Noninvasive diagnosis of liver cirrhosis using DNA sequencer-based total serum protein glycomics. *Nature Medicine*, **10**(4), 429–434.
Camilleri P, Harland GB and Okafo G (1995) High resolution and rapid analysis of branched oligosaccharides by capillary electrophoresis. *Analytical Biochemistry*, **230**(1), 115–122.
Carlson DM and Blackwell CJ (1968) Structures and immunochemical properties of oligosaccharides isolated from pig submaxillary mucins. *Journal of Biological Chemistry*, **243**, 616–626.
Domon B and Costello CE (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal*, **5**, 397–409.
Edge CJ, Rademacher TW, Wormald MR, Parekh RB, Butters TD, Wing DR and Dwek RA (1992) Fast sequencing of oligosaccharides: the reagent-array analysis method. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(14), 6338–6342.
Fukuda M, Kondo T and Osawa T (1976) Studies on the hydrazinolysis of glycoproteins. Core structures of oligosaccharides obtained from porcine thyroglobulin and pineapple stem bromelain. *Journal of Biochemistry*, **80**(6), 1223–1232.
Furukawa K, Takamiya K, Okada M, Inoue M and Fukumoto S (2001) Novel functions of complex carbohydrates elucidated by the mutant mice of glycosyltransferase genes. *Biochimica Et Biophysica Acta*, **1525**(1–2), 1–12.
Goodarzi MT and Turner GA (1998) Reproducible and sensitive determination of charged oligosaccharides from haptoglobin by PNGase F digestion and HPAEC/PAD analysis: glycan composition varies with disease. *Glycoconjugate Journal*, **15**(5), 469–475.
Guile GR, Harvey DJ, O'Donnell N, Powell AK, Hunter AP, Zamze S, Fernandes DL, Dwek RA and Wing DR (1998) Identification of highly fucosylated N-linked oligosaccharides from the human parotid gland. *European Journal of Biochemistry*, **258**(2), 623–656.
Guile GR, Rudd PM, Wing DR, Prime SB and Dwek RA (1996) A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. *Analytical Biochemistry*, **240**(2), 210–226.
Hakomori S (2001) Tumor-associated carbohydrate antigens defining tumor malignancy: basis for development of anti-cancer vaccines. *Advances in Experimental Medicine and Biology*, **491**, 369–402.
Harvey DJ (1999) Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrometry Reviews*, **18**(6), 349–450.
Harvey DJ (2000) Electrospray mass spectrometry and fragmentation of N-linked carbohydrates derivatized at the reducing terminus. *Journal of the American Society for Mass Spectrometry*, **11**(10), 900–915.
Harvey DJ (2001) Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics*, **1**(2), 311–328.
Harvey DJ (2003) Identification of sites of glycosylation. *Methods in Molecular Biology*, **211**, 371–383.

Hase S, Natsuka S, Oku H and Ikenaka T (1987) Identification method for twelve oligomannose-type sugar chains thought to be processing intermediates of glycoproteins. *Analytical Biochemistry*, **167**(2), 321–326.

IUPAC-IUB joint commission on biochemical nomenclature (JCBN). (1982) Abbreviated terminology of oligosaccharide chains. Recommendations 1980. *European Journal of Biochemistry / FEBS*, **126**(3), 433–437.

Iwase H, Tanaka A, Hiki Y, Kokubo T, Sano T, Ishii-Karakasa I, Hisatani K, Kobayashi Y and Hotta K (2001) Analysis of the microheterogeneity of the IgA1 hinge glycopeptide having multiple O-linked oligosaccharides by capillary electrophoresis. *Analytical Biochemistry*, **288**(1), 22–27.

Kamerling JP and Vliegenthart JFG (1992) *High-Resolution 1H-Nuclear Magnetic Resonance Spectroscopy of Oligosaccharide-Alditols Released from Mucin-Type O-Glycoproteins*, Plenum Press: New York.

Ko K, Tekoah Y, Rudd PM, Harvey DJ, Dwek RA, Spitsin S, Hanlon CA, Rupprecht C, Dietzschold B, Golovkin M, *et al*. (2003). Function and glycosylation of plant-derived antiviral monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 8013–8018.

Kornfeld R and Kornfeld S (1985) Assembly of asparagine-linked oligosaccharides. *Annual Review of Biochemistry*, **54**, 631–664.

Kuhn P, Tarentino AL, Plummer TH and Van Roey P Jr (1994) Crystal structure of peptide-N4-(N-acetyl-beta-D-glucosaminyl)asparagine amidase F at 2.2-A resolution. *Biochemistry*, **33**(39), 11699–11706.

Kui Wong N, Easton RL, Panico M, Sutton-Smith M, Morrison JC, Lattanzio FA, Morris HR, Clark GF, Dell A and Patankar MS (2003) Characterization of the oligosaccharides associated with the human ovarian tumor marker CA125. *The Journal of Biological Chemistry*, **278**(31), 28619–28634.

Kuraya N and Hase S (1996) Analysis of pyridylaminated O-linked sugar chains by two-dimensional sugar mapping. *Analytical Biochemistry*, **233**(2), 205–211.

Küster B, Wheeler SF, Hunter AP, Dwek RA and Harvey DJ (1997) Sequencing of N-linked oligosaccharides directly from protein gels: in-gel deglycosylation followed by matrix-assisted laser desorption/ionization mass spectrometry and normal-phase high-performance liquid chromatography. *Analytical Biochemistry*, **250**(1), 82–101.

Litynska A, Przybylo M, Ksiazek D and Laidler P (2000) Differences of alpha3beta1 integrin glycans from different human bladder cell lines. *Acta Biochimica Polonica*, **47**(2), 427–434.

Malhotra R, Wormald MR, Rudd PM, Fischer PB, Dwek RA and Sim RB (1995) Glycosylation changes of IgG associated with rheumatoid arthritis can activate complement via the mannose-binding protein. *Nature Medicine*, **1**(3), 237–243.

Mattu TS, Royle L, Langridge J, Wormald MR, Van den Steen PE, Van Damme J, Opdenakker G, Harvey DJ, Dwek RA and Rudd PM (2000) O-glycan analysis of natural human neutrophil gelatinase B using a combination of normal phase-HPLC and online tandem mass spectrometry: implications for the domain organization of the enzyme. *Biochemistry*, **39**(51), 15695–15704.

Merry AH, Neville DC, Royle L, Matthews B, Harvey DJ, Dwek RA and Rudd PM (2002) Recovery of intact 2-aminobenzamide-labeled o-glycans released from glycoproteins by hydrazinolysis. *Analytical Biochemistry*, **304**(1), 91–99.

Navazio L, Miuzzo M, Royle L, Baldan B, Varotto S, Merry AH, Harvey DJ, Dwek RA, Rudd PM and Mariani P (2002) Monitoring endoplasmic reticulum-to-Golgi traffic of a plant calreticulin by protein glycosylation analysis. *Biochemistry*, **41**(48), 14141–14149.

Ohyama C, Hosono M, Nitta K, Oh-Eda M, Yoshikawa K, Habuchi T, Arai Y and Fukuda M (2004) Carbohydrate structure and differential binding of prostate specific antigen to Maackia amurensis lectin between prostate cancer and benign prostate hypertrophy. *Glycobiology*, **14**(8), 671–679.

Parekh RB, Dwek RA, Rudd PM, Thomas JR, Rademacher TW, Warren T, Wun TC, Hebert B, Reitz B, Palmier M, *et al*. (1989) N-glycosylation and in vitro enzymatic activity of human recombinant tissue plasminogen activator expressed in Chinese hamster ovary cells and a murine cell line. *Biochemistry*, **28**(19), 7670–7679.

Parekh RB, Dwek RA, Sutton BJ, Fernandes DL, Leung A, Stanworth D, Rademacher TW, Mizuochi T, Taniguchi T, Matsuta K, *et al*. (1985) Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG. *Nature*, **316**(6027), 452–457.

Patel T, Bruce J, Merry A, Bigge C, Wormald M, Jaques A and Parekh R (1993) Use of hydrazine to release in intact and unreduced form both N- and O-linked oligosaccharides from glycoproteins. *Biochemistry*, **32**(2), 679–693.

Peracaula R, Royle L, Tabares G, Mallorqui-Fernandez G, Barrabes S, Harvey DJ, Dwek RA, Rudd PM and de Llorens R (2003a) Glycosylation of human pancreatic ribonuclease: differences between normal and tumor states. *Glycobiology*, **13**(4), 227–244.

Peracaula R, Tabares G, Royle L, Harvey DJ, Dwek RA, Rudd PM and de Llorens R (2003b) Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins. *Glycobiology*, **13**(6), 457–470.

Prime S and Merry T (1998) Exoglycosidase sequencing of N-linked glycans by the reagent array analysis method (RAAM). *Methods in Molecular Biology*, **76**, 53–69.

Robbins PW, Wirth DF and Hering C (1981) Expression of the streptomyces enzyme endoglycosidase H in *Escherichia coli*. *The Journal of Biological Chemistry*, **256**(20), 10640–10644.

Royle L, Mattu TS, Hart E, Langridge JI, Merry AH, Murphy N, Harvey DJ, Dwek RA and Rudd PM (2002) An analytical and structural database provides a strategy for sequencing o-glycans from microgram quantities of glycoproteins. *Analytical Biochemistry*, **304**(1), 70–90.

Royle L, Roos A, Harvey DJ, Wormald MR, van Gijlswijk-Janssen D, Redwan el RM, Wilson IA, Daha MR, Dwek RA and Rudd PM (2003) Secretory IgA N- and O-glycans provide a link between the innate and adaptive immune systems. *The Journal of Biological Chemistry*, **278**(22), 20140–20153.

Rudd PM, Morgan BP, Wormald MR, Harvey DJ, van den Berg CW, Davis SJ, Ferguson MA and Dwek RA (1997) The glycosylation of the complement regulatory protein, human erythrocyte CD59. *Journal of Biological Chemistry*, **272**(11), 7229–7244.

Strecker G, Pierce-Cretel A, Fournet B, Spik G and Montreuil J (1981) Characterization by gas–liquid chromatography–mass spectrometry of oligosaccharides resulting from the hydrazinolysis-nitrous acid deamination reaction of glycopeptides. *Analytical Biochemistry*, **111**(1), 17–26.

Strecker G, Wieruszeski JM, Plancke Y and Boilly B (1995) Primary structure of 12 neutral oligosaccharide-alditols released from the jelly coats of the anuran xenopus laevis by reductive beta-elimination. *Glycobiology*, **5**(1), 137–146.

Sutton-Smith M, Morris HR, Grewal PK, Hewitt JE, Bittner RE, Goldin E, Schiffmann R and Dell A (2002) MS screening strategies: investigating the glycomes of knockout and myodystrophic mice and leukodystrophic human brains. *Biochemical Society Symposium*, **69**, 105–115.

Takahashi N, Wada Y, Awaya J, Kurono M and Tomiya N (1993) Two-dimensional elution map of GalNAc-containing N-linked oligosaccharides. *Analytical Biochemistry*, **208**(1), 96–109.

Takeuchi M, Takasaki S, Miyazaki H, Kato T, Hoshi S, Kochibe N and Kobata A (1988) Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant Chinese hamster ovary cells. *The Journal of Biological Chemistry*, **263**(8), 3657–3663.

Taverna M, Tran NT, Merry T, Horvath E and Ferrier D (1998) Electrophoretic methods for process monitoring and the quality assessment of recombinant glycoproteins. *Electrophoresis*, **19**(15), 2572–2594.

Tekoah Y, Ko K, Koprowski H, Harvey DJ, Wormald MR, Dwek RA and Rudd PM (2004) Controlled glycosylation of therapeutic antibodies in plants. *Archives of Biochemistry and Biophysics*, **426**(2), 266–278.

Tomiya N, Awaya J, Kurono M, Endo S, Arata Y and Takahashi N (1988) Analyses of N-linked oligosaccharides using a two-dimensional mapping technique. *Analytical Biochemistry*, **171**(1), 73–90.

Tran NT, Taverna M, Merry AH, Chevalier M, Morgant G, Valentin C and Ferrier D (2000) A sensitive mapping strategy for monitoring the reproducibility of glycan processing in an

HIV vaccine, RGP-160, expressed in a mammalian cell line. *Glycoconjugate Journal*, **17**(6), 401–406.

Waddling CA, Plummer TH, Tarentino AL and Van Roey P, Jr (2000) Structural basis for the substrate specificity of endo-beta-N-acetylglucosaminidase F(3). *Biochemistry*, **39**(27), 7878–7885.

Yamashita K, Ohkura T, Tachibana Y, Takasaki S and Kobata A (1984) Comparative study of the oligosaccharides released from baby hamster kidney cells and their polyoma transformant by hydrazinolysis. *The Journal of Biological Chemistry*, **259**(17), 10834–10840.

Basic Techniques and Approaches

Glycosylphosphatidylinositol anchors – a structural perspective

Angela Mehlert
University of Dundee, Dundee, UK

1. Introduction

The fundamental role of glycosylphosphatidylinositol (GPI) anchors is to hold a protein onto the external surface of the plasma membrane of a cell, although many other functions have been suggested. A huge number of GPI-anchored proteins have been identified; they are a functionally diverse group and are ubiquitous in eukaryotic cells, particularly protozoa. There are also many instances in the protozoa where GPI structures are found on the parasites' cell surface but they are not attached to a protein (see McConville and Ferguson (1993), Branquinha *et al*. (1999), and Priest *et al*. (2003)).

2. Basic structure of GPIs

The first full structural characterization of a GPI, that of the variant surface glycoprotein (VSG) of *Trypanosoma brucei* and the result of about 6 years of work, was reported in 1988 (Ferguson *et al*., 1988), and the reason it took so long was the extraordinarily complex nature of the molecule. The GPI core is made up of three mannose residues linked to glucosamine that is joined to a *myo*-inositol residue. The inositol residue is part of a phosphatidylinositol (PI) or insitol-phosphoceramide (IPC) phospholipid that holds the protein next to the membrane. The signal peptide sequence required for a GPI to be attached to a protein consists of 15–20 hydrophobic residues followed by two small polar residues immediately before the GPI attachment site. An ethanolamine phosphate links the carboxy terminus of the protein to the third mannose residue of the GPI core. This basic structure may have various other groups attached, such as extra ethanolamine phosphate groups and carbohydrate side chains. Thus, the inherent complexity of GPIs, that include carbohydrate, lipid, phosphate, and peptide components, makes the structural analysis relatively difficult.

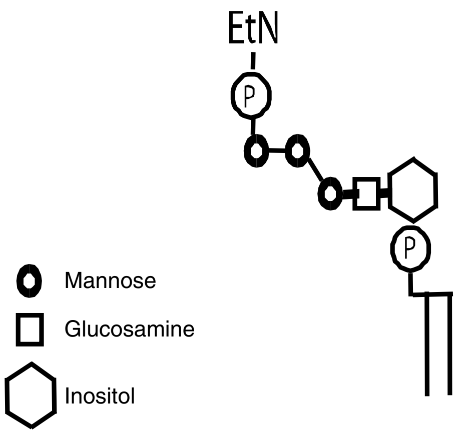


Figure 1 The core structure of a GPI anchor

3. Elucidation of the first GPI structures

The first anchors were characterized by basic structural chemistry techniques including ^{1}H NMR (nuclear magnetic resonance), gas chromatography-mass spectrometry (GC-MS) for composition and methylation linkage analysis, and specific enzymatic and chemical cleavages. The amount of protein needed for this characterization was substantial (over a hundred milligram of VSG was used) and although we now have access to much more sensitive instrumentation and specialized techniques, the availability of purified protein is still the main problem to be overcome in GPI analysis. However, with one exception, all protein-linked GPI anchors contain a core consisting of a minimum of three mannosyl residues in the following sequence:

EtN*P*-6Manαl-2Manαl-6Manαl-4GlcNαl-6(PI), see Figure 1, and this basic core is now relatively easy to confirm.

4. Procedures for GPI analysis

There are several chemical and enzymatic reagents that can be used for the selective cleavage of a GPI anchor structure. Some of these were originally used to help determine the early structures and are now applied to confirm the presence of a GPI anchor (Ferguson, 1993). They may also be used, in conjunction with chromatography or high performance thin layer chromatography (HPTLC), to obtain some structural information from native or [^{3}H]mannose-, [^{3}H]glucosamine-, [^{3}H]inositol-, or [^{3}H]fatty acid–radiolabeled GPI-anchored proteins or their GPI biosynthetic intermediates.

The most important analytical procedure is the nitrous acid deamination of the glucosamine residue of a GPI. This relatively gentle reaction (room temperature, pH 4.0) is highly selective for the glucosamine-inositol glycosidic bond because

it is dependent on the free amino group of the glucosamine residue. Apart from in GPIs, nonsubstituted glucosamines are uncommon in nature, the amino group usually being blocked by a sulfate or an acetate residue so there is little or no background. The deamination generates a free reducing terminus on the GPI glycan, a 2,5-anhydromannose. This reducing sugar can then be reduced to [1-^{3}H]2,5-anhydromannitol (AHM) by reduction with sodium borotritiide, thus introducing a radiolabel that aids further purification and analysis. Alternatively, the reducing terminus may be reductively aminated and attached to a fluorophore such as 2-amino-benzamide (2-AB). This technique may be easily performed using a 2-AB glycan labeling kit available commercially and has the advantage that it is simple to purify the fluorescently labeled glycans from the other products of the reaction by gel filtration or normal-phase HPLC. However, the cleanup of reduced and tritiated glycans requires removal of the substantial radiochemical contamination that is always present when using high specific activity $NaB^{3}H_4$. This cleanup procedure includes downward paper chromatography, followed by passage of the material through a mixed bed column in water (Ferguson, 1993).

5. Labeling of GPI anchors

When the GPI glycan is radioactively or fluorescently labeled and dephosphorylated with aqueous hydrogen fluoride (HF), it may be conveniently sequenced using exoglycosidases and HPTLC (Schneider and Ferguson, 1995). It is also possible to adapt the above protocol using only an aliquot of the glycan to label, the remaining material being reduced using $NaB^{2}H_4$. This allows a larger preparation to be handled while keeping the amount of radioactivity used at a reasonable level. The lipid moiety of the GPI is liberated by the nitrous acid reaction and this PI can be isolated by solvent extraction, purified on a micro silica column, and analyzed by negative ion tandem ES-MS (electrospray-mass spectrometry), as described by Thomson *et al*. (2002).

As mentioned earlier, the first anchors were characterized by NMR, which remains the only incontrovertible way to assign the anomericity of the glycosidic linkages. The easier option of specific cleavages achieved by either selective enzymatic or chemical techniques may be available for most of those linkages so far described, but if a negative result is obtained, or worse, a false-positive, mistakes can be made. However, obtaining the amount of material required for NMR analysis is not trivial. The use of a Shigemi tube and a cryoprobe can increase the sensitivity of the NMR up to tenfold, but the simplest structure to be solved would still require about 10 nmoles of purified material (personal observation). Also, there is no GPI NMR spectra reference database analogous to the one for N-linked oligosaccharides. However, if sufficient material and expertise is available, the first step in the analysis of a water soluble GPI would ideally be NMR in D_2O. NMR is a nondestructive technique, which means that further analysis may be done using the same sample, although it will probably contain contaminants that are present in even the best commercially available D_2O. Gel filtration in water will return the sample to a condition where it is suitable for further analysis.

6. Microscale analysis of GPIs

In reality, most analyses have to be done on a fraction of the material needed for NMR analysis and this has driven the search for other techniques. In the Ferguson lab in Dundee, progress has been made toward deriving a reliable microscale analysis regime, and methods are constantly being refined. Developments in mass spectrometry (MS) instrumentation and techniques have meant that although in the Fankhauser *et al*. paper (1993), the FAB (fast atom bombardment) MS of a permethylated GPI glycan could only be performed on 20 nmoles of material, but recently some proteomic-like analysis methods of GPI analysis have been published (Fontaine *et al*., 2003; Thomson *et al*., 2002; Priest *et al*., 2003). This approach is based on the use of partially purified proteins run on SDS-PAGE (sodium dodecyl sulphate-polyacrylamide gel electrophoresis), then Western blotted onto PVDF (polyvinylidene difluoride) membranes. These techniques overcome the problem of purifying large amounts of material and take advantage of the nonsubstituted glucosamine of GPIs that, as explained earlier, allows for specific labeling of the 2,5-AHM residue and release of the PI moiety. The ethanolamine phosphate connection, from the third mannose residue to the C-terminus of the protein, can then be exploited to release the labeled glycans from the protein and, therefore, from the PVDF membrane, by aqueous HF dephosphorylation. After digestion, the GPI will remain in the HF solution and the PVDF membrane can be removed. It appears that if there is enough protein for visualization by amido black on the membrane, it should be possible to deduce the basic structure of the molecule.

Unfortunately, using aqueous HF precludes the analysis of any other groups attached to the glycans through phosphate groups, for instance, other ethanolamine phosphate groups, will be lost. Also, the assumption has to be made that the ethanolamine phosphate bridge between the C-terminal of the protein and the GPI glycan is made to the third mannose, which has proved the case for all of the GPI structures solved so far.

Matrix assisted laser desorption ionization-time of flight (MALDI-TOF) analysis of sugars using various matrices can also be a useful technique. We have had most success with 2,5-dihydroxybenzoic acid as matrix, usually including trifluoroacetic acid in the solution to aid ionization. Glycolipids and mucin molecules may also be successfully analyzed by this technique (Almeida *et al*., 2000). Other cases where the amount of material has been limited have been resolved using permethylation of the neutral glycans prepared by HF digestion and then analyzed using positive ES-MS and ES-MS-CID (collision induced dissociation)-MS (Priest *et al*., 2003).

7. Proteomics type analysis of GPI

We have recently worked out a technique to help define whether a protein is GPI anchored by adapting the SDS-PAGE Western blotting technique, cutting out the band in question along with 2 control blank pieces of blot and a positive control (usually VSG). The PVDF membrane pieces are hydrolyzed in 6 M aqueous HCl

with a known amount of internal standard, as in our standard inositol analyses (Ferguson, 1993). The hydrolyzed material is dried and derivatized with tri-methylsilyl reagent, then analyzed by GC-MS using a SIM (selective ion monitoring) method, which scans for the ions characteristic of inositols (m/z 305 and 318). This method is very destructive on the GC column and the MS ion source will become contaminated from the acid hydrolysis of the PVDF membrane; however, it has proved useful in some cases where the amount of material available precluded the use of other techniques.

References

Almeida IC, Camargo MM, Procopio DO, Silva LS, Mehlert A, Travassos LR, Gazzinelli RT and Ferguson MAJ (2000) Highly purified glycosylphosphatidylinositols from *Trypanosoma cruzi* are potent proinflammatory agents. *EMBO Journal*, **19**, 1476–1485.

Branquinha MH, Vermelho AB, Almeida IC, Mehlert A and Ferguson MAJF (1999) Structural studies on the polar glycoinositol phospholipids of *Trypanosoma (Schizotrypanum) dionisii* from bats. *Molecular and Biochemical Parasitology*, **102**, 179–189.

Fankhauser C, Homans SW, Thomas-Oates JE, McConville MJ, Desponds C, Conzelmann A and Ferguson MAJ (1993) Structures of the glycosylphosphatidylinositol membrane anchors from *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, **268**, 26365–26374.

Ferguson MAJ, Homans SW, Dwek RA and Rademacher TW (1988) Glycosylphosphatidylinositol moiety that anchors *Trypanosoma brucei* Variant surface glycoprotein to the membrane. *Science*, **239**, 753–759.

Ferguson MAJ (1993) GPI membrane anchors: isolation and analysis. In *Glycobiology: A Practical Approach*, Fukuda M and Kobata A (Eds.), IRL Press at Oxford University Press: Oxford, pp. 349–383.

Fontaine T, Magnin T, Mehlert A, Lamont D, Latge J-P and Fergsuon MAJ (2003) Structures of the glycosylphosphatidylinositol membrane anchors from *Aspergillus fumigatus* membrane proteins. *Glycobiology*, **13**, 169–177.

McConville M and Ferguson MAJ (1993) The structure, biosynthesis and function of glycosylated phosphatidylinositols in the parasitic protozoa and higher eukaryotes. *Biochemical Journal*, **294**, 305–324.

Priest JW, Mehlert A, Arrowood MJ, Riggs MW and Ferguson MAJ (2003) Characterization of a low molecular weight glycolipid antigen from *Cryptosporidium parvum*. *Journal of Biological Chemistry*, **278**, 52212–52222.

Schneider P and Ferguson MAJ (1995) Microscale analysis of glycophosphatidylinositol structures. *Methods in Enzymology*, **250**, 614–630.

Thomson LM, Lamont DJ, Mehlert A, Barry JD and Ferguson MAJ (2002) Partial structure of Glutamic acid and Alanine-rich Protein, a major surface glycoprotein of the insect stages of *Trypanosoma congolense*. *Journal of Biological Chemistry*, **277**, 48899–48904.

Introductory Review

Classification of proteins into families

Nicola J. Mulder and Rolf Apweiler
European Bioinformatics Institute, Cambridge, UK

1. Introduction

In order to provide a review on the classification of proteins into families, it is necessary to describe what a protein family is and why and how they should be classified into families. A protein family has a number of different definitions, but the basic concept agreed upon is that it is a group of evolutionarily related proteins. The Dictionary of Cell and Molecular Biology (Lackie and Dow, 1999) **http://www.mblab.gla.ac.uk/∼julian/Dict.html**) describes a gene family as: "a set of genes coding for diverse proteins which, by virtue of their high degree of sequence similarity, are believed to have evolved from a single ancestral gene". Therefore, the basis for classifying proteins into families is their sequence similarity. A protein may be related to another very specifically at the subfamily level, it can be related to more diverse proteins at a family level, and it is then related to even more diverse proteins at the superfamily level. The number of common properties of the proteins at each level increases toward the subfamily level.

Why the interest in protein classification into families? Genome sequences of a large range of different organisms are flooding into the public databases. The sequence data is largely entering the databases unannotated and requiring further analysis to functionally characterize the proteins. DNA and protein sequences have traditionally been analyzed by comparing new sequences to existing ones of known function. This is all very well if the existing ones have a known function, but the flood of new sequences has resulted in an increased proportion of unknown to known sequences in the databases. Therefore, new sequences have a high likelihood of not receiving annotation from a sequence similarity search. It thus follows that the more characterized sequences in the database, the more efficiently new sequences will be functionally annotated. This presents a chicken and egg situation – where to start in the circle! However, as mentioned previously, protein sequences may be classified on different family levels in which certain functional characteristics are conserved. The possibility of classifying proteins into families thus presents an opportunity for inference of a conserved function on all members of the family and thus increasing the pool of characterized proteins in the database.

There are two main methods for classification of proteins into families: sequence clustering and protein signatures. Methods of clustering related protein sequences

by their similarities are well established and quick, but have their drawbacks. An important factor is the low probability of detecting distant relationships between new and existing sequences of the same protein family. Bioinformatics has evolved rapidly since its initiation, which has led to the development of new tools to solve these problems. A number of scientists in this discipline have independently developed the concept of protein signatures as a means to protein functional annotation by exploiting known similarities between related protein sequences and using mathematics to describe them. A signature is a "description" of an entity and it defines the characteristics associated only with that entity. Sequence clustering and protein signatures will be discussed in detail in this chapter.

2. Sequence clustering

Sequence clustering methods are generally entirely automated, and make the assumption that members of a protein family will cluster together on the basis of sequence similarity. An example of a database that uses this method is ProDom (Corpet *et al*., 2000). The general philosophy of ProDom construction is that autonomously evolving protein domains can be recognized as shuffled in present day sequences by sequence comparison methods. It takes all proteins in UniProt (the successor of Swiss-Prot, TrEMBL (Boeckmann *et al*., 2003), and PIR) (Apweiler *et al*., 2004), removes fragments, identifies the smallest remaining sequence, and uses this as a query sequence to search the protein database using PSI-BLAST (Altschul *et al*., 1997). The sequences that hit are made into a new ProDom domain family and removed from the protein database. The remaining sequences are once again sorted by size, and the process is repeated until completion. In this way, ProDom groups all the nonfragment sequences in Swiss-Prot and TrEMBL into more than 150 000 families. Another major alignment database is PIR-ALN (Srinivasarao *et al*., 1999), a database of annotated protein sequence alignments derived automatically from the PIR sequence database, with alignments at the superfamily and domain level. Other widely used cluster databases are ProtoMap (Yona *et al*., 2000), Systers (Krause *et al*., 2000), and CluSTr (Kriventseva *et al*., 2003). All these databases offer an automatic classification of all UniProt proteins into groups of related proteins on the basis of pairwise similarities. For more details on sequence clustering, *see* Article 92, **Classification of proteins by clustering techniques**, Volume 6.

3. Protein signatures and databases

There is a certain amount of inherent information in a protein sequence, but from one sequence it is difficult to discover what this information describes. However, if a number of related sequences are aligned, conserved regions can be identified, which may be an indication of function. These conserved areas of a protein family, domain, or functional site can be used to develop a description of the family using several different methods, including regular expressions (for patterns of conserved residues), profiles, and hidden Markov models (HMMs). The

Article 91, **Classification of proteins by sequence signatures**, Volume 6 gives a thorough introduction into the different methods used by the major protein signature databases available to the public.

4. Regular expressions

Regular expressions (or patterns) describe a group of amino acids that constitute a usually short, characteristic motif within a protein sequence. These motifs arise because these region(s) of a protein are important for activity.

PROSITE (Hulo *et al*., 2004) is a database that uses such patterns. These are built from sequence alignments of related sequences, which are taken from either a well-characterized protein family derived from the literature or from the result of sequence searches against Swiss-Prot and TrEMBL. The alignments are scanned for conserved regions involved in the catalytic activity or substrate binding. Patterns derived from these regions specify which amino acid(s) may or may not occur at each position. The syntax of a PROSITE pattern can be illustrated by the following hypothetical pattern: <M-G-x(3)-[IV]2-x-{FWY}, which is restricted at the N-terminus of a sequence (<) and translated as Met- Gly-any residue-any residue-any residue-[Ile or Val]-[Ile or Val]-any residue-{any residue but Phe or Trp or Tyr}. Patterns have many advantages, but they also have their limitations across whole sequences and for more divergent proteins.

5. Profiles

A profile is a table of position-specific amino acid weights and gap costs. The table describes the probability of finding a particular amino acid at a given position in the sequence (Gribskov *et al*., 1990). The numbers in the table (scores) are used to calculate similarity scores between a profile and a sequence for a given alignment. For each set of sequences, a threshold score is calculated to differentiate between true and nontrue matches to the profile.

PROSITE creates profiles for their database to complement the patterns (Bucher *et al*., 1996). For these, they also start with multiple sequence alignments, and use a symbol comparison table to convert residue frequency distributions into weights, resulting in a table of position-specific weights (Gribskov *et al*., 1990). A similarity score is calculated between the profile and sequences in Swiss-Prot, and the profile is refined until only the intended set of protein sequences scores above the threshold score for the profile. For a detailed description of PROSITE, *see* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6.

PRINTS (Attwood *et al*., 2003) uses the concept of profiles in a different way to develop their database of "fingerprints". A fingerprint is a group of conserved motifs used to characterize a protein family. Instead of focusing on small conserved areas only, the occurrence of these conserved areas across the whole sequence is taken into account. Profiles are built from a multiple sequence alignment for small conserved regions in the sequence, and together make up a fingerprint.

During creation of the fingerprints, each motif is used to scan the protein sequence database and the hit lists are correlated to add sequences to the original alignment. New motifs are then generated and the process is repeated until convergence. Recognition of individual elements in the fingerprint is mutually conditional, and true members match all elements in order while members of a subfamily may match part of the fingerprint. Many fingerprints have been created to identify proteins at the superfamily as well as the family and subfamily levels. For this reason, many of the fingerprints are related to each other in an ordered hierarchical structure. *See* Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6 for a detailed description of the PRINTS database.

6. Hidden Markov models

HMMs (Krogh *et al*., 1994) are statistical models that are based on probabilities rather than on scores. The HMMER package, written by Sean Eddy (**http://hmmer.wustl.edu/**), is based on Bayesian statistical models and is the major implementation of HMMs. It allows users to create an HMM from a sequence alignment and to search a database of sequences against the HMM. Many databases, for example, Pfam (Bateman *et al*, 2004), SMART (Letunic *et al*., 2004), TIGRFAMs (Haft *et al*., 2003), PIR SuperFamily (Wu *et al*., 2004), and SUPERFAMILY (Madera *et al*., 2004) use HMMs implemented in their own interpretations.

Pfam (Bateman *et al*., 2004) is a large collection of protein families available via the web and in flat file format. There are two sections in the Pfam database, PfamA, which is a set of high-quality manually annotated models, and PfamB, which has higher coverage, but is fully automated with no manual intervention. PfamA entries initiate from a seed alignment of the protein sequence region covered by the family or domain, which is trimmed manually to ensure accurate domain boundaries. A full alignment and an HMM are then created from the final alignment and annotation is added. PfamB is created from automatic clustering of the protein sequences in Swiss-Prot and TrEMBL by ProDom with PfamA members removed. Pfam has a high coverage of sequence space and is used extensively for the annotation of genomes. The chapter Article 86, **Pfam: the protein families database**, Volume 6 gives an in-depth description of Pfam.

SMART (a Simple Modular Architecture Research Tool: Letunic *et al*., 2004) is a database that develops HMMs to facilitate the identification and annotation of genetically mobile domains and the analysis of domain architectures. These are built from hand-curated multiple sequence alignments of representative family members based on tertiary structures where possible, otherwise those found by PSI-BLAST (Altschul *et al*., 1997). The models created are used to search the database for additional members to be included in the sequence alignment. This iterative process is repeated until no further homologs are detected. SMART focuses on models for domains found in signaling, extracellular, and chromatin-associated proteins, and provides useful tools for the analysis of protein domain composition.

TIGRFAMs (Haft *et al*., 2003) are a collection of protein families created to assist in sequence annotation, particularly from microbial genomes. The focus is

on HMMs that group homologous proteins conserved with respect to function. The models are produced in a similar way to those in Pfam and SMART, but should only hit equivalogs, that is, proteins that have been shown to have the same function. However, where the biology of a protein family does not support the construction of equivalog models, superfamily HMMs are developed and a hierarchy may exist. TIGRFAMs are a useful complement to Pfam where the latter may only have generalized models for a protein family not necessarily restricted to functional equivalents. A detailed description of TIGRFAMs is available in the chapter Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6.

PIR SuperFamily (Wu *et al*., 2004), known as *PIRSF*, is a protein classification system based on evolutionary relationships between whole proteins. Preliminary PIRSF clusters are computationally defined using both pairwise-based and cluster-based parameters. In each cluster, proteins sharing the same domain composition and similar lengths are manually identified and grouped as curated families. HMMs are created from these families to populate the PIRSF database. A new sequence must be matched by the HMM and satisfy the length restriction to be a member of a family. The focus on classification of whole proteins, not individual domains, allows annotation of specific biological functions of proteins as well as of generic biochemical functions. PIRSF is described in more detail in the chapter Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6.

SUPERFAMILY (Madera *et al*., 2004) is a database of HMMs based on protein structure rather than sequence. The HMMs in the database represent all proteins of known structure, and are based on the SCOP (Andreeva *et al*., 2004) classification of proteins. Each model corresponds to a SCOP domain, and aims at representing an entire superfamily. Many Pfam domains are also based on SCOP domains; however, structural families often differ from sequence-based ones since similar sequences may have similar structures, but the reverse is not always true. SUPERFAMILY has been applied to whole genomes and is used for genome comparisons and analysis of the distribution of structural families in the genomes.

7. Integration of the databases

Each of the methods for protein family classification has their advantages and disadvantages. Patterns are easy to generate and useful for small conserved regions like active sites or binding sites, but fail to provide information about the rest of the sequence, and may reject significant family members. Profiles and HMMs are efficient at identifying divergent members, but HMMs, in particular, are slow to develop and use. PRINTS fingerprints are strong in detecting families on different levels, but are weak in domains and sites. The ideal protein classification system would use all of the databases together.

Many of them already have interactions and exchange of data between them. For example, some use ProDom families as a starting point for their signatures, and many cross-reference each other where applicable. The obvious solution has recently been provided by several groups, which have made an effort to integrate some or all of these databases into a single coherent protein signature resource.

The integrated database resources include MetaFam (Silverstein *et al*., 2001), CDD (Wheeler *et al*., 2001) and InterPro (Mulder *et al*., 2003). MetaFam automatically create supersets of overlapping families from Blocks+, DOMO, Pfam, PIR-ALN, PRINTS, Prosite, ProDom, and SBASE (Murvai *et al*., 2001) using set theory to compare these databases. They provide reference domains covering the total area. CDD is a database of domains derived from SMART, Pfam, and contributions from NCBI LOAD (Library Of Ancient Domains). These integrated resources all use automatic methods, and in some cases simply provide the capability to search the individual databases from one page. InterPro, on the other hand, is produced by the members of the individual protein signature databases, and the integration process is manual. In the database, the similarities and differences between the contributing member database signatures are rationalized, so a user is clear on how to fathom multiple hits to a protein sequence. This makes InterPro a powerful resource that is greater than the sum of the individual components of which it is comprised. InterPro is discussed in detail in the chapter Article 83, **InterPro**, Volume 6.

8. Discussion

The classification of proteins into families is the first step in the process of functionally annotating unknown proteins in the database. The methods and databases described in this chapter are crucial to this process, and although they have developed independently they have involuntarily converged to provide a formidable and reliable combination for use in the automatic annotation of protein sequences. InterPro unites them to exploit their strengths and compensate for potential weaknesses and provides the bench scientists and genome sequencing centers with a gateway for analyzing their sequences of interest.

InterPro entries, through the protein signatures that describe them, group proteins into families or those with common domains. Once protein sequences are classified into families, if one or more members are well characterized, it is possible to make assumptions about the function of all the other members. Having a defined method for grouping the proteins means that new protein sequences entering the database can automatically be assigned to their relevant protein families without having to recreate the methods themselves. The result is a powerful tool for classifying and characterizing proteins that is scalable for application to whole genomes of many diverse organisms. As new protein families emerge, methods for classifying them can be developed until nearly all protein families in biological space are represented. It is important to note, however, that protein classification methods provide useful and often very accurate predictions of protein function, but these should ultimately be verified by experimental evidence.

Related articles

Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 85,

The PRINTS protein fingerprint database: functional and evolutionary applications, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 92, **Classification of proteins by clustering techniques**, Volume 6

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family. *Nucleic Acids Research*, **32**(1), D226–D229.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, **32**(1), D115–D119.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, *et al*. (2003) PRINTS and its automatic supplement pre-PRINTS. *Nucleic Acids Research*, **31**(1), 400–402.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**(1), D138–D141.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

Bucher P, Karplus K, Moeri N and Hofmann K (1996) A flexible motif search technique based on generalized profiles. *Computational Chemistry*, **20**(1), 3–23.

Corpet F, Servant F, Gouzy J and Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, **28**, 267–269.

Gribskov M, Luthy R and Eisenberg D (1990) Profile analysis. *Methods in Enzymology*, **183**, 146–159.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**(1), 134–137.

Krause A, Stoye J and Vingron M (2000) The SYSTERS protein sequence cluster set. *Nucleic Acids Research*, **28**(1), 270–272.

Kriventseva EV, Servant F and Apweiler R (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Research*, **31**(1), 388–389.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, **235**(5), 1501–1531.

Lackie JM and Dow JAT (1999) *The Dictionary of Cell and Molecular Biology*, Third Edition, Academic Press: London.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**(1), D142–D144.

Madera M, Vogel C, Kummerfeld SK, Chothia C and Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**(1), D235–D239.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**(1), 315–318.

Murvai J, Vlahovicek K, Barta E and Pongor S (2001) The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments. *Nucleic Acids Research*, **29**(1), 58–60.

Silverstein KA, Shoop E, Johnson JE and Retzel EF (2001) MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics*, **17**(3), 249–261.

Srinivasarao GY, Yeh LS, Marzec CR, Orcutt BC and Barker WC (1999) PIR-ALN: a database of protein sequence alignments. *Bioinformatics*, **15**(5), 382–390.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, *et al*. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **29**(1), 11–16.

Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, *et al*. (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research*, **32**(1), D112–D114.

Yona G, Linial N and Linial M (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research*, **28**(1), 49–55.

Specialist Review

Metalloproteins

Kirill Degtyarenko
European Bioinformatics Institute, Cambridge, UK

1. Introduction

Metalloproteins are an immensely diverse class of biological coordination compounds. The most important biochemical processes, including respiration, oxygenic photosynthesis, regulation of transcription and translation, nitrogen fixation, and metabolism of xenobiotics, depend on metalloproteins. Metalloenzymes were the first biological catalysts on Earth. About one-third of all structurally characterized proteins contain metals, while over 50% of all proteins are estimated to be metalloproteins.

From the first studies on hemoglobin onward, spectroscopic and other physical methods have played a crucial role in research into metalloproteins. Many of them possess distinctive colors, which explains why metalloproteins attracted attention early in the history of biochemistry. Indeed, polypeptides absorb light mostly in the ultraviolet region, and the colors of proteins are due to nonpeptide chromophores. The specific absorption maxima (in nanometers) were historically used in the names of chromophores and/or proteins, as in cytochrome b_{562}, cytochrome c_{554}, cytochrome P450, coenzyme F430, P680, and P700 centers of photosystems and so on, while other metalloproteins were given such colorful names as “azurin” and “purple acid phosphatase” (in this latter example, purple is the color of the enzyme, not of the acid). Incidentally, metalloenzyme urease was the first enzyme ever to be crystallized; later, myoglobin became the first protein to have a known three-dimensional (3D) atomic structure. Table 1 puts some major landmarks in the history of physics and chemistry into the context of metalloprotein science.

Many aspects of metalloprotein structure and function remain obscure. One of these concerns metal selectivity. It is not clear why some proteins display an almost absolute preference for one metal over another, for example, in tungsten and molybdenum enzymes (L’vov *et al*., 2002), while others easily incorporate “nonnative” metals. In the metalloprotein world, sequence homology does not guarantee similarity, at least at the level of the active centers. It is not surprising then that the traditional bioinformatics “approach” is to view metals as an irritating complication best to avoid.

Paradoxically, apparently complicated metalloproteins may hold the key to the solution of one of the most challenging problems of computational biology: the protein-folding problem. It is known that metal ions play a crucial role in proteinfolding and stabilization. Moreover, the evidence grows that successful

Table 1 Landmarks in the history of metalloproteins

Year		Nobel prize
1862	Ernst Felix Hoppe-Seyler (1825–1895) observes the absorption spectrum of hemoglobin.	
1864	Hoppe-Seyler names and crystallizes hemoglobin.	
1879	Hoppe-Seyler proposes that chemical structures of hemin and chlorophyll are similar.	
1886	Charles Alexander MacMunn (1852–1911) suggests that tissues contain heme pigments other than hemoglobin.	
1892	Alfred Werner (1866–1919) develops the coordination theory of inorganic compounds.	Chemistry 1913
1894	Gabriel Bertrand (1867–1962) describes laccase.	
1895	Wilhelm Conrad Röntgen (1845–1923) discovers X-radiation.	Physics 1901
1895	Gabriel Bertrand describes tyrosinase.	
1903	Mikhail Tswett (1872–1919) separates the plant pigments including chlorophylls with the help of his new method that he calls "chromatography".	
1912	Max von Laue (1879–1960) discovers the diffraction of X rays by crystals.	Physics 1914
1912	William Henry Bragg (1862–1942) develops the X-ray spectrometer; his son, William Lawrence Bragg (1890–1971), discovers the law of X-ray diffraction (the Bragg law).	
1913	W.H. Bragg and W.L. Bragg solve the first crystal structure: sodium chloride.	Physics 1915
1913	Richard Martin Willstätter (1872–1942) determines the chemical structure of chlorophyll.	Chemistry 1915
1924	Otto Heinrich Warburg (1883–1970) describes the *Atmungsferment* (cytochrome *c* oxidase).	Medicine 1931
1925	David Keilin (1887–1963) names cytochromes.	
1927	The Svedberg (1884–1971) measures the molecular weight of hemoglobin by ultracentrifugation.	Chemistry 1926
1926	James Batcheller Sumner (1887–1955) obtains the first crystalline enzyme, urease.	Chemistry 1946
1928	Chandrasekhara Venkata Raman (1888–1970) discovers the light scattering effect called after him.	Physics 1930
1929	Hans Fischer (1881–1945) synthesizes hemin.	Chemistry 1930
1939	Axel Hugo Theodor Theorell (1903–1982) purifies cytochrome *c*.	Medicine 1955
1941	Theorell crystallizes horseradish peroxidase.	
1944	Evgeny Zavoisky (1907–1976) discovers electron paramagnetic resonance (EPR).	
1945	Independent discovery of nuclear magnetic resonance (NMR) by Felix Bloch (1905–1983) and Edward Mills Purcell (1912–1997).	Physics 1952
1955	Osamu Hayaishi and Howard S. Mason independently discover biological oxygenation.	
1956	Rudolph A. Marcus (b. 1923) publishes the first paper in his series "On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer".	Chemistry 1992
1956	George Feher (b. 1924) develops the ENDOR (Electron Nuclear Double Resonance) method.	
1957	Dorothy Crowfoot Hodgkin (1910–1994) determines the X-ray structure of vitamin B_{12}.	Chemistry 1964
1957	Jens Christian Skou (b. 1918) discovers the ion-transporting enzyme Na^+, K^+-ATPase.	Chemistry 1997
1957	Rudolf Ludwig Mössbauer (b. 1929) discovers recoil-free γ-ray resonance absorption (Mössbauer effect).	Physics 1961
1960	John Cowdery Kendrew (1917–1997) reports the first three-dimensional atomic structure of a protein, myoglobin; Max Ferdinand Perutz (1914–2002) reports the X-ray structure of hemoglobin.	Chemistry 1962
1962	Leonard E. Mortenson discovers ferredoxin in *Clostridium pasteurianum*.	
1962	Ryo Sato (1923–1996) and Tsuneo Omura (b. 1930) identify P450 as a hemoprotein.	

Table 1 (*continued*)

Year		Nobel prize
1967	W.M. Fitch and E. Margoliash calculate the phylogenetic tree by comparing amino acid sequences of cytochromes *c*.	
1968	Kurt Wüthrich (b. 1938) publishes the high-resolution proton magnetic resonance spectra of myoglobin.	Chemistry 2002
1969	Joe McCord and Irvin Fridovich suggest that the function of "hemocuprein" is the catalysis of the reaction: $O_2^{\bullet-} + O_2^{\bullet-} + 2H^+ \rightarrow O_2 + H_2O_2$ and coin the term "superoxide dismutase".	
1972	Thomas Spiro obtains the resonance Raman spectra of myoglobin and cytochrome *c*.	
1976	Shuji Katsuki, William Arnold and Ferid Murad (b. 1936) discover the biological effect of nitric oxide (NO): activation of guanylate cyclase.	Medicine 1998 (Murad)
1977	Wüthrich and Richard R. Ernst (b. 1933) publish the first 2D ^{1}H NMR spectrum of a protein.	Chemistry 1991 (Ernst)
1978	Hans Freeman and colleagues determine the X-ray structure of poplar plastocyanin.	
1984	Johann Deisenhofer (b. 1943), Robert Huber (b. 1937) and Hartmut Michel (b. 1948) solve the first X-ray structure of a membrane protein, the photosynthetic reaction center.	Chemistry 1988
1987	Salvador Moncada (b. 1944) and Louis J. Ignarro (b. 1941) independently prove the suggestion of Robert F. Furchgott (b. 1916) that the "endothelium-derived relaxing factor" is identical to NO.	Medicine 1998 (Furchgott, Ignarro)
1993	Norbert Krauß and coauthors solve the X-ray structure of photosystem I at 6 Å.	
1995	Simultaneously, the X-ray structures of cytochrome *c* oxidase are published by So Iwata (enzyme from *Paracoccus denitrificans*) and Tomitake Tsukihara (bovine heart enzyme).	
2003	The structures of cytochrome b_6f complex are solved in the teams of Jean-Luc Popot (*Chlamydomonas reinhardtii*) and William A. Cramer (*Mastigocladus laminosus*).	

folding of many metalloproteins only occurs in the presence of specialized metal carrier proteins called metallochaperones. Therefore, the elucidation of the real chemical and biological environment of protein folding is required if we want to get closer to this Holy Grail of computational biology.

Throughout the text, cross-references to biological databases such as UniProt (Apweiler *et al*., 2004), EMBL (Kulikova *et al*., 2004), InterPro (*see* Article 83, **InterPro**, Volume 6), and PDB (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7) are used in the form (Database:Accession Number). The referenced database entries themselves contain comprehensive literature and database references. Additional Worldwide Web resources relevant to this chapter are compiled in Table 6.

2. Functional classification of metal centers in proteins

Metal centers in proteins can be classified according to their function:

- "permanent" centers
  - catalytic

  - electron transfer
  - exciton transfer
  - gas sensing
  - gas storage
- structural
  - "transient" centers
  - metal sensing
  - metal storage and transport.

It is a matter of taste whether proteins with "transient" metal centers qualify to be called metalloproteins. The distinction is not as rigid as it sounds: some metal centers can switch their roles. The iron-responsive element binding protein (IRE-BP) functions either as an active aconitase (aconitate hydratase; EC 4.2.1.3), when cells are iron-replete, or as an active mRNA-binding protein, when cells are iron-depleted (Rouault and Klausner, 1996). Many metalloproteins have more than one functional metal center.

There are four classes of functional groups collectively referred to as polypeptide-derived, or *endogenous*, ligands:

- side chain groups: amide (Asn, Gln), amino (Lys), carboxyl (Asp, Glu), guanidino (Arg), hydroxyl (Ser, Thr), imidazole (His), phenol (Tyr), selenol (Sec), sulfide (Met), and thiol (Cys)
- carbonyl and amide of the main chain
- amino group at the N-terminus
- carboxylate group at the C-terminus.

Some posttranslational modifications add to this list; for instance, the carboxyl group of N^6-carboxy-L-lysine provides a bridging ligand for the two Ni atoms in the urease active center.

Ligands not derived from a polypeptide are called *exogenous* (Holm *et al*., 1996). Exogenous ligands range from simple inorganic entities (e.g., oxide, hydroxide, sulfide, water and other solvent-derived molecules, or such physiological ligands as dioxygen or nitric oxide) to polydentate organic compounds, for example, porphyrins or corrins.

3. Metal centers in metalloproteins

3.1. Mononuclear centers

The simplest metal center consists of one metal atom and its first coordination shell ligands. The functional roles of mononuclear centers are diverse: catalytic (Fe in lipoxygenase, InterPro:IPR000907), electron-transfer (Cu in blue copper proteins, InterPro:IPR001235), metal transport (Ni in urease metallochaperone UreE, UniProt:P18317), metal sensing (copper efflux regulator CueR, PDB:1Q05), and structural (various Ca and Zn centers). This is reflected in the diversity of

their coordination geometries, with coordination numbers varying between 2 and 8. Vanadium haloperoxidases (InterPro:IPR008934) have an unusual mononuclear center, vanadate (VO_4^{3-}) (Figure 1a).

3.2. Dinuclear centers

Dinuclear (binuclear) centers contain two or more metal atoms within a single coordination shell. Both homodinuclear (dicobalt, dicopper, diiron, dimanganese, dimagnesium, dinickel, dizinc) and heterodinuclear (copper,zinc; iron,manganese; iron,nickel; iron,zinc) centers have been described. In dinuclear centers, two metal atoms can be linked by one or more bridging ligands. The dinuclear centers participate in catalysis (Cu–Zn in Cu,Zn superoxide dismutase; Ni–Ni in urease), electron-transfer (Fe_2S_2 in ferredoxins), oxygen storage (Cu–Cu in hemocyanins), and H_2 sensing (Ni–Fe in regulatory hydrogenase) (Haumann *et al*., 2003).

3.3. Polynuclear centers

Polynuclear centers contain more than two metal atoms within a single coordination shell. The functional roles of polynuclear centers include structural (iron–sulfur cluster loop domain, InterPro:IPR003651), electron transfer (Fe_3S_4 and Fe_4S_4 clusters in bacterial-type ferredoxins), catalysis (nitrogenase iron–molybdenum cofactor; Mn_4 cluster in the oxygen-evolving complex of the photosystem II) and oxygen sensing (Fe_4S_4 cluster in FNR regulator; UniProt:P03019).

3.4. Metal–organic molecule complex centers

Many metalloproteins employ organic molecules as polydentate exogenous ligands. The metals in these centers could have as little as one or even no endogenous ligands.

Naturally occurring tetrapyrroles form tight complexes with metal ions, such as Co (vitamin B_{12}), Fe (hemes), Mg (chlorophylls), and Ni (coenzyme F430). The coordination geometries of metals in tetrapyrrole complexes are square planar, square pyramidal, or octahedral.

Molybdenum and tungsten enzymes have a pterin cofactor that coordinates to the metal atom via its dithiolene group. The pterin cofactor common to Mo- and W-containing enzymes is frequently referred to by the trivial name *molybdopterin*, but this term refers to the metal-free organic cofactor alone. Fischer *et al*. (1998) suggested a systematic nomenclature for these cofactors and their metal-free pterin units. The cofactor is considered to be a derivative of ene-dithiol pyranopterin (H_2Dtpp). In eukaryotic enzymes, the terminal phosphate group is attached to the cofactor via the methylene group (R = mP), whereas in enzymes from prokaryota it can be linked to a nucleotide of adenine, cytosine, guanine, or inosine (R = mADP, mCDP, mGDP, or mIDP). Thus, the molybdenum cofactor of the aldehyde

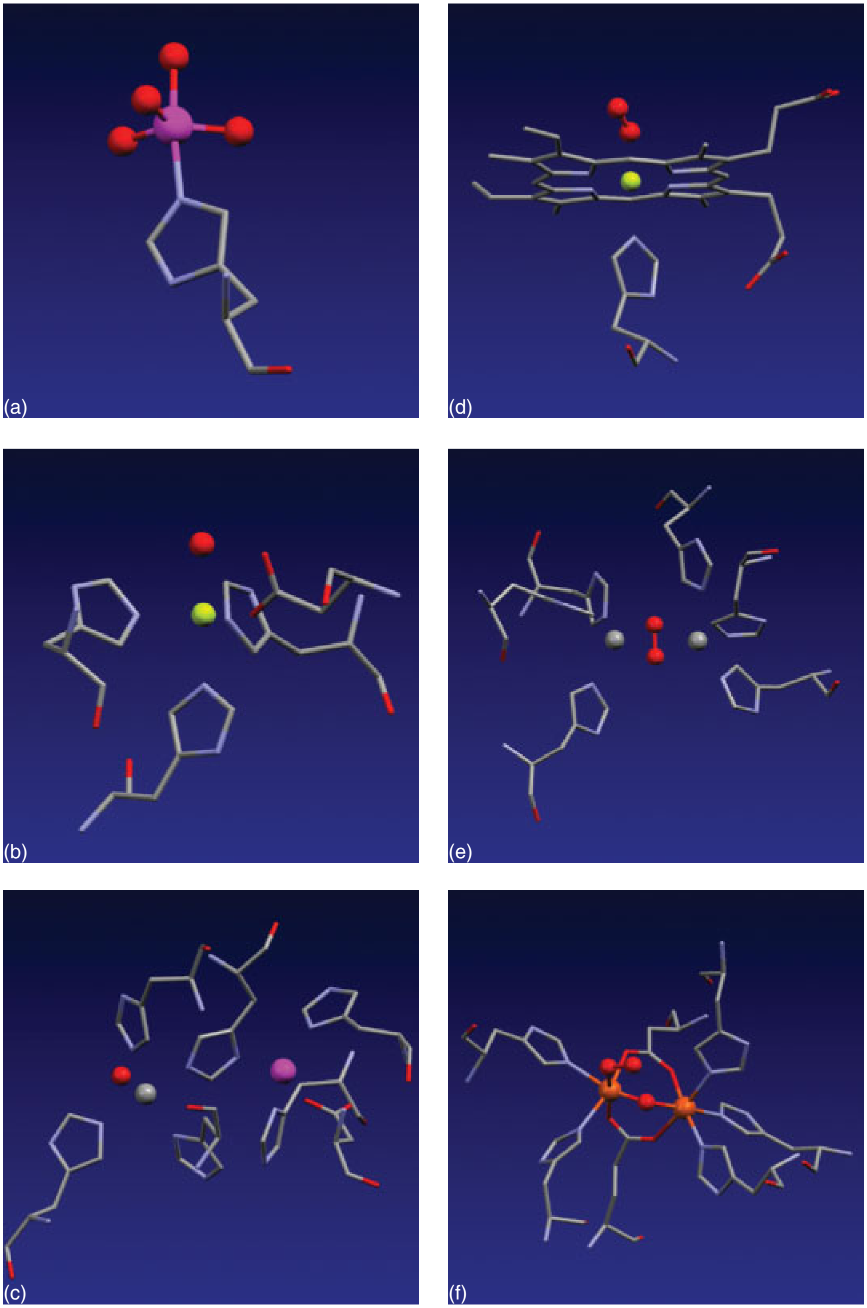


Figure 1 Examples of metal centers in proteins. (a) Mononuclear vanadate center in vanadium-containing chloroperoxidase from *Curvularia inaequalis* (PDB:1IDQ). (b) Mononuclear iron center in cambialistic superoxide dismutase from *Propionibacterium shermanii* (PDB:1AR5). (c) Dinuclear copper–zinc center in superoxide dismutase from *Xenopus laevis* (PDB:1XSO). (d) Heme *b* in sperm whale oxymyoglobin (PDB:1A6M). (e) Dinuclear iron center in oxyhemerythrin from *Themiste dyscrita* (PDB:1HMO). (f) Dinuclear copper center in oxyhemocyanin from *Octopus dofleini* (PDB:1JS8).

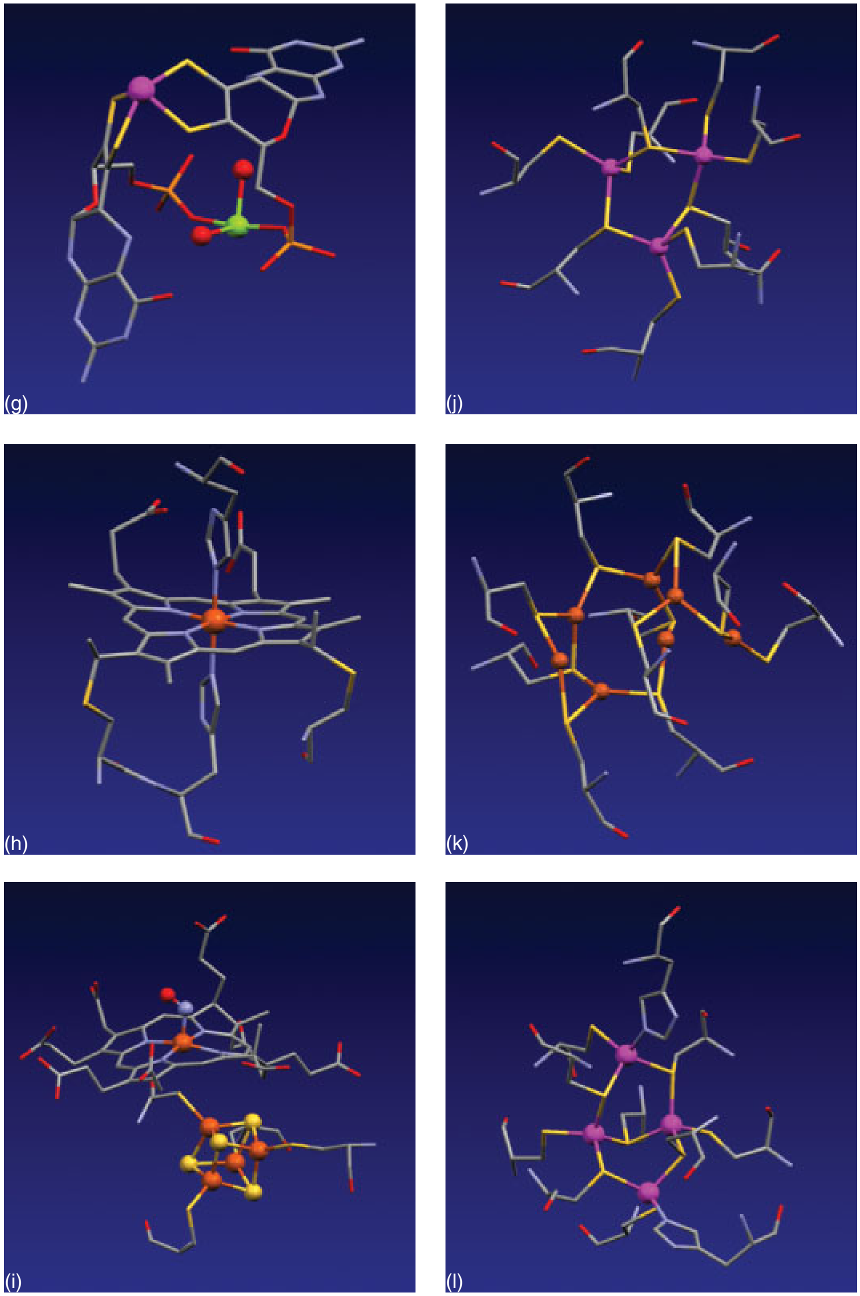


Figure 1 (*continued*) (g) Tungsten cofactor, W{Mg(H_2O)$_2$}(Dtpp)$_2$ of aldehyde ferredoxin oxidoreductase from *Pyrococcus furiosus* (PDB:1AOR). (h) Heme *c* 4 in cytochrome c_3 from *Desulfovibrio desulfuricans* (PDB:3CYR). (i) Siroheme–Fe_4S_4 center (NO complex) of sulfite reductase hemoprotein from *Escherichia coli* (PDB:6GEP). (j) Cd_3 cluster in metallothionein from *Notothenia coriiceps* (PDB:1M0J). (k) Cu_7 cluster in metallothionein from *Saccharomyces cerevisiae* (PDB:1AQR). (l) Zn_4 cluster in metallothionein from *Synechococcus sp*., PCC 7942 (PDB:1JJD)

oxidoreductase from *Desulfovibrio gigas* (PDB:1HLR) can de abbreviated as Mo(Dtpp-mCDP), and the tungsten cofactor of the aldehyde ferredoxin oxidoreductase from *Pyrococcus furiosus* (PDB:1AOR) as $W(Dtpp)_2$ (Figure 1g).

Methanol dehydrogenase (EC 1.1.99.8) and quinoprotein ethanol dehydrogenase (UniProt:Q9Z4J7) contain calcium complexes of pyrroloquinoline quinone (PQQ). Thiamine diphosphate (TPP), a cofactor of transketolase (EC 2.2.1.1), pyruvate dehydrogenase (EC 1.2.4.1), pyruvate:ferredoxin oxidoreductase (PDB:1B0P), and the pyruvate decarboxylase family (InterPro:IPR000399), binds Ca^{2+} or Mg^{2+} via its diphosphate group.

Metal–organic molecule complexes can form combined centers with mono-, di- or polynuclear inorganic centers. Examples include the siroheme–Fe_4S_4 center in sulfite reductase hemoprotein from *Escherichia coli* (Figure 1i), the Cu–S–Mo(Dtpp-mCDP) center in carbon monoxide dehydrogenase from *Oligotropha carboxidovorans* (PDB:1N5W), and the heme a_3–Cu_B center of ba_3-type cytochrome *c* oxidase from *Thermus thermophilus* (PDB:1EHK).

4. Gas storage and transport

After the advent of dioxygen (O_2) as a by-product of photosynthesis, living organisms had to solve a whole set of new problems: storage, transportation, sensing, detoxification, and utilization. Nature came up with at least three different solutions to the dioxygen storage/transport problem (Kurtz, 1999; Table 2). Note that the names of the oxygen transporters, "hemoglobin", "hemocyanin", and "hemerythrin", do not refer to the heme group (only found in globins), but are derived from the Greek word *αίμα* (blood).

4.1. Globins

Globins are heme-containing proteins involved in dioxygen binding and/or transport. One can say that the whole field of bioinorganic chemistry starts with globins. Hemoglobin was the first protein obtained in crystalline form by Hoppe-Seyler in 1864; a century later, the crystal structures of myoglobin (Figure 2a) and hemoglobin earned the Nobel Prize in Chemistry for Kendrew and Perutz. The globin superfamily includes vertebrate heterotetrameric ($\alpha_2\beta_2$) hemoglobins, vertebrate myoglobins, invertebrate globins (erythrocruorins and chlorocruorins), symbiotic plant globins (leghemoglobins), and bacterial and fungal flavohemoproteins (flavohemoglobins). Most globins contain a heme *b* (Fe–protoporphyrin-IX) group except for chlorocruorins, found in the blood of polychaete worms; their prosthetic group (chlorocruoroheme) is a derivative of the heme *b*, in which the vinyl group at position 2 is substituted by a formyl group.

The imidazole ring of the "proximal" histidine residue provides the fifth heme iron ligand. In most globins, the other axial heme iron position remains essentially free for O_2 coordination. O_2 binding (Figure 1d) results in a transition from

Table 2 Gas-binding proteins

Protein	Metal center	InterPro (UniProt)	PDB
Globins			
Hemoglobin	heme *b*	IPR002338, IPR002337	1A4F
Myoglobin	heme *b*	IPR002335	1MBO
Erythrocruorin	heme *b*	IPR002336	1ECD
Leghemoglobin	heme *b*	IPR001032	1GDI
Chlorocruorin	chlorocruoroheme	(Q26506)	
Indoleamine 2,3-dioxygenase-like myoglobin	heme *b*	(Q01966)	
Hemerythrin family			
Hemerythrin	Fe–Fe	IPR002063	1HMO
Myohemerythrin	Fe–Fe	IPR002063	1A7E
Hemocyanins			
Arthropod hemocyanin	Cu–Cu	IPR000896	1OXY
Molluscan hemocyanin	Cu–Cu		1JS8
Nitrophorins			
Rhodnius prolixus nitrophorin	heme *b*	IPR002351	1KOI
Cimex lectularius nitrophorin	heme *b*	(O76745)	
Gas sensors			
CO sensor CooA	heme *b*	(P72322)	1FT9
Soluble guanylate cyclase	heme *b*	IPR001054	
O_2 sensor FixL	heme *b*	IPR004358	1DP6
O_2 sensor FNR	Fe_4S_4	(P03019)	
Ethylene receptor ETR1	Cu	(P49333)	

high-spin to low-spin iron with accompanying changes in the Fe–N bond lengths and coordination geometry.

"Myoglobins" found in some gastropods do not belong to the globin superfamily but are homologous to the heme enzyme indoleamine 2,3-dioxygenase (EC 1.13.11.42) (Suzuki *et al*., 1998).

4.2. *Hemerythrin and myohemerythrin*

The *nonheme* iron proteins hemerythrin and myohemerythrin belong to a large group of diiron–carboxylate proteins. Hemerythrin has been isolated from the blood of marine invertebrates (annelids, brachiopods, and sipunculids). Hemerythrin typically exists as a homooctamer or heterooctamer composed of α- and β-type subunits of 13–14 kDa. Myohemerythrin is a monomeric O_2-binding protein found in the muscles of marine invertebrates. Hemerythrin and myohemerythrin possess one dinuclear iron center per subunit. The iron atoms are coordinated by five histidine residues and bridged by two carboxylate residues and a hydroxyl or oxo group (Figure 1e).

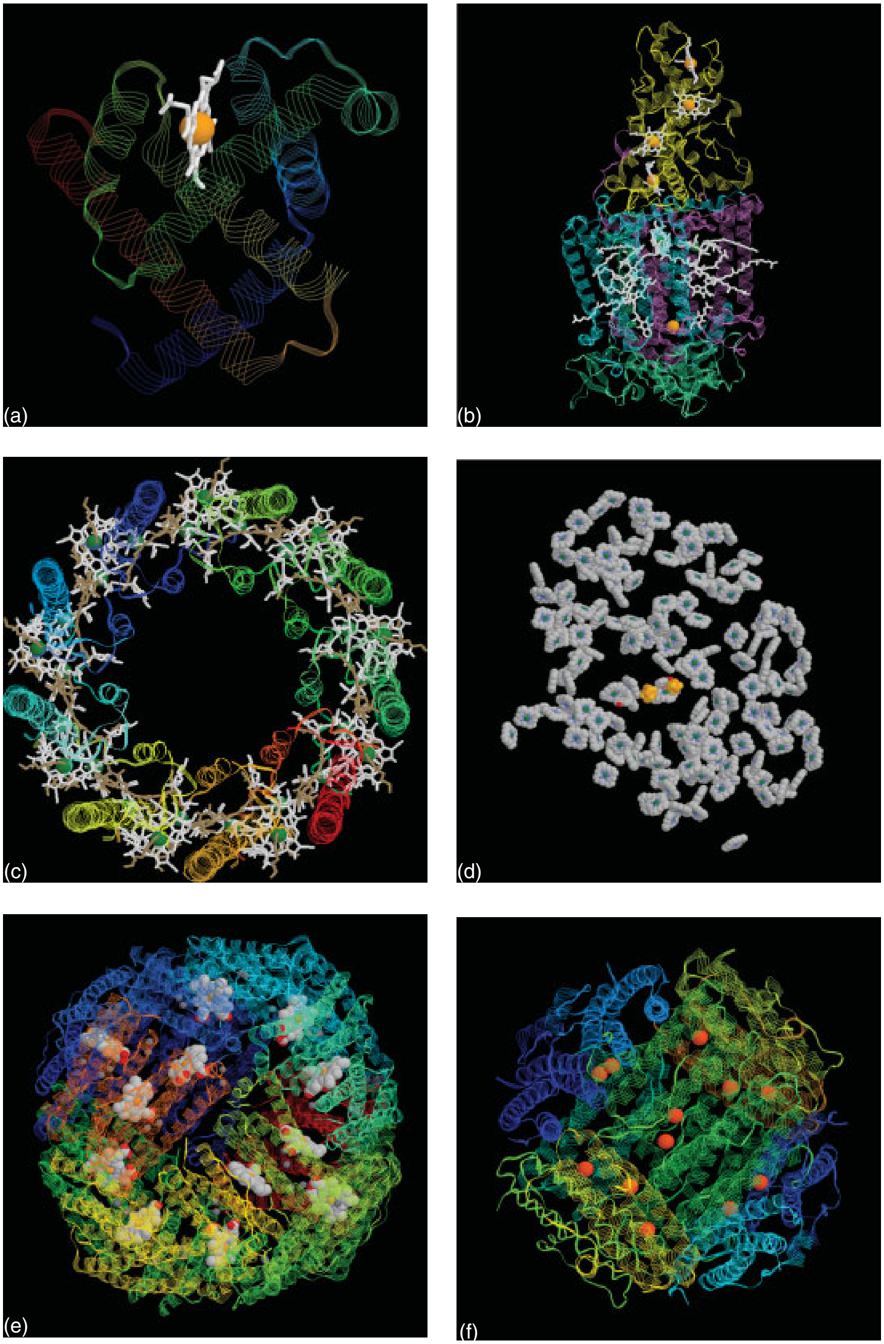


Figure 2 (a) Sperm whale myoglobin (PDB:1MBN). (b) Arrangement of cofactors in PRC from *Rhodopseudomonas viridis* (PDB:1PRC). (c) Light-harvesting complex 2 (B800-850) from *Rhodopseudomonas acidophila* (PDB:1KZU). (d) Arrangement of cofactors in PSI from *Pisum sativum* (PDB:1QZV). (e) Bacterioferritin from *Escherichia coli* (PDB:1BFR). (f) Dodecameric ferritin from *Listeria innocua* (PDB:1QGH)

4.3. Hemocyanins

Hemocyanins are the copper-containing oxygen transport proteins found in some arthropods and molluscs. Deoxyhemocyanin contains a dinuclear Cu^{+}–Cu^{+} center and is colorless. Upon oxygenation, the copper center is oxidized to Cu^{2+}–Cu^{2+} (dioxygen is bound as ${O_2}^{2-}$) and the protein becomes blue. At the level of primary, tertiary, and quaternary structure, arthropod and molluscan hemocyanins show significant differences. Arthropod hemocyanins exist as hexamers or oligo(hexamers) of similar polypeptides. Each ~75-kDa monomer consists of three structural domains; the central, mostly helical domain binds copper. Molluscan hemocyanins are hollow cylindrical decamers, dodecamers, or oligo(decamers) of large (~450 kDa) subunits. Each subunit consists of seven or eight globular "functional units" of ~400 residues containing one dinuclear copper site.

In spite of the great differences, the active sites of arthropod and molluscan hemocyanins are remarkably similar. This may be indicative of their distant relationship (van Holde *et al*., 2001). The oxygen-binding site contains two copper atoms, each coordinated by three histidine residues (Figure 1f). In molluscan hemocyanins, a thioether bridge is formed between one of the copper ligating histidines and an adjacent cysteine residue.

4.4. Nitrophorins

Dioxygen is not the only gas to have specific transporters. For instance, nitric oxide (NO) is transported by the nitrophorins, hemoproteins found in the saliva of blood-feeding insects. Two unrelated classes of nitrophorins are described. The saliva of the blood-sucking bug *Rhodnius prolixus* contains four homologous nitrophorins, designated NP1 to NP4 in order of their relative abundance in the glands. As isolated, nitrophorins contain NO ligated to the heme iron(III). Histamine, which is released by the host in response to tissue damage, is another nitrophorin ligand. Nitrophorins transport NO to the feeding site. Dilution, binding of histamine, and increase in pH (from pH ~5 in the salivary gland to pH ~7.4 in the host tissue) facilitate the release of NO into the tissue where it induces vasodilatation. Crystal structures of *R. prolixus* nitrophorins reveal a lipocalin-like fold.

The salivary nitrophorin from the hemipteran *Cimex lectularius* is a heme–thiolate protein that has no sequence similarity to *R. prolixus* nitrophorins. Apparently, the two classes of insect nitrophorins have arisen as a product of convergent evolution.

5. Gas sensors

A number of hemoproteins evolved to function as sensors for O_2, CO, and NO. Among them are the NO receptor soluble guanylate cyclase (EC 4.6.1.2), the O_2 sensor/histidine kinase FixL, and the CO-sensing transcriptional activator CooA (Rodgers, 1999). Nonheme gas sensors include fumarate and nitrate reduction (FNR) regulatory protein (Kiley and Beinert, 1998).

5.1. *CO sensor CooA*

The CO-sensing transcriptional activator CooA is found in the bacterium *Rhodospirillum rubrum* that can grow using CO as the sole energy source. CooA regulates the CO-dependent expression of the *coo* operon (EMBL:U65510) genes, including carbon monoxide dehydrogenase (EC 1.2.99.2) and hydrogenase (EC 1.12.7.2). CooA is a homodimer with each monomer consisting of a heme sensor and DNA-binding domains. Only CO-bound CooA binds to its target DNA. The unique features of CooA include ligation of the heme iron by the N-terminal proline residue and exchange of the axial ligands upon oxidation. The heme in CooA is hexacoordinate, with axial ligands Cys-75 and Pro-2 (Fe^{3+}), His-77 and Pro-2 (Fe^{2+}), and His-77 and CO (Fe^{2+}–CO) (Aono, 2003).

6. Metalloenzymes

The crystallization of urease (EC 3.5.1.5) by James Sumner in 1926 provided firm proof of the protein nature of enzymes and brought him the Nobel Prize in Chemistry 20 years later. However, it was not until 1975 that it was revealed that the active site of this enzyme contains two nickel atoms (Karplus *et al*., 1997).

Metalloproteins are present in all six enzyme classes (Table 3). Magnesium-dependent enzymes are probably the most abundant, which is not surprising given that Mg^{2+} is the most abundant divalent cellular cation (Maguire and Cowan, 2002). Mg^{2+} may be a part of the substrate (e.g., Mg^{2+}–isocitrate complex is *the* substrate of isocitrate lyase, EC 4.1.3.1) or serves as a cofactor, as in the dinuclear center of restriction endonuclease *Eco*RV (UniProt:P04390) (Cowan, 2002). Magnesium and zinc ions function as Lewis acids toward substrates. Arguably, oxidoreductases (EC 1) are the most interesting enzyme class for the bioinorganic chemist since they show the greatest diversity of metal centers. Transition metals such as V, Mn, Fe, Co, Ni, Cu, Mo, and W are used in the redox-active centers of oxidoreductases. However, they are also found in the active sites of nonredox enzymes, for example, Co or Mn centers in metalloproteinases (*see* Article 80, **Peptidases, families, and clans**, Volume 6).

Manganese and copper enzymes contain a variety of inorganic (mono-, di-, and polynuclear) centers. Iron, cobalt, and nickel enzymes contain both inorganic centers and tetrapyrrole complexes (hemes, cobamide, and coenzyme F430). Mn, Fe, Ni, Cu, and Zn are found in heterodinuclear and heteropolynuclear centers. With the notable exception of the molybdenum–iron–sulfur cofactor (FeMoco) of molybdenum nitrogenase (UniProt:P07328), molybdenum and tungsten enzymes bind metal–Dtpp or metal–$Dtpp_2$ complexes. Vanadium is present in the vanadium–iron–sulfur cluster of vanadium nitrogenase (UniProt:P16855) or as vanadate (VO_4^{3-}) in vanadium haloperoxidases (InterPro:IPR008934). Until recently, cadmium has been considered only as a toxic metal; however, it has been shown to be a physiological cofactor of the carbonic anhydrase (EC 4.2.1.1) from the marine diatom *Thalassiosira weissflogii* (Lane and Morel, 2000).

Table 3 Examples of structurally characterized metalloenzymes

EC	Enzyme name	Cofactor	PDB
1.1.1.1	Alcohol dehydrogenase	Zn	2OHX
1.1.99.8	Methanol dehydrogenase	Ca-PQQ	1G72
1.1.3.22	Xanthine oxidase	Mo(Dtpp-mP), FAD, Fe_2S_2,	1FIQ
1.2.1.2	Formate dehydrogenase	Mo(Dtpp-mGDP), Fe_4S_4, heme	1KQF
1.2.4.1	Pyruvate dehydrogenase	Mg–TPP	1L8A
1.2.4.4	2-oxoisovalerate dehydrogenase	Mg–TPP	1QS0
1.2.7.1	Pyruvate:ferredoxin oxidoreductase	Mg–TPP, Fe_4S_4	1B0P
1.2.7.5	Formaldehyde ferredoxin oxidoreductase $W(Dtpp)_2$	Mg–$W(Dtpp)_2$, Fe_4S_4	1B4N
1.2.99.2	Carbon monoxide dehydrogenase	Fe_4S_4, $NiFe_3S_4$	1JQK
1.2.99.2	Carbon monoxide dehydrogenase	Fe_4S_4, $NiFe_4S_5$	1JJY
1.2.99.2	Carbon monoxide dehydrogenase	FAD, Fe_2S_2, Cu-S-Mo(Dtpp-mCDP)	1N5W
1.3.99.1	Fumarate reductase	FAD, Fe_2S_2, Fe_3S_4, Fe_4S_4	1KF6
1.4.3.6	Amine oxidase (copper-containing)	Cu, TPQ	1OAC
1.4.3.13	Protein-lysine 6-oxidase	Cu, TPQ, Ca^{2+}	1N9E
1.4.99.4	Aralkylamine dehydrogenase	heme *c*, CTQ	1JJU
1.7.2.1	Nitrite reductase (NO-forming)	Cu	1OE1
1.7.2.1	Nitrite reductase (NO-forming)	hemes *c* and d_1	1QKS
1.7.3.4	Hydroxylamine oxidase	hemes *c* and P460	1FGJ
1.7.99.6	Nitrous-oxide reductase	Cu–Cu, Cu_4S	1QNI
1.8.1.2	Sulfite reductase (NADPH)	siroheme–Fe_4S_4, FAD, FMN	1AOP
1.8.3.1	Sulfite oxidase	Mo(Dtpp-mP), heme	1SOX
1.9.3.1	Cytochrome-*c* oxidase (aa_3-type)	Cu–Cu, heme *a*, heme *a*–Cu, Mg^{2+}	2OCC
1.9.3.1	Cytochrome-*c* oxidase (ba_3-type)	Cu–Cu, heme *b*, heme a_s–Cu	1EHK
1.10.2.2	Ubiquinol–cytochrome-*c* reductase (cytochrome bc_1 complex)	Fe_2S_2, hemes *b* and *c*	1EZV
1.10.99.1	Plastoquinol–plastocyanin reductase (cytochrome b_6f complex)	Fe_2S_2, hemes *b* and *c*, Chl-*a*	1Q90 1VF5
1.11.1.5	Cytochrome-*c* peroxidase	heme *b*	1EBE
1.11.1.6	Catalase	heme *b*	1GWH
1.11.1.6	Catalase	Mn–Mn	1JKU
1.11.1.7	Myeloperoxidase	heme *m*	1D2V
1.11.1.10	Chloride peroxidase	heme *b*	2CPO
1.11.1.10	Chloride peroxidase	VO_4^{3-}	1IDQ
1.11.1.11	L-Ascorbate peroxidase	heme *b*	1OAF
1.11.1.13	Manganese peroxidase	heme *b*	1MNP
1.11.1.14	Diarylpropane peroxidase	heme *b*	1B80
1.14.13.25	Methane monooxygenase hydroxylase	Fe–Fe	1MTY
1.14.13.39	Nitric oxide synthase	heme *b*, FAD, FMN, tetrahydrobiopterin, Zn^{2+}	1NSE
1.14.19.2	Acyl-[acyl-carrier-protein] desaturase	Fe–Fe	1OQ9
1.14.20.1	Deacetoxycephalosporin-C synthase	Fe^{2+}	1UOG
1.14.99.1	Prostaglandin-endoperoxide synthase	heme *b*	1CVU
1.15.1.1	Superoxide dismutase	Cu–Zn	1XSO
1.15.1.1	Superoxide dismutase	Fe	1ISA
1.15.1.1	Superoxide dismutase	Mn	1D5N
1.15.1.2	Superoxide reductase	Fe	1DQI
1.17.4.2	Ribonucleoside-triphosphate reductase (Class I)	Fe–Fe	1MXR
1.17.4.2	Ribonucleoside-triphosphate reductase (Class II)	B_{12}	1L1L
1.17.4.2	Ribonucleoside-triphosphate reductase (Class III)	Fe_4S_4	

(*continued overleaf*)

Table 3 (*continued*)

EC	Enzyme name	Cofactor	PDB
1.18.6.1	Nitrogenase (molybdenum)	FeMoco, Fe_8S_7, Fe_4S_4	1N2C
1.12.7.2	Hydrogenase (nickel–iron)	Ni–Fe, Fe_3S_4, Fe_4S_4	1H2A
1.12.7.2	Hydrogenase (iron)	Fe–Fe, Fe_2S_2, Fe_4S_4	1FEH
1.21.3.1	Isopenicillin-N synthase	Fe^{2+}	1HB2
2.1.1.13	Methionine synthase	B_{12}	1BMT
2.2.1.1	Transketolase	Ca–TPP	1TRK
2.4.2.14	Glutamine phosphoribosyldiphosphate amidotransferase	Fe_4S_4	1AO0
2.8.1.6	Biotin synthase	Fe_2S_2, Fe_4S_4	1R30
2.8.4.1	Methyl-coenzyme M reductase (coenzyme-B sulfoethylthiotransferase)	coenzyme F430	1HBN
3.1.3.2	Purple acid phosphatase (animal-type)	Fe–Fe	1UTE
3.1.3.2	Purple acid phosphatase (plant-type)	Fe–Zn	4KBP
3.1.8.1	Phosphotriesterase	Zn–Zn	1EZ2
3.1.21.4	Restriction endonuclease *Eco*RV (Type II site-specific deoxyribonuclease)	Mg^{2+}	1RVC
3.2.1.147	Thioglucosidase (myrosinase)	Zn^{2+}	1E6X
3.3.2.6	Leukotriene-A_4 hydrolase	Zn^{2+}	1HS6
3.4.11.1	Leucyl aminopeptidase	Zn–Zn	1LAP
3.4.11.9	X-Pro aminopeptidase	Mn–Mn	1AZ9
3.4.11.18	Methionyl aminopeptidase	Co–Co	1C24
3.4.24.27	Thermolysin	Zn^{2+}, Ca^{2+}	1GXW
3.5.1.5	Urease	Ni–Ni	1FWJ
3.5.1.88	Peptide deformylase	Fe^{2+}	1LM4
3.5.2.3	Dihydroorotase	Zn–Zn	1J79
3.5.4.1	Cytosine deaminase	Fe^{3+}	1K6W
4.1.2.5	Threonine aldolase	Ca^{2+}	1M6S
4.1.2.17	L-Fuculose-phosphate aldolase	Zn^{2+}	2FUA
4.1.2.20	2-Dehydro-3-deoxyglucarate aldolase	Mg^{2+}	1DXE
4.2.1.3	Aconitate hydratase	Fe_4S_4	7ACN
4.2.1.24	Porphobilinogen synthase	Mg^{2+}	1B4K
4.2.1.28	Propanediol dehydratase	B_{12}	1IWB
4.2.1.84	Nitrile hydratase	Fe^{3+}	2AHJ
4.2.1.84	Nitrile hydratase	Co^{3+}	1IRE
4.2.99.18	Endonuclease III	Fe_4S_4	2ABK
4.99.1.1	Ferrochelatase	Fe_2S_2	1HRK
5.1.3.4	L-Ribulose-phosphate 4-epimerase	Zn^{2+}	1JDI
5.1.99.1	Methylmalonyl-CoA epimerase	Co^{2+}	1JC5
5.3.1.5	Xylose isomerase	Co–Co	1A0C
5.3.1.5	Xylose isomerase	Mn–Mn	1A0D
5.3.3.2	Isopentenyl-diphosphate D-isomerase	Mn^{2+}	1I9A
5.4.2.2	Phosphoglucomutase	Mg^{2+}	3PMG
5.4.2.6	β-Phosphoglucomutase	Mg^{2+}	1O08
5.4.99.1	Methylaspartate mutase	pseudo-B_{12}	1I9C
5.4.99.2	Methylmalonyl-CoA mutase	B_{12}	4REQ
5.5.1.7	Chloromuconate cycloisomerase	Mn^{2+}	2CHR
5.5.1.8	Geranyl-diphosphate cyclase	Mg^{2+}	1N1B
6.1.1.10	Methionine–tRNA ligase	Zn^{2+}	1F4L
6.3.1.2	Glutamate–ammonia ligase	Mn–Mn	1F52
6.3.5.5	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)	Mn	1CE8
6.5.1.2	DNA ligase (NAD^+)	Zn^{2+}	1DGS

Some metal sites in enzymes are not involved in catalytic activity but play structural or regulatory roles; other metal sites facilitate electron transfer and are therefore analogous to those of electron-transfer proteins.

Here, I will only briefly dwell on two groups of metalloenzymes: the superoxide dismutase and P450 enzymes, which perfectly illustrate convergent and divergent evolution of biological catalysts.

6.1. *Superoxide dismutase*

The same biochemical problem could be presented with several metalloenzyme solutions. Superoxide dismutase (SOD; EC 1.15.1.1) catalyzes the redox disproportionation (dismutation) of the superoxide radical, $O_2^{\bullet-}$:

$$O_2^{\bullet-} + O_2^{\bullet-} + 2H^+ \rightarrow O_2 + H_2O_2 \quad (1)$$

Several groups of SODs are known, distinguished by the metal prosthetic group: Cu–Zn, Fe, Mn, and Ni.

Fe- and Mn-SODs constitute a structural family (InterPro:IPR001189). Fe- and Mn-SODs are unequally distributed throughout the kingdoms of living organisms and are located in different cellular compartments. Fe-SOD and Mn-SOD from some organisms (e.g., *E. coli*) exhibit almost absolute metal specificity, while other enzymes, such as "cambialistic" SOD from *Propionibacterium shermanii* (UniProt:P80293), are active with either metal. A variation on this theme is Fe,Zn-SOD from *Streptomyces coelicolor* (UniProt:O51917) that apparently contains an extra zinc-binding center. The active-site metal (Fe or Mn) is pentacoordinate, with the metal ligands (N^{ε} of three conserved histidine residues, O^{δ} of the conserved aspartic acid residue, and a water molecule) arranged in a distorted trigonal bipyramidal geometry (Figure 1b).

Cu,Zn-SOD (InterPro:IPR001424) is found in the cytosol of eukaryotic cells and in the periplasm of different gram-negative bacteria. The active site contains Cu, coordinated by four histidine ligands and a water molecule in a distorted square planar geometry, and Zn ligated by one aspartic acid and three histidine residues in a tetrahedral geometry. The two metal atoms are bridged by the imidazole ring of one histidine residue (Figure 1c).

Neelaredoxin from *D. gigas* (UniProt:O50258) is a mononuclear iron SOD, unrelated to the Fe/Mn-SOD family but homologous to the superoxide reductase from *P. furiosus* (UniProt:P82385). Ni-SOD are found in *Streptomyces seoulensis* (UniProt:P80734) and *S. coelicolor* (UniProt:P80735).

6.2. *P450 enzymes*

P450 enzymes (P450s), also known as cytochromes P450, constitute a superfamily of heme–thiolate proteins, widely distributed throughout all kingdoms of life. The animal genomes contain ~80–90 P450 genes and higher plant genomes contain up to 300 genes, while many bacterial species contain only one P450 gene. Of all known heme–thiolate enzymes (and heme enzymes in general), the P450s are by far the most functionally diverse. Usually, they act as terminal

oxidases in multicomponent electron-transfer chains, called P450-containing systems. Eukaryotic microsomal P450s and some bacterial P450s receive electrons from an FAD- and FMN-containing enzyme, NADPH:P450 reductase (CPR; EC 1.6.2.4). The microsomal CPR is a membrane-bound protein that interacts with different P450s. In *Bacillus megaterium* (UniProt:P14779) and *Bacillus subtilis* (UniProt:O08394; UniProt:O08336), CPR is a C-terminal domain of CYP102, a single polypeptide self-sufficient soluble P450 system.

Cytochrome b_5 (InterPro:IPR001199) can serve as an effector (activator or inhibitor) of some P450s. One (of several) hypotheses is that cytochrome b_5 is involved in the transfer of the second electron to P450, either from CPR or NADH:cytochrome b_5 reductase (EC 1.6.2.2) (Schenkman and Jansson, 2003).

Mitochondrial P450s, such as the steroid 11β-monooxygenase (EC 1.14.15.4) and the cholesterol monooxygenase (side-chain-cleaving) (EC 1.14.15.6), and bacterial P450s like the camphor 5-monooxygenase (EC 1.14.15.1) are reduced by small soluble Fe_2S_2 proteins of the adrenodoxin family (InterPro:IPR001055). In turn, adrenodoxin and its bacterial homologs are reduced by corresponding FAD-containing reductases.

A novel class of self-sufficient P450 monooxygenases has been recently described (De Mot and Parret, 2002), exemplified by P450 RhF from *Rhodococcus* sp. (UniProt:Q8KU27). This protein contains an Fe_2S_2 and FMN-binding domain homologous to phthalate oxygenase reductase (InterPro:IPR000951) fused to the N-terminal P450 domain.

The most common reaction catalyzed by P450 enzymes is a monooxygenase reaction, that is, insertion of one atom of oxygen into the substrate while the other oxygen atom is reduced to water:

$$RH + O_2 + 2H^+ + 2e^- \rightarrow ROH + H_2O \qquad (2)$$

Many P450 monooxygenases, including most drug-metabolizing enzymes, are historically classified as "unspecific monooxygenase" (EC 1.14.14.1), although the specificity varies immensely between different P450 forms.

Some P450s do not require any other protein components for their function. Nitric oxide reductase (CYP55, P450nor) is involved in denitrification in several fungal species. This enzyme does not possess monooxygenase activity but is able to reduce NO to form nitrous oxide (N_2O) directly using NAD(P)H as electron donor. Allene oxide synthase (EC 4.2.1.92), prostacyclin synthase (EC 5.3.99.4), and thromboxane synthase (EC 5.3.99.5) are examples of P450s that do not require either NAD(P)H or molecular oxygen for their catalytic activity. Substrates for all these enzymes are fatty acid derivatives containing partially reduced dioxygen (as either hydroperoxy or epidioxy groups).

7. Electron-transfer metalloproteins

Although many transition metals are found in metalloproteins, only iron and copper are used in the pure electron transferases. These include mononuclear copper

Table 4 Structurally characterized electron-transfer metalloproteins

Protein	Metal center	InterPro (UniProt)	PDB
Mononuclear copper proteins (Cupredoxins)		IPR008972	
Azurin	$Cu(N^{\delta}{}_{His})_2(O_{Gly})(S^{\delta}{}_{Met})(S^{\gamma}{}_{Cys})$	IPR003246	2AZA
Amicyanin	$Cu(N^{\delta}{}_{His})_2(S^{\delta}{}_{Met})(S^{\gamma}{}_{Cys})$	IPR002386	1AAC
Plastocyanin	$Cu(N^{\delta}{}_{His})_2(S^{\delta}{}_{Met})(S^{\gamma}{}_{Cys})$	IPR002387	1PLC
Stellacyanin	$Cu(N^{\delta}{}_{His})_2(O^{\varepsilon}{}_{Gln})(S^{\gamma}{}_{Cys})$	(P29602)	1JER
Mononuclear iron proteins			
Desulforedoxin	$Fe(S^{\gamma}{}_{Cys})_4$	(P00273)	1DXG
Rubredoxin	$Fe(S^{\gamma}{}_{Cys})_4$	IPR004039	1BRF
Iron–sulfur proteins			
Adrenodoxin	$(Fe_2S_2)(S^{\gamma}{}_{Cys})_4$	IPR001055	1AYF
Plant-type ferredoxin	$(Fe_2S_2)(S^{\gamma}{}_{Cys})_4$	IPR010241	1AWD
Thioredoxin-like ferredoxin	$(Fe_2S_2)(S^{\gamma}{}_{Cys})_4$	(O66511)	1F37
Rieske protein	$(Fe_2S_2)(N^{\delta}{}_{His})_2(S^{\gamma}{}_{Cys})_2$	IPR005806	1RFS
Fe_3 bacterial-type ferredoxin	$(Fe_3S_4)(S^{\gamma}{}_{Cys})_3$	IPR001080	1F2G
Fe_4 bacterial-type ferredoxin	$(Fe_4S_4)(S^{\gamma}{}_{Cys})_4$	IPR001450	1IQZ
High potential iron protein	$(Fe_4S_4)(S^{\gamma}{}_{Cys})_4$	IPR000170	1CKU
Fe_7 bacterial-type ferredoxin	$\{(Fe_3S_4)(S^{\gamma}{}_{Cys})_3\}, \{(Fe_4S_4)(S^{\gamma}{}_{Cys})_4\}$	IPR000813	1FDB
Fe_8 bacterial-type ferredoxin	$\{(Fe_4S_4)(S^{\gamma}{}_{Cys})_4\}_2$	(P00198)	2FDN
Cytochromes			
Cytochrome *b*	$Fe(por)(N^{\varepsilon}{}_{His})_2$	IPR005797	1BCC
Cytochrome b_6	$Fe(por)(N^{\varepsilon}{}_{His})_2$	IPR005797	1VF5
Cytochrome b_5	$Fe(por)(N^{\varepsilon}{}_{His})_2$	IPR001199	1CYO
Cytochrome b_{562}	$Fe(por)(N^{\varepsilon}{}_{His})(S^{\delta}{}_{Met})$	(P00192)	256B
Class I (mitochondrial-type) cytochromes *c*	$Fe(por)(N^{\varepsilon}{}_{His})(S^{\delta}{}_{Met})$	IPR003088	1HRC
Class IIb cytochrome *c*	$Fe(por)(N^{\varepsilon}{}_{His})(S^{\delta}{}_{Met})$	(P00150)	
Class III (bacterial-type) multiheme cytochromes *c*	$Fe(por)(N^{\varepsilon}{}_{His})_2$	IPR002322	1I77 1HH5 19HC
Class IV (photosynthetic reaction center) multiheme cytochrome *c*	$Fe(por)(N^{\varepsilon}{}_{His})(S^{\delta}{}_{Met})$, $Fe(por)(N^{\varepsilon}{}_{His})_2$	IPR003158	1DXR
Cytochrome c_{554}	$Fe(por)(N^{\delta}{}_{His})(N^{\varepsilon}{}_{His})$, $Fe(por)(N^{\varepsilon}{}_{His})$, $Fe(por)(N^{\varepsilon}{}_{His})_2$	(Q57142)	1FT5
Cytochrome c_1	$Fe(por)(N^{\varepsilon}{}_{His})(S^{\delta}{}_{Met})$	IPR002326	1BGY
Cytochrome *f*	$Fe(por)(N^{\alpha}{}_{Tyr})(N^{\varepsilon}{}_{His})$	IPR002325	1CI3

proteins (cupredoxins), the mononuclear iron proteins rubredoxin and desulforedoxin, iron–sulfur proteins (ferredoxins), and heme proteins (cytochromes) (Table 4).

7.1. Cytochromes

Cytochromes can be defined as electron- or proton-transfer proteins having one or several heme groups. Cytochromes possess a wide range of properties and function in a large number of different redox processes. Cytochromes can be classified according to their heme iron coordination, by heme type, and further by sequence

similarity. In contrast to other heme proteins, the heme iron in cytochromes is coordinated with two endogenous ligands in both reduced and oxidized states.

Cytochromes *b* can be defined as electron-transfer proteins having one or two heme *b* groups, noncovalently bound to the protein. Some hemoproteins, such as P450s and nitric oxide synthase (NOS), have been termed "cytochromes" as well (Kiel, 1995). Although both P450s and NOS can be considered as *b*-type cytochromes, their main function is catalysis.

Cytochromes b_5 are ubiquitous electron-transfer proteins found in animals, plants, fungi, and purple phototrophic bacteria. The microsomal and mitochondrial variants are membrane-bound, while bacterial ones and those from erythrocytes and other animal tissues are water-soluble. The eukaryotic microsomal cytochrome b_5 is an electron donor for stearoyl-CoA 9-desaturase (EC 1.14.19.1), linoleoyl-CoA desaturase (EC 1.14.19.13) and a number of P450 monooxygenases (Schenkman and Jansson, 2003).

Diheme cytochromes *b* include the cytochrome *b* subunit of the cytochrome bc_1 complex (ubiquinol–cytochrome-*c* reductase, EC 1.10.2.2) and the cytochrome b_6 subunit of the cytochrome b_6f complex (plastoquinol–plastocyanin reductase, EC 1.10.99.1).

Cytochromes *c* (InterPro:IPR000345) have one or more hemes bound to the protein by one or, more commonly, two thioether bonds involving sulfhydryl groups of cysteine residues. The fifth heme iron ligand is always provided by a histidine residue. Ambler (1991) recognized four classes of cytochromes *c*:

Class I includes the low-spin soluble cytochrome *c* of mitochondria and bacteria, with the heme-attachment site toward the N-terminus, and the sixth ligand provided by a methionine residue about 40 residues further on toward the C-terminus. Mitochondrial cytochrome *c* is among the most studied proteins. Since many eukaryotes have only one mitochondrial cytochrome *c*, the amino acid sequence comparison of Class I cytochromes *c* has been a natural choice for phylogenetic analysis since the 1960s. More recently, the key role of the cytochrome *c* in regulation of programmed cell death has been realized (Hancock *et al*., 2001).

Class II are four-α-helix bundle proteins that include the high-spin cytochrome c' (Class IIa; probably involved in gas sensing, not electron transfer) and a number of low-spin cytochromes (Class IIb), for example, cytochrome *c*-556 (Bertini *et al*., 2004). The heme-attachment site is close to the C-terminus.

Class III comprises the low redox potential multiple-heme cytochromes: triheme (cytochrome c_7), tetraheme (cytochrome c_3), nonaheme, and hexadecaheme (high-molecular-weight cytochrome, HMC), with only 30–40 residues per heme group. The hemes are all bis-His coordinated (Figure 1h), except for one high-spin heme in HMC, which is mono-His coordinated.

Class IV was originally created to hold the complex proteins that have other prosthetic groups as well as heme *c*, for example, flavocytochrome *c* and cytochrome cd_1. Alternatively, Moore and Pettigrew (1990) have suggested that Class IV includes tetraheme proteins containing both bis-His and His-Met coordinated hemes, with a 3D structure exemplified by that of the cytochrome *c* subunit of the photosynthetic reaction center (PRC), and which form a structurally homogeneous family.

Many other cytochromes *c* that do not belong to the classes I–IV are known. Among them are the cytochrome c_1 subunit of the cytochrome bc_1 complex, the cytochrome *f* subunit of the cytochrome b_6f complex, and the tetraheme cytochrome c_{554} from *Nitrosomonas europaea*. A number of oxidoreductases contain one or more hemes *c*: diheme cytochrome cd_1 nitrite reductase (EC 1.7.2.1), tetraheme flavocytochrome c_3 fumarate reductase (EC 1.3.99.1), pentaheme cytochrome *c* nitrite reductase (EC 1.7.2.2), and octaheme hydroxylamine oxidoreductase (EC 1.7.3.4).

7.2. *Rubredoxin and desulforedoxin*

In the $Fe(Cys)_4$ proteins, tetracoordinate iron is bound to four cysteine residues; these proteins, which are involved in electron transfer, contain the simplest form of an iron center. $Fe(Cys)_4$ proteins are often (incorrectly) included in the group of iron–sulfur proteins even though they do not bind inorganic sulfur.

Rubredoxin is a low-molecular-weight iron-containing bacterial protein involved in electron transfer, for example, as an electron donor in alkane 1-monooxygenase (EC 1.14.15.3) and superoxide reductase (EC 1.15.1.2) systems. In rubredoxin, iron is tetrahedrally coordinated. Desulforedoxin represents another class of $Fe(Cys)_4$ proteins with a distorted tetrahedral coordination of the iron. The distortion arises from adjacent cysteine residues in the sequence of desulforedoxin. In desulforedoxin mutants with one or two amino acid residues inserted between the two adjacent cysteines, the metal centers become spectroscopically and structurally similar to that of rubredoxin (Goodfellow *et al*., 2003).

7.3. *Ferredoxins*

Ferredoxins are iron–sulfur proteins that mediate electron transfer in a range of metabolic reactions; they fall into several subgroups according to the nature of their iron–sulfur cluster(s). One group, originally found in the bacterium *Clostridium pasteurianum*, has been termed “bacterial-type”, in which the active center is an Fe_4S_4 cluster. Bacterial-type ferredoxins may in turn be subdivided into further groups, on the basis of their sequence properties. Most contain at least one conserved domain, including four cysteine residues that bind an Fe_4S_4 cluster (InterPro:IPR001450). In some bacterial-type ferredoxins, one of the four conserved cysteine residues is lost. These ferredoxins bind an Fe_3S_4 cluster instead of an Fe_4S_4 cluster. In *P. furiosus*, Fe_4S_4 ferredoxin (UniProt:P29603), one of the conserved cysteine residues is substituted with an aspartic acid. During the evolution of bacterial-type ferredoxins, intrasequence gene duplication, transposition and fusion events occurred, resulting in the appearance of proteins with multiple iron–sulfur centers, such as dicluster $(Fe_4S_4)(Fe_4S_4)$ and $(Fe_3S_4)(Fe_4S_4)$ ferredoxins, polyferredoxins, and iron–sulfur subunits of various bacterial oxidoreductases.

High potential iron proteins (HiPIPs) form a unique family of Fe_4S_4 ferredoxins that function in anaerobic electron-transport chains. Some HiPIPs have a redox

potential higher than any other known iron–sulfur protein; for instance, HiPIP from *Rhodospirillum salinarum* has a redox potential of 500 mV (Heering *et al*., 1995). The iron–sulfur cluster in HiPIP accepts $(Fe_4S_4)^{2+}$ and $(Fe_4S_4)^{3+}$ oxidation states, while (Fe_4S_4) clusters of "bacterial-type" ferredoxins cycle between $(Fe_4S_4)^+$ and $(Fe_4S_4)^{2+}$ oxidation states.

The group of ferredoxins originally found in chloroplast membranes has been termed "chloroplast-type" or "plant-type". Another group, exemplified by the mitochondrial protein adrenodoxin, has been termed "mitochondrial-type" although homologous proteins were found in bacteria. In both "chloroplast-type" and "mitochondrial-type" ferredoxins, the active center is an Fe_2S_2 cluster, where the irons are tetrahedrally coordinated both by inorganic sulfurs and by sulfurs provided by four conserved cysteine residues.

In chloroplasts, Fe_2S_2 ferredoxins function as electron carriers in the photosynthetic electron-transport chain and as electron donors to various cellular proteins, such as glutamate synthase (EC 1.4.7.1), nitrate reductase (EC 1.7.7.2), and sulfite reductase (EC 1.8.7.1). In hydroxylating bacterial dioxygenase systems, they serve as intermediate electron-transfer carriers between reductase flavoproteins and an oxygenase (Mason and Cammack, 1992). In mitochondrial monooxygenase systems, adrenodoxin transfers an electron from NADPH:adrenodoxin reductase to a membrane-bound P450. In bacteria, putidaredoxin and terpredoxin serve as electron carriers between corresponding NADH-dependent ferredoxin reductases and soluble P450.

The Fe_2S_2 clusters of the Rieske proteins of electron-transfer chains (cytochrome bc_1 and cytochrome b_6f complexes) and of the ferredoxin component of aromatic-ring-hydroxylating dioxygenases such as benzene 1,2-dioxygenase (EC 1.14.12.3), toluene 1,2-dioxygenase (EC 1.14.12.11), naphthalene 1,2-dioxygenase (EC 1.14.12.12), and toluene 1,2-dioxygenase (EC 1.14.12.11) have similar chemical structures (Mason and Cammack, 1992). The Fe_2S_2 cluster is coordinated by two cysteine and two histidine residues.

7.4. *Cupredoxins*

In contrast to cytochromes and ferredoxins, all known copper electron-transfer proteins belong to one superfamily (InterPro:IPR008972). They contain a so-called "type I" mononuclear copper center with absorption at ~600 nm, which gives rise to the characteristic blue color (hence the names "azurin", "-cyanin" and the collective term for type I proteins, "blue copper proteins"). The common structural feature of type I Cu centers is two imidazole nitrogens and one thiolate sulfur (Gray *et al*., 2000). Cupredoxins are found in all kingdoms of life and participate in a variety of electron-transfer processes.

Plastocyanin mediates the electron transfer from the cytochrome b_6f complex to the photosystem I (PSI). It has been proposed that, with increasing bioavailability of copper, plastocyanin has evolved to replace its functional analog, Class I cytochrome c called cytochrome c_6. The cyanobacteria and algae are able to synthesize both plastocyanin and cytochrome c_6 depending on the bioavailabilities of copper and iron, respectively (Hervás *et al*., 2003).

8. Photosynthetic metalloproteins

Chlorophylls represent a group of tetrapyrrole pigments found in photosynthetic organisms that contain magnesium as the centrally chelated metal and contain a fifth five-membered ring E. Two functional classes of the protein–chlorophyll complexes may be defined: antenna systems (light-harvesting complexes) and photosynthetic reaction centers (PRCs) (Table 5). The antenna complexes broaden the spectral region of the light that can be used for photosynthesis and transfer the excitation energy from one to another and to the reaction center. With the exception of the Fenna–Matthews–Olson complex, all photosynthetic proteins reviewed here

Table 5 Structurally characterized chlorophyll proteins

Protein	Cofactors	Quaternary structure	PDB
Fenna–Matthews–Olson complex			
Bacteriochlorophyll *a* protein, *Prosthecochloris aestuarii*	BChl-*a*	α_3	4BCL
Bacteriochlorophyll *a* protein, *Chlorobium tepidum*	BChl-*a*	α_3	1KSA
Light-harvesting complexes of purple bacteria			
LH2, *Rhodospirillum molischianum*	BChl-*a*, lycopene	$(\alpha\beta)_8$	1LGH
LH2, *Rhodopseudomonas acidophila* strain 10050	BChl-*a*, rhodopin glucoside	$(\alpha\beta)_9$	1NKZ
LH3, *Rhodopseudomonas acidophila* strain 7050	BChl-*a*, rhodopinal glucoside	$(\alpha\beta)_9$	1IJD
Photosynthetic reaction centers of purple bacteria			
PRC, *Rhodobacter sphaeroides*	BChl-*a*, BPheo-*a*, Fe, ubiquinone-10, spheroidene	LMH	1PCR
PRC, *Thermochromatium tepidum*	BChl-*a*, BPheo-*a*, Fe, heme *c*, menaquinone-9, spirilloxanthin	CLMH	1EYS
PRC, *Rhodopseudomonas viridis*	BChl-*b*, BPheo-*b*, Fe, heme *c*, menaquinone-9, ubiquinone-9, 1,2-dihydroneurosporene	CLMH	1PRC
Photosystem I			
PSI, *Synechococcus elongatus*	Chl-*a*, Fe_4S_4, β-carotene, phylloquinone	12 subunits	1JB0
PSI, *Pisum sativum*	Chl-*a*, Fe_4S_4, phylloquinone	16 subunits	1QZV
Photosystem II and light-harvesting complex II			
PSII, *Thermosynechococcus vulcanus*	Chl-*a*, Pheo-*a*, heme *b*, heme *c*, β-carotene	14 subunits	1IZL
PSII, *Thermosynechococcus elongatus*	Chl-*a*, Pheo-*a*, heme *b*, heme *c*, β-carotene, plastoquinone	19 subunits	1S5L
LHCII, *Spinacia oleracea*	Chl-*a*, Chl-*b*, lutein, neoxanthin, violaxanthin	α_3	1RWT

are multisubunit transmembrane complexes (*see* Article 42, **Membrane-anchored protein complexes**, Volume 5).

8.1. Fenna–Matthews–Olson complex

Green sulfur bacteria (Chlorobiaceae) contain iron–sulfur type reaction centers and a large peripheral antenna complex called a *chlorosome*. The chlorosome collects light energy and transfers it to the membrane-bound reaction center through the amphypathic Fenna–Matthews–Olson (FMO) complex, also known as bacteriochlorophyll *a* (BChl-*a*) protein. The structure of the FMO complex from *Prosthecochloris aestuarii* was the first antenna structure to be solved. The complex consists of three identical subunits binding seven BChl-*a* molecules each. The magnesium of BChl-*a* is pentacoordinate; BChls 1, 3, 4, 6, and 7 are ligated by histidine residues, BChl 5 is coordinated to the backbone oxygen of leucine, while a bound water molecule serves as a ligand for BChl 2.

8.2. Light-harvesting complexes of purple bacteria

The photosynthetic antenna systems collect and deliver excited-state energy through excitation transfer to the reaction center. Photosynthetic purple bacteria have up to three types of antenna, or light-harvesting (LH), complexes besides the reaction center. LH1 complexes are closely associated with the reaction center and form a "core" complex, while LH2 and LH3 complexes are peripheral to the "core". The peripheral complexes have absorption maxima at shorter wavelengths than LH1 and the reaction center.

The LHs from purple bacteria are oligomers with the $\alpha\beta$ heterodimer as a basic unit. In LH2 and LH3, the $\alpha\beta$ heterodimers bind three bacteriochlorophyll *a* (BChl-*a*) molecules and at least one molecule of carotenoid. The role of the carotenoids in LH2 is twofold: they absorb light in the visible region of the spectrum and transfer it to BChl-*a* (accessory light-harvesting) and protect the LH2 from damage by singlet oxygen species (Isaacs *et al*., 1995). The $\alpha\beta$ heterodimers are arranged in a ring-shaped aggregate with strict rotational symmetry. The α- and β-subunits each possess single transmembrane helices, which are arranged in two concentric rings where the α-subunits form the inner circle and the β-subunits form the outer one. However, *Rs. molischianum* LH2 exists as an $(\alpha\beta)_8$ complex with eightfold symmetry, whereas *Rps. acidophila* LH2 and LH3 are $(\alpha\beta)_9$ complexes with ninefold symmetry (Figure 2c).

8.3. Photosynthetic reaction centers of purple bacteria

Photosynthetic reaction centers (PRCs) are membrane-spanning complexes of polypeptide chains and cofactors that catalyze the first steps in the conversion of light energy to chemical energy during photosynthesis. The PRC from *Rhodopseudomonas viridis* was the first transmembrane protein with a crystallographically

determined structure that brought the Nobel Prize in Chemistry to Deisenhofer, Huber, and Michel in 1988 (Figure 2b). Two major taxonomic groups of photosynthetic bacteria, purple sulfur bacteria (Chromatiaceae) and purple nonsulfur bacteria (Rhodospirillaceae), contain PRCs of similar structure. PRCs of photosynthetic purple bacteria consist of at least three protein subunits termed L (light), M (medium), and H (heavy); in *R. viridis* and *Thermochromatium tepidum*, the tetraheme (Class IV) cytochrome *c* is the fourth and largest protein subunit. The subunits L and M bind bacteriochlorophylls (BChl), bacteriopheophytins (BPheo), quinones, a ferrous iron ion, and a carotenoid as prosthetic groups. The PRCs from different organisms differ in the types of cofactors used (see Table 5). For instance, PRCs of *Rhodobacter sphaeroides* and *T. tepidum* contain BChl-*a* and BPheo-*a*, while the PRC of *R. viridis* contains BChl-*b* and BPheo-*b*.

8.4. Photosystem II

Photosystem II (PSII) is a multisubunit protein complex spanning the thylakoid membrane of plants, algae, and cyanobacteria that uses light energy to oxidize water to dioxygen:

$$2H_2O \rightarrow O_2 + 4H^+ + 4e^- \quad (3)$$

The light energy is absorbed by chlorophylls and carotenoid pigments of the antenna systems associated with PSII, which vary in different organisms (Barber, 2002). In higher plants, PSII exists as a complex of the "core" with the peripheral light-harvesting complex II (LHCII). PSII assembles as a dimer. Within the PSII core complex, Chl-*a* and β-carotene are bound mainly to the two largest PSII subunits, CP47 (PsbB) and CP43 (PsbC). Chl-*a* and β-carotene absorb the photons and transfer the excitation energy to the reaction center consisting of subunits D1 and D2. The subunits D1 and D2 bind Chl-*a*, Pheo-*a*, quinones, and an iron atom, and are structurally similar to the subunits L and M of the PRC of purple bacteria. The primary electron donor is a "special pair" of Chl-*a* molecules called P680. The O_2 is evolved after removal of four electrons from a catalytic center, comprising a spin-coupled tetramanganese cluster forming a part of the water-oxidizing complex. Apart from the Mn_4 cluster, the water-oxidizing complex contains one Ca^{2+}, one or two Cl^-, and several bridging oxygen ligands (Carrell *et al*., 2002). The Mn_4 cluster is bound at the luminal side of D1.

The crystal structures of cyanobacterial PSII have been solved. It contains the Mn_4 cluster, 36 Chl-*a*, two Pheo-*a*, two hemes, one nonheme Fe, and several β-carotene molecules per monomeric unit (see Table 5). In higher plants and green algae, 25 genes for the PSII core proteins have been identified.

8.5. Photosystem I

Like PSII, PSI is a multisubunit transmembrane protein of the thylakoid membrane. The PSII oxidizes water to O_2, four protons, and four electrons. The electrons are

Table 6 Metalloproteins on the Web

Protein families	
Cytochrome *c* oxidase	Cytochrome Oxidase Home Page: **http://www-bioc.rice.edu/~graham/CcO.html**
EF-hand domain	The Calcium-Binding Proteins Data Library: **http://structbio.vanderbilt.edu/cabp_database/**
Ferritin	Ferritin Molecular-Graphics Tutorial, **http://www.chemistry.wustl.edu/~edudev/LabTutorials/Ferritin/FerritinTutorial.html**
Globins	Globin Gene Server: **http://globin.cse.psu.edu/**
Lipoxygenases	LOX-DB: **http://www.dkfz-heidelberg.de/spec/lox-db/**
Metallothioneins	Metallothionein: **http://www.unizh.ch/~mtpage/MT.html**
P450, NOS, cytochrome b_5, ferredoxins	Directory of P450-containing Systems: **http://www.icgeb.org/p450/**
Plant peroxidases	International Working Group on Plant Peroxidases: **http://www.unige.ch/LABPV/perox.html**
Protein complexes	
Photosystem I	Photosystem I: **http://www.chemie.fu-berlin.de/fb_chemie/ikr/ag/saenger/projects/phosys/**
Photosystem II	Photosystem II: **http://www.bio.ic.ac.uk/research/barber/psIIimages/PSII.html**
Cytochrome bc_1 complex	The bc_1 complex site: **http://www.life.uiuc.edu/crofts/bc-complex_site/**
Light-harvesting complexes 1 and 2	Quantum Biology of the PSU: **http://www.ks.uiuc.edu/Research/psu/psu.html**
General	
Metal sites extracted from PDB	Metalloprotein site Database and Browser: **http://metallo.scripps.edu/**
Ontology of bioinorganic proteins	COMe, the bioinorganic motif database: **http://www.ebi.ac.uk/come/**
Glossary of terms used in bioinorganic chemistry	IUPAC Recommendations 1997: **http://www.chem.qmul.ac.uk/iupac/bioinorg/**
Nomenclature of electron-transfer proteins	NC-IUB Recommendations 1989: **http://www.chem.qmul.ac.uk/iubmb/etp/**

transferred from PSII to PSI through a pool of plastoquinones, the cytochrome b_6f complex, and a soluble metalloprotein, either the blue copper protein plastocyanin or the Class I cytochrome *c* called cytochrome c_6 (Hervás *et al*., 2003). PSI catalyzes the charge separation across the photosynthetic membrane by transferring the electrons from plastocyanin/cytochrome c_6 to Fe_2S_2 ferredoxin or flavodoxin; finally, the electrons are transferred to the flavoenzyme ferredoxin–$NADP^+$ reductase (EC 1.18.1.2).

PSI combines a PRC with an antenna system. The two largest subunits of PSI, PsaA and PsaB, bind most of the chlorophyll molecules of the antenna system. PSI of higher plants has an additional peripheral light-harvesting complex (LHCI). PSI is the largest membrane protein for which the 3D structure has been solved (*see* Article 99, **Large complexes by X-ray methods**, Volume 6); it also has the largest number of cofactors found in any known protein complex (Figure 2d). The excitation energy is transferred through Chl and the carotenoid molecules to the primary electron donor in the center of the complex. Similar to PSII, the primary electron donor is a Chl-*a* dimer called P700. From P700, a transmembrane electron transfer occurs via a chain of cofactors with descending midpoint potentials (Fromme *et al*., 2003).

The PSI from the cyanobacterium *Synechococcus elongatus* (PDB:1JB0) exists as a trimer with each monomer consisting of 12 protein subunits binding 96 Chl-*a*, 22 β-carotene molecules, three Fe_4S_4 clusters and two phylloquinones. The monomeric PSI from *Pisum sativum* (PDB:1QZV) consists of 12 core subunits and four peripheral subunits (LHCI). It contains 167 Chl-*a* molecules, three Fe_4S_4 clusters and two phylloquinones.

9. Metal storage and transport

9.1. Ferritins

Ferritins are spherical shell-like proteins that play a key role in iron metabolism. Their function is to remove Fe^{2+} ions from solution in the presence of O_2 and to deposit iron into the protein interior in a mineral form. Most known ferritins (InterPro:IPR009040) belong to a large group of diiron–carboxylate proteins. Some bacterial ferritins bind heme; in this case, they are called bacterioferritins (InterPro:IPR002024). The bacterioferritin from *E. coli* (UniProt:P11056) binds heme *b* while the *Desulfovibrio desulfuricans* protein (UniProt:Q93PP9) binds Fe–coproporphyrin III. The tertiary and quaternary structure of ferritins is highly conserved: 24 four-α-helix bundle subunits are arranged as a hollow, symmetrical shell with outside and inside diameters about 125 and 80 Å, respectively (Figure 2e). Up to 4500 iron atoms can be stored in its central cavity as a hydrated iron(III) oxide mineral: mainly ferrihydrite in animal ferritins and hydrated iron(III) phosphate in bacterial ferritins. Subunits pack tightly together except that at the threefold axes there are narrow channels traversing the shell. The threefold channels are the sites of iron transfer into the cavity of the ferritins.

Steps in iron storage within ferritin molecules consist of (1) Fe^{2+} oxidation, (2) Fe^{3+} migration and the nucleation, and (3) growth of the iron core mineral (Harrison and Arosio, 1996). The first step can be represented as

$$Fe^{2+} + O_2 + H_2O \rightarrow Fe^{3+} + H_2O_2 \quad (4)$$

and ferritin can be considered an enzyme catalyzing this reaction.

Another class of ferritins is represented by the dodecameric ferritin from *Listeria innocua* (PDB:1QGH; Figure 2f). This protein contains about 100 iron atoms per dodecamer; it is extremely stable in comparison to the classic 24-meric ferritins and preserves its quaternary structure at a pH as low as 2.0 (Chiaraluce *et al*., 2000).

9.2. Metallochaperones

Metallochaperones are cytosolic metal-binding proteins responsible for the intracellular trafficking of metal ions.

Three types of copper chaperones have been found (Markossian and Kurganov, 2003):

1. Atx1-like copper chaperones bind Cu^+ and interact directly with the Cu-exporting ATPase (EC 3.6.3.4); this family includes the yeast protein Atx1 (UniProt:P38636), its human homolog Hah1 (UniProt:O00244), and CopZ proteins from *Enterococcus hirae* (PDB:1CPZ) and *B. subtilis* (PDB:1K0V).
2. The copper chaperone for SOD (CCS; UniProt:O14618) delivers Cu^+ to Cu,Zn-SOD. This protein consists of three structural domains; the N-terminal domain I is homologous to Atx1, the central domain II is homologous to Cu,Zn-SOD, and the short C-terminal domain III is unique for CCS, which is crucial for copper insertion into apo-SOD.
3. Cox17 (InterPro:IPR007745), Cox11 (InterPro:IPR007533), and Sco1/Sco2 (InterPro:IPR003782) proteins are the copper chaperones for the cytochrome *c* oxidase (COX; EC 1.9.3.1). Cox17 is a soluble protein while Cox11 and Sco1 are located in the mitochondrial inner membrane. Cox17 and Sco1 are thought to provide copper exclusively to the dinuclear Cu_A site of the COX subunit 2 and Cox11 is implicated in the assembly of the mononuclear Cu_B site of the COX subunit 1.

Metallochaperones specific for other metals/metalloprotein targets are known. They include the nickel-binding urease accessory protein UreE (InterPro: IPR007864); the zinc transport protein ZntA (UniProt:P37617); the iron–sulfur cluster assembly proteins NifU and IscU (InterPro:IPR002871); the NafY protein facilitating the transfer of FeMoco to the aponitrogenase (PDB:1P90); and the heme chaperone CcmE required for cytochrome *c* biogenesis (UniProt:P33928).

9.3. Metallothioneins

Metallothioneins (MTs) are a ubiquitous class of diamagnetic low-molecular-weight proteins containing metal–thiolate clusters (Romero-Isart and Vašák, 2002). Various functions such as metal transport, storage, and detoxification have been proposed for MTs; however, the precise role of MTs remains to be established. MTs vary in length, tissue distribution, and metal specificity. Fungal MTs bind Cu^+ while MTs purified from mammals contain Cu, Zn, and Cd.

The 3D structure of the apoprotein (thionein) is mostly disordered. However, animal holo-MTs show a characteristic fold consisting of two globular domains, each binding a metal cluster.

A typical vertebrate MT such as the *Notothenia coriiceps* MT has three metal atoms bound by nine cysteine thiolates in the N-terminal β-domain (PDB:1M0J; Figure 1j), and four metal atoms bound by 11 cysteines in the C-terminal α-domain (PDB:1M0G). This pattern is reversed in an MT from the sea urchin *Strongylocentrotus purpuratus*, which has $Cd_4(S^{\gamma}{}_{Cys})_{11}$ in the N-terminal α-domain (PDB:1QJK) and $Cd_3(S^{\gamma}{}_{Cys})_9$ cluster in the C-terminal β-domain (PDB:1QJK). Crustacean MTs contain two domains, each binding a $Cd_3(S^{\gamma}{}_{Cys})_9$ cluster (UniProt:P29499, UniProt:P55949).

Although only a few residues shorter than the crustacean protein, the yeast MT (PDB:1AQR) binds seven Cu^+ ions in a *single* $Cu_7(S^{\gamma}{}_{Cys})_{10}$ cluster (Figure 1k).

Finally, the protein SmtA from *Synechococcus* sp. (PDB:1JJD) contains the single $Zn_4(S^{\gamma}{}_{Cys})_9(N^{\varepsilon}{}_{His})_2$ cluster (Figure 1l).

10. Conclusion

Metalloproteins have existed as long as protein-based life. Not only are metalloproteins involved in all physiological processes, they affect the whole biosphere. Their spectacular diversity has been demonstrated by structural studies; they nevertheless deserve much more attention from the bioinformatics and proteomics communities.

Related articles

Article 35, **Structural biology of protein complexes**, Volume 5; Article 42, **Membrane-anchored protein complexes**, Volume 5; Article 80, **Peptidases, families, and clans**, Volume 6; Article 83, **InterPro**, Volume 6; Article 95, **History and future of X-ray structure determination**, Volume 6; Article 99, **Large complexes by X-ray methods**, Volume 6; Article 104, **X-ray crystallography**, Volume 6; Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7

Acknowledgments

I thank Rolf Apweiler and Marcus Ennis for their helpful comments and suggestions on the manuscript.

Further reading

Ballou DP (Ed.), (1999) *Essays in Biochemistry: Metalloproteins*. Princeton University Press: Princeton.

Bugg T (1997) *An Introduction to Enzyme and Coenzyme Chemistry*. Blackwell Science: Oxford.

Cowan JA (1997) *Inorganic Biochemistry: An Introduction*, *Second Edition*, Wiley-VCH: New York.

Cowan JA (Ed.), (1995) *The Biological Chemistry of Magnesium*. VCH: New York.

Kaim W and Schwederski B (1994) *Bioinorganic Chemistry: Inorganic Elements in the Chemistry of Life*. John Wiley & Sons Ltd: Chichester.

Lippard SJ and Berg JM (1994) *Principles of Bioinorganic Chemistry*. University Science Books: Mill Valley.

Messerschmidt A, Huber R, Poulos T and Wieghardt K (Eds.), (2001) *Handbook of Metalloproteins*. John Wiley & Sons Ltd: Chichester.

References

Ambler RP (1991) Sequence variability in bacterial cytochromes *c*. *Biochimica et Biophysica Acta*, **1058**, 42–47.

Aono S (2003) Biochemical and biophysical properties of the CO-sensing transcriptional activator CooA. *Accounts of Chemical Research*, **36**, 825–831.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119. **http://www.uniprot.org/**

Barber J (2002) Photosystem II: a multisubunit membrane protein that oxidises water. *Current Opinion in Structural Biology*, **12**, 523–530.

Bertini I, Faraone-Mennella J, Gray HB, Luchinat C, Parigi G and Winkler JR (2004) NMR-validated structural model for oxidized *Rhodopseudomonas palustris* cytochrome c_{556}. *Journal of Biological Inorganic Chemistry*, **9**, 224–230.

Carrell G, Tyryshkin M and Dismukes C (2002) An evaluation of structural models for the photosynthetic water-oxidizing complex derived from spectroscopic and X-ray diffraction signatures. *Journal of Biological Inorganic Chemistry*, **7**, 2–22.

Chiaraluce R, Consalvi V, Cavallo S, Ilari A, Stefanini S and Chiancone E (2000) The unusual dodecameric ferritin from *Listeria innocua* dissociates below pH 2.0. *European Journal of Biochemistry*, **267**, 5733–5741.

Cowan JA (2002) Structural and catalytic chemistry of magnesium-dependent enzymes. *BioMetals*, **15**, 225–235.

De Mot R and Parret AHA (2002) A novel class of self-sufficient cytochrome P450 monooxygenases in prokaryotes. *Trends in Microbiology*, **10**, 502–508.

Fischer B, Enemark JH and Basu P (1998) A chemical approach to systematically designate the pyranopterin centers of molybdenum and tungsten enzymes and synthetic models. *Journal of Inorganic Biochemistry*, **72**, 13–21.

Fromme P, Melkozernov A, Jordan P and Krauss N (2003) Structure and function of photosystem I: interaction with its soluble electron carriers and external antenna systems. *FEBS Letters*, **555**, 40–44.

Goodfellow BJ, Rusnak F, Moura I, Ascenso CS and Moura JJG (2003) NMR solution structures of two mutants of desulforedoxin. *Journal of Inorganic Biochemistry*, **93**, 100–108.

Gray HB, Malmström BG and Williams RJP (2000) Copper coordination in blue proteins. *Journal of Biological Inorganic Chemistry*, **5**, 551–559.

Hancock JT, Desikan R and Neill SJ (2001) Does the redox status of cytochrome *c* act as a fail-safe mechanism in the regulation of programmed cell death? *Free Radical Biology and Medicine*, **31**, 697–703.

Harrison PM and Arosio P (1996) The ferritins: molecular properties, iron storage function and cellular regulation. *Biochimica et Biophysica Acta*, **1275**, 161–203.

Haumann M, Porthun A, Buhrke T, Liebisch P, Meyer-Klaucke W, Friedrich B and Dau H (2003) Hydrogen-induced structural changes at the nickel site of the regulatory [NiFe] hydrogenase from Ralstonia eutropha detected by X-ray absorption spectroscopy. *Biochemistry*, **42**, 11004–11015 (published erratum in Biochemistry, 42, 13786).

Heering HA, Bulsink BM, Hagen WR and Meyer TE (1995) Influence of charge and polarity on the redox potentials of high-potential iron-sulfur proteins: evidence for the existence of two groups. *Biochemistry*, **34**, 14675–14686.

Hervás M, Navarro JA and De la Rosa MA (2003) Electron transfer between membrane complexes and soluble proteins in photosynthesis. *Accounts of Chemical Research*, **36**, 798–805.

Holm RH, Kennepohl P and Solomon EI (1996) Structural and functional aspects of metal sites in biology. *Chemical Reviews*, **96**, 2239–2314.

Isaacs NW, Cogdell RJ, Freer AA and Prince SM (1995) Light-harvesting mechanisms in purple photosynthetic bacteria. *Current Opinion in Structural Biology*, **5**, 794–797.

Karplus PA, Pearson MA and Hausinger RP (1997) 70 years of crystalline urease: what have we learned? *Accounts of Chemical Research*, **30**, 330–337.

Kiel JL (1995) *Type-b Cytochromes: Sensors and Switches*, CRC Press: Boca Raton.

Kiley PJ and Beinert H (1998) Oxygen sensing by the global regulator, FNR: the role of the iron – sulfur cluster. *FEMS Microbiology Reviews*, **22**, 341–352.

Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, *et al*. (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, **32**, D27–D30. **http://www.ebi.ac.uk/embl/**

Kurtz DM Jr (1999) Oxygen-carrying proteins: three solutions to a common problem. *Essays in Biochemistry*, **34**, 85–100.

Lane TW and Morel FM (2000) A biological function for cadmium in marine diatoms. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4627–4631.

L'vov NP, Nosikov AN and Antipov AN (2002) Tungsten-containing enzymes. *Biochemistry (Moscow)*, **67**, 196–200.

Maguire ME and Cowan JA (2002) Magnesium chemistry and biochemistry. *BioMetals*, **15**, 203–210.

Markossian KA and Kurganov BI (2003) Copper chaperones, intracellular copper trafficking proteins. Function, structure, and mechanism of action. *Biochemistry (Moscow)*, **68**, 827–837.

Mason JR and Cammack R (1992) The electron-transport proteins of hydroxylating bacterial dioxygenases. *Annual Review of Microbiology*, **46**, 277–305.

Moore GR and Pettigrew GW (1990) *Cytochromes c. Evolutionary, Structural, and Physicochemical Aspects*, Springer: Berlin.

Rodgers KR (1999) Heme-based sensors in biological systems. *Current Opinion in Chemical Biology*, **3**, 158–167.

Romero-Isart N and Vašák M (2002) Advances in the structure and chemistry of metallothioneins. *Journal of Inorganic Biochemistry*, **88**, 388–396.

Rouault TA and Klausner RD (1996) Iron – sulfur clusters as biosensors of oxidants and iron. *Trends in Biochemical Sciences*, **21**, 174–177.

Schenkman JB and Jansson I (2003) The many roles of cytochrome b_5. *Pharmacology & Therapeutics*, **97**, 139–152.

Suzuki T, Kawamichi H and Imai K (1998) A myoglobin evolved from indoleamine 2,3-dioxygenase, a tryptophan-degrading enzyme. *Comparative Biochemistry and Physiology Part B*, **121**, 117–128.

van Holde KE, Miller KI and Decker H (2001) Hemocyanins and invertebrate evolution. *Journal of Biological Chemistry*, **276**, 15563–15566.

Specialist Review

Peptidases, families, and clans

Neil D. Rawlings and Alan J. Barrett
Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

Peptidases are the enzymes that hydrolyze peptide bonds. The terms "proteolytic enzyme", "protease" and "proteinase" are roughly synonymous with peptidase, but they are redundant, and less satisfactory in several respects, so we deprecate their use. About 2% of genes encode peptidases in all kinds of organisms, and this equates to over 500 active human peptidases. These enzymes catalyze innumerable important biological processes. Posttranslational proteolysis is crucial in the biosynthesis of many proteins, whether it is removal of the methionine that initiates translation of mRNA, of an N-terminal signal or targeting peptide, or of an activation peptide that unmasks the latent activity of an enzyme, protein hormone, or receptor. Physiological processes that depend upon the activities of peptidases include blood coagulation, complement activation, apoptosis, intracellular protein turnover, activation and catabolism of bioactive peptides, tissue remodeling, and the digestion of food. But proteolysis by the endogenous human peptidases is also involved in diseases such as metastatic invasion by cancer cells, Alzheimer's disease, emphysema, and rheumatoid arthritis. The peptidases of foreign organisms contribute to human diseases such as those caused by viruses (e.g., human immunodeficiency virus, poliovirus, influenza virus), bacteria (e.g., *Pseudomonas aeruginosa*, *Clostridium botulinum*, *Bacillus anthracis*), fungi (*Candida*), and protozoa (*Trypanosoma*, *Leishmania*, *Entamoeba*). It is therefore natural that peptidases represent one of the principle drug targets of the pharmaceutical industry.

Because of their large number and importance to human health, peptidases require a major information resource, and the MEROPS database seeks to fill this role. Any database requires the best possible classification and nomenclature for the objects it deals with, and it is the classification of peptidases described herein that is used in the MEROPS database. Also, as necessary, the system is modified and developed in the MEROPS database, so it is sometimes referred to as the MEROPS system, although it can perfectly well stand alone.

The modern system for the classification of peptidases is based on proposals we made in 1993, but before describing that it is appropriate to mention two earlier systems that are still of value. Hartley (1960) made a major contribution to the classification of peptidases by recognizing four types that we now know as serine peptidases, cysteine peptidases, aspartic peptidases, and metallopeptidases. More

recently, two additional catalytic types have been recognized: threonine peptidases (Seemüller *et al*., 1995) and glutamic peptidases (Fujinaga *et al*., 2004). A few peptidases are still of unknown catalytic type.

The second system that must be mentioned is that of the Enzyme Nomenclature of the International Union of Biochemistry and Molecular Biology (Tipton and Boyce, 2000). This strives to be a comprehensive system in which all enzymes are classified on the basis of knowledge of the reactions they catalyze. It is entirely logical that enzymes be classified in terms of their properties as catalysts, and it is important that the Enzyme Nomenclature continues to be maintained and expanded. However, while the IUBMB system is necessary, it is not sufficient for the classification of peptidases in the genomic era. One reason is that an enzyme can be added to the system only when it has been thoroughly characterized enzymologically, which remains a slow process. This is particularly the case for peptidases, which have substrate specificities that are so complex that they are very difficult to define and describe. A second consideration is that classification based on known enzymological activity does not have predictive value for the properties of an enzyme when it is known only as a nucleotide sequence found in a genome. The MEROPS system complements the IUBMB Enzyme Nomenclature with an orthogonal classification of peptidases that is based on homology (as defined by Reeck *et al*., 1987), and reflects the similarities in molecular structure and biology that stem from the evolutionary relationships.

The modern system for classification of peptidases was introduced by the present authors (Rawlings and Barrett, 1993) and has been further developed since then. It is a hierarchical system in which individual peptidases are recognized and then grouped in families. The families are assembled in clans. Some families have subfamilies, and some clans have subclans. Figure 1 illustrates the different levels of the classification, with the example of the large clan PA. Specific issues that arise at each level of the classification are discussed below.

2. Working with individual peptidases

The lowest level of the classification is that of the individual peptidases. The classification is designed for a particular functional group of proteins, those that catalyze the hydrolysis of peptide bonds and their homologs, and as part of their curation of the MEROPS database the present authors have the responsibility of deciding when a novel peptidase needs to be recognized.

Factors tending to support a decision to create a new peptidase identifier include:

- The availability of amino acid sequence data.
- The existence of enzymological data showing that the peptidase activity is unlike that of any peptidase already in the family or subfamily.
- Evidence that the peptidase is the product of a human or mouse gene that is not already in MEROPS. Human and mouse are favored species in that we consider that every peptidase and homolog encoded by their genomes merits a separate identifier in MEROPS. We plan to add additional species to the list in the future.

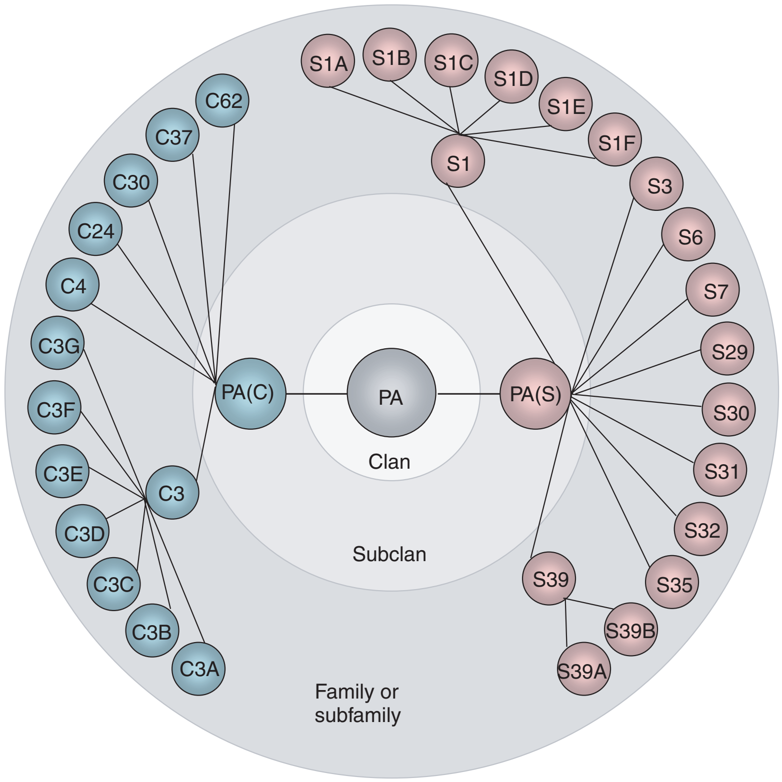


Figure 1 The hierarchical classification, from subfamily to clan. Clan PA contains both cysteine peptidases and serine peptidases, the relationship of which is detectable by strikingly similar protein folds. The letter "P" is used in forming the identifier of the clan because it contains the two types of protein nucleophile peptidases. Subclans PA(C) and PA(S) accommodate the families of cysteine peptidases and serine peptidases respectively. Some families contain subfamilies, numbering as many as seven in family C3; these are handled similarly to simple families

- Available biological data show that the peptidase has a distinctive function or distinctive spatial or temporal distribution in the organism.

Although not a requirement, it does speed up the process of recognition of a new peptidase entity if a good name has been proposed for it. If we are uncertain whether to create a new peptidase entity, we take a view as to whether it will be useful to the scientific community if we do so.

An identifier is assigned to each novel peptidase. This consists of a three-character form of the family name (see below), a dot and a three-digit serial number. If the peptidase cannot be placed in an existing family, a new family is created. Many peptidases have not yet been characterized to the extent that they can be assigned to precise identifiers, and these are provisionally described only as unassigned peptidases in the family or subfamily.

Once a new identifier has been created, we need to consider whether additional homologs are known that are close enough in sequence to merit assignment to

the same identifier, even without biochemical characterization. Typically, these appear to be variants of the same peptidase from other species. We maintain a sequence alignment and a similarity tree for each family and subfamily, and these make it simple to see sequences that cluster together. We may well assign the same MEROPS identifier to the entire cluster. Sequence alignments are made by using the programs MAFFT (Katoh *et al*., 2002) and ClustalW (Thompson *et al*., 1994), and trees are made with either the UPGMA algorithm as implemented in the QUICKTREE program (Howe *et al*., 2002) or the KITSCH algorithm from the PHYLIP package (Felsenstein, 1989). Ideally, only one sequence from an organism is assigned to a given identifier, but, in practice, lineage-specific expansion may have given rise to several close paralogs within a single species, any of which could be assigned to the identifier. Unless we have biochemical evidence to the contrary, all such sequences tend to be assigned to the same identifier. We are reluctant to assign sequences that differ by more than 50% of amino acid residues to the same identifier except in the light of evidence that other properties are similar, because the probability is that key characteristics will have changed at this level of divergence.

An example of the usefulness of similarity trees in distinguishing individual peptidases is provided in Figure 2. This is a tree for just the human and mouse members of family A1. Renin 2 has resulted from a recent duplication of the renin gene in mouse, and human pepsin A and mouse pepsin F, although encoded by orthologous genes (Puente *et al*., 2003), are too divergent to be considered the same peptidase. The decision to give pepsins A and F separate identifiers is supported by experimental evidence of differences in the activity and expression of the enzymes (Chen *et al*., 2001).

Properties that are recorded for every peptidase that has an identifier are (1) the identity of the *holotype* of the enzyme, (2) the accession number for the *type sequence*, (3) the extent of the *peptidase unit*, and (4) the positions of *active site* (catalytic and metal ligand) residues in the sequence.

We assume that every peptidase is encoded by the genomes of multiple organisms so that species variants exist that can be expected to show minor variations in properties. In order to be able to describe the properties of the peptidase with precision, we need a reference species variant. An example of such a definition is that for identifier M10.008 the *holotype* of which is “matrilysin (*Homo sapiens*)”.

Commonly, there are several deposits of sequences for a single peptidase that have minor differences as a result of sequencing errors, inclusion or exclusion of the initiating methionine, or genetic polymorphisms. In the interests of precision, we define a type sequence with a particular accession number for each peptidase. To continue with the example of matrilysin, this is “matrilysin (*Homo sapiens*), Uniprot Accession P09237”.

Many peptidases are multidomain proteins in which one can recognize a part of the molecule that carries the active site residues and is responsible for peptidase activity (termed the *peptidase unit*) as distinct from parts that may have other functions (Figure 3). Often, the peptidase unit is equivalent to the entire, mature sequence of the smallest active peptidase in the family. Figure 3 shows how the peptidase units have been defined for the peptidases related to collagenase 1 in family M10. The shortest sequence is that of matrilysin, and since this is an active

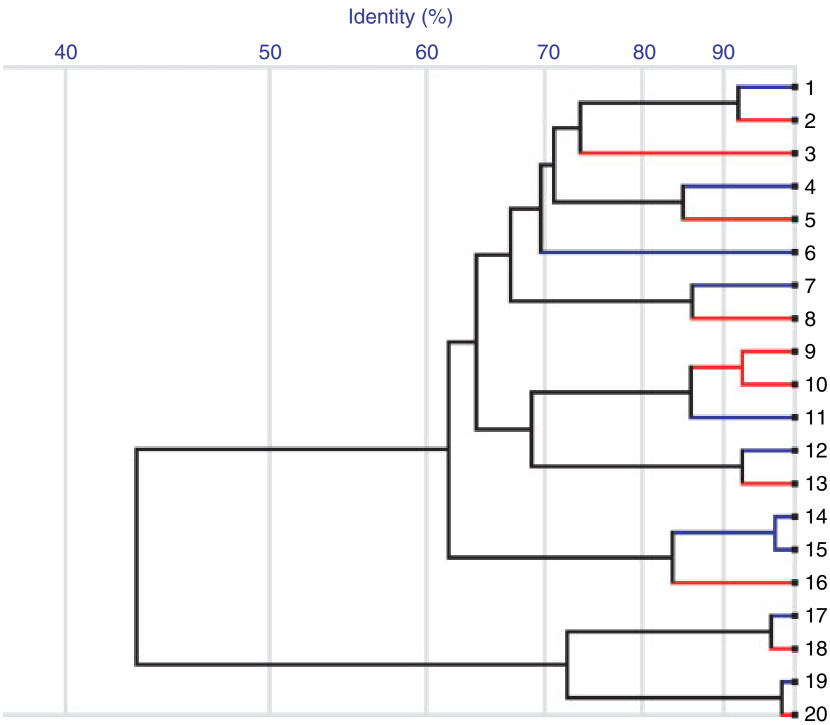


Figure 2 Distinct human and mouse peptidases resolved with the use of a sequence-similarity tree for family A1 of pepsin. The tree was made with only human and mouse sequences for simplicity. The ultimate branch is shown in blue for a mouse sequence, and red for human. Numbers across the top are percentage identity. Key to sequences (MEROPS identifier, species, protein name, UniProt accession, peptidase unit): (1) A01.010 (*Mus musculus*) cathepsin E (P70269) (residues: 69-394); (2) A01.010 (*Homo sapiens*) cathepsin E (P14091) (residues: 67-393); (3) A01.001 (*H. sapiens*) pepsin A (P00790) (residues: 63-388); (4) A01.006 (*M. musculus*) chymosin (-) (residues: 61-376); (5) A01.006 (*H. sapiens*) chymosin (-) (residues: 68-433); (6) A01.051 (*M. musculus*) pepsin F (Q9JKE6) (residues: 61-385); (7) A01.003 (*M. musculus*) gastricsin (-) (residues: 67-389); (8) A01.003 (*H. sapiens*) gastricsin (P20142) (residues: 61-388); (9) A01.972 (*H. sapiens*) napsin B pseudogene (napsin B pseudogene) (Q9UHB3) (residues: 69-393); (10) A01.046 (*H. sapiens*) napsin A (human-type) (O96009) (residues: 69-399); (11) A01.049 (*M. musculus*) napsin A (mouse-type) (O09043) (residues: 63-387); (12) A01.009 (*M. musculus*) cathepsin D (P18242) (residues: 71-405); (13) A01.009 (*H. sapiens*) cathepsin D (P07339) (residues: 71-410); (14) A01.007 (*M. musculus*) renin (P06281) (residues: 76-399); (15) A01.008 (*M. musculus*) renin-2 (P00796) (residues: 75-398); (16) A01.007 (*H. sapiens*) renin (P00797) (residues: 78-403); (17) A01.041 (*M. musculus*) memapsin-1 (Q9R1P7) (residues: 90-425); (18) A01.041 (*H. sapiens*) memapsin-1 (Q9Y5Z0) (residues: 93-428); (19) A01.004 (*M. musculus*) memapsin-2 (P56818) (residues: 77-416); (20) A01.004 (*H. sapiens*) memapsin-2 (P56817) (residues: 77-416)

peptidase we take the parts of the sequences of the other members of the family that align with the mature matrilysin sequence as their peptidase units. Matrilysin provides an example of the typical, simple style of definition of a peptidase unit: “matrilysin (*Homo sapiens*), Uniprot Accession P09237 (peptidase unit: 93-267)”. The two gelatinases in family M10 are more complex, in that they have domains

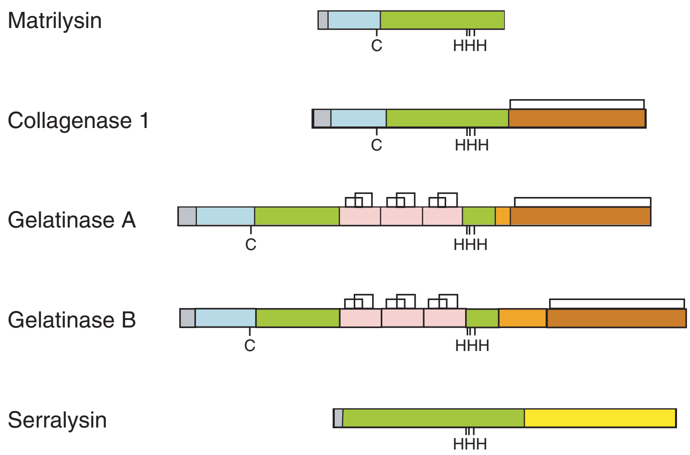


Figure 3 Peptidase units in family M10. The domain organizations of several members of peptidase family M10 are shown. The peptidase units are colored in green. The domains excluded from the peptidase units are signal peptides (gray), propeptides (blue), hemolysin-like domains (brown), fibronectin-like repeats (pink), and a calcium-dependent secretion domain (yellow). The positions of disulfide bridges are marked, as are the cysteine residues of the "cysteine switches" that maintain the latency of the proenzymes, and the zinc-binding histidine residues of the active sites

nested within the peptidase unit, and the special case of the interrupted peptidase unit is dealt with as: "gelatinase A (*Homo sapiens*), Uniprot Accession P08253 (peptidase unit: 108-214, 390-461)". Amino acid residue numbers can only be specified in the context of a particular version of the amino acid sequence, and the Uniprot accessions indicated here are termed the "type sequences" for the MEROPS identifiers M10.008 and M10.003, respectively. It is the sequence of each peptidase unit rather than that of the whole protein that is aligned and used in constructing trees for MEROPS. The Pfam (*see* Article 86, **Pfam: the protein families database**, Volume 6) and SMART databases (Letunic *et al*., 2004) manage protein domains in similar ways to peptidase units in MEROPS.

Data for the positions of catalytic residues in each peptidase sequence allow a check that any additional sequence that is assigned to the same identifier is that of a protein that has at least the potential to be catalytically active.

Since the classification of peptidases is strictly a classification of peptidase units, a problem arises when a single peptidase molecule contains more than one peptidase unit: no single location in the classification is right for the whole of such a compound peptidase molecule. To deal with the half-dozen or so of such cases, a special type of peptidase identifier starting in "X" exists for each compound peptidase, and an identifier of the standard type is used for each of the individual peptidase units. For example, the somatic form of peptidyl-dipeptidase A (angiotensin-converting enzyme), which contains two homologous peptidase units (M02.001 and M02.004), is XM02-001. In addition to the expected summary pages for M02.001 and M02.004, there is a special summary card for XM02-001.

3. Classification at the family and subfamily levels

We create a new peptidase family when there are data to show that a protein that cannot be placed in an existing family has peptidase activity. In founding a new family, we recognize a type-example peptidase for the family.

Every family contains peptidases of just a single catalytic type, and the catalytic type is a key property of the family that forms the basis of the family identifier. If a novel peptidase is founding a new family, experimental data as to its catalytic type are sought. Typically, amino acid residues that are conserved in all of a homologous set of peptidases will have been mutated experimentally, and effects on activity determined; a conserved residue that is essential for activity may well show the catalytic type. The initial letter of the new family identifier reflects the catalytic type ("A" for aspartic, "C" for cysteine, "G" for glutamic, "M" for metallo, "S" for serine, "T" for threonine, or rarely "U" for unknown), and this is followed by a serial number. A family identifier is permanently deleted if it is subsequently found that none of the proteins in the family is a peptidase after all. The identifier may also be deleted if the family is merged with another family, because the lower-numbered identifier of the two is used for the merged family.

The majority of families of peptidases (139 out of the 180 families in MEROPS Release 6.6) do not contain subfamilies, and these can be described as *simple families*, whereas the 31 families that are divided into subfamilies can be termed *compound families*. The similarity trees for the compound families show sets of sequences that are separated by deep divergences (typically falling to 33% of amino acid sequence identity) (Figure 4). When such a distinct subset contains at least one characterized peptidase, the set is treated as a subfamily. A type peptidase is then nominated for the subfamily and an identifier assigned. The identifier of a subfamily is that of the family with the addition of a serial capital letter (e.g., A22B).

The subfamilies of compound families are handled in just the same way as the simple families, but the compound families require special procedures. A family or subfamily may contain just a single peptidase so long as this has been thoroughly characterized.

Additional members of families are found by searches made with the programs FastA (Pearson and Lipman, 1988), BlastP, and TBlastN (Altschul *et al*., 1990). A sequence is considered to be homologous to the query sequence if the expect value (e) is 0.001 or less. Several types of searches are made:

1. The peptidase unit of the type-example peptidase of each family or subfamily is used as the query sequence in a BlastP search of the UniProt database.
2. Newly deposited sequences of peptidases identified during our weekly literature surveys are used in FastA searches of the library of sequences in the MEROPS database, PepUnit.Lib. If the new sequence is found to be significantly related to a sequence in PepUnit.Lib, but not to the type example of that subfamily, then a *transitive relationship* is considered to exist.
3. Sequences that are linked to a type-example sequence only by transitive relationships form a second set of query sequences that are used in BlastP searches of the UniProt database (Apweiler *et al*., 2004).

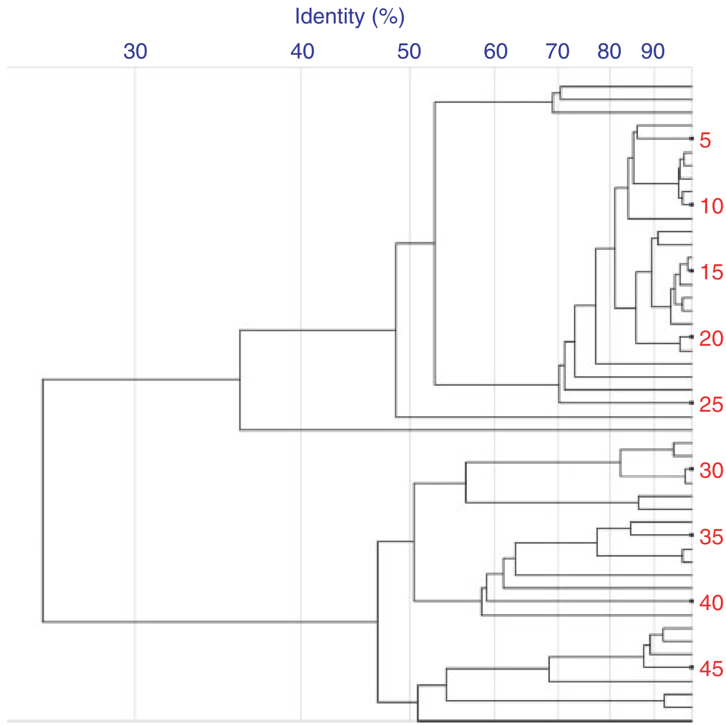


Figure 4 Tree for a compound family of peptidases. The tree for family A22 (from MEROPS Release 6.5) shows the deep divergence to below 30% sequence identity between subfamily A22A (tips 1–27) containing the presenilins and their homologs, and subfamily A22B (tips 28–49) containing signal peptide peptidase and its homologs. The full key to the sequences can be found in the MEROPS database (**http://merops.sanger.ac.uk**)

4. The peptidase unit of the type-example peptidase of each family or subfamily, and of each sequence linked to the type-example sequence only by a transitive relationship, serves as the query sequence in a TBlastN search of the EMBL nucleotide sequence database (Kulikova *et al*., 2004).
5. The peptidase unit of the type-example peptidase of each family or subfamily, and of each sequence linked to the type-example sequence only by a transitive relationship, serves as the query sequence in a FastA search against every newly completed genome sequence deposited at the DDBJ/EMBL/Genbank nucleotide sequence databases.

From time to time, a sequence is encountered that serves as a *linking sequence* between two families. This is a sequence that can be shown to be homologous to one or more members from each family. Rarely, a linking sequence is found in a BlastP or FastA search. More often, the BlastP or FastA results file includes members of another family that are above our expect value threshold. If this happens frequently, it may be that the two families should be merged. We then search for

linking sequences by use of shuffling tests made with the program PRSS34 from the FastA package (Pearson and Lipman, 1988).

4. Classification at the clan and subclan levels

A clan is intended to contain the full set of modern-day peptidases that have arisen from a single ancestral peptidase. These are commonly much more divergent than the peptidases in a single family, so most clans contain several families. The distant relationships at the clan level can sometimes be seen in the order of catalytic residues in the polypeptide chains, or in conserved amino acid sequence motifs around the catalytic residues, but similarities in protein fold provide the strongest justification for grouping families in a clan.

For a clan to be established, the tertiary structure of at least one member of a family must have been solved. Other families may be added to the clan if their peptidases have protein folds that are similar to that of the type peptidase of the clan. The comparison of tertiary structures requires great skill, and we rely upon the views of the authors of papers describing the tertiary structures, the classifications of protein structures by the SCOP (Andreeva *et al*., 2004) and CATH (*see* Article 89, **The CATH domain structure database**, Volume 6) databases, and results obtained from the DALI program (Holm and Sander, 1993).

Peptidase families are grouped in a clan only if the positions of the active site residues (and metal ligands of metallopeptidases) are consistent with common ancestry. There are a number of families that are included in the same SCOP superfamily or are found to be significantly similar by DALI, but for which the active sites are in different positions in the structures. This suggests that the peptidase active sites have evolved independently, although perhaps on homologous protein scaffolds.

The identifier of a clan consists of two letters, the first indicating the catalytic type of all the families in the clan and the second being a serial letter. The initial letter is taken from the set used for family identifiers (A, C, G, M, S, T, U) or may be P, because there are some clans that contain mixtures of protein nucleophile (i.e., C, S, T) families. If deleted, a clan identifier is not reused.

There are families for which no tertiary structure is known and which cannot reliably be assigned to clans. Instead, we assign these families to *pseudoclans* with names A-, C-, M-, S- and U-, depending on catalytic type. Assignment of a family to such a pseudoclan does not imply that its peptidases are related to others in the pseudoclan, but is simply a way of grouping them for practical convenience.

Four clans are divided into *subclans* in MEROPS Release 6.6 (Table 1). Three of these (PA, PB, PC) contain families of mixed protein nucleophile catalytic types. In all of these, the active sites are very similar in structure but the nucleophilic residue has changed. One clan of metallopeptidase also is divided into subclans. This contains the numerous families of peptidases that have the "HEXXH" motif at the active site, and is divided into the gluzincins (subclan MA(E)) and the metzincins (subclan MA(M)) (Hooper, 1994; Bode *et al*., 1993). The name of a

Table 1 Subclans of peptidases. Four clans are divided into subclans in MEROPS Release 6.5, three of them because they contain multiple catalytic types of protein nucleophile peptidases, and the fourth because it contains the different kinds of metallopeptidase families that have been described as gluzincins and metzincins

Clan	Subclan	Explanation of subclan identifier
PA	PA(C)	**C**ysteine-type peptidase
	PA(S)	**S**erine-type peptidase
PB	PB(C)	**C**ysteine-type peptidase
	PB(S)	**S**erine-type peptidase
	PB(T)	**T**hreonine-type peptidase
PC	PC(C)	**C**ysteine-type peptidase
	PC(S)	**S**erine-type peptidase
MA	MA(E)	Glu-zincin (most C-terminal zinc ligand is Glu, **E**)
	MA(M)	Met-zincin (contains the **M**et-turn structure, and the most C-terminal zinc ligand is not Glu)

subclan is formed from the clan name by placing a letter in parentheses at the end; this is not a serial letter, but reflects a characteristic of the active site.

5. The MEROPS database

The classification of peptidases is constantly evolving and expanding to accommodate new data, and the MEROPS database provides the medium through which the successive updates are published. New releases of MEROPS appear on the Internet quarterly, where they are freely accessible to all under general library license (**http://merops.sanger.ac.uk**) (Rawlings *et al.*, 2004a).

The forms of information that are provided in MEROPS can be briefly summarized as follows. For each clan of peptidases, there is a selective sequence alignment showing conservation of amino acids around the active site residues for the type example for each family in the clan. There are also images showing the similarities in secondary structure, and a bibliography. The summary page for the clan shows the type example, the list of families in the clan, and the distribution of peptidases across kingdoms of organisms.

For each family, there is an alignment of the protein sequences of active peptidase units, a tree derived from the alignment, and a figure showing whether a homolog can be found in each of the available whole-genome sequences. There are also images showing the similarities in secondary structure, a bibliography, and collections of human and mouse sequences. The summary lists the type example, links to structure databases such as SCOP (Andreeva *et al.*, 2004), CATH (*see* Article 89, **The CATH domain structure database**, Volume 6), and HSSP (Dodge *et al.*, 1998), databases for domain and family classification such as Pfam (*see* Article 86, **Pfam: the protein families database**, Volume 6), and InterPro (*see* Article 83, **InterPro**, Volume 6), and the PROSITE (*see* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6) database of sequence motifs as well as the HomStrad database of structural alignments (Mizuguchi *et al.*, 1998). There is also a list of peptidases in the family,

and a table of distribution amongst kingdoms of organisms. When a family is divided into subfamilies, these are shown.

At the peptidase level, there is a list of sequence database entries for every organism from which the peptidase is known, as well as a diagram showing the distribution of the peptidase amongst all the species from which any member of the family is known. If tertiary structures have been published, there is a Richardson diagram (Richardson, 1985) of this and a list of PDB entries (Berman *et al*., 2000). Also provided are a bibliography, alignments and tables of known human and mouse expressed sequence tags, and a list of substrates and cleavage sites. The summary shows other names, the full MEROPS classification, comments on pharmaceutical and physiological relevance, and human and mouse genetics.

Access to the data is facilitated by indexes of peptidase names, MEROPS identifiers, and source organisms. There are also CGI and Blast searches.

In Release 6.4, a classification of the proteins that inhibit peptidases was added to the MEROPS database. We found that we could classify the inhibitors by methods similar to those that we had developed for peptidases (Rawlings *et al*., 2004b), and we believe that this demonstrated that a similar approach could be applied to other functionally related groups of proteins.

6. Conclusions

The homology-based system for the classification of peptidases is described. This is well suited to the genomic era because it allows predictions about the functions of novel peptidases while they are known only by sequence. The system has now become established worldwide for identifying peptidase families and provides the structure of the public MEROPS database. The separate set of proteins that act as inhibitors of peptidases was recently classified in the same way, and we believe that this sort of system could be applied to other functional groups of proteins.

Related articles

Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 89, **The CATH domain structure database**, Volume 6

Acknowledgments

We thank colleagues, past and present, who have made essential contributions to the MEROPS database: Emmet O’Brien, Dominic Tolle and Fraser Morton. We also thank our colleagues at the Pfam database for helpful advice. We thank the Medical Research Council and the Wellcome Trust for funding.

Further reading

Holm L and Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, **25**, 231–234.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family. *Nucleic Acids Research*, **32**(1), D226–D229.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R and Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**(1), D115–D119.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The protein data bank. *Nucleic Acids Research*, **28**, 235–242.

Bode W, Gomis-Rüth FX and Stöcker W (1993) Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins'. *FEBS Letters*, **331**, 134–140.

Chen X, Rosenfeld CS, Roberts RM and Green JA (2001) An aspartic proteinase expressed in the yolk sac and neonatal stomach of the mouse. *Biology of Reproduction*, **65**, 1092–1101.

Dodge C, Schneider R and Sander C (1998) The HSSP database of protein structure sequence alignments and family profiles. *Nucleic Acids Research*, **26**, 313–315.

Felsenstein J (1989) PHYLIP - phylogeny inference package. *Cladistics*, **5**, 164–166.

Fujinaga M, Cherney MM, Oyama H, Oda K and James MN (2004) The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 3364–3369.

Hartley BS (1960) Proteolytic enzymes. *Annual Review of Biochemistry*, **29**, 45–72.

Holm L and Sander C (1993) Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, **233**, 123–38.

Hooper NM (1994) Families of zinc metalloproteases. *FEBS Letters*, **354**, 1–6.

Howe K, Bateman A and Durbin R (2002) QuickTree: building huge Neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.

Katoh K, Misawa K, Kuma K and Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K and Eberhardt R, *et al*. (2004) The EMBL nucleotide sequence database. *Nucleic Acids Research*, **32**(1), D27–D30.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**(1), D142–D144.

Mizuguchi K, Deane CM, Blundell TL and Overington JP (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science*, **7**, 2469–2471.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Puente XS, Sanchez LM, Overall CM and Lopez-Otin C (2003) Human and mouse proteases: a comparative genomic approach. *Nature Reviews. Genetics*, **4**, 544–558.

Rawlings ND and Barrett AJ (1993) Evolutionary families of peptidases. *The Biochemical Journal*, **290**, 205–218.

Rawlings ND, Tolle DP and Barrett AJ (2004a) MEROPS: the peptidase database. *Nucleic Acids Research*, **32**, Database issue, D160–D164.

Rawlings ND, Tolle DP and Barrett AJ (2004b) Evolutionary families of peptidase inhibitors. *The Biochemical Journal*, **378**, 705–716.

Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, Mclachlan AD, Margoliash E and Jukes TH (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**, 667.

Richardson JS (1985) Schematic drawings of protein structures. *Methods in Enzymology*, **115**, 359–380.

Seemüller E, Lupas A, Stock D, Löwe J, Huber R and Baumeister W (1995) Proteasome from *Thermoplasma acidophilum*: a threonine protease. *Science*, **268**, 579–582.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Tipton K and Boyce S (2000) History of the enzyme nomenclature system. *Bioinformatics*, **16**, 34–40.

Specialist Review

Transporter protein families

Qinghu Ren and Ian T. Paulsen
The Institute for Genomic Research, Rockville, MD, USA

The cellular membrane is a selectively permeable barrier between the cell and the extracellular environment. The movement of molecules and ions across cellular membrane is mediated by selective membrane transporters embedded in the phospholipid bilayer. Transporters function in the acquisition of organic nutrients, extrusion of toxic and waste compounds, maintenance of ion homeostasis, environmental sensing and cell communication, and other important cellular functions (Saier, 1999a). The complement of transporter systems displays great variations in different species (Figure 1). Between 2 and 16% of open reading frames (ORFs) in the prokaryotic and eukaryotic genomes are predicted to encode membrane transport proteins, emphasizing the important roles of transporters in the lifestyles of all species (Paulsen *et al*., 2000).

Various transport systems differ in their putative membrane topology, energy-coupling mechanism, and substrate specificities (Saier, 2000). The most common energy-coupling mechanisms are the utilization of adenosine triphosphate (ATP), phosphoenolpyruvate (PEP), or chemiosmotic energy in the forms of sodium ion or proton electrochemical gradients. Transporters of similar functions characteristically cluster together in phylogenetic analyses, hence substrate specificity appears to be a conserved evolutionary trait in transporters, and phylogeny can provide a rational basis for functional assignment (Paulsen *et al*., 2000; Paulsen *et al*., 1998b). The Transporter Classification (TC) system (**http://saier-144-164.ucsd.edu/tcdb/**) represents a systematic approach to classify membrane transporter families according to the mode of transport, energy-coupling mechanism, molecular phylogeny, and substrate specificity (Busch and Saier, 2004; Saier, 1999). The TC system is analogous to the Enzyme Commission (EC) system for classification of enzymes, except that it incorporates both functional and phylogenetic information. Transport mode and energy-coupling mechanism serve as the primary base for the classification owing to their relatively stable characteristics. There are four characterized classes of solute transporters in the TC system: channels, secondary transporters, primary active transporters, and group translocators (Figure 2). Transporters of unknown mechanism or function are included as a distinct class.

- *Class 1. Channels/pores:* Channels are energy-independent transporters that transport water, specific types of ions, or hydrophilic small molecules down a concentration or electric gradient with higher rates of transport and lower

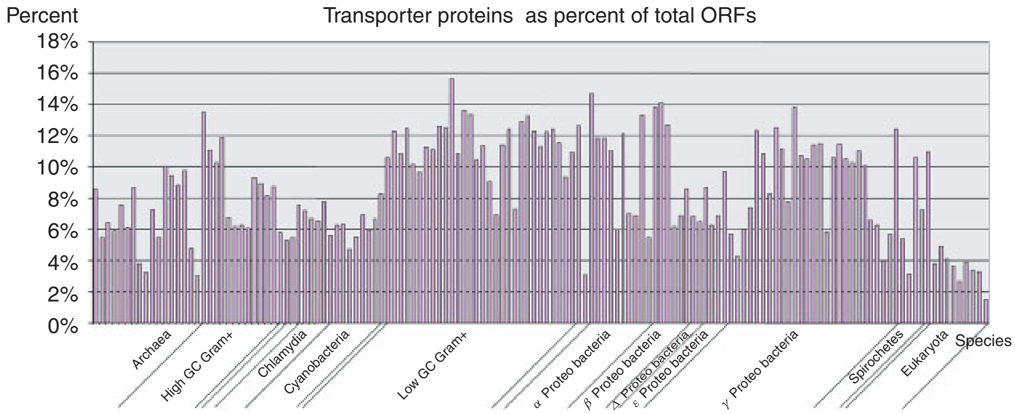


Figure 1 A comparison of the overall numbers of recognized transporter proteins relative to the total ORF numbers for the 141 species analyzed in the TransportDB database

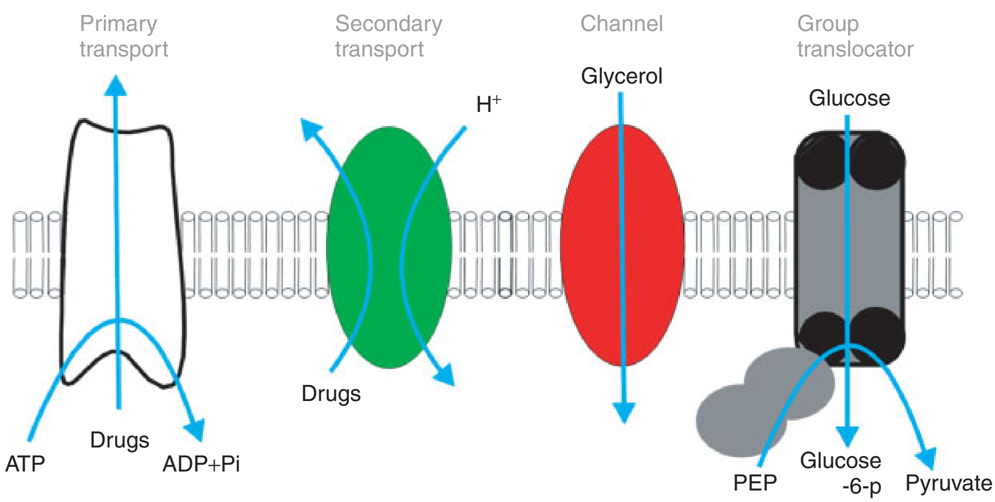


Figure 2 Representative examples of the four main types of transporters. Primary transporter, *Lactococcus lactis* LmrP multidrug efflux pump (Bolhuis *et al*., 1995); secondary transporter, *Staphylococcus aureus* QacA multidrug efflux transporter (Tennent *et al*., 1989); channel, *Escherichia coli* GlpF glycerol channel (Sweet *et al*., 1990); and group translocator, *E. coli* PtsG/Crr glucose PTS transporter (Henderson *et al*., 1977)

stereospecificity compared to other transporter classes. Most of the ion channels, referred to as *gated channels*, open only in response to specific chemical or electric signals.

- *Class 2. Electrochemical potential-driven transporters (secondary transporters):* Secondary transporters (also called carriers) utilize an ion or solute electrochemical gradient, for example, proton/sodium motive force, to drive the transport process. Uniporters move a single type of molecule down its concentration gradient. Antiporters and symporters couple the transport of ions or molecules against

their concentration gradient with the movement of one or more different ions or molecules down their gradient.

- *Class 3. Primary active transporters:* Primary active transporters couple transport process to a primary source of energy, such as a chemical reaction (e.g., ATP hydrolysis), and move substrates across membrane against a chemical concentration gradient or electric potential or both.
- *Class 4. Group translocators:* Group translocators modify their substrates during the transport process. For example, the bacterial phosphotransferase system (PTS) phosphorylates its sugar substrates using phosphoenolpyruvate as the phosphoryl donor and energy source and releases them into cytoplasm as sugar phosphates.
- *Class 9. Incompletely characterized transporter systems:* Transporter protein families for which insufficient information is available to allow classification in a defined class belong to this class.

Each transporter class is further classified into individual families or superfamilies according to their function, phylogeny, and/or substrate specificity (Saier, 2000). A specific transport protein can be represented by a TC number, which normally has five components as follows: V.W.X.Y.Z. V (a number) corresponds to the transporter type (i.e., channel, secondary transporters, primary active transporters, or group translocators); W (a letter) corresponds to the transporter subtype, which in the case of primary active transporters refers to the energy source used to drive transport; X (a number) corresponds to the transporter family; Y (a number) corresponds to the subfamily in which a transporter is found; and Z corresponds to the substrate or range of substrates transported. For example, the well-characterized *Escherichia coli* LacY lactose permease (Kaback *et al*., 2001) is represented in the TC system as 2.A.1.5.1, where "2" indicates it is a secondary transporter, "A" indicates it is a uniporter/symporter/antiporter, "1" indicates it belongs to the Major Facilitator Superfamily (MFS), "5" indicates it belongs to an oligosaccharide symporter subfamily within the MFS family, and "1" indicates it is a lactose/proton symporter. Transport system families included in the current TC system are available in database format on the web (TCDB, **http://saier-144-164.ucsd.edu/tcdb/**). TCDB provides detailed information and references for each transporter class, subclass, family, subfamily, and individual proteins. The website also provides a search tool that allows users to search by keyword, gene name, family, or protein sequence (Busch and Saier, 2004).

Membrane transport proteins typically consist of multiple transmembrane α-helical segments (*see* Article 38, **Transmembrane topology prediction**, Volume 7). The hydrophobic nature and solubility only in presence of detergents make them difficult to study experimentally (*see* Article 65, **Analysis and prediction of membrane protein structure**, Volume 7). Genomic and bioinformatics analyses provide an attractive alternative. More than 180 prokaryotic and eukaryotic genomes have been published to date, with over 900 additional genome-sequencing projects currently underway around the world. An in-depth look at transport proteins is vital to the understanding of the metabolic capability of organisms. We have previously undertaken systematically genome-wide comparisons of the transporter systems and families across a wide range of sequenced organisms (Paulsen *et al*., 2000; Paulsen *et al*., 1998b), and similar studies have been done on specific

genomes (Meidanis *et al*., 2002; Paulsen *et al*., 1998a). Convenient and effective bioinformatics methods for whole-genome transporter analysis have been developed (Ren *et al*., 2004). TransportDB (**http://www.membranetransport.org/**) is a relational database describing the predicted cytoplasmic membrane transport protein complement for organisms whose complete genome sequences are available. For each organism, its complete membrane transport complement was identified, classified into protein families according to the TC classification system, and functional predictions are provided. The web pages also include various features such as a search engine, a BLAST search tool, a comparative genomics page, and phylogenetic information of individual transporter families. Currently, TransportDB contains data from 141 species, including 115 bacteria, 17 archaea, and 9 eukaryotes. Transporter proteins from these organisms were identified and classified into 134 families, including 7 families of primary transporters, 80 families of secondary transporters, 32 channel protein families, 2 phosphotransferase systems, and 13 unclassified families. Some of these families are very large superfamilies with numerous members, such as the ATP-binding Cassette (ABC) Superfamily and the MFS, both of which are widely distributed in bacteria, archaea, and eukaryotes (Table 1). Some are very small families with only few members present in certain species.

The occurrence of transporter families in the three domains of life (*see* Article 40, **The domains of life and their evolutionary implications**, Volume 7) is shown in Figure 3. There are 41 eukaryotic-specific families, mainly the channel families in class 1, which exist exclusively in multicellular eukaryotic species such

Table 1 The top 10 largest transporter families in the TransportDB database. Most of them are widely distributed in bacteria, archaea, and eukaryotic species, with the exception of SSPTS (bacteria) and MC (eukaryotes)

Family ID	Family name	TC#	Count[a]	Occurrence[b]
ABC	The ATP-binding Cassette (ABC) Superfamily	3.A.1	913	B, A, E
MFS	The Major Facilitator Superfamily (MFS)	2.A.1	819	B, A, E
DMT	The Drug/Metabolite Transporter (DMT) Superfamily	2.A.7	511	B, A, E
RND	The Resistance-Nodulation-Cell Division (RND) Superfamily	2.A.6	433	B, A, E
SSPTS	Sugar Specific Phosphotransferase System	4.A	415	B
APC	The Amino Acid-Polyamine-Organocation (APC) Family	2.A.3	408	B, A, E
MOP	The Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase Superfamily	2.A.66	405	B, A, E
VIC	The Voltage-gated Ion Channel (VIC) Superfamily	1.A.1	394	B, A, E
P-ATPase	The P-type ATPase (P-ATPase) Superfamily	3.A.3	384	B, A, E
MC	The Mitochondrial Carrier (MC) Family	2.A.29	242	E

[a]Transporter count in 141 species analyzed
[b]B: bacteria; A: archaea; E: eukaryotes.

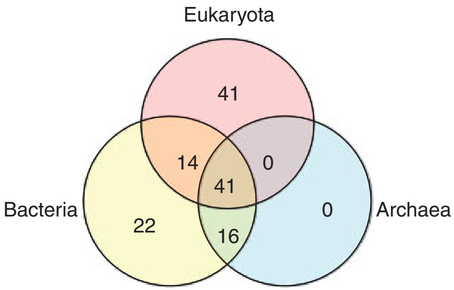


Figure 3 Distribution of transporter families among three domains of life. A total of 134 families have been identified, including 41 eukaryotes-specific families, mainly the various ion channels, which exist exclusively in higher eukaryotic species, and 38 prokaryotes-specific families, of which 22 families only exist in bacteria and 16 are shared by bacteria and archaea. About 41 families are shared by both prokaryotic and eukaryotic species

as *Drosophila melanogaster*, *Arabidopsis thaliana*, and humans, and usually are restricted to a single organismal type. Prokaryotic-specific transporter families (38 families), of which 22 families only exist in bacteria and 16 are shared by bacteria and archaea, are more prevalent than those specifically found in eukaryotes, possibly reflecting that prokaryotes endure more evolutionary pressure, which force them to maintain diversity so that they could be adaptive in response to the various environmental stress conditions. There are no archaea-specific transporter families known, partly reflecting the sparse functional characterization of archaeal proteins. About 41 families are ubiquitous in all three domains of life, highlighting the fundamental importance of these families. Most of them are found within subclass 2A. There are 14 families shared by bacteria and eukaryotes and 16 shared by bacteria and archaea. Some of these families shared only in two domains could be ubiquitous with more sequence data available for analysis.

In general, eukaryotic species, especially multicellular eukaryotic organisms, such as *Drosophila* (640 transporters), *Arabidopsis* (855 transporters), *Caenorhabditis elegans* (656 transporters), and humans (805 transporters), exhibit much larger numbers of total transporter proteins than prokaryotic and lower eukaryotic systems, but the transporter proteins in eukaryotes occupy a smaller percent of total ORFs (average 3.3%) compared to those in prokaryotic systems (average 5.4%) (Figure 1). Although multicellular eukaryotes exhibit fewer transporter families than prokaryotes, they have generated an extraordinary number of paralogs by gene duplication or expansion within certain families, such as ABC, MFS, and the Voltage-gated Ion Channel (VIC) Superfamily, probably evolved into tissue-specific and/or organelle-specific functions. Species with intracellular lifestyles, including intracellular prokaryotic pathogens such as *Mycoplasma* spp. and *Chlamydia* spp., symbionts such as *Buchnera* sp., *Wolbachia* sp., and *Wigglesworthia brevipalpis*, and eukaryotic parasites such as *Plasmodium falciparum* and *Encephalitozoon cuniculi*, usually exhibit a more limited repertoire of membrane transporters. Most of the sequenced archaea are also underrepresented in transporters relative to the total ORFs, in some cases because of their largely autotropic metabolism, or possibly

because of more limited experimental characterization of these organisms. Soil-/plant-associated organisms, such as *Mesorhizobium loti*, *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, *Bradyrhizobium japonicum*, *Pseudomonas aeruginosa*, and *Bacillus subtilis*, all of which have a free-living lifestyle, possess a robust array of transporter systems (Paulsen *et al*., 2004). The prevalent number of transporter families and total transporter proteins are consistent with the evolution of adaptive responses to changes in environment, which permits these organisms to thrive in diverse ecological niches.

Phylogenetic analyses of membrane transporters can be used to examine in detail specific subtypes or subfamilies across all the sequenced genomes. Previous studies have shown that substrate specificity is a highly conserved evolutionary trait and that transporters with similar substrate specificities tend to cluster together (Chang *et al*., 2004; Paulsen *et al*., 2000; Saier, 1996; Saier *et al*., 1999b). An example of this phenomenon is demonstrated in the phylogenetic tree of human ABC transporters displayed in Figure 4. The ABC transporter superfamily is one of the largest families

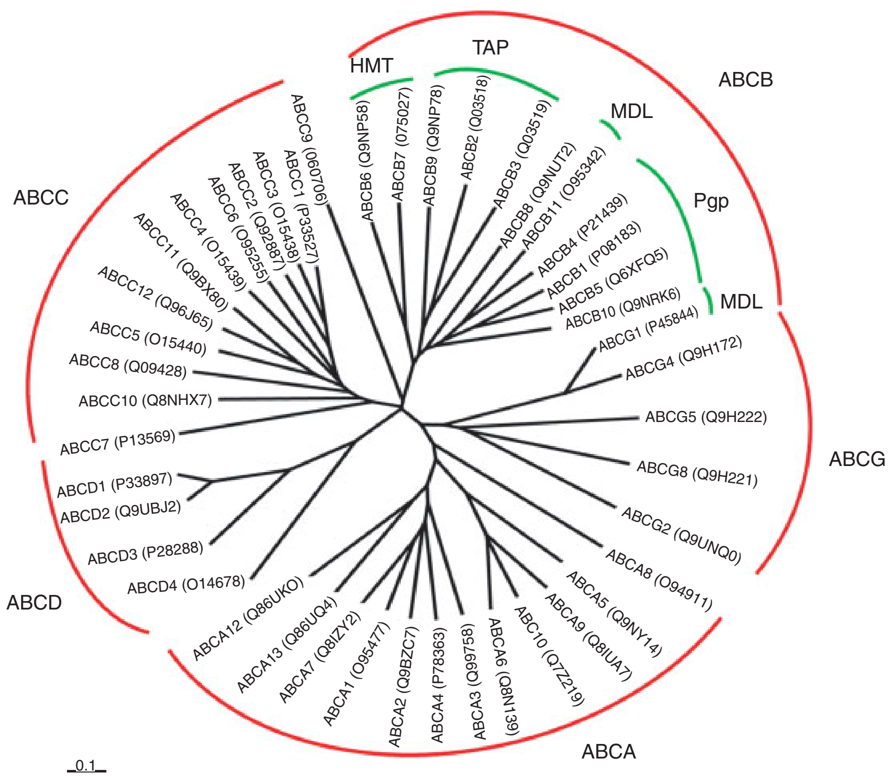


Figure 4 Unrooted phylogenetic tree of human ABC transporters. UniProt accession numbers are used to represent proteins. Five major subfamilies are labeled. The ABCB subfamily can be further classified into smaller groups with similar functions. Predicted ABC transporters are aligned to the Pfam model (*see* Article 86, **Pfam: the protein families database**, Volume 6) ABC_tran (accession PF00005). The Neighbor-joint tree was constructed using CLUSTALX (Thompson *et al*., 1997)

in the genomes of both prokaryotes and eukaryotes, with thousands of identified members. Most ABC transporters consist of four structural domains: two very hydrophobic transmembrane (TM) domains, which contain six membrane-spanning α-helices and two hydrophilic cytoplasmic ATP-binding domains, which transfer the energy for the transport process. ABC transporters are involved in the import or the export of a diverse substrate spectrum, including drugs, inorganic anions and cations, lipids, amino acids and peptides, sugars, carboxylates, and so on. ABC importers typically require an additional substrate-binding protein and are prokaryotic-specific. Prokaryotic ABC transport systems usually carry TM and ABC domains encoded by separate genes that are often organized in gene operon or cluster. ABC exporters are found in both prokaryotic and eukaryotic species, which are involved in the extrusion of waste, drugs, and extracellular toxins. The TM and ABC domains are usually fused in different combinations (Dean and Allikmets, 2001).

The whole-genome analysis of human transporters indicated that the human genome encodes 44 putative ABC transporters. Phylogenetic tree analyses were conducted on the ATP-binding domain of these transporters (Figure 4). Five major subfamilies can be identified with different functions and/or substrate specificities:

- The ABCA subfamily comprises 12 human transporters. The ABCA subfamily proteins are highly conserved in mammals and are present in *Drosophila*, *Arabidopsis*, *C. elegans*, *Anopheles gambiae*, *Dictyostelium discoideum*, *Cyanidioschyzon merolae*, and *Aspergillus fumigatus*, but are absent from yeast and intracellular parasites like *Cryptosporidium parvum*, *P. falciparum*, and *E. cuniculi* (Table 2), suggesting that ABCA family genes have been lost during evolution from the majority of the fungi and lower eukaryotic lineage. These transporters can be further divided into two distinct clusters on the basis of phylogenetic analyses: ABCA1-4, ABCA7, and ABCA 12-13 form one group, and ABCA5, 6, 8, 9, 10 belong to the other. Two members in this group, ABCA1 and ABCA4 have been studied in detail. Mutations in the human ABCA1 gene cause Tangier disease (Hoffman and Fredrickson, 1965), which is characterized by deficient efflux of lipids from macrophages and absence of high-density lipoprotein (HDL) cholesterol in the plasma, suggesting its roles in cholesterol and phospholipids efflux (Young and Fielding, 1999). Other studies indicate that the human ABCA1 may be involved in the process of clearance of cells dying by apoptosis. ABCA1 transcripts are upregulated in the macrophages recruited to areas of developmental cell death (Luciani and Chimini, 1996). The *in vivo* loss-of-function and *in vitro* gain-of-function experiments further established its role in the engulfment of dead cells. ABC1-deficient macrophages show an impaired ability of engulfing apoptotic preys, a phenotype that could be suppressed by the acquisition of phagocytic behavior by ABC1 transfectants (Hamon *et al*., 2000). The human ABCA4 gene encodes the rod and cone photoreceptor Rim protein functioning in the transport of retinol (Vitamin A) derivatives from the photoreceptor outer segment disks into the cytoplasma (Koenekoop, 2003). Mutations in the ABCA4 gene have been linked to multiple eye disorders (Allikmets, 2000).
- The ABCB subfamily can be divided further into four distinct functional groups. The largest group, Pgp subfamily, consists of ABCB1, ABCB4, ABCB5, and ABCB11. These proteins, encoding a broad specificity efflux pump

P-glycoprotein (Pgp), have been suggested to be responsible for multidrug resistance in human cells (Ambudkar *et al*., 2003). ABCB1 (MDR1) was the first human ABC transporter characterized (Juliano and Ling, 1976) and was demonstrated to transport hydrophobic substrates such as drugs, xenobiotics, and steroids (Ambudkar *et al*., 2003). In the epithelial cells of the gastrointestinal tract, liver, and kidney, and capillaries of the brain, testes, and ovaries, ABCB1 acts as a barrier to the uptake of toxic metabolites, and promotes their excretion in the bile and urine. ABCB1 also affects the pharmacokinetics of many commonly used drugs, including anticancer drugs (Hoffmeyer *et al*., 2000). The ABCB2, ABCB3, and ABCB9 form a cluster, TAP, and are involved in eukaryotic peptide export. The ABCB2 and ABCB3 form heterodimers to translocate peptides generated by the proteosome into the endoplasmic reticulum (Abele and Tampe, 1999). ABCC9 was found to be associated with lysosomes (Zhang *et al*., 2000). The HMT cluster, including ABCB6 and ABCB7, contains homologs to the yeast ATM1 gene, which is essential for the transport of iron-sulfur clusters from the mitochondrial matrix to the cytosol and, therefore, involved in the maintenance of mitochondrial ion homeostasis (Lill and Kispal, 2001). The MDL cluster includes another group of mitochondrial ABC transporters (ABCB8 and ABCB10). These are homologs to the yeast MDL1 and MDL2 genes, which have been suggested to be involved in the mitochondrial peptide export process (Young *et al*., 2001).

- The ABCC subfamily contains 12 members distributed to two major branches on the tree, which are involved in the efflux of organic anions and drugs conjugated with anionic molecules such as glutathione or glucuronide derivatives (ABCC1, ABCC4, ABCC5, etc.) (Kruh and Belinsky, 2003), chloride ion channel formation (ABCC7/CFTR) (Kleizen *et al*., 2000), and ion channel regulation (ABCC8 and ABCC9) (Gribble and Reimann, 2003). Cystic fibrosis, one of the most common inherited human diseases (Quinton, 1999), is caused by mutations in ABCC7 (CFTR) (Riordan *et al*., 1989), which forms an anion-selective channel functioning in epithelial chloride transport. ABCC8 (SUR1) is a high-affinity receptor for sulfonylurea, a class of drugs used to increase insulin secretion. It regulates the ATP-sensitive potassium (K-ATP) channels in pancreatic β-cells (Ashcroft, 2000). ABCC9 (SUR2) is an isoform of ABCC8 and expressed more ubiquitously (Isomoto *et al*., 1996). One additional member, ABCC13, is a pseudogene that encodes a truncated protein instead of a functional ABC protein, therefore, is not included in our analysis (Annilo and Dean, 2004).
- The ABCD subfamily includes four homologs to the adrenoleukodystrophy protein ALDp (ABCD1), which is defective in X chromosome–linked adrenoleukodystrophy (ALD), a neurodegenerative disorder with impaired peroxisomal oxidation of very long chain fatty acids (Mosser *et al*., 1993). It has been proposed that the function of these transporters is the export of very long chain fatty acids.
- The ABCG subfamily consists of five proteins homologous to the *Drosophila* white, brown, and scarlet genes, which are involved in the transport of eye pigment precursors, such as guanine and tryptophan (Chen *et al*., 1996). ABCG1 is highly induced in lipid-loaded macrophages, suggesting a role in cholesterol and phospholipids trafficking (Klucken *et al*., 2000).

Table 2 Comparisons of the ABC-type transporter subfamilies in eukaryotic species

	Human[a,b]	*Drosophila*[a,b]	*C. elegans*[a]	*Arabidopsis*[a,c]	*A. gambiae*[d]	*Dictyostelium*[e]	*P. falciparum*[a]	*E. cuniculi*[a,f]	*C. parvum*[a]	*C. merolae*[a]	*A. fumigatus*[a]	*S. cerevisiae*[a]
ABCA	12	11	6	12	6	12	0	0	0	2	1	0
ABCB	11	8	21	27	5	9	7	6	1	6	13	4
ABCC	12	13	6	16	14	14	3	0	10	2	12	7
ABCD	4	2	5	2	1	3	0	0	1	3	2	2
ABCG	5	14	9	44	12	21	1	5	1	8	15	10
Others	0	3	1	9	2	3	0	0	0	7	2	1
Total	44	51	48	110	40	62	11	11	13	27	45	24

[a]Ren *et al*. (2004).
[b]Dean *et al*. (2001).
[c]Sanchez-Fernandez *et al*. (2001).
[d]Roth *et al*. (2003).
[e]Anjard *et al*. (2002).
[f]Cornillot *et al*. (2002).

Two other subfamilies, ABCE (RNase L inhibitor) (Bisbal *et al*., 1995) and ABCF (yeast protein GCN20 that mediates the activation of the eIF-2 alpha kinase) (Marton *et al*., 1997), do not exhibit any transmembrane segments and are not involved in transport processes. Therefore, these families are not included in this analysis.

Table 2 compares the number of ABC transporters in different subfamilies. Humans do not have significantly higher number of transporters than other eukaryotes. *Arabidopsis* exhibits the largest inventory of ABC transporters, twice as many as in human, *Drosophila* and *C. elegans*, which reflects the extraordinary metabolite versatility of plants and the requirement of exporting the large complement of toxic secondary metabolites out of the compartments against the steep concentration gradients. *Arabidopsis* ABC transporters are also involved in chlorophyll biosynthesis, formation of Fe/S clusters, stomatal movement, and probably ion fluxes, suggesting their central roles in many plant growth and developmental processes (Martinoia *et al*., 2002; Rea *et al*., 1998). *Arabidopsis* also has a dominant number of transporters in each individual subfamily except ABCD. Intracellular parasites, such as *E. cuniculi*, *C. parvum*, and *P. falciparum*, possess the lowest number of ABC transporters, with the apparent loss or reduction in ABCA and ABCD subfamilies. ABCB, ABCC, and ABCG subfamilies are widely distributed across all eukaryotic species analyzed except the lack of ABCC transporters in *E. cuniculi*, suggesting an essential role for these transporters in all eukaryotes. The degree of gene expansion varies for different subfamilies in different species. For example, the ABCD subfamily is greatly expanded in *Arabidopsis,* with almost 10 times as many subfamily members as in humans, while humans and *Arabidopsis* have a similar number of ABCC subfamily transporters. *Anopheles gambiae* appears to have less proteins belonging to the ABCA and ABCB subfamilies compared to *Drosophila* and humans, but the ABCG subfamily genes are expanded, leading to twice as many subfamily members as in humans. The comparisons of ABC transporter subfamilies in different organisms indicate that the ABC superfamily evolved by a gene "birth-and-death" process (Dean *et al*., 2001). The constituents of the ABC subfamilies are generated by gene duplications. Some of the genes persist and develop highly specific functions, while others are deleted or become nonfunctional pseudogenes, like ABCC13 (Annilo and Dean, 2004).

The discussion of the ABC transporter family presented here is an example of a detailed phylogenomic analysis of one transporter family and the type of the insights that this style of analysis reveals. With the complete transporter data from 141 sequenced species, we have been able to conduct phylogenomic analyses on all 134 transporter families. The phylogenetic trees of all these families are available from the TransportDB website (**http://www.membranetransport.org/**). Such approaches will help to reconstruct the evolutionary history of a family, and to investigate instances of gene loss, lateral gene transfer, and gene duplications and expansions. These analyses could also assist in the prediction of the structure, mechanism, and function of unknown transporters.

In summary, a comprehensive classification system for transporter proteins (TCDB), and a complete transporter system comparison database for sequenced genome (TransportDB) have been built up, aimed to include all membrane transport systems found in all living organisms. The advantage of the incorporation of phylogeny into the classification system and the importance of a phylogenetic

approach are underlined by the fact that phylogeny provides reliable guide to structure, function, and mechanism of transporter proteins, and supplies valuable information on the evolutionary history of a family.

Related articles

Article 86, **Pfam: the protein families database**, Volume 6; Article 38, **Transmembrane topology prediction**, Volume 7; Article 40, **The domains of life and their evolutionary implications**, Volume 7; Article 65, **Analysis and prediction of membrane protein structure**, Volume 7

References

Abele R and Tampe R (1999) Function of the transport complex TAP in cellular immune recognition. *Biochimica et Biophysica Acta*, **1461**, 405–419.

Allikmets R (2000) Simple and complex ABCR: genetic predisposition to retinal disease. *American Journal of Human Genetics*, **67**, 793–799.

Ambudkar SV, Kimchi-Sarfaty C, Sauna ZE and Gottesman MM (2003) P-glycoprotein: from genomics to mechanism. *Oncogene*, **22**, 7468–7485.

Anjard C and Loomis WF (2002) Evolutionary analyses of ABC transporters of Dictyostelium discoideum. *Eukaryotic Cell*, **1**, 643–652.

Annilo T and Dean M (2004) Degeneration of an ATP-binding cassette transporter gene, ABCC13, in different mammalian lineages. *Genomics*, **84**, 34–46.

Ashcroft SJ (2000) The beta-cell K(ATP) channel. *The Journal of Membrane Biology*, **176**, 187–206.

Bisbal C, Martinand C, Silhol M, Lebleu B and Salehzada T (1995) Cloning and characterization of a RNAse L inhibitor. A new component of the interferon-regulated 2-5A pathway. *The Journal of Biological Chemistry*, **270**, 13308–13317.

Bolhuis H, Poelarends G, van Veen HW, Poolman B, Driessen AJ and Konings WN (1995) The Lactococcal lmrP gene encodes a proton motive force-dependent drug transporter. *The Journal of Biological Chemistry*, **270**, 26092–26098.

Busch W and Saier MH (2004) The IUBMB-endorsed transporter classification system. *Molecular Biotechnology*, **27**, 253–262.

Chang AB, Lin R, Keith Studley W, Tran CV and Saier MH Jr (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Molecular Membrane Biology*, **21**, 171–181.

Chen H, Rossier C, Lalioti MD, Lynn A, Chakravarti A, Perrin G and Antonarakis SE (1996) Cloning of the cDNA for a human homologue of the Drosophila white gene and mapping to chromosome 21q22.3. *American Journal of Human Genetics*, **59**, 66–75.

Cornillot E, Metenier G, Vivares CP and Dassa E (2002) Comparative analysis of sequences encoding ABC systems in the genome of the microsporidian Encephalitozoon cuniculi. *FEMS Microbiology Letters*, **210**, 39–47.

Dean M and Allikmets R (2001) Complete characterization of the human ABC gene family. *Journal of Bioenergetics and Biomembranes*, **33**, 475–479.

Dean M, Rzhetsky A and Allikmets R (2001) The human ATP-binding cassette (ABC) transporter superfamily. *Genome Research*, **11**, 1156–1166.

Gribble FM and Reimann F (2003) Sulphonylurea action revisited: the post-cloning era. *Diabetologia*, **46**, 875–891.

Hamon Y, Broccardo C, Chambenoit O, Luciani MF, Toti F, Chaslin S, Freyssinet JM, Devaux PF, McNeish J, Marguet D, *et al*. (2000) ABC1 promotes engulfment of apoptotic cells and transbilayer redistribution of phosphatidylserine. *Nature Cell Biology*, **2**, 399–406.

Henderson PJ, Giddens RA and Jones-Mortimer MC (1977) Transport of galactose, glucose and their molecular analogues by Escherichia coli K12. *The Biochemical Journal*, **162**, 309–320.

Hoffman HN and Fredrickson DS (1965) Tangier disease (familial high density lipoprotein deficiency). Clinical and genetic features in two adults. *The American Journal of Medicine*, **39**, 582–593.

Hoffmeyer S, Burk O, von Richter O, Arnold HP, Brockmoller J, Johne A, Cascorbi I, Gerloff T, Roots I, Eichelbaum M, *et al*. (2000) Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3473–3478.

Isomoto S, Kondo C, Yamada M, Matsumoto S, Higashiguchi O, Horio Y, Matsuzawa Y and Kurachi Y (1996) A novel sulfonylurea receptor forms with BIR (Kir6.2) a smooth muscle type ATP-sensitive K+ channel. *The Journal of Biological Chemistry*, **271**, 24321–24324.

Juliano RL and Ling V (1976) A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. *Biochimica et Biophysica Acta*, **455**, 152–162.

Kaback HR, Sahin-Toth M and Weinglass AB (2001) The kamikaze approach to membrane transport. *Nature Reviews. Molecular Cell Biology*, **2**, 610–620.

Kleizen B, Braakman I and de Jonge HR (2000) Regulated trafficking of the CFTR chloride channel. *European Journal of Cell Biology*, **79**, 544–556.

Klucken J, Buchler C, Orso E, Kaminski WE, Porsch-Ozcurumez M, Liebisch G, Kapinsky M, Diederich W, Drobnik W, Dean M, *et al*. (2000) ABCG1 (ABC8), the human homolog of the Drosophila white gene, is a regulator of macrophage cholesterol and phospholipid transport. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 817–822.

Koenekoop RK (2003) The gene for Stargardt disease, ABCA4, is a major retinal gene: a mini-review. *Ophthalmic Genetics*, **24**, 75–80.

Kruh GD and Belinsky MG (2003) The MRP family of drug efflux pumps. *Oncogene*, **22**, 7537–7552.

Lill R and Kispal G (2001) Mitochondrial ABC transporters. *Research in Microbiology*, **152**, 331–340.

Luciani MF and Chimini G (1996) The ATP binding cassette transporter ABC1, is required for the engulfment of corpses generated by apoptotic cell death. *The EMBO Journal*, **15**, 226–235.

Martinoia E, Klein M, Geisler M, Bovet L, Forestier C, Kolukisaoglu U, Muller-Rober B and Schulz B (2002) Multifunctionality of plant ABC transporters–more than just detoxifiers. *Planta*, **214**, 345–355.

Marton MJ, Vazquez de Aldana CR, Qiu H, Chakraburtty K and Hinnebusch AG (1997) Evidence that GCN1 and GCN20, translational regulators of GCN4, function on elongating ribosomes in activation of eIF2alpha kinase GCN2. *Molecular and Cellular Biology*, **17**, 4474–4489.

Meidanis J, Braga MD and Verjovski-Almeida S (2002) Whole-genome analysis of transporters in the plant pathogen Xylella fastidiosa. *Microbiology and Molecular Biology Reviews*, **66**, 272–299.

Mosser J, Douar AM, Sarde CO, Kioschis P, Feil R, Moser H, Poustka AM, Mandel JL and Aubourg P (1993) Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters. *Nature*, **361**, 726–730.

Paulsen IT, Kang KH, Hance M and Ren Q (2004) Genome analysis of membrane transport. In *Microbial Genomes*, Fraser CM *et al*. (Eds.), Humana Press: Totowa, pp. 113–126.

Paulsen IT, Nguyen L, Sliwinski MK, Rabus R and Saier MH Jr (2000) Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *Journal of Molecular Biology*, **301**, 75–100.

Paulsen IT, Sliwinski MK, Nelissen B, Goffeau A and Saier MH Jr (1998a) Unified inventory of established and putative transporters encoded within the complete genome of Saccharomyces cerevisiae. *FEBS Letters*, **430**, 116–125.

Paulsen IT, Sliwinski MK and Saier MH Jr (1998b) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *Journal of Molecular Biology*, **277**, 573–592.

Quinton PM (1999) Physiological basis of cystic fibrosis: a historical perspective. *Physiological Reviews*, **79**, S3–S22.

Rea PA, Li ZS, Lu YP, Drozdowicz YM and Martinoia E (1998) From vacuolar Gs-X pumps to multispecific Abc transporters. *Annual Review of Plant Physiology and Plant Molecular Biology*, **49**, 727–760.

Ren Q, Kang KH and Paulsen IT (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Research* **32** Database issue, D284–D288.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, *et al*. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, **245**, 1066–1073.

Roth CW, Holm I, Graille M, Dehoux P, Rzhetsky A, Wincker P, Weissenbach J and Brey PT (2003) Identification of the Anopheles gambiae ATP-binding cassette transporter superfamily genes. *Molecules and Cells*, **15**, 150–158.

Saier MH Jr (1996) Phylogenetic approaches to the identification and characterization of protein families and superfamilies. *Microbial & Comparative Genomics*, **1**, 129–150.

Saier MH Jr (1999a) *Classification of Transmembrane Transport Systems in Living Organisms*, Academic Press: San Diego, pp. 265–276.

Saier MH Jr (1999b) A functional-phylogenetic system for the classification of transport proteins. *Journal of Cellular Biochemistry*, (Suppl. 32–33), 84–94.

Saier MH Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and Molecular Biology Reviews*, **64**, 354–411.

Saier MH Jr, Eng BH, Fard S, Garg J, Haggerty DA, Hutchinson WJ, Jack DL, Lai EC, Liu HJ, Nusinew DP, *et al*. (1999) Phylogenetic characterization of novel transport protein families revealed by genome analyses. *Biochimica et Biophysica Acta*, **1422**, 1–56.

Sanchez-Fernandez R, Davies TG, Coleman JO and Rea PA (2001) The Arabidopsis thaliana ABC protein superfamily, a complete inventory. *The Journal of Biological Chemistry*, **276**, 30231–30244.

Sweet G, Gandor C, Voegele R, Wittekindt N, Beuerle J, Truniger V, Lin EC and Boos W (1990) Glycerol facilitator of Escherichia coli: cloning of glpF and identification of the glpF product. *Journal of Bacteriology*, **172**, 424–430.

Tennent JM, Lyon BR, Midgley M, Jones IG, Purewal AS and Skurray RA (1989) Physical and biochemical characterization of the qacA gene encoding antiseptic and disinfectant resistance in Staphylococcus aureus. *Journal of General Microbiology*, **135**(Pt 1), 1–10.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.

Young SG and Fielding CJ (1999) The ABCs of cholesterol efflux. *Nature Genetics*, **22**, 316–318.

Young L, Leonhard K, Tatsuta T, Trowsdale J and Langer T (2001) Role of the ABC transporter Mdl1 in peptide export from mitochondria. *Science*, **291**, 2135–2138.

Zhang F, Zhang W, Liu L, Fisher CL, Hui D, Childs S, Dorovini-Zis K and Ling V (2000) Characterization of ABCB9, an ATP binding cassette protein associated with lysosomes. *The Journal of Biological Chemistry*, **275**, 23287–23294.

Specialist Review

Structure comparison and protein structure classifications

Oliver Redfern, Christopher Bennett and Christine Orengo
University College London, London, UK

1. Introduction

Evolution has given rise to families of homologous proteins that appear to be related to a common ancestral gene. As mutations occur and new proteins evolve, amino acids sequences may change dramatically. However, protein structure can remain much more highly conserved and it is possible to identify distant evolutionary relatives by comparing their structures. This chapter outlines some of the concepts behind the methods used to classify known protein structures into families. There are numerous protein structure classifications and we discuss and contrast the implementations of the various structure comparison and classification methodologies used to create them.

Perhaps the most important rationale for creating structural databases is that as structure is so much more highly conserved during evolution than sequence, one can recognize very distant relatives, undetectable by sequence comparison. Although the recent development of very powerful sequence profiles and Hidden Markov Models have improved the recognition of distant homologs, visual inspection of the structures of remote homologs greatly aids validation. Some distant homologs are only detectable by structure comparison methods (see Section 6.2). Furthermore, recognition of individual domains within proteins is much more easily achieved by manual examination of the structure.

All structural classifications are based on the solved protein structures deposited in the Protein Structure Databank (PDB, Berman *et al*., 2002). This was established in 1971 and was originally based at the Brookhaven National Laboratories, USA. Since 1998, the PDB has been maintained by the Research Collaboratory of Structural Biology (RCSB), **http://www.rcsb.org/pdb/**. There is also a European node, the Macromolecular Structure Database (MSD), based at the European Bioinformatics Institute in the United Kingdom, **http://www.ebi.ac.uk/msd/**. As of July 2004, the PDB contains ~25 000 entries and currently around 500 new structures are deposited monthly. This number is expected to increase as the structural genomics initiatives become more productive.

Table 1 Table of the most widely used protein structure family database and descriptions of the methods used to create them

Database	Location and author	Coverage (July, 2004)	Structure comparison method	Type	Description	URL
CAMPASS	Cambridge University, UK, Sowdhamini	7580 domains in 1409 superfamilies	COMPARER (Sali and Blundell, 1990), SEA (Rufino and Blundell, 1994)	Structure-based sequence alignments of SCOP superfamilies.	CAMbridge database of Protein Alignment organized as Structural Superfamilies. Provides sequence alignments of structural domains within a superfamily.	**http://www-cryst.bioc.cam.ac.uk/~campass/**
CATH Gene3D	UCL, London, UK, Orengo	58 000 domains in 1459 superfamilies	SSAP (Taylor and Orengo, 1989), GRATH (Harrison *et al.*, 2002)	Automatic structural and sequence comparison methods are combined with manual validation of superfamily alignments and domain boundaries.	CATH is a hierarchical classification of protein domains structures, clustered by Class, Architecture, Topology and Homologous Superfamily.	**http://www.biochem.ucl.ac.uk/bsm/cath/**
CE	SDSC, La Jolla, CA, USA, Bourne	All chains in PDB	CE (Shindyalov and Bourne, 1998)	Fully automatic. Nearest neighbors.	Combinatorial Extension of the optimal path. A database of structural alignments and similarities between all structures in the PDB.	**http://cl.sdsc.edu/ce.html**

DHS	UCL, London, UK	1459 superfamilies in CATH	SSAP (Taylor and Orengo, 1989) CORA (Orengo, 1999)	Fully automatic multiple structure alignments of close relatives in CATH superfamilies.	Dictionary of Homologous Superfamilies. Multiple structure alignments of homologous domains as defined by superfamilies in the CATH database. These are further annotated with functional information from UniProt, ENZYME, GO, KEGG.	**http://www.biochem.ucl.ac.uk/bsm/dhs/**
ENTREZ/MMDB	NCBI, Bethesda, MD, USA, Bryant	All in PDB	VAST (Madej *et al.*, 1995)	Fully automatic. Nearest neighbors.	MMDB contains precalculated pairwise structural comparisons and alignment between all structures in the PDB.	**http://www.ncbi.nih.gov/Structure/MMDB/mmdb.shtml**
HOMSTRAD	Cambridge University, UK, Blundell	7500 domains in over 1400 superfamilies	COMPARER (Sali and Blundell, 1990)	Manual classification of close protein homologs.	HOMologous STRucture Alignment Database, Database of annotated structural alignments for homologous protein families, utilizing SCOP, Pfam, and SMART to identify relatives.	**http://www-cryst.bioc.cam.ac.uk/~homstrad/**
SCOP SUPERFAMILY	LMB-MRC, Cambridge, UK, Murzin	54 745 domains in 1294 superfamilies	Manual	Manual classification.	Structural Classification Of Proteins. Hierarchical classification by Class, Fold, Superfamily, Family.	**http://scop.mrc-lmb.cam.ac.uk/scop/**

The first structural classifications became available in the early 1990s (Swindells *et al*., 1998) when enough structures were available in the PDB to organize them into evolutionary families and fold groups. Table 1 lists the most comprehensive structural family databases and structural neighborhood resources accessible over the Web, together with a summary of any distinctive features used in generating, validating, and curating them. This chapter will focus on generic protocols that are common to all the major classifications and then highlight individual resources where approaches have been adopted that differ to the common protocols.

2. The hierarchy of domain structure classifications

Most databases adopt a hierarchical classification of protein domain structures (see Figure 1). The top level is the class of the protein, which is assigned depending on the percentage of residues that adopt α-helical or β-strand conformations. There are generally three types of class, mainly-α, mainly-β, or a mixture of the two, $\alpha-\beta$. In some databases (e.g., Structural Classification of Proteins, SCOP), the $\alpha-\beta$

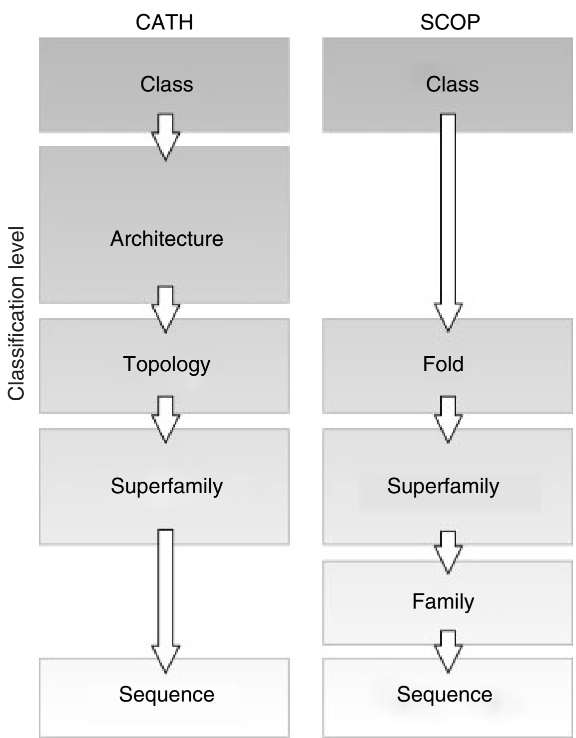


Figure 1 Schematic comparison of the SCOP and CATH hierarchical structure classifications. SCOP recognizes four major classes (mainly α, β, α/β, and $\alpha+\beta$), whereas CATH combines the 2 $\alpha-\beta$ classes. The fold level of SCOP is equivalent to the topology level in CATH; however, SCOP then groups homologs more closely into families, while CATH groups more distant relatives into superfamilies

class is further divided into alternating α/β and $\alpha+\beta$. Within each class, proteins are subclustered according to their fold. The fold describes the relative orientations of the secondary structures and their connectivity; that is, the order in which they appear along the polypeptide chain.

In the CATH database, an additional architecture level is assigned between class and fold, depending on the relative orientations of the secondary structures. Architecture effectively describes the overall shape adopted by the protein domain. For example, within the $\alpha-\beta$ class, two very popular architectures are the $\alpha-\beta$ barrel and the 3-layer $\alpha-\beta$ sandwich. Within each fold group, domains sharing a common evolutionary ancestor are then grouped into homologous superfamilies. SCOP divides these into families. Some resources further subcluster automatically within each superfamily to group domains according to significant levels of sequence similarity (e.g., 30, 60, 95%). Figure 1 compares the CATH and SCOP classifications.

3. Methods used to classify structures

There are a number of automated and semi-automated stages in the classification of structures. First, since most structural classifications are based on domain similarities and since nearly 40% of all known structures in the PDB comprise multidomain protein chains, it is necessary to divide the protein chain into its constituent domain regions. Next, homologous domains must be identified. For close relatives, this can be achieved using standard sequence alignment methods and 1D-profiles (see below).

More distantly related domains can often only be recognized by employing sensitive structure comparison methods. Those structures sharing other features besides structural similarity, for example, rare sequence motifs or functional properties, are likely to have diverged from a common ancestor and can be classified into the same homologous superfamily. Similarly, those sharing only structural similarity are clustered into fold groups. It is possible to reliably assign class using automated methods but architectures are usually described following visual inspection.

The concepts behind the algorithms used in the different stages of classification are outlined in more detail below.

4. Domain boundary assignment

Most multidomain proteins comprise two domains although up to eight domains are observed in some chains. In the majority of cases, each domain is associated with a compact globular arrangement of secondary structures comprising a single fold. Although there are some cases of folds being formed by two or more domain regions from different proteins, these are rare and are sometimes anomalies arising from the manner in which the proteins have been divided for structure determination.

It is likely that the proportion of multidomain structures solved will increase as the structural genomics initiatives progress. Analyses of sequences from completed

genomes by Teichmann and co-workers (Apic *et al*., 2001) have suggested that up to 60% of sequences in prokaryotes are multidomain. This proportion rises to nearly 80% in eukaryotes.

A significant minority of domains (>20%, Pearl *et al*., 2002) occurring in multidomain contexts are discontiguous in sequence, that is, they comprise separate regions of the polypeptide chain and probably arise from domain insertions during evolution. This can considerably complicate the automatic detection of domain boundaries and is particularly problematic for sequence-based methods. More accurate boundary placement can be achieved using the structural data since visual inspection allows boundary refinement in difficult cases.

There have been several successful approaches to recognizing domain boundaries automatically from structure (see Pearl and Orengo, 2003). In most multidomain proteins, the majority of residue contacts occur in the core of a domain, with relatively few contacts between domains. A number of algorithms (DOMAK, PUU) explore alternative divisions of the protein chain that optimize the number of contacts within domains and minimize contacts between domains. Since internal contacts frequently involve compactly packed hydrophobic residues, some methods (e.g., DETECTIVE, Swindells, 1995) attempt to identify domains by locating hydrophobic clusters, which are then expanded until the packing density falls.

More sophisticated approaches developed recently, such as STRDL (Wernisch *et al*., 1999) and DOMS (Taylor, 1999), use different methods but achieve similar success rates. STRDL (STRuctural Domain Limits) uses the Kernigham–Lin graph heuristic to partition a protein chain into sets of residues between which there are minimal interactions – deduced from a weighted Voronoi diagram. Putative partitions are scored on the basis of how well their physical properties fit those of known structural domains.

Most of the algorithms perform well for simple multidomain structures in which the separate domains have minimal contact. However, for more complex chains in which domains are more intimately connected or discontiguous, the performance is less notable. Despite recent improvements, most methods only correctly distinguish proteins as single or multidomain in 80–90% of cases, when tested using manually validated datasets from SCOP and CATH (Jones *et al*., 1998). The success rate falls to 70–80% when assessing the fidelity of residue ranges of assigned domains. Some domains belonging to the most highly populated families (e.g., TIM barrels) are split by many of the methods. A consensus approach (DBS, Jones *et al*., 1998) developed for the CATH database, which sought consensus between three independent methods (DOMAK, PUU, Detective) found that these methods only agreed within reasonable limits 55% of the time.

In particular, automatic methods for discerning domain boundaries are hindered by the lack of any clear quantitative definition of a domain. Domains were first suggested to be independent folding units by Richardson (1981). Subsequently, various qualitative definitions exist, for example, the following criteria are often used in assessing domain regions within a structure:

- The domain possesses a compact globular structure.
- Residues within a domain make more internal contacts than residues in the rest of the polypeptide.

- Secondary structure elements are usually not shared with other regions of the polypeptide.
- There is evidence for existence of this region as an evolutionary unit. For example, in other structural contexts (i.e., different multidomain structures) or as a single domain.

This latter concept of recurrence is used strictly in generating the SCOP classification and domains are identified manually by visually inspecting the structures. In the DALI Domain Database (DDD), an automatic algorithm (PUU) is used to identify putative domains, which are then validated by searching for recurrences in other structures.

Other resources (e.g., CATH, ENTREZ, RCSB) also exploit the fact that the majority of domains recur in different contexts and therefore use structure comparison techniques to locate fold matches within multidomain proteins using a nonredundant library of preclassified folds. These techniques are becoming increasingly successful as the fold libraries expand. Nearly 95% of newly determined domain structures are related to structures already classified in CATH. This proportion is likely to increase as structural genomics initiatives progress as they are specifically targeting novel folds, thereby increasing the fold repertoire (this is also discussed below).

Although no method achieves 100% accuracy in locating precise domain boundaries, the automated approaches significantly assist in locating domains within newly determined protein structures and thereby reduce the amount of manual inspection that is required. Putative boundary definitions can be refined by manual inspection using graphical representations. Programs such as RASMOL (**http://www.openrasmol.org/**), Protein Explorer (**http://www.umass.edu/microbio/rasmol/**) and PyMol (**http://pymol.sourceforge.net/**) are freely available and provide excellent 3D representations of protein structures that can be easily rotated.

5. Identifying close relatives by sequence-based methods

Currently, at least 90% of structures in the PDB are nearly identical (>90% sequence identity) to another structure. This redundancy results from interest in solving relatives with mutated residues in order to determine the impact of these mutations on the stability and functions of the proteins. For example, there are around 844 mutant lysozyme structures in the PDB. In some families, there are also identical or near-identical structures solved with different ligands bound.

Figure 2 shows the distribution of pairwise sequence identities between all nonredundant homologs from the CATH database (release 2.51, 2004). Relatives within a family tend to be highly similar or quite remote (<30% sequence identity) as structural biologists have frequently targeted proteins that appear to have no homologs in the PDB and may thus possess novel folds.

However, as the sequence and structure databases grow and sequence comparison methods become increasingly sophisticated, a significant proportion of newly determined structures each year can be putatively assigned to a structural family

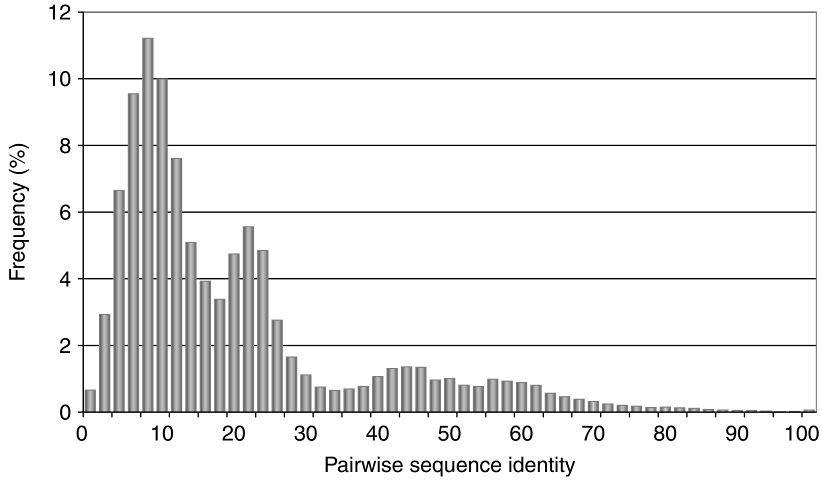


Figure 2 Graph showing the frequency of sequence identities between pairs of homologous domains. It can be seen that the majority of homologs in the PDB show less than 25% sequence identity

based purely on sequence information (discussed in more detail below). This is important as structure comparison can be more than an order of magnitude slower than sequence comparison and in order to keep pace with all the new structures being deposited, it is essential to use rapid sequence methods as a filter for identifying clear homologs.

Sequence-based methods for homolog recognition are well established and are described in some detail in other chapters of this volume. Most resources exploit fast-search algorithms like BLAST or FASTA to identify a putative family to which a new structure may be assigned. Subsequently, slower comparisons using various implementations of the Smith–Waterman or Needleman–Wunsch algorithms are run against all nonredundant structures in the family to obtain more accurate alignments for sequence subclustering within superfamilies.

However, in using these approaches, it is important to check the degree of overlap between any query and matched structure as there may be additional unmatched domains in one or other protein if they are multidomain structures. Details on thresholds used by the different groups in clustering homologs vary for different resources but require overlap values of 60% for remote homologs and up to 80% for close homologs.

For remote homologs the recent generation of 1D-profile–based methods such as PSI-BLAST (Altschul *et al*., 1997) and IMPALA (see Lee *et al*., 2003 for a review) are often used. These use multiple relatives from a family to derive a sequence profile that encodes patterns of highly conserved residues that may be important for stability and function. There are many pattern or profile-based techniques that have been developed to recognize relatives of sequence-based families (e.g., InterPro, SMART, PRINTS, Pfam) and readers are directed to chapters? for further details.

Hidden Markov models, as implemented in the SAM-T9x technology of the Karplus group (Karplus and Hu, 2001) have recently been shown to be even more

powerful than many of the 1D-profile methods at recognizing remote homologs. They also perform better in the recognition of discontiguous domains. Chothia and co-workers have pioneered the use of structural databases (e.g., SCOP) to assess the percentage of remote homologs that can be recognized using these approaches. Similar results have been obtained by benchmarking the methods using the CATH database (Salamov *et al*., 1999; Pearl *et al*., 2000). In a recent analysis using manually validated homologs from the CATH database (release 2.5.1), at least 80% of these remote homologs could be recognized with a low error rate (<0.1%).

Use of large intermediate sequence libraries can also help in pushing the limits of homolog detection. To create these libraries, several structural classifications have recruited hundreds of thousands of additional sequence relatives from the genomes and large sequence databases (e.g., UniProt, Apweiler *et al*., 2004), applying the techniques described above, with conservative thresholds. The SUPERFAMILY resource established in 2001 by the SCOP group expands the SCOP database (Gough *et al*., 2001). They have assigned 1232 superfamilies to predicted proteins from 154 completely sequenced genomes, the UniProt database, and other sequence collections. Close to 60% of all proteins have at least one match.

The Gene3D resource provides an equivalent library of 854 000 sequence relatives for CATH families. This gives 27 000 additional sequence families (clustered at 35% sequence identity). By building Hidden Markov models from representatives of each sequence family, it is possible to increase the percentage of remote CATH homologs recognized to 85% (Sillitoe *et al*., 2004). As the sequence databases expand, these methods will clearly become even more sensitive.

Importantly, a large proportion of newly determined protein structures can now be identified using sequence-based methods, both pairwise and profile based. Over the last few years, nearly 80% of new structures with relatives already classified in the structure databases, can be recognized in this way. However, structures solved by the structural genomics consortia have often been targeted by excluding very remote homologs, using the profile-based methods. Therefore, a smaller proportion of these structures can be recognized in this way and structure comparison methods must be used.

6. Structure-based methods for recognizing remote homologs and analogs

As two proteins diverge from a common ancestor, their sequences can change beyond recognition. However, their three-dimensional structures usually remain similar. This was originally demonstrated by Chothia and Lesk (1986) who plotted sequence similarity against structural similarity for homologs in the PDB. A more recent analysis (Orengo *et al*., 2001) of several hundred well-populated superfamilies in the CATH database, containing three or more sequence families, confirmed that even in very remote relatives (<20% sequence identity), at least 50% of the structure remains conserved. The most highly conserved positions usually correspond to residues in secondary structures, in the buried core of the protein.

Structural comparison methods were introduced in the 1970s, shortly after the advent of the Protein Databank (PDB). Although they can be used to superpose

entire multidomain chains, it is often useful to separate proteins into their constituent domains, as the connectivity and orientation of domains can vary widely and this can have negative effects on the quality of the structural alignment.

There are well over 50 different structure comparison algorithms cited to date but many are variations on a number of techniques. In general, they perform the alignment in two stages. Some measure of similarity of residues and/or secondary structure features between both proteins is calculated and a subsequent optimization strategy is employed to find the alignment that maximizes these similarities. The majority of methods use geometric properties of the C_α or C_β atoms and/or secondary structure information, such as distances or intramolecular vectors. Physicochemical properties, such as hydrophobicity, hydrogen bonding and disulfide bonding, are sometimes utilized. However, they can be of limited use and create unnecessary noise in the alignment.

6.1. *Calculating structural similarity*

Irrespective of the method used to align or superpose two protein structures, some quantitative measure of similarity is needed. The most widely used of these is the Root Mean Square Deviation (RMSD). This is simply the square root of the average squared distance between equivalent atoms as defined by

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} {d_i}^2}{N}}$$

Similar proteins folds tend to give an RMSD below 4.0; however, this can be higher for distant relatives with more than 400 residues. This highlights the main problem with using RMSD as a measure of similarity; namely, that it is dependent on the size of the proteins being aligned (in the equation above). It is therefore important to consider both the RMSD and the number of equivalent residue pairs when assessing the significance of the similarity. Despite its limitations, RMSD remains a widely used and valuable measure. As with sequence-searching tools, most structural comparison methods have developed more rigorous statistical methods for database searching, discussed below.

6.2. *Structure comparison algorithms*

There are now many structure alignment algorithms (see Sillitoe and Orengo, 2003, for a review), in this chapter, we review the main concepts involved and discuss in more detail those most frequently used to identify homologs in the structural classifications listed in Table 1.

6.2.1. Rigid body superposition methods

It is possible to treat two protein structures as rigid objects and think of superposition as finding the best way of fitting one on top of the other, so there is

a minimum difference between them. It should be noted that this is distinct from structural alignment methods, which map equivalent residues between two proteins. This was the rational of the first methods pioneered by Rossman and Argos in the 1970s. Rigid body superposition can be thought of in the following three stages:

1. translating structures to a common position in the coordinate frame, usually by translating their center of mass to the origin;
2. finding putative equivalent positions to start the optimization;
3. rotating one protein, relative to the other, around to three major axes to look for the "best fit".

The major difficulty with this method lies in identifying putative equivalent positions to begin the optimization. For close relatives (>35% sequence identity), standard sequence alignment methods can be used. However, for more distantly related proteins, this is unreliable and the algorithm often requires manual inspection to define known equivalent residues, such as catalytic residues in the active site.

Therefore, rigid body superposition is generally only used to compare closely related proteins or to superpose structures once alternative algorithms, able to handle extensive insertions and deletions, have determined through structural alignment equivalent positions (see below).

6.2.2. Secondary structure-based methods

One approach to handling insertions/deletions (indels) in distant homologs is to simply compare the secondary structures as most indels occur in the loops connecting secondary structures. Graph theoretical methods tend to dominate this approach to secondary structure comparison, as they are fast and effective. Most concentrate on the distances and angles between secondary structures in both proteins, which are compared to find equivalent pairs (see Figure 3).

6.2.2.1. GRATH, SSM

Graph theory is a significant branch of mathematics that has been applied to many different areas of biology and computer science. A graph consists of points, or *nodes*, in two-dimensional space connected by lines, or *edges*, that describe the relationship between them. A protein structure can be reduced to a graph where the nodes are secondary structures and the edges describe the geometric relationships between them (e.g., distances, angles). Artymiuk and co-workers were the first to use these techniques in 1993, although Harrison *et al*. (2002) have applied them more recently to detect fold similarities as part of classification in the CATH database (GRATH).

Linear vectors are used to represent the secondary structures and then edges are labeled with distances between the midpoints and angles, describing the tilt and rotation between the vectors. The resulting two protein graphs are then evaluated to look for common secondary structure "cliques or subgraphs", by identifying equivalent edges that are labeled with similar distances and angles (see Figure 3).

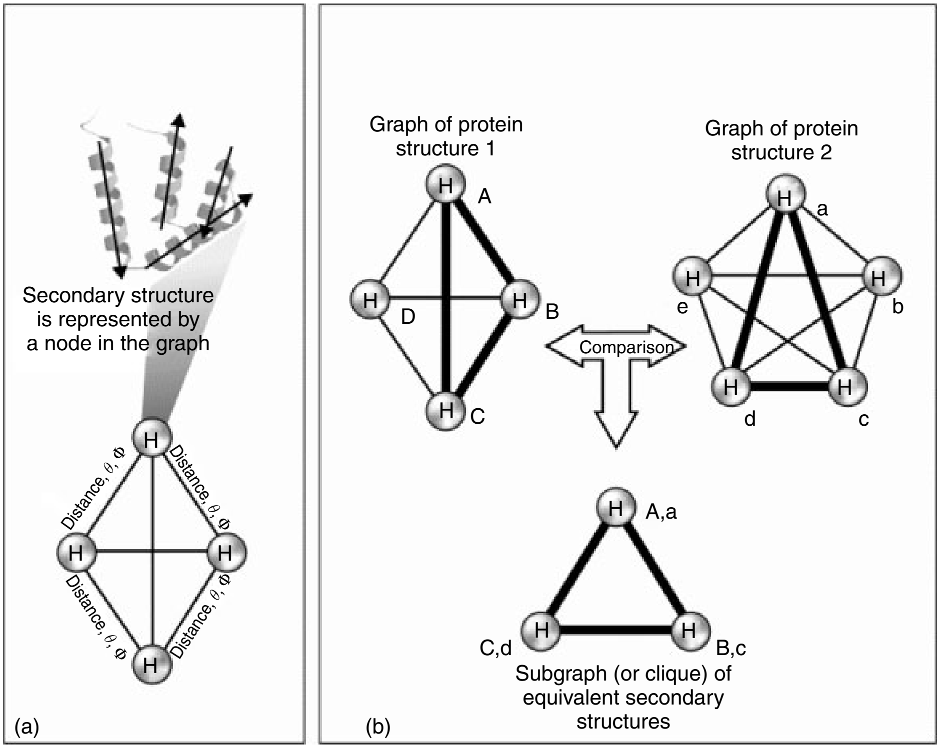


Figure 3 Illustration of graph theory–based structure comparison algorithms. (a) Linear vectors are calculated through each secondary structure and used to represent each node in a graph. The relationships between these vectors (e.g., angles and midpoint distances) then annotate the edges between them. (b) Two protein graphs are compared by looking for equivalent edges (highlighted in bold). Whereas SSM looks only for common subgraphs, GRATH looks for fully connected cliques. The resulting secondary structure graphs can represent a common topology shared by the two protein domains

At the European Bioinformatics Institute (EBI) in Cambridge, UK, Krissinel and co-workers (Krissinel and Henrick, 2004) have optimized a subgraph matching algorithm, on which they base their SSM method. Much like GRATH, it labels edges with distances and angles to determine equivalent relationships. However, a greater emphasis is placed on the similarity between the sizes of secondary structures. The major difference is that SSM does not search for fully connected cliques. This is compensated for by also looking for equivalent connectivity, that is, matched secondary structures must be in the same order along the protein chain.

These methods are extremely fast (particularly with <20 secondary structures) at searching databases of folds and very effective at identifying distant fold similarities. They are often used to find putative structural relatives that can then be aligned more accurately to the query structure using residue-based methods.

6.2.2.2. VAST

Entrez at the NCBI provides a Web resource of structural alignments and superpositions of around 10 000 domain substructures within the PDB using the

VAST (Vector Alignment Search Tool) algorithm (Madej *et al*., 1995). In a similar way to graph theory methods, VAST focuses on the relationship between secondary structures. They define "units" of tertiary structure similarity as pairs of secondary structure elements that share a similar type, relative orientation, and connectivity. Two protein domains are aligned to give the optimal superposition score across all pairs of secondary structure elements (SSEs). A statistical method is then employed to measure the likelihood that this similarity would be seen by chance, by calculating the probability that the score would be obtained when random pairs of SSE combinations between the two domains were superposed. The method yields a significance score that appears to be highly discriminatory at identifying structural relatives.

6.2.3. Residue distance and contact map–based methods

Some of the earliest structural comparison methods were based on distance plots. These are 2D matrices that are shaded according to the distances between residues in a protein. In a similar vein, contact maps, which record those residues that are in contact (within a threshold distance ∼8Å), can also be generated. These contacts can be based on C_α atoms only or other atoms in the residue side chains. The patterns that arise in the resulting matrix are often characteristic of a particular protein fold. For example, dense stretches of contacts indicate closely packed secondary structures. Protein structures can be aligned by overlaying their contact maps. However, as with rigid body methods, it is difficult to overlay the maps of distant homologs and various strategies have evolved to cope with extensive insertions/deletions.

6.2.3.1. DALI and CE

One approach to aligning distant structural relatives is to divide each protein into fragments. CE and DALI are popular examples of this method that involves finding equivalent fragments and then combining them to form an overall alignment, using some form of optimization strategy.

Holm and Sander developed the DALI algorithm (see Figure 4), which fragments the proteins into hexapeptides and then uses contact maps to compare them. Potentially equivalent fragments are identified by looking for similar patterns of distances between residues, within a specific threshold. These pairs are then concatenated to extend the alignment using a Monte Carlo optimization. RMSD is used to assess the quality of the extension as the concatenation progresses.

CE fragments the polypeptide chain into octapeptides and aligns residues on the basis of characteristics of their local geometry (as defined by vectors between C_α positions). Matching fragments are termed *aligned fragment pairs* (*AFPs*). Heuristics are used to define a set of optimal paths joining AFPs with gaps as needed. The path with the best RMSD is subjected to dynamic programming to achieve an optimal alignment. For specific families of diverse proteins, additional characteristics are used to weight the alignment.

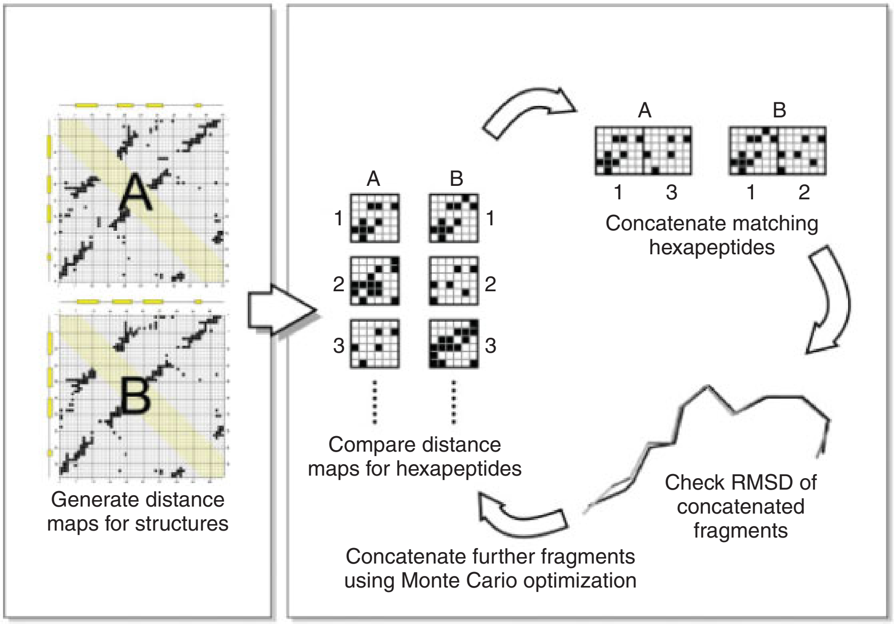


Figure 4 The DALI method of Holm and Sander. Proteins are fragmented into hexapeptides and their contact maps compared to find equivalent fragments. Fragments are concatenated and their RMSD checked to find valid extensions. Monte Carlo optimization is used to guide the extension process to a full alignment

6.2.3.2. SSAP

Another approach, based on comparing distances between residues, was developed by Taylor and Orengo (1989). They sought to deal with the structural embellishments observed between distant relatives by applying the dynamic programming techniques used in sequence alignment methods. In the SSAP algorithm, dynamic programming is in fact utilized twice: first, to compare residue views and second, to determine the equivalent residues.

At the heart of the comparison lies the concept of "residues views" (see Figure 5). These are simply vectors calculated between all C_β (side chain carbon) atoms and a specific residue within a structure. These vectors are compared between the two proteins by scoring for similarity. The number of potentially equivalent residues is also limited by selecting on physicochemical and secondary structure properties (e.g., accessibility, phi and psi angles). A "residue-level score matrix" is constructed for each pair of putatively equivalent residues with scores that reflect the similarity of a given pair of vectors. For example, vectors from residue (i) to other residues in protein A are compared to vectors from residue (j) to other residues. Dynamic programming is used to find the optimal high-scoring path through the matrix.

The second step is to amalgamate the information from the residue-level matrices into a summary score matrix. Pairs of residues are selected as being potentially equivalent on the basis of the score of the best path through their residue-level matrix. These optimal paths are collated in the summary matrix and an overall optimal path is calculated using dynamic programming. The SSAP algorithm is

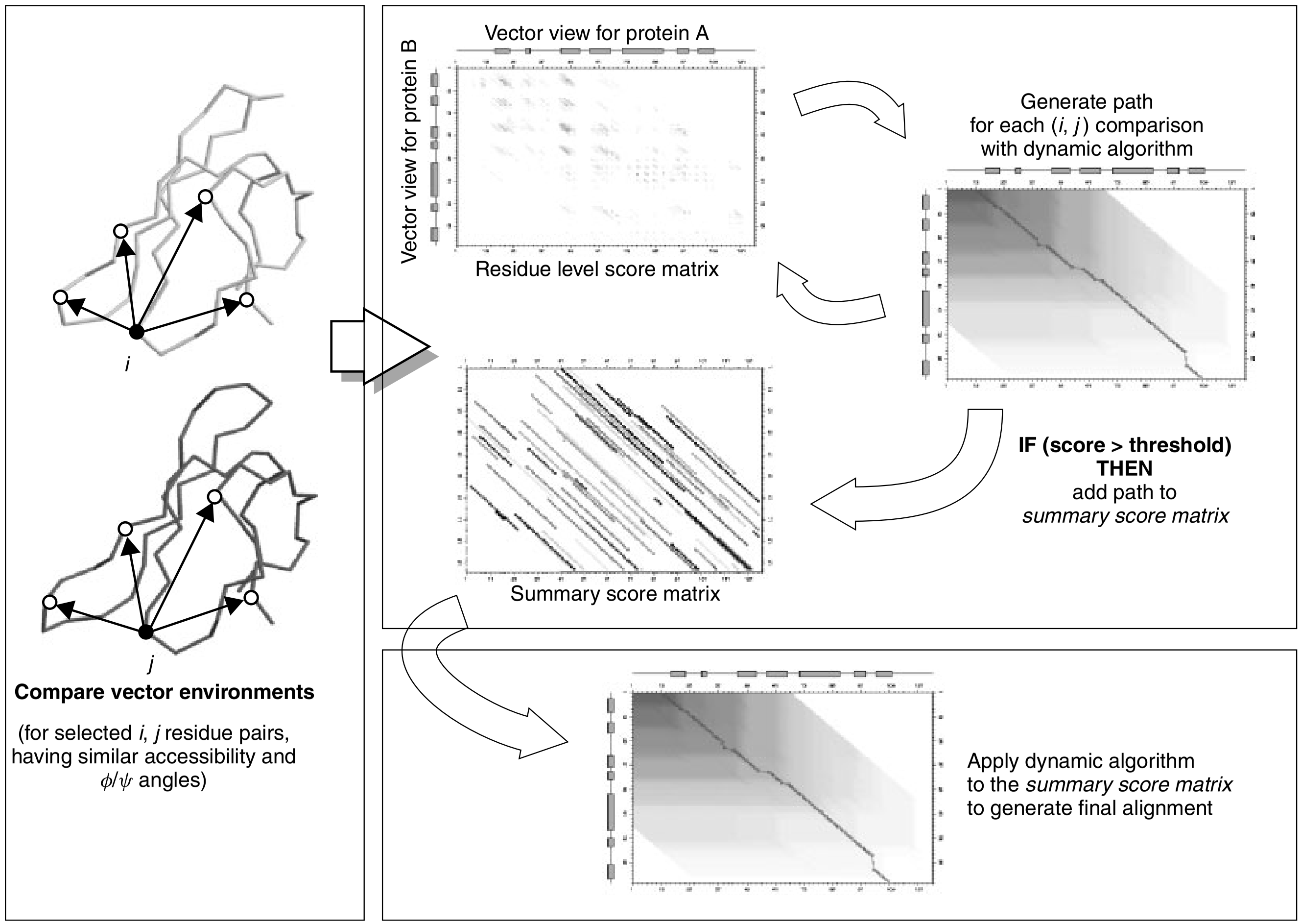


Figure 5 Flowchart of the SSAP algorithm. Vector environments are compared between pairs of potentially equivalent residues in each protein. A residue-level score matrix is constructed for each pair and optimal paths (putative alignments) are calculated by dynamic programming. High-scoring paths are then added to the summary score matrix. Dynamic programming is then applied to the summary matrix to generate the final optimal alignment of the two structures

used to recognize remote homologs for classification in the CATH database (see Table 1).

6.2.3.3. COMPARER

Sali and Blundell (1990) developed COMPARER by using intermolecular superposition and then subsequently assessing relationships between residues within each structure. Properties of residues, such as secondary structure type, side-chain orientations and torsional angles, are then compared between proteins and used to populate a 2D matrix. These are combined with intramolecular information (C_α distances, hydrogen bonding patterns, distances to the protein's center of mass) to find equivalent residues. Putative equivalences are optimized by rigid body superposition followed by a technique known as simulated annealing. This employs the probabilistic Boltzmann energy function, which calculates energy change as temperature decreases to find optimal solutions to superposition of the proteins. The final alignment is then optimized using dynamic programming.

6.3. Assessing structural similarity

Earlier we mentioned how RMSD can be used to assess how closely two structures have been superposed. However, it was highlighted that RMSD is length dependent and therefore the embellishments observed between large structural relatives cause the RMSD to be unrepresentatively high. Moreover, proteins in different folds use a common repertoire of super-secondary structure motifs (e.g., β-hairpins, $\alpha-\beta$ motifs) that can create local similarities that may bias the RMSD.

Workers in the field of structural comparison have taken similar approaches to those used in scanning large sequence databases, by employing statistical methods to provide a more representative score. A commonly used measure is the Z-score. Given a particular structural similarity between putative relatives, the Z-score gives a measure of the number of standard deviations a particular score is from the mean. DALI and CE both use this approach to rank their database scans.

Although Z-scores are often helpful, they are at their most powerful when data follows a normal distribution. However, most database scans follow an extreme value distribution where the highest scores occur at low frequencies. By taking a protein and performing multiple scans of a database, an exponential function can be "fitted" to the latter part of the distribution. This function can then be used to convert each score into a probability. This probability represents the likelihood that the score could have been given by unrelated proteins. By multiplying this probability by the number of structures in the database, it is easy to calculate how many unrelated proteins you would expect to "hit" with a specific score (the expectation or E-value). A low E-value (<1) is a good indication that the query structure has matched a significant structural relative. The GRATH algorithm uses E-values to detect fold similarities for classification in the CATH database.

6.4. Multiple structure comparison

As with using multiple sequence analysis to identify conserved amino acids within a protein family, multiple structure alignments can reveal conserved structural features that may be important for function or fold stability. Within structural superfamilies (see later), relatives tend to exhibit embellishments around a common conserved "core" (see Sillitoe *et al*., 2001 for a review). Capturing that core in a structural fingerprint can provide a more sensitive tool for identifying relatives of that family.

Several of the pairwise algorithms STAMP (Russell and Barton, 1992; COMPARER (Sali and Blundell, 1990)) have been adapted to align sets of structures. The most common strategy is to perform a progressive alignment where the multiple alignments are built from a series of pairwise alignments. Pairwise similarities are first calculated and subjected to hierarchical single-linkage clustering. The subsequent phylogenetic tree defines the order in which structures are added to the alignment. In order to preserve the quality of the alignment, it is common to take the most similar pair of structures first.

6.5. Databases of multiple structure alignments

The HOMSTRAD and CAMPASS databases (see Table 1) aim to provide multiple structural alignments for protein superfamilies. HOMSTRAD uses SCOP, PFAM, and other resources to look for close structural relatives where there is clear evidence from structural comparison (and often high sequence homology) that the proteins are related. Although individual pairwise sequence identities can be as low as 12% in families such as the globins, all the members can be single-linked with a sequence identity of at least 21%. These multiple alignments can be used to derive substitution matrices to calculate residue-exchange matrices or to encode conserved structural features in a template that can be used to identify further relatives. The CAMPASS database groups more distant structural homologs than HOMSTRAD by using the structural comparison algorithms COMPARER and SEA (Rufino and Blundell, 1994) to generate multiple alignments of SCOP superfamilies. Currently, they have around 7500 domains in over 1400 superfamilies.

Alignments from the HOMSTRAD database have been used to derive substitution matrices describing preferences for residue exchange in different structural environments. Although, to date, such matrices have not performed as well as those derived using sequence-based families (e.g., the BLOCKS database for the BLOSUM matrices), recent updates have performed well and as the structure databank expands with the structural genomics initiatives and more diverse representatives are determined, these approaches will become increasingly powerful.

As an alternative to chaining pairwise alignments, other methods (such as CORA; Orengo, 1999) instead build a "consensus structure" at each stage of the alignment to align subsequent structures. This is thought to better preserve the fidelity of the alignment. In the CATH database, the CORA algorithm is used to provide

structural alignments for subgroups of structures in each superfamily sharing a reasonably high degree of structural similarity as assessed by scores returned from the SSAP algorithm. Structural fingerprints or 3D-templates are also generated for each structural subgroup to help improve recognition of distant homologs. Currently, CATH generates ~700 multiple alignments for 498 well-populated domain superfamilies.

7. Structural classification: assigning domains to different levels in the structure classification hierarchy

As already discussed above, most structural classifications have been established at the domain level, since this is generally thought to be an important unit that can participate in the evolution of new functional proteins. Some domains are highly recurrent across the genomes and interesting insights into evolutionary mechanisms can often be gleaned by understanding how relatives in domain families combine with different domain partners to evolve new functions.

Therefore, for classifying new structures, domain boundaries are identified first. This is achieved either by fold recognition algorithms (e.g., DALI, GRATH, SSM, VAST) or if the protein contains novel domains, using ab initio approaches (e.g., PUU, DETECTIVE) and manual inspection. The major classifications show that the current set of structural entries in the PDB comprise between 50 000–60 000 domains (SCOP – 20 619 entries, 54 745 domains; CATH – 22 303 entries, 58 141 domains).

Subsequently, close relatives are identified using standard pairwise methods (e.g., BLAST, SSEARCH). Some classifications (e.g., CATH) also apply these pairwise methods to preliminary clustering of closely related multidomain structures before selecting structures from each multidomain sequence family identified for domain boundary recognition. Approximately 60% of domain structures deposited in the PDB over the last two years could be recognized using BLAST.

In CATH release 2.5.1, around 4000 families clustered at 35% sequence identity are currently recognized by these techniques. In the current SCOP database, release 1.65, close relatives have been clustered into 2327 families provided they share common functional properties. More remote homologs are then found by applying profile or HMM-based technologies (e.g., PSI-BLAST, SAM-T99). Currently, approximately 15–20% of newly deposited structures are classified using these approaches.

Finally, structure comparison methods are used to recognize very remote homologs and those proteins that share structural similarity but for which there is no other evidence of any evolutionary relationship. These structures, whose similarity may reflect convergence to an energetically or kinetically favored fold, are described as analogous structures and are discussed further below. In order to clearly distinguish homologs from analogs, it is often necessary to perform extensive manual validation of matches suggested by the structure comparison methods unless very highly significant scores are obtained. Usually, this analysis is done by searching the literature for evidence of common functional traits or visually inspecting the structures for rare structural motifs.

Various automatic methods for homolog validation have also been attempted. Russell and co-workers developed a method on the basis of Murzin's approach to classifying very remote homologs in the SCOP database, through identification of rare structural motifs conserved during evolution (Stark *et al*., 2003). Alternatively, Holm and co-workers have designed a neural network method that combines information on structure similarity, sequence similarity, and functional keywords (Dietmann and Holm, 2001). However, none of these approaches are completely reliable and therefore some degree of manual checking is generally performed for very remote homologs.

For the CATH database, once fold group and homologous superfamily have been assigned, architectures are assigned by manual inspection of structures. Subsequently, CATH employs an automated approach for explicitly defining protein class (Michie *et al*., 1996). This method calculates the percentage of residues in α-helical or β-strand conformations. It also takes account of the nature of secondary structure packing. Other resources do not explicitly assign class or, as in the case of SCOP, assign it by manual inspection.

7.1. Databases of structural neighbors

There are a number of on-line resources that provide lists of structural relatives for a given protein domain, rather than explicit classifications. FSSP (EBI), MMDB (NCBI), and the PDB (RCSB) do not generate protein families, but instead produce lists of structural neighbors for a given query sequence. FSSP provides a search tool that exploits the DALI algorithm to find structural relatives. Although a high structural similarity suggests homology, it is up to the user to decide how closely other domains are linked by evolution.

The MMDB exploits the vector-based VAST algorithm to automatically find similar structures within the PDB. It provides alignments annotated with automatic domain assignments and graphical structural superposition. The PDB resource itself makes use of the CE program to search for structural neighbors automatically. Again, it is up to the user to further group these into individual protein families.

Most of these resources use Z-scores to provide a measure of the significance for any structural match. Usually a value above 3 or 4 is a good indication of homology.

8. Analysis of protein structural families

Classifying proteins into families permits a more profound analysis of the mechanisms of evolution, such as identifying structural variations in the active sites of paralogs that have evolved to enable different substrates to bind. Analysis of structural families can also give insight into the biophysics of protein structure, such as secondary structure packing and favored topologies. In most families, we observe a highly conserved structural core, although in some families this is highly embellished to confer different enzymatic functions and protein–protein interaction partners (see a review by Orengo *et al*., 2001 and Todd *et al*., 2001).

It is important to remember that at present, transmembrane proteins are massively underrepresented in the PDB because they are difficult to crystallize and too large to be solved in solution by NMR. These proteins are hugely important in drug metabolism and it remains an experimental challenge to resolve the structure of these proteins. Once solved, they will provide valuable insight into how proteins have evolved differently to exploit the hydrophobic environment of cell membranes as opposed to the aqueous environment inhabited by most globular proteins.

Several groups have postulated that there are actually a limited number of superfamilies in nature, totaling only a few thousand (Chothia, 1992; Coulson and Moult, 2002, see Lee *et al*., 2003 for a review). With this in mind, the recent structural genomics initiatives aim to solve representative structures for all known protein families in the genomes. This will fill an important gap in our knowledge of evolution and the relationship between protein structure and function. Further structural representatives may allow currently distinct families in the sequence-based resources (e.g., Pfam) to be merged into superfamilies.

Several groups have shown that 50–60% of genes in completed genomes can currently be completely or partially assigned to fewer than 1500 structural superfamilies, using sensitive profile-based methods and HMMs (Gough *et al*., 2001, see Lee *et al*., 2003 for a review). This proportion increases to 70% if more powerful threading-based technologies are applied (McGuffin *et al*., 2004). From current analyses of the number of protein families identified in completed genomes (30 000–60 000 (see Lee *et al*., 2003 for a review)) this leaves many families to be structurally characterized. However, many of these families tend to be very small and specific to the organism or kingdom. They may comprise paralogs of common structural families that have acquired new and specific functions and have therefore diverged beyond the limit of sensitivity of sequence-based methods.

Extreme sequence divergence in paralogous genes may also explain the existence of analogous structures in some of the very highly populated fold groups in the structural classifications. For example, in CATH, some 55 fold groups comprise three or more different homologous superfamilies. This trend was first observed nearly a decade ago when the term "superfolds" was coined to characterize these highly recurrent folds (Chothia, 1992; Orengo *et al*., 1994). Currently in CATH, members of these fold groups comprise nearly one-third of all nonredundant (<35% sequence identity) domain structures. Although the diverse superfamilies observed within these fold groups may represent remote paralogs that have acquired new functions thereby removing some of the most important constraints on sequence conservation, they may also represent energetically favored folding arrangements.

It is reasonable to suppose that there are chemical–physical constraints on the ways in which proteins fold to optimize hydrophobic packing of core residues. This may also explain the preponderance of certain types of regular architecture. The most symmetric architectures in the CATH database (α-bundle, β-barrels, $\alpha\beta$-barrels, β-sandwiches and $\alpha\beta$-sandwiches) comprise nearly 60% of all nonredundant domains in CATH. Similarly, Shaknovitch and co-workers have shown that there are some folds in the database that are intrinsically able to support a much more diverse repertoire of sequences (Deeds *et al*., 2003).

Given the fact that we cannot obtain all the intermediate sequences arising during evolution of these popular folds, it is difficult to determine whether these structural

similarities represent divergence from a common ancestor or convergence to an energetically or kinetically favored fold.

Further reading

Siddiqui AS and Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, **4**(5), 872–884.

Wolf YI, Grishin NV and Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, **299**, 897–905.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Apic G, Gough J and Teichmann SA (2001) Domain combinations in Archaeal, Eubacterial and eukaryotic Proteomes. *Journal of Molecular Biology*, **310**, 311–325.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **1**(32), D115–D119.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, *et al*. (2002) The protein data bank. *Acta Crystallographica. Section D, Biological Crystallography*, **58**, 899–907.

Chothia C and Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5**(4), 823–826.

Chothia C (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**(6379), 543–544.

Coulson AF and Moult J (2002) A unifold, mesolfold and superfold model of proteins. *Proteins*, **46**, 61–71.

Deeds EJ, Dokholyan NV and Shakhnovich EI (2003) Protein evolution within a structural space. *Biophys Journal*, **85**(5), 2962–2972.

Dietmann S and Holm L (2001) Identification of homology in proteins structure classification. *Nature Structural Biology*, **11**, 953–957.

Gough J, Karplus K, Hughey R and Chothia C (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of Molecular Biology*, **313**(4), 903–919.

Harrison A, Pearl F, Mott R, Thornton J and Orengo C (2002) Quantifying the similarities within fold space. *Journal of Molecular Biology*, **323**(5), 909–926.

Jones S, Stewart M, Michie A, Swindells MB, Orengo C and Thornton JM (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science*, **7**(2), 233–242.

Karplus K and Hu B (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**(8), 713–720.

Krissinel E and Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica*, **D60**, 2256–2268

Lee D, Grant A, Buchan D and Orengo C, (2003) A structural perspective on genome evolution. *Curr Opin Struct Biol.* **13**(3) 359–69.

Madej T, Gibrat JF and Bryant SH (1995) Threading a database of protein cores. *Proteins* **23**(3), 356–369.

McGuffin LJ, Street S, Sorensen SA and Jones DT (2004) The genomic threading database. *Bioinformatics*, **20**(1), 131–132.

Michie AD, Orengo CA and Thornton JM (1996) Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology*, **262**(2) 168–185.

Orengo CA (1999) CORA: topological fingerprints for protein structural families. *Protein Science*, **8**(4), 699–715.

Orengo CA, Jones DT and Thornton JM (1994) Protein superfamilies and domain superfolds. *Nature*, **372**(6507), 631–634.

Orengo CA, Sillitoe I, Reeves G and Pearl FM (2001) Review: what can structural classifications reveal about protein evolution? *Journal of Structural Biology*, **134**(2-3), 145–165.

Pearl FM, Lee D, Bray JE, Buchan DW, Shepherd AJ and Orengo CA (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Science*, **11**(2), 233–244.

Pearl F and Orengo CA (2003) Protein structure classifications. In: Orengo *et al*. (Eds), *Bioinformatics: Genes, Proteins and Computers*, BIOS.

Pearl F, Todd AE, Bray JE, Martin AC, Salamov AA, Suwa M, Swindells MB, Thornton JM and Orengo CA (2000) Using the CATH domain database to assign structures and functions to the genome sequences. *Biochemical Society Transactions*, **28**(2), 269–275.

Richardson JS (1981) The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, **34**, 167–339.

Rufino SD and Blundell TL (1994) Structure-based identification and clustering of protein families and superfamilies. *Journal of Computer-aided Molecular Design*, **8**(1), 5–27.

Russell RB and Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**(2), 309–323.

Salamov AA, Suwa M, Orengo CA and Swindells MB (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Engineering*, **12**(2), 95–100.

Sali A and Blundell TL (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of Molecular Biology*, **212**, 403–428.

Shindyalov IN and Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, **11**(9), 739–47.

Sillitoe I, Dibley M, Bray J, Addou S and Orengo CA (2004) Assessing strategies for improved superfamily recognition. *Protein Science*, In Press.

Sillitoe I and Orengo CA (2003) Protein structure comparison. In Orengo *et al*. *(Eds)*, *Bioinformatics: Genes, Proteins and Computers* BIOS.

Stark A, Sunyaev S and Russell RB (2003) A model for statistical significance of local similarities in structure. *Journal of Molecular Biology*, **326**(5), 1307–1316.

Swindells MB (1995) A procedure for detecting structural domains in proteins. *Protein Science*, **4**(1), 103–112.

Swindells MB, Orengo CA, Jones DT, Hutchinson EG and Thornton JM (1998) Contemporary approaches to protein structure classification. *BioEssays*, **20**(11), 884–891.

Taylor WR (1999) Protein structural domain identification. *Protein Engineering*, **12**(3), 203–216.

Taylor WR and Orengo CA (1989) Protein structure alignment. *Journal of Molecular Biology*, **208**(1), 1–22.

Todd AE, Orengo CA and Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, **307**(4), 1113–1143.

Wernisch L, Hunting M and Wodak SJ (1999) Identification of structural domains in proteins by a graph heuristic. *Proteins*, **35**(3), 338–352.

Short Specialist Review

InterPro

Nicola J. Mulder and Rolf Apweiler
European Bioinformatics Institute, Cambridge, UK

1. Introduction

With the increasing availability of raw genome sequence data, the need for tools for protein functional annotation by automatic methods has arisen. The gap between the number of characterized and raw protein sequences in the public databases is widening, making it difficult to extract biological value from the vast quantity of new sequences. This has led to the emergence of protein signatures, which provide distinct advantages over the traditional sequence similarity methods of protein functional classification. There are different ways of deriving protein signatures (descriptive entities for a protein family, domain, or functional site), including regular expressions, profiles, or hidden Markov models (HMMs). PROSITE (Hulo *et al*., 2006) uses regular expressions and profiles, PRINTS (Attwood *et al*., 2003) uses an adaptation of profiles called *fingerprints*, and HMMs form the basis for Pfam (Finn *et al*., 2006), SMART (Letunic *et al*., 2006), TIGRFAMs (Haft *et al*., 2003), PIRSF (Wu *et al*., 2004), SUPERFAMILY (Madera *et al*., 2004), Gene3D (Yeats *et al*., 2006), and PANTHER (Mi *et al*., 2005). ProDom (Bru *et al*., 2005), on the other hand, uses sequence clustering to generate protein families. These are the major protein signature databases available, and while each of these is useful in its own right, the databases are far more effective when unified into a single resource. InterPro (Mulder *et al*., 2005) was created using these databases, as the value in uniting the efforts to provide a "one stop shop" for protein signature analysis was recognized.

2. Integration of signatures

All PRINTS fingerprints, PROSITE patterns and profiles, ProDom domains, and Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER HMMs that identify the same protein family or domain within a set of protein sequences are grouped together in a single InterPro entry. The criteria for grouping are that signatures should overlap in position on the sequence and significantly in the proteins matched. Where one signature identifies a subset of proteins found by another signature, these are put into separate InterPro entries and relationships are inserted between these entries. Two types of relationships can exist between

InterPro entries: the parent/child and contains/found in relationships. Parent/child relationships are used to describe a common ancestry between entries, or where one entry represents a subfamily of another, whereas the contains/found in relationship refers to the domain composition.

3. Content of an InterPro entry

Each entry contains one or more signatures from the individual member databases that describe the same group of proteins. All entries contain a list of matches against all proteins in UniProtKB (Wu *et al*., 2006) and annotation regarding the protein family/domain.

3.1. Protein matches

Protein matches are calculated using the InterProScan (Quevillon *et al*., 2005) software package, which is a combination of the search methods of the individual databases. The match lists may be viewed in tabular and different graphical formats. The table lists the protein accession numbers and the positions in the amino acid sequence where each signature from that InterPro entry hits. In the detailed graphical view, the sequence is split into lines for each hit by a unique signature and the bars are color coded by member database. The proteins can also be viewed graphically in an overview, which splits the protein sequence into different lines for each InterPro entry matched. Where structures are available for proteins, there is a link from the graphical view to the corresponding Protein Data Bank (PDB) structures and a separate line in the display showing the SCOP (Andreeva *et al*., 2004) and/or CATH (Orengo *et al*., 2003) curated matches on the sequence as white striped bars. Where the structures have not been solved but can be modeled by MODBASE (Pieper *et al*., 2004) or SWISS-MODEL (Kopp and Schwede, 2006), these are displayed as yellow and white or red and white striped bars, respectively. Protein matches are also now available for different splice variants, which are displayed together with matches for the master sequence.

A new protein match view is the InterPro Domain Architectures (IDAs) view, which summarizes the InterPro matches as oval beads along the sequence. All proteins with the same sequence of InterPro matches are shown once, with the number of proteins sharing this architecture displayed on the left. The IDA view as well as the table and detailed views can be displayed simultaneously for a protein sequence by inserting the UniProtKB protein accession number in the search box and selecting “Find protein matches” (Figure 1). In the protein match views, the user can also select protein sets by the protein or InterPro accession number(s), whether they have solved structures or splice variants, or by taxonomy. Results can be displayed in any of the above formats.

3.2. InterPro annotation

InterPro entries provide annotation in the form of a name, a short name, and an abstract describing the proteins in question. The abstracts provide the bulk of the

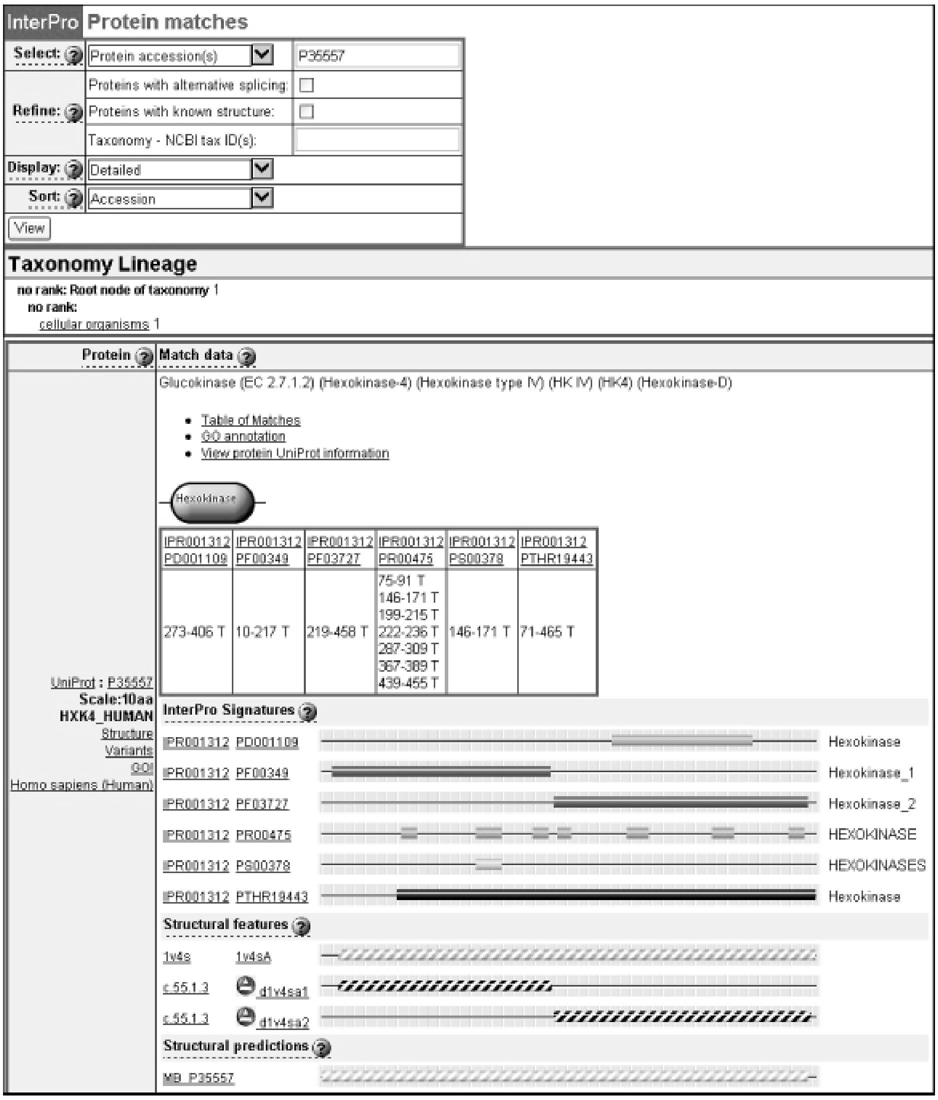


Figure 1 Example of InterPro search results for UniProtKB protein P35557, the human Glucokinase protein. This view shows the matches for this protein in various formats, including the domain architecture, table, and detailed views. The new searching options are shown at the top of the figure, where one can search by InterPro accession, protein accession, or architecture code; filter results by taxonomy ID, proteins with known structure, or splice variants; and display the output in different formats with user-selected ordering

annotation and cite the relevant publications. Additional annotation is in the form of cross-references to specialized databases, such as the MEROPS peptidase database, BLOCKS, and more recently, the IntAct protein–protein interaction database. Links to the PDB, SCOP and CATH structural databases are provided when proteins matching the entry have solved structures. A taxonomy field in each InterPro entry

aims to provide an indication of the taxonomic range of proteins matched by the entry.

Mapping of InterPro entries to GO (Gene Ontology) terms (Harris *et al*., 2004) provides a means for the automatic mapping of the proteins matched to the same GO terms. Entries are mapped to the relevant terms at the most appropriate level in the GO hierarchy when the characterized proteins in the match list share a common molecular function or biological process, or perform their function in the same cellular component (these are the three divisions in GO). The terms are then assumed to apply to all the other proteins matched. These InterPro2GO mappings are the largest providers of GO annotations to proteins.

4. Data access

InterPro is implemented in an Oracle relational database, and is available for text or sequence searches via the web. A simple text search allows the user to query InterPro entries or protein matches, but more complex queries can be performed through sequence retrieval system (SRS) (**http://srs.ebi.ac.uk/**). Query protein or DNA sequences can be submitted to InterProScan interactively or via email. A DNA sequence will be translated in all six frames before submission to InterProScan. If confidentiality or bulk searches are required, users may download a Perl stand-alone InterProScan package for running locally. The results will give an indication of what family the unknown protein belongs to, its domain composition, and potential GO annotations, where available. For more information, the user can read about related proteins or the protein family in the annotation of the InterPro entries it hits.

5. Discussion

Genome sequences of a number of different organisms of different kingdoms of life are currently being sequenced or have already been sequenced. This has a potentially large impact on biologists around the globe working on these various organisms. Instead of seeing the small picture, they are now able to get a bigger picture of where their gene of interest fits into the whole system. In other words, context information, like biological pathways, should be more accessible, since the sequences of all players in the pathways are available. However, this is only possible if those players have been identified and classified. Methods of sequence clustering and protein signatures are now well established and are very useful in the prediction of protein function. Each is different though, so their unification into InterPro and the value-added annotation in InterPro entries provide a powerful resource for sequence analysis and genome annotation. InterPro has already been used for the annotation of a multitude of complete genomes, including that of human (The International Human Genome Consortium, 2001) and mouse (Kawaji *et al*., 2002). It is used for large-scale automatic annotation of TrEMBL, and is continually showing its worth in new applications.

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 000, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 0; Article 86, **Pfam: the protein families database**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6; Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6

References

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family. *Nucleic Acids Research*, **32**(1), D226–D229.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, *et al* (2003) PRINTS and its automatic supplement pre-PRINTS. *Nucleic Acids Research*, **31**(1), 400–402.

Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S and Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research*, **33**, D212–D215.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al* (2006) Pfam: clans, web tools and services. *Nucleic Acids Research*, **34**, D247–D251.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al* Gene Ontology Consortium. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M and Sigrist CJA (2006) The PROSITE database. *Nucleic Acids Research*, **34**, D227–D230.

The International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Kawaji H, Schonbach C, Matsuo Y, Kawai J, Okazaki Y, Hayashizaki Y and Matsuda H (2002) Exploration of novel motifs derived from mouse cDNA sequences. *Genome Research*, **12**(3), 367–378.

Kopp J and Schwede T (2006) The SWISS-MODEL repository: new features and functionalities. *Nucleic Acids Research*, **34**, D315–D318.

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J and Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Research*, **34**, D257–D260.

Madera M, Vogel C, Kummerfeld SK, Chothia C and Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**(1), D235–D239.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, *et al* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284–D288.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, *et al* (2005) InterPro, progress and status in 2005. *Nucleic Acids Research*, **33**, D201–D205.

Orengo CA, Pearl FM and Thornton JM (2003) The CATH domain structure database. *Methods in Biochemical Analysis*, **44**, 249–271.

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis F, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, *et al* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, **32**, D217–D222.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R and Lopez R (2005) InterProScan: protein domains identifier. *Nucleic Acids Research*, **33**, W116–W120.

Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, *et al* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research*, **32**(1), D112–D114.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, *et al* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, **34**, D187–D191.

Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S and Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Research*, **34**, D281–D284.

Short Specialist Review

Functionally and structurally relevant residues in PROSITE motif descriptors

Christian J. A. Sigrist, Edouard De Castro, Petra S. Langendijk-Genevaux, Virginie Le Saux, Amos Bairoch and Nicolas Hulo
University of Geneva, Geneva, Switzerland

1. Introduction

PROSITE is an annotated collection of biologically meaningful motif descriptors dedicated to the identification of protein families and domains (*see* Article 91, **Classification of proteins by sequence signatures**, Volume 6). Its current release 18.43 (January 4, 2004) contains 1311 documentation entries describing 1775 motif descriptors, which are either patterns or profiles (Hulo *et al*., 2004; Sigrist *et al*., 2002). Both are derived from multiple alignments of homologous sequences, giving them the notable advantage of identifying distant relationships between sequences that would have passed unnoticed based solely on pairwise sequence alignment.

PROSITE is also a member database of InterPro (*see* Article 83, **InterPro**, Volume 6) (Mulder *et al*., 2003) and a main contributor to the InterPro annotation. More information on PROSITE is available at **http://www.expasy.org/prosite**.

2. PROSITE patterns

The first motif descriptors used by PROSITE were *regular expressions* or *patterns*. Patterns describe a group of amino acids that constitute a usually short, but characteristic motif that constitutes the signature of a function, a posttranslational modification, a domain, or a protein family. These motifs arise because specific region(s) of a protein, which are important, for example, for their binding properties or their enzymatic activity, are conserved in both structure and sequence. In a pattern, the motif is reduced to a consensus expression in which all but the most significant residue information is discarded. PROSITE patterns are qualitative, they either match or they do not. If there is a mismatch at any position, the pattern will not match, even though the mismatch is a conservative, biologically feasible substitution.

Despite their limitations, patterns are still very popular since:

- They are easy to formulate and to use.
- The scan of a protein database with patterns can be performed in reasonable time on any computer.
- They are well designed for the detection of biologically meaningful sites and the residues playing a structural or functional role are easily identifiable. This quality of patterns is further strengthened by the information contained in the */SITE* qualifier of the CC lines, which indicates, among other features, active sites, disulfide bonds, and/or metal and other binding sites in a PROSITE pattern.

3. PROSITE profiles

Recently, patterns have tended to be replaced with the more sophisticated *generalized profiles* (or *weight matrices*), which show an enhanced sensitivity when compared to patterns and are able to detect highly divergent families or domains with only a few well-conserved positions. Profiles also characterize protein domains over their entire length, not just the most conserved parts of them.

The generalized profiles (Bucher *et al*., 1996; Bucher and Bairoch, 1994; Hofmann, 2003) used in PROSITE are quantitative motif descriptors providing a scoring table of position-specific amino acid weights and gap costs. The automatic procedure used for deriving profiles from multiple alignments is capable of assigning appropriate weights to residues that have not yet been observed at a given alignment position, using the empirical knowledge about amino acid substitutability contained in a substitution matrix (Dayhoff *et al*., 1978; Henikoff and Henikoff, 1992) or methods involving hidden Markov modeling. The numerical weights of a profile are used to calculate a similarity score for any alignment between a profile and a sequence, or parts of a profile and a sequence. An alignment with a similarity score higher than or equal to a given cutoff value constitutes a match. This cutoff value is estimated by calibrating the profile against a randomized protein database (Pagni and Jongeneel, 2001; Sigrist *et al*., 2002). The normalization procedure used for PROSITE profiles makes the normalized scores independent of the database size, allowing comparison of scores from different scans. A mismatch at a highly conserved position can thus be accepted, provided that the rest of the sequence displays a sufficiently high level of similarity.

4. Functionally and/or structurally relevant residues in profiles

Although profiles generally cover domains or proteins over their entire length and hence make the identification of the biologically most relevant residues less obvious than with patterns, functionally and structurally important amino acids can be recovered from the alignment between the profile and the matched sequence, provided that their position in the profile is known. As a domain might have acquired different functions through substitutions of critical residues during

evolution, it is necessary, once these residues are identified, to check if they fulfill the condition expected for the functional and/or structural properties associated with the domain. If not, it means that the expected biological property is no longer associated with the domain or that residues at other positions in the sequence compensate for it. Hence, serine proteases of the trypsin type and haptoglobin share a structural domain whose catalytic residues are no longer present in the nonenzymatic haptoglobin (Kurosky *et al*., 1980) (see Figure 1). This example illustrates clearly that it is not sufficient to just know about the presence of a domain to make valid functional predictions, but that in most cases it is necessary to have information about particular residues to predict the biological function of a domain more accurately.

As PROSITE aims to be an annotated database consisting of biologically significant sites, patterns, and profiles that help reliably identify to which known protein family a new sequence belongs, we decided to use information about functionally and/or structurally meaningful residues of a profile to make more specific assumptions. Whereas positions of biologically relevant residues are stored together with the regular expression (CC \SITE lines of the pattern), we have created a new section of PROSITE (ProRule) to store the position of interesting residues contained in profiles. ProRule contains the position in the profile of residues potentially involved in disulfide bonds, catalytic activity, phosphorylation, prosthetic group attachment (heme, pyridoxal-phosphate, biotin, etc.), or metal, protein, or molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) binding, or forming a transmembrane region. For each of these amino acids at potentially interesting positions, ProRule also contains the condition(s) the residue must fulfill to play its structural and/or functional role. These conditions can, in some cases, be quite complex in order to provide the user with the most precise and relevant information possible. For example, as the catalytic center of the trypsin-type serine protease domain is formed by a triad of residues, we consider that the group of three amino acids must be present to make a confident prediction about the function of the domain. For the prediction of disulfide bridges, conditions have to be even more complex as it is known that disulfide bridges are not always invariant among homologous domains: conserved cysteine residues can exhibit different disulfide bond patterns from one protein type to the other (Calvete *et al*., 2000; Chong *et al*., 2002). In some instances, the situation is so complex that it is not possible to reduce the biological diversity to a set of conditions predicting unambiguously which bond will be made, and no rules can be made.

From the positional and conditional information contained in the rules, residues with a potential functional and/or structural function can be retrieved in a new sequence from its alignment with the profile, and their agreement with the expected function checked, providing additional useful information to the user. For the 455 profiles available in the current release of PROSITE, 94 ProRules (∼21%) contain information about functionally and/or structurally critical amino acids.

PROSITE users analyzing their sequences through the ScanProsite web page (**http://www.expasy.org/tools/scanprosite/**) can benefit from the additional information contained in ProRule and get not only the domain boundaries but also the position and potential function of critical residues that might be found in the

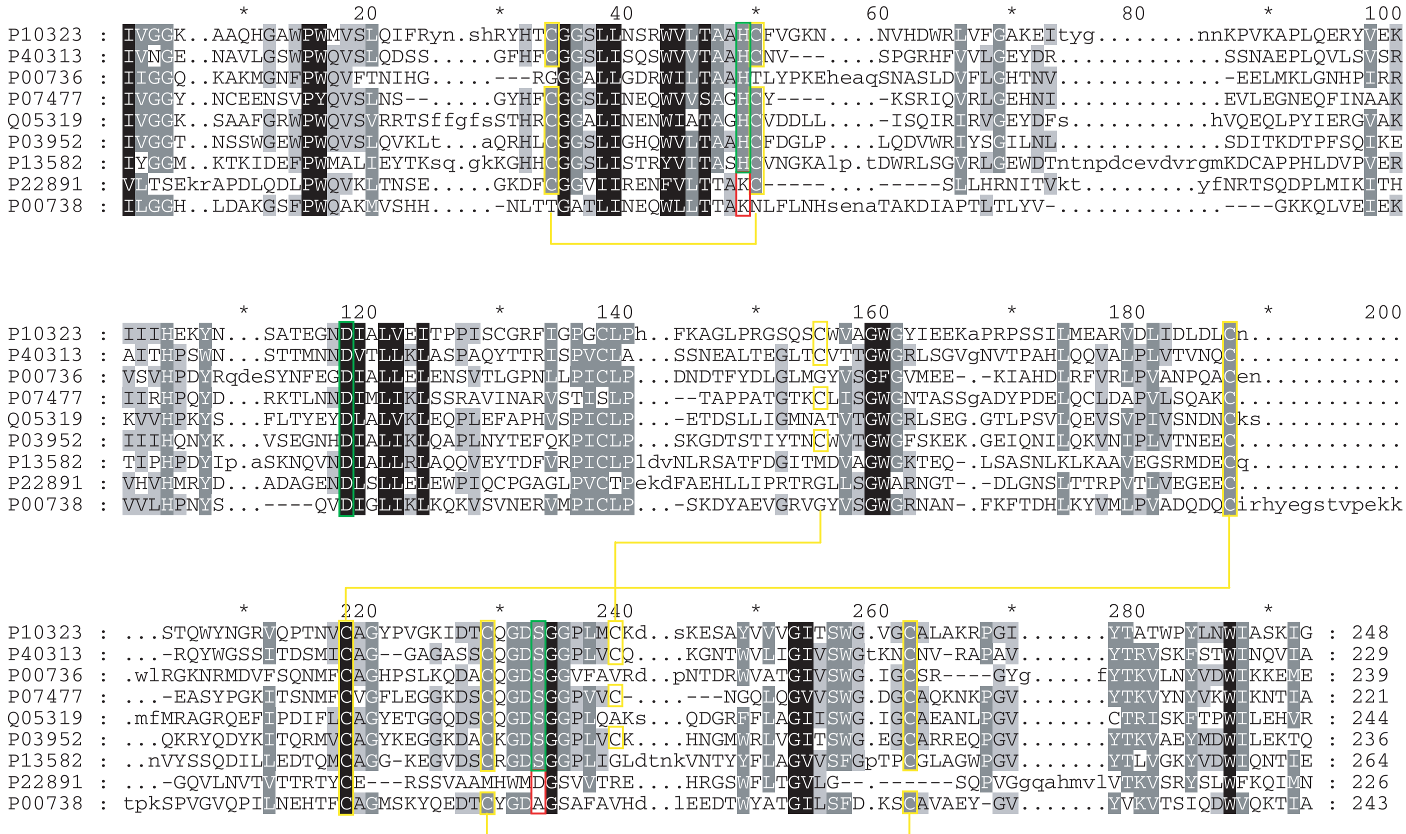


Figure 1 Multiple alignment of the trypsin-type serine protease domain (PS50240) of human acrosin (P10323), chymotrypsin-like protease CTRL-1 (P40313), complement C1r component (P00736), trypsin I (P07477), plasma kallikrein (P03952), vitamin K–dependent protein Z (P22891) and haptoglobin (P00738), and drosophila serine proteinases stubble (Q05319) and easter (P13582). Some proteins lack cysteine residues involved in disulfide bonds and the vitamin K–dependent protein Z and haptoglobin have lost two of the essential catalytic residues (histidine and serine) and have no enzymatic activity. The residues from the catalytic triad are shown in green boxes if they are compatible with the catalytic activity and in red boxes if not. The cysteines known to be involved in disulfide bonds (yellow lines) are in yellow boxes. The accession numbers (AC) correspond to records in the Swiss-Prot section (Boeckmann *et al*., 2003) of the UniProt Knowledgebase (Apweiler *et al*., 2004)

profiles. This supplementary service should facilitate protein analysis by making functional prediction more accurate.

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6

Acknowledgments

We wish to thank Andrea H. Auchincloss for the correction of the manuscript. PROSITE is supported by grant n° 3152A0-103922/1 from the Swiss National Science Foundation.

Further reading

Hofmann K (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Briefings in Bioinformatics*, **1**, 167–178.

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

Bucher P and Bairoch A (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, **2**, 53–61.

Bucher P, Karplus K, Moeri N and Hofmann K (1996) A flexible motif search technique based on generalized profiles. *Computers & Chemistry*, **20**, 3–23.

Calvete JJ, Moreno-Murciano MP, Sanz L, Jurgens M, Schrader M, Raida M, Benjamin DC and Fox JW (2000) The disulfide bond pattern of catrocollastatin C, a disintegrin-like/cysteine-rich protein isolated from Crotalus atrox venom. *Protein Science*, **9**, 1365–1373.

Chong JM, Üren A, Rubin JS and Speicher DW (2002) Disulfide bond assignments of secreted Frizzled-related protein-1 provide insights about Frizzled homology and netrin modules. *The Journal of Biological Chemistry*, **277**, 5134–5144.

Dayhoff MO, Schwartz RM and Orcutt BC (1978) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation: Washington.

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915–10919.

Hofmann K (2003) Profile searching. *Nature Encyclopedia of the Human Genome*, Vol. 4, Macmillan Publishers Ltd, Nature Publishing Group: London, pp. 726–730.

Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**, D134–D137.

Kurosky A, Barnett DR, Lee T-H, Touchstone B, Hay RE, Arnott MS, Bowman BH and Fitch WM (1980) Covalent structure of human haptoglobin: a serine protease homolog. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 3388–3392.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**, 315–318.

Pagni M and Jongeneel CV (2001) Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics*, **2**, 51–67.

Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A and Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, **3**, 265–274.

Short Specialist Review

The PRINTS protein fingerprint database: functional and evolutionary applications

Teresa K. Attwood and Alexander L. Mitchell

University of Manchester, Manchester, UK

European Bioinformatics Institute, Cambridge, UK

Anna Gaulton, Georgina Moulton and Lydia Tabernero

University of Manchester, Manchester, UK

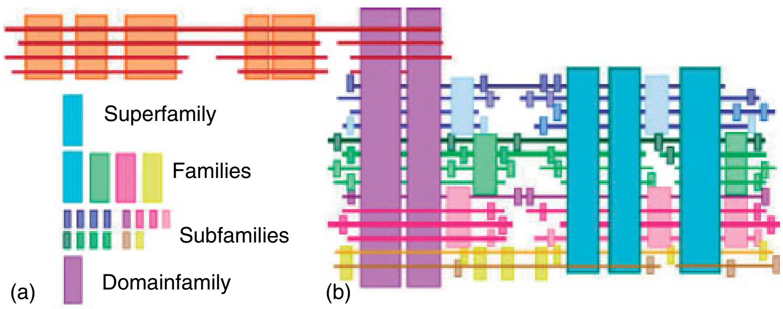
1. Introduction

The PRINTS database houses a collection of protein sequence "fingerprints" that, by encapsulating the conserved features of known family members, may be used to assign functional and structural attributes to uncharacterized sequences. As with several other pattern-recognition approaches, fingerprinting exploits multiple sequence alignments to derive its familial signatures. When building alignments, insertions are often required at particular positions to bring equivalent parts of adjacent sequences into register. As a result, islands of conservation tend to emerge from a backdrop of mutational change. These conserved regions or "motifs" (typically $<15-20$ residues in length) tend to correspond to the core structural or functional elements of the protein. Taken together, these motifs may therefore provide a signature for the aligned family. In 1990, this observation prompted the creation of the PRINTS database of diagnostic fingerprints for use in protein sequence classification and characterization.

2. Distinguishing features of fingerprinting

The fingerprinting method involves manual creation of a seed alignment, followed by identification and excision of conserved motifs (up to a maximum of 15 motifs, of maximum length 30 residues). The motifs are translated into frequency matrices (no mutation or other similarity data are used to weight the results)

and searched against the source database, a fragmentless version of the Swiss-Prot/TrEMBL component of UniProt (Apweiler *et al*., 2004). Fingerprint diagnostic performance is enhanced by iterative scanning; the motifs therefore mature with each database pass, as more sequences are matched and assimilated into the process. When the iterative process has converged, the fingerprints are annotated manually with biological information, literature, database cross references, and so on, before accession to the database. The fingerprinting technique departs from other pattern-matching methods by readily allowing the creation of signatures at the domain-, superfamily-, family-, and subfamily-specific levels (Figure 1). This degree of resolution is possible because the manual approach allows us to focus not only on regions of shared similarity (such as those that characterize superfamilies) but also on regions of difference. (such as those that resolve subfamilies from closely related siblings within a family, or those that distinguish



Highest scoring fingerprints for THRB_HUMAN			
Fingerprint	E-value	GRAPHScan	Motif3D
PROTHROMBIN (relations)	1.513982e-91	Graphic	3D-view
CHYMOTRYPSIN (relations)	1.566739e-16	Graphic	
KRINGLE (relations)	6.362489e-12	Graphic	3D-view
GLABLOOD (relations)	1.916728e-11	Graphic	3D-view

(c)

Figure 1 Resolving domain-, super- and subfamily relationships. At the top of the figure, gapped alignments of two theoretical families are represented: (a) a gene family and (b) a typical superfamily. Motifs are denoted by colored rectangles. Members of the superfamily are likely to share a similar overall structure and function; its constituent families (blue, green, pink, and yellow) have more specific functions; and their subfamilies (shades of each color) have highly specialized functions closely related to the family to which they belong. At the N-terminus of the superfamily lies a genetically mobile domain, or module (the purple motifs), a copy of which also lies at the C-terminus of the gene family. While the domain has the same overall structure and function in both biological contexts, the gene family and superfamily are otherwise unrelated. All families containing the domain collectively comprise a domain family. Fingerprint analysis of Swiss-Prot entry THRB HUMAN (accession number P00734), human prothrombin, is shown in (c). One family-level fingerprint is matched, PROTHROMBIN (PR01505), and one superfamily-level fingerprint, CHYMOTRYPSIN (PR00722). This suggests that the protein is prothrombin and belongs to the peptidase family S1A, the chymotrypsin serine proteases. The sequence also matches two domain signatures: KRINGLE (PR00018) and GLABLOOD (PR00001); these modules are commonly found in proteins involved in blood coagulation

families from their parent superfamilies). This is important because it is the subtle differences between close relatives that largely determine their functional specificities. This hierarchical approach has been used to analyze a range of proteins, especially those of pharmaceutical interest – for example, to resolve G-protein-coupled receptor (GPCR) superfamilies into their constituent families and receptor subtypes (Attwood, 2001), and to finely classify a variety of channel proteins, transporters, and enzymes. Fingerprinting therefore provides a useful complement to the hidden Markov model and profile-based methods, which tend to specialize in the diagnosis of superfamilies or domains. The diagnostic potency of fingerprints derives primarily from the use of multiple motifs and the mutual context provided by motif neighbors – the more motifs a signature contains, the better able it is to identify distant relatives, even when parts of the signature are absent (*e.g.*, a sequence matching five of eight motifs may still be diagnosed as a true match, if the motifs are matched in the correct order, and the distances between them are consistent with those expected of true neighboring motifs). The method is thus sufficiently robust to tolerate mismatches at the level both of individual residues within motifs, and of motifs within complete signatures (Attwood, 2002); it is therefore more flexible and powerful than single-motif approaches.

3. Distinguishing features of PRINTS

PRINTS' motifs are stored in the form of ungapped, local alignments. This allows different implementations to be established with alternative scoring methods. Thus, a Blocks-format version of the database that exploits Blocks-scoring methods is available at the Fred Hutchinson Cancer Research Center (Henikoff *et al*., 2000). In addition, the eMOTIF and eMATRIX resources at Stanford derive permissive regular expressions (Huang and Brutlag, 2001) and position-specific scoring matrices (Wu *et al*., 2000) from PRINTS' multiply aligned motifs, offering different levels of stringency from which to infer the significance of matches. Derivative databases are useful as they provide different perspectives on the same data: they allow validation of results where there are equivalent matches in more than one resource and they offer the chance to make diagnoses that may have been missed by the original implementation.

Another important feature of PRINTS is that it is one of the few signature databases that includes significant amounts of manually crafted annotation. The purpose of the annotation is both to provide background information on the gene or domain family being described and to rationalize the motifs in functional and structural terms. The annotation procedure is exhaustive and time consuming; consequently, PRINTS remains relatively small by comparison with other, largely automatically derived, signature databases. Overall, however, the precision of the results, coupled with the extent of its annotations, has justified the sacrifice of speed, and has set the database apart from the growing number of resources that contain little, if any, annotation. In this respect, PRINTS adopts a similar philosophy to PROSITE (described in Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6) (Falquet *et al*.,

2002) and, in 1992, the database curators agreed to collaborate in creating a composite resource – this goal was eventually realized in 1999 with the first release of InterPro (*see* Article 83, **InterPro**, Volume 6) (Mulder *et al*., 2003), which, by that time, also included profiles, Pfam (described in Article 86, **Pfam: the protein families database**, Volume 6) (Bateman *et al*., 2002) and ProDom (Servant *et al*., 2002).

4. An automatic supplement, prePRINTS

To address its limited size, we developed an automatic supplement, termed prePRINTS (Attwood *et al*., 2003). This uses ProDom clusters as seed alignments; motif detection and iterative scanning are performed automatically, and the results annotated using the PRECIS annotation tool (Mitchell *et al*., 2003; (**http://umber.sbs.man.ac.uk/cgi-bin/dbbrowser/precis/precis.cgi**). prePRINTS is a valuable extra source of fingerprints, serving as an "incubator", where entries are manually refined before accession to PRINTS.

5. Functional and evolutionary applications

In addition to its role in classifying protein sequences into families and providing source data for the Blocks, eMOTIF, eMATRIX, and InterPro databases, PRINTS has been used in a number of functional and evolutionary studies. For example, its extensive collection of hierarchical K+ channel fingerprints has been used to mine mammalian, *Caenorhabditis elegans*, and *Drosophila melanogaster* proteomes, and the identified proteins classified according to the seven known families. Subsequent phylogenetic analyses allowed identification of orthologous relationships and shed light on likely underpinning evolutionary events, such as gene duplications and gene losses (Moulton *et al*., 2003). The implicit evolutionary relationships within the PRINTS hierarchy have thus been exploited, and have been explicitly expressed, by using phylogenetic methods.

In another analysis, PRINTS' vast repertoire of GPCR fingerprints has been used to investigate the structural and functional relationships of superfamily-, family-, and subfamily-level motifs. The study revealed that while some family- and subfamily motifs map onto residues and regions that participate in ligand binding (Figure 2) and G-protein coupling, others contain regions that are involved in protein–protein interactions and receptor dimerization (Figure 3), suggesting that these fingerprints could be used to predict the locations of ligand-binding sites and other functional features in as yet uncharacterized receptor families.

More recently, fingerprints for two mitogen-activated protein kinase (MAPK) phosphatase subfamilies (tyrosine specific and dual specificity) were used in genome-wide screens, identifying more than 80 MAPK phosphatase domains, several from partial sequences or nonannotated proteins. One of the predicted orthologs from *Xenopus* was confirmed experimentally to bind to ERK1/2, suggesting a role in MAPK signaling. Structural analysis of the fingerprints using the 3D structure of MAPK phosphatases as template revealed motifs

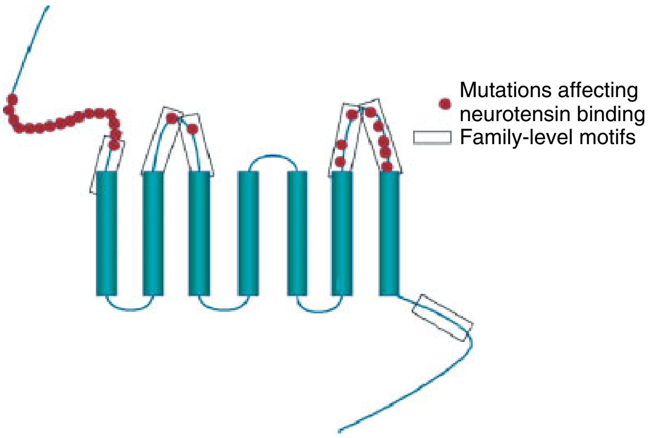


Figure 2 Schematic drawing of the neurotensin receptor, with family motifs mapped (rectangles), together with positions of mutations known to affect neurotensin binding (red circles). Most motifs contain residues important in ligand binding

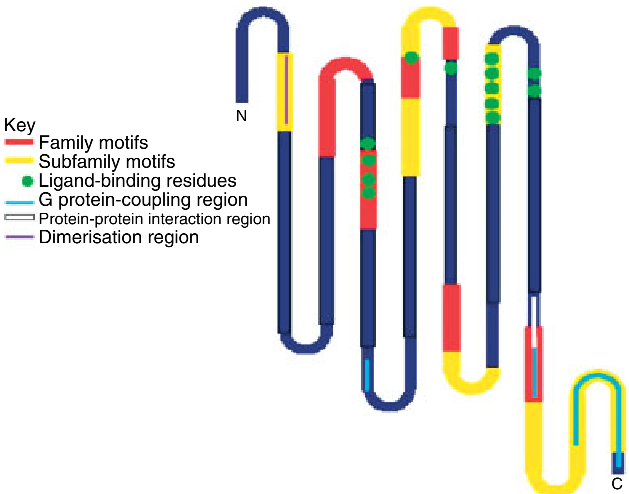


Figure 3 Schematic drawing of the bradykinin receptor, with family- (red) and subfamily (yellow) motifs mapped, together with positions of residues involved in ligand binding (green circles), G-protein coupling (light blue bars), receptor dimerization (purple bar), and other protein–protein interactions (white bar)

in the N-terminal non-catalytic regions that coincided with reported MAPK binding sites, and other motifs in the catalytic phosphatase domain suggestive of putative allosteric sites for modulation of protein–protein interactions (Figure 4). The fingerprints have therefore been useful not only for rapid annotation and

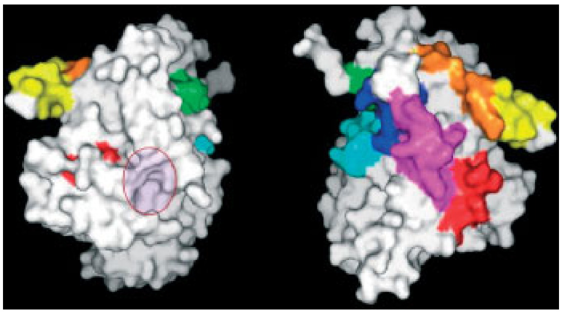


Figure 4 Surface representation of the catalytic domains of hematopoietic protein-tyrosine phosphatase; the motifs in the fingerprint are colored on the surfaces to highlight their clustering. The left-hand image shows the front view of the protein, the active site marked with a translucent oval; the right-hand image shows the back view, where the fingerprint motifs can be seen to form a continuous exposed surface, suggestive of an interaction surface

classification of MAPK phosphatase domains, but also for novel identification of putative functional motifs, providing a framework for future experimental studies.

6. Availability

Maintained at the University of Manchester, PRINTS 38.0 and prePRINTS 2.0, containing ~11 500 and 1510 motifs (1900 and 380 fingerprints) respectively, are available for searching at **http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/** and **http://www.bioinf.man.ac.uk/dbbrowser/prePRINTS/**.

7. Conclusion

The PRINTS protein fingerprint database provides a potent means of classifying uncharacterized sequences at the domain-, super-, and subfamily levels and also provides a rich source of manually crafted annotation. These features set it apart from the growing number of signature databases, many of which offer little or no annotation and tend to focus on the diagnosis of domain families and superfamilies. In addition to its utility as a sequence analysis tool, the resource also has value both in shedding light on evolutionary relationships and in predicting functional sites in complex gene families.

Related articles

Article 44, **Protein interaction databases**, Volume 5; Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 87,

The PIR SuperFamily (PIRSF) classification system, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**(1), D115–D119.

Attwood TK (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends in Pharmacological Sciences*, **22**, 162–165.

Attwood TK (2002) The PRINTS database: a resource for identification of protein families. *Briefings in Bioinformatics*, **3**, 252–263.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, *et al* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, **31**, 400–402.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Research*, **28**, 263–266.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K and Bairoch A (2002) The PROSITE database, its status in 2002. *Nucleic Acids Research*, **30**, 235–238.

Henikoff J, Greene EA, Pietrokovski S and Henikoff S (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research*, **28**, 228–230.

Huang JY and Brutlag DL (2001) The EMOTIF database. *Nucleic Acids Research*, **29**, 202–204.

Mitchell AL, Reich JR and Attwood TK (2003) PRECIS – An automatic tool for generating protein reports engineered from concise information in SWISS-PROT. *Bioinformatics*, **19**, 1664–1671.

Moulton G, Attwood TK, Parry-Smith DJ and Packer JC (2003) Phylogenomic analysis and evolution of the potassium channel gene family. *Receptors and Channels*, **9**, 363–377.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**, 315–318.

Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D and Kahn D (2002) ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics*, **3**, 246–225.

Wu TD, Nevill-Manning CG and Brutlag DL (2000) Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**, 233–244.

Pfam: the protein families database

Robert D. Finn
Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

Over a million protein sequences have been deposited in the UniProt sequence database (Apweiler *et al*., 2004). The number of new sequences deposited shows no signs of abating. Thus, protein sequence analysis may seem like an endless task. However, the majority of protein sequences appear to fall into a few thousand protein families (Chothia, 1992). Therefore, if we can place new sequences into existing families, we can make our task simpler. Very often, these protein families are representative of proteins at the domain level. Domains are discrete structural units that are frequently found in different molecular contexts. Different combinations of domains give rise to functional diversity (Vogel *et al*., 2004), which includes their ability to form different protein interactions.

2. The Pfam database

Pfam is a database of protein domain families (Bateman *et al*., 2004). Each family is represented by multiple sequence alignments and profile hidden Markov models (HMMs). In addition, each family has associated annotation, literature references, and links to other databases. The entries in Pfam are freely available via the web and in flatfile format (Pfam is available in Europe at **http://www.sanger.ac.uk/Software/Pfam/** (United Kingdom), **http://www.cgb.ki.se/Pfam/** (Sweden), and **http://pfam.jouy.inra.fr/** (France), and in the United States at **http://pfam.wustl.edu/**). Pfam is a founding member database of InterPro (*see* Article 83, **InterPro**, Volume 6) and, therefore, also available via the InterPro site at **http://ebi.ac.uk/interpro**.

The use of Pfam by molecular biologists as a protein information resource and analysis tool is widespread. The multiple sequence alignments around which Pfam families are built are important for understanding both protein structure and function. The alignments are also the basis for techniques such as secondary structure prediction, fold recognition, and phylogenetic analysis and can guide mutation design. In addition to the identification of domains in novel protein sequences,

Pfam can be used in conjunction with the Wise2 package to predict genes and annotate genomic DNA.

There are two multiple sequence alignments for each Pfam family; the seed alignment containing a manually curated set of representative members of the family and the full alignment that contains all detectable members in the sequence database. The underlying sequence database, called *Pfamseq*, is based on UniProt (Apweiler *et al*., 2004). An HMM is built from the seed alignment using the HMMer2 software package (**http://hmmer.wustl.edu/**) and used to search the Pfamseq database. All matches (scoring above a curated set of thresholds) are aligned against the HMM to make the full alignment. Full alignments can be large, with the top 20 families now each containing over 7000 sequences. For example, the largest full alignment in Pfam (the HIV GP120 glycoprotein) has over 40 000 members, yet the seed alignment only has 24 representative members.

Pfam is designed to be both accurate and comprehensive (Sonnhammer *et al*., 1998). As such, the Pfam database is structured in two parts, referred to as *Pfam-A* and *Pfam-B* (Bateman *et al*., 2004; Sonnhammer *et al*., 1998). The Pfam-A entries are manually curated families with high-quality sequence alignments and annotations. At the time of writing, the latest release of Pfam (13.0) contains 7426 Pfam-A families. Seventy-four percent of all sequences in UniProt contain at least one match to Pfam-A, and these families cover approximately 53% of all residues in the protein databases. However, as Pfam aims to be comprehensive, the sequences not covered by Pfam-A families are clustered according to the PRODOM database (Servant *et al*., 2002), and are referred to as *Pfam-B entries* (see Pfam-B flatfile). The Pfam-B supplement contains ~90 000 small protein families of lower quality.

So why use a domain database? Traditional approaches to large-scale sequence annotation use a pairwise sequence comparison method such as BLAST (Altschul *et al*., 1990) to find similarity to proteins of known function (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7). Annotations are transferred from the protein of known function to the predicted protein. The pairwise similarity search does not give a clear indication of the domain structure of the proteins, and mistakes in annotation can result from not considering the domain organization of proteins. For example, a protein may be misannotated as an enzyme when the similarity is only to a regulatory domain.

3. Pfam proteome coverage

Pfam was originally designed to aid the annotation of the genomes, in particular the *Caenorhabditis elegans* genome. More recently, Pfam-based annotations have been applied to many genome projects, including human, mouse, and fly. The coverage of Pfam across the sequence database is high. An alternative measurement of coverage is the coverage of a proteome. Figure 1 shows graphically the results from Pfam/HMMer analysis on the proteomes from eight representative genomes: the eubacteria *Escherichia coli*, *Aquifex aeolicus*, *Streptomyces coelicolor*, the archaeon *Methanococcus jannaschii*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, and human *Homo sapiens*.

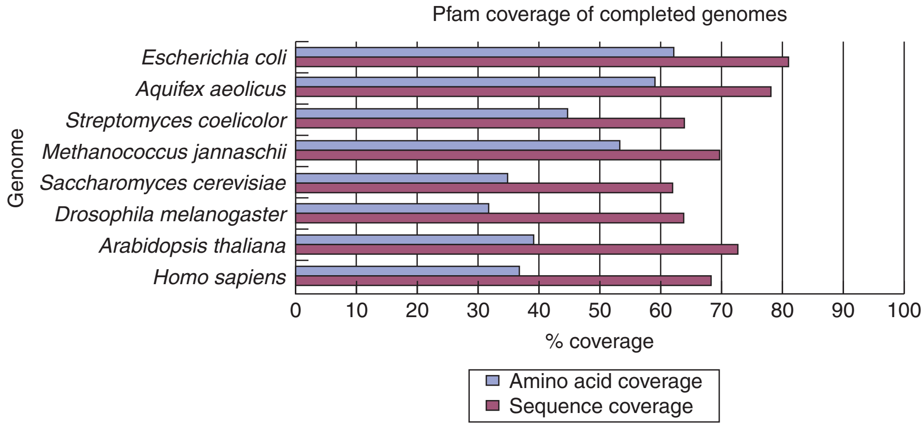


Figure 1 Pfam-A coverage of a representative selection of completed genomes, from all kingdoms of life. The per-residue coverage is indicated by the blue bars and the per-sequence coverage is indicated by the red bars

In the example proteomes, the sequence coverage by Pfam-A families is always greater than 60%, but in the case of *E. coli*, exceeds 80%. The amino acid coverage can be as low as 34.8% for yeast, yet as high as 62.2% in *E. coli.* Thus, Figure 1 shows that the genomic coverage by Pfam differs greatly.

4. Genome comparisons

As more genomes become completed, it is interesting and informative to identify which protein families are specific to a particular genome and which families are common to genomes. Pfam readily allows such analyses to be performed. For example, if the protein families in human and the model organism *C. elegans* are compared, we find that there are 2109 families that are found in both organisms, 777 families that are specific to human, and 160 that are specific to *C. elegans* (see Figure 2). In the case of common domains, *C. elegans* may provide an appropriate model system for studying human proteins of interest. Similar analyses may be useful in identifying domains that are responsible for symbiosis in bacteria and drug targeting for intracellular pathogens.

Pfam is not only aimed at sequence and genome annotation. Pfam also contains functional information about the families, where available. For example, secondary structure, solvent accessibility, and active site information can be found in the alignments. Recently, a whole new resource has been made available for studying domain interactions, called *iPfam* (Finn *et al*., 2005) (**http://www.sanger.ac.uk/Software/Pfam/iPfam**).

5. Domain interactions

Large-scale protein interaction analyses indicate that the majority of proteins bind transiently or are part of large molecular complexes (Gavin *et al*., 2002).

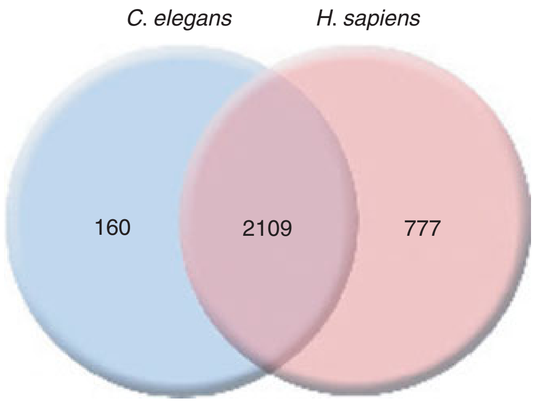
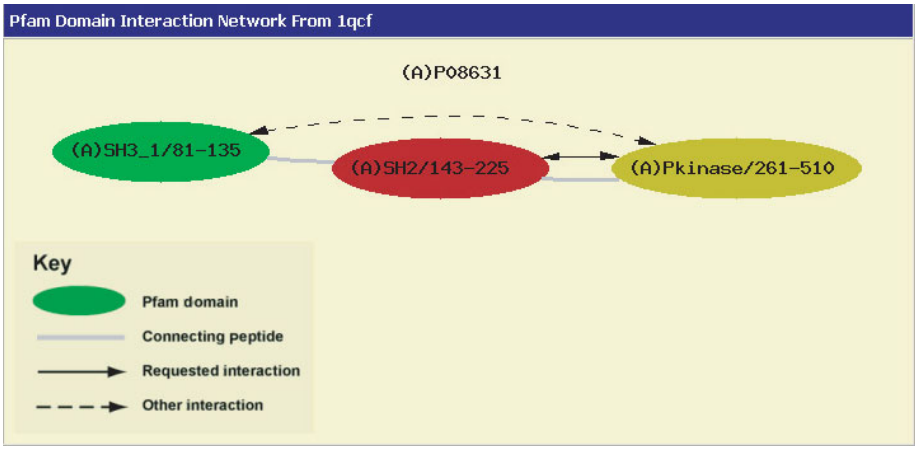


Figure 2 Comparison of domain distribution between *H. sapiens* (pink) and *C. elegans* (blue). The Venn diagram shows which families are in common between the two organisms (2109) and which families are specific to each organism. This data was produced from the search engine at **http://www.sanger.ac.uk/Software/Pfam/genome_dist.pl**

To understand protein interactions at a molecular level, information at the resolutions of domains, residues, and atoms are required. This sort of detailed interaction information is normally only found in the PDB database of protein structures. Approximately, one-third of the 7426 Pfam-A entries have one or more structural representatives. The *i*Pfam resource describes domain–domain interactions observed in known protein structure at the resolution of protein(s), domains, residues, and atoms. At the time of writing, there are over 2500 different domain–domain interactions described. Currently, the *i*Pfam data is laid out in two different ways. First, for a specific structure containing a domain–domain interaction. In this case, the structure, domains, and interactions are displayed as a 2D graph (Figure 3a). The interaction can be explored further at the residue- and atom-resolution levels, either as sequence or structure. The second way the data is presented in *i*Pfam allows the visualization of the conservation of domain–domain interactions between different examples of the same domain (Figure 3b). Such a display quickly allows the user to discern which residues are always found at a particular domain–domain interface and, when multiple domains are compared, whether interacting interfaces overlap.

6. Future prospects

Pfam continues to add new entries and review existing families to improve the coverage of sequence space. New strategies for building more comprehensive families are being developed. Web services allowing machine access to the data will be available shortly, to enable a more automated access to annotational and functional information contained within Pfam.



(a)



(b)

Figure 3 (a) Graphical representation of domain interactions that occur in the structure with PDB identifier 1QCF. In this three-domain protein, the Pkinase (PF00069) domain interacts with both the SH2 (PF00017) and SH3 (PF00018) domains, whereas the SH2 and SH3 do not form a domain–domain interaction with each other. (b) Comparison of SH2 domain interactions with SH3 domains. Each unique sequence is shown in the alignment. Residues of each SH2 domain that form an interaction with a SH3 domain are indicated by colored boxes. Each row of colored boxes represents a separate domain interaction, with the PDB identifier and chain noted above the row. In this example, several distinct surfaces are used by SH2 to interact with SH3

Related articles

Article 83, **InterPro**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7

Acknowledgments

Thanks go to all members of the Pfam consortium, past and present, who have contributed to the database. Thanks also go to Alex Bateman and David Studholme for their constructive comments on this article. RDF is supported by an MRC grant.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Chothia C (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Finn RD, Marshall M and Bateman A (2005) iPfam: Visualization of protein-protein interactions in PDB at domain and aminoacid resolutions. *Bioinformatics*. **21**(3), 410–412.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. (2002) *Nature*, **415**, 141–147.

Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D and Kahn D (2002) ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics*, **3**, 246–251.

Sonnhammer EL, Eddy SR, Birney E, Bateman A and Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, **26**, 320–322.

Vogel C, Bashton M, Kerrison ND, Chothia C and Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, **14**, 208–216.

Short Specialist Review

The PIR SuperFamily (PIRSF) classification system

Winona C. Barker
National Biomedical Research Foundation, Washington DC, USA

Raja Mazumder, Anastasia N. Nikolskaya and Cathy H. Wu
Georgetown University Medical Center, Washington DC, USA

1. Introduction

The Protein Information Resource (PIR) supports genomic and proteomic research by providing protein databases and analysis tools freely accessible to the scientific community (Wu *et al*., 2003a). The PIR-International Protein Sequence Database of functionally annotated protein sequences grew out of the *Atlas of Protein Sequence and Structure* edited by Dayhoff (1965–1978). The previously independent protein sequence database activities of PIR, the European Bioinformatics Institute, and the Swiss Institute of Bioinformatics are now unified to provide UniProt, the Universal Protein Resource (Apweiler *et al*., 2004), as a central resource of protein sequence and function.

The major purpose of the PIRSF (Protein Information Resource SuperFamily) classification system is to facilitate sensible propagation and standardization of protein annotation, specifically in the UniProt database. Basing classification on full-length proteins allows annotation of biological functions, biochemical activities, and sequence features that are family specific. In contrast, the domain architecture of a protein provides insight into general functional and structural properties, as well as into complex evolutionary mechanisms. Thus, the PIRSF can be used by the scientific community as a tool for such diverse activities as genome annotation, study of evolutionary mechanisms, comparative genomic studies, choosing of candidate proteins for structural genomics projects, and analysis of proteomic expression data.

2. PIRSF concept and properties

The protein superfamily concept articulated by Dayhoff (1976) was the first classification based on sequence similarity, where sequences were categorized into

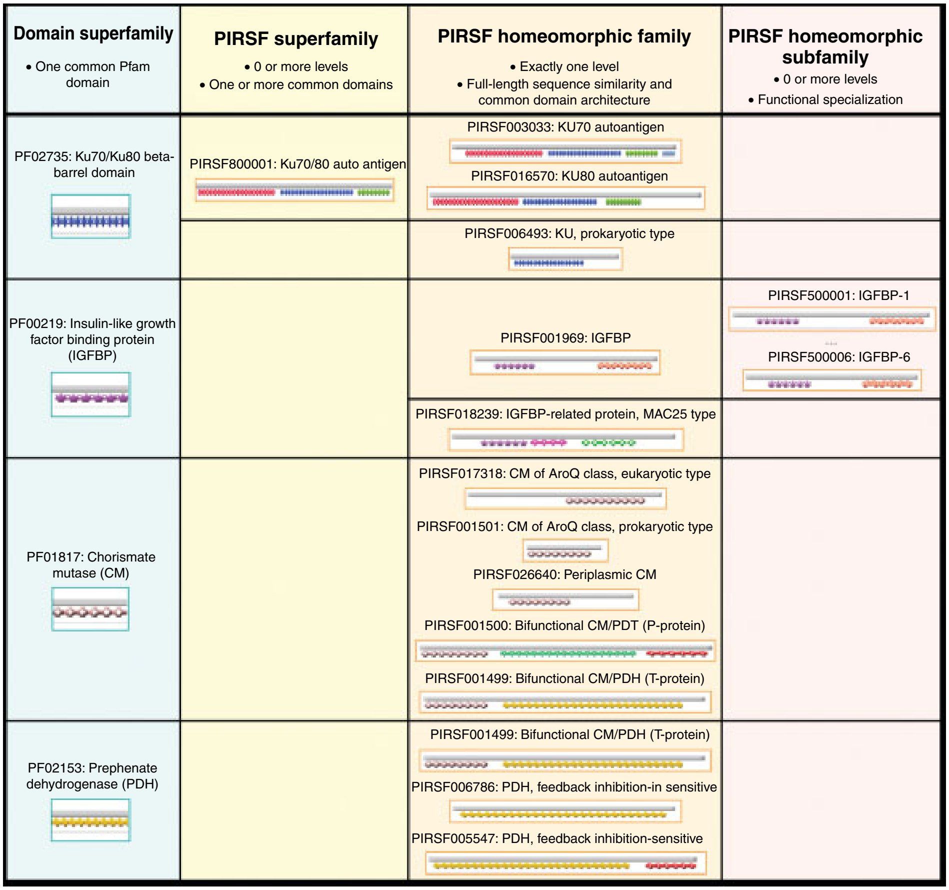


Figure 1 PIRSF classification system based on evolutionary relationship of full-length proteins

superfamilies and further subdivided into closely related families, very closely related subfamilies, and nearly identical entries using different percent identity thresholds. This hierarchical structure was challenged by the discovery of multidomain proteins in which the several domains may have different evolutionary histories. To permit the representation of these complex relationships, the PIRSF system is formally defined as *a network classification system based on evolutionary relationship of full-length proteins* (Wu *et al*., 2004a) (Figure 1).

The primary PIRSF classification unit is the *homeomorphic family* whose members are both *homologous* (evolved from a common ancestor) and *homeomorphic* (sharing full-length sequence similarity and a common domain architecture). Common domain architecture is indicated by the same type, number, and order of core domains, although variation may exist for repeating domains and/or auxiliary domains. Each protein can be assigned to only one homeomorphic family, which may have zero or more parent *superfamilies* and zero or more child *subfamilies*. The parent superfamilies connect related families and orphan proteins on the basis of one or more common domains, which may or may not extend over the entire

lengths of the proteins. The child subfamilies are homeomorphic groups that may represent functional specialization. The flexible number of parent–child levels from superfamily to subfamily reflects natural clusters of proteins with varying degrees of sequence conservation, rather than arbitrary similarity thresholds. While a protein will belong to one and only one homeomorphic family, multidomain proteins may belong to multiple superfamilies (hence, the network structure). A domain superfamily, which consists of all proteins that contain a particular domain, is usually represented by the corresponding Pfam domain (Bateman *et al*., 2004) for convenience.

3. PIRSF case examples

Figure 1 illustrates three PIRSF examples, one with a superfamily (PIRSF800001: Ku70/80 autoantigen), one with subfamilies (PIRSF001969: insulin-like growth factor binding protein, IGFBP), and one with multiple domain parents (PIRSF001499: bifunctional chorismate mutase/prephenate dehydrogenase, T-protein).

Ku is a multifunctional protein found in both eukaryotes and prokaryotes. The eukaryotic Ku protein is an obligate heterodimer of two subunits, Ku70 and Ku80. Ku70 and Ku80, predicted to have evolved by gene duplication from a homodimeric ancestor, share three domains and are placed into one superfamily. They differ, however, at their carboxyl ends, so they form two homeomorphic families (PIRSF003033 and PIRSF016570). The central Ku β-barrel domain (PF02735) is also conserved (without the N- and C-terminal extensions) in several prokaryotes (PIRSF006493). The prokaryotic Ku was predicted (Aravind and Koonin, 2001), and later experimentally confirmed, to form a homodimer involved in repairing double-strand breaks in DNA in a mechanistically similar fashion to eukaryotic Ku-containing proteins.

IGFBPs are a group of vertebrate-secreted proteins that bind to IGF with high affinity and modulate the biological actions of IGFs. These proteins share a common domain architecture, with an N-terminal IGF binding protein domain (PF00219) and a C-terminal thyroglobulin type-1 repeat domain, and are classified into a homeomorphic family (PIRSF001969). These proteins, however, have a highly variable mid-region and diverse functions ranging from the traditional IGF-related role to acting as growth factors independent of the IGFs. Six member types, IGFBP-1 through 6, have been delineated on the basis of gene organization, sequence similarity, and binding affinity to IGFs (Baxter *et al*., 1998). Each of the member types corresponds to a clearly distinguishable clade of the family phylogenetic tree and consists of orthologous sequences from various vertebrate species. Accordingly, six homeomorphic subfamilies (PIRSF500001-500006) are defined. The IGFBP family is related to the MAC25 gene family (PIRSF018239), which shares only the N-terminal IGFBP domain and has a relatively low IGF-binding activity.

Chorismate mutase (CM) (EC 5.4.99.5) and prephenate dehydrogenase (PDH) (EC 1.3.1.12) are two enzymes of the shikimate pathway present in bacteria, archaea, fungi, and plants. The corresponding domains (PF01817, PF02153) occur in proteins with different domain architectures, reflecting extensive domain shuffling. These proteins are classified into several PIRSF homeomorphic families

according to their domain architectures and are assigned family names indicative of their known or predicted functions.

4. PIRSF for functional prediction, database annotation, and protein nomenclature

Sequence analysis and protein classification based on full-length proteins can lead to educated predictions for both generic biochemical and specific biological functions. For example, members of PIRSF005547 are predicted to be feedback inhibition-sensitive PDHs due to the presence of the ACT domain (see Figure 1). The prediction is based on the ACT domain being essential for phenylalanine-mediated feedback inhibition and ligand binding in *Escherichia coli* P-protein (PIRSF001500) (Pohnert *et al*., 1999). Stand-alone versions of the PDH lacking the ACT domain (PIRSF006786) are predicted to be feedback inhibition-insensitive. Another example is PIRSF017318, where the members contain a catalytic CM domain and a unique N-terminal regulatory domain (Figure 1) and are subject to allosteric inhibition by tyrosine and activation by tryptophan. Even though this regulatory domain is not yet defined by any domain database, PIRSF classification points to its existence and provides information about its function.

PIRSF classification can facilitate accurate, consistent, and rich functional annotation of proteins (Wu *et al*., 2003b) and assist with the development of standard protein nomenclature and ontology. A classification-driven rule-based annotation approach (*see* Article 36, **Large-scale, classification-driven, rule-based functional annotation of proteins**, Volume 7) is central to UniProt protein annotation, in particular for transferring functional annotation such as functional sites and protein names from characterized members of a protein family to poorly studied proteins of this family. The multiple levels of sequence diversity improve annotation, allowing for classification of distantly related orphan proteins (usually at the levels of superfamilies and families) and accurate propagation of annotation (usually within families and subfamilies). While annotation can be conveniently propagated for homeomorphic families of closely related proteins, the subfamily level is most appropriate for larger and more diverse families. Standardized nomenclature can be developed on the basis of protein name rules that define standard protein names, synonyms, acronyms or abbreviations, Enzyme Commission (EC) name and number, and even “misnomers” for commonly misannotated proteins. The PIRSF names assigned at different classification levels, from more general to specific names, reflect different degrees of functional granularity and can be used to develop a protein ontology.

5. Accessing PIRSF

The PIRSF database is freely accessible from the PIR website at **http://pir.georgetown.edu/pirsf/**, and supported by the iProClass bioinformatics framework for data integration and associative analysis (Wu *et al*., 2004b). In particular,

PIRSF family classification is directly mapped to other family, function, and structure classification schemes, such as Pfam, InterPro (Mulder *et al*., 2003), and SCOP (Andreeva *et al*., 2004). PIRSF family reports present membership information, other classification schemes, structure and function cross-references, and graphical display of domain and motif architecture. To indicate their curation status, PIRSF families are labeled as "full", "preliminary", or "none", respectively, representing those that are fully curated with description and bibliography, partially curated with membership and domain architecture, or automatically classified and not yet curated. The reports for all curated PIRSF families include additional family annotation and links to multiple sequence alignments and evolutionary trees dynamically generated from seed members. Fully curated PIRSF families are integrated into the InterPro database. More than 20 PIRSF fields are searchable, including other database unique identifiers and annotations such as family name, keywords, and length. For example, one can identify all PIRSFs sharing one or more common Pfam domains, or all PIRSFs in a SCOP fold superfamily. Users can also perform a BLAST search (Altschul *et al*., 1997) of a query sequence to get a list of best-matching families and of all protein sequences above a given threshold.

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 36, **Large-scale, classification-driven, rule-based functional annotation of proteins**, Volume 7

Acknowledgments

The project is supported by grant U01-HG02712 from National Institutes of Health and grant DBI-0138188 from National Science Foundation.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, D226–D229.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Gasteiger E, Huang H, Martin MJ, Natale DA and O'Donovan C (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Aravind L and Koonin EV (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Research*, **11**, 1365–1374.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S and Sonnhammer EL (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Baxter RC, Binoux MA, Clemmons DR, Conover CA, Drop SL, Holly JM, Mohan S, Oh Y and Rosenfeld RG (1998) Recommendations for nomenclature of the insulin-like growth factor binding protein superfamily. *Endocrinology*, **139**, 4036.

Dayhoff MO (1965–1978) *Atlas of Protein Sequence and Structure. 5 Volumes, 3 Supplements*, National Biomedical Research Foundation: Washington.

Dayhoff MO (1976) The origin and evolution of protein superfamilies. *Federation Proceedings*, **35**, 2132–2138.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**, 315–318.

Pohnert G, Zhang S, Husain A, Wilson DB and Ganem B (1999) Regulation of phenylalanine biosynthesis. Studies on the mechanism of phenylalanine binding and feedback inhibition in the Escherichia coli P-protein. *Biochemistry*, **38**, 12212–12217.

Wu CH, Yeh L-S, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS and Suzek BE (2003a) The Protein Information Resource. *Nucleic Acids Research*, **31**, 345–347.

Wu CH, Huang H, Yeh LS and Barker WC (2003b) Protein family classification and functional annotation. *Computational Biology and Chemistry*, **27**, 37–47.

Wu CH, Nikolskaya A, Huang H, Yeh L-S, Natale D, Vinayaka CR, Hu Z, Mazumder R, Kumar S and Kourtesis P (2004a) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Research*, **32**, D112–D114.

Wu CH, Huang H, Nikolskaya A, Hu Z and Barker WC (2004b) The iProClass integrated database for protein functional analysis. *Computational Biology and Chemistry*, **28**, 87–96.

Short Specialist Review

Equivalog protein families in the TIGRFAMs database

Jeremy D. Selengut and Daniel H. Haft
The Institute for Genomic Research, Rockville, MD, USA

The process of genomic annotation – the assignment of function to nucleotide and protein sequences – involves leveraging a small amount of experimental data and applying it to the maximum number of sequences possible without making significant false statements. Experimental characterizations are often found for particular enzymes in only one or a handful of species, but these activities must be identified over the whole breadth of the sampled tree of life. Owing to the conservative nature of protein evolution, sequences derived from a common ancestor have homology and this homology is generally more pronounced to the extent that the function of the progeny of this ancestor is unchanged. An "equivalog" is defined (Haft *et al*., 2001) as a family of proteins conserved in function since their last common ancestor. This term differs from the more often used term "ortholog", which is not limited to groups with conserved function and is strictly limited to speciation events and excludes horizontal gene transfers (Fitch, 1970). Whereas orthologs are broadly useful objects for determining the process of protein evolution and discriminating the relationships between sequences, equivalogs are most useful in genomic annotation where the conservation of function is a critical factor.

The combination of multiple sequence alignments and phylogenetic analysis (tree-building) is a powerful tool for identifying families of proteins sharing a common ancestry. Multiple sequence alignments of such families can be represented mathematically by Hidden Markov Models (HMMs) such that novel query sequences can be compared and accepted or rejected as members of the family on the basis of a scoring function and a manually determined "trusted" threshold. HMMs that represent equivalog families in the TIGRFAMs database (Haft *et al*., 2003) are generally built either from the bottom up (by identifying a characterized protein, identifying homologs, and pruning away sequences for which the hypothesis of conserved function is not supported) or from the top down (by identifying a family containing proteins of diverse functions and breaking it up into smaller, functionally uniform pieces, see Figure 1).

Since the density of experimentally characterized sequences is sparse, criteria have been developed for supporting or rejecting the hypothesis of conserved function. These criteria are conservative so as to limit the false application of functional information. TIGRFAMs models do, however, contain a "noise cutoff", which is a second, less stringent threshold value. Sequences that score between

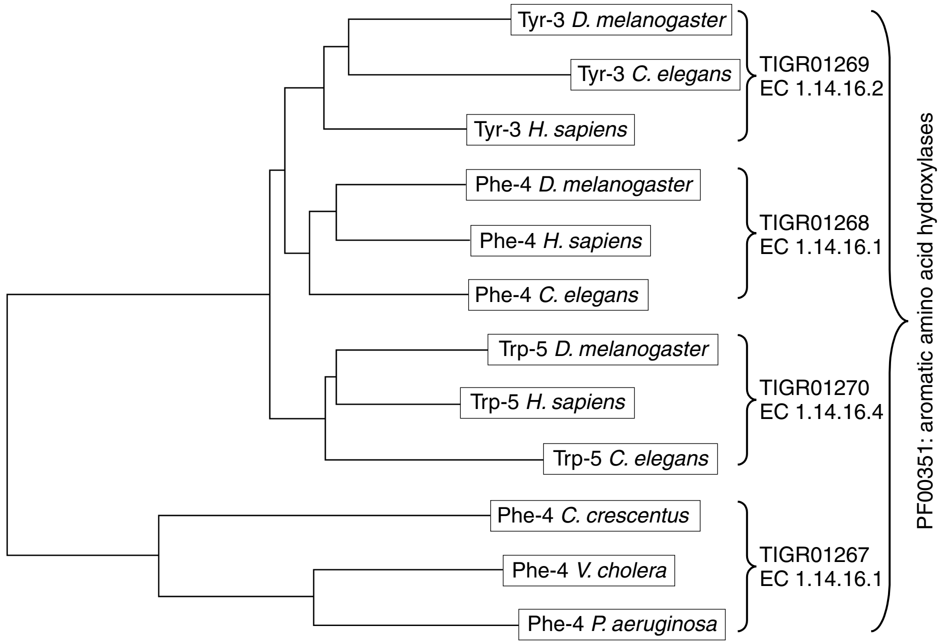


Figure 1 A protein family broken up into equivalogs. The Pfam HMM PF00351 represents a diverse family of aromatic amino acid hydroxylases. A neighbor-joining phylogenetic tree of the family is shown. The position of the root of the tree is supported by alignment versus sequences of greater divergence (not shown). The tree can be broken down into four nonoverlapping pieces representing monomeric eukaryotic tyrosine-3-monoxygenases (Tyr-3), phenylalanine-4-monoxygenases (Phe-4), tryptophan-5-monoxygenases (Trp-5), and tetrameric bacterial phenylalanine-4-monoxygenases (Phe-4). Each piece has been used as the source of sequences in the building of equivalog TIGRFAMs HMMs

the noise and trusted cutoffs will not be annotated automatically, but may be found to have the indicated function based on independent lines of (manually reviewed) evidence. A primary requirement for a TIGRFAMs equivalog model is a properly rooted phylogenetic tree. This is achieved by including in the multiple sequence alignment sequences from a distinct but closely related family. The resulting tree is examined to find the highest branching node encompassing the sequences with experimentally determined functions. The family represented by this highest branching node may be phylogenetically narrow. The family may be broadened (by walking up the nodes of the tree) to include new sequences within certain constraints. No sequence should be included if it has evidence of a different function. With limited exceptions, each species should only be represented once in the family. Supportive evidence such as gene clustering with genes of related function is sought for sequences brought into the family in this way. The tree is also examined to identify sequences with excessively long branch lengths (which may represent functional divergence) to determine if these should be excluded.

Protein families differ from each other in the extent to which function is heterogeneous across the family and in how great the sequence differences are between subsets that perform different functions. Closely related substrates, such as collections of six-carbon sugars, may be acted on by closely related enzymes or

transporters. Proteins from paralogous families in which counts of member proteins differ widely between genomes, characterizations are sparse, and functions often differ between closely related proteins tend to be poor candidates for equivalog model construction. HMMs representing such families are classified as "subfamily", and they identify and provide a warning for cases in which strong sequence similarity easily leads to the error of overly specific annotation.

TIGRFAMs emphasizes models that provide a functionally descriptive name for whole proteins. In contrast, Pfam (*see* Article 86, **Pfam: the protein families database**, Volume 6) (Bateman *et al*., 2004) and SMART (Letunic *et al*., 2004) models tend to mark out larger protein families and make less specific identifications. In some cases, such a model includes proteins with end-to-end homology, but a variety of functions. In other cases, members may share regions of sequence homology, often corresponding to conserved structural domains and referred to as homology domains, while other regions lack homology. Where the emphasis of a Pfam or SMART model may be to extend the sensitivity and selectivity of homology detection beyond that of pairwise BLAST comparisons, TIGRFAMs models often discriminate among rather closely related proteins (Figure 2).

Equivalog HMMs are useful in automatic genome annotation. The information that would be manually attached to each gene by an annotator is instead attached to an equivalog HMM. The HMM is then used to transfer the information automatically to the appropriate genes. This results in not only a significant savings in time but also an improvement in accuracy and data consistency. Groups of equivalog HMMs can be linked together to create computable objects representing proteins that act together in a biological process (such as a metabolic pathway). Such compound objects can be used in metabolic reconstruction and applied to genomic sequence data even prior to annotation. The assignment of process-related annotation information is more labor intensive than assigning molecular functions, requiring an appreciation of the genomic context of the gene in question. TIGR has recently implemented a system to represent such objects (Genome Properties) in part to improve the richness of gene annotation (Haft *et al*., 2004).

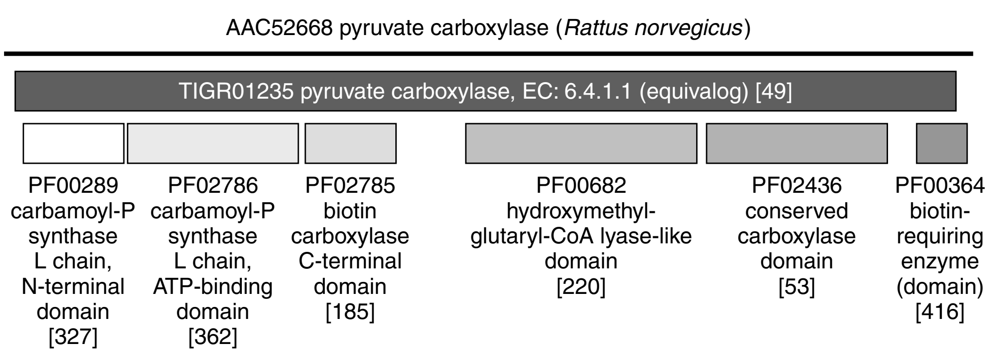


Figure 2 A single TIGRFAMs model, TIGR01235, matches a long protein to which is assigned a particular function, pyruvate decarboxylase. All query sequences that score above the model's trusted cutoff score can be annotated automatically as instances of pyruvate decarboxylase. Meanwhile, six different Pfam HMMs describe nonoverlapping domains, all of which identify additional proteins outside the TIGR01235 family

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6

References

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141. **http://www.sanger.ac.uk/Software/Pfam/**.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**, 99–113.

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT and White O (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, **29**, 41–43. **http://www.tigr.org/TIGRFAMs**.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373. **http://www.tigr.org/TIGRFAMs**.

Haft DH, Selengut JD, Brinkac LM, Zafar N and White O (2004) Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, Sep 3, **http://www.tigr.org/Genome_Properties**.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**, D142–D144. **http://smart.embl-heidelberg.de/**.

Short Specialist Review

The CATH domain structure database

Frances Pearl, Christopher Bennett and Christine Orengo
University College London, London, UK

1. Introduction

The CATH protein structural domain database was established in 1993 (Orengo *et al.*, 1997; Pearl *et al.*, 2003) and made available on the World Wide Web in 1994 (**http://www.biochem.ucl.ac.uk/bsm/cath/**). The database currently contains nearly 60 000 structural domains derived from protein structures deposited in the Protein Data Bank (PDB, Berman *et al.*, 2000).

CATH is a hierarchical database in which structures are classified in a manner that reflects a hierarchy of protein folding (see Figure 1). There are four major levels in the hierarchy. At the top, the **C**lass describes the composition of the protein according to the percentage of residues adopting α-helical or β-strand conformations or a mixture of the two ($\alpha-\beta$). The next level, **A**rchitecture, describes the general shape of the structure given by the arrangement of secondary structures in 3D space. The third, **T**opology level, then considers both the arrangement in 3D and the connectivity of the secondary structures. Finally, the **H**omologous Superfamily groups together domain structures that have clearly evolved from a common ancestor. Within each homologous superfamily, close relatives are grouped into families according to levels of sequence similarity (35, 60, 95, 100%) by single-linkage clustering. A list of nonredundant CATH representatives is generated by selecting a single relative from each family clustered at 35% sequence identity (Sreps).

The protocol for classifying new structures in the CATH domain database involves several steps, some of which are fully automated, while others require some manual validation (see Figure 2). In this chapter, we describe the various algorithms used for classification. We also present a short overview of current populations of architectures, fold groups, and superfamilies.

2. Protocol for classifying protein structures in the CATH domain database

As a first step, each protein chain downloaded from the PDB is deconstructed into its constituent domains. Currently, protein chains from both crystal structures with

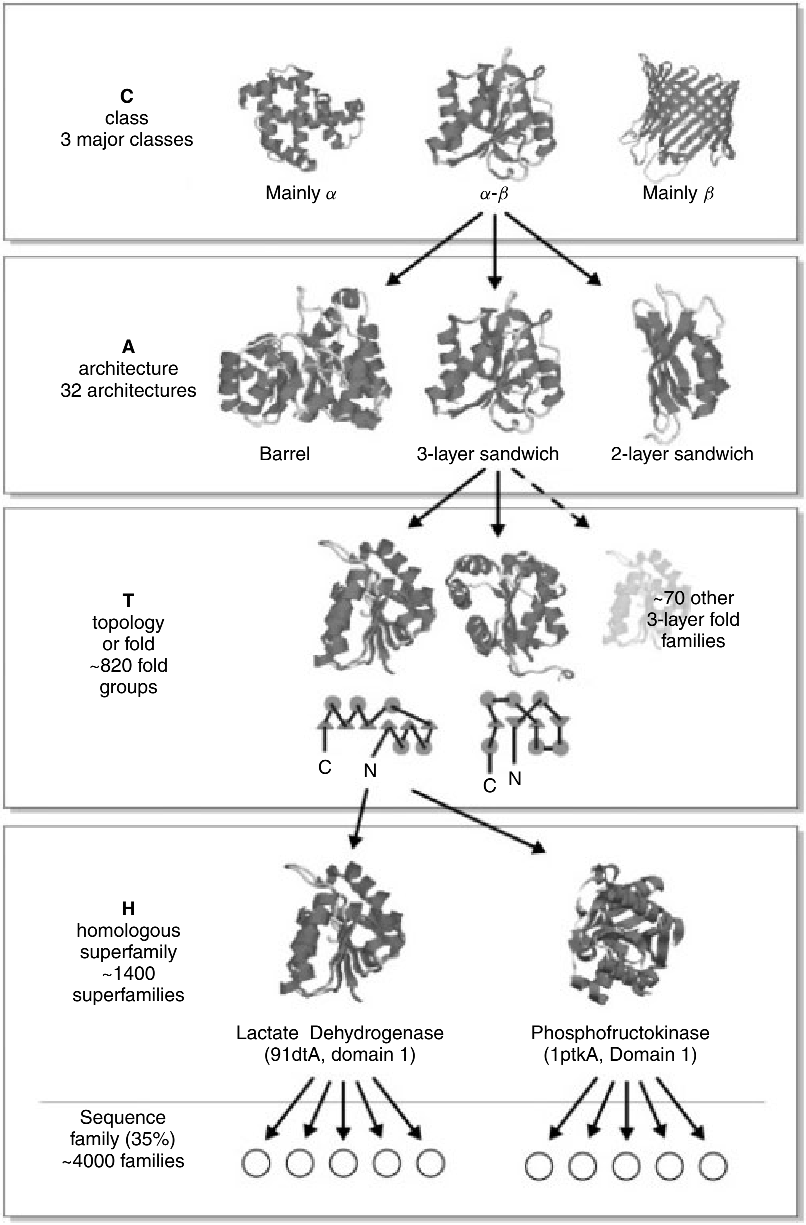


Figure 1 Illustration of the hierarchical nature of the CATH (version 2.5.1) classification scheme. Folds are first assigned to a class contingent on their secondary structure content, and then to architecture that is dependent on their gross geometric shape (e.g., barrel, sandwich, etc.). Each architecture is then divided into fold groups, where each fold has a unique topology. This is dependent on the connectivity between the secondary structures within a fold. Structural domains within a fold group are then clustered into homologous superfamilies, where there is evidence that all members of a superfamily have evolved from a common ancestor. Finally, the homologous superfamilies are grouped into sequence families where all members share sequence similarity > 35% to at least one other member in the family

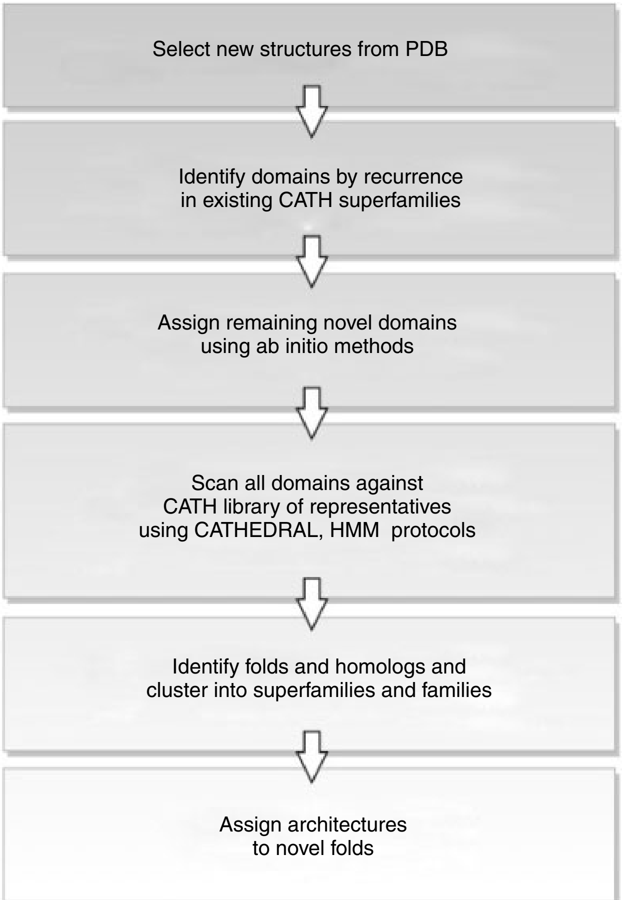


Figure 2 Schematic illustration of the steps required in the classification of protein structural domains

better than 4 Å resolution and NMR structures are considered. Since its inception, CATH has been a domain-based database as domains are thought to be important evolutionary units (*see* Article 82, **Structure comparison and protein structure classifications**, Volume 6). They are much more easily recognized using structural data rather than sequence data; thus, domain structure resources often provide important complementary data to the sequence resources. Structure is also much more highly conserved than sequence and thus the structural data allow more distant evolutionary relationships to be detected than the purely sequence-based resources.

Having determined the domain boundaries of a protein chain, sequence and structure comparison methods are then used to classify the domain into the appropriate fold group and homologous superfamily. Over the last few years, several new methods and protocols have been developed to allow CATH to keep pace with the growing number of structures deposited each year, especially since the onset of the structural genomics initiatives. Currently, 400 new structures are deposited in the

PDB each month. Fortunately, a substantial proportion of these can now be classified using largely automated approaches (see below). The individual steps taken to update CATH, summarized in Figure 2, are discussed in more detail below.

2.1. *Domain boundary assignment*

Two approaches are used for recognizing domain boundaries in multidomain proteins. New structures are first scanned against a library of nonredundant domain folds from CATH in order to recognize any known domain folds CATHEDRAL, see below. This approach exploits the principle of domain recurrence. It has become increasingly clear from genomic analyses that in many protein families domains have been extensively duplicated during evolution (Apic *et al*., 2001) and combined with different domain partners. In a recent CATH release, 75% of protein chains with no significant sequence similarity (sequence identity <35%) to entries in CATH had domains that could be recognized using CATHEDRAL.

Once known folds have been identified and their boundaries refined by manual validation, an ab initio method is used to try to determine whether any remaining regions in the multidomain structure comprise more than one domain.

2.1.1. Methods that exploit domain recurrence

2.1.1.1. Structure-based methods

New PDB structures are scanned against a library of nonredundant structures (Sreps) from the CATH database using a rapid structure comparison method (CATHEDRAL). In parallel, sequences of the new structures are scanned against a library of Hidden Markov Models (HMMs) from representative structural superfamilies.

CATHEDRAL is an acronym that stands for CATH's Existing Domain Recognition Algorithm. The CATHEDRAL suite exploits a fast structure comparison algorithm (GRATH, Harrison *et al*., 2003) that compares secondary structures between proteins using graph theory. CATHEDRAL uses a robust statistical framework based on analyzing the extreme value distribution of scores generated by GRATH'ing a new structure against the CATH fold library, to determine the statistical significance of any fold match.

For each significant match, CATHEDRAL then performs more sensitive residue-based SSAP structure alignment. This is done to obtain an accurate residue-based alignment in order to locate the domain boundary reliably. Matching structures are superposed on the equivalent residues identified by SSAP, using rigid body methods, in order to calculate an root mean squared deviation (RMSD). Several stringent SSAP and RMSD criteria have to be met for a structure match to be confirmed. An iterative process then continues with the remaining regions of the chain scanned against the database until the complete chain has been assigned domains.

In some superfamilies that are highly structurally variable (e.g., the $\alpha\beta$-hydrolases), the SSAP criteria imposed by CATHEDRAL are sometimes too stringent occasionally resulting in a missed similarity. Evidence of a match to

a sensitive sequence profile or HMM can also confirm membership of a particular superfamily. Therefore, all new structures are also scanned against a library of HMMs derived from nonredundant representatives (Sreps) of each superfamily, using the SAM-T99 technology (Karplus and Hu, 2001). Statistically significant matches to any superfamily are identified using the DomainFinder algorithm (Pearl *et al*., 2002) and where necessary, pair-wise SSAPs are run against relatives in that superfamily to obtain the best structural match and locate the boundary more reliably.

2.1.1.2. Ab initio domain assignments

These approaches use structural information alone, such as residue contact data, and compactness measurements. A number of successful algorithms have been developed. Two, DOMAK (Siddiqui and Barton, 1995) and PUU (Holm and Sander, 1994) explore alternative divisions of the protein chain that optimize the number of contacts within domains and minimize contacts between domains. Since internal contacts frequently involve compactly packed hydrophobic residues, other methods (e.g., DETECTIVE, Swindells, 1995) attempt to identify domains by locating hydrophobic clusters, which are then expanded until the packing density falls.

In the CATH classification, the Domain Boundary Suite (DBS; Jones, 1998) integrates these three independent programs (DOMAK, PUU, DETECTIVE) and the resulting boundary assignments are compared automatically. Where all the programs agree, the consensus domain boundaries are accepted automatically. If no consensus is reached, then domain boundaries are derived manually, guided by the predictions.

Using DBS nearly 80% of single domain proteins can be reliably identified. However, as the size and complexity of proteins deposited in the PDB has increased with improvements in technology, the proportion of multidomain proteins that DBS can automatically assign has decreased. Therefore, for those domains with no, partial or conflicting domain assignments, manual intervention is required and the literature is consulted.

2.2. Methods to identify homologs (chains and domains)

As described above, all new structures are scanned against libraries of structural representatives from CATH using CATHEDRAL and their sequences are scanned against HMM libraries to detect significant similarities. Once multidomain proteins have been decomposed into their constituent domains, these are also rescanned against the CATH sequence and structure libraries, to confirm fold and homology assignments. Statistical significance of any match is assessed using the E-value and residue overlap returned by the match. Conservative thresholds are used to minimize the error rate for automatic recognition of homologs and these have been tailored to give the best coverage for any given superfamily.

For remote homologs, the scores often fall below the threshold values and in these cases extensive manual validation is performed to ensure that there is conclusive evidence that the proteins evolved from a common ancestor. This includes detection

of rare sequence and structural motifs and/or conservation of function or of some functional feature; for example, evidence of a conserved chemical mechanism.

2.3. *Recognizing fold similarities and assigning architectures for novel folds*

Any new structure that matches an existing CATH representative but for which there is no evidence of an evolutionary relationship is assigned to the same fold group but as a new superfamily within that fold group. Automatic criteria are used to assign these fold similarities; however, where the empirical and statistical values are borderline, automatic fold assignment is considered unsafe and expert (human) opinion is used. Fold space is better described as a continuum and as recent papers (Harrison *et al*., 2002; Melo *et al*., 2002) have demonstrated, there is often much structural overlap between different folds. Frequently the literature is consulted to resolve these difficult cases.

More interestingly, there are occasionally relatives in structurally variable families that have changed so extensively that they could be considered to have evolved a completely new fold (Grishin, 2001). These relatives are flagged in the CATH relational database (see below) as having an outlying structure and full details on the degree of structural similarity between this structure and other relatives in the superfamily are given in the Dictionary of Homologous Superfamilies (DHS, see below).

New structures that do not match any existing CATH structures are assigned as novel folds and are manually inspected to assign their architecture. There are currently 34 well-described architectures in CATH (5 mainly α, 18 mainly β, and 11 $\alpha-\beta$) and few new architectures have been assigned over the last few years (Pearl and Orengo, 2003).

3. The dictionary of homologous superfamilies (DHS)

The CATH dictionary of homologous superfamilies (DHS, **http://www.biochem.ucl.ac.uk/bsm/dhs**) contains detailed information on each CATH superfamily. For example, sequence and structure similarity scores between relatives in each superfamily and multiple structural alignments for structurally coherent subgroups. These alignments have been annotated with ligand data, where available from the MSD database. Representative plots are presented for each superfamily showing equivalent secondary structures and describing variations in their separations and orientations (2Dsec, EquivSec). Functional descriptions are also provided where they can be extracted from a range of functional databases (EC, Swiss-Prot, GO, COG, KEGG).

4. The CATH relational database and CATH server

All the information on the CATH hierarchy, including domain boundary data, similarity scores within fold groups and superfamilies, is stored in an Oracle 9i

relational database. CATH family data has been integrated with derived structural data (e.g., information on residues in contact with ligands) from the MSD datamart. A reduced version of the data is also available in a Postgres relational database. Web pages are constructed from the database and can be accessed at (**http://www.biochem.ucl.ac.uk/bsm/cath/**). Multiple structure alignments from the DHS are available by ftp from **ftp://ftp.biochem.ucl.ac.uk/pub/cathdata/**. The CATH library of HMMs built from representatives structures (Sreps) is also available from the same ftp site.

New structures can be sent as queries to the CATH server (**http://www.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl**), which scans the structures against the CATH fold library using CATHEDRAL to recognize putative domains and then performs pairwise structure comparison using SSAP on representatives from any fold group matched.

5. CATH structural annotations for completed genomes (Gene3D)

Sequences from completed genomes have been scanned against the CATH HMM library in order to identify sequences or partial sequences that can be reliably assigned to CATH structural superfamilies. Conservative thresholds have been applied and conflicting matches resolved using the DomainFinder algorithm (Pearl *et al*., 2001). All the matches have been stored in the Gene3D relational database and can be viewed over the Web (**http://www.biochem.ucl.ac.uk/bsm/cath/Gene3D/**). Currently, between 40 and 70% of genes in completed genomes can be annotated (20–45% by residue), depending on the genome. A CATH intermediate sequence library has also been generated containing CATH sequence relatives in Swiss-Prot/TrEMBL, which currently contains over 600 000 sequences.

6. CATH statistics

CATH version 2.6 contains 22 303 PDB entries comprising 45 903 chains and 60 435 domains. These have been classified into 1569 superfamilies, 907 fold groups and 34 well defined architectures. However, fold groups and superfamilies are not evenly populated. While most of the fold groups (85%) are very small containing a single superfamily, approximately 50 of the largest fold groups in CATH (7% of the total number of fold groups), comprising 3 or more diverse superfamilies, account for nearly 53% of the nonredundant structures in CATH (Sreps). This power law–like behavior is also mirrored in the population of superfamilies. Figure 3 shows that the top 150 most highly populated superfamilies in CATH account for 50% (49%) of the nonredundant structures.

Similar behaviour is observed with the architectures. Although there are currently 34 well-defined architectures in CATH, six regular architectures (α-bundle, $\alpha-\beta$ barrel, 2- and 3-layer$\alpha-\beta$ sandwiches, β-sandwiches, β-barrels) are more highly populated containing 36% of the fold groups and 56% of the nonredundant structures.

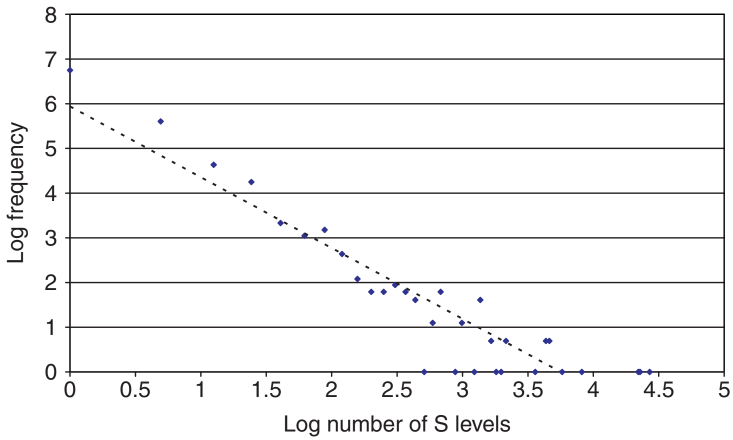


Figure 3 Graph showing the power law–like relationship between the number of sequence families found within the homologous superfamilies

The preponderance of certain architectures and fold groups within the hierarchy may reflect energetically favorable arrangements of secondary structures. For example, the most highly populated architectures all have very regular arrangements of secondary structures, which might be expected to optimize packing of hydrophobic residues in the buried core. Alternatively, it may be the consequence of the extremely divergent evolution of those structural superfamilies that have been extensively duplicated and distributed throughout the genomes (see Apic *et al*., 2001; Lee *et al*., 2003).

In particular, many of the most highly populated fold groups (e.g., TIM barrels, Rossmann folds) are common throughout all kingdoms of life and have clearly been extensively duplicated in the genomes (see Apic *et al*., 2001; Lee *et al*., 2003). Many of them are performing generic functions such as provision of redox equivalents or energy for a catalytic reaction. Paralogs frequently adopt new functions (Todd *et al*., 2001) thus removing some of the constraints on structural change and allowing the protein to diverge beyond the point at which an evolutionary relationship can be easily detected using current methods.

Interestingly, a significant proportion of newly determined structures can be assigned to existing superfamilies and fold groups within CATH, suggesting that we may now have representatives of most of the major protein superfamilies in nature. Clearly, the membrane-bound structures are still an important exception and still underrepresented in the structural databases. A recent analysis by Todd *et al*. (2004) has shown that nearly 80% of newly determined structures, solved within the last year, could be classified in CATH using sophisticated sequence profiles and HMMs. A further 15% could be assigned to structural families and fold groups using structure-based comparison (CATHEDRAL, SSAP). The remaining 5% of structures were novel folds the majority of which could be assigned to existing architectures within the classification.

Further reading

Grant A, Lee D and Orengo C (2004) Progress towards mapping the universe of protein folds. *Genome Biology*, **5**(5), 107.

References

Apic G, Gough J and Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, **310**(2), 311–325.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The protein data bank. *Nucleic Acids Research*, **28**, 235–242.

Grishin NV (2001) Fold change in evolution of protein structures. *Journal of Structural Biology*, **134**(2–3), 167–185.

Harrison A, Pearl F, Mott R, Thornton J and Orengo C (2002) Quantifying the similarities within fold space. *Journal of Molecular Biology*, **323**(5), 909–926.

Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J and Orengo C (2003) Recognizing the fold of a protein structure. *Bioinformatics*, **19**(14), 1748–1759.

Holm L and Sander C (1994) Parser for protein folding units. *Proteins*, **19**(3), 256–268.

Jones S, Stewart M, Michie A, Swindells MB, Orengo C and Thornton JM (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science*, **7**(2), 233–242.

Karplus K and Hu B (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**(8), 713–720.

Lee D, Grant A, Buchan D and Orengo C (2003) A structural perspective on genome evolution. *Current Opinion in Structural Biology*, **13**(3), 359–369.

Melo F, Sanchez R and Sali A (2002) Statistical potentials for fold assessment. *Protein Science*, **11**(2), 430–448.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**(8), 1093–1108.

Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J and Orengo CA (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, **31**(1), 452–455.

Pearl FM, Lee D, Bray JE, Buchan DW, Shepherd AJ and Orengo CA (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Science*, **11**(2), 233–244.

Pearl FM, Martin N, Bray JE, Buchan DW, Harrison AP, Lee D, Reeves GA, Shepherd AJ, Sillitoe I, Todd AE, *et al*. (2001) A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Research*, **29**(1), 223–227.

Pearl FM and Orengo CA (2003) *Protein Structure Classifications. Bioinformatics: Genes, Proteins and Computers*, Orengo C, Jones D and Thornton J (Eds.), BIOS Scientific Publishers, 103–120.

Siddiqui AS and Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, **4**(5), 872–884.

Swindells MB (1995) A procedure for detecting structural domains in proteins. *Protein Science*, **4**(1), 103–112.

Todd AE, Orengo CA and Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *The Journal of Molecular Biology*, **307**(4), 1113–1143.

Todd AE, Marsden RL, Thornton JM and Orengo CA (2004)Progress of structural genomics initiatives: An analysis of solved target structures. *The Journal of Molecular Biology*, submitted.

Short Specialist Review

COGs: an evolutionary classification of genes and proteins from sequenced genomes

Eugene V. Koonin
National Institutes of Health, Bethesda, MD, USA

1. Introduction

The rapid accumulation of genome sequences is both a tremendous opportunity and a major challenge to biologists attempting to extract the maximum functional and evolutionary information from the new genomes (Koonin and Galperin, 2002). With hundreds of genomes being sequenced, each consisting of thousands of genes (Bernal *et al*., 2001), an informational overflow is inevitable unless an approach is developed to classify these genes in such a manner that the number of individual entities to be examined is drastically reduced. Since all genomes are products of evolution, a natural classification of genes should be based on their evolutionary relationships. More specifically, two fundamental concepts of evolutionary biology are central to such classifications: orthology and paralogy, which describe the two distinct types of homologous relationships between genes. Orthologs are homologous genes derived by vertical descent from a single ancestral gene in the last common ancestor of the compared species. Paralogs, in contrast, are homologous genes, which, at some stage of evolution of the respective gene family, have evolved by duplication of an ancestral gene (Fitch, 1970; Fitch, 2000). Orthology and paralogy are intimately linked because, if a duplication(s) occurred after the speciation event that separated the compared species, orthology becomes a relationship between sets of paralogs (co-orthologs), rather than individual genes (Sonnhammer and Koonin, 2002). Disentangling orthologous and paralogous relationships among genes is critical for functional and evolutionary genomics. Orthologs typically occupy the same functional niche in different species, whereas paralogs evolve toward functional diversification. Therefore, robustness of genome annotation depends on accurate identification of orthologs. Similarly, knowing which homologous genes are orthologs and which are paralogs is required for constructing evolutionary scenarios that involve, in addition to vertical inheritance, lineage-specific gene loss and horizontal gene transfer.

2. Clusters of Orthologous Groups of proteins (COGs)

These principles outlined above became the foundation of an evolutionary classification of proteins from sequenced genomes, which was developed at the National Center for Biotechnology Information in 1997 (Tatusov *et al*., 1997). This classification system was named Clusters of Orthologous Groups of proteins, or COGs, to emphasize the fact that a classification unit included not only orthologous proteins but also paralogs resulting from duplications that apparently occurred in one or more species subsequent to their divergence (in-paralogs). The procedure for constructing COGs is based on clustering the matrix of the all-against-all comparison of the proteins encoded in the analyzed genomes; the comparison is performed using the BLAST program (Altschul *et al*., 1997) although, in principle, other methods for sequence comparison can be employed. The clustering procedure was designed to explicitly exploit the notion of orthology: orthologous protein sequences from two species normally are more similar to each other than any of them is to any other sequence from the same genomes. Thus, to form COGs, clusters of mutually consistent, genome-specific best hits (most similar sequences) are identified, after which in-paralogs are added by assigning to COGs the sequences that remained unassigned at the first step. This procedure allowed incorporating into COGs both slowly evolving and fast-evolving proteins, without imposing arbitrary statistical cutoffs on sequence similarity. Additionally, a semiautomatic procedure was developed for handling mutlidomain proteins, in which two (or more) distinct portions of the protein (domain) have different evolutionary histories and belong to different COGs.

The original collection of COGs included six prokaryotic genomes and one genome of a unicellular eukaryote, the yeast *Saccharomyces cerevisiae* (Tatusov *et al*., 1997). Simultaneously with the procedure for COG construction, a method was devised for updating the collection, that is, assigning proteins from newly sequenced genomes to COGs. This method, implemented in the COGNITOR program, is based on the same principle as the original COG-making procedure, that is, genome-specific best hits. Proteins from a new genome that consistently show best hits to proteins from the same COG are included in this COG. The construction of the COGs was followed by manual annotation of the known or predicted functions of each COG by expert biologists. The COG system was updated several times since its inception (Tatusov *et al*., 2000; Tatusov *et al*., 2001; Tatusov *et al*., 2003) and currently includes 63 prokaryotic and 7 eukaryotic genomes (as of January 1, 2004). The procedure for COG construction required that each COG included proteins from at least three sufficiently distant species. This conservative approach notwithstanding, ~60 to ~85% of the proteins encoded in prokaryotic genomes were included in the COGs (Tatusov *et al*., 2003). The coverage of eukaryotic genomes is somewhat lower but exceeds 50% for each included species (Koonin *et al*., 2004). The data on the COG membership of proteins from selected prokaryotes and eukaryotes are shown in Table 1. That the majority of the proteins from each of the analyzed genomes could be included in clusters of orthologs is an important observation in itself because it quantitatively demonstrates the prevalence of evolutionary conservation in genome evolution and emphasizes the utility of comparative genomics in functional and evolutionary studies.

Table 1 Coverage of bacterial, archaeal, and eukaryotic genomes in COGs

Species	Number of annotated proteins	Number (and percentage) of proteins in COGs	Number of COGs that include the given species
Bacteria			
Escherichia coli K12	4279	3623 (85%)	2131
Helicobacter pylori J99	1491	1106 (74%)	921
Bacillus subtilis	4112	3125 (76%)	1771
Mycoplasma genitalium	484	385 (80%)	362
Mycobacterium tuberculosis H37Rv	3927	2843 (72%)	1450
Mycobacterium tuberculosis CDC1551	4187	2756 (66%)	1434
Mycobacterium leprae	1605	1180 (74%)	927
Treponema pallidum	1036	737 (71%)	639
Archaea			
Archaeoglobus fulgidus	2420	1953 (81%)	1244
Methanocaldococcus jannaschii	1758	1448 (82%)	1117
Methanosarcina acetivorans	4540	3142 (69%)	1462
Sulfolobus solfataricus	2977	2207 (74%)	1084
Eukaryota			
Saccharomyces cerevisiae	6338	3012 (48%)	1299
Encephalitozoon cuniculi	1996	1105 (55%)	696

The COG system, accompanied by the COGNITOR program, has become a widely used tool for computational genomics. The most important applications of the COGs are functional annotation of newly sequenced genomes and genome-wide evolutionary analyses. It has been shown that a considerable accuracy of assignment of proteins to COGs (<5% false-positives) can be achieved even through automatic application of COGNITOR (Natale *et al*., 2000). These results can be further improved through relatively straightforward manual intervention. With regard to evolutionary analysis, the COGs immediately provide valuable information on phyletic patterns of genes, that is, the patterns of presence–absence of genes in species and taxa (this can be done on-line using the phyletic pattern search tool). Strikingly, only ∼1% of the COGs are ubiquitous, that is, represented in all analyzed species; the rest are missing in some genomes (Figure 1). In many cases, lineage-specific gene loss, horizontal gene transfer, or a combination thereof best explains the observed phyletic patterns. Even a superficial examination of the phyletic patterns clearly demonstrates the major importance of these processes in evolution, at least in prokaryotes. The phyletic patterns can be used for more sophisticated evolutionary analyses with maximum parsimony and maximum likelihood methods (Mirkin *et al*., 2003; Snel *et al*., 2002). The COGs also present excellent starting material for genome-wide construction of phylogenetic trees and other evolutionary reconstructions (Wolf *et al*., 2002).

The COGs face the major challenge of keeping up with the ever-increasing rate of genome sequencing. If this challenge is met, the utility of the system for

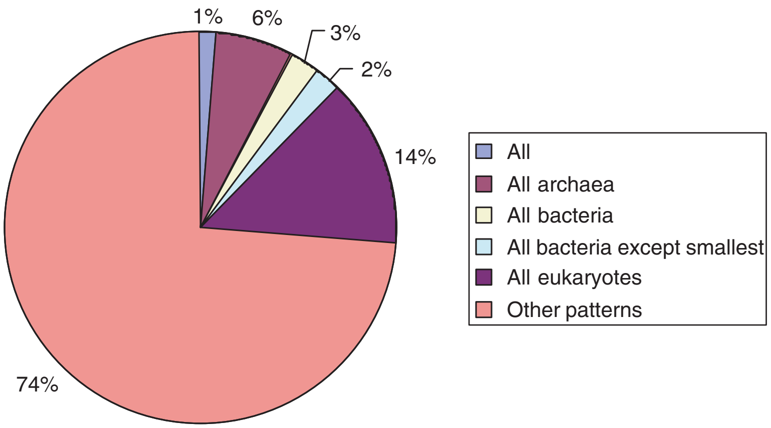


Figure 1 Phyletic patterns of COGs. *All*, represented in all unicellular organisms included in the COG system; *All archaea, All bacteria, All eukaryotes*, represented in each species from the respective domain of life (and possibly in some species from other domains); *All bacteria except the smallest*, represented in all bacteria except, possibly, parasites with small genomes (mycoplasma, chlamydia, rickettsia, and spirochetes)

both functional annotation of genomes and large-scale evolutionary analyses should increase progressively with the inclusion of new species from diverse taxa.

The COGs are accessible at **http://www.ncbi.nlm.nih.gov/COG/** and **ftp://ftp.ncbi.nih.gov/pub/COG/**.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bernal A, Ear U and Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, **29**, 126–127.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**, 99–106.

Fitch WM (2000) Homology a personal view on some of the problems. *Trends in Genetics*, **16**, 227–231.

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, **5**, R7.

Koonin EV and Galperin MY (2002) *Sequence – Evolution – Function. Computational Approaches in Comparative Genomics*, Kluwer Academic Publishers: New York.

Mirkin BG, Fenner TI, Galperin MY and Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, **3**, 2.

Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L and Koonin EV (2000) Genome annotation using clusters of orthologous groups of proteins (COGs) – towards understanding the first genome of a Crenarchaeon. *Genome Biology*, **5**, RESEARCH0009.

Snel B, Bork P and Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, **12**, 17–25.

Sonnhammer EL and Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**, 619–620.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, **28**, 33–36.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, **29**, 22–28.

Wolf YI, Rogozin IB, Grishin NV and Koonin EV (2002) Genome trees and the tree of life. *Trends in Genetics*, **18**, 472–479.

Short Specialist Review

PANTHER: Protein families and subfamilies modeled on the divergence of function

Paul D. Thomas
Director, Evolutionary Systems Biology Group, SRI International, Menlo Park, CA, US

1. Overview

Related protein sequences have diverged both in sequence and in function over evolutionary time. The aim of Protein Analysis Through Evolutionary Relationships (PANTHER) is to group protein sequences into subfamilies of shared function. This aim is accomplished by first grouping sequences into families, and then, if there are proteins with different functions within the family, dividing the family into subfamilies using both computational tools and human expertise. PANTHER therefore allows detailed analysis of the relationship between sequence evolution and function evolution across a large number of protein families. The main applications of PANTHER have been in the following areas: annotation of protein function for genome projects (Venter *et al*., 2001; Mi *et al*., 2003); high-level overview and browsing of whole genomes by function and sequence relationships (Kerlavage *et al*., 2002; Mi *et al*., 2005); analysis of genome-wide experiments, such as mRNA and protein expression data, to find evidence of coordinated activity of particular biological processes and pathways (Clark *et al*., 2003; Freeman *et al*., 2005); and analysis of genetic mutations or variations (single nucleotide polymorphisms, or SNPs) that cause an amino acid substitution in a particular protein (Thomas and Kejariwal, 2004; Brunham *et al*., 2005). PANTHER is freely available at **http://www.pantherdb.org**.

Introduction to protein families, orthologs, paralogs, and functional divergence. The function of a protein molecule is determined by its amino acid sequence. Most known proteins, either from the same organism or from different organisms, can be grouped into "families" of sequences that are recognizably similar to each other, despite millions, and in some cases billions, of years since their divergence from a common ancestral sequence. This sequence similarity persists because some parts of the sequence cannot be changed without lowering the fitness of the corresponding gene, usually by impairing the function of the protein. These essential parts of the protein are therefore "conserved" during evolution.

Protein families have been built up over time by two processes: *copying* of an ancestral sequence, and then *divergence* from that sequence. In general,

there are two ways to generate a copy of a sequence. The first is speciation, in which two genes in two different species descend from the same single gene in the ancestral species (resulting in orthologous genes). The second is gene duplication, in which a gene is copied within a gamete during meiosis, and this duplication then spreads throughout the population, usually due to selective advantage (resulting in paralogous genes). In many cases, orthologous proteins retain the ancestral function, so most of the sequence changes that accumulate over evolutionary time have little effect on the protein's function (i.e. they are selectively neutral). For duplicated genes, on the other hand, because there are now two copies of the same gene, one copy is now freed of prior functional constraints and can therefore take on new or modified functions through mutations in its sequence. As a paralog adapts to its new function, it will tend to change in sequence relatively rapidly (a larger proportion of mutations will be beneficial). Given enough time it will thus form a distinct subfamily that is clearly separated from the ancestral subfamily. The sequence changes that were essential for the new or modified function will be conserved within the subfamily but differ from other subfamilies. The presence of distinct subfamilies of protein sequences, then, allows inference of a *divergence in function* between these subfamilies.

These relationships can be represented as a tree such as the one in Figure 1, where branch lengths represent sequence divergence from the inferred common ancestor. Figure 1 shows a portion of a larger tree of G protein coupled receptors (GPCRs) with two related but distinct functional subfamilies, Dopamine D1 Receptor (D1R) and Serotonin 4 Receptor (5-HT4). While both D1R and 5-HT4 proteins signal through the large G protein signaling pathway, they do so in response to very different extracellular ligands (dopamine vs. serotonin), resulting in a very different physiological response. D1R is found in both vertebrates and invertebrates (and therefore predates the protostome-deuterostome divergence), while 5-HT4 is found only in vertebrates. One possible evolutionary history is that the ancestral D1R gene was duplicated in the vertebrate lineage. One copy then diverged in function, and therefore in sequence, to recognize serotonin and become the 5-HT4 gene. The

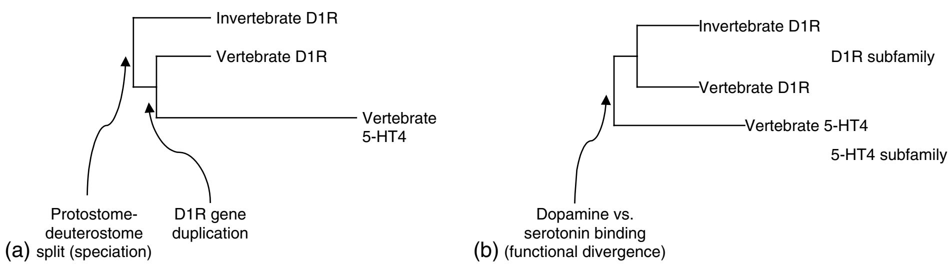


Figure 1 Phylogenetic tree (a) compared to sequence similarity tree (b) for dopamine D1 receptors (D1R) and serotonin type 4 receptors (5-HT4). The tree was built from protein sequences from a number of different invertebrate and vertebrate organisms (see **http://www.pantherdb.org/panther/family.do?clsAccession=PTHR19266** for complete data set) and is summarized here omitting the specific sequences in each group. While the phylogenetic tree represents evolutionary events, the sequence similarity tree represents the selective constraints on different subfamilies, which serves as a basis for correlating sequence with function

vertebrate D1R gene retained dopamine binding and was therefore constrained to conserve the amino acids responsible for this specificity. Assuming this history, a "true" phylogenetic tree would show two copying events: first a speciation (the protostome-deuterostome split), then a gene duplication in the vertebrate lineage followed by rapid sequence evolution along the 5-HT4 branch. The resulting tree (Figure 1A) does not group the sequences by function. However, if we build the tree using sequence similarities instead (Figure 1B), vertebrate D1R appears in the same subtree as invertebrate D1R, while 5-HT4 proteins from different vertebrates form a separate subfamily where all members share more sequence similarity between themselves than they do with the D1R subfamily. For representing sequence-function relationships, a sequence similarity tree is therefore often more useful than a true "phylogenetic" tree.

PANTHER aims to represent these subfamily relationships for all protein families, and currently has high coverage of known metazoan and yeast proteins. For each of over 5000 protein families, a sequence similarity tree is generated and divided by expert biologists into functional subfamilies, modeling the relationships between sequence and function on a large scale.

2. PANTHER: From protein sequences, to subfamilies annotated by function, process and pathway

The entire process for constructing PANTHER families and subfamilies has been described in detail (Thomas *et al*., 2003; Mi *et al*., 2005), and is summarized briefly here. The first step is to cluster proteins into *families* according to their sequence relationships. Our definition of a protein family is designed to facilitate the goal of finding functional subfamilies: a family should be as large as possible, but the sequences should be similar enough to build an accurate multiple sequence alignment. A multiple sequence alignment specifies "equivalent" positions in different proteins (i.e. deriving from a common amino acid in the ancestral sequence for all aligned sequences), so accurate alignment is critical for modeling of molecular evolutionary events.

The second step is to divide protein families into *subfamilies* based on divergence in sequence and function. To accomplish this, we first build sequence similarity trees for each family, and then annotate each protein in the tree with information about its function, species, and protein domains. A biologist curator then reviews the tree structure and protein annotations, and "cuts" the tree into subtrees (subfamilies) that result from inferred functional divergence events. Subfamilies often correspond to groups of orthologous sequences, but we allow the curator to define subfamilies on a case-by-case basis since their inferences will depend on many different factors, such as the details of the biology, the evolutionary trajectory, and the current state of knowledge about the given proteins.

The expert curator then associates each subfamily with controlled vocabulary (ontology) terms that describe the molecular function(s) of the proteins in the subfamily and the biological processes they are known or inferred to participate in. This is important because it provides a way for biologists to perform genome-wide analyses, by looking at a relatively small number of biologically relevant

categories rather than, for example, each of the more than 20 000 distinct human protein-coding genes. The PANTHER ontology terms are a selected subset of the terms available in the Gene Ontology (Gene Ontology Consortium, 2000). In addition to ontology terms, PANTHER also associates protein sequences with a growing number of biological pathways, including both signaling and metabolic pathways. For these proteins, expert curators capture in detail catalysis reactions and protein–protein interactions, using the process notation of Kitano (2003).

Finally, the sequences in each group (both families and subfamilies) are used to build statistical models that represent the sequence diversity in each group. The statistical modeling follows the Hidden Markov Model (HMM) formalism developed by Haussler (Krogh *et al*., 1994). In short, an HMM provides, for each aligned position in a group of sequences, the probabilities of observing each of the 20 amino acids at that position, as well as the probabilities that the position will either be deleted or harbor insertions relative to other sequences in the family. The resulting "library" of family and subfamily HMMs can classify new protein sequences according to the best family and subfamily matches, and can also estimate the importance of a given position for function, based on a quantitative measure of evolutionary conservation (Thomas and Kejariwal, 2004).

References

Brunham LR, Singaraja RR, Pape TD, Kejariwal A, Thomas PD and Hayden MR (2005) Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. *PLoS Genetics*, **1**, e83.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MJ, Tanenbaum DM, Civello D, Lu F, Murphy B, *et al*. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous trios. *Science*, **302**, 1960–1963.

Freeman WM, Brebner K, Amara SG, Reed MS, Pohl J and Phillips AG (2005) Distinct proteomic profiles of amphetamine self-administration transitional states. *The Pharmacogenomics Journal*, **5**, 203–214.

Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, Mayberry F, Mural R, Nodell M, Yandell M, Zhang J, *et al*. (2002) The Celera Discovery System. *Nucleic Acids Research*, **30**, 129–136.

Kitano H (2003) A graphical notation for biochemical networks. *Biosilico*, **1**, 169–176.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, **235**, 1501–1531.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284–D288.

Mi H, Vandergriff J, Campbell M, Narechania A, Majoros W, Lewis S, Thomas PD and Ashburner M (2003) Assessment of genome-wide protein function classification for drosophila melanogaster. *Genome Research*, **13**, 2118–2128.

Thomas PD, Campbell MC, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A and Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research*, **13**, 2129–2141.

Thomas PD and Kejariwal A (2004) Coding single nucleotide polymorphisms associated with complex vs. mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15398–15403.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Short Specialist Review

Functional prediction through phylogenetic inference and structural classification of proteins

Kimmen Sjölander and Chelsea Specht
University of California, Berkeley, CA, US

1. Functional classification of genes is a pressing need in the genome era

With tens of thousands of new genes being identified monthly, experimental determination of gene function for all new genes is not possible. Thus, computational prediction of gene function is an essential tool for modern biologists. The primary method of gene functional annotation employs transfer of annotation from the top hit in a database search. Because homology-based methods of function prediction have been shown to be prone to systematic errors (Galperin and Koonin, 1998; Koski and Golding, 2001), methods in recent years have focused on inclusion of information from evolutionary studies and structural analysis. In this paper, we present a perspective on the use of phylogenetic and structural analyses for improving the inference of protein function.

2. Structural and functional divergence in protein superfamilies

Evolutionary processes, particularly gene duplication and domain shuffling, produce protein superfamilies that often span a bewildering variety of functions and exhibit extreme structural and sequence divergence. Evolutionary pressures to conserve function (e.g., binding pocket positions that determine enzyme substrate specificity or receptor–ligand interaction) or structure (e.g., disulfide bridges in secreted proteins) can be relieved as gene duplication events permit the family as a whole to innovate novel functions and structures; one copy maintains the original function (with consistent constraints at key sites) while the other is enabled to achieve novel function (termed *neofunctionalization*) or to specialize for different tissue types or temporal needs (termed *subfunctionalization*). If not considered appropriately, the

resulting sequence divergence can cause complications in determining homology and interpreting phylogenetic signal. If the underlying assumptions of phylogenetic analyses are taken into consideration, however, they can provide powerful tools for inferring protein function.

3. Phylogenetics and protein superfamilies

The first assumption of a phylogenetic reconstruction is that the data being analyzed are *homologous*, that is, related by evolution. One of the major challenges in phylogenetic reconstruction of protein superfamilies is that not all positions are homologous across all family members; some regions of proteins may have been inserted or deleted over time (e.g., insertions at surface loops). These positions may not be easily distinguishable from distantly related regions that are homologous, albeit with low sequence similarity. It is critical to differentiate homologous and nonhomologous regions at the outset, perhaps by alignment masking protocols, else the phylogenetic analysis may be adversely affected. This remains a challenge in practice.

Other issues in phylogenetic analysis of protein superfamilies arise due to lineage-specific or site-specific rate variation. While most phylogenetic tree inference methods may allow site-specific rate variation, few attempt to model different types of conservation within different lineages at individual positions. Site-specific mutation rates can vary across sites within a lineage, and can be entirely different from rates observed in other lineages. Both lineage and site specificity are considered independently (Muse and Gaut, 1994; Yang, 1998; Yang and Nielsen, 1998; Nielsen and Yang 1998; Yang *et al*., 2000) and together (Yang and Nielsen, 2002) in a variety of models. These models may be appropriate when some positions are conserved within a particular lineage (e.g., substrate specificity determining residues) but vary across lineages. In other cases, residues observed to interact in the buried core of proteins often exhibit correlated or compensatory mutations. If appropriately incorporated in a phylogenetic analysis, these types of conservation can be used to help determine phylogenetic relatedness.

For example, G-protein coupled receptors (GPCRs) span dozens of different subtypes developed by multiple gene duplication events. Small modifications at key positions determine the ligands recognized by these receptors, for example, serotonin, dopamine, histamine, acetylcholine, chemical odorant signals, chemokines, and opiates. Residues at these positions determine the ligand-specificity of each protein, and thereby its molecular function. Phylogenetic methods that can identify these positions *a priori* could upweight these positions in tree topology estimation. The SATCHMO (Edgar and Sjolander, 2003) algorithm is an example of this type of phylogenetic inference approach.

Phylogenies, or evolutionary trees, are the fundamental structures that enable us to visualize evolution in a hierarchical and comparative framework. Molecular phylogenetic methods have been primarily developed for and validated with respect to inferring the evolutionary history of organisms and gene families based on DNA sequence data. When protein evolution has been examined, it has been in the context of a single group of orthologous genes. In recent years, the challenges of inferring

molecular function for hundreds of thousands of unknown genes have driven the application of phylogenetic approaches to elucidating the evolutionary history of protein superfamilies. Within a family of proteins, sequences with known functions can be grouped with proteins of unknown function in a phylogenetic context, and proteins that group together phylogenetically may be predicted to share a common function. This approach is called *phylogenomic inference of protein function* (Brown and Sjölander, 2006).

4. The objective of protein superfamily phylogenetic analysis

Unlike the primary objective of organismal phylogenetic analysis – elucidating the evolutionary history of an organism or lineage of organisms – the primary objective for most phylogenetic analyses of protein families is protein function prediction. Phylogenomic inference of protein function – the analysis of a protein sequence in the context of a larger protein family with potentially numerous functional subtypes – is a prime example of this type of analysis.

Phylogenomic inference of protein function has several distinct challenges. First, accurate phylogenomic inference depends on homologs with the same domain structure; this is not normally known for sequences found in database search. Most database searches return sequences that may have only local homologs (i.e., have a different domain architecture from the seed). (Brown and Sjölander, 2006). Second, many protein families span large evolutionary distances, resulting in significant structural divergence among homologs and reduced alignment accuracy (Baker and Sali, 2001). While closely related sequences (e.g., with pairwise identities $>50\%$) are normally clustered into subtrees consistently by most phylogenetic tree methods, the branching order between these conserved clades can be very different from one phylogenetic tree method to the next. This is particularly problematic if function inference is based on the branching order observed in a single phylogenetic tree, and other phylogenetic trees for the same dataset are not examined.

In contrast to the issues involved in reconstructing phylogenies for groups of single orthologous genes, phylogenetic reconstruction of protein superfamilies has a greater degree of uncertainty and variability, due to choices made in gathering homologs, constructing a multiple sequence alignment (MSA), alignment masking, choice of phylogenetic method, and phylogenetic tree interpretation. We discuss each of these points in turn.

4.1. Selecting homologs

Accurate phylogenetic reconstruction requires adequate sampling of all the major evolutionary subgroups. Engineered sequences should be excluded, and sequences should be filtered to remove those not sharing the same overall domain structure. The Berkeley Phylogenomics Group FlowerPower server enables selection of

sequences that can be predicted to share the same domain structure as an input seed sequence.

4.2. *Constructing a multiple sequence alignment*

Phylogenetic analysis and alignment construction are integrally related: most phylogenetic analyses are based on input MSAs, which encode the primary source of evolutionary signal upon which the phylogenetic tree construction method operates. The MSA represents the primary statement of homology, such that each column of DNA sequence or amino acid data is considered to be derived from a common ancestor as a single evolutionary event. All methods for alignment therefore inherently make inferences about evolution, and errors in the input alignment can be positively misleading to the phylogenetic analysis.

Progressive alignment methods, such as ClustalW (Thompson *et al*., 1994), construct a "guide tree" developed by distance-based methods to determine the order in which sequences are aligned; alignments are fixed within subgroups as sequences are progressively aligned. Progressive methods are among the most popular, but are primarily appropriate for closely related sequences with few required *indel* characters. Assessment of sequence alignment accuracy relative to structural superposition of 3D solved protein structures shows alignment accuracy drops sharply for pairs of sequences with <30% identity. Datasets including such highly divergent sequences will require more advanced and computationally intensive alignment methods. More recently, iterative methods have been developed that allow sequences to realign and thereby improve the overall accuracy (e.g., MUSCLE (Edgar, 2004), ProbCons (Do *et al*., 2005), and MAFFT (Katoh *et al*., 2005).

4.3. *Masking protocols*

The general consensus among computational biologists is that masking protocols ought to be employed, so that ambiguously aligned columns, or regions of questionable homology, are excluded from the phylogenetic analysis. The relative pros and cons of different masking protocols are not well understood.

4.4. *Selection of phylogenetic tree method*

There are two main classes of molecular phylogenetic analysis: character-based (e.g., maximum parsimony (MP), maximum likelihood (ML) and Bayesian approaches) and distance methods (e.g., neighbor-joining). Distance methods work by estimating a matrix of pairwise distances between input sequence data. These distances are then used to construct a tree topology that is most consistent with all observed pairwise distances. Character-based approaches, by contrast, retain the character state information in estimating a phylogenetic tree that is based on some optimality criterion (i.e., MP searches for the shortest tree consistent with the data, and ML methods attempt to find the most likely tree given the data and some model

of evolution). Distance methods have the primary advantage of being computationally efficient, and are therefore often used by biologists estimating phylogenies for large datasets. However, they are not believed to be as accurate as character-based approaches in estimating tree topology and branch lengths (Yang, 1994), both of which are essential to accurate functional inference by phylogenetic analysis.

Despite inherent controversy, recent years have seen Bayesian methods making a large contribution to phylogenetic inference (Yang and Rannala, 1997; Li *et al*., 2000; Huelsenbeck *et al*., 2001). Controversy surrounds the selection of a prior and its overall contribution to the resulting tree topology; however, the use of Markov-chain Monte Carlo algorithms to rapidly explore tree space makes model-based phylogenetic reconstruction and determination of statistical topological support attainable, even for large and variable datasets. The ability of Bayesian methods to elucidate the evolutionary history of protein superfamilies has not been fully explored, but we believe this class of phylogenetic method has great potential in this area due to the probabilistic nature of the analyses combined with computational speed.

4.5. Assessing phylogenetic method accuracy

The accuracy of phylogenetic methods has traditionally been assessed by simulation studies. In a simulation study, the parameters of a phylogenetic model are predetermined (forming the "true" tree topology), and data is generated from those parameters. These data are then used as input to a phyologenetic tree estimation, and the estimated tree is compared to the "true" tree. The sensitivity of a phylogenetic method to violations of assumptions can then be assessed. Simulation studies have demonstrated the sensitivity of both distance and character-based methods to variations in evolutionary rates across taxa, to site-specific mutation rates and to shifts in evolutionary rates over time (see Swofford *et al*., 1996; Felsenstein, 2004).

4.6. How informative are these simulation studies to phylogenetic reconstruction of protein superfamilies, especially in cases of significant structural divergence due to gene duplication?

Structural studies have shown that as sequence similarity decreases, the degree of structural superposability (of solved 3D structures) also drops. Extremes of structural and sequence divergence is common in protein superfamilies. However, to our knowledge no simulation studies have assessed the impact of the types of evolutionary innovations and dramatic changes observed in multigene families on the accuracy of phylogenetic reconstruction. In fact, most simulation studies and comparisons of methods for phylogeny reconstruction are performed based on two assumptions; that the data being analyzed come from a single orthologous gene family and that the alignment is correct. The first assumption is automatically nullified for protein superfamilies, and studies of alignment accuracy under significant sequence divergence show that the second assumption must also be

assumed to be false. The degree to which errors in the input MSA affect the phylogenetic tree topology accuracy for large protein superfamilies is not known.

4.7. *Using multiple approaches*

Because different tree methods can produce dramatically different tree topologies for the same input MSA, careful attention must be paid to the assumptions made during phylogenetic inference. We recommend applying more than one discrete tree-building method to the data in order to better understand the data and their response to phylogenetic analysis. The wide range of available models provides tools for exploring data while developing a robust phylogenetic hypothesis. If well-supported nodes are in conflict when different tree-building algorithms are used, manipulations of the data (removing characters) and the taxa/proteins (removing taxa) may be necessary to test for underlying conflict in the data set. Davis *et al.* (1998) and Gatesy *et al.* (1999) give reviews of various means of testing for phylogenetic incongruence, data decisiveness, and hidden support in phylogenetic data in a parsimony framework, while Kishino and Hasegawa (1989) review support and conflict in a likelihood framework.

5. Assessing the accuracy of phylogenetic methods and phylogenetic trees

Both parametric and nonparametric means are used for assessing the amount of uncertainty in a phylogenetic tree. The likelihood ratio test (Goldman, 1993) provides a means of the testing assertions about the parameters of the chosen model of evolution, using the ML estimate and comparing it with alternative topologies that cannot be rejected based on their likelihood score. Interior branch tests (Nei *et al.*, 1985; Li, 1989; Tajima, 1992; Rzhetsky and Nei 1992) use the variance on the estimate of the length of a branch in the interior of a tree to determine the reliability of the branch. Both of these are limited by the fact that they are parametric in nature and may underestimate the uncertainty for a given topology. Resampling techniques such as the bootstrap (resampling with replacement: Felsenstein, 1985) and jackknife (resampling without replacement: Mueller and Ayala, 1982; Farris *et al.*, 1996), decay indices (Bremer, 1992; Baker *et al.*, 1998) and data decisiveness tests (Davis *et al.*, 1998) use empirical information about character variation to infer confidence in tree topology. In Bayesian analyses, support for relationships is determined by posterior probabilities of the identified nodes. These values tend to be inflated relative to bootstrap/jackknife values and may not be directly correlated with respect to indication of nodal support (Douady *et al.*, 2003).

6. Computational efficiency

The complexity of protein superfamily evolution begs for more complex models of molecular evolution. Unfortunately, most protein families have large numbers

of taxa (in the hundreds or thousands) as well as significant sequence divergence (many pairs with $<15\%$ identity) while the sequence length is relatively short (at most a few hundred residues in length). This leaves very little actual information to estimate complex phylogenetic models. In addition, model complexity increases the risk of overparameterization, leading to decreased support for any phylogenetic hypothesis.

Many of the phylogenetic questions typically addressed by functional and structural biologists require supercomputers or clusters to run the algorithms necessary to adequately align the data and construct a reliable phylogenetic tree in a reasonable amount of time. Bayesian methods promise to reduce the time necessary for a model-based approach using Markov-chain Monte Carlo simulations of character space.

While these evolutionary reconstructions of protein superfamilies are undoubtly computationally demanding, we believe that the need for function prediction accuracy warrants the time and resources spent on phylogenetic reconstruction. The field as a whole will benefit from simulation studies specifically designed to assess the expected accuracy of existing methods for protein superfamily reconstruction.

7. The role of structure prediction and analysis in protein function prediction

Structural analysis efforts and structural classification databases often have a similar intent: enabling biologists to infer molecular function on the basis of structural similarities. For example, SCOP and CATH enable biologists to make inferences of protein function based on structural similarity, and the DALI and VAST servers provide on-the-fly identification of structural neighbors for newly solved structures. A variety of webservers enable prediction of structural domains through the use of profile or hidden Markov model approaches (e.g., 3d-pssm, the NCBI Conserved Domain Database, and Superfamily). Several methods that integrate structural and phylogenetic analyses have been successfully used to predict functional epitopes in proteins (Lichtarge *et al*., 1996; Glaser *et al*., 2003). A number of protein sequence databases serve as repositories of vast amounts of data and associated bioinformatics predictions; these can be extremely valuable in inferring function and structure for proteins (e.g., UniProt, InterPro, SMART, and PhyloFacts).

8. Website references

http://phylogenomics.berkeley.edu/UniversalProteome/	PhyloFacts Universal Proteome Explorer
http://phylogenomics.berkeley.edu/resources/	Other Berkeley Phylogenomics Group resources, including FlowerPower, SATCHMO, and SCI-PHY
http://www.pantherdb.org	Celera Panther tools

http://smart.embl-heidelberg.de/	SMART
http://www.ncbi.nlm.nih.gov/	NCBI
http://www.uniprot.org/	UniProt
http://www.ebi.ac.uk/interpro/	InterPro
http://www.sanger.ac.uk/Software/Pfam/	PFAM HMM library at the Sanger Institute
http://www.rcsb.org/pdb/	Protein Data Bank (PDB)
http://scop.berkeley.edu/	Structural Classification of Proteins (SCOP)
http://cathwww.biochem.ucl.ac.uk/	CATH
http://ekhidna.biocenter.helsinki.fi/dali/	DALI/FSP
http://www.sbg.bio.ic.ac.uk/3dpssm/	3d-pssm
http://supfam.org/SUPERFAMILY/	Superfamily

Acknowledgments

This work was supported in part by Grant #0238311 from the National Science Foundation, and by Grant #R01 HG002769-01 from the National Institutes of Health to KS.

References

Baker D and Sali A (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Baker R, Yu XB and DeSalle R (1998) Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. *Molecular Phylogenetics and Evolution*, **9**, 427–436.

Bremer K (1992) Branch support and tree stability. *Cladistics*, **10**, 295–304.

Brown D and Sjölander K (2006) Functional classification using phylogenomics inference. *PLOS Computational Biology*, **2**, e77.

Davis JI, Simmons MP, Stevenson DW and Wendel JF (1998) Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from monocotyledons using mitochondrial atpA sequences. *Systematic Biology*, **47**, 282–310.

Do CB, Mahabhashyam MS, Brudno M and Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF and Douzery EJP (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, **20**, 248–254.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edgar RC and Sjölander K (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.

Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, **12**, 99–124.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein J (2004) *Inferring Phylogenies*, Sinauer Associates: Sunderland, MA.

Galperin MY and Koonin EV (1998) Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biology*, **1**, 55–67.

Gatesy J, O'Grady P and Baker R (1999) Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, **15**, 271–313.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E and Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Goldman M (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Katoh K, Kuma K, Miyata T and Toh H (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform Ser Workshop Genome Inform*, **16**, 22–33.

Kishino H and Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, **29**, 170–179.

Koski LB and Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, **52**, 540–542.

Li S, Pearl DK and Doss H (2000) Phylogenetic tree construction using Markov Chain Monte Carlo. *Journal of the American Statistical Association*, **95**, 493–508.

Li WH (1989) A statistical test of phylogenies estimated from sequence data. *Molecular Biology and Evolution*, **6**, 424–435.

Lichtarge O, Bourne HR and Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, **257**, 342–358.

Mueller LD and Ayala FJ (1982) Estimation and interpretation of genetic distance in empirical studies. *Genetical Research*, **40**, 127–137.

Muse S and Gaut BS (1994) A likelihood method for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, **11**, 715–724.

Nei M, Stephens JC and Saitou N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution*, **2**, 66–85.

Nielsen R and Yang Z (1998) Likelihood models for detecting positively selected amino acids sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Rzhetsky A and Nei M (1992) Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, **35**, 367–375.

Swofford DLGJ, Olsen PJ, Waddell DM Hillis (1996) *Phylogenetic Inference in Molecular Systematics*, Second Edition, Hillis DM, Moritz C and Mabel BK (Eds.), Sinauer Associates: Sunderland, MA.

Tajima F (1992) Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. *Molecular Biology and Evolution*, **9**, 168–181.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Yang Z (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, **43**, 329–342.

Yang Z (1998) On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, **47**, 125–133.

Yang Z and Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, **46**, 409–418.

Yang Z and Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**, 908–917.

Yang Z, Nielsen R, Goldman N and Pedersen A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang Z and Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

Classification of proteins by sequence signatures

Kay Hofmann
Memorec Biotec GmbH, Köln, Germany

1. Advantages of sequence signatures

Whenever a novel protein sequence moves into the research focus, the question of a sequence-based protein classification arises. Typically, the aim of the classification is the prediction of protein properties, based on information already available for better-characterized members of the same class (Hofmann, 1998). A number of different and sometimes even orthogonal classification systems exist. One of them, referred to as "molecular taxonomy," deals with evolutionary relationships of the protein. Other systems classify proteins by their domain content, biological function, subcellular localization, or other protein properties that can be predicted from the sequence.

The traditional way of classifying an uncharacterized protein starts by running a database search for related sequences. In advantageous situations, this search will uncover convincing sequence similarities to a well-characterized protein, maybe from another species, but possibly also a paralog from the same species as the query protein. In this case, the researcher can speculate that some of the important properties may be shared between the two proteins. This relatively simple method has some shortcomings and also offers some pitfalls. Frequently, the result of a similarity search is not as clear-cut, for example, if multiple significant similarities to proteins of different (or even contradicting) functions are reported. (Attwood, 2000; Hofmann, 1998).

Even an apparently clear similarity situation can be misleading: If a query protein turns out to be closely related to 10 different proteins, all well-characterized protein kinases from different organisms, one is tempted to assume that the query is also a kinase. However, there still is the chance that the similarity between the query and the kinases is confined to a sequence region not including the catalytic domain. In this case, the simple transfer of information would be unwarranted. As many database sequences have obtained their annotation records through fully automated procedures, some of them apparently prone to this kind of mistake, current sequence databases abound with misannotated entries. This unfortunate situation further complicates the protein classification by sequence similarity (Bork and Koonin, 1998).

An obvious way of avoiding this kind of problem is to replace (or at least to augment) the simple sequence-to-sequence comparison with a method based on sequence signatures. A sequence signature aims to be a suitable descriptor for a certain functionality, for example, the function of a protein to act as a kinase. Sequence signatures must be able to store all the information necessary to detect whether a given sequence contains the feature represented by the signature.

2. Types of sequence signatures

As the nature of potentially interesting sequence features can be very diverse, ranging from small protein modification sites to large functional homology domains, there also exist a number of different signature types.

2.1. Signatures based on regular expressions

"Regular expressions" is a concept borrowed from text matching that can be readily applied to sequence searches as well. In the biological context, a regular expression can be viewed as an extension of the widely used consensus sequence representation, where some positions require a particular sequence residue (a strict consensus), while other positions have a more relaxed requirement or no requirement at all. Two examples for easy regular expressions of proteins are given below, one describing a simplified version of the N-linked glycosylation motif and the other one a C-terminal motif for ER-localization:

$$\mathrm{N}-x-[\mathrm{ST}] \quad \text{or} \quad \mathrm{K}-x(0,1)-\mathrm{K}-x-x\rangle$$

The first pattern matches a fixed-length sequence region of three residues, the first one requiring an asparagine (N), the second matched by any residue, while the third being either serine (S) or threonine (T). The second pattern adds two powerful features, a variable-length spacer $x(0,1)$, meaning "either 0 or 1 position are matched by anything" and, also, a provision for anchoring a match to the end of a sequence (indicated here by the angular bracket). Regular expressions are particularly well suited for the description of short sequence stretches, such as modification sites and binding- or targeting motifs. This use of regular expressions has been heralded by the PROSITE database (Bairoch, 1993; *see also* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6), whose syntax has been used in the examples above. More recently, the ELM project (for eukaryotic linear motifs) has started to use regular expressions, augmented by other types of information (Puntervoll *et al*., 2003).

In order to be useful for prediction purposes, a sequence signature must have a certain minimal information content. If very short sequence regions are described, a large proportion of strongly conserved residues is required; otherwise, the pattern would be expected to match many database sequences by chance alone. Many motifs of biological importance are too weakly defined to be useful for database

searches, among them are the glycosylation motif mentioned above and most recognition sites for protein kinases. Even if the information content is too low to tell a relevant sequence from an irrelevant one, those regular expressions can still be useful to find a particular site within a given sequence. Biological motifs largely devoid of invariant positions, but rather defined by a larger region of weak positional preferences, are more easily described by other types of sequence signatures.

2.2. Signatures based on weight matrices

While regular expressions always give a clear yes/no decision, most other sequence signatures use a scoring approach instead, where the score indicates how close the sequence under study comes to the "ideal" motif. If a binary decision is required, those quantitative descriptors must be combined with a threshold score, which indicates whether the match is good enough to be considered real.

The "weight matrix" is a simple version of a quantitative sequence signature (Staden, 1984). A weight matrix is a two-dimensional table that contains scores for each possible residue at each position. In the simplest case, these scores can be obtained by counting how frequently a particular residue is observed at a particular position while analyzing a set of well-established occurrences of the motif. Once a weight matrix is "trained," it can be used to score query sequences. The total score assigned to a sequence or a sequence fragment typically corresponds to the sum of the residue scores over all positions of the weight matrix.

As weight matrices make no provisions for sequence insertions or deletions, they are most useful for describing short- to medium-sized sequence motifs without gaps. In particular, when describing DNA motifs such as transcription factor binding sites, weight matrices have been used very successfully (Stormo, 1988).

2.3. Signatures based on profiles

Sequence profiles, an extension of the original weight matrix concept (Gribskov *et al*., 1987), are more suitable for the description of longer protein motifs or domains, which may contain insertions and deletions. As the weight matrix, a profile is a two-dimensional table of residue scores, and the overall quality of a sequence aligned to a profile is calculated as the sum of all residue scores occurring in that alignment. There are a number of important differences though.

As profiles are geared toward protein comparisons, the residue scores are not just the corresponding residue frequencies but are derived from those frequencies by application of a substitution matrix. As an advantage of this procedure, a profile also contains meaningful score values for residues that have never been observed in the set of "training sequences" used for profile construction. Most useful substitution matrices contain both positive values (for identical residues or conservative substitutions) and negative values (for all other substitutions), with a distinctly negative average. Thus, like a two-sequence comparison by the Smith–Waterman algorithm (Smith and Waterman, 1981), a sequence-to-profile comparison can be run in local mode, that is, not requiring all of the profile to match to the sequence.

Another feature borrowed from Smith–Waterman alignments is the treatment of sequence gaps by applying two specific penalty scores, a more expensive "gap creation penalty" and a cheaper penalty for extending an already existing gap by one residue. In classical two-sequence comparisons, those penalties are uniform throughout the alignment. By contrast, profiles can store different sets of penalties for each position. As profiles are typically derived from preexisting multiple alignments of related sequences, gap penalties can be lowered at profile positions where insertions or deletions are known to occur; they can be raised in profile regions where members of the relevant sequence class are known to align without gaps.

Sequence profiles can be seen as a hybrid between weight matrices and Smith–Waterman sequence comparisons, combining the best of both worlds. The sequence-to-profile score is calculated by a modified version of the dynamic programming algorithm widely used for sequence-to-sequence comparisons. There are a number of different profile methods being used. The original version introduced by Gribskov in 1987 (Gribskov *et al*., 1987) had some restrictions, which are alleviated in the more versatile "generalized profiles" of Bucher *et al*. (1996).

2.4. *Signatures based on HMMs*

In 1994, a particular class of linear Hidden Markov Model (HMM) was introduced as a sequence motif descriptor (Baldi *et al*., 1994; Krogh *et al*., 1994). Conceptually, HMMs are very different from the simpler signature types described above. However, the resulting motif or domain descriptor behaves very similar to a sequence profile. It has even been shown that the most widely used class of HMMs can be transformed into a generalized profile (and back) without any loss of information (Bucher *et al*., 1996). This type of Hidden Markov Model is thus often referred to as "profile-HMM" and forms the basis of several very useful protein domain databases (Eddy, 1998).

An exhaustive explanation of the HMM concept would be beyond the scope of this chapter, but there is an extensive literature on that subject, for example, Eddy (1998). In brief, a profile-HMM can be understood as a kind of "machine" that generates sequences. This machine has a large number of "dials," which must be fine-tuned so that the generated sequences are typical representatives of a given motif class. There are several kinds of components within this machine, some of them having the task of adding a further residue to the sequence (with a particular "dial" adjusting the probability of what residue to create). Other components make decisions whether to switch into an insertion or deletion mode, again associated with a specific parameter indicating the probabilities.

All of the HMM parameters are fine-tuned in a probabilistic fashion by training the HMM with a given set of known instances of the domain or motif to be described. When scoring a query sequence against a trained HMM, the direct result is a probability, indicating how likely it is that this particular sequence would have been generated by the HMM. The absolute value of this probability is meaningless but is typically compared to the probability that this sequence is

generated by a background-HMM, trained to produce a random sequence of the same characteristics as the query sequence. It is the ratio of those probabilities that decides whether a query sequence should be considered as containing the motif described by the trained HMM. The range of applications of profile-HMMs is similar to that of sequence profiles, that is, the description of medium-sized or longer protein domains that may contain gaps.

2.5. Other signatures

In addition to the sequence signatures described above, there are a number of hybrid forms being used, and there also exist some signature types addressing very specific tasks. A protein "block," as introduced by Henikoff and Henikoff (1991), is a particular type of weight matrix that uses the same substitution matrix concept as profiles but does not allow for insertions or deletions. A protein "fingerprint," introduced by Attwood and Beck (1994), is an extension of the "block" approach, where the simultaneous occurrence of multiple blocks in a single sequence is scored, neglecting the distance between the blocks. A signature type specifically addressing the task of defining RNA-motifs is the "stochastic context-free grammar" (SCFG), in particular, its implementation as a "covariance model," which forms the basis of an RNA motif database (Eddy, 2002).

3. Making use of sequence signatures

Sequence signatures have been very successfully applied as comprehensive descriptors of protein motifs and domains. Searching matches to a query sequence in a database of sequence signatures typically yields results that are far easier to interpret than those from a sequence database search. In addition, the more sophisticated signatures, such as generalized profiles and HMMs, are also more sensitive than sequence searches and can detect sequence relationships at very long evolutionary distances (Hofmann, 2000). The main reason for the enhanced sensitivity lies in the additional information on the relative importance of the motif residues that can be stored in a sequence signature. A sequence-to-sequence alignment algorithm cannot tell which regions are important or where gaps can be tolerated easily. A profile or HMM is derived from a multiple alignment and does have access to this kind of information. Obviously, signature databases cannot offer the same degree of coverage as the sequence databases. Nevertheless, there is a large number of signature databases available, many of them described in more detail elsewhere in this book (*see* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6, Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6, Article 86, **Pfam: the protein families database**, Volume 6, and others). In particular, the creation of InterPro (*see* Article 83, **InterPro**, Volume 6) as a kind of meta-database has alleviated the need to search a large number of signature databases with different signature types and query interfaces.

4. Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6; Article 92, **Classification of proteins by clustering techniques**, Volume 6

References

Attwood TK (2000) The quest to deduce protein function from sequence: the role of pattern databases. *The International Journal of Biochemistry & Cell Biology*, **32**, 139–155.

Attwood TK and Beck ME (1994) PRINTS–a protein motif fingerprint database. *Protein Engineering*, **7**, 841–848.

Bairoch A (1993) The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Research*, **21**, 3097–3103.

Baldi P, Chauvin Y, Hunkapiller T and McClure MA (1994) Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 1059–1063.

Bork P and Koonin EV (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nature Genetics*, **18**, 313–318.

Bucher P, Karplus K, Moeri N and Hofmann K (1996) A flexible motif search technique based on generalized profiles. *Computers & Chemistry*, **20**, 3–23.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy SR (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.

Gribskov M, McLachlan AD and Eisenberg D (1987) Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 4355–4358.

Henikoff S and Henikoff JG (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, **19**, 6565–6572.

Hofmann K (1998) Protein classification & functional assignment. *Trends in Bioinformatics*, **16** (Suppl 1), 18–21.

Hofmann K (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Briefings in Bioinformatics*, **1**, 167–178.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, **235**, 1501–1531.

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, **31**, 3625–3630.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Staden R (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, **12**, 505–519.

Stormo GD (1988) Computer methods for analyzing sequence recognition of nucleic acids. *Annual Review of Biophysics and Biomolecular Structure*, **17**, 241–263.

Classification of proteins by clustering techniques

Evgenia V. Kriventseva
BASF AG, Ludwigshafen, Germany

1. Introduction

Genome sequencing provides us with an abundance of genes lacking experimentally determined function. To overcome this limitation, Bioinformatics tools are used to generate hypotheses on gene functions. To use just pairwise sequence homology as a hint on gene function can be misleading (Bork and Koonin, 1998). A more reliable basis for functional inference relies on the concept of protein families, collections of proteins with common evolutionary origin. During the past 15 years, different approaches, such as protein-signature recognition and sequence clustering (Kriventseva *et al*., 2001) were developed for the identification of protein families.

Many well-studied protein families contain conserved regions, which can be summarized by a characteristic sequence profile. There are a number of databases storing this kind of protein domain definitions using several different methodologies and a varying degree of biological annotation (see Table 1). PROSITE (*see* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6) is the oldest of the sequence-motif databases that relies on regular expressions and generalized profiles to identify groups of related proteins. Another broadly used approach to model the consensus sequence of amino acids probabilistically is called *HMM* (*Hidden Markov Models*) (Eddy, 1998). This approach has been successfully used in databases like Pfam (*see* Article 86, **Pfam: the protein families database**, Volume 6), SMART (Letunic *et al*., 2004), TIGRrFams (*see* Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6), and SUPERFAMILY (Madera *et al*., 2004).

When structural information is available, it is possible to define protein families and domains at the level of their physical shapes, which are more conserved than sequences. The most prominent examples of such databases are CATH (Orengo *et al*., 2003) and SCOP (Andreeva *et al*., 2004). Most of the sequence domain signature databases and structure classification databases are now integrated in InterPro (*see* Article 83, **InterPro**, Volume 6). Although extensive, these collections of protein family and domain descriptors depend on manual curation and are therefore not comprehensive. Protein-clustering techniques complement these efforts in an automated and comprehensive manner. This review will describe the most popular clustering algorithms and databases of clustered proteins.

Table 1 A list of databases relevant to this article

Resource	URL
Sequence-motif methods	
InterPro	**http://www.ebi.ac.uk/interpro/**
Pfam	**http://www.sanger.ac.uk/Pfam/**
PRINTS	**http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/**
PROSITE	**http://www.expasy.org/prosite/**
SMART	**http://SMART.embl-heidelberg.de**
SUPERFAMILY	**http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY**
TIGRFAMs	**http://www.tigr.org/TIGRFAMs/index.shtml**
Structural alignment and cluster databases	
Protein Structure Classification (CATH)	**http://www.biochem.ucl.ac.uk/bsm/cath/**
SCOP (Structural Classification of Proteins)	**http://scop.mrc-lmb.cam.ac.uk/scop/**
Sequence-cluster databases	
CluSTr	**http://www.ebi.ac.uk/clustr/**
DOMO	**http://www.infobiogen.fr/services/domo**
PIR-ALN	**http://pir.georgetown.edu/pirwww/dbinfo/piraln.html**
iProClass	**http://pir.georgetown.edu/iproclass/**
ProDom	**http://www.toulouse.inra.fr/prodom.html**
Protein Information Resource (PIR)	**http://pir.georgetown.edu/pirwww/pirhome.shtml**
PROT-FAM	**http://www.mips.biochem.mpg.de/desc/protfam/**
ProtoMap	**http://protomap.cornell.edu/index.html**
ProtoNet	**http://www.protonet.cs.huji.ac.il/**
SBASE	**http://hydra.icgeb.trieste.it/sbase/**
SYSTERS	**http://systers.molgen.mpg.de/**
Tribe-MCL	**http://www.ebi.ac.uk/research/cgg/tribe/**
Phylogenetic classifications	
COGs and KOGs	**http://www.ncbi.nlm.nih.gov/COG**

2. Clustering techniques

Clustering can be defined as a way of identifying a natural grouping of the data. Groups are identified on the basis of distances between elements. In the case of protein classification, the workspace can be represented as a graph whose vertices correspond to the sequences. Vertices may be linked by weighted directional edges. The edges are weighted to represent the degree of similarity between two vertices calculated as a pairwise similarity scores between sequences. Strongly connected sets of vertices should then represent clusters of related proteins in sequence space (Tatusov *et al*., 1997).

There are different mathematical solutions on how to classify objects into groups or clusters, optimizing similarity among elements inside a cluster or dissimilarity among different clusters. The latter approach is applicable if the final number of clusters could be estimated in advance (i.e., k-mean clustering). Since the number of protein families in a set is usually not known in advance, the agglomerative hierarchical clustering procedures are better suited for classifying protein sequences into families, providing snapshots on different levels of specificity.

A hierarchical clustering procedure starts with considering elements in the input set as singleton clusters. Then the procedure merges interactively similar clusters and, using different similarity levels, forms the hierarchy of clusters. This procedure results in a tree-like structure, where sibling clusters partition their parent cluster. The hierarchical clustering problem is essentially equivalent to the reconstruction of evolutionary trees also known as "phylogeny".

There are various formulations of the target function for the hierarchical clustering favoring different topologies of the resulting clusters, for example, the single-linkage algorithm produces clusters of linear topology where an element should be similar to at least one other element in the cluster, while complete linkage performs better on compact clusters where all elements of a cluster are similar to each other. These methods produce nonoverlapping clusters; one protein can belong at a certain similarity level to only one group. Unfortunately, nature is more complex and multidomain proteins cannot be correctly classified in this way. A few attempts have been made to redefine the problem to allow overlapping solutions (e.g., Gaussian mixtures models (Dubey *et al*., 2004), fuzzy k-mean (Sarkar and Leong, 2001), etc.).

There are two main complications in automatic clustering procedures: different protein families have different levels of sequence similarity and multidomain proteins pull together proteins with different domains. Hierarchical clustering, defining clusters at different levels of sequence similarity, is one of the approaches to solve these problems. For example, the CluSTr (Kriventseva *et al*., 2003) methodology works with a prior defined set of cutoffs, whereas the SYSTERS (Krause *et al*., 2002) and SL-KL (Kawaji *et al*., 2004) methods automatically select different levels of protein similarity. The following section will show in more details how different approaches try to address these problems.

Single-linkage clustering, also known as *Nearest Neighbor Clustering* (Sokal and Rohlf, 1995), can be described as follows: if an object A, in our case a protein A, is similar to a protein B and the protein B is similar to a protein C then all three proteins are clustered together. This is a simple clustering algorithm, where the results are easy to interpret and are well corresponding to the underlying evolutionary model of protein divergence from a common ancestor. For these reasons, the algorithm was used in SYSTERS (Krause *et al*., 2002), CluSTr (Kriventseva *et al*., 2003), GeneRage (Enright and Ouzounis, 2000), and BLASTClust (**ftp://ftp.ncbi.nlm.nih.gov/blast/**; BLASTClust is a part of the NCBI BLAST package, performing BLAST comparisons followed by single-linkage clustering).

Complete linkage clustering, also known as *Furthest Neighbor Clustering* (Sokal and Rohlf, 1995), can be described as follows: two objects A and B belong to the same cluster if the similarity between them is greater than a chosen threshold. With the complete linkage clustering, all objects in a cluster are similar to each other. The complete linkage strategy produces compact clusters without chaining.

The average linkage clustering or "centroid sorting method" used by the ProtoNet project (Sasson *et al*., 2003) can be seen as an intermediate between single and complete linkage clustering, picking out more homogeneous clusters than those obtained by the single-linkage method.

A few new clustering algorithms were brought to Bioinformatics recently, for example, the MCL (Van Dogen 2000 thesis) clustering based on probabilistic flow simulation in a defined similarity matrix. Although interesting, the method yields poorly interpretable results. Another method proposed by Abascal and Valencia (Abascal and Valencia, 2003) adopts a normalization cut (Ncut) algorithm (Shi and Malik, 1997). In comparison to other methods, this procedure focuses on the space surrounding a given protein by analyzing only local sequence relations. A method inspired by the physical properties of ferromagnetics called superparamagnetic clustering (Blatt *et al*., 1996) is based on partition of "spin–spin" correlations defined on short-range interactions rather than distances themselves.

3. Current protein-clustering approaches

There are two main approaches for protein classification: the comparison of full-length sequences and the comparison of just parts of sequences, the protein domains. ProDom (Servant *et al*., 2002), SBASE (Vlahovicek *et al*., 2003), DOMO (Gracy and Argos, 1998), and Picasso (Heger and Holm, 2001) provide an automated partition of protein sequences into protein domains based on common regions in sequence alignments. The ProDom approach considers the shortest full-length sequence in a database as a single-domain protein and uses PSI-BLAST (Altschul and Koonin, 1998) to find homologous sequences, which are then clustered in the same ProDom entry. In the Picasso approach, clusters are defined on the basis of multiple overlaps in aligned sequences using all-against-all BLAST comparison. For each such multiple alignment, a profile is created, which is used for the further unification of the clusters by profile–profile comparison.

A group of protein family–cluster databases is connected to the Protein Information Resource (PIR) (McGarvey *et al*., 2000). PIR-ALN is a database of protein sequence alignments derived from sequences and annotations in the PIR protein sequence database. Information about PIR protein families, protein superfamilies, and homology domains can be found in ProtFam database hosting the alignments of all PIR superfamilies and homology domains (Mewes *et al*., 2002). iProClass is a recent extension of ProClass (Wu *et al*., 2004) classifying proteins using PIR superfamilies and PROSITE motifs. PIR-ASDB is an Annotation and Similarity DataBase (McGarvey *et al*., 2000) containing precomputed sequence neighbors of all PIR-PSD entries based on all-against-all FASTA searches (Pearson, 1990).

CluSTr (Kriventseva *et al*., 2003), SYSTERS (Krause *et al*., 2002), ProtoMap (Yona *et al*., 2000), and ProtoNet (Sasson *et al*., 2003) classify proteins on the basis of full-length sequence comparisons and provide a hierarchical organization of protein clusters by performing analysis at different levels of sequence similarity. The SYSTERS approach is based on all-against-all gapped BLAST comparison followed by single-linkage clustering. To avoid arbitrarily chosen static thresholds, the procedure defines superfamilies at the level corresponding to the largest increase in a cluster size by traversing the merging tree in the hierarchical single-linkage procedure. Each of the superfamilies is further cut into family clusters by detecting weak connections.

ProtoMap uses a combination of Smith–Waterman (Smith and Waterman, 1981), FASTA (Pearson, 1990), and BLAST (Altschul and Koonin, 1998) searches. In the first classification step, confident cores of clusters are determined at high stringency cutoff. The resulting classes are then merged to form bigger and more diverse clusters by considering weaker, less significant similarities. The procedure is repeated at different levels of similarity. The ProtoNet project builds on the work started in the ProtoMap project. It uses gapped BLAST to calculate similarities, and then applies an average linkage clustering algorithm with three different metrics to merge proteins into clusters: harmonic, geometric, and arithmetic averages of protein distances.

In comparison to the algorithms described above, the CluSTr approach is based on rigorous Smith–Waterman comparisons. To estimate the statistical significance of raw Smith–Waterman scores, a Monte-Carlo simulation (Comet *et al*., 1999) resulting in a Z-score is used. This allows incremental updating of data, thus avoiding time-consuming recalculations. The clusters are built using a single-linkage algorithm at different levels of protein similarity, allowing users to choose different levels of confidence.

The GeneRage (Enright and Ouzounis, 2000) algorithm starts with BLAST all-against-all comparisons, resulting in a binary matrix of similarities. The matrix is then analyzed for identifying multidomain proteins. This is achieved through iterative processing of the matrix elements with successive rounds of Smith–Waterman dynamic programming alignments. Single-linkage clustering is then applied to build the clusters. Clusters containing multidomain families are split into their constituents.

TRIBE-MCL (Enright *et al*., 2002) uses the MCL clustering algorithm. The approach is built on the basis of all-against-all BLAST alignments with following averaging of the similarity scores to make the distance matrix symmetric. Then Markov Clustering is performed by simulating a random walk within a similarity graph. This is done by taking iteratively the power of a matrix, called *expansion* (i.e., matrix squaring), followed by entry-wise powers, called *inflation*. Expansion causes stochastic flow to dissipate within clusters, whereas inflation eliminates flow between different clusters.

The CLICK (CLuster Identification via Connectivity Kernels) (Sharan *et al*., 2003) clustering algorithm was originally developed for the grouping of genes with similar expression patterns into clusters. The algorithm identifies kernels in the similarity graph, defined as highly interconnected subgraphs. Several heuristic procedures are then used to expand the kernels into the full clustering. CLICK has been implemented and tested on a variety of biological datasets, ranging from gene expression, cDNA oligo-fingerprinting to protein sequence similarity.

4. Conclusions

Very recently some major advances in protein-cluster analysis have been made. The exponentially increasing volume of functionally uncharacterized sequences drove the research. While sequence signature recognition methods developed through expert analysis of protein families and domains gain more and more importance,

they need to be complemented by comprehensive clustering techniques. Although the currently available resources have already provided a starting point for in-depth studies of uncharacterized protein families and domains, further work is required to ensure that more of the biological properties of data are recognized correctly.

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 83, **InterPro**, Volume 6; Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6; Article 86, **Pfam: the protein families database**, Volume 6; Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6

References

Abascal F and Valencia A (2003) Automatic annotation of protein function based on family identification. *Proteins*, **53**, 683–692.

Altschul SF and Koonin EV (1998) Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. *Trends in Biochemical Sciences*, **23**, 444–447.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, Database issue, D226–D229.

Blatt M, Wiseman S and Domany E (1996) Superparamagnetic clustering of data. *Physical Review Letters*, **76**, 3251–3254.

Bork P and Koonin EV (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nature Genetics*, **18**, 313–318.

Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP and Codani JJ (1999) Significance of Z-value statistics of Smith-Waterman scores for protein alignments. *Computers & Chemistry*, **23**, 317–331.

Dubey A, Hwang S, Rangel C, Rasmussen CE, Ghahramani Z and Wild DL (2004) Clustering protein sequence and structure space with infinite Gaussian mixture models. In *Pacific Symposium on Biocomputing*, Altman RB, Dunker AK, Huntee L and Klein TE (Eds.), World Scientific Publishing: Singapore, pp. 399–410.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Enright AJ and Ouzounis CA (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Enright AJ, Van Dongen S and Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**, 1575–1584.

Gracy J and Argos P (1998) DOMO: a new database of aligned protein domains. *Trends in Biochemical Sciences*, **23**, 495–497.

Heger A and Holm L (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.

Kawaji H, Takenaka Y and Matsuda H (2004) Graph-based clustering for finding distant relationships in a large set of protein sequences. *Bioinformatics*, **20**, 243–252.

Krause A, Haas SA, Coward E and Vingron M (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Research*, **30**, 299–300.

Kriventseva EV, Biswas M and Apweiler R (2001) Clustering and analysis of protein families. *Current Opinion in Structural Biology*, **11**, 334–339.

Kriventseva EV, Servant F and Apweiler R (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Research*, **31**, 388–389.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**, Database issue, D142–D144.

Madera M, Vogel C, Kummerfeld SK, Chothia C and Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**, Database issue, D235–D239.

McGarvey PB, Huang H, Barker WC, Orcutt BC, Garavelli JS, Srinivasarao GY, Yeh LS, Xiao C and Wu CH (2000) PIR: a new resource for bioinformatics. *Bioinformatics*, **16**, 290–291.

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Weil B (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, **30**, 31–34.

Orengo CA, Pearl FM and Thornton JM (2003) The CATH domain structure database. *Methods of Biochemical Analysis*, **44**, 249–271.

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, **183**, 63–98.

Sarkar M and Leong TY (2001) Fuzzy K-means clustering with missing values. *Proceedings AMIA Symposium*, 588–592.

Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N and Linial M (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Research*, **31**, 348–352.

Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D and Kahn D (2002) ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics*, **3**, 246–251.

Sharan R, Maron-Katz A and Shamir R (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.

Shi J and Malik J (1997) Normalized cuts and image segmentation. *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, Puerto Rico, June 1997*, pp. 731–737.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Sokal RR and Rohlf FJ (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, W. H. Freeman and Company: New York.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Vlahovicek K, Kajan L, Murvai J, Hegedus Z and Pongor S (2003) The SBASE domain sequence library, release 10: domain architecture prediction. *Nucleic Acids Research*, **31**, 403–405.

Wu CH, Huang H, Nikolskaya A, Hu Z and Barker WC (2004) The iProClass integrated database for protein functional analysis. *Computational Biology and Chemistry*, **28**, 87–96.

Yona G, Linial N and Linial M (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research*, **28**, 49–55.

Getting the most out of protein family classification resources

Nicola J. Mulder
European Bioinformatics Institute, Cambridge, UK

1. Introduction

Protein family classification resources have been discussed in the previous chapters. They include the following protein signature and protein clustering databases: PROSITE (*see* Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6) (Hulo *et al*., 2004), PRINTS (*see* Article 85, **The PRINTS protein fingerprint database: functional and evolutionary applications**, Volume 6) (Attwood *et al*., 2003), Pfam (*see* Article 86, **Pfam: the protein families database**, Volume 6) (Bateman *et al*., 2004), SMART (Letunic *et al*., 2004), TIGRFAMs (*see* Article 88, **Equivalog protein families in the TIGRFAMs database**, Volume 6) (Haft *et al*., 2003), PIR SuperFamily (*see* Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6) (Wu *et al*., 2004), SUPERFAMILY (Madera *et al*., 2004), ProDom (Corpet *et al*., 2000), and CluSTr (Kriventseva *et al*., 2003). These databases are automatically derived with manual input, in most cases, and are designed for predicting protein family classification for new and existing protein sequences. SCOP (Andreeva *et al*., 2004) and CATH (Orengo *et al*., 2003) classify protein structures, and Merops (Rawlings *et al*., 2004), Transport Classification (Busch and Saier, 2003), and so on, are specialized protein family resources. The former are based on protein 3D structures data deposited in the Protein Data Bank (PDB; Westbrook *et al*., 2003), while the latter are curated databases developed by experts in a particular protein family. These are useful resources for describing the structural or specialized families, and can be viewed via their web pages (see Table 1). They will not be discussed in this article.

All of the aforementioned resources have their individual websites and databases available for searching, and are intuitive to use. Some of them are integrated into InterPro, which serves the data and software from a single source, and thus makes using the individual data and software quick and easy to access. All of them are represented in SRS (sequence retrieval system; Zdobnov *et al*., 2002), which provides an interface for simple and complex queries and facilitates linking between results and new data. This article demonstrates how to access and utilize the protein family classification data and tools using InterPro (*see* Article 83, **InterPro**,

Table 1 Useful databases for protein family, domain, and motif analysis

Database	Description	URL
UniProt	Protein sequence database	**http://www.ebi.ac.uk/uniprot/**
PROSITE	Database of patterns and profiles describing protein families and domains	**http://www.expasy.ch/prosite/**
Prints	Compendium of protein fingerprints	**http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/**
Pfam	Collection of multiple sequence alignments and hidden Markov models	**http://www.sanger.ac.uk/Software/Pfam/index.shtml**
SMART	A Simple Modular Architecture Research Tool – collection of protein families and domains	**http://smart.embl-heidelberg.de/**
TIGRFAMs	Protein families based on hidden Markov models	**http://www.tigr.org/TIGRFAMs/index.shtml**
ProDom	Automatic compilation of homologous domains	**http://prodes.toulouse.inra.fr/prodom/doc/prodom.html**
PIR SuperFamily	Database of HMMs describing full-length proteins with common domain composition	**http://pir.georgetown.edu/pirsf/**
SUPERFAMILY	Database of HMMs based on SCOP structural superfamily classes	**http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/**
InterPro	Integrated resource for protein families, domains, and functional sites	**http://www.ebi.ac.uk/interpro/**
SRS	Sequence retrieval system for accessing many database and tools	**http://srs.ebi.ac.uk/**
Merops	Resource for peptidases and the proteins that inhibit them	**http://merops.sanger.ac.uk/**
Transport Classification	Classification system for membrane transport proteins known as the Transport Commission (TC) system	**http://tcdb.ucsd.edu/tcdb/**
PDB	Database of protein 3D structures	**http://www.rcsb.org/pdb/**
SCOP	Structural Classification of Proteins	**http://scop.mrc-lmb.cam.ac.uk/scop/**
CATH	Database of protein Class, Architecture, Topology and Homology	**http://www.biochem.ucl.ac.uk/bsm/cath_new/**

Volume 6) as an example. InterPro consists of entries containing signatures that describe the same set of proteins. Entries that are subsets of others are related to each other as parent/child (family and subfamily) or contains/found in (domain composition) relationships. The goal of InterPro is the classification of proteins into families.

2. Protein sequence classification using InterProScan

A query protein sequence may be run against all the protein signatures in InterPro to identify protein domains or diagnostic features of protein families. The sequence-based search, InterProScan, uses tools provided by the member databases. Sequences can be submitted through a web interface at **http://www.ebi.ac.uk/interpro/InterProScan** and results of hits can be obtained

in InterPro in both a graphical and tabular view. The results from all submissions, including interactive searches, are also received by email that provides the XML formatted results and a link to the HTML output. A mail server is available for sequence searches at interproscan@ebi.ac.uk. The mail server and web server InterProScan are linked to SRS (Zdobnov *et al.*, 2002), which indexes the InterPro matches and data XML files. The advantage of this is the access to additional features in SRS.

2.1. Protocol

2.1.1. Requirements

Workstation connected to the Internet, email program, and web browser.

2.1.2. Steps

1. Open the InterProScan sequence search page: **http://www.ebi.ac.uk/InterProScan/**.
2. Paste your sequence into the box in any format or upload your sequence from a file using the "upload file" button. Enter your email address in the appropriate box (this is required in case the search takes too long and times out).
3. Select PROTEIN or DNA depending on what type of sequence you submit. If it is the latter, you can select your translation table.
4. Press "Submit Job" button to proceed with the search.
5. The results are returned in a graphical format, click on "Table View" button to get the results in a tabular format.
6. Follow the link to the InterPro entries in the database by clicking on the InterPro image (clicking on the SRS image goes to the entry as displayed in SRS), and retrieve further information on the families in question. From the entry view, a user can determine if a relationship exists between the entries matched. If a parent/child relationship exists, click on the "Tree" button to see the hierarchical family tree.

2.1.3. Results

The results (Figure 1a) display matches to the InterPro member database signatures and the corresponding InterPro entries. The table view (Figure 1b) shows the positions of the signatures within the sequence, and shows the entry relationships: parent/child or contains/found in, and Gene Ontology (GO, The Gene Ontology Consortium, 2001) terms mapped to the entries hit. In some cases, there may be matches to signatures that have no corresponding InterPro entry, which are usually ProDom domains that will not be integrated. They may, however, provide additional information for the user, and so are maintained in the InterProScan data files. The results should give the user an idea of the family and domain composition of the query protein sequence. If the entries matched are mapped to GO terms, this implies that the query protein can be mapped to these GO terms, thus providing more functional annotation.

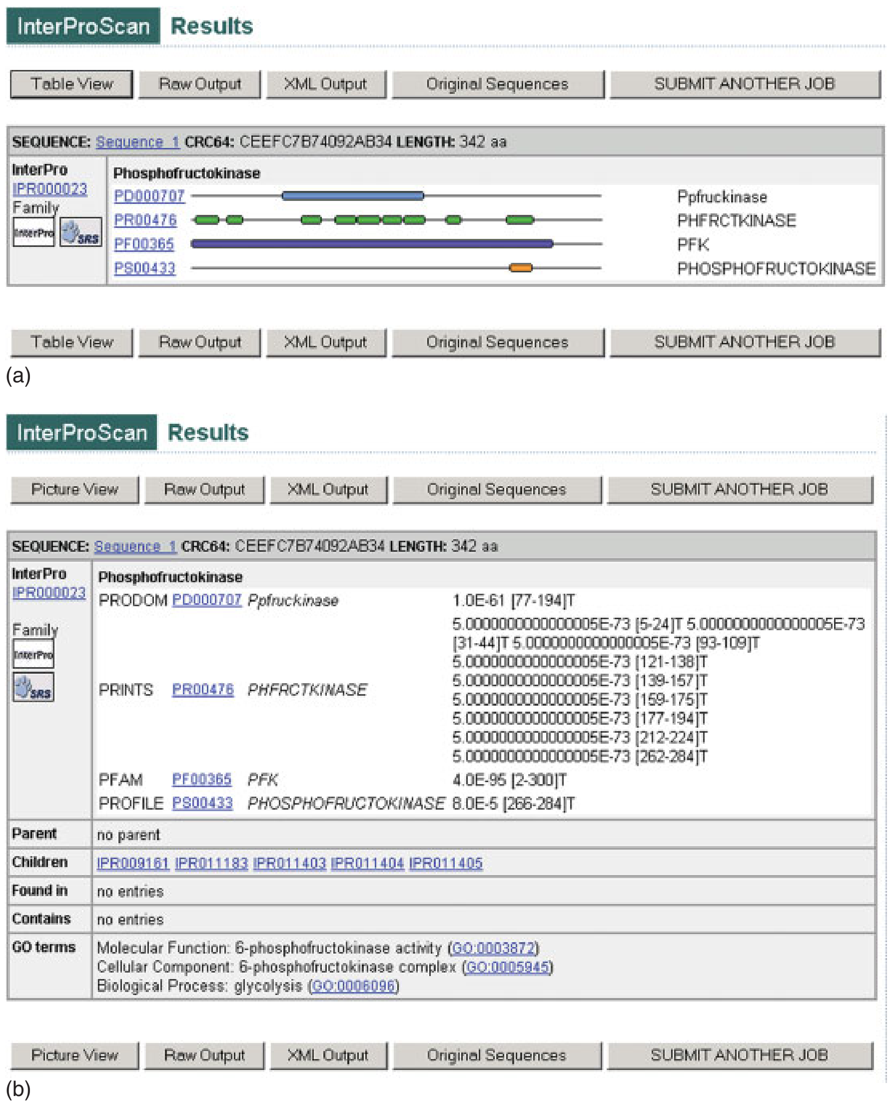


Figure 1 Example of a sequence search result from InterProScan. (a) Graphical view and (b) table view showing entry relationships and GO terms. Links to the original sequence(s) submitted or other formats are provided on each results page

3. Searching InterPro

3.1. InterPro web server

The InterPro web server provides a simple text search tool, as well as a more complex SRS search tool. In the former, it is possible to search entries, in which InterPro fields are searched, find protein matches, where a protein ID or

accession number is the input, and go directly to an InterPro entry by searching for an InterPro accession number. The SRS-based query form facilitates searching based on one field and term in InterPro and one in the protein entries. The output format is selectable. Both search options allow a user to search InterPro on the basis of their terms. In order to browse entries, the InterPro home page (**http://www.ebi.ac.uk/interpro/**) has links to lists of all entries of each type. The lists display the entry names and accession numbers linked to the entries.

3.2. Protocol

3.2.1. Requirements

Workstation connected to the Internet and web browser.

3.2.2. Steps

1. Open the InterPro text search page: **http://www.ebi.ac.uk/interpro/search.html**.
2. To view the InterPro matches for a known protein sequence from UniProt (Apweiler *et al*., 2004), paste a protein accession number in the search box and choose "Find protein matches".
3. Query the database for your protein family or domain of interest by putting the relevant text in the search box and choose "Search entries".
4. Click on one of the entries in the results and view the InterPro entry. In the entry, click on "Table: For all matching proteins " in the first field to list all proteins matching the entry.

3.2.3. Results

The search results are provided in different formats, depending on the output requested. Either way, however, it is possible to reach the relevant InterPro entries. Each entry has the following information in it: protein matches, signatures, relationships, GO terms, abstract, example list, taxonomy field, structural or database links, and publications. The annotation provides a lot of information about the proteins or the search term(s) in question.

3.3. InterPro through SRS

SRS is a system for retrieval of data from a variety of sources. It indexes a number of databases with the ability to launch applications from search results. Results can also be linked to other databases, which allows for very complex queries. InterPro is indexed in SRS and InterProScan is one of the applications available. A minor difference between searching InterPro in SRS and via the InterPro web server is that the former uses the InterPro XML file and the latter is served directly out of the database. At new releases, the data is synchronized.

3.4. Protocol

3.4.1. Requirements

Workstation connected to the Internet and web browser.

3.4.2. Steps

1. Go to SRS at **http://srs.ebi.ac.uk**.
2. Click on "library" tab on the top menu to get to the library page.
3. To begin, choose a "library", for example, InterPro, you would like to search. If InterPro is not visible on the page, you may need to expand the relevant heading (Protein function, structure and interaction databases). Select InterPro as the library, enter a term in the search box at the top and click "Quick Search".
4. Alternatively, select the library and choose Standard or Extended Query form on the left. In the "Standard Query form", four different search terms can be entered, while the "Extended Query form" allows a user to search all fields available in the library.
5. Once you have the search results, go to the "Results" link at the top of the page. Select the required result, and on the "view" scrollbar you can choose a view for them. The results can be ordered by a chosen field, and applications can be launched from them.
6. It is possible to customize a view for your results. Go to "Views" at the top of the page and follow the instructions in the SRS help pages.

3.4.3. Results

The SRS search tools produce the results in the user-specified format. If multiple searches are performed, chosen results sets can be combined on the "results" page. Results can also be linked to other databases, for example, InterPro results can be linked to GO to find the associated terms.

4. Troubleshooting

InterProScan is based on scanning methods native to the InterPro member databases, so all data and applications rely on that provided by these databases. It is distributed with preconfigured method cutoffs recommended by the member database experts and that are believed to report relevant matches.

If the HTML link to InterProScan results returns an error message, it may be that either the web server is inaccessible at the time or that the results are more than 24-hours old, after which time they are removed. In these cases, it is necessary to run the sequence through InterProScan again. For any problems, please email interhelp@ebi.ac.uk.

If no results are obtained, it may be that the sequence is too short, but it is more likely that the member databases have not yet created a protein signature diagnostic for that protein family. Currently, signatures in InterPro cover over

80% of all proteins in UniProt, and the remaining proteins are less likely to have signatures since they belong to very small families. If no results are found, it may be worth running the sequence at the ProDom or Pfam websites. ProDom groups all the nonfragment sequences in UniProt into more than 150 000 families and has therefore almost complete coverage of protein sequence space. Only the bigger, "quality" families in ProDom are integrated into InterPro. A potential hit at the ProDom website that was not found by InterProScan will not provide much annotation or biological information but will give an idea of any related proteins.

Pfam has an automatic supplement to the curated HMMs known as PfamB, which is based on ProDom and provides another means of finding related proteins. If still no information is derived from these searches for the query sequence, it may be worth trying the protein structure classification databases SCOP (Andreeva *et al*., 2004) and CATH (Orengo *et al*., 2003).

5. Discussion

Elucidation of protein function is important not only in genomics but also for individual researchers identifying new genes. Protein family classification resources provide useful tools for achieving this. These tools are available via the Web and should be utilized alone, or more importantly, in combination, for analysis of uncharacterized protein sequences. InterPro and SRS facilitate this and give the user significant options for selection of inputs and outputs to narrow down their searches and save a lot of valuable time. This article has described the main approaches of searching InterPro through the InterPro web server and through SRS as an example of an extremely useful family classification resource. Other databases are also available at EBI through SRS using similar access protocols, and those databases not available at EBI through SRS or other search interfaces are usually available via their own web servers. The Web is awash with data and resources for the scientific community to explore protein functional classification methods.

Further reading

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**(1), 315–318.

Zdobnov EM and Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**(9), 847–848.

References

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family. *Nucleic Acids Research*, **32**(1), D226–D229.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**(1), D115–D119.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, *et al*. (2003) PRINTS and its automatic supplement pre-PRINTS. *Nucleic Acids Research*, **31**(1), 400–402.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**(1), D138–D141.

Busch W and Saier MH Jr (2003) The IUBMB-endorsed transporter classification system. *Methods in Molecular Biology*, **227**, 21–36.

Corpet F, Servant F, Gouzy J and Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, **28**, 267–269.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**(1), 134–137.

Kriventseva EV, Servant F and Apweiler R (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Research*, **31**(1), 388–389.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**(1), D142–D144.

Madera M, Vogel C, Kummerfeld SK, Chothia C and Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**(1), D235–D239.

Orengo CA, Pearl FM and Thornton JM (2003) The CATH domain structure database. *Methods in Biochemical Analysis*, **44**, 249–271.

Rawlings ND, Tolle DP and Barrett AJ (2004) MEROPS: the peptidase database. *Nucleic Acids Research*, **32**(1), D160–D164.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Research*, **11**, 1425–1433.

Westbrook J, Feng Z, Chen L, Yang H and Berman HM (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Research*, **31**(1), 489–491.

Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, *et al*. (2004) PIRSF: family classification system at the protein information resource. *Nucleic Acids Research*, **32**(1), D112–D114.

Zdobnov EM, Lopez R, Apweiler R and Etzold T (2002) The EBI SRS server–recent developments. *Bioinformatics*, **18**(2), 368–373.

Introductory Review

What use is a protein structure?

Simon E. V. Phillips
University of Leeds, Leeds, UK

1. Introduction

Understanding biological organisms in terms of the underlying molecular processes is the main aim of molecular biology. Biological molecules obey the same laws of physics and chemistry as other simpler molecules, so it should be possible to understand how cells and organisms work in molecular detail. Complexity is a defining feature of biological macromolecules, manifested in terms of information content, as in DNA sequence, and structure, as in proteins (and some RNA). Proteins are the major workhorses in the system, and their highly complex three-dimensional structures allow high-fidelity molecular recognition between the thousands of components that must operate together. In order to understand fully how such systems work, it is essential to have accurate, atomic resolution structures for the components. This requires X-ray crystallography (*see* Article 104, **X-ray crystallography**, Volume 6) and NMR (*see* Article 105, **NMR**, Volume 6) experiments since structure prediction from gene sequence, while advancing rapidly, is currently insufficiently reliable and accurate.

While some proteins, such as collagen fibers in tendon, simply contribute to the overall macroscopic structure of an organism, the majority are active components. Enzymes catalyze chemical reactions, both in metabolism and signaling, receptors recognize signals and process them, membrane proteins pass molecules and/or signals across membranes, muscle proteins form linear motors that drive motion, photosynthetic proteins transduce light energy into chemical energy, transport proteins carry cargo molecules, and immune system proteins search for, and destroy, foreign molecules. It is a natural human aspiration to understand these systems, both in healthy and diseased states, but this knowledge also gives us the power to intervene. This could be for therapy, such as in the design of new drugs, or by taking proteins out of their natural system and engineering them for other uses in vitro.

When Dorothy Crowfoot (later Hodgkin) and Max Perutz (*see* Article 95, **History and future of X-ray structure determination**, Volume 6) set out, in the 1930s, to determine the structures of insulin and hemoglobin respectively, they did so because of their fascination with the problem. They also recognized that understanding these protein structures was likely to have implications for medicine. While the relevance of insulin to diabetes was already clear at the time, it was some years later when Linus Pauling proposed that the inherited disease sickle-cell

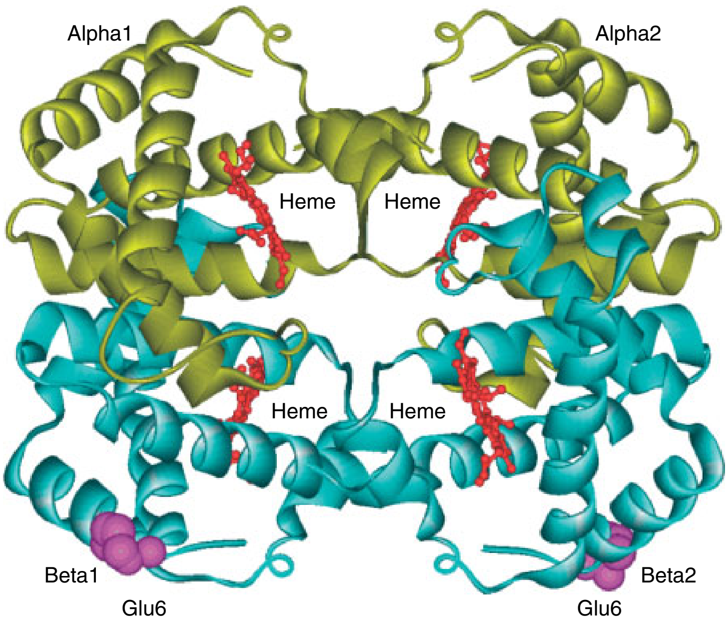


Figure 1 Human deoxyhemoglobin tetramer (HbA), with the protein chains shown as ribbons and the heme groups in red ball and stick representation. Alpha chains are gold and beta chains cyan. Oxygen binding to the heme groups causes a rotation of the alpha1/beta1 dimer relative to alpha2/beta2. The site of the mutation to HbS is shown with Glu6 colored magenta in space-filling representation. Coordinates from Protein Data Bank (PDB – **http://www.pdb.org**) code 4hhb

anemia was caused by a defect in the hemoglobin molecule, and was therefore a "molecular disease" (Pauling *et al*., 1949). When Perutz finally determined the crystal structure of deoxyhemoglobin, and built an atomic model in 1968, the result was stunning in its beauty and complexity, but he expressed disappointment that how it worked in cooperative oxygen transport was not immediately apparent from looking at its structure (Figure 1). The later crystal structure of the oxy- form showed large changes in quaternary structure (arrangement of the four subunits) that was related to the observed cooperativity and triggered by small structural changes when oxygen binds to the heme groups. It was clear that, although knowledge of the three-dimensional structure is essential, it is also critical to combine information on function gained from other studies.

Structural studies are vital to guide functional studies, and the two proceed in a symbiotic relationship (for an early account of hemoglobin structure and function see Perutz, 1978). In the case of insulin, where the structure in the crystal is a hexamer, the critical discovery came when the regions contacting the insulin receptor were identified and it became clear that the active form was the free monomer. Pauling's prediction of defective hemoglobin (HbS) in sickle-cell anemia could now be interpreted in structural terms, as it was known that HbS differs from normal hemoglobin (HbA) in that the glutamate residue at position 6 of the ß-chain has been mutated to valine (Figure 1). The change from a charged to a hydrophobic residue gives rise to an extra hydrophobic patch on the surface, which tends to stick

to hydrophobic patches on other HbS molecules. HbA is very soluble, and present in the red cells at very high concentrations, but it does not restrict the cells' flexibility and ability to squeeze through the finest capillaries of the vascular system. HbS molecules, on the other hand, when in the deoxy- form, tend to stick together via their extra hydrophobic patches to form stiff fibers that cause the cells to become inflexible and jam in the capillaries, resulting in the painful and dangerous crises of sickle-cell disease. The crystal structure of HbS, together with electron microscopic studies of the fibers, revealed the molecular basis of the disease, and it was possible to consider the prospect of designing a treatment for the disease at the molecular level. We now know of many diseases caused by such point mutations. Another dramatic early example was the observation that a common mutation in human tumor cells is in the ras p21 protein, involved in signaling pathways, where inspection of the crystal structure shows that the change lies in a protein loop responsible for binding part of the substrate guanosine triphosphate (GTP).

The cases above show how knowledge of the structure of a protein is essential to understanding its function (or malfunction), but how can that knowledge and understanding be applied to modify its behavior?

2. Structure-based drug design

Much of the success of modern medicine stems from the availability of a wide range of drugs. Until recently, many of these had their origins in traditional remedies from which pharmaceutical companies identified and produced the active agents, as well as chance observations such as penicillin, and random screening of large numbers of natural products. Most drugs are effective because they bind to a specific receptor, often (but not always) a protein molecule, and modify or block its function. Drugs are generally organic molecules with molecular weights of up to a few hundred, relatively small compared to proteins but large enough to have complex shapes and chemical characteristics. They bind to their target proteins with high affinity and specificity, typically in clefts whose shapes and chemical characteristics (charge, hydrophobicity, hydrogen bonding) are accurately complementary to those of the drug. Since the physical principles of the noncovalent interactions between the drug and the protein are well understood, it is possible to use computer graphics modeling to design small molecules that should bind to a site in a given protein structure. Such a procedure clearly depends on knowledge of the structure of the target protein, and has become known as structure-based drug design. True de novo design of a drug remains challenging since it requires accurate predictions of binding energies, and these are complicated by protein dynamics, solvent, and entropic effects. The method is, however, very powerful, where some experimentally determined structures are also known for complexes of the protein with putative drug molecules that may not bind very tightly but can serve as "lead" compounds. Such lead compounds can sometimes be initially identified by virtual screening using a computer, where the known structure of a protein drug target can be tested against huge libraries of small molecules in a search for possible binding partners. Computer graphics is then used to improve the design, a new molecule is synthesized, the structure determined of its complex with the

protein, and its affinity measured. The cycle can be continued until the tight-binding, specific compounds are found that can go to the next stages of drug testing. This has become standard practice in the pharmaceutical industry, and the methodology is also applicable in the design of herbicides and pesticides. The whole procedure is, however, critically dependent on the availability of accurate, three-dimensional structures of the protein targets, that must be determined experimentally and cannot be simply predicted from gene sequences.

This has led to the structure determination of many proteins known to be drug targets, as well as those predicted to be new targets from genome information. In the case of some traditional drugs, the target protein structures have been determined in an effort to make improvements to the drug design. For instance, this is the case for aspirin, where structures are known for its targets cyclooxygenase and phospholipase-A2, and their complexes with various aspirin derivatives. Perhaps more interesting are drugs designed more recently on the basis of target protein structures, such at those aimed at human immune deficiency virus (HIV) protease and reverse transcriptase, influenza neuraminidase, and human kinases. Influenza is a dangerous virus, where antigenic variation renders vaccination difficult. Neuraminidase (NA) is one of two major proteins on the influenza virus surface. It cleaves sialaic acid from cell surface receptors, probably facilitating release of progeny virions, and its inhibition reduces the severity of the infection. Following the determination of the crystal structure of the globular catalytic head domain of NA (Colman *et al*., 1983), the active site was used as a design target for inhibitors, leading to one of the first cases of true structure-based design. Zanamavir (Relenza^{TM}) is now licensed as an anti-influenza drug, and the crystal structure of its complex with NA is shown in Figure 2. Protein kinases are major targets in cancer therapy, as they are critical to cellular signaling systems (e.g., in differentiation and proliferation) and there are several hundred different ones in a typical cell. Kinases are related to each other, but have sufficient differences to allow design of drugs that target some of them selectively. In chronic myelogenous leukemia (CML), a chromosome translocation occurs, generating a novel Ab1 tyrosine kinase with

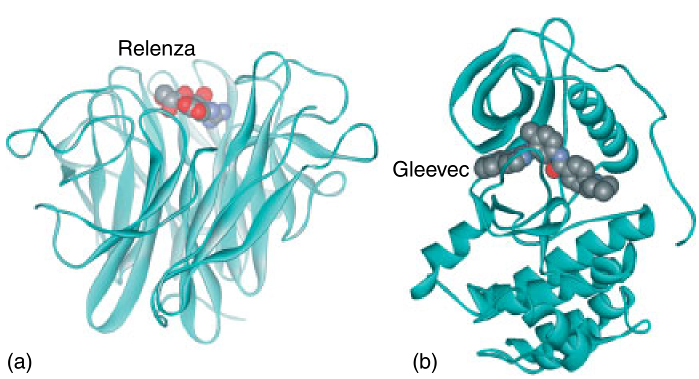


Figure 2 Drug binding to proteins (a) Zanamavir/Relenza^{TM} bound to influenza neuraminidase (PDB code 1a4g) (b) ST571/Gleevec/Glivec^{TM} bound to c-Kit kinase (PDB code 1t46). Drug molecules are shown as space-filling representations in CPK colors

enhanced activity that is a target for intervention. A structure-based design exercise, based on several kinase structures, yielded a compound (STI571) that potently inhibits Ab1, c-Kit kinase, and platelet-derived growth factor receptor (PDGFR), but not other kinases. This has been licensed as the drug Gleevec/Glivec^{TM}, which has shown high activity against CML. Figure 2 shows the crystal structure of Gleevec/Glivec^{TM} bound to the c-Kit kinase.

Many targets for drug intervention, however, are complex protein–protein interactions in signaling pathways or, indeed, in Pauling's original molecular disease, sickle-cell anemia. Drug design is more challenging in these cases, although many structures are known, since most protein–protein interfaces lack the clefts necessary to bind small drug molecules tightly. This is an area of intense research, but progress is slow and no simple cure has yet been found for sickle-cell disease.

3. Protein engineering

While drug design remains the most widespread practical use of protein structures, the ability to use site-directed mutagenesis to engineer modified proteins has resulted in the design of novel proteins for therapeutic and other uses. The most widespread use of enzymes is in "biological" washing powders and some industrial processes. The obvious target for protein engineering here is in improving the stability of the enzymes, especially at higher temperatures. Protein structures are used to guide mutations that confer stability, and obvious alterations include adding additional disulfide bridges or removing long, flexible loops. Other industrial targets include improving stability of useful enzymes in organic solvents. Enzymes are increasingly used in industrial synthesis of small organic molecules with sufficient complexity that chemical synthesis results in mixtures of stereoisomers. Enzymes, on the other hand, are stereospecific, and can synthesize complex molecules, such as those needed to make drugs, as single isomers. It is often the case, however, that no specific enzyme exists for the required synthesis, but it is possible to use knowledge of the protein structure to aid in the design of an engineered enzyme with a more suitable product.

Therapeutic uses of protein engineering based on structural knowledge include the reengineering of insulin and human growth hormone to increase their potency. Knowledge of the hormone structure, and its interaction with the receptor, are important in the design. For instance, normal human insulin exists as a hexamer, but the free monomer is the active species, so the protein was engineered to reduce its tendency to form hexamers by mutating the interfaces between subunits, resulting in a more effective hormone for treatment of diabetes. A major area for protein redesign based on known structures is in the therapeutic uses of antibodies. For instance, antibodies that target antigens on human tumor cells can be raised in rats, but they are of limited use. If administered to a patient, there is some remission of the disease, but the rat antibodies are foreign proteins, so the patient's own immune response soon targets them and reduces their effectiveness. The structure determinations of many antibodies, and especially of their complexes with protein antigens (e.g., Amit *et al*., 1986; Figure 3), showed that only a limited number of amino acid

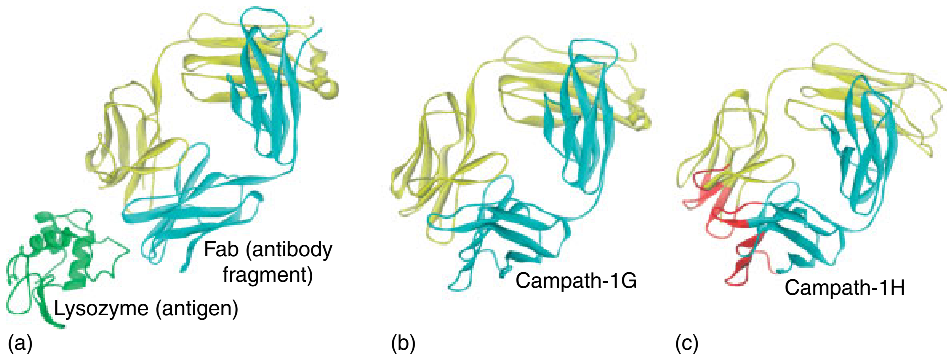


Figure 3 Structure of Fab fragments of antibodies, with the light chains gold and heavy chains cyan. (a) Mouse antilysozyme antibody D1.3 complexed to its antigen. Note the only regions in contact with lysozyme are the CDR loops (PDB code 1fdl). (b) Rat CAMPATH-1G antibody before humanization (PDB code 1bfo). (c) Human CAMPATH1-H antibody with the CDR loops grafted from CAMPATH1-G shown in red (PDB code 1bey)

residues in the six complementarity determining loops (CDRs) of the antibody actually interact with its target antigen. The rest of the antibody molecule simply acts as a constant framework to support the CDRs. This led to the design of "humanized" antibodies, where the specific CDRs of a therapeutic rat antibody were grafted by protein engineering on to the framework of a human antibody, guided by the known structures. In the case of a rat antibody, CAMPATH1-G (Figure 3), that targets human lymphocytes and can be used to treat a number of autoimmune diseases, a humanized version was prepared, CAMPATH1-H (Figure 3), that provided extended remission for patients (Riechmann *et al*., 1988) by evading their immune responses.

4. Protein assemblies and molecular machines

Many of the current uses of protein structures relate to relatively simple systems, but it is clear that most proteins in higher cells act as components of larger complexes. Some of these constitute true molecular machines, such as linear motors in muscle fibers and rotary motors in the ATPase or the bacterial flagellum. Viruses and bacteriophages often have elaborate machinery for entering cells and the photosynthetic complexes in plants contain multiple components that act together to convert light to chemical or electrical energy. Structures are now becoming available for such complexes (*see* Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6), with the possible opportunity of reengineering them for specific purposes, such as the clean conversion of solar energy or the construction of nanoscale molecular devices. The uses of protein structures are still in their infancy.

References

Amit AG, Mariuzza RA, Phillips SEV and Poljak RJ (1986) Three-dimensional structure of an Antigen-antibody complex at 2.8Å resolution. *Science*, **233**, 747–753.

Colman PM, Varghese JN and Laver WG (1983) Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature*, **303**, 41–44.

Pauling L, Itano HA, Singer SJ and Wells IC (1949) Sickle cell anemia: a molecular disease. *Science*, **110**, 543–548.

Perutz MF (1978) Hemoglobin structure and respiratory transport. *Scientific American*, **239**, 92–125.

Riechmann L, Clarke M, Waldmann H and Winter G (1988) Reshaping human antibodies for therapy. *Nature*, **332**, 323–327.

Introductory Review

History and future of X-ray structure determination

Joachim Jaeger and Caitriona Dennis
Center for Medical Science, Wadsworth Center, NYS DOH, Albany, NY, USA

1. Early history (1910–1940)

Single crystal X-ray diffraction originated in Germany at the beginning of the twentieth century. Von Laue and colleagues conducted the first diffraction experiments on rock salt and other alkali halides by shining X-radiation, discovered by Roentgen in 1895, onto the crystalline samples (Figure 1). These studies and his mathematical formulation of single crystal X-ray diffraction earned Von Laue the Nobel Prize for Physics in 1914. Independently, W. L. Bragg and W. H. Bragg carried out similar X-ray studies in Leeds, England, where they found that the diffraction phenomenon could be treated – mathematically – as a reflection by successive parallel planes passing through crystal lattice points.

According to both Laue and Bragg, scattered intensities arising from crystals result in sharply ordered diffraction maxima (diffraction spots). The incident beam approaching the lattice plane at an angle θ is being reflected from that plane at an equal angle ("Glanzwinkel"). Laue discovered that the conditions for observing a maximum in diffracted intensities require that the path difference between reflected beams from adjacent lattice planes be an integral number of wavelengths ($h = \mathbf{S}\ a$, $k = \mathbf{S}\ b$, or $l = \mathbf{S}\ c$). Incorporating Bragg's mathematical equation, the distance between lattice planes could be defined as $d = 1/|S|$ and $|S| = 2|\sin\theta|/\lambda$. Thus, it followed that $n\lambda = 2d\sin\theta$, known as Bragg's Law. The Braggs (father and son) were awarded the Nobel Prize for Physics in 1915.

During the 1920s and 1930s, the focus of diffraction studies shifted to more complex systems such as macromolecular fibers and protein crystals. In Leeds, W. T. Astbury and coworkers pioneered structural studies on large fibrous proteins such as hair, wool, and quills and on DNA fibers. After taking the first fiber diffraction images of DNA, he correctly predicted the overall dimensions of the molecule and found that the nucleotide bases were stacked at intervals of 3.3 Å perpendicular to its long axis. It was left, however, to Watson and Crick to later elucidate the detailed atomic structure of the DNA double helix.

The problems encountered in the early days of macromolecular crystallography related to the fact that biological matter interacts only weakly with electromagnetic radiation as the sample is not densely packed with electrons. Proteins and nucleic

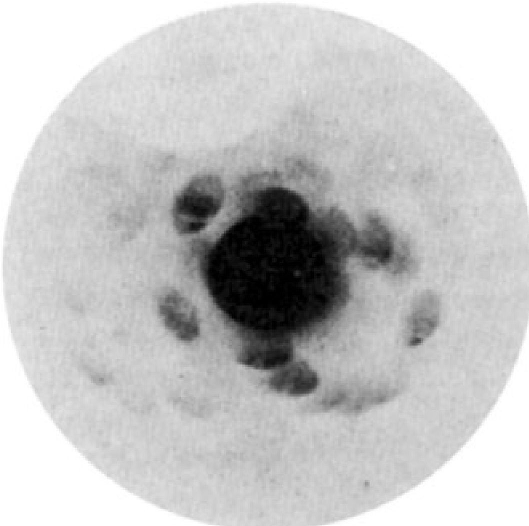


Figure 1 First diffraction pattern from NaCl crystals recorded by Von Laue and colleagues

acid typically consist of light elements such as hydrogen, carbon, nitrogen, oxygen, phosphorous, and sulfur atoms, which scatter weakly. In order to enhance the diffraction signal, therefore, it is necessary to increase the intensity of the incident beam as well as the scattering volume of the sample. This is achieved by shining highly focused, collimated X-radiation onto crystalline samples, that is, well-ordered, symmetrical, three-dimensional arrays. These preliminary studies led to further investigation of the diffraction properties of pepsin crystals by J. D. Bernal and D. Crowfoot from the Cavendish Laboratories at Cambridge, during the mid-1930s. They recognized that biomacromolecular samples must be kept in an aqueous, more native-like environment ("mother liquor") rather than as dry-mounted crystals.

2. The first breakthroughs: hemoglobin and myoglobin (1940–1970)

In 1937, Max Perutz performed the first experiments to find out whether it might be possible to determine the atomic structure of hemoglobin by X-ray diffraction. It took Perutz until 1953 to achieve the most critical breakthrough in actually visualizing such complex structures. He succeeded in incorporating heavy atoms, namely, those of mercury, into definite positions in the hemoglobin molecule. The diffraction pattern, in this method, is altered to some extent and the changes are utilized to determine a direct image of the hemoglobin molecule, within three-dimensional crystals. This method provided for the first time a solution for an often-insurmountable problem in X-ray structure determination, known as the *phase problem*.

Using the same technique, which became known as Multiple Isomorphous Replacement (MIR), Kendrew succeeded in incorporating heavy atoms (mercury and gold) into myoglobin crystals. By 1958, the structures of myoglobin and hemoglobin had been determined after more than 20 years of dedicated labor, and for these groundbreaking discoveries, Perutz and Kendrew were awarded the Nobel Prize in Chemistry in 1962. Within the next 5 years, the first structures of the enzymes lysozyme, carboxypeptidase, RNase S, chymotrypsin, subtilisin, and papain were also determined at near-atomic resolution.

3. Recent history (1970–1990)

In the 1970s and 1980s, X-ray crystallography continued to flourish. This was due to further development of the theoretical background and dramatic improvements in experimental designs and computational methods. In the computational area of research, the technique of "molecular replacement" represented a major breakthrough in overcoming the phase problem. Provided a homologous or sufficiently similar structure is already available, phases could be derived initially from a randomly placed but structurally related model. The model orientation and position relative to the origin would then be corrected (and optimized) by dissecting the molecular replacement procedure into a rotational and translational search. The search procedures are based on principles derived by Patterson, Crowther, Rossmann, and Huber.

In the 1980s and 1990s, computer hardware had advanced so much that it was possible to carry out most of the mathematically challenging work, including structure refinement and graphics computer-based model building. This period saw the introduction of highly interactive graphics programs such as FRODO and O by T. A. Jones, or refinement routines such as PROLSQ by J. H. Konnert and W. A. Hendrickson, TNT by D. E. Tronrud and XPLOR by A. T. Bruenger. The latter routines employed fast Fourier transforms (FFT) and least squares (LSQ) methods for structural refinement. Major improvements in the efficiency of structure refinement came about with the implementation of molecular dynamics (temperature bath-coupled, simulated annealing) algorithms in CNS. These techniques proved more powerful than least squares methods due to an increase in the radius of convergence. The introduction of the Free R-factor, a reliability index calculated between the refined structure and a set of omitted, thereby, unbiased set of observed X-ray reflections, substantially increased the reliability and accuracy of refinement procedures. The manual (computer-based) intervention in generating and fitting atomic structures in conjunction with a certain degree of automation in the subsequent refinement procedure caused a dramatic increase in the number of structures deposited in the Protein Structure Databank (PDB).

Besides many improvements in computational methods, experimental techniques were advanced as well. Historically, diffraction intensities had been recorded on X-ray sensitive films, followed by amplitude estimation by eye or by digitizing on a rotational scanner. In the 1980s, X-ray-sensitive area detectors with big apertures, however, became available and have since had a major impact on the accuracy and the speed of X-ray data collection. Imaging plate detectors

use polyester plates coated with highly dispersed barium fluorohalide phosphor crystals (BaF(Br,I):Eu2+). The recorded X-ray diffraction intensities are read out through photoactivation with a laser and recorded by a photomultiplier. A further improvement in data acquisition speed and accuracy came about with the implementation of charge coupled device detectors (CCD) in the early 1990s. CCD detectors allow the direct readout of a current that is generated by an incident photon onto an X-ray-sensitive phosphor that is backed by arrays (up to 4096 × 4096 pixels) of tiny, siliconoxide-based capacitors coupled to tapered optical fibers. The pixels have a diameter of about 10 µm, and by using paneled CCD designs (2 × 2 or 3 × 3 arrays), the apertures are sufficient for macromolecular X-ray data collection. The readout time (or "dead-time") of such a device is less then 1 s, thereby substantially decreasing collection time to hundreds of minutes rather than days.

Synchrotron rings have had and will continue to have a major impact on the analysis of macromolecular structure. The charged particles within the storage ring generate electromagnetic radiation when diverted from a straight path, through a magnetic field generated by a bending magnet. Insertion devices, such as multipole wiggler magnets or undulators, further modify the path of the electrons (or positrons) and allow modification of the X-ray wavelength. The wavelength can be selected not only to reduce radiation damage but can also be tuned to meet specific scattering properties of the biological sample or bound ligands. This has been particularly successful in utilizing anomalous diffraction characteristics (or X-ray fluorescence) of heavier chemical elements such as Se, Fe, Zn, Cu, and so on. Single or multiplewavelength anomalous dispersion (SAD/MAD, pioneered by Wayne Hendrickson) and single isomorphous replacement in conjunction with anomalous scattering (SIRAS) have now become the phasing method of choice. These phasing methods are based on the anomalous effect, which in turn is a wavelength-dependent phenomenon. Anomalous scattering usually occurs in heavy atoms such as sulphur, bromine, iodine, and, prominently, in main and transition group metals. The effect is caused by the interaction of X-ray photons with outer shell electrons in the anomalous scatterer. Some photons may be absorbed and reemitted at lower energy ("fluorescence"), but more importantly, some photons are absorbed in a wavelength-dependent manner and immediately reemitted at the same energy. Such scattered photons gain an additional imaginary component to its phase, indicating that it is being retarded compared to a normally scattered photon. Accordingly, the atomic form factors of heavy atoms should be separated into the following components:

$$f_{\text{tot}} = f_{\text{norm}}(\theta) + f_{\text{anom}}(\lambda) = f_{\text{norm}}(\theta) + f'_{\text{anom}}(\lambda) + i f'_{\text{anom}}(\lambda) \quad (1)$$

The differences in intensities (Bijvoet differences) and in phase shifts (retardation) can be utilized to overcome the phase problem. The differences in anomalous amplitudes are usually considerably smaller than those obtained from MIR and require a high degree of accuracy in measuring individual intensities by using more intense X-ray beams and highly sensitive detectors.

4. Present and future (2000–2010)

Recent advances in computational approaches using Bayesian statistics and maximum likelihood methods have helped to facilitate initial phase determination and have proved very useful in overcoming problems in phase refinement. Maximum likelihood methods as implemented in the programs CNS and REFMAC have also been applied to X-ray structure refinement. Automated Patterson search procedures in conjunction with partially or fully automated fragment-based model building and structure refinement routines have helped a great deal in simplifying and speeding up the entire process of structure determination and refinement.

Structural proteomics programs brought about major technical advances in high-throughput methods such as the use of robotics in crystal growth, screening, mounting, X-ray data collection, and evaluation. The brilliance of modern synchrotron beams combined with very short exposure times have dramatically increased the speed of structure determination. The future of single crystals diffraction methods lies in the development of even more powerful, tunable radiation sources such as the X-ray free electron laser (XFEL). Single exposure times will decrease well below the second scale, opening up opportunities for time-resolved diffraction studies, an area that has shown a lot of promise by using polychromatic X-radiation but has been hampered by inherent technical difficulties with Laue diffraction.

To this day, however, diffraction methods remain the most commonly used, successfully applied, and productive techniques to elucidate macromolecular structure. In 2004, the Protein Data Bank (**www.rcsb.org**) lists more than 28 000 entries with approximately 85% of the deposited structures determined by single crystal diffraction. Major advances in multidimensional NMR techniques are contributing significantly to our understanding of macromolecular structure, particularly with respect to their dynamic properties. With the recent advances in cryogenic techniques at liquid helium temperatures, electron diffraction and cryogenic imaging techniques have begun to make major contributions to structural studies at the molecular level. Combination of these techniques along with correlating structure, function, and dynamics provides a more complete understanding of biomacromolecular structure.

Specialist Review

Fundamentals of protein structure and function

James R. Bradford, Jennifer A. Siepen and David R. Westhead
University of Leeds, Leeds, UK

1. Introduction

The amino acid sequence of a protein encodes its three-dimensional (3D) structure, which in turn determines its biological function. Proteins have evolved over millions of years to produce a diverse set of structures and can be described as globular, fibrous, or transmembrane. Globular proteins are generally soluble in water, adopt globular 3D shapes, and carry out functions that range from enzyme catalysis to transport and storage. Fibrous proteins usually perform a structural role, with examples including collagen, α-keratin, and elastin. Integral transmembrane proteins exist in lipid membranes and their functions include forming channels through which molecules or ions can pass, and participating in signal transduction by conveying chemical messages into and out of the cell. Invariably, these functions require that proteins interact with each other or with different molecules such as DNA. Therefore, to understand the relationship between a protein's structure and function fully, knowledge of that protein's evolutionary history and its interactions is essential.

2. Primary structure

2.1. Amino acids

Proteins are composed of 20 "standard" amino acids, which contain an amine group, a carboxylic acid, a hydrogen atom, and a characteristic side chain (R group) all bonded to a central carbon atom, commonly called the alpha carbon (C_α), as shown in Figure 1. Amino acids exist as a zwitterion in physiological pH, with the carboxylic acid deprotonated and the amine group protonated. As a result, they are very soluble in polar solvents such as water but insoluble in most organic solvents. All of the amino acids (with the exception of glycine) are chiral molecules, because they have four chemically distinct groups attached to their chiral center (C_α). Chiral molecules can exist in two different forms (L and D forms), all amino acids that

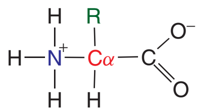


Figure 1 The typical chemical structure of an amino acid. R represents the variable side chain

form natural proteins have the same relative configuration about their C_α atoms and adopt the L form.

Each of the 20 amino acids are characterized by distinctive physicochemical properties that include polarity, acid/base properties, aromatic properties, bulk, conformational flexibility, cross-linking ability, hydrogen-bonding capability, and chemical reactivity. These properties are interrelated and largely responsible for the great range of protein properties. Amino acids are often classified into three groups, hydrophobic, polar, and charged, as shown in Figure 2, on the basis of the chemical nature of their side chains.

2.2. *Peptide bond*

During protein synthesis, amino acids polymerize in a head-to-tail arrangement, extending from the N (amino) terminus to the C (carboxy) terminus, by the formation of peptide bonds between two successive amino acids. A peptide bond is formed between adjacent residues by a condensation reaction (the elimination of water) between the amino group of one residue and the carboxyl group of another (Figure 3). This polymer, which can contain approximately 40 to over 4000 amino acids, is called a *polypeptide*. The main chain or backbone of the polypeptide is composed of the C_α atom, to which the side chain, an NH group and a carbonyl group (C'=O) are attached and a peptide bond joins C' of one residue to N of the next. In the 1950s, Linus Pauling and Robert Corey (1951) observed that peptide groups, with few exceptions, assume the planar trans conformation, in which successive C_α atoms are on opposite sides of the intervening peptide bond.

2.3. *Polypeptide backbone*

The conformation of the polypeptide backbone may be described by two repeating dihedral angles, shown in Figure 4, about the C_α–N bond (phi, ϕ) and the C_α–C bond (psi, ψ), the third repeating angle about the N–C bond (omega, ω) is generally fixed at 180°. Sterically forbidden conformations of the ϕ and ψ angles are those in which the nonbonding interatomic distances are less than the corresponding van der Waals distances. The Ramachandran diagram (Figure 5), so called after its inventor G. N. Ramachandran (1968), shows the allowed ϕ and ψ angles in proteins in which all the common types of regular secondary structure (discussed later) that are found in proteins fall. The conformational angles of most residues in 3D X-ray determined structures (*see* Article 104, **X-ray crystallography**, Volume 6) are observed to lie in the allowed regions of the Ramachandran diagram. The only exception is glycine: possessing only a hydrogen atom as its side chain confers

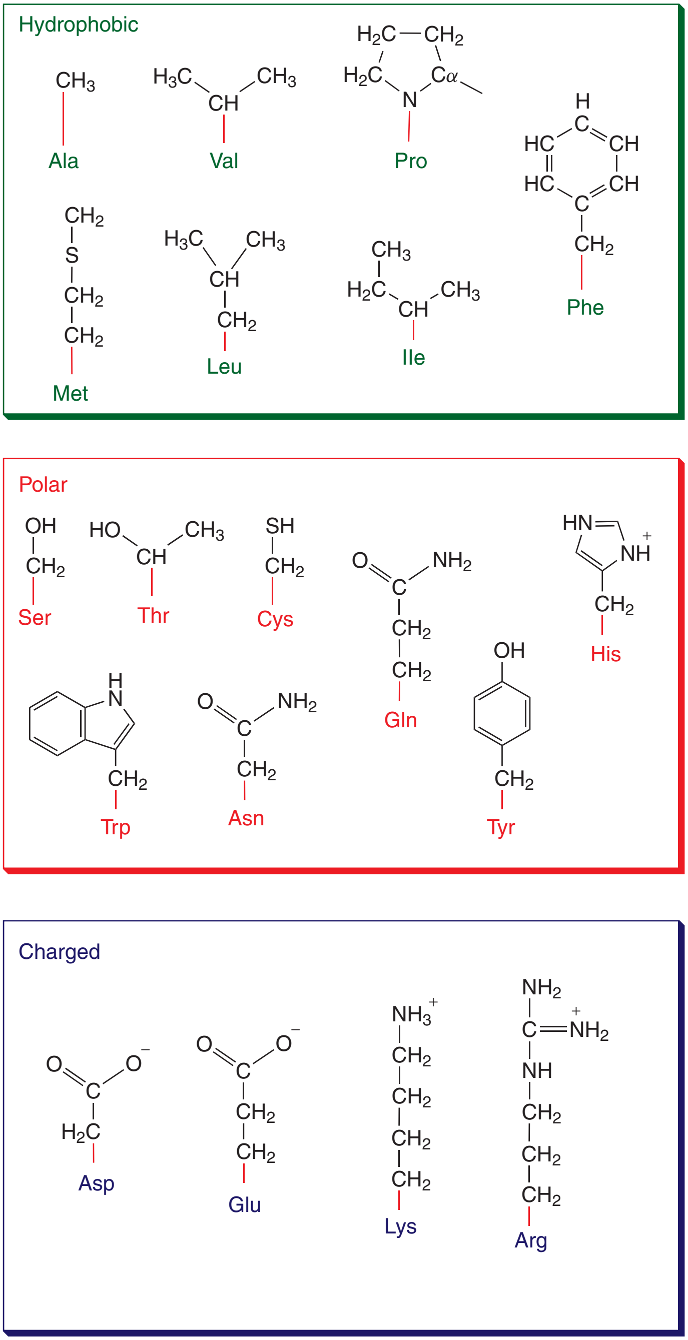


Figure 2 The chemical structures of the 20 amino acids

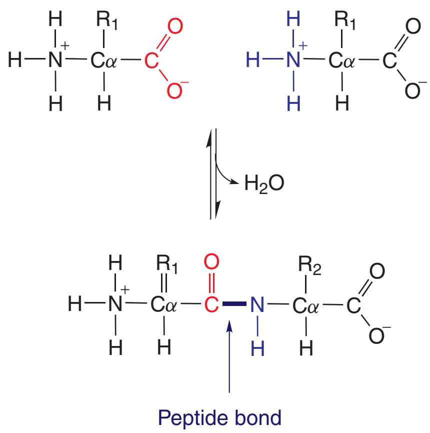


Figure 3 The formation of a peptide bond

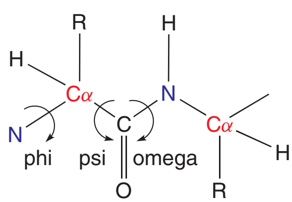


Figure 4 Backbone torsion angles

to it greater than normal flexibility on the polypeptide chain by reducing steric hindrance.

3. Secondary structure

Polypeptide chains form hydrogen bonds between the N–H and C′=O of different residues to form regular secondary structure elements, where consecutive residues have similar ϕ and ψ angles. The two most common secondary structure elements are the alpha (α) helix, formed from a stretch of consecutive residues in the polypeptide, and the (β) beta sheet, where hydrogen bonding occurs between different regions of the polypeptide chain.

3.1. α-Helix

Hydrogen bonding between the C′=O group of residue n and the N–H of residue $n + 4$ results in an α-helix with ϕ and ψ angles of approximately -60^{o} and -50^{o} respectively. An α-helix contains 3.6 residues per turn and each turn is 5.4-Å high (the pitch of the helix); they vary in length from 4 to over 40 residues, with the

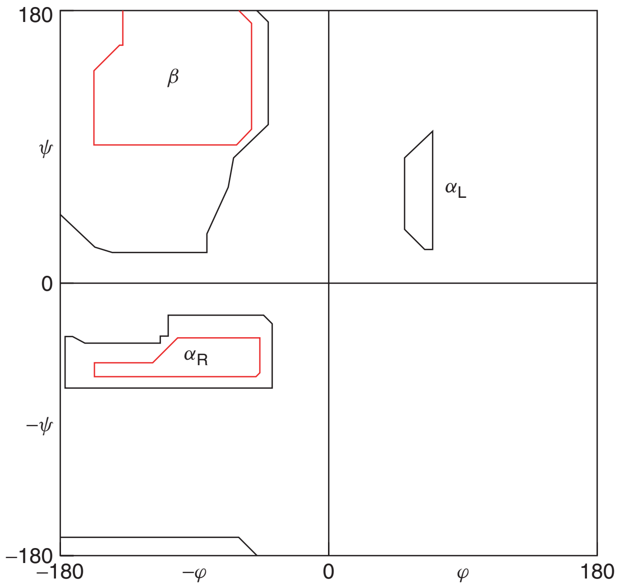


Figure 5 A typical Ramachandran diagram. The red lines enclose the sterically allowed regions for three conformational patterns: right handed α-helices (α_R), left-handed α-helices (α_L) and β-strands (β). The black lines indicate the estimated maximum tolerable limits of steric strain

average helix 10 residues in length (three turns). The α-helix can be right- or left-handed, however, because all amino acids adopt the L form, the alpha helices are nearly always right handed. Other helical structures exist that are either more or less tightly coiled than the α-helix. The 3_{10} helix (C′=O of residue n is hydrogen bonded to the N–H of residue $n + 3$) and the π-helix (C′=O of residue n is hydrogen bonded to the N–H of residue $n + 5$) are less energetically favorable than the α-helix because the packing of the backbone atoms is less stable. They are a rare occurrence in protein structures and are usually present at the end of α-helices or as single turn helices.

3.2. β-Sheet

Continuous stretches of the polypeptide chain, usually 5–10 residues, can form β-strands and these strands hydrogen bond to other β-strands from a different region of the chain. When every adjacent strand is orientated in the same direction, in terms of the N- and C-terminus, a parallel β-sheet is formed, whereas opposing strands form an antiparallel β-sheet. Mixed sheets contain a combination of parallel and antiparallel strands. The hydrogen bonding patterns for the two arrangements are very different and shown in Figure 6. The backbone in a β-sheet is in an almost fully extended conformation, with ϕ and ψ angles of approximately -119^o and 113^o respectively for parallel sheets and -139^o and 135^o respectively for antiparallel β-sheets. The dihedral angles of the β-sheet backbone result in a pleated appearance with a slight right-handed twist. Successive C_α atoms lie a little above and little

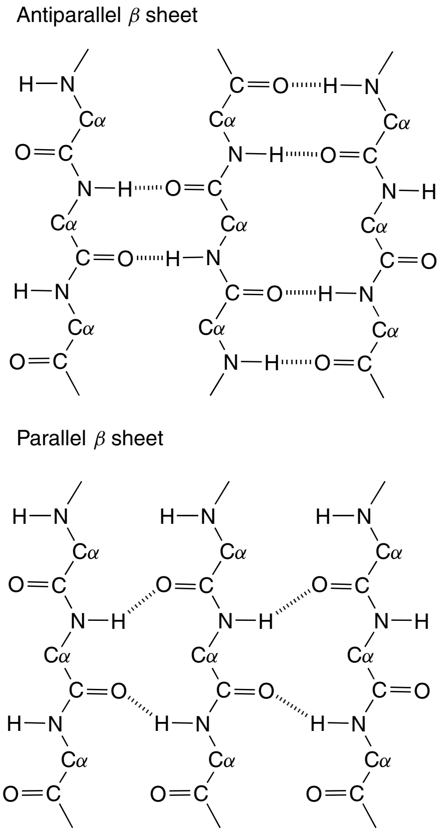


Figure 6 Hydrogen bonding in a parallel and antiparallel β-sheet

below the plane of the sheet and the side chains alternate below and above the sheet.

3.3. Other secondary structures

Proteins structures usually consist of a stable, densely packed core made up of well-defined secondary structure elements. The secondary structure elements are often connected by nonrepetitive elements such as tight turns, β-bulges, and random coil structures. These structures, which are often irregular and varied in length, are usually found exposed to the solvent on the surface of the protein and as a result they are more likely to contain residues that are polar or charged.

Tight turns (also called *hairpin loops* or *reverse turns*) are important aspects of protein structure both structurally and functionally. A tight turn often connects two adjacent antiparallel β-strands and their location on the protein surface suggests their involvement in molecular recognition. There are five types of tight turns depending on the number of atoms forming the turn. The δ-turn is the smallest of the tight turns and involves only two amino acid residues. The γ-turn involves

Table 1 Normalized frequencies of amino acid residues in different secondary structures. Data taken from Levitt M (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*, 17, 4277–4285

Amino acid	α-helix	β-sheet	Turn	Preferred location
Ala	1.29	0.90	0.78	α-helix
Cys	1.11	0.74	0.80	
Leu	1.30	1.02	0.59	
Met	1.47	0.97	0.39	
Glu	1.44	0.75	1.00	
Gln	1.27	0.80	0.97	
His	1.22	1.08	0.69	
Lys	1.23	0.77	0.96	
Val	0.91	1.49	0.47	β-sheet
Ile	0.97	1.45	0.51	
Phe	1.07	1.32	0.58	
Tyr	0.72	1.25	1.05	
Trp	0.99	1.14	0.75	
Thr	0.82	1.21	1.03	
Gly	0.56	0.92	1.64	Turn
Ser	0.82	0.95	1.33	
Asp	1.04	0.72	1.41	
Asn	0.90	0.76	1.28	
Pro	0.52	0.64	1.91	
Arg	0.96	0.99	0.88	

three amino acid residues, a β-turn four amino acid residues, the α-turn five amino acid residues and the largest of the tight turns, the π-turn, involves six amino acid residues. A turn that contains more than seven residues is called a *loop*.

There is a strong preference for specific types of amino acids to form specific secondary structure elements. The amino acid preferences for the α-helix and the β-sheet, summarized in Table 1, are an important contributor to early methods of secondary structure prediction from amino acid sequence.

4. Tertiary and quaternary structure

Secondary structure elements and connecting loops combine to form a protein's tertiary structure or fold. This tertiary structure, formed from only a single polypeptide chain, may be sufficient for the protein to carry out its biological function, although in many cases, two or more tertiary structures must come together to attain the protein's biologically relevant oligomeric state. This complex of polypeptide chains is called the quaternary structure of a protein.

Both tertiary and quaternary structures comprise of smaller functional units. A single polypeptide chain can be divided into one or more functional units called domains. Domains may be defined in a variety of ways: as stable structures able to fold and exist independently, as geometrically distinct parts of a 3D structure, as units with a single identifiable function (for example, DNA binding), or simply as

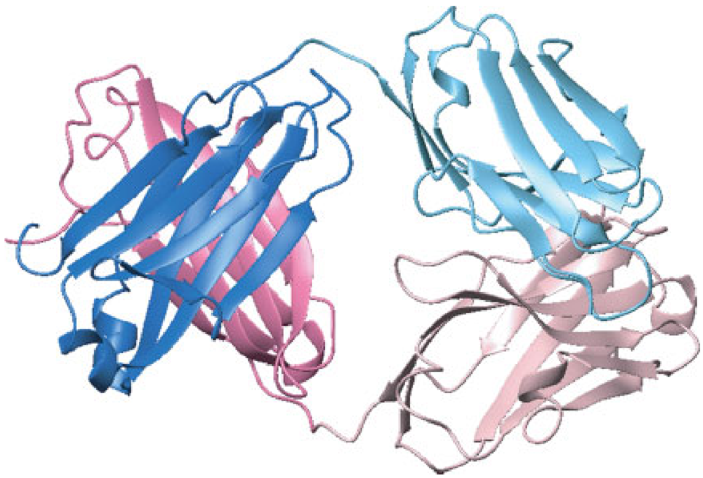


Figure 7 The crystal structure of the monoclonal anti-lysozyme antibody D1.3. The light chain (shown in blue) has two distinct domains, the constant domain (dark blue) and the variable domain (light blue). The heavy chain, shown in pink, also has two distinct domains shown in dark pink (constant domain) and light pink (variable domain). PDB code: 1fld

units observed to occur independently in many otherwise unrelated proteins. The variable and constant domains of the immunoglobulins are illustrated in Figure 7.

In quaternary structure, a polypeptide chain is itself a functional unit called a subunit. Subunits and domains are often equivalent in different species. For example, in mammals a single polypeptide chain of seven domains has evolved to carry out the seven different reactions required for fatty acid synthesis, whereas the same process requires seven different protein subunits in plants (*see* Article 41, **Investigating protein–protein interactions in multisubunit proteins: the case of eukaryotic RNA polymerases**, Volume 5). Indeed, this observation has been used by some as a means to predict protein–protein interactions (*see* Article 45, **Computational methods for the prediction of protein interaction partners**, Volume 5).

Folding of globular proteins is primarily driven by the "hydrophobic effect", where hydrophobic residues aggregate at the core of the protein away from water and polar residues form the solvent accessible surface. Stability of a fold is also achieved by balancing all the other opposing forces that occur between different parts of the protein structure. Attractive electrostatic (also called *ionic bond*, *salt linkage*, *salt bridge,* or *ion pair*) forces exist between two protein groups of opposite charge. There are three main types of electrostatic force: charge–charge (between oppositely charged R-groups), charge–dipole (e.g., interaction of ionized R-groups with the dipole of the water molecule), and dipole–dipole (noncovalent associations between electrically neutral molecules such as van der Waals forces). Hydrogen bonds are extremely important forces in protein structures; they are a permanent electrostatic interaction between a weakly acidic donor group (hydrogen atom with a partial positive charge) and an acceptor atom that bears a lone pair of electrons (negatively charged). In biological systems, both the donor and acceptor can be

highly electronegative, such as nitrogen and oxygen atoms and occasionally sulfur. Van der Waals forces are nonspecific and arise when any two atoms are 3 to 4 Å apart. Although they are weaker and less specific than hydrogen and electrostatic bonds, they become significant when numerous atoms in one of a pair of molecules come close to many atoms of the other. Attractive van der Waals forces are a result of induced dipoles that arise from fluctuations in the charge densities occurring between adjacent uncharged nonbonded atoms. Finally, disulfide bonds are covalent interactions between two cysteine residues, they are frequently found in extracellular proteins where they stabilize the protein against the hostility of the environment such as fluctuating pH and/or temperature.

Integral transmembrane proteins are characterized by predominantly hydrophobic segments that span lipid membranes connected by hydrophilic loops found outside the membrane; thus, forces that control their folding are different to those experienced by globular proteins. Most membrane-spanning segments such as those of G-protein-coupled receptors and the voltage-gated potassium channel are α-helical, while others such as porins form β-barrels. However, examples of X-ray structures of integral transmembrane proteins are rare because their lipid solubility makes them difficult to crystallize.

5. Relationship of sequence to structure

In general, proteins with similar sequences adopt similar structures. Sander and Schneider (1991) used a database of proteins in their normal biologically active confirmations to derive an equation to calculate the percentage sequence identity threshold above which most structures are similar. This equation was subsequently updated by Rost (1999) using more recent data and is shown below:

$$t(L) = 480L^{-0.32(1+e^{-L/1000})} \qquad (1)$$

where L is the length of the alignment and t is the percentage identity threshold. Equation (1) tells us that two protein sequences are likely to share the same basic 3D structure if more than 32% of residues are identical over an alignment length of 80 residues or more. Below this threshold, and certainly below 25%, sequence analysis will not always reveal evolutionary relationships between two proteins. Indeed, many borderline cases exist between 20 and 30% sequence identity; a region described as the "twilight zone" by Doolittle (1986). There have even been examples where two proteins of sequence identity less than 10% are related. Thus, structural comparisons are better than sequence alignments in detecting distant evolutionary relationships between proteins.

These results only apply to naturally occurring proteins of known structure. Some artificially engineered proteins have been proved to have different structures with sequence similarity of as much as 50%. Also, in processes such as amyloid aggregation, proteins can adopt various conformations, following the rules of thermodynamics, under different conditions of temperature, pressure, pH, concentration, or ionic strength.

6. Relationship of structure to function

A protein's 3D structure often reflects the function of that protein. For example, the helix-turn-motif structure of many prokaryotic transcription factors allows them to bind the major groove of DNA. Also, the segments of proteins that span membranes, such as voltage-gated potassium channels, are often α-helical. There are numerous ways of describing the function of a protein. For example, trypsin is a serine endoprotease but can also be described as an enzyme that participates in the degradation of other proteins. A number of attempts have been made to classify protein function but probably the most comprehensive has been the Gene Ontology (GO) scheme (**http://www.geneontology.org**). Ontologies are structured, controlled vocabularies that GO collaborators hope to apply to gene products across different databases and organisms. Three ontologies are being developed to describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species independent manner. For example, actin takes part in muscle contraction (biological process) by binding to myosin (molecular function) within a muscle fiber (cellular component). The ontologies are nonhierarchical and flexible because the relationship of gene product to each of these three categories is usually one-to-many, and descriptions of a gene product in the literature are often incomplete.

Knowledge of a protein's 3D structure is essential to understanding its function. Indeed, for those proteins whose function has yet to be assigned, there has been considerable effort to infer function from 3D structure alone. The most useful protein structures are complexes, where proteins are observed in interaction with other proteins, nucleic acids or smaller molecules, because they reveal the location of the key functional areas including the interaction site. Calculating the geometric and physical–chemical properties at these sites helps determine the molecular mechanism of function at the atomic level. Even without a complex structure, useful information can still be sought from an individual protein structure such as which residues are located in the structural core or in contact with solvent.

7. Evolution of structure and function

To understand the relationship between a protein's structure and function fully, consideration of that protein's possible evolutionary history is essential. Proteins evolve through mutations in their genes, sometimes in combination with more dramatic genetic events. The most common type of mutation is a point mutation involving substitution of a single nucleotide. Deletion or insertion of one or more base pairs can also occur. Only mutations that result in neutral or advantageous changes are permitted, so residues playing a critical role in structural stability or function are generally conserved. Exceptions to this rule include the major histocompatibility (MHC) molecules of the immune system, which rely on high mutation rates at their binding site to adapt to foreign proteins. Buried residues

evolve more slowly than residues on the protein surface because the efficient packing of the protein's hydrophobic core must be kept intact. Usually, it is the secondary structural elements such as helices and strands that reside in the hydrophobic core, alternating with loops found predominantly at the surface. Therefore, loops evolve more rapidly than secondary structure elements. The functional response to the effect of mutations at different locations on the protein can often be explained by the changes in structure that the mutations cause.

Point mutations usually lead to only minor improvements in function. More substantial genetic changes such as gene duplication, domain recombination, domain swapping, or migration are less common although they often confer a major and immediate selective advantage on the organism. Gene duplication produces two copies of a protein domain that either evolve separately or fuse to produce a homodimer, which then undergoes mutation and selection. Most gene duplications probably result in the production of a pseudogene that is without function and not expressed, but occasionally a new and advantageous function may emerge. The regulatory protein α-lactalbumin is thought to be the product of duplication of a gene encoding the enzyme lysozyme. Subsequent mutations changed key functional residues but retained the overall structure of the protein, resulting in two homologs carrying out completely different functions.

Two unrelated domains can also combine to enhance a protein's function in a process called domain recombination. Many regulatory domains of enzymes such as the plasma proteases have arisen through a combination of duplication and recombination events. Some monomers have acquired oligomeric properties by domain swapping – the mutual exchange of domains or structural elements between each molecule of a dimer. These structures, including some members of the ribonuclease family, are often in dynamic equilibrium with their monomers, thus providing a mechanism for changing quaternary structure via small changes in sequence. Occasionally, functional residues, especially catalytic residues in enzymes such as the α/β hydrolases, can migrate to different structural elements making function prediction on the basis of homology difficult in these cases.

Proteins are subject to both divergent and convergent evolution. Proteins related by divergent evolution derive from a common ancestor and are called *homologs*. In extreme cases, divergent evolution can operate until sequence similarity between homologs is almost undetectable. For example, members of the globin family, including hemoglobin, myoglobin, and plant leghemoglobin, adopt the same overall structure and carry out oxygen transport by the same mechanism, but often share less than 20% identical residues. This demonstrates that structure evolves much slower than sequence and with the wealth of new structures now available, many evolutionary relationships that could not be identified by sequence have recently been uncovered. Proteins related by convergent evolution derive from different ancestors but arrive at the same structural solution to achieve a particular function. As a result of convergent evolution, both chymotrypsin and bacterial subtilisin share the same Ser-His-Asp catalytic triad and function as serine endopeptidases via the same catalytic mechanism. However, they are completely unrelated at the level of overall tertiary structure.

8. Protein interactions

8.1. Molecular recognition

There are at least three types of protein contact: crystal contacts, contacts between protein subunits, and contacts between two interacting molecules. Crystal contacts are nonspecific and an artifact of the crystallographic process used to elucidate protein structure; they are not biologically relevant and will not be dealt with here. Contacts between protein subunits provide the basis for quaternary structure as we have seen. Therefore, for the rest of this chapter, we are concerned with contacts between two interacting molecules, at least one being a protein. All metabolic and regulatory processes such as enzyme catalysis, molecular transport, signaling, the immune response, and gene expression involve protein interactions.

The cell is usually a crowded environment, so a protein has many potential binding partners. All these molecules have a certain specificity and affinity for each other. Specificity is the capacity to select between binding partners, especially those that have similar surface properties, whereas affinity is the capacity to interact with the chosen binding partner. Most proteins, such as enzymes and their inhibitors, have only one or two binding partners (highly specific), although some have multiple binding partners (multispecific). Multispecificity can arise when members of a protein family recognize a specific pattern on the target protein. For example, SH2 and SH3 domains bind to proteins with phosphotyrosine and proline-rich sequences, respectively. Different protein functions require different levels of specificity and affinity, so a tight control of these factors is necessary. Two proteins that have high affinity towards each other may not necessarily have high specificity, as with the interactions involving SH2 and SH3 domains.

When a molecule interacts with another molecule, an interface between the two molecules is formed. This interface is defined as the area of the protein surface that is buried (removed from contact with water) as the interaction takes place. Interfaces between participating molecules must be complementary in both shape and physical chemistry. Protein interactions are therefore driven by molecular recognition. Molecular recognition is defined as the interpretation of all the shape and physical–chemical information required by a molecule to select and interact with its binding partner.

The molecular recognition process begins with a protein and its binding partner completely surrounded by a solvent (usually water) but in close proximity. Proteins interact with their solvent, so some of these interactions must be broken if a complex is to be formed, a process called *desolvation*. Unfortunately, van der Waals, electrostatic, and hydrogen bonds that form between a protein and its binding partner are generally not as favorable as those between protein and water, so the energy barrier desolvation provided is difficult to overcome. As one would expect, desolvating nonpolar parts of the protein surface is easier than desolvating polar and charged areas. Desolvation of nonpolar areas releases water molecules and therefore increases solvent entropy. This hydrophobic effect is a major driving force behind complex formation, just as it is behind protein folding, and partially compensates for the desolvation penalty. The proteins themselves will also be subject to energy changes due to two molecules effectively becoming one.

Translational and rotational entropy will decrease, and there will also be changes in vibrational entropy. Some conformational entropy will also be lost because of the restriction of dihedral angles. The complex conformation usually puts significant strain on the individual proteins involved, which means that their internal energies will be higher in the complex than when apart. Nevertheless, the resulting complex conformation is usually the most energetically favorable that can exist.

8.2. Types of protein interaction

There are three main types of protein interaction: protein–small molecule, protein–DNA, and protein–protein. Small molecules, in this sense, include ATP, GTP, metabolites, and some inhibitors or drugs. Proteins binding DNA include promoters, repressors, and transcription factors; their binding sites are characterized by large, positively charged electrostatic patches on the protein surface, matching the negatively charged DNA phosphate backbone groups. Protein–protein interactions usually involve two or more polypeptides of lengths greater than 20 amino acids. Recently, Nooren and Thornton (2003) have described two ways of classifying protein–protein interactions based on the components and lifetime of the complex. Interactions between identical chains occur in homo-oligomers (Figure 8), whereas those between nonidentical chains occur in hetero-oligomers. Most homo-oligomers are obligate, meaning that their individual components cannot exist as stable structures independently in vivo. By contrast, most hetero-oligomers are nonobligate (Figure 9) in which each component can exist as a stable structure under physiological conditions. Interfaces on obligate complexes are generally large and hydrophobic. Nonobligate complexes exhibit a more polar interface, because

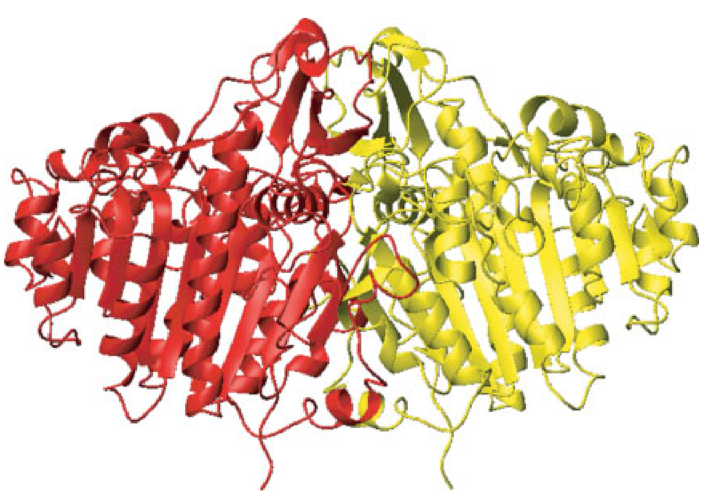


Figure 8 Alkaline phosphatase: an example of a homodimeric, obligate interaction. PDB code: 1b8j



Figure 9 Complex of acetylcholinesterase (red) with its inhibitor fasciculin II (yellow): an example of a heterodimeric, transient interaction. PDB code: 1fss

the monomers have to function independently. A large exposed hydrophobic patch would be unstable, possibly leading to aggregation events.

Interactions can be further classified as permanent, weak transient, or strong transient according to the lifetime of the complex. Permanent interactions are very stable and mostly, but not always, occur in obligate complexes. Weak transient interactions occur between two proteins that need to associate and dissociate continuously in vivo. The local concentration of the protein components, ions and solutes, and changes in pH and temperature all influence weak transient interactions. Strong transient interactions are more stable so require a molecular signal to initiate association or dissociation. For example, guanosine triphosphate (GTP) binding to G proteins causes their dissociation into Gα and G$\beta\gamma$ subunits, but the dephosphorylation of GTP to guanosine diphosphate (GDP) initiates the trimer to reform. A transient interaction may become permanent under certain cellular conditions but usually the type of interaction is inferred by the function of the protein. For instance, intracellular signaling requires rapid association and dissociation so interactions are transient.

The different types of protein interactions, and the diverse range of protein biological functions illustrate that a protein's function and structural properties are intrinsically linked. This tailoring of structure to function has been brought about by evolution over millions of years. A detailed understanding of how this diverse set of structures are assembled from amino acid sequences will provide yet further insights into evolutionary history and function. The articles in the remainder of this

section describe the recent developments and improvements in methodology that will accelerate this process of discovery in the coming years.

Further reading

Branden C and Tooze J (1999) *Introduction to Protein Structure*, Second Edition, Garland Science: New York.

Jones S and Thornton JM (1996) Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences (USA)*, **93**, 13–20.

Kinch LN and Grishin NV (2002) Evolution of protein structures and functions. *Current Opinion in Structural Biology*, **12**, 400–408.

Lesk AM (2001) *Introduction to Protein Architecture*, First Edition, Oxford University Press: Oxford, UK.

Patthy L (1999) *Protein Evolution*, First Edition, Blackwell Science: Oxford, UK.

Voet JG (2004) *Biochemistry*, Third Edition, John Wiley & Sons: New York.

References

Doolittle RF (1986) *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences*, University Science Books: Mill Valley, CA.

Nooren IMA and Thornton JM (2003) Diversity of protein-protein interactions. *The EMBO Journal*, **22**, 3486–3492.

Pauling L and Corey RB (1951) Atomic coordinates and structure factors for two helical configurations of two polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, **37**, 235–240.

Ramachandran GN and Sasisekharan V (1968) Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, **23**, 283–438.

Rost B (1999) Twilight zone of protein sequence alignments. *Protein Engineering*, **12**, 85–94.

Sander C and Schneider R (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Structural genomics – expanding protein structural universe

Andrzej Joachimiak
Midwest Center for Structural Genomics and Structural Biology Center, Argonne, IL, USA
University of Chicago, Chicago, IL, USA

1. Introduction

Microbial, viral, animal, and plant genome projects have generated enormous interest and have provided a critical milieu of data into the elucidation of all cellular components, their interactions, and regulation. The Human Genome and other genome sequencing projects have also accelerated the pace of discovery in biology and medicine. In the past several years, accumulation of sequence data has increased significantly. 1421 genome projects are currently underway (**http://www.genomesonline.org/**). The sequences of 263 genomes have been completed, annotated, and made available to the public. The phylogenetic diversity of the genomes that have been sequenced is increasing. Sequencing of 489 eukaryotic genomes is underway, and the sequence of a diatom genome, a unicellular photosynthetic, and eukaryotic algae was recently published (Armbrust *et al*., 2004). For the first time, the entire set of proteins and RNA coded by the genome have become available for experimental and theoretical studies. Genomic analysis has addressed numerous questions such as a minimal set of genes required for performing essential biochemical functions (Mushegian and Koonin, 1996; Koonin, 2000), the total number of protein families coded by genomes (Pearl *et al*., 2005; Grant *et al*., 2004), and the relationship between protein sequences and biochemical functions (Andrade *et al*., 1999; Todd *et al*., 2001). At the same time, biologists have raised questions about the structural repertoire of proteins coded by genomes and a basic set of proteins folds utilized by living organisms (Wolf *et al*., 2000; Qian *et al*., 2001; Holm and Sander, 1996; O'Toole *et al*., 2003).

In the past 50 years, biologists have focused primarily on the proteome components of a few model organisms. Since 1980, recombinant DNA technology has allowed in-depth investigation of virtually all cellular components, and these capabilities were further expanded by the availability of complete genome sequences. As a result, the most fundamental processes of the living cell are quite well characterized (e.g., translation, most fundamental metabolic pathways, transcription regulation, replication, RNA splicing, chaperone-mediated protein folding) (Harrison, 2004). Some important gaps remain in our knowledge, on average, nearly 50%

of all genes in a typical genome are described as uncharacterized or hypothetical and 10–15% of genes have no sequence relative (ORFans/singletons) (Liu *et al*., 2004; Jordan *et al*., 2001; Jordan *et al*., 2004; Siew and Fischer, 2004).

The complete understanding of the functions of biological systems requires detailed knowledge of proteins and their interactions with other proteins and cellular components. To carry out their functions, including molecular recognition and catalysis, proteins require a correctly folded 3D structure (Erlandsen *et al*., 2000; Dobson, 2003). Therefore, the knowledge of a protein structure is essential. However, a protein structure not only contributes to comprehending underlying chemical principles of protein function but also adds to our understanding of the evolution of biological systems, since the protein sequence does not always allow the prediction of structure and function (Pearl *et al*., 2005; Bray *et al*., 2000).

The multitude and complexity of molecular interactions, and their temporal and transient nature, requires basic information about the structures of individual components as well as of the complete functional assemblies (Saibil, 2000; Halic *et al*., 2004; Dormitzer *et al*., 2004). The individual proteins are very important in their own right as agents of health and disease or as tools with economically significant applications (Eckhardt *et al*., 2004; Zweiger and Scott, 1997). In the last four decades, structural biology, thanks to the generous public and private support of hypothesis driven research, has been quite successful at addressing fundamental mechanistic and functional questions in biology – we understand the most basic principles of protein structure. The 3D structures of many important classes of proteins, including enzymes and receptors, and the structures of at least one member of approximately half of the largest protein superfamilies are available. Structures of several important large protein assemblies are known as well (Ban *et al*., 2000; Abrahams *et al*., 1994; Murakami *et al*., 2002; Simpson *et al*., 2000; Schlunzen *et al*., 2001).

Although many proteome components of the fundamental processes of the living cell have been characterized structurally, the number of specific structures from individual species is very low, except for a very few cases such as *Escherichia coli*. Homology modeling coverage is higher, and can reach 40–50% of all proteins (Goldsmith-Fischman and Honig, 2003; Zhang and Skolnick, 2004; Eswar *et al*., 2003); however, the models need better accuracy to be applicable for broad biotechnology and biomedical applications. One of the major challenges in structural biology is, therefore, to determine structures of novel proteins encoded by genomes and to increase structural coverage of proteomes in a rapid and cost-effective manner. The vast sequence data provided by genomes made clear that "traditional" structural biology approaches were insufficient to address such a large-scale problem, and at the current pace it would take 40–50 years to complete such a project.

2. Early structural genomics efforts

In the late 1990s, a number of structural biology and bioinformatics groups in North America, Europe, and Asia, stimulated by the rapid progress of genome programs, have initiated small pilot projects. These efforts have tested the idea of using predicted protein sequences from genomes to provide a strategy to

determine structures of all, or at least most, of the structures of proteins encoded by genomes. The important point for the North American efforts was the Argonne's 1998 Structural Genomics meeting (Shapiro and Lima, 1998; Gaasterland, 1998). This meeting brought together researchers and representatives of funding agencies who thought that improvements in technology, combined with the successes of the genome sequencing projects, had set the stage for a large-scale structure determination project. The overall strategy for covering protein structures space was discussed.

X-ray crystallography of macromolecules, in particular, has seen remarkable progress in recent years. These advances were made possible by the public investment in the development of third-generation synchrotron sources (Cassman and Norvell, 1999), initially in European Union, United States, and Japan, and later in other countries, and dedicated beamlines receiving X rays from insertion devices and equipped with modern X-ray detectors (Phillips *et al*., 2002; Westbrook and Naday, 1997). A recent survey of the Protein Data Bank (PDB) (Westbrook *et al*., 2003) showed that contribution of X-ray structures obtained using synchrotrons to the PDB increased from 19% in 1995 to 75.9% in 2004 (Figure 1) (Jiang and Sweet, 2004). The analysis also confirmed that synchrotron sources extend the size of macromolecules for which structures can be solved with equal or better quality, than for other X-ray sources. Moreover, synchrotron facilities accelerated data collection (Jiang and Sweet, 2004; Walsh *et al*., 1999a,b; Hendrickson, 2000), lowered the dependence on large crystals for high-quality data (Jiang and Sweet, 2004), and facilitated new approaches in structure determination (Hendrickson, 1991; Moffat and Ren, 1997; Terwilliger, 1994). The third-generation synchrotron sources also brought a new dimension to protein crystallography by allowing very efficient execution of the structure determination methods that utilize a so-called anomalous signal (multiwavelength (MAD) and single-wavelength (SAD) anomalous diffraction). The PDB survey showed that in 1997 only 12% of all experimentally determined structures were phased using the SAD/MAD approach; this number exceeded 70% in 2003 and 2004 (Figure 2).

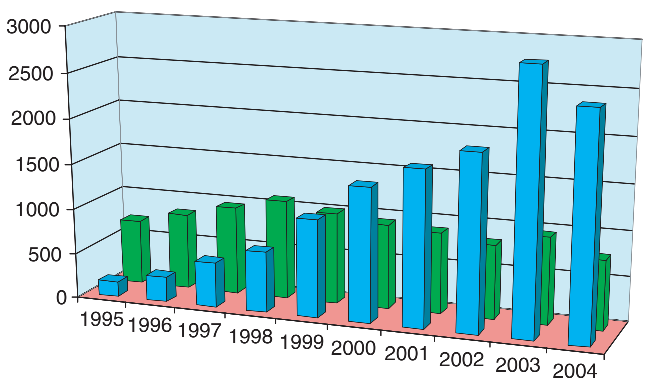


Figure 1 Contribution of structures from synchrotron (blue) and "home" (green) sources to PDB (Jiang and Sweet, 2004)

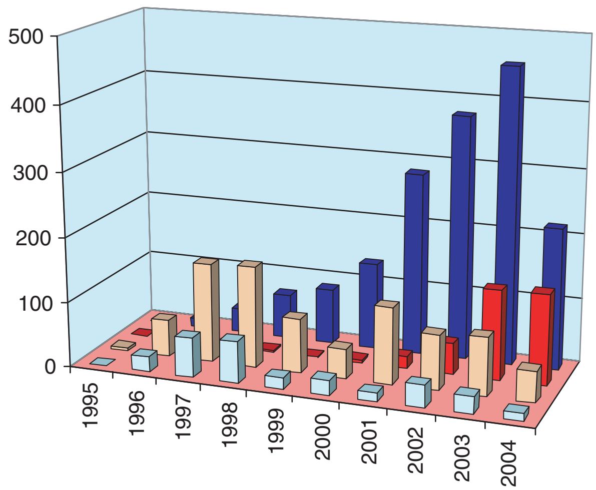


Figure 2 Experimental determination of structures using traditional methods (SIR (light blue), MIR (pale)), and methods utilizing anomalous signal (SAD (red) and MAD (blue)) (Jiang and Sweet, 2004)

The advances in the hardware at synchrotrons could not have been exploited fully if it were not for complementary advances in the software for crystallographic data collection (Otwinowski and Minor, 1997; Terwilliger, 2002; Terwilliger and Berendzen, 1999; Terwilliger, 2004a; Holton and Alber 2004; Brunzelle *et al*., 2003), structure determination and refinement (Terwilliger, 2004a; Brunzelle *et al*., 2003; Bricogne *et al*., 2003; Morris *et al*., 2003; Potterton *et al*., 2002; Adams *et al*., 2002), and high-capacity computing and data storage resources (Brooksbank *et al*., 2005; Ackerman *et al*., 2001; Westbrook *et al*., 2003; Hallin and Ussery, 2004; Manjasetty *et al*., 2003). In fact, many of these developments were driven by the advances in Structural Genomics pilot projects (Norvell and Machalek, 2000; O'Toole *et al*., 2004; Chance *et al*., 2004) and development of synchrotron facilities.

Similar progress has been made in recent years in the field of NMR that continues to be an important contributor to the structural biology with the development of new powerful high-field spectrometers (Montelione *et al*., 2000). NMR is also being applied to difficult proteins such as large soluble and small integral membrane proteins (Kyogoku *et al*., 2003; De Angelis *et al*., 2004). In this review, I will specifically focus on the structural genomics applications in synchrotron-based X-ray crystallography.

In addition to genome sequencing, X-ray crystallography, and NMR developments, there were several major technology breakthroughs that were important for undertaking a major Structural Genomics project:

- well-established PCR-based recombinant DNA technology,
- robust high-level protein expression systems in vivo and in vitro,
- robotic systems for molecular biology,
- robotic systems for protein purification, crystallization, and crystal imaging,

- crystallization formulations for efficient screening protein crystallization conditions and application of crystal cryofreezing to data collection,
- bioinformatics tools, databases, and DNA sequence comparison capabilities.

These developments provided a robust foundation for a large-scale Structural Genomics initiative (Chance *et al*., 2004; Terwilliger, 2000) and allowed for the creation of an integrated structure determination process or a "pipeline" from gene to structure.

3. Emergence of international structural genomics initiative

A Structural Genomics initiative has emerged as a multidisciplinary and international effort to maximize the information value from delineation of the genome content of humans and other organisms (Terwilliger, 2000; Burley, 2000). The initial Structural Genomics goals were: (1) to expand the universe of protein folds by rapidly determining novel structures, (2) to produce homology-based 3D models of all proteins, to extend the coverage of sequence space, and (3) to elucidate novel biological function directly from 3D structure. The ultimate goal was to make it feasible to predict an accurate 3D protein structure on the basis of sequence alone. This will allow inference of variations in functional and 3D structural properties of proteins that result from genetic differences. Additionally, genomic analysis would provide insight into distinctive metabolic capabilities or deficiencies related to the functional capabilities of particular organisms. Development of methods to rapidly scan genetic sequences for potential biological function would impact many fields including health, biotechnology, agriculture, and bioremediation.

The scale of Structural Genomics goals is enormous, clearly comparable with the accomplishments of the past 40 years of structural biology. Maximizing the output of the structure determination pipeline and minimizing the use of valuable resources is a primary requirement for achieving all Structural Genomics goals. At the beginning of this initiative, there were less then 2500 nonredundant structures in PDB (Westbrook *et al*., 2003), and the bioinformatics analysis of available genomic sequences predicted the need to determine 15 000–20 000 unique protein structures (Vitkup *et al*., 2001; Liu and Rost, 2002; Chandonia and Brenner, 2005). This number equaled the scale of all deposits in the PDB. The international structural genomics efforts were initiated, in part, to achieve a more rapid understanding of sequence/structure relationships than would have been accomplished in a problem-focused approach. Several pilot centers have been established to test and improve existing and develop new technologies (Table 1).

The initial Structural Genomics pilot projects focused their effort on developing and establishing working, high-throughput protein structure determination pipelines (Figure 3) (Chance *et al*., 2002; Lesley *et al*., 2002; O'Toole *et al*., 2004). These pipelines include the following important components:

- Genome sequence analysis, family classification, and target selection.
- Efficient production of expression clones, proteins, and crystals suitable for structure determination.

Table 1 Structural genomics consortia (**www.isgo.org**)

The Berkeley Structural Genomics Center (BSGC)
Center for Eukaryotic Structural Genomics (CESG)
The Joint Center for Structural Genomics (JCSG)
The Midwest Center for Structural Genomics (MCSG)
The New York Structural Genomics Research Consortium (NYSGRC)
The Northeast Structural Genomics Consortium (NEGS)
The Southeast Collaboratory for Structural Genomics (SECSG)
Structural Genomics of Pathogenic Protozoa Consortium (SGPP)
The TB Structural Genomics Consortium (TB)
Bacterial Targets at IGS-CNRS, France (BIGS)
Biological Information Research Center (JBIRC)
Cyber Cell Project
deCode Genetics Israel Structural Proteomics Center (ISPC)
Korean Structural Proteomics Research Organization (KSPRO)
Marseilles Structural Genomics Program (MSGP)
Montreal-Kingston Bacterial Structural Genomics Initiative
North West Structural Genomics Centre (NWSGC)
Ontario Centre for Structural Proteomics
Oxford Protein Production Facility (OPPF)
Paris-Sud Yeast Structural Genomics (YSG)
Protein Structure Factory (PSF)
RIKEN Structural Genomics Initiative (RIKEN)
Structural Genomics Consortium (SGC)
Structural Genomics Consortium for Research on Gene Expression System (SGCGES)
Structural Genomix (SGX)
Structural Proteomics in Europe (SPINE)
Structure 2 Function Project (S2F)
Syrrx
XMTB – Mycobacterium Tuberculosis Structural Proteomics Project

- Rapid protein structure determination using synchrotron-based X-ray crystallography or NMR.
- Advanced robotic and computational systems for semiautomated, rapid data collection, processing, phasing, modeling, and refinement.
- Dissemination to public domain of Structural Genomics results and computational tools for structural, evolutionary, functional, and biomedical analyses. Important requirement included immediate deposition and release of protein structure coordinates.
- Creation of dedicated databases (**http://targetdb.pdb.org/**) (Chen *et al*., 2004) and dissemination of data, materials, and results and training in new technologies.

Initially, different strategies were pursued by various Structural Genomics consortia to determine the most effective experimental strategy (Norvell and Machalek, 2000). Identification of key bottlenecks has been a major effort in Structural Genomics, and this subject was a theme of numerous workshops and conferences (**http://www.nigms.nih.gov/psi/**). Improving selection of targets, better expression of soluble proteins, growth of X-ray quality crystals, automation of X-ray crystallography and NMR structure determination process, and automation and robotics as well as data management have been identified as key bottlenecks.

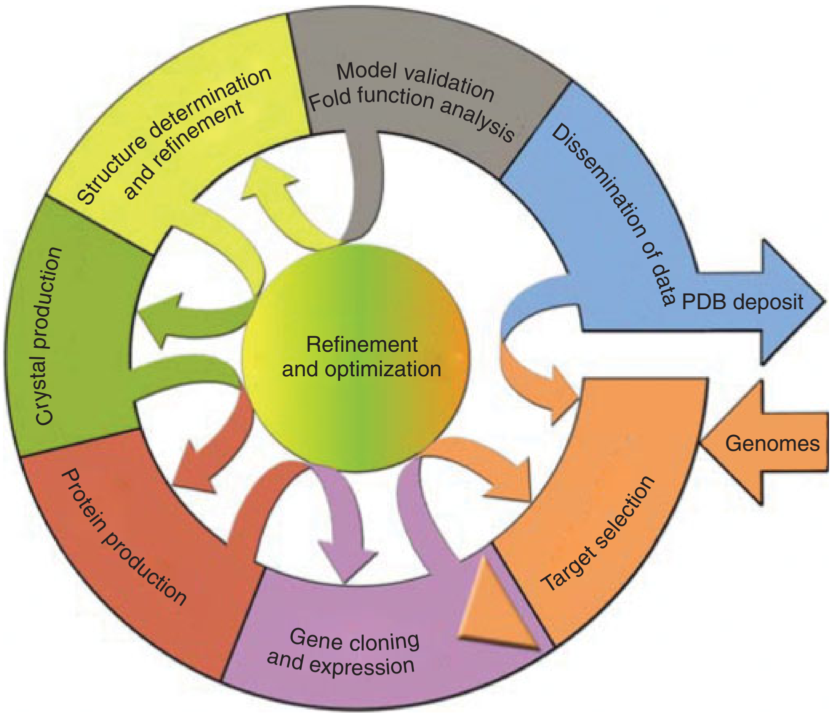


Figure 3 Schematics of the SG structure determination pipeline

At the same time, the Structural Genomics efforts coincided with the information explosion, and this benefited the initiative enormously. It allowed for rigorous and thorough correlative scrutiny of expanding primary and tertiary structural databases for functional significance, and to elucidate novel biological functions by comparison of structural information and genetic sequence databases.

To coordinate the international efforts, the international task forces have been organized to address issues ranging from informatics, target overlap, intellectual property, to data release criteria and their implementation (**www.isgo.org**). Table 1 lists the currently active Structural Genomics programs worldwide.

4. Creation of structural genomics pipelines

Some Structural Genomics pilot consortia have divided the structure determination pipeline into several distinct steps that helped to design integrated systems (Chance *et al*., 2004; Chance *et al*., 2002; Lesley *et al*., 2002). These steps are briefly described below.

4.1. Target selection

Target selection is one of the most important element of Structural Genomics programs (Liu *et al*., 2004; Canaves *et al*., 2004; Watson *et al*., 2003; Bray *et al*.,

2004; Frishman, 2005). Comparative sequence analysis allows the organization of all sequences into families (Chandonia and Brenner, 2005). This is essential for the rational selection of targets and to reduce the number of structures that need to be determined. The structural biologists can select the most suitable family representatives for structure analysis, and select families with particular features (Canaves *et al*., 2004; Frishman, 2005; Kimber *et al*., 2003; Savchenko *et al*., 2003). Since it is not possible to crystallize all proteins, the analysis of protein sequences, to the extent it is possible, is needed to filter protein sequences that are known to crystallize poorly (proteins with internal transmembrane regions, highly disordered proteins, large proteins with domain repeats, etc.). Majority of current Structural Genomics programs focus on soluble proteins or domains of large proteins and use 30% sequence identity cutoff, over the whole protein sequence, with a protein of a known structure as a major selection criterion.

There are a number of other important issues in the target selection strategy. The size of the family to which the protein belongs is important (Watson *et al*., 2003). Large families are preferred over those with few members, since these ensure that many more proteins can be modeled once a template structure is obtained. However, solving the structures of ORFans/singletons, proteins found in a single organism that lack detectable homology to any other protein may be valuable for future structure prediction.

The species distribution of a protein family is also considered in target prioritization (Galperin and Koonin, 2004). Families whose members span phylogenetically distant lineages in all three kingdoms of life are of high interest since they may have an unknown but fundamental function that the solved structure might help define.

Defining the level of detail, or granularity of the structural coverage is an important issue for target selection. Some families may be covered at course granularity, for others more structures and fine granularity coverage would be needed (>=30% sequence identity) in order to produce more accurate models for studying the biological function (Vitkup *et al*., 2001; Chandonia and Brenner, 2005).

Human targets are also important since these structures may have significant biomedical implications.

Various approaches that are used for the prediction of distant evolutionary relationships between all remaining sequences and proteins in the PDB are important. Evolutionary trace analysis is a powerful method for identifying which residues are functionally important and can aid target selection (Yao *et al*., 2003; Joachimiak and Cohen, 2002). Since tertiary structure is better conserved than sequence, structure determination often reveals very distant evolutionary relationships that may yield valuable functional insights.

At the end of the process, a prioritized target list is produced and forms the basis for the experimental work. Analysis of the Structural Genomics success and failure data may help further prioritize certain targets if their sequence suggest they might be more amenable to structure determination – for example, they are more likely to be soluble and crystallize (Canaves *et al*., 2004; Kimber *et al*., 2003).

At present, this analysis is essential but in the future the Structural Genomics plans ultimately to address all protein targets. Currently, there are some protein classes that may not be fully compatible with the Structural Genomics pipelines,

primarily because they are unlikely to be expressed in folded and soluble form in bacterial expression systems.

4.2. Gene cloning and protein expression

To meet the Structural Genomics production levels, a comprehensive 96-well-plate high-throughput technology must be used to generate clones using PCR technology and express soluble proteins (Lesley *et al*., 2002; Dieckman *et al*., 2002). Selected target genes are passed through the automated generation of protein expression clones using liquid handlers and robotic systems. The process consists of a series of interlinked protocols representing liquid manipulations and incubations on various stations of the robotic systems. Ligation-independent cloning protocols have been implemented that simplified sample manipulations and reduced the number of steps (Stols *et al*., 2002; Moy *et al*., 2004). These automated methods provide the capability for high-throughput production of expression clones for Structural Genomics. For the majority of Structural Genomics targets, the proteins are produced at a high level using an inexpensive fermentation process in *E. coli* strains. Considerable effort has also been put into developing a more robust cell-free expression systems (Kigawa *et al*., 2004; Sawasaki *et al*., 2002; Folkers *et al*., 2004).

4.3. Protein production

A critical issue in Structural Genomics, and in structural biology in general, is the availability of many high-quality samples. "Structural-Biology-Grade" proteins must be generated in a quantity and quality suitable for structure determination experiments using X-ray crystallography or NMR. The protein must meet strict quality and quantity criteria. Typically, protein purity of >95% is required. Protein samples must be suitable for incorporation of heavy atoms or isotopic enrichment to aid structure determination. These criteria put certain restrictions on the methods and procedures of sample preparation. A number of Structural Genomics consortia have implemented automated protocols on robotic workstations capable of performing multidimensional chromatography (Chance *et al*., 2004; Lesley *et al*., 2002; Kim *et al*., 2004; Firbank, 2005).

For example, the Midwest Center for Structural Genomics (MCSG) (**www.mcsg.anl.gov**) has developed protein purification procedures that are based on the following principles (Kim *et al*., 2004):

- All proteins are expressed as a fusion with a uniform, cleavable affinity tag.
- Proteins are purified using affinity chromatography followed by buffer-exchange chromatography.
- The affinity tag is cleaved off by a specific tagged protease.
- The protein is further purified using affinity chromatography followed by buffer-exchange chromatography, compatible with protein concentration and crystallization methods.

These protocols can be executed on robotic purification workstations as shown in Figure 4.

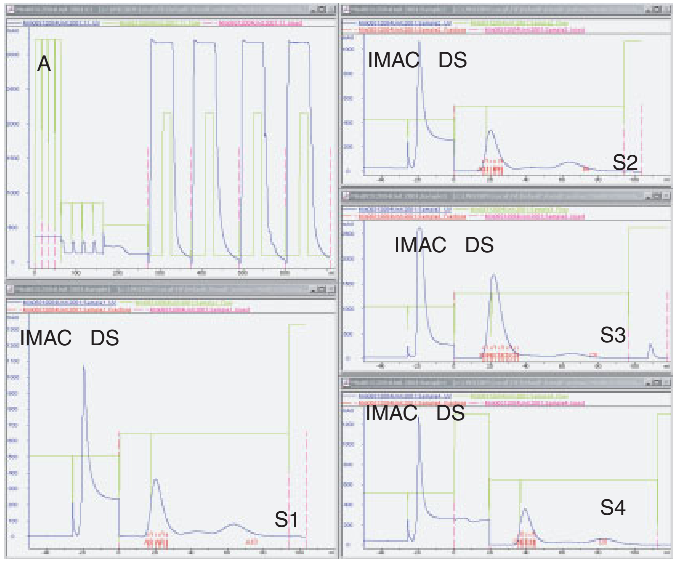


Figure 4 Example chromatograms of IMAC-I and buffer-exchange steps using AKTAexpress from the results file of a four-protein run (APC26434, APC26429, APC26423, APC1976). Load four samples on IMAC and column wash with buffer containing 50 mM HEPES pH 8.0, 500 mM NaCl, 10 mM imidazole 10 mM β-ME and 5% glycerol. (S1, S2, S3, and S4) Wash samples with buffer containing 20-mM imidazole and elution of His_6-tagged target proteins with buffer containing 250-mM imidazole followed by desalting step. In each chromatogram, UV absorbance at 280 nm is plotted versus milliliters of buffer solution flow

Table 2 Protein characterization methods used in SG

Protein parameter	Method of characterization
Purity	SDS-PAGE stained with Coomassie Brilliant Blue and lab-on-the-chip bioanalyzer
Concentration	Protein assay and UV spectrometry
Polydispersity	Dynamic light scattering
Estimated molecular weight in solution	Size-exclusion chromatography, SAXS
Suspected chemical heterogeneity and bound ligands	Mass spectrometry
Bound ligands spectra	UV/Vis spectrometry
Folding, aggregation, homogeneity	HSQC NMR

Purified proteins are typically characterized using a variety of biophysical methods, some of which are listed in Table 2.

4.4. Crystal production

With thousands of proteins and hundreds of different crystallization conditions each, it is essential to use automated crystallization and crystal visualization approaches (DiDonato *et al*., 2004; McPherson, 2004). Typically in the initial trials, proteins are

crystallized using commercially available formulations and crystallization screens. Nanoliter technology has been very successful in rapidly obtaining good quality crystals and at the same time reducing the amount of protein needed for crystallization (Hosfield *et al*., 2003). In an effort to optimize the number of crystallization conditions and maximize their effectiveness, the crystallization data are being analyzed and procedures are adjusted. Recent analysis suggests that large proteins (>60 kDa) and proteins with high pI crystallize less readily using current standard set of crystallization conditions. Such correlations are useful for further optimizing crystallization conditions and aid protein engineering for crystallization (Canaves *et al*., 2004).

Only a fraction of X-ray quality crystals are obtained in the initial crystallization screens. Majority of suitable crystals are produced after more sophisticated optimization of crystallization conditions. In X-ray crystallography, this is one of the major bottlenecks of the structure determination pipeline and is one of the most challenging tasks. A number of strategies have been developed for crystal optimization with the aid of software for experiment design and robotic solution makers and crystallization workstations (Firbank, 2005; Chayen and Saridakis, 2002). Since X-ray diffraction data collection experiments are typically carried out at cryogenic temperatures, crystals must also be cryoprotected prior to conducting the experiments (Pflugrath, 2004; Garman, 2003).

4.5. Structure determination and refinement

For Structural Genomics using X-ray crystallography, the majority of diffraction data are collected at the synchrotron sources, often with an aid of robotic workstations (Hendrickson, 2000; Brunzelle *et al*., 2003; Chance *et al*., 2002; Lesley *et al*., 2002). Typically, a protein with less than 30% sequence identity to a previously deposited structure requires phase-sensitive measurement. Since 2000, so-called anomalous scattering for protein structure determination, either as SAD/MAD, has become the most important means to structure determination and automation (Hendrickson, 2000; Hendrickson, 1991). The success of the SeMet MAD approach (Hendrickson *et al*., 1990), in which methionine is replaced in vivo by SeMet, has made SAD/MAD the most popular technique (Walsh *et al*., 1999a,b). Methionine is found in almost all proteins, and represents 2% of all amino acids. Milligram quantities of proteins in which methionines are quantitatively substituted with SeMet can be routinely produced in vivo (Stols *et al*., 2004; Kim *et al*., 2004) as bacteria, and eukaryotic cells can grow on media containing SeMet (Carfi *et al*., 2002). The substitution can also be accomplished using in vitro expression systems (Yokoyama, 2003). The SAD/MAD method has matured to the point that it has outpaced traditional experimental structure determination approaches as the method of choice for determining the structure of novel proteins (Jiang and Sweet, 2004) (Figures 2 and (5)). Although SAD/MAD relies on a small signal, it has advantages over the MIR method – all data can usually be measured from a single crystal, cryogenically cooled to obtain very accurate phase information. In retrospect, the SAD/MAD techniques currently used by synchrotron facilities truly transformed the field of protein crystallography and Structural Genomics; in fact, SAD/MAD is the key to success of Structural Genomics using X-ray crystallography.

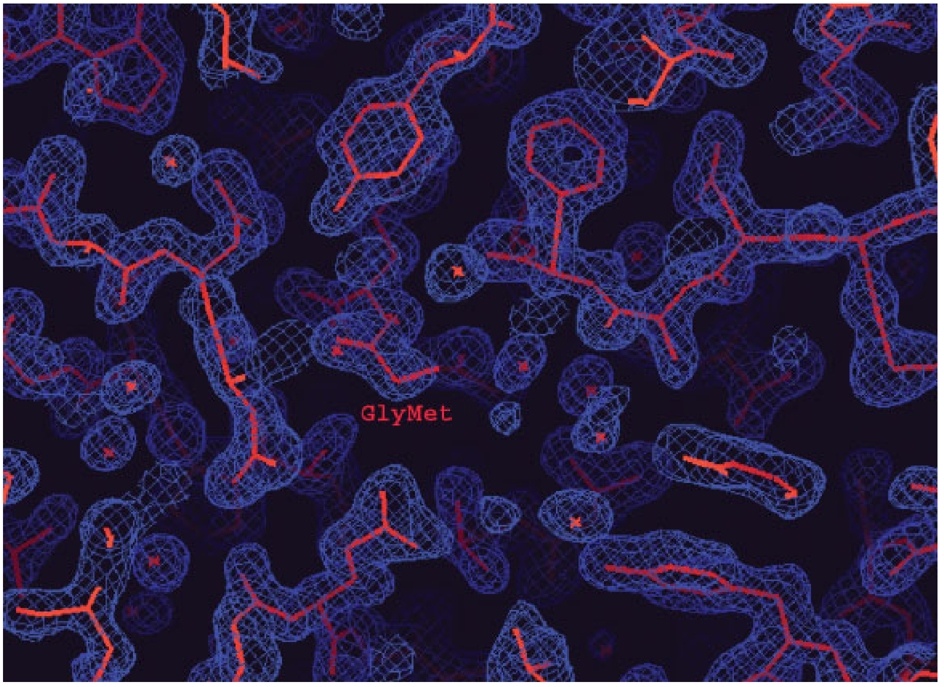


Figure 5 The electron density map for the APC009 protein obtained with 1 Se atom per 297 residues. The phases were obtained from a SAD experiment at the Se-edge. The experimental maps were calculated at 1 s after density modification (Williams *et al*., 2004)

Advances in hardware as well as in data integration and phasing software have resulted in the rapid evolution of experimental protocols. For example, the highly integrated HKL2000_PH suite allows cross talk with other programs and data analysis. This approach is changing protein crystallography from several individual steps linked manually, to a single, multistep, integrated process. The final product, a high-quality electron density map and an initial structural model, is obtained more efficiently (Otwinowski and Minor, 1997).

Some Structural Genomics consortia have developed comparable approaches to structure determination that integrate data collection, reduction, phasing, and model building – significantly accelerating structure determination and reducing the number of data sets required for a single structure determination. These applications convert diffraction data to an interpretable electron density map, and for smaller structures, into an initial model. The initial model is typically an auto-build into electron density maps using a number of excellent software packages (Terwilliger, 2004a,b; Lamzin and Perrakis, 2000; Levitt, 2001). Structures are being refined (Brunger *et al*., 1998; Potterton *et al*., 2002), and final models are still being inspected by experienced crystallographers and are adjusted manually; furthermore, there are efforts to automate these steps as well. Although this article does not address structure determination using NMR, this method and its application to structural genomics is discussed in this volume (Craven, 2005; Kalverda, 2005).

4.6. Model validation and fold/functional analysis

The final protein models are validated using well-established crystallography approaches (Laskowski, 2001; Redfern *et al*., 2005) that are being integrated into pipelines, and structures are deposited to PDB, including structure factors. The protein fold is compared with all proteins that are available in the public databases, and the protein is classified into family and superfamily. There are a number of search tools that can take a protein structure and scan it against the fold databases, and retrieve the closest matches, usually with some measure of the significance of the match (Pearl *et al*., 2005; Redfern *et al*., 2005; Holm and Sander, 1998; Andreeva *et al*., 2004).

One of the important premises of Structural Genomics is that the structure could provide important resources for understanding biochemical function of proteins. A 3D structure provides much more information than a sequence of amino acids. It reveals which residues are on the surface and available for interactions, and which are buried. The surface of a protein, its shape, and electrostatic properties, may reveal binding clefts, charged patches, and conserved regions, potentially involved in function. Clusters of conserved residues may constitute an enzyme's active site, or metal binding site, which may be recognized by comparison with a library of such sites (Redfern *et al*., 2005; Binkowski *et al*., 2003; Sanishvili *et al*., 2003; Kim *et al*., 2003). Crystal contacts may highlight specific interaction patches of biological significance. Structures often include ligands serendipitously revealing cofactors or possible substrates. The structures may also reveal distant evolutionary relationships, which often provide some functional clues, even if the function may have been altered during evolution.

5. Structural genomics contributions (thus far)

Technology development in Structural Genomics programs has been used to increase the efficiency of the protein structure determination pipeline and has also been applied to proteins that pose greater experimental challenges (Heinemann, 2001; Terwilliger, 2004b). Several Structural Genomics consortia have already built pipelines capable of determining over 100 structures per year. Specifically, technological improvements have:

- reduced the time typically needed to solve a new protein structure from months to days;
- reduced the average cost of solving the structure;
- substantially reduced attrition rates at intermediate steps;
- allowed to study new classes of proteins; and
- broadened structural coverage of the protein fold space.

The new technologies, which are now being used in academia and industry, have accelerated new discoveries and have allowed us to comprehensively address large problems in biology. The improvements that occurred during the Structural Genomics pilot stage have also permitted the Structural Genomics pipeline to

accommodate more challenging proteins. For example, the existing pipeline has been used to coexpress interacting proteins for small protein complexes, express membrane-associated proteins without membrane anchors (Zhang *et al*., 2004), express small membrane proteins and soluble domains of membrane proteins (Moy *et al*., 2004), and express domains of large human proteins.

Structural Genomics initiatives are expanding the universe of protein fold space by rapidly determining structures of proteins that were intentionally selected on the basis of low sequence similarity to proteins of known structure. Often, these proteins have no associated biochemical or cellular functions (Watson *et al*., 2003).

The majority of the Structural Genomics programs are now in the fourth year of operation. Thus far, the Structural Genomics efforts have produced nearly a quarter of all novel structures deposited into the PDB. The Structural Genomics pilot projects have already determined 1915 protein structures (vast majority in the past two years). More importantly, over 65% of these structures are unique, based on sequence identity (<30%) as compared with 13% of all deposits to PDB. By the end of 2004, there were totally 26 406 protein structures in PDB, and 4054 (15.4%) were unique. Structural Genomics programs in the past 4 years contributed 1054 or 23.3% of unique structures. More importantly, in some Structural Genomics consortia, the cost per structure has been reduced by nearly 5 times.

Structural Genomics has determined structures of many biomedically important proteins, including many proteins from pathogens. An example is provided in Figure 6, with the structure of sortase B from *B. anthracis* – an important transpeptidase involved in the attachment of proteins to cell surface (Zhang *et al*., 2004). Figure 7 shows structures of 20 DNA- and RNA-binding proteins determined just by one of the Structural Genomics centers in the United States – the MCSG. The scope of this article does not permit more detailed listing of biomedically important structures.

As expected, the determination of many unique structures have lead to the expansion of the protein universe (Hou *et al*., 2003). A number of new folds have revealed proteins with a "knot" (Zarembinski *et al*., 2003; Nureki *et al*., 2004). However, Structural Genomics showed with many examples that proteins with very low sequence similarity have this same fold (and very similar structure). In fact, this confirmed the hypothesis that the structure-based classification of proteins contains far fewer protein families than the sequence–based classification. Clearly, protein structure is better conserved than amino acid sequence, and can reveal distant evolutionary relationships that are undetectable by sequence comparisons. In fact, several structures of ORFans/singletons and proteins from very small families showed familiar folds, contradicting the hypothesis that these families may represent a rich reservoir of new folds (Koonin *et al*., 2002). Structural Genomics data indicate that sequence-based methods have major limitations for identifying proteins with potentially new folds for highly diverged sequences. Structural Genomics results have also raised new questions about the evolution of proteins and protein folds.

Structural Genomics confirmed observation that within a large family of proteins, only truly conserved residues are associated with protein function. In fact, proteins with very different folds and no sequence similarity show virtually identical active

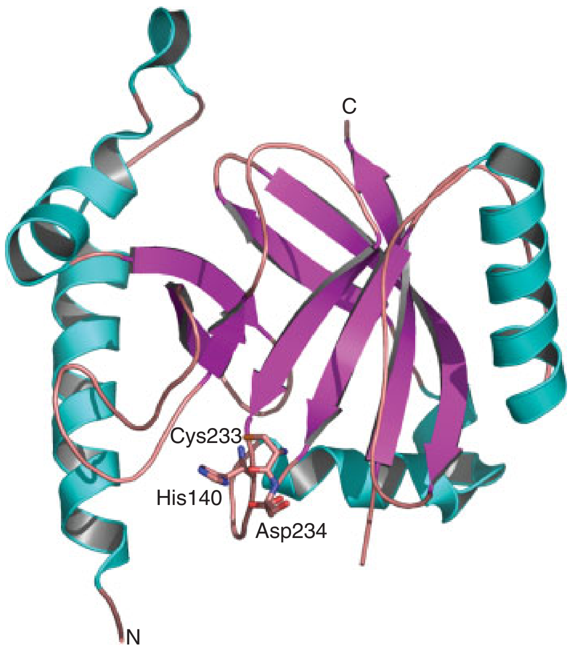


Figure 6 Crystal structure of Sortase B from *B. anthracis*. N- and C-termini and active site residues are labeled (Zhang *et al*., 2004)

sites. Proteins were discovered with different active sites which appear to carry out the same reaction with similar kinetic parameters (Zhang *et al*., 2003a,b).

A number of cofactors have been observed as a part of protein structure and in some cases, we have been able to infer the biochemical and molecular function of the protein directly from the structure (Zhang *et al*., 2002).

The Structural Genomics initiative has already made available numerous profits to structural biology by providing:

- expanded coverage of protein fold space with a dictionary of high-resolution protein structures determined experimentally by synchrotron-based X-ray crystallography and NMR;
- new structures of proteins with medical importance;
- new reagents for broad use;
- a comprehensive library of recombinant protein expression clones representing protein structures and functions;
- new high-throughput technologies for the currently powerful but tedious and labor-intensive protocols of molecular biology, protein expression, purification, crystallization, and structure determination;
- some functional information derived from structure; and
- comprehensive Structural Genomics databases.

As the field of structural genomics matures and accumulates a large volume of data, it will represent nonreductionist cell biology research, making it possible to

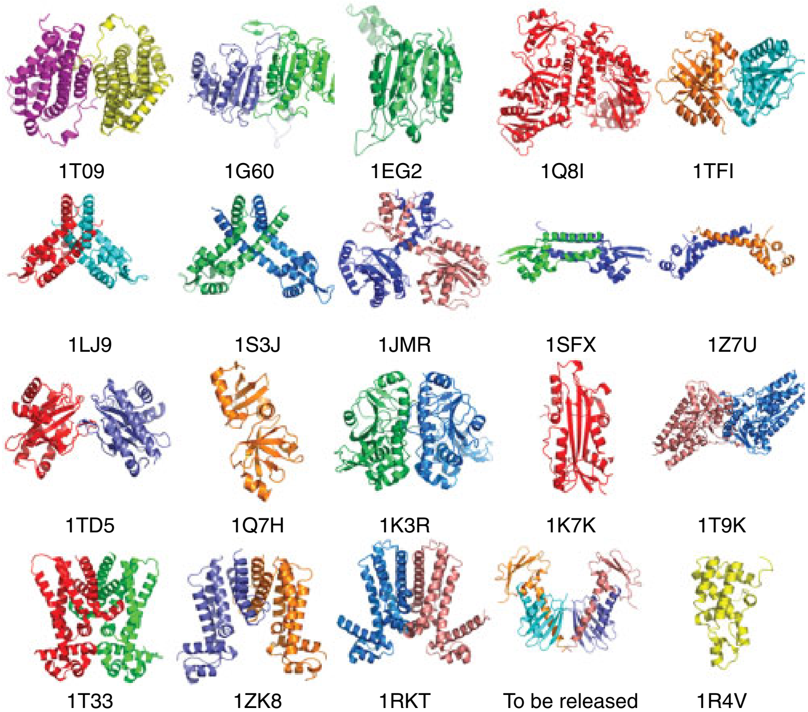


Figure 7 Models of DNA- and RNA-binding proteins determined at the MCSG

develop hypotheses at the integrated cell level. For this to occur, a high degree of completeness in our knowledge about protein sequence, structure, biochemical function, and biochemical relationships will be necessary. Importantly, many of the tools required for future genomics research are already being developed by structural genomics efforts. It is expected, as illustrated by recent progress of the Structural Genomics initiative, that this field will continue to accelerate the process of scientific discovery. There are a number of challenges that need to be addressed in order to fulfill the ultimate goal of Structural Genomics, to provide complete structural coverage of all major protein superfamilies. These challenges include:

- improved target selection to deliver maximum benefit from each solved structure and provide nearly complete coverage of the protein fold space;
- high-throughput pipelines for all protein classes;
- improved protein expression and solubility, especially for eukaryotic and membrane proteins;
- designing new expression systems to improve protein expression, folding, optimization of target and solubility, including eukaryotic proteins;
- new insights into the structure–function relationship using computational tools;
- improvement of knowledge-based sequence-to-structure and structure-to-function predictions;
- evolutionary links that can give clues to unknown structure/function relationship;

- development of new function-recognition techniques;
- improving methods for target selection and analysis, including domain parsing, identification of signal sequences, and prediction of transmembrane and disordered regions;
- improving crystallization of proteins using orthologs, gene shuffling, surface mutagenesis, chemical modification, and reducing heterogeneity caused by partial ligand occupancy, with the aid of genomic database mining;
- automating crystal screening and optimization of data collection using robotics and automation;
- integrated structure determination and refinement;
- automating structure verification, fold, and functional analysis.

6. Conclusion

The Structural Genomics programs have expanded rapidly and accelerated the structure determination process. Structural Genomics will grow from its goal to produce a collection of individual structures to address the more complex structural biological questions. It will take advantage of all relevant biology data to maximize structural output and impact. Ultimately, structural genomics will become a major component of, and contributor to the next structural biology challenge – visualization of the key stages in reconstruction of all major subcellular macromolecular assemblies and the whole cell. Addressing such complex systems will require exceedingly sophisticated high-throughput structural biology approaches.

Acknowledgments

I would like to thank Lindy Keller for help in preparation of this manuscript, Rongguang Zhang, Youngchang Kim, Monika, and Boguslaw Nocek in preparation of figures. This work was supported by National Institutes of Health Grant GM62414 and by the US Department of Energy, Office of Biological and Environmental Research, under contract W-31-109-Eng-38.

Author's note

I dedicate this article to my mentors Paul. B. Sigler and M. Wiewiorowski

References

Abrahams JP, Leslie AGW, Lutter R and Walker JE (1994) Structure at 2.8-Angstrom Resolution of F1-Atpase from Bovine Heart-Mitochondria. *Nature*, **370**(6491), 621–628.

Ackerman MJ, Yoo T and Jenkins D (2001) From data to knowledge–the Visible Human Project continues. *Medinfo*, **10**(Pt 2), 887–890.

Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK and Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 11), 1948–1954.

Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, *et al*. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**(5), 391–412.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, Database issue, D226–D229.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, *et al*. (2004) The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science*, **306**(5693), 79–86.

Ban N, Nissen P, Hansen J, Moore PB and Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science*, **289**(5481), 905–920.

Binkowski TA, Adamian L and Liang J (2003) Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns. *Journal of Molecular Biology*, **332**(2), 505–526.

Bray JE, Marsden RL, Rison SC, Savchenko A, Edwards AM, Thornton JM and Orengo CA (2004) A practical and robust sequence search strategy for structural genomics target selection. *Bioinformatics*, **20**(14), 2288–2295.

Bray JE, Todd AE, Pearl FM, Thornton JM and Orengo CA (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Engineering*, **13**(3), 153–165.

Bricogne G, Vonrhein C, Flensburg C, Schiltz M and Paciorek W (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallographica. Section D, Biological Crystallography*, **59**(Pt 11), 2023–2030.

Brooksbank C, Cameron G and Thornton J (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Research*, **33**, Database issue, D46–D53.

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, *et al*. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica. Section D, Biological Crystallography*, **54**(Pt 5), 905–921.

Brunzelle JS, Shafaee P, Yang X, Weigand S, Ren Z and Anderson WF (2003) Automated crystallographic system for high-throughput protein structure determination. *Acta Crystallographica. Section D, Biological Crystallography*, **59**(Pt 7), 1138–1144.

Burley SK (2000) An overview of structural genomics. *Nature Structural Biology*, **7**, 932–934.

Canaves JM, Page R, Wilson IA and Stevens RC (2004) Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. *Journal of Molecular Biology*, **344**(4), 977–991.

Carfi A, Gong H, Lou H, Willis SH, Cohen GH, Eisenberg RJ and Wiley DC (2002) Crystallization and preliminary diffraction studies of the ectodomain of the envelope glycoprotein D from herpes simplex virus 1 alone and in complex with the ectodomain of the human receptor HveA. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 5), 836–838.

Cassman M and Norvell JC (1999) Support for structural genomics and synchrotrons. *Science*, **286**(5438), 239–240.

Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, *et al*. (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Science*, **11**(4), 723–738.

Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T and Marinkovic N (2004) High-throughput computational and experimental techniques in structural genomics. *Genome Research*, **14**(10B), 2145–2154.

Chandonia JM and Brenner SE (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins*, **58**(1), 166–179.

Chayen NE and Saridakis E (2002) Protein crystallization for genomics: towards high-throughput optimization techniques. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 6, Pt 2), 921–927.

Chen L, Oughtred R, Berman HM and Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**(16), 2860–2862.

Craven CJ (2005) NMR. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, *First Edition*. Wiley.

De Angelis AA, Nevzorov AA, Park SH, Howell SC, Mrse AA and Opella SJ (2004) High-resolution NMR spectroscopy of membrane proteins in aligned bicelles. *Journal of the American Chemical Society*, **126**(47), 15340–15341.

DiDonato M, Deacon AM, Klock HE, McMullan D and Lesley SA (2004) A scaleable and integrated crystallization pipeline applied to mining the Thermotoga maritima proteome. *Journal of Structural and Functional Genomics*, **5**(1–2), 133–146.

Dieckman L, Gu M, Stols L, Donnelly MI and Collart FR (2002) High throughput methods for gene cloning and expression. *Protein Expression and Purification*, **25**(1), 1–7.

Dobson CM (2003) Protein folding and misfolding. *Nature*, **426**(6968), 884–890.

Dormitzer PR, Nason EB, Prasad BVV and Harrison SC (2004) Structural rearrangements in the membrane penetration protein of a non-enveloped virus. *Nature*, **430**(7003), 1053–1058.

Eckhardt F, Beck S, Gut IG and Berlin K (2004) Future potential of the Human Epigenome Project. *Expert Review of Molecular Diagnostics*, **4**(5), 609–618.

Erlandsen H, Abola EE and Stevens RC (2000) Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites. *Current Opinion in Structural Biology*, **10**(6), 719–730.

Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, *et al*. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Research*, **31**(13), 3375–3380.

Firbank S (2005) X-Ray Crystallography. *Encyclopedia of Genetics, Genomics Proteomics and Bioinformatics*, John Wiley & Sons.

Folkers GE, van Buuren BN and Kaptein R (2004) Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. *Journal of Structural and Functional Genomics*, **5**(1–2), 119–131.

Frishman D (2005) Target Selection for Structural Genomics. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, Ed. Wiley.

Gaasterland T (1998) Structural genomics taking shape. *Trends in Genetics*, **14**(4), 135.

Galperin MY and Koonin EV (2004) Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Research*, **32**(18), 5452–5463.

Garman E (2003) 'Cool' crystals: macromolecular cryocrystallography and radiation damage. *Current Opinion in Structural Biology*, **13**(5), 545–551.

Goldsmith-Fischman S and Honig B (2003) Structural genomics: computational methods for structure analysis. *Protein Science*, **12**(9), 1813–1821.

Grant A, Lee D and Orengo C (2004) Progress towards mapping the universe of protein folds. *Genome Biology*, **5**(5), 107.

Halic M, Becker T, Pool MR, Spahn CMT, Grassucci RA, Frank J and Beckmann R (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, **427**(6977), 808–814.

Hallin PF and Ussery DW (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics*. **20**(18), 3682–3686.

Harrison SC (2004) Whither structural biology?. *Nature Structural & Molecular Biology*, **11**(1), 12–15.

Heinemann U (2001) Secure web book to store structural genomics research data. *Ernst Schering Research Foundation Workshop* (34), 101–121.

Hendrickson WA (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**(5028), 51–58.

Hendrickson WA (2000) Synchrotron crystallography. *Trends in Biochemical Sciences*, **25**(12), 637–643.

Hendrickson WA, Horton JR and LeMaster DM (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *The EMBO Journal*, **9**(5), 1665–1672.

Holm L and Sander C (1996) Mapping the protein universe. *Science*, **273**(5275), 595–603.

Holm L and Sander C (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Research*, **26**(1), 316–319.

Holton J and Alber T (2004) Automated protein crystal structure determination using ELVES. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(6), 1537–1542.

Hosfield D, Palan J, Hilgers M, Scheibe D, McRee DE and Stevens RC (2003) A fully integrated protein crystallization platform for small-molecule drug discovery. *Journal of Structural Biology*, **142**(1), 207–217.

Hou J, Sims GE, Zhang C and Kim SH (2003) A global representation of the protein fold space. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(5), 2386–2390.

Jiang J and Sweet RM (2004) Protein Data Bank depositions from synchrotron sources. *Journal of Synchrotron Radiation*, **11**(Pt 4), 319–327.

Joachimiak MP and Cohen FE (2002) Taking MAD to the extreme: ultrafast protein structure determination. *Genome Biology*, **3**(12), RESEARCH0077.

Jordan, IK, Kondrashov, FA, Rogozin, IB, Tatusov, RL, Wolf, YI, and Koonin, EV (2001) Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biology* **2**(12), RESEARCH0053.

Jordan IK, Wolf YI and Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology*, **4**(22), 1–11.

Kalverda AP (2005) The Importance of Protein Structural Dynamics and the Contribution of NMR Spectroscopy. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, *First Edition*, Wiley.

Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A and Yokoyama S (2004) Preparation of Escherichia coli cell extract for highly productive cell-free protein expression. *Journal of Structural and Functional Genomics*, **5**(1-2), 63–68.

Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S and Kim R (2003) Structure-based functional inference in structural genomics. *Journal of Structural and Functional Genomics*, **4**(2–3), 129–135.

Kim Y, Dementieva I, Zhou M, Wu R, Lezondra L, Quartey P, Joachimiak G, Korolev O, Li H and Joachimiak A (2004) Automation of protein purification for structural genomics. *Journal of Structural and Functional Genomics*, **5**(1–2), 111–118.

Kimber MS, Vallee F, Houston S, Necakov A, Skarina T, Evdokimova E, Beasley S, Christendat D, Savchenko A, Arrowsmith CH, *et al*. (2003) Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens. *Proteins*, **51**(4), 562–568.

Koonin EV (2000) How many genes can make a cell: the minimal-gene-set concept. *Annual Review of Genomics and Human Genetics*, **1**, 99–116.

Koonin EV, Wolf YI and Karev GP (2002) The structure of the protein universe and genome evolution. *Nature*, **420**(6912), 218–223.

Kyogoku Y, Fujiyoshi Y, Shimada I, Nakamura H, Tsukihara T, Akutsu H, Odahara T, Okada T and Nomura N (2003) Structural genomics of membrane proteins. *Accounts of Chemical Research*, **36**(3), 199–206.

Lamzin VS and Perrakis A (2000) Current state of automated crystallographic data analysis. *Nature Structural Biology*, **7**, 978–981.

Laskowski RA (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, **29**(1), 221–222.

Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T, *et al*. (2002) Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(18), 11664–11669.

Levitt DG (2001) A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps. *Acta Crystallographica. Section D, Biological Crystallography*, **57**(Pt 7), 1013–1019.

Liu J, Hegyi H, Acton TB, Montelione GT and Rost B (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins*, **56**(2), 188–200.

Liu J and Rost B (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**(7), 922–933.

Manjasetty BA, Hoppner K, Mueller U and Heinemann U (2003) Secure web book to store structural genomics research data. *Journal of Structural and Functional Genomics*, **4**(2–3), 121–127.

McPherson A (2004) Protein crystallization in the structural genomics era. *Journal of Structural and Functional Genomics*, **5**(1-2), 3–12.

Moffat K and Ren Z (1997) Synchrotron radiation applications to macromolecular crystallography. *Current Opinion in Structural Biology*, **7**(5), 689–696.

Montelione GT, Zheng D, Huang YJ, Gunsalus KC and Szyperski T (2000) Protein NMR spectroscopy in structural genomics. *Nature Structural Biology*, **7**, 982–985.

Morris RJ, Perrakis A and Lamzin VS (2003) ARP/wARP and automatic interpretation of protein electron density maps. *Methods in Enzymology*, **374**, 229–244.

Moy S, Dieckman L, Schiffer M, Maltsev N, Yu GX and Collart AF (2004) Genome-scale expression of proteins from Bacillus subtilis. *Journal of Structural and Functional Genomics*, **5**(1–2), 103–109.

Murakami KS, Masuda S, Campbell EA, Muzzin O and Darst SA (2002) Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex. *Science*, **296**(5571), 1285–1290.

Mushegian AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(19), 10268–10273.

Norvell JC and Machalek AZ (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Structural Biology*, **7**, 931.

Nureki O, Watanabe K, Fukai S, Ishii R, Endo Y, Hori H and Yokoyama S (2004) Deep knot structure for construction of active site and cofactor binding site of tRNA modification enzyme. *Structure (Camb)*, **12**(4), 593–602.

O'Toole N, Grabowski M, Otwinowski Z, Minor W and Cygler M (2004) The structural genomics experimental pipeline: insights from global target lists. *Proteins*, **56**(2), 201–210.

O'Toole N, Raymond S and Cygler M (2003) Coverage of protein sequence space by current structural genomics targets. *Journal of Structural and Functional Genomics*, **4**(2–3), 47–55.

Otwinowski Z and Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. In *Methods in Enzymology*, Carter CWJ and Sweet RM (Eds.), Academic Press: New York.

Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, *et al*. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Research*, **33**, Database issue, D247–D251.

Pflugrath JW (2004) Macromolecular cryocrystallography–methods for cooling and mounting protein crystals at cryogenic temperatures. *Methods*, **34**(3), 415–423.

Phillips WC, Stewart A, Stanton M, Naday I and Ingersoll C (2002) High-sensitivity CCD-based X-ray detector. *Journal of Synchrotron Radiation*, **9**(Pt 1), 36–43.

Potterton E, McNicholas S, Krissinel E, Cowtan K and Noble M (2002) The CCP4 molecular-graphics project. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 11), 1955–1957.

Qian J, Luscombe NM and Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology*, **313**(4), 673–681.

Redfern O, Bennett C and Orengo C (2005) Protein structure classification. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, *First Edition*, John Wiley & Sons.

Saibil HR (2000) Conformational changes studied by cryo-electron microscopy. *Nature Structural Biology*, **7**(9), 711–714.

Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie GA, Thornton JM, Arrowsmith CH, Savchenko A, *et al*. (2003) Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from Escherichia coli. *The Journal of Biological Chemistry*, **278**(28), 26039–26045.

Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, *et al*. (2003) Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. *Proteins*, **50**(3), 392–399.

Sawasaki T, Ogasawara T, Morishita R and Endo Y (2002) A cell-free protein synthesis system for high-throughput proteomics. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(23), 14652–14657.

Schlunzen F, Zarivach R, Harms J, Bashan A, Tocilj A, Albrecht R, Yonath A and Franceschi F (2001) Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature*, **413**(6858), 814–821.

Shapiro L and Lima CD (1998) The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure*, **6**(3), 265–267.

Siew N and Fischer D (2004) Structural biology sheds light on the puzzle of genomic ORFans. *Journal of Molecular Biology*, **342**(2), 369–373.

Simpson AA, Tao YZ, Leiman PG, Badasso MO, He YN, Jardine PJ, Olson NH, Morais MC, Grimes S, Anderson DL, *et al*. (2000) Structure of the bacteriophage phi 29 DNA packaging motor. *Nature*, **408**(6813), 745–750.

Stols L, Gu M, Dieckman L, Raffen R, Collart FR and Donnelly MI (2002) A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Expression and Purification*, **25**(1), 8–15.

Stols L, Millard CS, Dementieva I and Donnelly MI (2004) Production of selenomethionine-labeled proteins in two-liter plastic bottles for structure determination. *Journal of Structural and Functional Genomics*, **5**(1–2), 95–102.

Terwilliger TC (1994) Structures and technology for biologists. *Acta Crystallographica. Section D, Biological Crystallography*, **50**(Pt 1), 17–23.

Terwilliger TC (2000) Structural genomics in North America. *Nature Structural Biology*, **7**, 935–939.

Terwilliger TC (2002) Automated structure solution, density modification and model building. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 11), 1937–1940.

Terwilliger T (2004a) SOLVE and RESOLVE: automated structure solution, density modification and model building. *Journal of Synchrotron Radiation*, **11**(Pt 1), 49–52.

Terwilliger TC (2004b) Structures and technology for biologists. *Nature Structural and Molecular Biology*, **11**(4), 296–297.

Terwilliger TC and Berendzen J (1999) Automated MAD and MIR structure solution. *Acta Crystallographica. Section D, Biological Crystallography*, **55**(Pt 4), 849–861.

Todd AE, Orengo CA and Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, **307**(4), 1113–1143.

Vitkup D, Melamud E, Moult J and Sander C (2001) Completeness in structural genomics. *Nature Structural Biology*, **8**(6), 559–566.

Walsh MA, Dementieva I, Evans G, Sanishvili R and Joachimiak A (1999a) Taking MAD to the extreme: ultrafast protein structure determination. *Acta Crystallographica. Section D, Biological Crystallography* **55**(Pt 6), 1168–1173.

Walsh MA, Evans G, Sanishvili R, Dementieva I and Joachimiak A (1999b) MAD data collection – current trends. *Acta Crystallographica. Section D, Biological Crystallography*, **55**(Pt 10), 1726–1732.

Watson JD, Todd AE, Bray J, Laskowski RA, Edwards A, Joachimiak A, Orengo CA and Thornton JM (2003) Target selection and determination of function in structural genomics. *IUBMB Life*, **55**(4–5), 249–255.

Westbrook J, Feng Z, Chen L, Yang H and Berman HM (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Research*, **31**(1), 489–491.

Westbrook EM and Naday I (1997) Charge-coupled device-based area detectors. *Methods in Enzymology*, **276**, 244–268.

Williams WA, Zhang RG, Zhou M, Joachimiak G, Gornicki P, Missiakas D and Joachimiak A (2004) The membrane-associated lipoprotein-9 GmpC from Staphylococcus aureus binds the dipeptide GlyMet via side chain interactions. *Biochemistry*, **43**(51), 16193–16202.

Wolf YI, Grishin NV and Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, **299**(4), 897–905.

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L and Lichtarge O (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology*, **326**(1), 255–261.

Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. *Current Opinion in Chemical Biology*, **7**(1), 39–43.

Zarembinski TI, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith CH, Edwards AM and Joachimiak A (2003) Deep trefoil knot implicated in RNA binding found in an archaebacterial protein. *Proteins*, **50**(2), 177–183.

Zhang R, Andersson CE, Savchenko A, Skarina T, Evdokimova E, Beasley S, Arrowsmith CH, Edwards AM, Joachimiak A and Mowbray SL (2003a) Structure of Escherichia coli ribose-5-phosphate isomerase: a ubiquitous enzyme of the pentose phosphate pathway and the Calvin cycle. *Structure (Camb)*, **11**(1), 31–42.

Zhang RG, Andersson CE, Skarina T, Evdokimova E, Edwards AM, Joachimiak A, Savchenko A and Mowbray SL (2003b) The 2.2 A resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction. *Journal of Molecular Biology*, **332**(5), 1083–1094.

Zhang RG, Dementieva I, Duke N, Collart F, Quaite-Randall E, Alkire R, Dieckman L, Maltsev N, Korolev O and Joachimiak A (2002) Crystal structure of Bacillus subtilis ioli shows endonuclase IV fold with altered Zn binding. *Proteins*, **48**(2), 423–426.

Zhang Y and Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(20), 7594–7599.

Zhang R, Wu R, Joachimiak G, Mazmanian SK, Missiakas DM, Gornicki P, Schneewind O and Joachimiak A (2004) Structures of sortase B from Staphylococcus aureus and Bacillus anthracis reveal catalytic amino acid triad in the active site. *Structure (Camb)*, **12**(7), 1147–1156.

Zweiger G and Scott RW (1997) From expressed sequence tags to ‘epigenomics’: an understanding of disease processes. *Current Opinion in Biotechnology*, **8**(6), 684–687.

Specialist Review

Target selection for structural genomics

Dmitrij Frishman
Technical University of Munich, Wissenschaftszentrum Weihenstephan, Freising, Germany

1. Introduction: why target selection?

The notion of "target selection for structural genomics" emerged in 1998, when structural genomics itself was officially born (Shapiro and Lima, 1998). As the name implies, it involves generating a putative list of proteins, out of a much larger candidate pool that represent high-priority targets for structure determination. In principle, the need for rational target selection is caused by two competing factors: the quickly growing capacity to solve a large number of structures and, at the same time, impossibility of exhaustive structure determination for all sequences existing in nature. The three basic questions to be answered are what proteins to analyze, how many, and why.

Technical and financial realizability of high-throughput structure determination marked the transition of structural biology from conventional, hypothesis-driven experimental work to a discovery-oriented, observational approach. While in-depth laboratory-scale research on a few selected systems of interest rarely needs external help in deciding which structure to do next, large structural genomics consortia operating in a factorylike fashion critically rely on bioinformatics-driven strategies in this regard.

On the other hand, with the currently available experimental capacities and even with the boldest prognosis as to how these capacities will grow in the future, there is little doubt that it will never be possible to obtain three-dimensional structures for all known protein sequences, the count of which stands at 1 400 000 at the time of writing and doubles roughly every 18 months. For this reason, even the biggest structural genomics consortia potentially capable of solving hundreds of structures within several years of grant funding are still confronted with the question: which particular proteins are the most promising and rewarding candidates for answering a particular biological question.

The mission of bioinformatics-driven target selection is to support the major long-term objectives of structural genomics, which are (1) discovering all types of protein folds existing in nature, (2) providing some sort of structural coverage for all known proteins, (3) predicting function for genomic proteins, and (4) enhancing structure determination technology. Each of these goals is so ambitious

that no single structural genomics consortium, or even the entire structural genomics community, can address them exhaustively. Hence, there is no generally applicable "correct" target-selection strategy: each structural genomics consortium develops its own approach, which suits best its specific scientific goals and experimental capacities. Nevertheless, there are several recurrent themes in target selection that are discussed below.

2. Target list

The centerpiece of the target-selection process is the target list that contains the prioritized list of proteins selected for analysis by a given consortium, and also indicates the experimental stages of work a particular protein has undergone so far, from being selected for analysis via cloning, expression, solubilization, and purification to crystallization, X-ray or NMR-spectroscopic analysis, and structure deposition. For example, the target list of the Joint Center for Structural genomics (**http://www.jcsg.org**; Figure 1) enumerates over 6000 proteins from several model organisms selected for structure determination.

Every target list is necessarily a dynamic data structure, which has to be regularly updated to reflect progress in experimental work and changes in molecular biological databases. Quite often, certain targets get abandoned either because

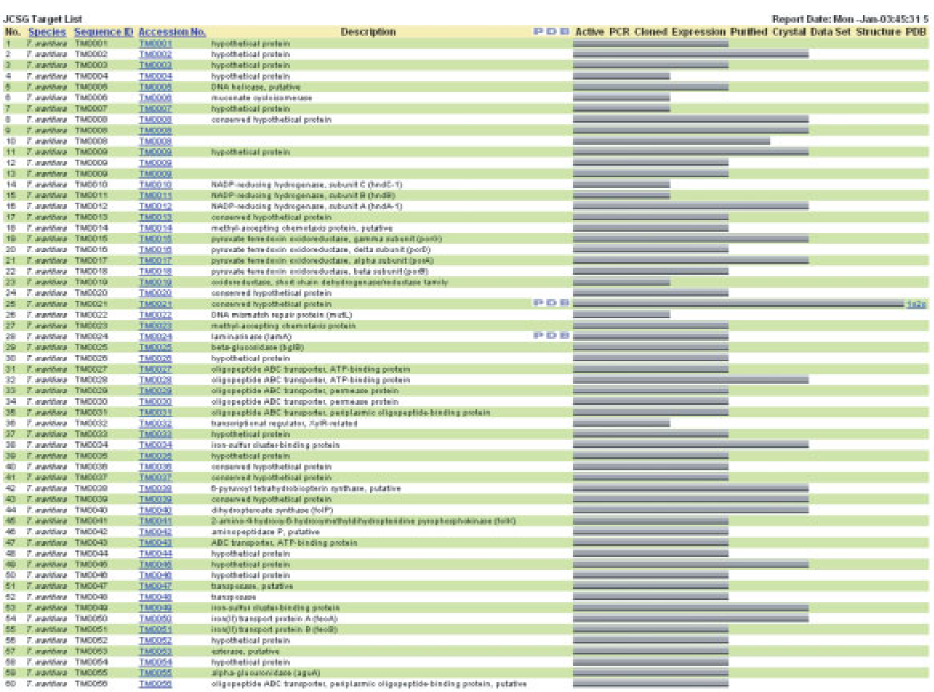


Figure 1 Target database of the Joint Center for Structural genomics (**http://www.jcsg.org**), showing experimental status of each protein selected for structure determination

they prove to be unsolvable or because new information appearing in the public databases invalidates initial rationale for their selection or simply because other research groups also working on this target are more successful. At the same time, new targets may be added to the list on the basis of the latest available evidence.

3. Basic criteria for target selection

Comprehensive functional and structural characterization of gene products by various bioinformatics tools is an important prerequisite for rational target selection. The following criteria are most frequently used as a basis for target selection decisions:

- Is this a soluble or a membrane-bound protein?
  Predicting the location of transmembrane regions and general topology of integral α-helical membrane proteins is possible with reasonable accuracy (Moller *et al*., 2001) and should be supplemented by identification of signal peptides for better results. Prediction of transmembrane β-barrel structures remains difficult because of scarce structural data, but progress is being made (Zhai and Saier, 2002).
- What is known about its function?
  The most commonly used method to predict protein function is based on similarity searches against the entire protein sequence database. Significant similarity to a protein of known function, which may be defined as 40% sequence identity over at least 60% of the protein length, usually serves as a basis for a reasonably confident function prediction, while weak similarity matches allow for only very coarse functional assignments. The presence of conserved sequence motifs as defined by PFAM, SMART, or PROSITE databases (reviewed by Liu and Rost, 2003) usually provides a very strong indication of a protein's membership in a specific functional family. Many additional functional attributes, such as enzyme classes or subcellular localization (e.g., Eisenhaber and Bork, 1998), can be automatically derived using similarity matches to thoroughly annotated proteins of known function. More sophisticated bioinformatics methods based on genomic context explore conservation of gene order, patterns of gene occurrence in different genomes (phylogenetic profiling), or similarity of expression profiles to predict protein function, including its potential interaction partners, in a similarity-free fashion (Galperin and Koonin, 2000).
- What is the taxonomic distribution of this protein?
  Is it found only in one domain of life or in all three? Specifically, are there orthologs from thermophilic species that are usually easier to crystallize? Does it have mammalian orthologs, especially those that are disease relevant? Two particularly interesting classes of targets are gene products encoded in only one organism (so-called genomic orphans (Fischer, 1999)), as well as evolutionary conserved proteins of unknown function. An additional issue is the size of the paralogous family in the genome of interest.
- What is known about its structure?
  Three-dimensional (3D) structures have been determined for less than 0.3% of proteins in completely sequenced genomes (Liu and Rost, 2002). However,

a sizable number of gene products can be linked to proteins whose crystal structure has been determined using similarity searches against the sequences of proteins with known structures. Even greater sensitivity in fold assignment can be achieved with modern threading techniques, such as GENTHREADER (McGuffin and Jones, 2003). These methods attempt to identify sequence-structure fitness using knowledge-based potentials or other statistical parameters derived from the protein structure database. Sequence comparisons with individual structural domains described in the SCOP (Andreeva *et al*., 2004) and CATH (Orengo *et al*., 2001) databases allow predicting putative folding class of the protein and also shed light on its domain structure. Where no similarity to a known structure can be detected, many informative structural features, such as the location of secondary structure elements or the presence of the coiled coil motif, can be delineated directly from protein sequence.

- Does it have structural features that might complicate structure determination?
  Apart from hydrophobic transmembrane segments, further complicating factors are low complexity and disordered regions (Linding *et al*., 2003). Proteins of large size (over 300 amino acids long) are generally not suitable for NMR-spectroscopic analysis.
- Is this protein already being targeted by other groups, and if yes – what is its experimental status?
  Most of the publicly funded structural genomics consortia joined the policy of open dissemination of their target lists and made them available for download in a standard XML-based format. In addition, the Protein Data Bank, which holds all experimentally determined atomic coordinate sets for proteins, provides a central target database for 17 structural genomics projects in the United States, Germany, and Japan (**http://targetdb.pdb.org**). A detailed analysis of the targets currently pursued by these 17 initiatives can be found in O'Toole *et al*. (2003). Similarity comparison of candidate proteins with the latest release of the target database should be routinely performed.

The quality of target-selection decisions critically depends on the general quality of genome annotation. Several software systems have been developed in recent years to automate routine analysis of protein sequences encoded in genomes (e.g., Frishman *et al*., 2001). However, since unsupervised application of automatic bioinformatics methods unavoidably results in a large number of misassignments and wrong predictions, many structural genomics consortia subject automatically produced annotation to careful manual curation.

4. General strategies of target selection

4.1. Organism-based approach

Historically, structural genomics on a single microbial organism was the first to be proposed and implemented, to some extent because of its direct analogy to genome sequencing projects. This proposition is attractive both from the structural

and functional points of view. Exhaustive X-ray or NMR-spectroscopic analysis of all (water-soluble) proteins from a single free-living organism has the potential to provide us with an unbiased view of the protein structure variety sufficient to support cellular life (Frishman and Mewes, 1997). At the same time, structure determination could help to decipher the function of many genomic proteins that are otherwise functionally uncharacterized (Kim, 2000). Efforts to obtain complete structural complements of several prokaryotic species (*Mycobacterium tuberculosis*, *Pseudomonas aerophilum*, *Haemophilus influenzae*, *Methanococcus jannaschii*, *Methanobacterium thermoautotropicum*, *Thermus thermophilis*, and others) are now underway (reviewed in Frishman, 2003). A number of eukaryotic organisms are also being addressed, including *Saccharamyces cerevisiae, C. elegans, D. melanogaster, A. thaliana,* and *H. sapiens.*

What is the current status of functional and structural characterization of a completely sequenced genome and how far are we from completing the structural portrait of a freely living organism? The function of a sizable fraction of gene products remains unknown. For the best studied prokaryotic genome, *Escherichia coli*, the percentage of proteins that still do not have even a coarse functional assignment is impressively low (19.5% according to Serres *et al*., 2004), but for most other genomes it is much higher: 48% for *M. tuberculosis* (Camus *et al*., 2002), 30% for *S. cerevisiae* (**http://mips.gsf.de**), and roughly 40% for *Arabidopsis thaliana* (Klaus Mayer, personal communication). The number of transmembrane proteins is remarkably constant in genomes of varying complexity, between 20 and 30% of the proteome (Frishman and Mewes, 1997). Typically, up to a half of all proteins can be attributed to a known structural domain, as represented in the SCOP database (Frishman *et al*., 2001). Importantly, low complexity and disordered regions are most prominent in higher eukaryotic genomes, where at least short predicted regions of biased composition are detected in nearly every gene product, while in prokaryotic genomes they affect 2–3% of proteins.

Computationally derived statistics of the type discussed above serve as a starting point for devising target-selection strategies for the single-organism approach (Frishman, 2002; Liu and Rost, 2002; Grigoriev and Choi, 2002). The target list must reveal the minimal set of sufficiently long sequence segments that possess yet unknown 3D structures. Given the high abundance of duplication modules, both on the level of whole genes or parts of genes, intrinsic to all complete genomes a crucial step in creating the list of putative structural targets involves grouping together proteins sharing similar sequence segments. This is typically achieved through single-linkage clustering of amino acid sequences based on pairwise similarity comparisons. The problem is that joint mosaic occurrence of multiple conserved protein modules leads to the well-known phenomenon of domain chaining, such that totally unrelated protein sequences may end up in the same cluster. Partitioning such single-linkage clusters into single domain clusters may represent a significant algorithmic challenge (*see* e.g., Enright and Ouzounis, 2000).

The complexity of single-linkage clusters can be reduced by considering predicted structural attributes of protein domains and filtering out those sequence segments that do not pass the minimal criteria for valid targets (Frishman, 2002). For example, out of five sequence-similar proteins from the *M. tuberculosis* genome shown in Figure 2, three will be immediately rejected either because they are

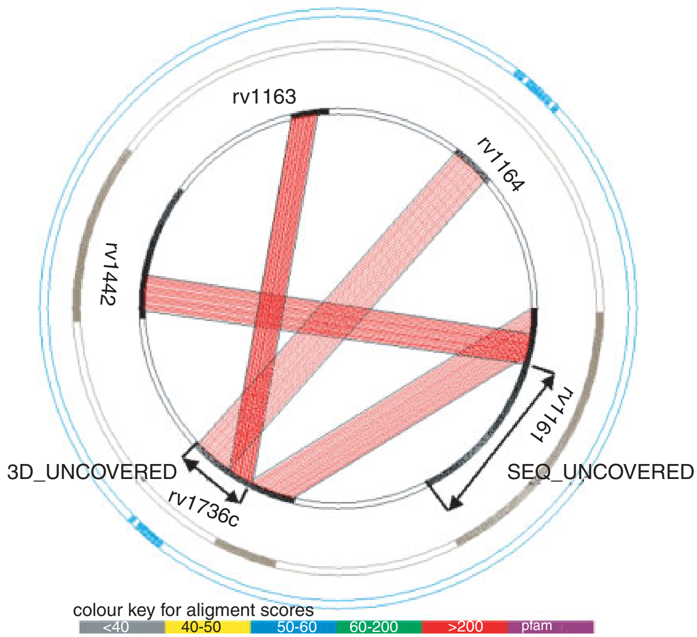


Figure 2 Circular representation of a single-linkage cluster. On the inner circle, five *M. tuberculosis* gene products (rv1164, rv1161, rv1736, rv1442, and rv1163) are shown as black sectors. The N- to C-terminal direction is clockwise. Sequence regions aligned by BLAST after an all-against-all comparison of *M. tuberculosis* proteins are joined by stripes colored according to the BLAST alignment score. On the next (middle) circle shown in brown color similarity hits in the database of protein sequences with known three-dimensional coordinates are mapped. The outer circle, shown in blue, indicates the location of predicted transmembrane regions. Further, parameters used for cluster analysis are the fraction of each protein sequence uncovered by structural information (3D_UNCOVERED) and possessing no similarity to other sequences in the cluster (SEQ_UNCOVERED) (Reproduced from Frishman D (2004) Knowledge-based selection of targets for structural genomics. *Protein Engineering*, **15**, 169–183 by permission of Oxford University Press)

predicted to be membrane-bound or because they show a significant sequence similarity to proteins of a known structure. The remaining allowable targets for structure determination are either the protein *rv1163* or the middle domain of *rv1736c*. Proceeding in this fashion, it is possible to automatically derive a putative target list for any completely sequenced genome that has been appropriately annotated. For example, single-linkage clustering of 4187 proteins encoded in the *M. tuberculosis* proteins results in 380 single-linkage clusters containing 2220 sequences and 1967 singlets – sequences not having any paralogs in the genome (Figure 3). At subsequent steps of the procedure, global redundancy between clustered sequences as well as partial domain overlaps are being resolved using an iterative procedure that involves reclustering of proteins each time a certain member of a single-linkage cluster is excluded.

Such knowledge-based approach based on excision of structurally characterized and transmembrane regions as well as gradual simplification of single-linkage clusters leads to 1544 gene products, or roughly 30% of the proteome, qualifying

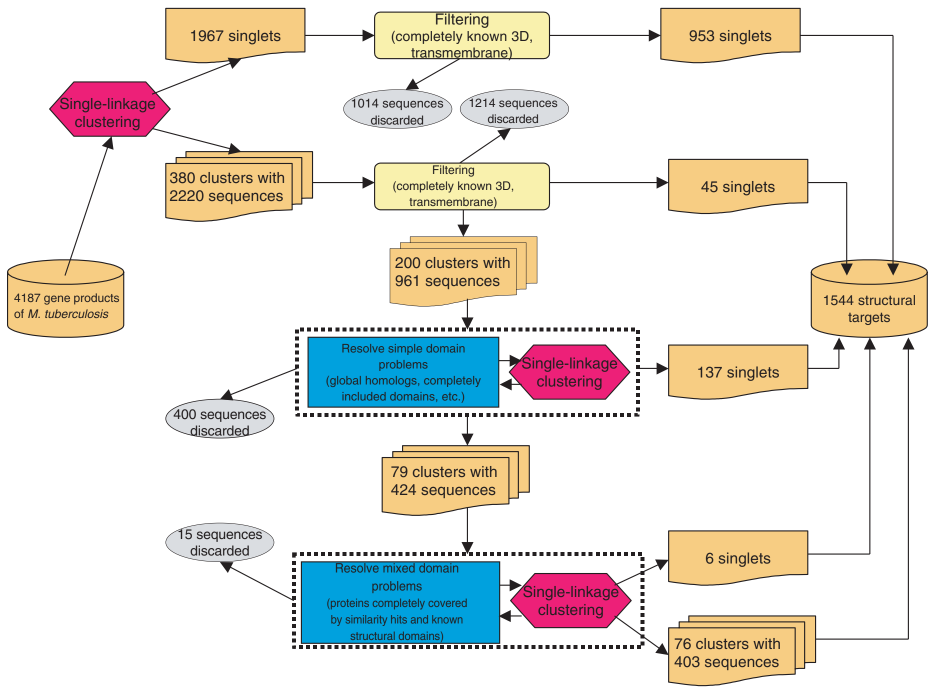


Figure 3 Flowchart of the target-selection algorithm, exemplified using the *E. coli* genome. The entire protein complement of *E. coli*, comprising 4277 gene products, is subjected to single-linkage clustering and gets split into 2235 clustered sequences and 2042 singlets (sequences not having any paralogs in the genome). Both subsets are filtered to exclude transmembrane proteins and those proteins that are completely structurally characterized. The resulting singlets are declared structural targets. The remaining 701 sequences are subjected to two stages of iterative reclustering. At the first stage, simple domain problems, such as cannibalization of a short domain by a longer one are resolved. At the second stage, more complex situations involving proteins with domain similarities to protein of known structure are treated. At each iteration, redundant sequences are excluded from further consideration and clusters reduced to just one sequence become singlets and are added to the target pool. Sequences remaining in the single-linkage clusters after the application of the algorithm are also declared structural targets since they are guaranteed to possess at least one unique and sufficiently long domain not covered by structural information

as targets for structure determination. In general, the number of targets in a completely sequenced prokaryotic genome varies between 30 and 60% and is mainly determined by two factors: size and number of paralogous families, on one hand, and the number of gene products that can be assigned to known structures by similarity, on the other hand. Higher degree of duplication (more and larger paralogous families) and higher percentage of proteins for which structural information is already available both lead to reduction of the target number.

Liu and Rost (2002) arrived at similar conclusions. They reported that approximately 48% of all genomic proteins contain structurally uncharacterized regions that do not possess experimentally intractable structural features (transmembrane and low complexity sequence spans, coiled coil), and are thus of interest for structural genomics. For about 54% of individual amino acid residues in complete genomes, there is no structural information.

In addition to computationally derived features of proteins, many other project-specific functional criteria are applied. For example, the *M. tuberculosis* structural genomics project (Goulding *et al*., 2003) is aiming at identifying novel folds, but also considers as particularly attractive targets essential genes, proteins specific to this organism, proteins related to known antituberculosis drug targets, as well as virulence proteins. Additionally, it is required that structural targets have no mammalian homologs, contain 1% of methionines, and are smaller than 50 kDa.

4.2. The quest for novel folds

Perhaps, the most attractive group of targets is constituted by proteins possessing currently unknown folds. Technically, a novel fold may be defined as one that does not show any significant similarity to any previously structurally characterized domain based on automatic (Holm and Sander, 1998), semiautomatic (Orengo *et al*., 2001), or purely manual (Andreeva *et al*., 2004) structure comparison. While every new structure is undoubtedly useful, those shedding light on not-yet-seen topologies are clearly the most exciting ones. Unfortunately, while the number of atomic coordinate sets deposited with the PDB each year is dramatically increasing, the percentage of novel folds among newly determined structures in recent years is already distressingly low (1–5%, Koppensteiner *et al*., 2000; Orengo *et al*., 2001) and continues to diminish.

If a protein under study displays a significant sequence similarity to a known structure, it can be confidently classified as a known fold. The opposite is not necessarily true: the absence of sequence similarity to a PDB protein does not prove that the fold is novel. 3D structure is evolutionarily more conserved than sequence, and hence proteins belonging to the same fold class may have completely dissimilar primary structures.

There are essentially only two ways to further advance the prediction of novel folds from sequence. The first possibility is to enhance the identification of sequences that *do not* possess a novel fold, thus narrowing down the pool of potential candidates for yet unknown topologies. This may be achieved by developing more sensitive threading techniques (McGuffin and Jones, 2003) and providing careful statistical estimates for novel fold confidence (Mallick *et al*., 2000). The second possibility is to use unconventional techniques not directly based on sequence comparison between a query sequence and a library of folds. For example, the approach developed by Portugaly *et al*. (2002) takes advantage of the exhaustive classification of all protein sequences available in the SWISS-PROT database represented as a directed graph. Clusters of related sequences on this graph roughly correspond to protein families. An assumption is made that the likelihood of producing a new fold is determined by the distance on the graph between the sequence cluster it belongs to and the nearest cluster occupied by a solved structure.

4.3. Covering sequence space with protein structures

The most general target-selection strategy provides for determining enough structures to characterize the entire population of currently known proteins across

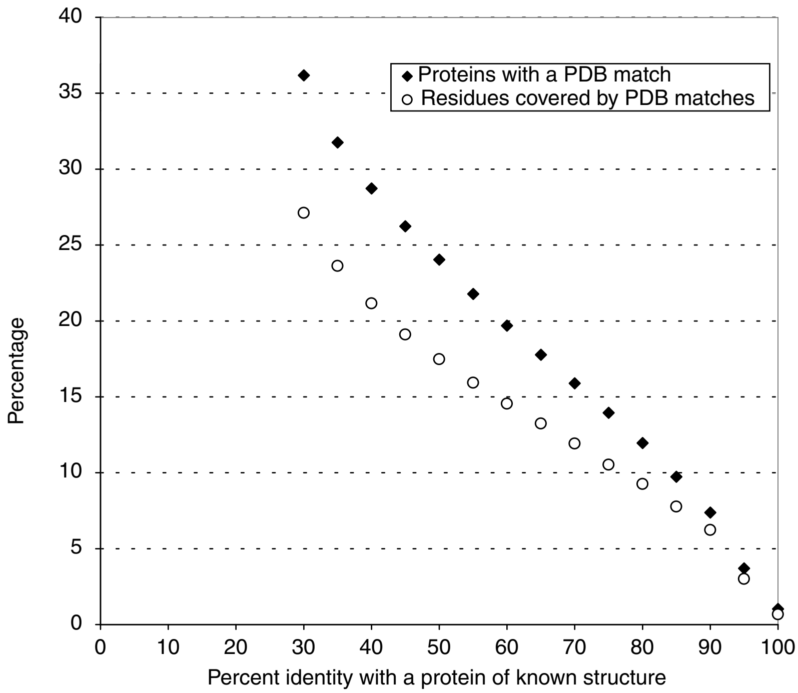


Figure 4 Percentage of proteins and individual amino acid residues in the entire sequence database covered by structural information as a function of minimal allowed sequence identity with proteins of known structure

all organisms. This idea may be viewed as a generalization of the organism-based and fold-centric approaches. The key premise of this strategy is the ability to model structures of a large number of sequences if they display sufficiently high similarity to proteins of known structure (Sanchez *et al*., 2000). The target-template sequence identity, often called *modeling distance*, determines the quality of resulting models. For coarse models, 30% of sequence identity may be sufficient, while for accurate comparative modeling at least 60% are required. As seen in Figure 4, roughly 20% of all proteins known today are amenable to homology modeling at a modeling distance of 60%.

The desired model quality is thus the key factor determining the number of structure determinations needed to cover all protein sequences with structural information. Because of the mosaic multidomain composition of proteins, especially eukaryotic ones, the most appropriate approach is to provide such coverage on the domain basis, rather than for full-length amino acid chains. Computational tests show, for example, that at least 13 000 new structures are necessary to bring each nontransmembrane sequence domain defined in the PFAM database within a modeling distance of 30% from a known PDB structure (Vitkup *et al*., 2001). However, since not all domains are encompassed by the PFAM database, the actual number of new structures required to cover all sequences is much higher. Using recently developed advanced alignment techniques, it is possible to generate

accurate alignments and hence homology models of satisfactory quality even at very low (below 20%) sequence identity level (Heger and Holm, 2003), further reducing the number of structures that need to be experimentally solved to achieve full structural coverage of sequence space.

4.4. Increasing the success rate of structural genomics

In addition to various functional and structural attributes discussed above, an important issue in rational target selection is the feasibility of structure determination for a particular protein of interest. In spite of the recent progress in structure determination technology, the success rate of structural genomics remains quite low. Christendat *et al*. (2000) reported that approximately 20% of targets with molecular weight under 20 kDa survive all stages of sample preparation and produce well-diffracting crystals, while for larger proteins this figure does not exceed 5%.

Many structural genomics projects have a clear technological focus, aiming primarily at building a high-throughput structural proteomics pipeline and investigating its main technical bottlenecks. Thus, the target list of the *M. thermoautotropicum* project, comprising 424 out of 900 nonmembrane proteins without known 3D structures, was created not on the basis of functional or structural priorities, but was rather randomly chosen to represent an unbiased sample of soluble proteins from a single organism (Christendat *et al*., 2000). The *T. maritima* project (Lesley *et al*., 2002) pursues all gene products without a structural homolog, including the difficult ones (e.g., membrane proteins), in an attempt to derive a generalized methodology for expression and refolding.

Most of the structural genomics groups began systematic collection of information on successful and failed experiments at all stages of structure determination. On the basis of retrospective analysis of experimental outcomes, attempts are being made to delineate rules for selecting structural targets that are more likely to yield to crystallographic or NMR analysis. Sequence properties that especially strongly influence experimental success include the degree of evolutionary conservation, content of charged residues, presence or absence of hydrophobic patches, number of interaction partners, and length (Goh *et al*., 2004).

4.5. Target selection in real life

Real-life target selection is rarely based on pure bioinformatics. While computational analysis of genomes and proteins does provide an invaluable information basis for structural genomics, many decisions in selecting targets are being driven by pragmatic considerations – availability of experimental material, equipment, and expertise within a given consortium, amount of funding, competitive pressure – as well as scientific intuition, curiosity, and historical background of individual researchers.

Most of the structural genomics projects (**http://www.nigms.nih.gov/funding/psi/psi_research_centers.html**) pursue mixed strategies in picking structural targets and, in fact, the boundaries between these strategies are only relative. For example,

it has been estimated that the genomes of unicellular organisms encode a substantial fraction of all folds adopted in nature – from 25% for small genomes, such as *Mycoplasma genitalium*, to up to 80% in large bacterial genomes and yeast (Wolf *et al*., 2000). This means that projects aimed at systematic determination of structures from model organisms are bound to make a significant contribution towards closing white spots on the protein fold map in the fold-centric and model-centric approaches.

5. Outlook

The technology to select targets for structure determination and the structure determination technology itself evolve in a coordinated fashion. As the latter improves, structural genomics will increasingly be focusing on very difficult targets that are largely out of reach today. The most obvious example of an upcoming target class of major importance is membrane proteins, which at present constitute less than 1% of all structures in the PDB databank. At least two small-scale pilot projects already pursue this class of proteins: MepNet (**http://www.mepnet.org**) and ProAMP (**http://binfo.bio.wzw.tum.de/proj/proamp**). Another nonstandard class of targets, also strongly underrepresented in PDB, is constituted by disordered proteins. These functionally important molecules, highly unstructured as separate entities, are presumed to acquire well-defined shape upon binding with other proteins. It is obvious that these and other challenging targets necessitate substantially different target-selection logic.

In the future, structural genomics will not just be cataloging the shapes of individual biomolecules, but also providing structural insights into the interacting complexes and entire molecular machines, ultimately leading to system-level understanding of the workings in the living cell. Correspondingly, a new quality of computational support will be required, away from sequence-by-sequence analysis towards exhaustive description of cellular functional modules.

Further reading

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV and Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, **29**, 2994–3005.

Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T and Riley M (2001) A functional update of the Escherichia coli K-12 genome. *Genome Biology*, **2**, RESEARCH0035.

References

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, D226–D229.

Camus JC, Pryor MJ, Medigue C and Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*, **148**, 2967–2973.

Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, *et al*. (2000) Structural proteomics of an archaeon. *Nature Structural Biology*, **7**, 903–909.

Eisenhaber F and Bork P (1998) Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology*, **8**, 169–170.

Enright AJ and Ouzounis CA (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Fischer D (1999) Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. *Protein Engineering*, **12**, 1029–1030.

Frishman D (2003) What we have learned about Prokaryotes from structural genomics. *OMICS, A Journal of Integrative Biology*, **7**, 211–224.

Frishman D and Mewes HW (1997) Protein structural classes in five complete genomes. *Nature Structural Biology*, **4**, 626–628.

Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A and Mewes HW (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.

Frishman D (2002) Knowledge-based selection of targets for structural genomics. *Protein Engineering*, **15**, 169–183.

Goh CS, Lan N, Douglas SM, Wu B, Echols N, Smith A, Milburn D, Montelione GT, Zhao H and Gerstein M (2004). Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *Journal of Molecular Biology*, **336**, 115–130.

Galperin MY and Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology*, **18**, 609–613.

Goulding CW, Perry LJ, Anderson D, Sawaya MR, Cascio D, Apostol MI, Chan S, Parseghian A, Wang SS, Wu Y, *et al*. (2003) Structural genomics of Mycobacterium tuberculosis: a preliminary report of progress at UCLA. *Biophysical Chemistry*, **105**, 361–370.

Grigoriev IV and Choi IG (2002) Target selection for structural genomics: a single genome approach. *OMICS: A Journal of Integrative Biology*, **6**, 349–362.

Heger A and Holm L (2003) More for less in structural genomics. *Journal of Structural and Functional Genomics*, **4**, 57–66.

Holm L and Sander C (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Research*, **26**, 316–319.

Kim SH (2000) Structural genomics of microbes: an objective. *Current Opinion in Structural Biology*, **10**, 380–383.

Koppensteiner WA, Lackner P, Wiederstein M and Sippl MJ (2000) Characterization of novel proteins based on known protein structures. *Journal of Molecular Biology*, **296**, 1139–1152.

Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T, *et al*. (2002) Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11664–11669.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ and Russell RB (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Liu J and Rost B (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.

Liu J and Rost B (2003) Domains, motifs and clusters in the protein universe. *Current Opinion in Chemical Biology*, **7**, 5–11.

Mallick P, Goodwill KE, Fitz-Gibbon S, Miller JH and Eisenberg D (2000) Selecting protein targets for structural genomics of Pyrobaculum aerophilum: validating automated fold assignment methods by using binary hypothesis testing. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 2450–2455.

McGuffin LJ and Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.

Moller S, Croning MD and Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Orengo CA, Sillitoe I, Reeves G and Pearl FM (2001) Review: what can structural classifications reveal about protein evolution? *Journal of Structural Biology*, **134**, 145–165.

O'Toole N, Raymond S and Cygler M (2003) Coverage of protein sequence space by current structural genomics targets. *Journal of Structural and Functional Genomics*, **4**, 47–55.

Portugaly E, Kifer I and Linial M (2002) Selecting targets for structural determination by navigating in a graph of protein families. *Bioinformatics*, **18**, 899–907.

Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N and Sali A (2000) Protein structure modeling for structural genomics. *Nature Structural Biology*, **7**(Suppl), 986–990.

Serres MH, Goswami S and Riley M (2004) GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Research*, **32**, D300–D302.

Shapiro L and Lima CD (1998) The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure*, **6**, 265–267.

Vitkup D, Melamud E, Moult J and Sander C (2001) Completeness in structural genomics. *Nature Structural Biology*, **8**, 559–566.

Wolf YI, Grishin NV and Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, **299**, 897–905.

Zhai Y and Saier MH Jr (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Science*, **11**, 2196–2207.

Specialist Review

Protein–ligand docking and structure-based drug design

Peter R. Oledzki, Alasdair T. R. Laurie and Richard M. Jackson
University of Leeds, Leeds, UK

1. Introduction

Drug discovery is a time consuming and expensive process, and increasing its speed and efficiency through the application of computational methods has received considerable attention (Jorgensen, 2004). The first stages of the drug design process are lead discovery and lead optimization. Traditionally, many lead compounds have been discovered serendipitously, by chemically modifying and improving existing drugs (the so-called me-too approach) or by isolating the active ingredients in herbal remedies. However, increasing importance is being placed on the discovery of selective biochemical entities that act on distinct molecular targets, such as receptors, enzymes, and ion channels. In addition, many pharmaceutically relevant macromolecular 3D structures from physical (X-ray crystallography and Nuclear Magnetic Resonance) and computational (comparative/homology modeling) sources have become available. The potential of "structure-based" drug design (SBDD) methods were recognized as early as the 1980s, and interest has since grown considerably as demonstrable progress has been made. The critical factors for successful drug design are the identification and optimization of lead compounds whilst averting unwanted side effects.

Recently, experimental high-throughput screening (HTS) methods, involving screening large chemical libraries against a protein target, have become common in the pharmaceutical industry. However, large scale HTS is expensive and it is beneficial to restrict the size of a chemical library to compounds that are most likely to be successful. Virtual screening (VS) of a chemical library is one way for potentially successful compound identification if the structure of the target is known. The term virtual screening has been used to define *in silico* methods that are used to identify ligands that have activity against a specified protein target or a similar activity to an identified "lead" compound. Screening compounds for similar activity to specified ligands with only information derived from those ligands (i.e., no structural information for the macromolecular target) is termed *ligand-based design* and will not be covered in this review. Here, the term virtual screening is used in regard to applications to macromolecular–ligand docking. Both HTS and

VS are limited by the size of the chemical library they use. *De novo* drug design attempts to overcome this limitation by increasing the exploration of chemical search space; however, a limitation of this method is that ligands produced in this way are not always chemically synthesizable. Both VS and *de novo* drug design require a three-dimensional representation of the protein target and are therefore referred to as *SBDD methods*.

2. Prerequisites for structure-based drug design

There are two prerequisites for SBDD. Firstly, a three-dimensional representation of the protein target must be available. Preferably, this structure should be derived from experimental information such as X-ray crystallography or NMR. However, for the appropriate application of SBDD methods, an understanding of the effects and limitations that structural data imposes on the quality of results must be fully taken into account (Acharya and Lloyd, 2005; Davis *et al*., 2003). Alternatively, homology/comparative modeling can be used to obtain a suitable structure. The role and utility of comparative models in drug design have been reviewed at length (Hillisch *et al*., 2004; Jacobson and Sali, 2004; Wieman *et al*., 2004). A model of a protein structure may be created if there is sequence similarity between the target and another protein for which the experimental structure has already been resolved. Drug design is frequently faced with the problem of discovering a ligand for a protein that has no experimentally determined structure. The most prominent examples of this are the G-protein coupled receptors (GPCRs), with 50% of recently developed drugs being targeted against GPCRs (Klabunde and Hessler, 2002). The advent of structural genomics is ultimately expected to yield enough experimental structures to cover fold space (Brenner and Levitt, 2000). Thus, it should become increasingly possible to generate realistic models for any given protein sequence using comparative modeling techniques (Marti-Renom *et al*., 2000).

The second prerequisite for SBDD is the location of the ligand-binding site. Binding sites can be identified by cocrystallization of a protein with a ligand experimentally. Otherwise, using sequence or structural similarity of the site of interest to a known binding site is a common method for the identification of protein functional sites (Campbell *et al*., 2003). Alternatively, there are many computational methods available for the *de novo* identification and mapping of small-molecule binding sites (Sotriffer and Klebe, 2002), with Q-SiteFinder (Laurie and Jackson, 2005) being a specific example.

3. Protein–ligand docking

The docking problem can be defined as “Given the atomic coordinates of two molecules, predict their “correct” bound association” (Brooijmans and Kuntz, 2003). The docking problem relates to the “lock and key” mechanism first proposed by Fischer over 100 hundred years ago. The theory of molecular recognition was further developed by Koshland to encompass conformational changes during the binding event, which has been termed *induced fit* (Koshland, 2004). The

docking problem involves protein–protein interactions, protein–DNA interactions, and protein/DNA–ligand interactions. However, only protein–ligand docking will be discussed here due to its direct relevance to SBDD. Other applications of molecular docking are beyond the scope of this review.

Docking is traditionally viewed as a two-stage problem, with both stages interrelated. The first component of the docking problem relates to the sampling of conformational space to predict the correct orientation and position of a ligand. For rigid-body docking six degrees of freedom apply, three for rotation and three for translation. The second stage of docking relates to ranking of solutions and the determination of favorable conformations within the protein-binding site. This classification of docking solutions is achieved using scoring functions. Scoring functions are an area of great importance and currently the focus of intense research activity and will be discussed in further detail below.

The field of protein–ligand docking has been the subject of interest for many years, with one of the first docking algorithms DOCK being developed as early as 1982 (Kuntz *et al*., 1982). The first generation of automatic docking algorithms only used rigid-body calculations with no simulation of ligand or protein flexibility. However, the importance of conformational change is now well recognized, and docking methods now incorporate ligand flexibility and in some cases protein flexibility. There are numerous protein–ligand docking algorithms with several recent reviews of the methods (Kitchen *et al*., 2004; Kroemer, 2003; Taylor *et al*., 2002).

4. Docking search procedures incorporating conformational change

Many different approaches have been used to solve the problem of ligand flexibility. Docking algorithms sample ligand conformational space using genetic algorithms (Jones *et al*., 1997), Monte Carlo simulations (Goodsell *et al*., 1993; Venkatachalam *et al*., 2003), tabu list searching (Westhead *et al*., 1997), molecular dynamics (Wang *et al*., 1999), energy minimization methods (McMartin and Bohacek, 1997), stochastic tunneling (Mancera *et al*., 2004), probabilistic sampling (Jackson, 2002), and incremental construction (Rarey *et al*., 1996). The sampling method must cover sufficient conformational space to find the correct position and orientation (binding mode) of a ligand within a protein pocket. Therefore the objective of all search procedures is to sample conformational space sufficiently in a time efficient manner. Each individual method achieves this; however, there is always a trade-off between adequately searching conformational space and computational expense. Search methods implemented in docking can be separated into systematic methods, stochastic simulation methods and various hybrid approaches.

Systematic methods try to explore all the degrees of freedom permitted by a molecule, but are prone to combinatorial explosion. Using a fragment-based approach, ligands can be grown in a stepwise or incremental manner within the active site. Rigid ligand fragments can be docked in the site and then linked

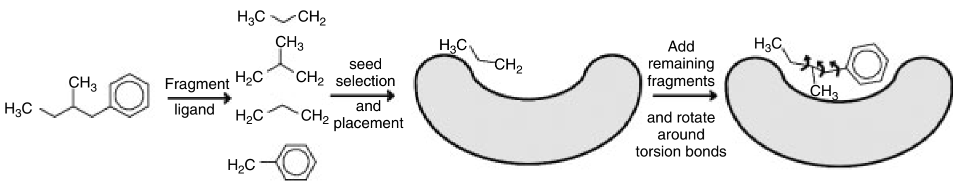


Figure 1 Incremental construction for docking consists of fragmenting a ligand, placing a seed/anchor fragment, and building up the remaining ligand fragments

covalently together; this follows the similar *de novo* ligand-design strategy of place and join (see below). Alternatively the ligand can be separated into core and flexible fragments. The core fragments can be docked into the binding site and used as starting positions for growing the remaining fragments in an energetically favorable manner. This approach has been used in FlexX (Rarey *et al*., 1996), DOCK 4.0 (Ewing *et al*., 2001), and FlexLigDock (Oledzki, Lyon and Jackson, in preparation) (see Figure 1). Other systematic methods consist of pregenerated multiple ligand conformers such as in FLOG (Miller *et al*., 1994) and FRED (1997); these are then docked individually, reducing the search problem to a rigid body docking procedure.

Stochastic search procedures change single ligands or populations of ligands in a random fashion in relation to a predefined probability function. Implementation of a Monte Carlo search (Goodsell and Olson, 1990) and genetic algorithms can be found in the popular docking programs Autodock (Goodsell *et al*., 1996) and GOLD (Jones *et al*., 1997). Other simulation methods, such as using molecular dynamics for sampling conformational space have been less popular for docking. However, many energy minimization techniques have been shown to complement other sampling methods in docking as a refinement step.

The concept of using pharmacophores within SBDD has recently grown in popularity (Dror *et al*., 2004; Guner, 2002; Mason *et al*., 2001). A pharmacophore is defined as "being a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" (Gund, 1977). This information can be derived from sets of known active ligands for a target protein. The minimal structural requirements essential for receptor recognition, receptor binding, and biological response, are represented in a pharmacophore. Typically, a pharmacophore consists of three or more feature points such as hydrogen-bonding groups and charged or hydrophobic groups, along with a set of geometrical rules that describe how these features are related in 3D space. A pharmacophore can be used to constrain docking calculations to produce ligand conformations more likely to recreate biologically relevant protein–ligand contacts representing experimentally observed binding modes. Many recent docking algorithms have included pharmacophore constraints to speed up docking calculations and to improve binding mode prediction (Goto *et al*., 2004; Hindle *et al*., 2002; Seifert, 2005; Yang and Shen, 2005).

Structural changes in the receptor upon ligand binding are a common phenomenon (Najmanovich *et al*., 2000). Therefore, the degree to which the protein moves upon ligand binding mirrors the drop in docking accuracy (Erickson *et al*.,

2004). The various effects of protein flexibility on drug discovery have been recently reviewed (Davis *et al*., 2003; Teague, 2003). The issue of addressing protein flexibility is a complex problem, and was initially solved by using soft docking (Jiang and Kim, 1991). Softness allows imprecision in both the protein and ligand structures and models their adaptive changes to each other. Partial side-chain flexibility has also been incorporated into docking procedures (Jones *et al*., 1995; Leach, 1994) among others that have been comprehensively reviewed (Carlson, 2002; Teodoro and Kavraki, 2003). Nevertheless, most of these methods do not include explicit sampling of side chains, such as those implemented in protein–protein docking (Jackson *et al*., 1998) or allow for backbone rearrangements as this would be impracticable for VS of large chemical libraries due to combinatorial explosion. Using multiple receptor conformations (MRCs) (either experimentally or computationally generated) for docking seems to be the most practical approach to date (Teodoro and Kavraki, 2003; Claussen *et al*., 2001). An important advantage of this approach is that the structural space of the binding pocket can be represented in a VS process, even in the case of loop displacements (Cavasotto and Abagyan, 2004). However, the best way to implement protein flexibility in an efficient manner for the VS process is still an unresolved problem (Carlson, 2002).

5. *De novo* drug design tools

De novo drug design involves creating new ligands in a binding site one atom or fragment at a time. This increases exploration of chemical space relative to VS tools, since in the latter there is a predefined compound library. However, a potential disadvantage is that suggested lead compounds may not necessarily be chemically synthesizable. *De novo* drug design tools fall into two main categories, or may use a combination of the techniques: (1) fragment based and (2) atom based. Examples of both types are discussed below.

Fragment-based approaches use molecular fragments to build up molecules. They have an advantage in that molecules have a greater chance of being synthesizable with a realistic structure, but the novelty of the molecules produced is limited by the size of the fragment library. There are two subsets of a fragment-based approach. In the “place and join” method, the molecular fragments are placed independently before attempting to connect them together. The alternative is an incremental construction approach, in which fragments are aligned using the orientation of the growing ligand (Figure 2).

GroupBuild (Rotstein and Murcko, 1993) and LigBuilder (Wang *et al*., 2000) are examples of incremental construction algorithms (Figure 2b). GroupBuild uses a library of 14 organic fragments (formic acid, formaldehyde, formamide, amine, benzene, cyclohexane, cyclopentane, ethane, ethylene, water, methanol, methane, sulfone, and thiophene). A core fragment is docked into the binding site (or a part of a known inhibitor may be used). Fragments are added by replacing hydrogen atoms within the ligand. Fragments are rotated around the vector projection of the replaced bond. Only chemically feasible bonds are allowed. The algorithm uses a greedy strategy, randomly retaining one of the best 25% that scoring presents after

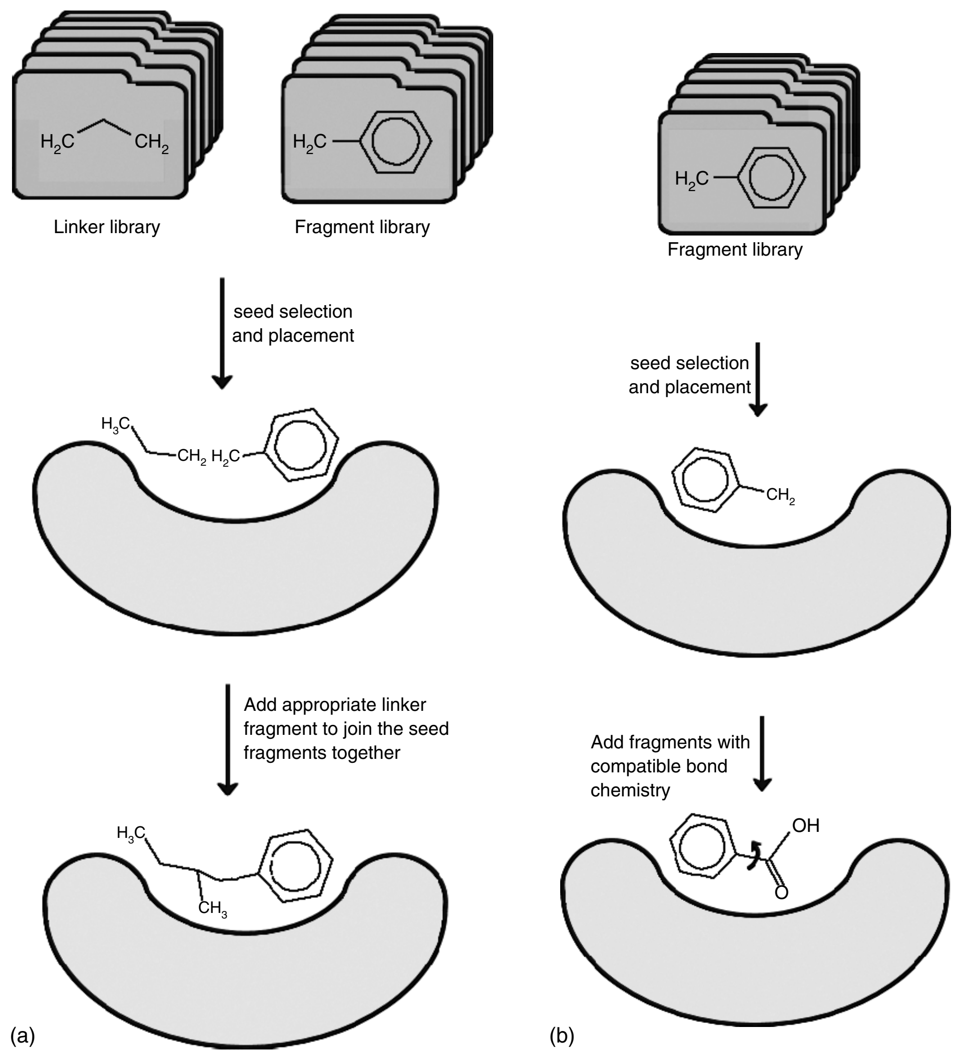


Figure 2 (a) The “place and join” method identifies favorable binding regions for seed fragments or atoms and then attempts to join them together using a suitable linker fragment. (b) The *de novo* incremental construction approach uses a library of fragments, selects an anchor/seed placement and adds fragments from the library with compatible bond chemistry

each round of fragment addition. LigBuilder uses a library of 56 organic fragments. It uses a similar method of incremental construction to GroupBuild. In addition, the program is capable of performing atomic mutations.

SPROUT (Gillet *et al*., 1993) and MCSS/HOOK (Caflisch *et al*., 1993; Eisen *et al*., 1994; Miranker and Karplus, 1991) are examples of “place and join” methods. SPROUT initially identifies target sites in the receptor. A template molecule is then fitted so that it occupies at least one target site. More templates are added such that they (1) interact with the other target sites, (2) do not cause

steric interference with the binding site, and (3) can be joined to the existing templates. Templates are joined together either by fusion or bridge (single bond) formation. SPROUT does not use linker fragments. MCSS/HOOK uses a different "place and join" approach with linker fragments (Figure 2a). MCSS (multiple copy simultaneous search) was developed to determine energetically favorable binding modes of functional groups in a ligand-binding site. Thousands of copies of the functional group of interest are placed in random positions within the binding site. The algorithm uses a Monte Carlo energy minimization. Regions of favorable interactions are indicated by a concentrated localization of functional group copies to regions within the binding site. HOOK uses a database of linker fragments to join together functional groups that are placed in the binding site by MCSS. It fuses bonds within the linker fragment (called *hooks*) with free CH_3–R bonds of the functional groups. HOOK scores the chemical complementarity and steric fit of the resulting molecules within the binding site.

De novo ligands generated by atom-by-atom methods are not limited by the size of a fragment library, but may be less likely to produce realistic or synthesizable ligands. LEGEND (Nishibata and Itai, 1993) forms molecules atom by atom. The first atom is placed at a set distance from a hydrogen bond acceptor in the binding site. New atoms are added, testing all possible dihedral angles and are assigned an atom type and hybridization state. If the atom occupies a forbidden position on the grid, it is rejected. The algorithm terminates at a user defined molecule size. SEEDS (Honma *et al*., 2001), uses LEGEND output to search for commercially available compounds with structural similarity.

Some *de novo* drug design methods do not fall into these three classes. For example, Monte Carlo *De Novo* Ligand Generator (MCDNLG) (Gehlhaar *et al*., 1995), does not use a fragment library, database searching or atom-by-atom approach. The binding cavity is filled with an array of atoms of random type and hybridization. In each Monte Carlo step, parameters such as atom types and bond angles are allowed to change. The process is guided by intramolecular and protein–ligand energy functions as well as penalties on unreasonable chemical structures (such as sp^3-hybridized atoms with double bonds). It produces novel leads, but they are not necessarily chemically accessible.

Testing *de novo* design algorithms is often difficult, as novel molecules must be chemically synthesized and screened against the target to verify that they genuinely interact. Many *de novo* design algorithms are not tested experimentally, unless they are used commercially, which slows their acceptance by the pharmaceutical industry. SPROUT has yielded a number of new leads that have been tested *in vivo* (Han *et al*., 2000; Milne *et al*., 1996), which proves the usefulness of the *de novo* approach to lead generation.

6. Scoring functions

Energy functions are used to estimate binding energies of ligand conformations. "The binding free energy of a protein–ligand complex can be intuitively understood as the difference in how much the ligand likes to be in the protein binding pocket and how much it likes to be in the solvent at infinite separation from the protein"

(Muegge and Rarey, 2001). For ease of calculation the binding free energy has been separated into various terms. Thus the free energy (equation 1) can be expressed as the sum of different parts, sometimes referred to as a *master equation* (Ajay and Murcko, 1995). However, the theoretical justification for this has been challenged, as binding energies are a global property of a system and cannot be expressed as a sum of components (Mark and van Gunsteren, 1994). The free energy can be divided into components that account for the difference in solvation energies of the ligand, protein, and complex (ΔG^{lig}_{solv}, ΔG^{prot}_{solv}, $\Delta G^{complex}_{solv}$), the interaction between the protein and the ligand in complex (ΔG_{int}), entropy changes for the ligand and protein ($T\Delta S$), and conformational changes (reorganization energy) in the ligand and protein ($\Delta\lambda$). This last term involves the loss of internal rotations, and translational/rotational degrees of freedom and change in vibrational free energy upon complex formation. The binding free energy is given by

$$\Delta G_{binding} = \Delta G^{complex}_{solv} - \Delta G^{lig}_{solv} - \Delta G^{prot}_{solv} + \Delta G_{\text{int}} - T\Delta S + \Delta\lambda \qquad (1)$$

The terms in equation (1) (Muegge and Rarey, 2001) have been expressed elsewhere in slightly different forms (e.g. Ajay and Murcko, 1995; Gohlke and Klebe, 2002). They represent binding by partitioning through the use of various empirical factor models. They are known as *partitioning models* because the components are often calculated independently.

Calculating accurate free energies of binding can be problematic, because large individual energy contributions balance to obtain a small overall binding energy, which can lead to statistical and systematic errors (Kollman, 1993; Muegge *et al*., 1995). In addition, other terms such as entropy can be difficult to determine and only approximations are available. These simplified energy functions give a fast estimation of binding affinities for determining favorable docking and *de novo* solutions, and are generally referred to as *scoring functions*. The wide variety of scoring functions implemented within docking algorithms underpins the fact that no single scoring function has been developed to correctly rank docking solutions in all cases. There has been significant focus on developing scoring functions in recent times. There have been several reviews (Gohlke and Klebe, 2001, 2002; Jansen and Martin, 2004; Tame, 1999), numerous comparative studies of scoring performance (Ferrara *et al*., 2004; Wang *et al*., 2003, 2004) and scoring function/docking algorithm combinations (Bissantz *et al*., 2000; Ha *et al*., 2000; Xing *et al*., 2004). The scoring functions of ligands ideally permit the discrimination of a solution with the correct ligand-binding mode from alternative solutions. Decoy structures are commonly used to validate the capabilities of scoring functions to detect the correct binding modes of ligands (Wang *et al*., 2003, 2004; Perola *et al*., 2004). The scoring not only needs to be able to differentiate between active and inactive compounds, but to also reduce the number of false–positives and negatives by correctly recognizing the actives. In addition to binding mode discrimination, ideally the scoring function will be capable of ranking a set of ligands according to experimental binding affinity.

A variety of scoring functions exist which can be categorized into four main groups, (1) scoring functions that are derived from first principles, (2) empirical

scoring functions, (3) knowledge-based scoring functions, and (4) other functions that use chemical scores, contact scores, or shape complementary scores.

Force field scoring functions derived from first principles have certain advantages: they are fast to compute, they are transferable and conceptually they use terms that have a physical basis. A major disadvantage of first principles force fields is that they tend to measure only potential energy with some force fields containing additional terms for solvation (ΔG_{solv}^{lig}, ΔG_{solv}^{prot}, $\Delta G_{solv}^{complex}$) and entropy ($T\Delta S$) using different theoretical models. Standard non-bonded interaction energies are calculated by most force fields with two main energy terms, the van der Waals (vdW) and the electrostatic contributions. The vdW term is generally calculated using a Lennard-Jones 6–12 potential; however, this function has been softened in some docking scoring schemes. Most often, the electrostatic term is described by Coulomb's law with scaling factors used to model the dielectic screening effects of the solvent (water). In addition, geometric-dependent, hydrogen-bonding terms may be utilized as in the case of Q-fit (Jackson, 2002) and GOLD (Jones *et al*., 1997).

Empirical scoring functions use multivariate regression methods that fit coefficients of physically motivated contributions to the binding free energy to reproduce measured binding affinities of a training set of known 3D protein–ligand complexes (Bohm, 1994; Horton and Lewis, 1992). The empirical scoring function employed by LUDI (Bohm, 1994) was further enhanced for the docking method FlexX. The free energy of binding for a protein–ligand complex is estimated in FlexX as the sum of free energy contributions from the number of rotatable bonds in the ligand, hydrogen bonds, ion–pair interactions, hydrophobic and π-stacking interactions of aromatic groups, and lipophilic interactions. The advantages of empirical scoring functions are their speed and reasonably good predictive power for binding free energies. However, conceptual disadvantages are that these functions are trained on protein–ligand complexes with medium-to-strong binding energies. Therefore due to the nature of their derivation they have no effective penalty for bad conformations and their ability to predict new structurally different complexes not in their training set, has been questioned (Ferrara *et al*., 2004).

Knowledge-based scoring is a deductive approach that uses existing protein–ligand complexes to create potentials and forces that can be used for scoring binding affinities. The idea is that forces that determine the ligands interaction with a protein are complicated; therefore, the information in protein–ligand complexes is taken as the main source of information. Deriving scoring functions for ligand–protein docking from crystal structures was inspired by the success of similar methodologies for protein folding and protein structure prediction (Sippl, 1990). Protein–ligand atom pair potentials are derived from crystallographic complexes under the assumption that the optimal placements of ligand atoms with respect to protein atoms are represented. The continual growth of structural data for protein–ligand complexes now provides several million ligand–protein atom distances. These distances can be used to create statistical preferences. One of the first investigations that incorporated knowledge-based potentials was Verkhivker et al. for the binding affinity prediction of HIV-1 protease complexes (Verkhivker *et al*., 1995). Many different knowledge-based scoring functions have

been developed in the last few years. Drugscore (Gohlke *et al*., 2000) and DFIRE (Zhang *et al*., 2005) are examples of freely available knowledge-based scoring functions for academic use.

Consensus scoring, using multiple scoring functions in conjunction with each other, have recently become a popular method of enhancing scoring function performance (Charifson *et al*., 1999; Clark *et al*., 2002; Wang and Wang, 2001; Yang *et al*., 2005). This follows from the popular concept of data fusion from Information Theory. For example, one scoring function can be used to search for the correct binding mode and another function can be used to refine the placement and correctly score the protein–ligand complex. However, increased performance with consensus scoring can only be achieved if both scoring functions differ in how they approach the scoring problem, and create a synergistic effect.

7. Virtual screening with docking

The VS of databases containing thousands or millions of compounds using docking has become a routine part of modern SBDD. VS has been shown to be a powerful method for identification of lead compounds due to its ability to rank a series of candidate ligands correctly according to binding affinity (Abagyan and Totrov, 2001; Alvarez, 2004; Shoichet *et al*., 2002). There are three main reasons why VS has become accepted and successful: (1) There is an increased amount of experimental protein structural information. (2) There is successful enrichment of compound databases with VS. (3) The speed of VS decreases the time for drug design cycles, reducing the cost of drug development.

The overall aim of a VS experiment is to reduce a database of candidate ligands to an enriched list of bioactive ligands. A variety of factors are used to evaluate and validate the quality of the methodology. For any scoring scheme, the yield is the percentage of active ligands in the top ranked compounds, and can be calculated by $(A_h/T_h) \times 100$, where A_h is the number of active compounds in a selected subset of the ranked database, and T_h, represents some predetermined number of compounds to be screened in a database of size T. The coverage of the screen, that is, the percentage of active (true positives) ligands retrieved from the database is defined as $(A_h/A) \times 100$, where A is the number of actives in the whole library database. Another factor considered is the false–positive rate given by $(T_h - A_h)/(\mathrm{T} - \mathrm{A})$. Finally the factor most widely cited by VS experiments is the enrichment factor given by $(A_h/T_h)/(A/T)$.

8. Structure-based drug design success

Over the last decade SBDD methods have improved sufficiently to be used routinely in many large-scale drug discovery projects within the pharmaceutical industry where structural information for the target is available. The recent success of SBDD in creating clinically important molecules is well documented (Hardy and Malikayil, 2003). SBDD has been utilized to create more than 50 chemical entities for clinical trials, while numerous drugs have also been approved. Specific examples

of successful drugs currently on the market are the HIV protease inhibitors Viracept (Kaldor *et al*., 1997) and Agenerase (Kim *et al*., 1995).

There are several examples where a *de novo* approach has yielded potent inhibitors. This includes, BREED (Pierce *et al*., 2004), which was used to generate a HIV protease inhibitor with nanomolar binding affinity, and TOPAS (Schneider *et al*., 2000), which was used to generate an inhibitor for a human cannabinoid receptor. A comprehensive study using a variety of *de novo* drug design tools was applied to cyclin-dependent kinase-4 (cdk4). Cdk4 plays a key role in cell division, and inhibitors could be useful anticancer agents. Honma (Honma, 2003) used LEGEND to generate around 1000 lead compounds for a homology model of cdk4. These compounds were usually not chemically synthesizable or commercially available, so the SEEDS program was used to find structurally related commercially available compounds. 382 compounds were screened against cdk4, and five compounds in the diarylurea class were found to bind with micromolar affinity. Combinatorial chemistry was used to synthesize several hundred urea compounds and one was selected (IC_{50} = 100 nM) for further optimization. X-ray crystallography revealed the binding mode of a closely related compound and a chemical library was designed to take into account the orientations of the amino acids in the binding site. LUDI and Leapfrog (2005) were used to suggest the types and sizes of the fragments for substitution. The process successfully elucidated nanomolar inhibitors.

The comparison of VS with HTS technology has provided evidence to show that these two techniques are complementary as they are capable of finding differing bioactive molecules. VS was shown to select more druglike hits than HTS for the target protein tyrosine phosphatase 1B (Doman *et al*., 2002). While another study reported the failure of HTS to identify inhibitors for DNA gyrase, an *in silico* screening and lead optimization procedure using a structure-based approach was successfully implemented to identify an inhibitor 10 times more potent than an existing DNA gyrase inhibitor novobiocin (Boehm *et al*., 2000). A large percentage of false–positives are generally found in hit lists from HTS, therefore follow-up assays are required to distinguish active from inactive substances. However, in an attempt to reduce the number of false–positives a consensus approach of using HTS in conjunction with VS was taken. The method implemented gave a sixfold enrichment over the HTS hit rate and the HTS hits ranked in the top 2% of the VS included 42% of the true hits, but only 8% of the false–positives (Jenkins *et al*., 2003).

There are a number of successful examples of structure-based inhibitor identification that cover a wide range of protein targets and identification methodologies (Alvarez, 2004). Generally inhibitors in the micromolar range are discovered; however, the potency and number of potential inhibitors reported from such screens can vary. This is dependent on methodologies, refinement of hits and the number of potential inhibitors tested experimentally. For example, a VS using pharmacophore constraints and docking on an in-house library collection for checkpoint kinase-1 identified 103 compounds for testing. From theses compounds 36 inhibitors were identified from four different classes that had a range of IC(50) value ranges from 110 nM to 68 μM (Lyne *et al*., 2004). Another study performed a VS on 5-aminoimidazole-4-carboxamide ribonucleotide (AICAR) transformylase

to reveal 44 potential inhibitors; *in vitro* inhibition assays were performed for 16 soluble compounds. Micromolar inhibition was identified for 8 compounds which contained novel scaffolds; these scaffolds were used in further docking studies to identify 11 additional compounds with one particular inhibitor having an IC(50) value of 600 nM. These 19 inhibitors served as novel scaffolds/templates for further development of potent and specific inhibitors (Li *et al*., 2004).

Homology models are traditionally less accurate than experimental models and when utilized in VS can supply inferior enrichment rates (McGovern and Shoichet, 2003). The docking tool DragHome (Schafferhans and Klebe, 2001) has been developed to try and address such issues. It combines homology model data with information about known ligands to the protein to create a pharmacophore. This idea was further developed by the same group in modeling binding sites including ligand information explicitly (MOBILE) (Evers *et al*., 2003) and has proved to be successful in VS (Evers and Klebe, 2004). There are many other examples of successful VS using homology modeling (Alvarez, 2004). Another notable and highly successful study performed by Novartis (Vangrevelinghe *et al*., 2003) used DOCK 4.0 to screen an in-house library of 400 000 molecules for inhibitors of human casein kinase II. A homology model built from the Zea mays enzyme, which has 82% sequence identity in the ATP binding site, was used. A dozen compounds were selected for testing, of which four exhibited better than 50% inhibition at a concentration of 10 mM, including an indoloquinazolinone derivative that had an IC(50) of 80 nM, making it the most potent reported inhibitor of this enzyme.

9. Summary

The number of studies involving protein–ligand docking and SBDD have grown significantly over recent years. This is probably because of the number of reported successes in the "lead" identification process using these methods. These developments have been stimulated by the wealth of experimental structural information. However, comparative modeling is increasingly playing a role. Improvements in small-molecule binding site prediction tools, advances in *de novo* drug design and docking algorithms have led to a more systematic approach to drug design.

Scoring functions play an important role in the success of SBDD, as scoring is used to identify bioactive molecules. Improvements in scoring functions and in consensus scoring schemes have increased our ability to identify bioactive molecules. All these enhancements have permitted VS to be used more routinely in the structure-based drug discovery process.

References

Abagyan R and Totrov M (2001) High-throughput docking for lead generation. *Current Opinion in Chemical Biology*, **5**(4), 375–382.

Acharya KR and Lloyd MD (2005) The advantages and limitations of protein crystal structures. *Trends in Pharmacological Sciences*, **26**(1), 10–14.

Ajay and Murcko MA (1995) Computational methods to predict binding free energy in ligand-receptor complexes. *Journal of Medicinal Chemistry*, **38**, 4953–4967.

Alvarez JC (2004) High-throughput docking as a source of novel drug leads. *Current Opinion in Chemical Biology*, **8**(4), 365–370.

Bissantz C, Folkers G and Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, **43**(25), 4759–4767.

Boehm HJ, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, Kostrewa D, Kuehne H, Luebbers T, Meunier-Keller N, *et al*. (2000) Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *Journal of Medicinal Chemistry*, **43**(14), 2664–2674.

Bohm HJ (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, **8**(3), 243–256.

Brenner SE and Levitt M (2000) Expectations from structural genomics. *Protein Science*, **9**(1), 197–200.

Brooijmans N and Kuntz ID (2003) Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, **32**, 335–373.

Caflisch A, Miranker A and Karplus M (1993) Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *Journal of Medicinal Chemistry*, **36**(15), 2142–2167.

Campbell SJ, Gold ND, Jackson RM and Westhead DR (2003) Ligand binding: functional site location, similarity and docking. *Current Opinion in Structural Biology*, **13**(3), 389–395.

Carlson HA (2002) Protein flexibility is an important component of structure-based drug discovery. *Current Pharmaceutical Design*, **8**(17), 1571–1578.

Cavasotto CN and Abagyan RA (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *Journal of Molecular Biology*, **337**(1), 209–225.

Charifson PS, Corkery JJ, Murcko MA and Walters WP (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, **42**(25), 5100–5109.

Clark RD, Strizhev A, Leonard JM, Blake JF and Matthew JB (2002) Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling*, **20**(4), 281–295.

Claussen H, Buning C, Rarey M and Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. *Journal of Molecular Biology*, **308**(2), 377–395.

Davis AM, Teague SJ and Kleywegt GJ (2003) Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie (International ed. in English)*, **42**(24), 2718–2736.

Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT and Shoichet BK (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of Medicinal Chemistry*, **45**(11), 2213–2221.

Dror O, Shulman-Peleg A, Nussinov R and Wolfson HJ (2004) Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design. *Current Medicinal Chemistry*, **11**(1), 71–90.

Eisen MB, Wiley DC, Karplus M and Hubbard RE (1994) HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins*, **19**(3), 199–221.

Erickson JA, Jalaie M, Robertson DH, Lewis RA and Vieth M (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, **47**(1), 45–55.

Evers A, Gohlke H and Klebe G (2003) Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *Journal of Molecular Biology*, **334**(2), 327–345.

Evers A and Klebe G (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of Medicinal Chemistry*, **47**(22), 5381–5392.

Ewing TJ, Makino S, Skillman AG and Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, **15**(5), 411–428.

Ferrara P, Gohlke H, Price DJ, Klebe G and Brooks CL, III (2004) Assessing scoring functions for protein-ligand interactions. *Journal of Medicinal Chemistry*, **47**(12), 3032–3047.

FRED (1997) Documentation on World Wide Web URL, **http://www.eyesopen.com/products/applications/fred.html**.

Gehlhaar DK, Moerder KE, Zichi D, Sherman CJ, Ogden RC and Freer ST (1995) De novo design of enzyme inhibitors by Monte Carlo ligand generation. *Journal of Medicinal Chemistry*, **38**(3), 466–472.

Gillet V, Johnson AP, Mata P, Sike S and Williams P (1993) SPROUT: a program for structure generation. *Journal of Computer-Aided Molecular Design*, **7**(2), 127–153.

Gohlke H, Hendlich MM and Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, **295**(2), 337–356.

Gohlke H and Klebe G (2001) Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology*, **11**(2), 231–235.

Gohlke H and Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International ed. in English)*, **41**(15), 2644–2676.

Goodsell DS, Lauble H, Stout CD and Olson AJ (1993) Automated docking in crystallography: analysis of the substrates of aconitase. *Proteins*, **17**(1), 1–10.

Goodsell DS, Morris GM and Olson AJ (1996) Automated docking of flexible ligands: applications of AutoDock. *Journal of Molecular Recognition*, **9**(1), 1–5.

Goodsell DS and Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. *Proteins*, **8**(3), 195–202.

Goto J, Kataoka R and Hirayama N (2004) Ph4Dock: pharmacophore-based protein-ligand docking. *Journal of Medicinal Chemistry*, **47**(27), 6804–6811.

Gund P (1977) Three-dimensional pharmacophoric pattern searching.. *Progress in Molecular and Subcellular Biology*, **5**, 117.

Guner OF (2002) History and evolution of the pharmacophore concept in computer-aided drug design. *Current Topics in Medicinal Chemistry*, **2**(12), 1321–1332.

Ha S, Andreani R, Robbins A and Muegge I (2000) Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *Journal of Computer-Aided Molecular Design*, **14**(5), 435–448.

Han Q, Dominguez C, Stouten PF, Park JM, Duffy DE, Galemmo RA, Jr., Rossi KA, Alexander RS, Smallwood AM, Wong PC, *et al*. (2000) Design, synthesis, and biological evaluation of potent and selective amidino bicyclic factor Xa inhibitors. *Journal of Medicinal Chemistry*, **43**(23), 4398–4415.

Hardy L and Malikayil A (2003) The impact of structure-guided drug design on clinical agents. *Current Drug Discovery*, 16–20.

Hillisch A, Pineda LF and Hilgenfeld R (2004) Utility of homology models in the drug discovery process. *Drug Discovery Today*, **9**(15), 659–669.

Hindle SA, Rarey M, Buning C and Lengaue T (2002) Flexible docking under pharmacophore type constraints. *Journal of Computer-Aided Molecular Design*, **16**(2), 129–149.

Honma T, Hayashi K, Aoyama T, Hashimoto N, Machida T, Fukasawa K, Iwama T, Ikeura C, Ikuta M, Suzuki-Takahashi I, *et al*. (2001) Structure-based generation of a new class of potent Cdk4 inhibitors: new de novo design strategy and library design. *Journal of Medicinal Chemistry*, **44**(26), 4615–4627.

Honma T (2003) Recent advances in de novo design strategy for practical lead identification. *Medicinal Research Reviews*, **23**(5), 606–632.

Horton N and Lewis M (1992) Calculation of the free energy of association for protein complexes. *Protein Science*, **1**(1), 169–181.

Jackson RM (2002) Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of Computer-Aided Molecular Design*, **16**(1), 43–57.

Jackson RM, Gabb HA and Sternberg MJ (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *Journal of Molecular Biology*, **276**(1), 265–285.

Jacobson M and Sali A (2004) Comparative protein structure modeling and its applications to drug discovery. *Annual Reports in Medicinal Chemistry*, **39**, 259–276.

Jansen JM and Martin EJ (2004) Target-biased scoring approaches and expert systems in structure-based virtual screening. *Current Opinion in Chemical Biology*, **8**(4), 359–364.

Jenkins JL, Kao RY and Shapiro R (2003) Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. *Proteins*, **50**(1), 81–93.

Jiang F and Kim SH (1991) "Soft docking": matching of molecular surface cubes. *Journal of Molecular Biology*, **219**(1), 79–102.

Jones G, Willett P, Glen RC, Leach AR and Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, **267**(3), 727–748.

Jones G, Willett P and Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, **245**(1), 43–53.

Jorgensen WL (2004) The many roles of computation in drug discovery. *Science*, **303**(5665), 1813–1818.

Kaldor SW, Kalish VJ, Davies JF II, Shetty BV, Fritz JE, Appelt K, Burgess JA, Campanale KM, Chirgadze NY, Clawson DK, *et al*. (1997) Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *Journal of Medicinal Chemistry*, **40**(24), 3979–3985.

Kim EE, Baker CT, Dwyer MD, Murcko MA, Rao BG, Tung RD and Navia MA (1995) Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally available inhibitor of the enzyme. *Journal of the American Chemical Society*, **117**, 1181–1182.

Kitchen DB, Decornez H, Furr JR and Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews. Drug Discovery*, **3**(11), 935–949.

Klabunde T and Hessler G (2002) Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem*, **3**(10), 928–944.

Kollman P (1993) Free energy calculations - applications to chemical and biological phenomena. *Chemical Reviews*, **7**, 2395.

Koshland DE Jr (2004) Crazy, but correct. *Nature*, **432**(7016), 447.

Kroemer RT (2003) Molecular modelling probes: docking and scoring. *Biochemical Society Transactions*, **31**(Pt 5), 980–984.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R and Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, **161**(2), 269–288.

Laurie AT and Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**(9), 1908–1916.

Leach AR (1994) Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology*, **235**(1), 345–356.

LeapFrog Manual. (2005) *SYBYL Version 6.8*, Tripos: St. Louis, MO.

Li C, Xu L, Wolan DW, Wilson IA and Olson AJ (2004) Virtual screening of human 5-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of autodock to identify novel nonfolate inhibitors. *Journal of Medicinal Chemistry*, **47**(27), 6681–6690.

Lyne PD, Kenny PW, Cosgrove DA, Deng C, Zabludoff S, Wendoloski JJ and Ashwell S (2004) Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *Journal of Medicinal Chemistry*, **47**(8), 1962–1968.

Mancera RL, Kallblad P and Todorov NP (2004) Ligand-protein docking using a quantum stochastic tunneling optimization method. *Journal of Computational Chemistry*, **25**(6), 858–864.

Mark AE and van Gunsteren WF (1994) Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *Journal of Molecular Biology*, **240**(2), 167–176.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F and Sali A (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, **29**, 291–325.

Mason JS, Good AC and Martin EJ (2001) 3-D pharmacophores in drug discovery. *Current Pharmaceutical Design*, **7**(7), 567–597.

McGovern SL and Shoichet BK (2003) Information decay in molecular docking screens against holo, apo and modelled conformations of enzymes. *Journal of Medicinal Chemistry*, **3**, 2895–2907.

McMartin C and Bohacek RS (1997) QXP: powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design*, **11**(4), 333–344.

Miller MD, Kearsley SK, Underwood DJ and Sheridan RP (1994) FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, **8**(2), 153–174.

Milne GW, Wang S and Nicklaus MC (1996) Molecular modeling in the discovery of drug leads. *Journal of Chemical Information and Computer Sciences*, **36**(4), 726–730.

Miranker A and Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins*, **11**(1), 29–34.

Muegge I, Ermler U, Fritzsch G and Knapp E (1995) Free energy cofactors at the quinone Q(a) site of the photosynthetic reaction center of rhodobacter sphaeroids calculated by minimizing the statistical error. *The Journal of Physical Chemistry*, **99**, 17917.

Muegge I and Rarey M (2001) Small molecule docking and scoring. *Reviews in Computational Chemistry*. **17**, 1–60.

Najmanovich R, Kuttner J, Sobolev V and Edelman M (2000) Side-chain flexibility in proteins upon ligand binding. *Proteins*, **39**(3), 261–268.

Nishibata Y and Itai A (1993) Confirmation of usefulness of a structure construction program based on three-dimensional receptor structure for rational lead generation. *Journal of Medicinal Chemistry*, **36**(20), 2921–2928.

Perola E, Walters WP and Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, **56**(2), 235–249.

Pierce AC, Rao G and Bemis GW (2004) BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *Journal of Medicinal Chemistry*, **47**(11), 2768–2775.

Rarey M, Wefing S and Lengauer T (1996) Placement of medium-sized molecular fragments into active sites of proteins. *Journal of Computer-Aided Molecular Design*, **10**(1), 41–54.

Rotstein SH and Murcko MA (1993) GroupBuild: a fragment-based method for de novo drug design. *Journal of Medicinal Chemistry*, **36**(12), 1700–1710.

Schafferhans A and Klebe G (2001) Docking ligands onto binding site representations derived from proteins built by homology modelling. *Journal of Molecular Biology*, **307**(1), 407–427.

Schneider G, Lee ML, Stahl M and Schneider P (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *Journal of Computer-Aided Molecular Design*, **14**(5), 487–494.

Seifert MH (2005) ProPose: steered virtual screening by simultaneous protein-ligand docking and ligand-ligand alignment. *Journal of Chemical Information and Modeling*, **45**(2), 449–460.

Shoichet BK, McGovern SL, Wei B and Irwin JJ (2002) Lead discovery using molecular docking. *Current Opinion in Chemical Biology*, **6**(4), 439–446.

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, **213**(4), 859–883.

Sotriffer C and Klebe G (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco*, **57**(3), 243–251.

Tame JR (1999) Scoring functions: a view from the bench. *Journal of Computer-Aided Molecular Design*, **13**(2), 99–108.

Taylor RD, Jewsbury PJ and Essex JW (2002) A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, **16**(3), 151–166.

Teague SJ (2003) Implications of protein flexibility for drug discovery. *Nature Reviews. Drug Discovery*, **2**(7), 527–541.

Teodoro ML and Kavraki LE (2003) Conformational flexibility models for the receptor in structure based drug design. *Current Pharmaceutical Design*, **9**(20), 1635–1648.

Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D and Furet P (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *Journal of Medicinal Chemistry*, **46**(13), 2656–2662.

Venkatachalam CM, Jiang X, Oldfield T and Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling*, **21**(4), 289–307.

Verkhivker G, Appelt K, Freer ST and Villafranca JE (1995) Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering*, **8**(7), 677–691.

Wang J, Dixon R and Kollman PA (1999) Ranking ligand binding affinities with avidin: a molecular dynamics-based interaction energy study. *Proteins*, **34**(1), 69–81.

Wang R, Gao Y and Lai L (2000) LigBuilder: a Multi-purpose program for structure based drug design. *Journal of Molecular Modeling*, **6**, 498–516.

Wang R, Lu Y and Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry*, **46**(12), 2287–2303.

Wang R and Wang S (2001) How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences*, **41**(5), 1422–1426.

Wang R, Lu Y, Fang X and Wang S (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *Journal of Chemical Information and Computer Sciences*, **44**(6), 2114–2125.

Westhead DR, Clark DE and Murray CW (1997) A comparison of heuristic search algorithms for molecular docking. *Journal of Computer-Aided Molecular Design*, **11**(3), 209–228.

Wieman H, Tondel K, Anderssen E and Drablos F (2004) Homology-based modelling of targets for rational drug design. *Mini Reviews in Medicinal Chemistry*, **4**(7), 793–804.

Xing L, Hodgkin E, Liu Q and Sedlock D (2004) Evaluation and application of multiple scoring functions for a virtual screening experiment. *Journal of Computer-Aided Molecular Design*, **18**(5), 333–344.

Yang JM and Shen TW (2005) A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **59**(2), 205–220.

Yang JM, Chen YF, Shen TW, Kristal BS and Hsu DF (2005) Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling*, **45**(4), 1134–1146.

Zhang C, Liu S, Zhu Q and Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *Journal of Medicinal Chemistry*, **48**(7), 2325–2335.

Short Specialist Review

Large complexes by X-ray methods

Poul Nissen
University of Aarhus, Aarhus, Denmark

1. Introduction

In the recent years, atomic structures of large complexes such as the 20S proteasome (Lowe *et al*., 1995; Jap *et al*., 1993), the groEL chaperone (Braig *et al*., 1994), the ribosome (Ban *et al*., 2000; Wimberly *et al*., 2000; Schluenzen *et al*., 2000; Yusupov *et al*., 2001), RNA polymerase (Cramer *et al*., 2000), and the bluetongue virus core (Grimes *et al*., 1998) have been determined by the use of X-ray crystallography. Combined with biochemical and genetic data, the characterization of these macromolecular assemblies can provide a high level of functional understanding (*see* Article 94, **What use is a protein structure?**, Volume 6). This has been very clearly demonstrated for the ribosome whose complex function in protein synthesis is now addressed at the very atomic level – a goal that would have seemed impossible to reach only few years ago. Furthermore, other methods of obtaining structural information without crystals, in particular, single-particle cryo-electron microscopy with 3D reconstruction (*see* Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6 and Article 106, **Electron microscopy as a tool for 3D structure determination in molecular structural biology**, Volume 6), can further extend the use of crystal structures in the interpretation of other functional states by docking models. Genome sequencing, microarray data, and proteome analysis by mass spectrometry (*see* Article 44, **Protein interaction databases**, Volume 5) as well as many other sources of information indicate that there is a high level of complexity in the cell depending on multicomponent assemblies and networks of interaction between macromolecules. Clearly, tomorrow's biological research will increasingly be directed towards an in-depth understanding of cell biology at the molecular level. Interesting questions are therefore: (1) how do we obtain crystals of very large assemblies so that their atomic structures can be determined, (2) how do we solve these structures, and (3) how can the methods be improved in the future?

2. Crystallization of large structures

Crystallization of large structures and complexes is no different from any ordinary protein *per se*, provided that it is available in a purified form, in sufficient amounts,

and as a stable and homogenous sample (*see* Article 104, **X-ray crystallography**, Volume 6). These criteria may, however, be difficult to meet in practice. As an example, only ribosomes from a few species and in a few functional states have so far provided well-diffracting crystals, even though ribosome samples are easily prepared with large yields from most organisms. Well-diffracting crystals of ribosome preparations represent particularly stable forms, such as the *Haloarcula marismortui* and *Deinococcus radiodurans* 50S subunit (von Bohlen *et al*., 1991; Ban *et al*., 2000; Harms *et al*., 2001). In the case of the *Thermus thermophilus* 30S subunit, a fortuitous crystal packing arrangement seems to have played a critical role in getting high resolution (Wimberly *et al*., 2000; Clemons *et al*., 2001).

The reproducibility and diffraction properties of the crystals *H. marismortui* 50S subunit were dramatically improved by a back-extraction procedure, where partitioning was obtained by precipitation in polyethylene glycol (PEG) and next solubilization at saturating conditions in the final crystallization buffer (Ban *et al*., 2000). This procedure has likely improved the homogeneity of the sample and thereby also the crystal order. The *T. thermophilus* 30S subunit was purified by an elegant procedure based on hydrophobic interaction chromatography, which enriched a subfraction of subunits without the S1 protein bound. Thereby, the reproducibility of the crystals was greatly improved. A different approach of 30S crystal improvement was based on the use of a large tungsten cluster (Schluenzen *et al*., 2000), which may have imposed a similar effect by displacing contaminating S1 due to overlapping binding sites. Crystallization of the *T. thermophilus* 70S ribosome was facilitated by a complex salt wash procedure (Cate *et al*., 1999), which has probably removed loosely bound contaminants. The *Escherichia coli* 70S ribosome, on the other hand, was stabilized by the use of polyamines (which are also natural constituents of the cytoplasm) as a critical component of the crystallization buffer (Vila-Sanjurjo *et al*., 2003). These crystals have later been optimized and now diffract to very high resolution (J. Cate, personal communication).

However, well-diffracting crystals of ribosomes in complex with, for example, the GTP/GDP binding translation factors remain to be seen. Cryo-EM studies of such complexes now routinely employ the use of single-particle classification as a prerequisite, which is to indicate that they are indeed difficult to prepare as homogenous samples (Valle *et al*., 2002; Klaholz *et al*., 2004).

Once crystals have formed, their diffraction properties may greatly improve by controlled dehydration, such as in the case of the yeast RNA polymerase II complex (Cramer *et al*., 2001; Gnatt *et al*., 1997). In general, the dehydration procedures shrink unit cell dimensions and thereby enforce intermolecular contacts. In the case of very large complexes, this may alter crystal contacts covering rather large areas and thus have a very profound effect on the crystal order. This method should always be thoroughly investigated.

The use of a double-mutant form was observed to yield the best crystals of the *E. coli* groEL complex (Braig *et al*., 1994). This strategy has also been successful for very difficult cases, such as the *E. coli* lactose permease (Abramson *et al*., 2003), which resisted crystallization for many years. The mutational strategy is particularly well suited for the crystallization of multicomponent assemblies obtained by a recombinant approach where mutant forms can readily be constructed and analyzed. In this case, the possibility of using multiple species (normally considered a good

strategy) would also be less attractive due to the laborious effort of establishing the expression system alone.

3. Data collection

Very large structures automatically lead to very large unit cell dimensions, and thus very close separation of reflections on the X-ray detector. Detectors have a finite size and, therefore, the collection of high-resolution data is challenged by the necessity to also increase the crystal-to-detector distance to separate the reflections well and to reduce the background noise. Furthermore, the reflections from crystals with large unit cells are relatively weak since fewer unit cells repeat through the crystal. The advent of third-generation synchrotrons and the recent developments of large CCD detectors with fast read-out and small point-spread on the detector surface have certainly facilitated the structure determination of large structures. However, only few beam line stations are yet able to meet all the requirements for high-resolution data collection from crystals with large unit cells. New developments in synchrotrons and beam lines will provide X-ray beams with extremely low divergence and high energy resolution, and this can be used to reduce the reflection spot size sufficiently so as to allow for more diffraction orders to be resolved per given unit of detector surface area. The development of new and larger detectors with fast read-out will be just as important.

The weak diffraction from crystals with large unit cells calls for long exposure times during collection of high-resolution data. Therefore, radiation damage will often become a serious issue, which may require the use of a few to even hundreds of crystals for a data set, even when crystals are properly flash-frozen at 100 K. An implementation of crystal mounting robotics, as scheduled for most synchrotron beam lines in the near future, will facilitate the handling and bookkeeping of many crystals and allow the user to pick only the best crystals for an optimal data collection strategy.

A particular problem arises from virus crystals that are notoriously difficult to flash-freeze for cryoprotection during data collection. Therefore, only a few or even just a single frame can be obtained from each crystal, and as many as a thousand crystals may be necessary to mount, expose, and process for a workable data set at reasonable resolution to be obtained (Grimes *et al*., 1998). This is tedious and compromises data quality. It is likely that crystals of other large assemblies in the future will suffer from similar problems, and it will be a high priority to devise protocols for the cryoprotection of such crystals.

4. Structure determination

Macromolecular crystallography provides high-throughput approaches for structure determination when phasing is based on, for example, seleno-methionine protein (*see* Article 97, **Structural genomics – expanding protein structural universe**, Volume 6 and Article 104, **X-ray crystallography**, Volume 6). However, solving very large structures means that exceptionally potent sources of isomorphous

and anomalous differences must be used to provide a signal that allows for site identification, heavy-atom refinement, and phasing. Partial seleno-methionine incorporation was obtained for RNA polymerase II, but it was insufficient for phasing (Bushnell *et al*., 2001). Likewise, it has been rationalized that seleno-methionine incorporation would be insufficient for phasing of the ribosomal subunits (Brodersen *et al*., 2003). Therefore, either heavy-atom clusters or many individual sites of strong anomalous scatterers (L-edge) are required to obtain a significant signal for site identification and phasing (reviewed by Abrahams and Ban, 2003). In particular, clusters containing either mercury (TAMM, tetrakis-acetoxymercuri methan), tantalum ($Ta_6Br_{12}{}^{2+}$), or tungsten (phosphotungstic acid derivatives) have been useful for initial phasing of very large structures, as successfully implemented for crystals of the nucleosome core (O'Halloran *et al*., 1987), the rubisco protein (Schneider, 1994), the α_2-macroglobulin (Andersen *et al*., 1995), the 20S proteasome (Knablein *et al*., 1997), the ribosome and ribosomal subunits (Ban *et al*., 1998; Clemons *et al*., 1999; Schluenzen *et al*., 2000; Cate *et al*., 1999), the bacterial RNA polymerase (Zhang *et al*., 1999), and yeast RNA polymerase II (Fu *et al*., 1999). The individual atoms of a cluster scatter in phase at low resolution (below ~6 Å) where the clusters behave as "superatoms" with hundreds to several thousand electrons. The cluster sites can therefore be located by difference Patterson maps calculated at very low resolution despite a background of tens to hundreds of thousand non-hydrogen atoms. The identification of cluster sites can also be facilitated and checked by difference Fourier maps at very low resolution if a molecular replacement solution can be obtained using a molecular envelope derived from electron microscopy as a search model. This was successfully applied for the 50S ribosomal subunit (Ban *et al*., 1998). Unfortunately, heavy-atom cluster derivatives most often turn out to be useless for phasing at higher resolution due to cluster disorder that causes the isomorphous and anomalous differences to vanish below the noise level above medium resolution (4–6 Å) where the clusters resolve into individual, yet delocalized, heavy atoms.

When low to medium resolution phases have been derived from clusters, the identification of single heavy-atom sites can conveniently be approached by difference Fourier maps (O'Halloran *et al*., 1987; Schneider, 1994), which then allow for phasing at higher resolution (<4 Å). For ribosome crystals, osmium and iridium amines were particularly helpful for phasing as single-atom derivatives (Ban *et al*., 1999; Cate *et al*., 1999; Ban *et al*., 2000; Wimberly *et al*., 2000). Osmium hexa- and pentamine as well as iridium hexamine interact gently and yet extensively with complex RNA structures, similar to the binding of hydrated Mg^{2+} ions (Cate *et al*., 1996). Further, these elements display abnormally high anomalous signal ("white lines") at the L-absorption edges and at appropriate wavelengths for data collection (1.14 and 1.21 Å for Os and Ir, respectively, at the stronger LIII edge). In hindsight, it was demonstrated for the 30S ribosomal subunit that it could be solved directly from an osmium hexamine derivative with automated site determination based on the anomalous differences (Brodersen *et al*., 2003). For the large protein structures, mercurials have been particularly useful, together with platinates – similar to ordinary proteins. Interestingly, phasing by multiwavelength anomalous dispersion (MAD; *see* Article 104, **X-ray crystallography**, Volume 6) has not

been successful for large complexes (for an exception, see Cate *et al*., 1999), probably due to radiation damage building up through data collection. Instead, single or multiple isomorphous replacements with anomalous scattering (SIRAS/MIRAS) and single-wavelength anomalous difference phasing (SAD) have been used (*see* Article 104, **X-ray crystallography**, Volume 6).

5. Density modification

With experimental phases available at high resolution, the structure determination re-enters the routines of general crystallography where density modification extends and refines the phases and produces the final, experimental map for model building. Even at medium resolution (∼5 Å), density modification proved valuable when cautiously implemented for ribosome data (Ban *et al*., 1999; Clemons *et al*., 1999; Cate *et al*., 1999). In fact, it can be shown that density modification is particularly powerful for large structures (Abrahams and Ban, 2003). In the case of the 50S ribosomal subunit, an excellent experimental map at 2.4 Å resolution was based on a solvent flipping procedure initiated using rather poor heavy-atom-derived phases extending only barely beyond 4 Å resolution (Ban *et al*., 2000).

Symmetric assemblies, such as viruses, the 20 S proteasome, and groEL, display high levels of noncrystallographic symmetry (NCS), which is helpful in the site identification of heavy-atom derivatives. More importantly, NCS allows for powerful averaging to bootstrap and refine phases. The structure determination of groEL represents a good example (Braig *et al*., 1994). A single mercury derivative was identified by the use of NCS parameterization of the heavy-atom search. The phases were refined by density modification with sevenfold NCS averaging. An initial model was constructed and then used to derive a mask of the groEL monomer. Phases were then randomized and refined by the sevenfold NCS averaging only and using the final high-resolution data set. A high level of NCS has the additional advantage that it relaxes the need for complete and accurate data collection. The structure determination of the ∼40 mio Da bluetongue virus core based on incomplete data extending to 3.5 Å resolution collected from more than a thousand individual crystals mounted at room temperature represents a stunning example (Grimes *et al*., 1998).

Related articles

See Article 35, **Structural biology of protein complexes**, Volume 5; Article 95, **History and future of X-ray structure determination**, Volume 6; Article 97, **Structural genomics – expanding protein structural universe**, Volume 6; Article 100, **Large complexes and molecular machines by electron microscopy**, Volume 6; Article 104, **X-ray crystallography**, Volume 6

Acknowledgments

Fruitful discussions with Nenad Ban, Peter B. Moore, and Thomas A. Steitz on the application and implementation of techniques involved in the structure determination of the large ribosomal subunit are greatly acknowledged. Ditlev E. Brodersen is thanked for critical reading of the manuscripts.

References

Abrahams JP and Ban N (2003) X-ray crystallographic structure determination of large asymmetric macromolecular assemblies. *Methods in Enzymology*, **374**, 163–188.

Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR and Iwata S (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**, 610–615.

Andersen GR, Koch TJ, Dolmer K, Sottrup-Jensen L and Nyborg J (1995) Low resolution X-ray structure of human methylamine-treated alpha 2-macroglobulin. *The Journal of Biological Chemistry*, **270**, 25133–25141.

Ban N, Freeborn B, Nissen P, Penczek P, Grassucci RA, Sweet R, Frank J, Moore PB and Steitz TA (1998) A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. *Cell*, **93**, 1105–1115.

Ban N, Nissen P, Hansen J, Capel M, Moore PB and Steitz TA (1999) Placement of protein and RNA structure into a 5 Å-resolution map of the 50S robosomal subunit. *Nature*, **400**, 841–847.

Ban N, Nissen P, Hansen J, Moore PB and Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.

Braig K, Otwinowski Z, Hegde R, Boisvert DC, Joachimiak A, Horwich AL and Sigler PB (1994) The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature*, **371**, 578–586.

Brodersen DE, Clemons WM Jr, Carter AP, Wimberly BT and Ramakrishnan V (2003) Phasing the 30S ribosomal subunit structure. *Acta Crystallographica. Section D, Biological Crystallography*, **59**, 2044–2050.

Bushnell DA, Cramer P and Kornberg RD (2001) Selenomethionine incorporation in *Saccharomyces cerevisiae* RNA polymerase II. *Structure (Cambridge)*, **9**, R11–R14.

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR and Doudna JA (1996) Crystal structure of a group I ribozyme domain: principle of RNA packing. *Science*, **273**, 1678–1685.

Cate JH, Yusupov MM, Yusupova GZ, Earnest TN and Noller HF (1999) X-ray crystal structures of 70S ribosome functional complexes. *Science*, **285**, 2095–2104.

Clemons WM Jr, Brodersen DE, McCutcheon JP, May JL, Carter AP, Morgan-Warren RJ, Wimberly BT and Ramakrishnan V (2001) Crystal structure of the 30S ribosomal subunit from Thermus thermophilus: purification, crystallization and structure determination. *Journal of Molecular Biology*, **310**, 827–843.

Clemons WM Jr, May JL, Wimberly BT, McCutcheon JP, Capel MS and Ramakrishnan V (1999) Structure of a bacterial 30S ribosomal subunit at 5.5 Å resolution. *Nature*, **400**, 833–840.

Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AM, David PR and Kornberg RD (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, **288**, 640–649.

Cramer P, Bushnell DA and Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, **292**, 1863–1876.

Fu J, Gnatt AL, Bushnell DA, Jensen GJ, Thompson NE, Burgess RR, David PR and Kornberg RD (1999) Yeast RNA polymerase II at 5 Å resolution. *Cell*, **98**, 799–810.

Gnatt A, Fu J and Kornberg RD (1997) Formation and crystallization of yeast RNA polymerase II elongation complexes. *The Journal of Biological Chemistry*, **272**, 30799–30805.

Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S, Mertens PP and Stuart DI (1998) The atomic structure of the bluetongue virus core. *Nature*, **395**, 470–478.

Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F and Yonath A (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679–688.

Jap B, Puhler G, Lucke H, Typke D, Lowe J, Stock D, Huber R and Baumeister W (1993) Preliminary X-ray crystallographic study of the proteasome from Thermoplasma acidophilum. *Journal of Molecular Biology*, **234**, 881–884.

Klaholz BP, Myasnikov AG and Van Heel M (2004) Visualization of release factor 3 on the ribosome during termination of protein synthesis. *Nature*, **427**, 862–865.

Knablein J, Neuefeind T, Schneider F, Bergner A, Messerschmidt A, Lowe J, Steipe B and Huber R (1997) Ta6Br(2+)12, a tool for phase determination of large biological assemblies by X-ray crystallography. *Journal of Molecular Biology*, **270**, 1–7.

Lowe J, Stock D, Jap B, Zwickl P, Baumeister W and Huber R (1995) Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 Å resolution. *Science*, **268**, 533–539.

O'Halloran TV, Lippard SJ, Richmond TJ and Klug A (1987) Multiple heavy-atom reagents for macromolecular X-ray structure determination. Application to the nucleosome core particle. *Journal of Molecular Biology*, **194**, 705–712.

Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, *et al*. (2000) Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, **102**, 615–623.

Schneider G (1994) Ta(6)Br(14) is a useful cluster compound for isomorphous replacements in protein crystallography. *Acta Crystallographica. Section D, Biological Crystallography*, **50**, 186–191.

Valle M, Sengupta J, Swami NK, Grassucci RA, Burkhardt N, Nierhaus KH, Agrawal RK and Frank J (2002) Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. *The EMBO Journal*, **21**, 3557–3567.

Vila-Sanjurjo A, Ridgeway WK, Seymaner V, Zhang W, Santoso S, Yu K and Cate JH (2003) X-ray crystal structures of the WT and a hyper-accurate ribosome from *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8682–8687.

von Bohlen K, Makowski I, Hansen HA, Bartels H, Berkovitch-Yellin Z, Zaytzev-Bashan A, Meyer S, Paulke C, Franceschi F and Yonath A (1991) Characterization and preliminary attempts for derivatization of crystals of large ribosomal subunits from Haloarcula marismortui diffracting to 3 Å resolution. *Journal of Molecular Biology*, **222**, 11–15.

Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T and Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH and Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.

Zhang G, Campbell EA, Minakhin L, Richter C, Severinov K and Darst SA (1999) Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 Å resolution. *Cell*, **98**, 811–824.

Short Specialist Review

Large complexes and molecular machines by electron microscopy

J. M. Plitzko and H. Engelhardt
Max-Planck-Institute of Biochemistry, Martinsried, Germany

1. Electron microscopy of native biological material

Over the past 70 years, researchers have used transmission electron microscopy (EM) to study biological objects such as cells, organelles, viruses, and isolated cellular components mostly chemically fixed and stained. However, during the last two decades, EM developed a great potential for studying the ultrastructure of native macromolecules and more recently even of living organisms at the nanometer scale. Biological material, however, is ill-suited for ultrahigh vacuum, and a constant "bombardment" of electrons creates an inhospitable environment in the electron microscope. Not until the introduction of sophisticated preparation and microscopy techniques, which are still in the state of being improved, has EM allowed in vivo studies of quasi-living organisms, native organelles, and macromolecular complexes. The commonly used preparations such as thin plastic-embedded sections or negatively stained protein complexes have been expanded and replaced by cryo-techniques, that is, the rapid freezing of biological material in the original environment or in their buffer solution on EM-grids (Adrian *et al*., 1984). Embedding samples in vitrified ice preserves the structure of interest with the demand that all the successive investigations have to be done at low temperature (~90 K) and under extremely low electron dose conditions to prevent structural changes of the specimen or smelting of the ice due to irradiation. Therefore, all low temperature methods and techniques in EM have been extended with the prefix "cryo". Nowadays, cryo-EM of isolated protein complexes and cryo-electron tomography of larger objects such as intact cells allow close-to-live investigations of biological structures at very high resolution.

2. Disclosure of the three-dimensional structure: the single-particle approach

One powerful advantage of electron microscopy is the fact that it produces projection images, unlike X-ray crystallography that yields diffraction data having

lost any information on position (phase problem). Thus, images may be treated in real space as well as in Fourier space and there is no difference whether they contain information of a single molecule without any internal symmetry or of a two-dimensional (2D) crystalline array. This given fact is exploited in the so-called single-particle approach, which enables the electron microscopist to determine three-dimensional (3D) density maps of individual macromolecules (Frank, 1996). Depending on a variety of conditions, the 3D models show near-atomic, or, in the characteristic case, intermediate resolution (≥7 Å) (Jiang *et al*., 2003; Spahn *et al*., 2001).

Projection images from macromolecules contain information on the 3D structure as 2D density profiles are characteristic of the particular orientation of the particles in the electron beam. To disentangle molecular substructures in the third dimension, a number of different images of known projection geometry have to be combined. Protein complexes not too large in size are usually oriented arbitrarily in vitrified ice, so that at least one EM image contains many different projections of the isolated protein species. Once the relative orientations of the particles are known, that is, the three Eulerian angles defining the rotation of a particle around the axes in 3D space, their projections can be virtually placed on the surface of a sphere centered to the origin of a coordinate system. The Eulerian angles exactly define the positions of the projections that can now be backprojected centrally such as to superimpose the 2D densities into a common 3D volume of the particle. The most popular and utilized reconstruction algorithm in real space is the so-called weighted backprojection (Radermacher *et al*., 1992). Very large macromolecular complexes, filamentous aggregates, or proteins embedded in biological membranes usually occur in preferred orientations in frozen samples. Here, and in cases in which nonredundant structures such as singular protein assemblies are to be investigated, the specimen is tilted around a fixed axis at a certain angular increment in the EM to obtain a series of projections. This approach of *electron tomography* is applicable almost universally if the specimen is not too thick for the electrons to penetrate.

Since the electron exposure must be restricted to a value so low so as not to destroy the sensitive biological material, the images contain a high level of so-called shot noise (statistical variation in the number of electrons detected at each pixel in the image). Additionally, electron scattering from elements low in atomic number (Z), like the constituents of most biological materials, for example, carbon, oxygen, and hydrogen produce a very weak image contrast. On the basis of these facts, biological structures are barely visible in single EM images (Figure 1a). Furthermore, the exact alignment of the particles becomes a challenge, which is tackled by averaging a large number of equivalent projections of separate molecules, thus reducing the noise level. To this end, the images of some 10^3 to 10^5 particles are sorted into distinct classes of views (Figure 1b) by multivariate statistical analysis (Van Heel and Frank, 1981). Once a large set of views is available, a preliminary 3D reconstruction can be computed and refined iteratively (Figure 1c). Typically, molecular complexes must be large in size (>200 kDa) to provide sufficient signal for the alignment at high resolution. If the particles possess a high internal symmetry, for example, virus capsids or regular structures like 2D crystals, the averaging and reconstruction process is simplified and the amount of necessary images and computation time is reduced.

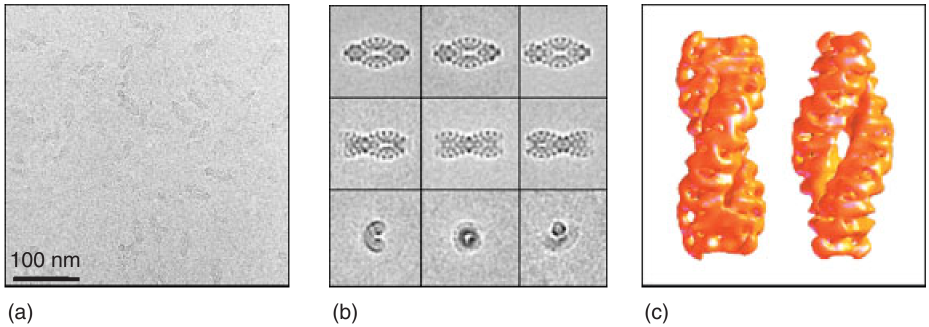


Figure 1 Single-particle investigation of the TPP II complex from *Drosophila melanogaster* embedded in vitrified ice. (a) Cryo-electron micrograph; (b) classification and averaging of a selection of different views; and (c) 3D representation of the protease TPP II complex (Reproduced from Rockel B, Peters J, Kuhlmorgen B, Glaeser RM and Baumeister W (2002) A giant protease with a twist: the TPP II complex from *Drosophila melanogaster* studied by electron microscopy. *EMBO Journal*, **22**, 5979–5984 by permission of Nature Publishing Group)

To conceive the structural basis of the working mechanism of molecular machines, we have to study the structures, positions, and interactions of their building blocks, that is, their distinct subunits. Since the structure of protein–protein contacts is only resolved at near-atomic resolution, it is hardly possible to identify molecular borders in 3D models of cryo-EM data by visual inspection. But the ability to identify elements of secondary structure, and α-helices in particular, at a resolution of about 6 to 8 Å makes it possible to fit a previously determined (atomic) model of protein subunits or subcomplexes into 3D density maps of larger assemblies obtained by cryo-EM (Li *et al*., 2002 and Bottcher *et al*., 1997). This approach, called *docking*, leads to unraveling the organization of macromolecular machines that cannot be structurally characterized otherwise. Right now researches continue to develop quantitative criteria that can guide this operation (Volkmann and Hanein, 1999).

3. The preparation problem

Single-particle cryo-EM is a powerful but still a slow method compared to X-ray crystallography, provided that for the latter, well-diffracting crystals are available. The data collection and processing can extend over several months or longer even for specialists in the field. However, with the advent of computer-controlled electron microscopes, data acquisition and analysis have become subjects for the development of elaborate automation procedures that promise to considerably reduce the time for creating high-resolution 3D density maps of large macromolecular complexes.

The real bottleneck for investigations of known or still unknown molecular machines is another experimental task. Commonly used purification procedures tend to select stable and abundant protein complexes only. Very large molecular machines, rare or transient macromolecular assemblies, complexes consisting of membrane spanning and soluble components, and all those held together by forces too weak to withstand the purification procedures, escape isolation or even

detection. The lesson is twofold: we need new approaches for isolating or reconstituting fragile assemblies, and cryo-EM of isolated single particles will probably not tell us the whole truth. To address the full complexity of functional macromolecular assemblies, noninvasive 3D imaging techniques have to take over to fulfill the task of studying the supramolecular architecture in its native, cellular context (Plitzko *et al*., 2002). The technique of *cellular cryo-electron tomography* (cryo-ET) is a “nouvelle route” with great potential in the field of structural biology that promises to provide a noninvasive and deep insight into the functional organization of the cellular proteome (Sali *et al*., 2003).

4. Cellular cryo-electron tomography

In almost the same manner as physicians obtaining tomograms from the interior of patients by means of computer tomography (CT) or magnetic resonance imaging (MRI), today’s structural biologists have begun to look inside cells using electron tomography (ET). Unlike in CT and MRI, where the patient remains still while the detector and the source rotates around him, ET applies a fixed electron beam and detector. The sample, containing native cells in frozen state (or other biological material), is tilted over a large angular range in the electron microscope. For every tilting step, an image is acquired resulting in a tilt series, which is used to reconstruct the structure of interest in 3D space. Typically, 100 to 200 images have to be recorded over an angular range of about $\pm 70^\circ$ or larger. Unfortunately, the ideal range of $\pm 90^\circ$ is not yet accessible due to the geometry of the sample and the specimen holders, with consequences for the resolution in the z-direction (Crowther *et al*., 1970).

Although it sounds straightforward to do electron tomography, and the first experiments were indeed carried out in the late 60s and early 70s of the last century (Hart, 1968; DeRosier and Klug, 1968; Hoppe, 1974), the successful application to native biological objects required substantial automation in both data acquisition and microscope control. The total exposure time for samples embedded in vitrified ice is limited to approximately 5 min for an entire tilt series at low-electron-dose conditions. Within this time frame, up to hundreds of images with a sufficient signal-to-noise ratio have to be recorded. This is only possible by automated computer control concerning corrections such as lateral shift of the region of interest and changes in focus upon every tilting step, with the aim of minimizing the necessary irradiation. In this way, only a few percent of the total electrons are used for the adjustments (Dierksen *et al*., 1995), a value that could never be achieved by manual operation. Nevertheless, the automated acquisition of a whole tilt series, for example, with an angular increment of 1.5°, requires 3 to 4 hr even today.

With automated cryo-electron tomography, it is possible to obtain molecular-resolution tomograms of structures as large and complex as whole prokaryotic or thin eukaryotic cells (Koster *et al*., 1997; Medalia *et al*., 2002; Gruenewald *et al*., 2003). Such tomograms are essentially 3D images of the cell’s entire proteome. They reveal information about the spatial relationship of macromolecules in the cytoplasm and provide us with the possibility of identifying large protein assemblies and molecular machines in their functional environment (Figure 2).

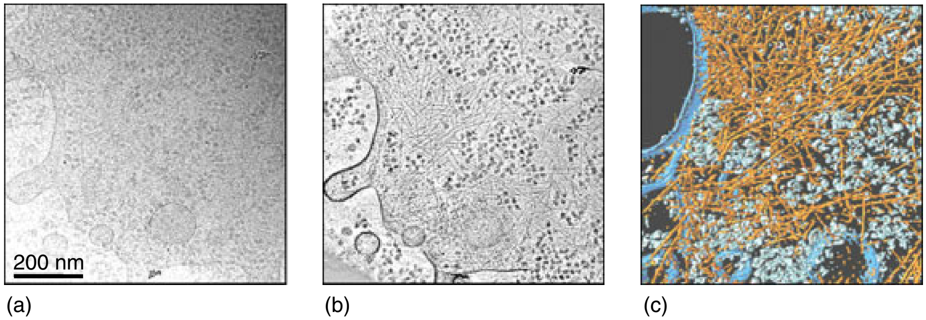


Figure 2 Electron tomographic investigation of the slime mold *Dictyostelium discoideum* embedded in vitrified ice. (a) Cryo-electron micrograph at 0°; tilt (conventional 2D projection); (b) tomographic reconstruction from a complete tilt series (120 images); and (c) visualization by segmentation. Large macromolecular complexes, for example, ribosomes are shown in green color, the actin filament network in orange-red and the cells' membrane in blue (Reprinted from Medalia O, Weber I, Frangakis AS, Gerisch G and Baumeister W (2002) Macromolecular architecture in eukaryotic cells visualized by cryo-electron tomography. *Science*, **298**, 1209–1213, **http://www.sciencemag.org** by permission of American Association for the advancement of Science)

However, the corresponding exploitation of tomograms is still hampered by two major technical problems, the inherent residual noise present in any cryo-EM data and the missing information resulting from the incomplete tilt range. Current efforts in electron microscopical image processing and technical development is focused on improving these situations. Moreover, the cytoplasm is very densely packed, with macromolecules literally touching each other, a phenomenon known as *macromolecular crowding* (Ellis, 2001). Feature extraction (*segmentation*) by visual inspection is therefore limited to large and easily recognizable structures, such as membranes or the cytoskeleton (Figure 2c).

The identification of specific molecules highly depends on the size of the complexes in question and on the resolution of the tomogram (which in turn depends on a number of technical features from the electron source to the camera recording the signal). Computer-based pattern recognition techniques have to be used for a systematic search of the reconstructed volume (Boehm *et al*., 2000; Frangakis *et al*., 2002). The prerequisite of such a molecular signature–based approach is a high-resolution structure of the template molecules obtained by X-ray crystallography or cryo-EM (Sali *et al*., 2003). It is obvious that the quest for molecules inside cellular tomograms is computationally demanding, suggesting parallel processing, since molecular complexes occur in any possible position (x, y, z) and orientation (Eulerian angles: θ, φ, ψ). Ideally, this search will be exhaustive, detecting all copies of the target molecules and their possible interactions inside the cell. In the future, the perfection of such pattern recognition tools will be crucial for taking full advantage of the power of cryo-ET.

With the current (nonisotropic) resolution of 4 to 6 nm, one can address larger complexes (>400 kDa) within a cellular context. However, there is still potential to widen the scope of cellular tomography. Ongoing instrumental developments, such as liquid helium cooling, improved detector designs and dual-axis tilting, reducing the missing angular information, make a resolution near 2 nm a realistic goal for the future.

Further reading

http://www.3dem-noe.org/.
http://www.interaction-proteome.org/.

References

Adrian M, Dubochet J, Lepault J and McDowall AW (1984) Cryo-electron microscopy of viruses. *Nature*, **308**, 32–36.

Boehm J, Frangakis AS, Hegerl R, Nickell S, Typke D and Baumeister W (2000) Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14245–14250.

Bottcher B, Wynne SA and Crowther RA (1997) Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, **386**, 88–91.

Crowther RA, DeRosier DJ and Klug A (1970) The reconstruction of a three-dimensional structure from its projections and its applications to electron microscopy. *Proceedings of the Royal Society, London*, **317**, 319–340.

DeRosier DJ and Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. *Nature*, **217**, 130–134.

Dierksen K, Typke D, Hegerl R, Walz J, Sackmann E and Baumeister W (1995) Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophysical Journal*, **68**, 1416–1422.

Ellis RJ (2001) Macromolecular crowding: obvious but underappreciated. *Trends in Biochemical Sciences*, **26**, 597–604.

Frangakis AS, Böhm J, Förster F, Nickell S, Nicastro D, Typke D, Hegerl R and Baumeister W (2002) Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(22), 14153–14158.

Frank J (1996) *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, Academic Press: London.

Gruenewald K, Desai P, Winkler DC, Heymann JB, Belnap DM, Baumeister W and Steven AC (2003) Three-dimensional structure of herpes simples virus from cryo-electron tomography. *Science*, **302**, 1396–1398.

Hart RG (1968) Electron microscopy of unstained biological material: the polytropic montage. *Science*, **159**, 1464–1467.

Hoppe W (1974) Towards three-dimensional electron microscopy at atomic resolution. *Zeitschrift Naturwissenschaften*, **61**, 239–249.

Jiang W, Li Z, Zhang Z, Baker ML, Prevelige PE Jr and Chiu W (2003) Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nature Structural Biology*, **10**, 131–135.

Koster AJ, Grimm R, Typke D, Hegerl R, Stoschek A, Walz J and Baumeister W (1997) Perspectives of molecular and cellular electron tomography. *Journal of Structural Biology*, **120**(3), 276–308.

Li HL, DeRosier DJ, Nicholson WV, Nogales E and Downing KH (2002) Microtubule structure at 8 Å resolution. *Structure (London, England)*, **10**(10), 1317–1328.

Medalia O, Weber I, Frangakis AS, Gerisch G and Baumeister W (2002) Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science*, **298**, 1209–1213.

Plitzko JM, Frangakis AS, Nickell S, Förster F, Gross A and Baumeister W (2002) In vivo veritas: electron cryotomography of cells. *Trends Biotech*, **20**(Suppl 8), S40–S44.

Radermacher M (1992) In *Weigthed Back-Projection Methods*, Frank J (Ed.) Electron Tomography Plenum Press: New York, pp. 91–115.

Rockel B, Peters J, Kuhlmorgen B, Glaeser RM and Baumeister W (2002) A giant protease with a twist: the TPP II complex from *Drosophila melanogaster* studied by electron microscopy. *EMBO Journal*, **22**, 5979–5984.

Sali A, Glaeser R, Earnest T and Baumeister W (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.

Spahn CMT, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G and Frank J (2001) Structure of the 80S ribosome from saccharomyces cerevisiae - tRNA-ribosome and subunit-subunit interactions. *Cell*, **107**(3), 373–386.

Van Heel M and Frank J (1981) Use of multivariate statistics in analyzing the images of biological macromolecules. *Ultramicroscopy*, **6**, 187–194.

Volkmann N and Hanein D (1999) Quantitative fitting of atomic models into observed densities derived by electron microscopy. *Journal of Structural Biology*, **125**(2–3), 176–184.

Short Specialist Review

The importance of protein structural dynamics and the contribution of NMR spectroscopy

Arnout P. Kalverda, Gary S. Thompson and W. Bruce Turnbull
University of Leeds, Leeds, UK

1. Introduction

Over the last 25 years, the traditional crystallographic picture of biomolecules as static entities has been superseded by the realization that proteins are conformationally flexible and display dynamic motions on a variety of timescales (*see* Article 95, **History and future of X-ray structure determination**, Volume 6 and Article 104, **X-ray crystallography**, Volume 6). Solution state Nuclear Magnetic Resonance (NMR) spectroscopy (*see* Article 105, **NMR**, Volume 6) experiments play a major role in revealing the conformational dynamics that are essential to many biochemical processes, including protein folding, molecular recognition, enzymatic catalysis, allosteric interactions, and interdomain motions. In such studies, the greatest strength of NMR lies in being able to monitor simultaneous processes that may occur on picosecond to millisecond timescales (Figure 1), and at multiple atom-specific sites within a biomolecule, without recourse to site-specific labeling. Although the technique is limited to systems of tractable mass (currently <100 kDa), requires biomolecules enriched in certain stable isotopes (typically ^{2}H, ^{13}C, and ^{15}N), and employs specialized equipment and experiments, the rich insights that can be derived from NMR experiments more than justify investment in the technique.

2. NMR methods for measuring protein dynamics

The NMR methods used for measuring dynamics in biomolecules can be classified into four main categories in order of current importance: relaxation, chemical exchange effects, hydrogen exchange measurements, and averaging of dipolar couplings.

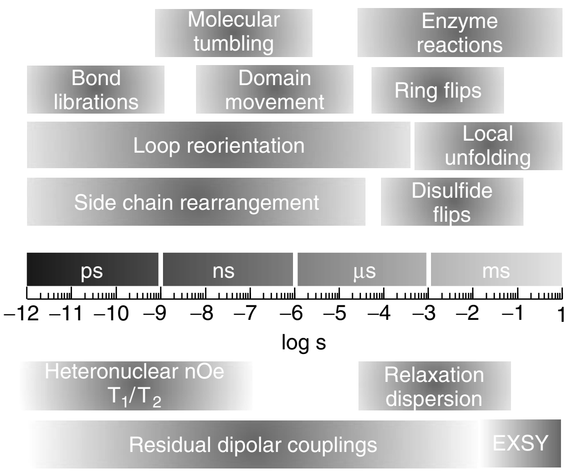


Figure 1 Timescales for different dynamic processes exhibited by proteins, and NMR methods applicable for studying these processes. EXSY: exchange spectroscopy

In NMR, unlike most forms of spectroscopy, excited states require stimulation to return (relax) to equilibrium (*see* Article 105, **NMR**, Volume 6). Ultimately, this stimulation is provided through the Brownian motion of the molecules. The frequency spectrum of motions arising from overall molecular tumbling, and also faster internal motions, thus determines the relaxation rate. Longitudinal (T_1) and transverse (T_2) relaxation and nuclear Overhauser effect measurements can provide information on molecular motions occurring on nanosecond–picosecond timescales (Palmer, 2001). Together, these parameters can be used to calculate conformational entropies for individual residues (Palmer, 2001).

Where two or more different conformations of a protein interconvert on a microsecond–millisecond timescale, chemical exchange effects may arise in the NMR spectra. In such instances, methods such as relaxation dispersion experiments can be used to determine the exchange rate for these processes (Loria *et al*., 1999).

Hydrogen atom exchange between a protonated protein and a deuterated solvent can be used to monitor the breaking and forming of hydrogen bonds associated with dynamics on timescales of seconds to months, for example, protein breathing motions or unfolding (Krishna *et al*., 2004).

Residual dipolar couplings (RDCs) provide a new method that may allow dynamics to be measured over a wide variety of timescales (Lukin *et al*., 2003), including microsecond–millisecond. RDCs arise when molecules become partially aligned with the magnetic field, a phenomenon that may occur in the presence of a molecular barrier, such as a liquid crystal. The magnitudes of RDCs are reduced in dynamic systems, as internal molecular motions average away the effects of the molecular alignment (Griesinger *et al*., 2004). RDCs have been employed successfully for defining the extent of domain motions in multidomain proteins. Together, these different methods allow the study of molecular motions that occur over a very wide range of timescales, including those that are critical to a protein's function.

3. Enzyme active site dynamics

Often, changes in protein conformation and/or dynamics occur in response to ligand, or substrate, binding. Thus, conformational dynamics must play an integral role in enzymatic processes. The binding site must be accessible for its substrate to dock, yet the catalytic reaction may have to be shielded from solvent, and finally, the enzyme must be able to release the reaction products to allow catalytic turnover. NMR relaxation measurements often reveal the presence of microsecond–millisecond dynamics in enzyme active sites. In some cases, it can been shown that these dynamics are correlated with the processes involved and it is even possible to monitor them during catalytic turnover (Eisenmesser *et al*., 2002).

The ubiquitous enzyme dihydrofolate reductase (DHFR, EC 1.5.1.3) has been studied extensively by NMR and crystallography (Schnell *et al*., 2004). NMR studies of DHFR:folate complexes show that the Met20 loop, which protects the active site during catalysis, is in exchange between two states (occluded and closed) that were observed previously in X-ray structures. On forming the full DHFR:Folate:$NADP^+$ product complex, the Met20 loop closes around the nicotinamide ring of $NADP^+$ in its binding pocket (Figure 2). This change in conformation, from occluded to closed forms, is mirrored by a reduction in the amplitude of picosecond–nanosecond motion of backbone atoms in the Met20 loop. NMR dynamics measurements of this system have also indicated that the dynamics of the Met20 loop control product release and turnover. Active site motions, which occur on the same timescale as the key catalytic hydride transfer step, have also been detected by NMR methods.

The conformation of catalytic residues that stabilize a transition state may be of higher energy than the ground state conformation of the protein. Dynamic motions of catalytically important residues can thus control catalysis by repositioning side chains in their most productive conformations. Furthermore, dynamic fluctuations in the active site can potentially be harnessed to drive the reaction, and more distal

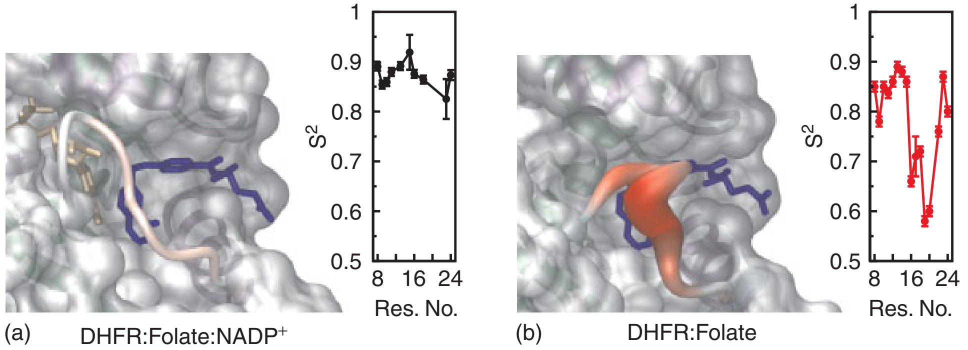


Figure 2 Structure and dynamics of the DHFR binding site showing the Met20 loop in (a) the closed [1RX2], and (b) the occluded conformations [1RX7]. The low-order parameters (S^2) shown for the occluded conformer demonstrate that the loop is very flexible (Schnell *et al*., 2004). The degree of flexibility is indicated on the protein structure by the width of the sausage representation of the Met20 loop and also by its color (white to red = ordered to disordered). Folate and $NADP^+$ are depicted in blue and gold, respectively

changes in dynamics can help reduce the entropic cost of forming the catalytic state (see next section).

4. Entropy, allostery, and preexisting equilibria

Comparative studies of a receptor in the presence and absence of ligand allow measurement of changes in dynamics that occur during complex formation. Amino acid residues forming a binding site are often rigidified upon association with a ligand. Although this process makes an unfavorable entropic contribution to the free energy change, its effect is often reduced by more distant residues becoming more mobile (Stone, 2001). In cases where proteins have multiple binding sites, differential dynamics on ligand binding can also contribute to allosteric effects. For example, when one Ca^{2+} ion binds to calbindin D9k, the remaining, unoccupied binding site becomes more rigid. The entropic cost of binding a Ca^{2+} ion to the second site is thus lowered, and hence positive cooperativity or 'dynamic allostery' results (Maler *et al*., 2000). Conversely, for cytidyltransferase, both binding sites retain some flexibility on binding the first ligand but decrease their dynamics on binding to a second ligand molecule, with a concomitant negative cooperativity (Stevens *et al*., 2001).

Protein activation pursuant to phosphorylation, or interactions with a ligand or another protein, is widely believed to result from changes in conformation. However, as proteins exist in solution as a dynamic ensemble of conformational substates, the active conformer may already be present, albeit in low concentration. NMR studies demonstrating the presence of millisecond timescale dynamics or domain orientations that differ from those in crystal structures can provide evidence for the existence of multiple interconverting conformations. For example, NMR studies have revealed that hemoglobin, the archetypal allosteric system, has an interdomain orientation, in solution, that is midway between the two allosterically distinct states (Lukin *et al*., 2003). A shift in equilibrium between the preexisting, different conformational states upon oxygen binding could be involved in the allosteric effect.

Protein conformational dynamics influence the thermodynamics of essentially all biomolecular interactions. Understanding these dynamic processes is essential if we are to succeed in rational drug design, and NMR spectroscopy continues to be the most powerful technique available to achieve this goal.

Related articles

Article 35, **Structural biology of protein complexes**, Volume 5; Article 36, **Biochemistry of protein complexes**, Volume 5; Article 45, **Computational methods for the prediction of protein interaction partners**, Volume 5; Article 94, **What use is a protein structure?**, Volume 6; Article 96, **Fundamentals of protein structure and function**, Volume 6; Article 102, **New approaches to bridging the timescale gap in the computer simulation of biomolecular processes**, Volume 6

References

Eisenmesser EZ, Bosco DA, Akke M and Kern D (2002) Enzyme dynamics during catalysis. *Science*, **295**, 1520–1523.

Griesinger C, Peti W, Meiler J and Brüschweiler R (2004) Projection angle restraints for studying structure and dynamics of biomolecules. In *Protein NMR Structure Techniques*, Second Edition, *Methods in Molecular Biology Volume 278*, Downing AK (Ed.), Humana Press: Totawa, pp. 107–122.

Krishna MMG, Hoang L, Lin Y and Englander SW (2004) Hydrogen exchange methods to study protein folding. *METHODS*, **34**, 51–64.

Loria JP, Rance M and Palmer AG III (1999) A relaxation-compensated Carr-Purcell-Meiboom-Gill sequence for characterizing chemical exchange by NMR spectroscopy. *Journal of the American Chemical Society*, **121**, 2331–2332.

Lukin JA, Kontaxis G, Simplaceanu V, Bax A and Ho C (2003) Quaternary structure of hemoglobin in solution. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 517–520.

Maler L, Blankenship J, Rance M and Chazin WJ (2000) Site-site communication in the EF-hand Ca^{2+}-binding protein calbindin D_{9k}. *Nature Structural Biology*, **7**, 245–250.

Palmer AG III (2001) Nmr probes of molecular dynamics: Overview and comparison with other techniques. *Annual Review of Biophysics and Biomolecular Structure*, **30**, 129–155.

Schnell JR, Dyson JH and Wright PE (2004) Structure, dynamics and catalytic function of dihydrofolate reductase. *Annual Review of Biophysics and Biomolecular Structure*, **33**, 119–140.

Stevens SY, Sanker S, Kent C and Zuiderweg ERP (2001) Delineation of the allosteric mechanism of a cytidylyltransferase exhibiting negative cooperativity. *Nature Structural Biology*, **8**, 947–952.

Stone MJ (2001) NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Accounts of Chemical Research*, **34**, 379–388.

Short Specialist Review

New approaches to bridging the timescale gap in the computer simulation of biomolecular processes

Leo S. D. Caves
University of York, York, UK

Chandra S. Verma
Bioinformatics Institute, Singapore

1. Introduction

A recent study of water permeation across a biological membrane (de Groot and Grubmüller, 2001) is something of a landmark in the field of atomistic computer simulations in molecular (structural) biology (Karplus and McCammon, 2002). Although significant work has been going on in the field for decades (Duan and Kollman, 1998; McCammon *et al*., 1977), this study represents the first time that a significant biological event was observed in "real time" in a computer simulation. While there are several examples of the successful use of simulations to provide insight into the mechanisms underlying fundamental biomolecular processes (Karplus and Gao, 2004; Garcia-Viloca *et al*., 2004; Marszalek *et al*., 1999), investigations of events occurring on a biologically relevant timescale are still in their infancy (Snow *et al*., 2002). The main bottleneck in the application of molecular dynamics (MD) simulations is the timescales that can be simulated, which are typically orders of magnitude smaller than those characteristic of the processes of interest to most biochemists and biologists.

2. Addressing the timescale problem: new approaches in biomolecular simulation

The agreement of MD simulations with experimental data (where available) for processes on short timescales is heartening (de Groot and Grubmüller 2001; Snow *et al*., 2002; Meller and Elber 1998). However, there is a need for alternative methods, complementary to MD, that extend the reach of simulation and that widen the scope of applications to encompass biomolecular processes on timescales that begin to approach physiological relevance (Elber *et al*., 2003).

One class of methods, that fully exploits the fruits of structural biology, makes use of multiple structures (such as the *reactant*, *product*, and/or intermediate states) that are representative of a biomolecular process of interest. These methods address the pathway of conformational changes within (and between) the biomolecule(s) associated with the process. One approach (e.g., Fischer and Karplus, 1992) addresses the problem of the pathway from the perspective of the classical (adiabatic) reaction coordinate, and focuses on features of its underlying potential energy surface. Such approaches are essentially static, and dynamic information must be inferred by the application of appropriate rate theories. By contrast, a second class of methods attempt to address the timescale of the process by explicit consideration of the dynamics associated with the pathway (Elber *et al*., 2003; Straub *et al*., 2002; Bolhuis *et al*., 2002).

The former approaches can generate a pathway that is representative of a process at (absolute) zero temperature and focus on locating stationary points and other features of the potential energy along the reaction coordinate. However, while they yield accurate enthalpic barriers, they do sacrifice the recovery of entropic information and thus any direct considerations of the free energy of the process. The free-energy profile (the *potential of mean force*) is typically recovered by subsequent conformational sampling about the adiabatic reaction coordinate by constraint or restraint methods (Roux, 1995).

The latter class of methods recognizes and embraces the fact that, in general, biological processes operate in a fairly limited range of finite temperatures. The consequence of an explicit consideration of temperature in a process is that a single reaction pathway picture is not appropriate and processes are best characterized by a distribution (or ensemble) of pathways, which are typically considered (and optimized) as a whole (in contrast to a focus on individual critical points). Timescales in such methods are typically recovered by the appropriate consideration of energetic properties (such as the classical action) of the whole pathways associated with the process (Elber *et al*., 2003).

3. Success stories in the simulation of biomolecular processes

A specific algorithm (of the first class mentioned above), the conjugate peak refinement (CPR) method, which determines the saddle points associated with the process of conformational change between two given states of a biomolecule, has been implemented in the TRAVEL module of the CHARMM simulation software (Fischer and Karplus, 1992). We and others have applied CPR to the generation of pathways of biomolecular conformational processes in systems as diverse as ice and proteins, and have found very good agreement with experimental activation enthalpies (Verma *et al*., 1996; Fischer *et al*., 1998; Bondar *et al*., 2004). In one particular case (Fischer *et al*., 1998), the method was extended to estimate the activation entropy. In this application, it was fortuitous that the processes exhibited a relatively "smooth" energy profile. In general, a biomolecular pathway is likely to be characterized by a "rugged" energy profile, which makes robust estimates of thermodynamic parameters (via methods such as importance sampling) substantially more difficult.

4. Simulations in the era of large-scale computing: sampling and sophistication

The rugged character of the energy landscape of biomolecules (Frauenfelder *et al*., 2003), which confers their rich and complex dynamical behavior, poses a sampling problem in biomolecular simulation (Clarage *et al*., 1995). There has been a long-standing debate (without consensus) in the field of MD simulations as to whether it is best to generate one long trajectory or several short ones (Caves *et al*., 1998; Auffinger *et al*., 1995). The advent of high-performance clustered and distributed computing, and thus the unprecedented opportunity to perform very long simulations (e.g., Snow *et al*., 2002), offers a way forward in this regard.

Already, initial concern over whether the current force fields, which have traditionally been parameterized and tested over short timescales, would behave well at longer timescales seems to be dying. Other uses of the larger capacity for simulation resulting from ever-increasing computing power are in improvements to the representation of the underlying system interactions and environmental conditions such as pH (Janin and Simonson, 2004). However, the traditional tension endures: the cost of the increased physical fidelity obtained by greater sophistication of the model representation must be traded against the cost of effective and representative sampling (the extent and duration of the simulations).

5. Looking to the future: exploiting the convergence of timescales simulation and experimental studies of protein dynamics

Commensurate with the increasing reach of simulation methods into longer time-scale processes, is the increasing resolution of time-resolved (single molecule) spectroscopies (Michalet *et al*., 2003). In this context, we have embarked on a program of research to learn how best to exploit this convergence of simulation and experimental timescales (Figure 1). Initially, we are working from the perspective of a protein as a complex many-body dynamical system. As such, we take inspiration in our approach from the concepts of dynamical systems theory and in particular the embedding theorem (Takens, 1981; Packard *et al*., 1980). Briefly, the embedding theorem states that a scalar time series of a dynamical system contains sufficient information to allow for the reconstruction of its (multidimensional) phase space (Sauer *et al*., 1991).

Thus, if we have a scalar signal from a protein, such as the distance between the lobes of a two-domain protein – the kind of information that is amenable to extraction using current state-of-the-art single molecule spectroscopy – then from the methods of *embedology* (Sauer *et al*., 1991), this signal should contain sufficient information to characterize the inherent dynamics of the whole protein. This could, in principle, lead to knowledge of the system phase space geometry and topology, which are fundamental to the characterization and understanding of a dynamical system.

This is suggestive of a potentially rich means for comparing the view of protein dynamics derived from experimental methods, with that emerging from the ever-increasing reach of molecular dynamics simulation(s). In a model study of the protein

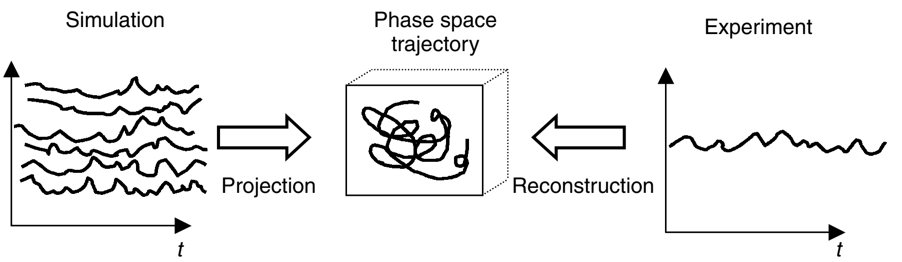


Figure 1 A schematic representation of the potential exploitation of the simulated and experimental data on biomolecular dynamics from a dynamical systems perspective. The left panel represents the multidimensional time series of properties (such as atomic coordinates) characteristic of trajectories generated from MD simulations. A low-dimensional view of the phase (conformational) space sampled in such trajectories is often generated by multivariate projection methods, such as principal component analysis (PCA). The right panel represents a time series of a scalar property of a biomolecule, such as that available from time-resolved spectroscopies. The high(er) dimensional phase space of the biomolecular system may be reconstructed by embedding the time series in a high(er) dimensional space. As the timescale gap closes, a consideration of the phase space behavior in both approaches has the potential for a more intimate comparison of the experimental and simulated data on protein dynamics leading to greater insight into the rich phenomena of biomolecular dynamics

crambin (Caves and Verma, 2002) using simulated data, we have provided an empirical demonstration of the embedding theorem in the context of biomolecular dynamics. We now look forward to applications contrasting experimental and simulated data on the same system.

Further reading

Böckmann R and Grubmüller H (2002) Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase. *Nature Structural Biology*, **9**, 198–202.

References

Auffinger P, Louise-May S and Westhof E (1995) Multiple molecular-dynamics simulations of the anticodon loop of tRNA(ASP) in aqueous-solution with counterions. *Journal of the American Chemical Society*, **117**, 6720–6726.

Bolhuis PG, Chandler D, Dellago C and Geissler P (2002) Transition path sampling: throwing ropes over mountain passes, in the dark. *Annual Review of Physical Chemistry*, **59**, 291–318.

Bondar A-N, Elstner M, Suhai S, Smith JC and Fischer JC (2004) Mechanism of primary proton transfer in bacteriorhodopsin. *Structure*, **12**, 1281–1288.

Caves LSD, Evanseck JD and Karplus M (1998) Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Science*, **7**, 649–666.

Caves LSD and Verma CS (2002) Congruent behaviour of complete and reconstructed phase space trajectories from biomolecular dynamics simulation. *Proteins: Structure, Function, Genetics*, **47**, 25–30.

Clarage JB, Romo T, Andrews BK, Pettitt BM and Phillips GN (1995) A sampling problem in molecular dynamics simulations of macromolecules. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 3288–3292.

de Groot BL and Grubmüller H (2001) Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science*, **294**, 2353–2357.

Duan Y and Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**(5389), 740–744.

Elber R, Ghosh A, Ghosh A and Stern HA (2003) Bridging the gap between reaction pathways, long time dynamics and calculation of rates. *Advances in Chemical Physics*, **126**, 93–129.

Fischer S and Karplus M (1992) Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chemical Physics Letters*, **194**, 252–261.

Fischer S, Verma CS and Hubbard RE (1998) Rotation of buried water inside a protein: calculation of the rate and vibrational entropy of activation. *The Journal of Physical Chemistry B*, **102**, 1797–1805.

Frauenfelder H, McMahon BH and Fenimore PW (2003) Myoglobin: the hydrogen atom of biology and a paradigm of complexity. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(15), 8615–8617.

Garcia-Viloca M, Gao J and Karplus M (2004) How enzymes work: analysis by modern rate theory and computer simulations. *Science*, **303**, 186–195.

Janin J and Simonson T (2004) Theory and simulation: from protons to genomes. *Current Opinion in Structural Biology*, **14**(2), 117–259.

Karplus M and Gao YQ (2004) Biomolecular motors: the F1-ATPase paradigm. *Current Opinion in Structural Biology*, **14**(2), 250–259.

Karplus M and McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, **9**, 646–652.

Marszalek PE, Lu H, Li H, Carrion-Vazquez M, Oberhauser AF, Schulten K and Fernandez JM (1999) Mechanical unfolding intermediates in titin modules. *Nature*, **402**(6757), 100–103.

McCammon JA, Gelin BR and Karplus M (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.

Meller J and Elber R (1998) Computer simulations of carbon monoxide photodissociation in myoglobin: structural interpretation of the B states. *Biophysical Journal*, **74**, 789–802.

Michalet X, Kapanidis AN, Laurence T, Pinaud F, Doose, S, Pflughoefft M and Weiss S (2003) The power and prospects of fluorescence microscopies and spectroscopies. *Annual Review of Biophysics and Biomolecular Structure*, **32**, 161–182.

Packard NH, Crutchfield JP, Farmer JD and Shaw RS (1980) Geometry from a time series. *Physical Review Letters*, **45**, 712.

Roux B (1995) The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, **91**, 275.

Sauer TJ, Yorke JA and Casdagli M (1991) Embedology. *Journal of Statistical Physics*, **65**, 579–616.

Snow CD, Nguyen H, Pande VS and Gruebele M (2002) Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, **420**(6911), 102–106.

Straub JE, Guevara J, Huo S, Huo JP and Lee JP (2002) Long time dynamic simulations: exploring the folding pathways of an alzheimer's amyloid beta-peptide. *Accounts of Chemical Research*, **35**, 473–481.

Takens F (1981) Detecting strange attractors in turbulence. In *Lecture Notes in Mathematics*, Rand DA and Young LS (Eds.), Springer: Berlin, pp. 366–381.

Verma CS, Fischer S, Caves LSD, Roberts GCK and Hubbard RE (1996) Calculation of the reaction pathway for the aromatic ring flip in methotrexate complexed to dihydrofolate reductase. *The Journal of Physical Chemistry*, **100**, 2510–2518.

Short Specialist Review

Modeling membrane protein structures

Yalini Arinaminpathy and Mark S.P. Sansom
University of Oxford, Oxford, UK

Integral membrane proteins (*see* Article 15, **Handling membrane proteins**, Volume 5) constitute ~25% of all genes (Wallin and von Heijne, 1998) and play key functional roles in cells, for example, as receptors, ion channels, and transporters. Hence, they dominate lists of predicted drug targets (Terstappen and Reggiani, 2001). Despite their importance, comparatively little is known of the structure of membrane proteins. Of the ~24 000 structures in the current PDB, only ~50 are membrane proteins. This disparity is largely due to the experimental difficulties in overexpression of integral membrane proteins and in growth of crystals suitable for structural determination by X-ray diffraction. Alternatives to X-ray diffraction have been used to determine membrane protein structures, including NMR from detergent/protein micelles and electron microscopy from two-dimensional membrane crystals. However, the resolution of the resultant structures does not match that of X-ray diffraction. A further limitation is that overexpression of *mammalian* membrane proteins remains a challenge. Thus, the majority of the structures of, for example, ion channels and transporters are for bacterial homologs. For all of these reasons, there is an almost overwhelming need to generate accurate structural models of membrane proteins, especially to aid in dissection of structural/functional relationships by mutagenesis and related methods. Models of membrane proteins can also be used to study their conformational dynamics. Both modeling *per se* and simulations of membrane protein dynamics have been made feasible by recent advances in computing power and software.

Modeling the structures of membrane proteins is somewhat simpler than that for soluble proteins as the constraints introduced by the membrane restrict the number of possible architectures. Indeed, other than the bacterial outer membrane proteins (Wimley, 2003), the vast majority of membrane proteins are composed of bundles of transmembrane (TM) α-helices. Membrane protein modeling can be broken down into three levels: prediction from sequence of the location of TM helices; homology modeling of membrane protein folds; and model evaluation and refinement.

Prior to attempting to model the three-dimensional structure of a membrane protein, it is important to identify the TM domain(s) within the larger membrane protein and to identify the location of the membrane-spanning α-helices. A number of TM helix and topology prediction methods are available. A recent study (Cuthbertson, Sansom & Doyle, ms. in preparation) suggests that most accurate

methods include TMHMM2 (Krogh *et al*., 2001), MEMSAT2 (Jones, 1998), and SPLIT4 (Juretic *et al*., 2002) and that a consensus method may be a little more accurate. However, although these methods exceed 80% accuracy in their overall per residue accuracy of TM helix prediction, they are still inevitably approximate in predicting the exact start and end positions of TM helices.

If there is detectable sequence similarity between two proteins, structural similarity can be assumed. If the three-dimensional structure of a (usually bacterial) homolog of a membrane protein of interest is known, then one may use homology modeling. Homology modeling is composed of four sequential stages. The first and most crucial step is template selection. The structural similarity between the template and target declines as the sequence identity between the two proteins decreases. If there is less than ∼25% identity between two sequences, identifying a suitable template may be problematic, although prior biological knowledge and the use of more sensitive sequence similarity search methods (e.g., PSI-BLAST) may help. Once a template has been found, the sequences must be optimally aligned. Alignment quality has a major impact on the accuracy of the resultant three-dimensional models. An optimal alignment is more difficult to achieve as the sequence identity falls, with alignment errors becoming common in the "twilight zone" of less than ∼30% identity. Multiple sequence comparisons and experimental data, for example, from mutagenesis studies may be used in optimizing an alignment.

Once an alignment has been constructed, a three-dimensional model of the protein may be built using the template structure to determine the fold of the model. This can be done using a number of homology modeling programs, for example, MODELLER (Sali and Blundell, 1993) (**http://salilab.org/modeller/modeller.html**). In general, spatial restraints are mainly determined from the alignment of the target sequence to the template protein structure, but may also be derived from, for example, NMR, cross-linking, or site-directed mutagenesis data. Models are generated by minimizing the violations of all the specified restraints. Loop regions may be modeled by searching a database of known loops. Spatial restraints are combined with potential energy terms (e.g., from the CHARMM force field) in order to enforce correct (local) stereochemistry. Together, the restraints and energy terms yield an objective function. A model is generated by optimizing this function in Cartesian space.

It is important to try to assess the quality of a model before further interpreting it. If errors are found in the model, it may be necessary to revisit earlier steps in the modeling process. For example, one may wish to adjust the initial alignment or perhaps even search for an alternative template (if available). Stereochemical tests of model quality within, for example, MODELLER may be supplemented by more general "protein-checking" programs such as PROCHECK (Laskowski *et al*., 1993). In addition to these general programs, we may examine key features of the model in the context of our increased understanding of aspects of membrane protein architecture (Ulmschneider and Sansom, 2001). With an increasing number of structures being determined, it is becoming possible to use structural bioinformatics tools to identify general features of membrane proteins. For example, statistical analyses of known structures have revealed a preference for glycine residues at the interfaces between tightly packed TM helices (Javadpour *et al*., 1999). Also, proline

residues occur with a high frequency within TM helices, where they may function as a molecular hinge (Cordes *et al*., 2002). Also important are "belts" of amphipathic aromatic residues (i.e., Trp, Tyr) at the membrane/water interface (Killian and von Heijne, 2000) and (especially in bacterial membrane proteins) a preponderance of basic residues (Lys, Arg) immediately adjacent to the intracellular ends of TM α-helices (von Heijne, 1992). Such rules derived from an increasing dataset of structures can be used to help evaluate the plausibility of membrane protein models.

After a membrane protein model has been generated, it may be further evaluated and explored by performing molecular dynamics simulations (Domene *et al*., 2003) of the model protein embedded in a lipid bilayer. Such simulations enable one to explore the conformational dynamics of the protein in relation to its biological function. The degree of conformational drift in such simulations may also aid in the evaluation of the stability of a model.

Modeling studies have been used to study a number of membrane proteins, including members of the aquaporin (de Groot *et al*., 2001) and potassium channel (Capener *et al*., 2000) families. As an example of such studies, we will take the case of the TM domain of the bacterial glutamate receptor, GluR0, which being a combination of sequence comparison and functional studies (Chen *et al*., 1999) has demonstrated to be homologous with a bacterial K channel, KcsA (Zhou *et al*., 2001). GluR0 (which forms a potassium-selective glutamate-activated channel) has a 31% sequence identity to KcsA within the core TM domain formed by the α-helical M1, P, and M2 regions. An alignment was generated for the core M1, P, and M2 regions, taking notice of the fingerprint TVGYG motif associated with K^+ selectivity, which is present in both KcsA and GluR0. MODELLER was employed to construct an ensemble of 25 models of the tetrameric TM domain GluR0, using secondary structure restraints to ensure that the M1, P-helix, and M2 regions were α-helical. The 25 models were ranked by analysis of their stereochemistry and deviation from the template, KcsA. The highest-ranking structure was selected for further analysis and as a starting structure for molecular dynamics (*see* Article 102, **New approaches to bridging the timescale gap in the computer simulation of biomolecular processes**, Volume 6) simulations (Arinaminpathy *et al*., 2003).

The similarity of surface side chains in the KcsA crystal structure and the model of the TM domain of GluR0 support the plausibility of the model (see Figure 1). For example, both the crystal structure and the model have bands of amphipathic aromatic side chains (i.e., Trp, Tyr) at the presumed location of the lipid headgroup/water interface. Furthermore, both the crystal structure and the model have rings of acidic side chains at the mouth of the pore, a feature associated with potassium channels. MD simulations based on the GluR0 model confirm that it forms a relatively stable TM domain and that it will allow permeation of K^+ ions through the membrane. Thus, the model may be used to aid interpretation of aspects of GluR0 function.

In summary, as our database of membrane protein structures expands, molecular modeling will play an increasingly important role in enabling extrapolation from known structures of bacterial membrane proteins to generate models of their human homologs. As the pace of membrane protein structure grows, it will be important to develop high throughput approaches for modeling and simulation of membrane proteins.

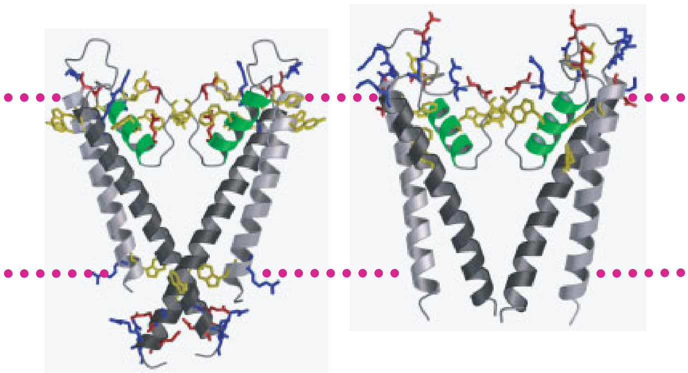


Figure 1 The structure of an integral membrane protein (the TM domain of a bacterial potassium channel, KcsA) is compared with a homology model (the TM domain of a bacterial glutamate receptor, GluR0). In each case only two of the four subunits are shown, for clarity. Basic (blue), acidic (red) and amphipathic aromatic (yellow) side chains are shown. The dotted horizontal lines indicate the approximate location of the lipid head groups of the membrane in which the proteins are embedded

References

Arinaminpathy Y, Biggin PC, Shrivastava IH and Sansom MSP (2003) A prokaryotic glutamate receptor: homology modelling and molecular dynamics simulations of GluR0. *FEBS Letters*, **553**, 321–327.

Capener CE, Shrivastava IH, Ranatunga KM, Forrest LR, Smith GR and Sansom MSP (2000) Homology modelling and molecular dynamics simulation studies of an inward rectifier potassium channel. *Biophysical Journal*, **78**, 2929–2942.

Chen GQ, Cui CH, Mayer ML and Gouaux E (1999) Functional characterization of a potassium-selective prokaryotic glutamate receptor. *Nature*, **402**, 817–821.

Cordes FS, Bright JN and Sansom MSP (2002) Proline-induced distortions of transmembrane helices. *Journal of Molecular Biology*, **323**, 951–960.

de Groot BL, Engel A and Grubmüller H (2001) A refined structure of human aquaporin-1. *FEBS Letters*, **504**, 206–211.

Domene C, Bond P and Sansom MSP (2003) Membrane protein simulation: ion channels and bacterial outer membrane proteins. *Advances in Protein Chemistry*, **66**, 159–193.

Javadpour MM, Eilers M, Groesbeek M and Smith SO (1999) Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophysical Journal*, **77**, 1609–1618.

Jones DT (1998) Do transmembrane protein superfolds exist? *FEBS Letters*, **423**, 281–285.

Juretic D, Zoranic L and Zucic D (2002) Basic charge clusters and predictions of membrane protein topology. *Journal of Chemical Information and Computer Sciences*, **42**, 620–632.

Killian JA and von Heijne G (2000) How proteins adapt to a membrane-water interface. *Trends in Biochemical Sciences*, **25**, 429–434.

Krogh ALB, von Heijne G and Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

Laskowski RA, Macarthur MW, Moss DS and Thornton JM (1993) Procheck – a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291.

Sali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815.

Terstappen GC and Reggiani A (2001) In silico research in drug discovery. *Trends in Pharmacological Sciences*, **22**, 23–26.

Ulmschneider MB and Sansom MSP (2001) Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta*, **1512**, 1–14.

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive inside rule. *Journal of Molecular Biology*, **225**, 487–494.

Wallin E and von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archean, and eukaryotic organisms. *Protein Science*, **7**, 1029–1038.

Wimley WC (2003) The versatile β-barrel membrane protein. *Current Opinion in Structural Biology*, **13**, 404–411.

Zhou Y, Morais-Cabral JH, Kaufman A and MacKinnon R (2001) Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 Å resolution. *Nature*, **414**, 43–48.

X-ray crystallography

Susan Firbank
Wadsworth Centre, New York State Department of Health, Albany, NY, USA

1. Introduction

Before the reporting of the first protein crystal structure, that of myoglobin, in the late 1950s (Kendrew *et al*., 1958), little was known or even imagined about the form of proteins. Since then, approximately 20 000 X-ray structures have been deposited in the Protein Data Bank (PDB) (**http://www.rcsb.org/pdb**). They have revealed how proteins are folded and how they carry out their functions: enzymes have been trapped with substrates bound at the active site, illustrating the reactive mechanisms; receptors are found with their ligands bound revealing the conformational changes induced upon binding; and protein complexes have been observed in their functional arrangements. Traditionally, structures determined have been of proteins for which the functions are known. Today, however, structural-genomics projects are providing new scope and opportunity for the crystallographic community to assist in the characterization of a long list of, so-called, hypothetical proteins – proteins for which only the sequence is known. It is hoped that where sequence analysis fails, the structure will allow functions to be assigned to the uncharacterized proteins. Nevertheless, regardless of the purpose of a particular structure determination, the methods that must be applied are the same.

2. The crystallographic method

Structure determination by X-ray crystallography can at first be considered analogous to the examination of objects under the light microscope. As the light passes through the sample, it is diffracted and then refocused with lenses to allow visualization of the sample object. A similar process can be carried out to examine proteins, although electromagnetic radiation in the visible range cannot be used since the required information is on a scale much smaller than the wavelength of visible light. X rays with wavelengths in the low nanometer range are used for structural analysis of proteins, since they are of the same order of magnitude as the atom separation within proteins: ~ 0.1 nm, or 1 angstrom (Å). While conceptually similar, the process required to obtain the structure of a protein by X-ray diffraction is in practice more complicated than the visualization of objects using the light microscope.

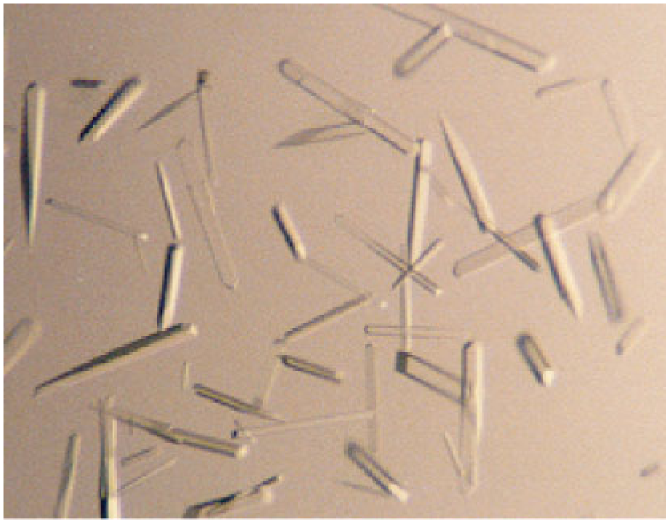


Figure 1 Examples of protein crystals. The longest dimension of these crystals is <0.5 mm. Individual crystals can be used in diffraction experiments. Data can be collected either at room temperature, but are often also collected at 100 K to reduce damage to the crystals by the X-ray radiation

The first hurdle to be overcome is that the scattering of X rays from a single protein molecule is not great enough to allow detection, and so, in order to increase the intensity of the diffraction obtained, the protein is crystallized (Figure 1) to generate repeating units of protein molecules, thus amplifying the signal. The scattered waves add up when they are in phase owing to constructive interference whilst waves that are out of phase cancel each other out owing to destructive interference (Figure 2). Detection of the resulting waves reveals an array of spots with measurable intensities (Figure 3) where the spacing of the spots is dependent upon the size of the repeating unit within the crystal and the intensities are a consequence of the arrangement of electrons and hence the atomic positions in the protein molecule.

The maximum resolution of a protein diffraction pattern, measured in angstroms, is ultimately dependent upon the degree of order within the crystal and controls the amount of atomic detail that can be obtained. As the order within the crystal increases, so does the maximum angle of diffraction, with the spots at greater angles providing higher-resolution information. An increase in resolution results in objects with smaller separations being resolvable. Thus, as resolution increases, the resolution distance, in angstroms, decreases. A perfectly ordered crystal would result in measurable diffraction intensities to the maximum resolution possible, which is equal to the incident X-ray wavelength divided by 2. Proteins, however, are flexible molecules, and thus do not form perfectly ordered crystals. At higher resolutions, the diffracted waves no longer interfere constructively, thus resulting in a loss of intensity, restricting the amount of detail obtainable from the structure. If a crystal diffracts to only a maximum resolution of 6 Å (0.6 nm), helices are visible, whereas β-sheets cannot be identified; at about 2.5 Å (0.25 nm), the side

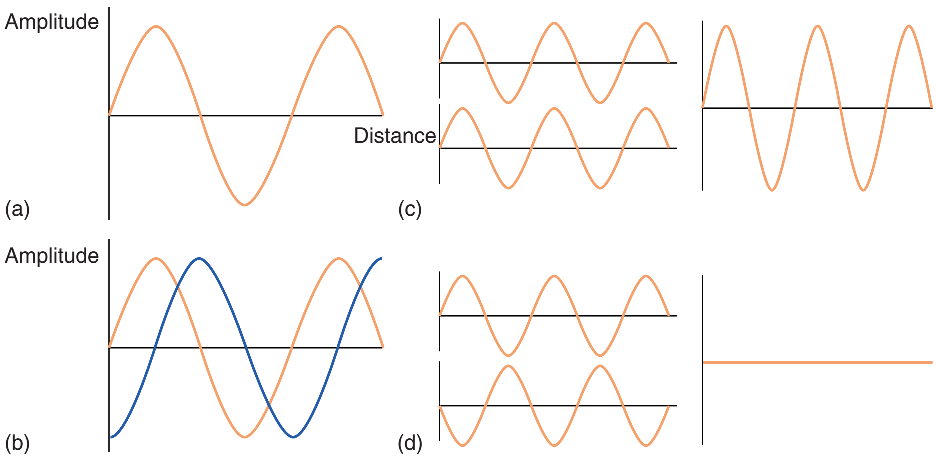


Figure 2 (a) A sine wave. The sine wave can be described in terms of its amplitude *A*, that is, its maximum displacement from zero, and wavelength λ–the distance between adjacent peaks, that is, one whole waveform. (b) Two sine waves. The two waves shown have identical wavelengths and amplitudes. They are not, however, placed identically with respect to the origin, which indicates that the waves are not in phase. The blue wave has one-fourth of a wavelength phase delay from the red one. (c) Constructive interference. The two waves on the left have the same amplitude, wavelength, and phase. If these two waves are combined, the resulting wave will be like that shown on the right: the phase and wavelength will remain the same, while the amplitude will double. This is constructive interference. (d) Destructive interference. The two waves illustrated on the left are one-half of a wavelength out of phase. Combination of these two waveforms will result in the wave on the right, which has an amplitude of zero. This is destructive interference

chains are identifiable, and towards 1 Å (0.1 nm), even hydrogen atoms begin to be visible.

Once diffracting crystals have been obtained, a data set can be measured, whereby a series of diffraction images are collected from the crystal. The crystal is rotated to allow data collection from different orientations so that diffraction intensities will give information about the whole protein molecule. In the case of the light microscope, a lens is used to refocus the diffracted rays. There is, however, no lens for X rays and the reconstruction of the protein structure must be obtained computationally. Unfortunately, the observed diffraction intensities only provide part of the required information. In order to determine the structure of the crystallized protein, both the amplitude and relative phase of the diffracted waves must be known (see Figure 2). The amplitudes of the spots can be calculated from the intensities, since the intensity of a spot is proportional to the square of the amplitude. Identification of the relative phase of each spot, however, is not so straightforward.

There are conceptually three methods that can be used to estimate the phases: one depends upon determining the phases by trial and error, the second estimates the phases experimentally, that is, by the collection of further diffraction data, and the third applies phases from a known structure thought to be similar to the target structure.

The first of these methods – the direct method – is used in small-molecule crystallography, but requires data at resolutions better than 1.2 Å (0.12 nm), and preferably less than 1000 atoms in the structure and therefore cannot be applied to

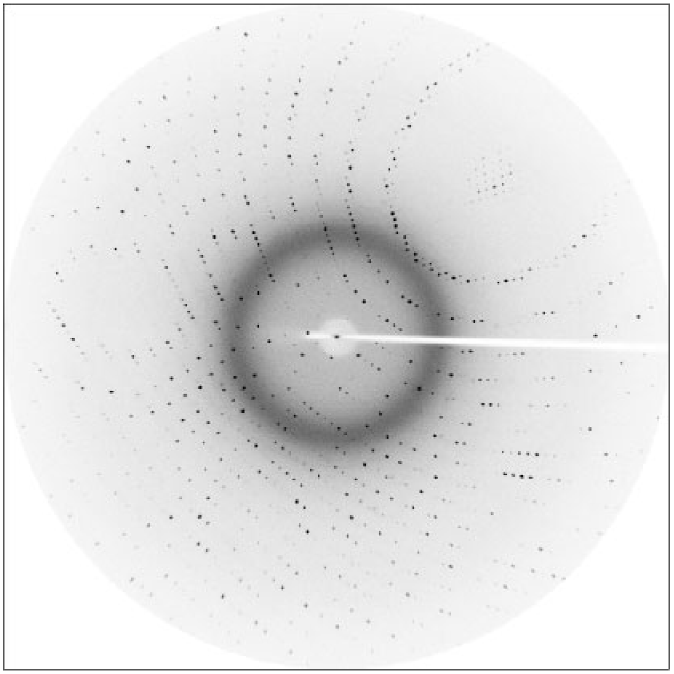


Figure 3 An example of a diffraction image. The individual spots are the result of constructive interference of the X-rays as they are diffracted by the molecules in the crystal. Many such images are required in order to determine the structure of a protein

the majority of protein structures. The second is more commonly used in protein crystallography and depends upon collecting a data set from either an identical or virtually identical crystal. In methods such as multiple isomorphous replacement (MIR), heavy atoms, for example, mercury, are soaked into the crystals, allowing the atoms to bind at specific sites within the crystals. The presence of such heavy atoms, with high electron numbers, perturbs the diffraction intensities. The differences between corresponding intensities in different data sets can then be calculated, enabling estimation of the possible values of the phase angles. A similar approach, multiple wavelength anomalous dispersion (MAD) makes use of anomalous scatterers within the crystal, either native or incorporated, which scatter differently depending on the wavelength of the incident beam. Anomalous scattering results from the absorption and reemission of the X rays by the electrons in the crystal. In light atoms, that is, those in the higher rows of the periodic table, the electron energies are too high to allow excitation of electrons to occur. Thus, the anomalous scattering effect of such light atoms is insignificant. With atoms heavier than sulfur, providing that X rays of a suitable wavelength are selected, the X rays may be absorbed by an electron and then subsequently emitted with a phase shift from the incident beam.

Data collection can be optimized, depending on the particular anomalous scatterers within the crystal, to obtain the maximum difference in scattering behavior

during successive rounds. Again, the differences between data sets can be calculated and used to find an estimation of the phase angles.

The third method, molecular replacement, depends upon the successful prediction that the target structure is similar to that of a known structure. Proteins with high sequence homology usually have a similar protein structure, and if the structure of such a homologous protein is already known, the model can often be fitted to the collected diffraction data and the phases estimated. The success of this method largely depends upon the identification and availability of a structurally similar protein. The method is particularly useful when data is collected from the same protein molecule with a different packing arrangement, or from highly homologous proteins – perhaps from two different organisms.

Following the successful measurement of reflection intensities and estimation of phases, an electron density map can be calculated and examined. The protein molecule can be built into the electron density map (Figure 4), usually with the help of the sequence of the protein. Refinement is then carried out, a process that aims to optimize the positions of the atoms within the model such that they fit with the observed diffraction intensities and estimated phases.

Thus the basic requirements for the determination of a protein structure by crystallography are as follows: a sufficient quantity of pure protein to set up crystallization trials, preferably in the milligram range; diffraction quality crystals; and a method to determine the phases (Scheme 1). Each of these processes can be problematic. The expression of sufficient quantities of protein to allow a high level of purification such that many crystallization experiments can be set up is a fundamental requirement and often the first hurdle to be overcome. Advances in cloning, expression, and purification techniques, along with the application of high throughput methods are fundamental to the pursuit of a structural-genomics approach, but are covered elsewhere (*see* Article 97, **Structural genomics – expanding protein**

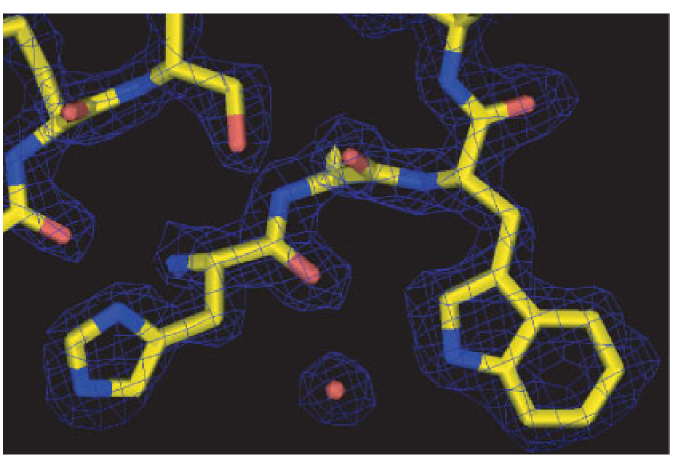
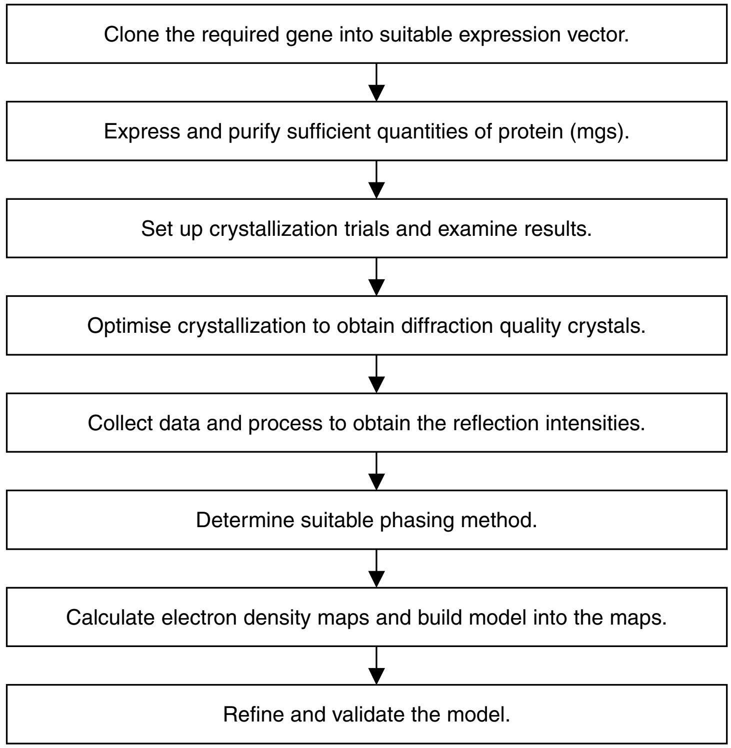


Figure 4 A stick model of the protein is built into the electron density maps, represented by the mesh-like lines. Here a good fit can be seen of the residues within the map; the histidine and tryptophan side chains are clearly identifiable. A spherical water molecule can be seen toward the bottom of the figure, modeled by the small red sphere



Scheme 1 The basic stages of structure determination by X-ray crystallography. At most points, problems may be encountered that force a return to an earlier stage in the procedure

structural universe, Volume 6). At the same time, there is progress in minimizing the amount of protein required. The development of automated crystallization robots has greatly improved the potential for crystallization of a protein. The technology is now available to allow crystallization on a nanoliter scale, thus allowing crystallization experiments with approximately 1000 times less than a classical crystallization trial. The crystallization of proteins is a trial and error procedure, often requiring hundreds of individual crystallization experiments to be set up. With nanoliter scale drops, many more individual experiments can be set up, allowing screening of a much greater range of conditions, thereby increasing the chance of obtaining crystals. Once promising conditions have been identified, they can be optimized, and, if required, scaled up to the microliter range. Robots allow this to be carried out far more rapidly, efficiently, and reproducibly than by hand.

After obtaining crystals, diffraction experiments are required, to determine if the crystals diffract sufficiently or if further optimization of the crystallization is required. If the crystals diffract well enough, then data collection can proceed and a strategy for determining the phases must be established. Since one of the aims of structural genomics is to determine the structure of proteins with novel folds, it is unlikely that molecular replacement, the use of a known structure to estimate phases, will serve as the main phasing tool. With the increase in computational

power, however, molecular replacement routines can be carried out rapidly and increasingly automatically, which raises the potential to search with a large number of structures, even if homology is low. Any high-scoring solutions can then be examined to see if they are likely to provide good phases. Traditional methods – such as the soaking in heavy metals – can be used, but are not so amenable to automation. The use of anomalous scatters is perhaps the most promising automated phasing method. Currently, the majority of proteins requires the incorporation of an anomalous scatterer into the protein during expression, for example, with the replacement of methionine residues with selenomethionine residues. Research is currently underway to enable the phasing of proteins using the anomalous signal from the native protein – that is, from sulfur atoms in side chains, or from phosphate ions. Such methods are ideal for structural genomics, since no additional gene manipulation, soaking experiments, or growth of cells on expensive and toxic selenium media are required, fitting with the high throughput maxim.

A range of programs exist to assist with the building and refinement of the model, once all of these goals have been achieved. Increasingly, with high-resolution and high-quality data, building and refinement can also be largely automated but in many cases this is not the case and the process requires many hours spent working at the computer. There is also the problem that the number of parameters to be fitted generally exceeds the number of observations – that is, those derived from the measured intensities. In order to obtain the best model, therefore, standard geometrical restraints are normally applied during refinement, incorporating realistic values for bond lengths and angles. This enables sensible refinement of the model to be carried out with the limited amount of experimental observations available.

As with any experiment, a form of validation should be carried out. The low ratio of observations to the number of parameters to be fitted can give rise to over fitting of the model – where the model appears to fit the data well, but in fact is incorrect. In order to reduce over fitting, a random subset of reflections is excluded from refinement to act as a control. If the model is built correctly, it should still fit with the excluded subset. A second method to validate the structure is to compare it with other structures: bond lengths and angles are known from small-molecule structures, and the torsion angles found within proteins are mostly found to have a limited range of values as found in the Ramachandran plot. Consistency with such accepted ranges is a good indication that the structure is good.

The determination of a structure, however, may only be the start of the battle in the structural-genomics world. Whilst the function of some proteins can be predicted by sequence analysis, in other cases the function is unknown. Correctly predicting the function of a protein on the basis of its structure is far from straightforward, although some methods are available and they are improving as we learn more about the structure–function relationships of proteins. In some cases, small molecules are copurified and crystallized with the protein of interest, perhaps giving insight into the function. In other cases, the fold of the protein places it in a known class of proteins not predictable from the sequence level. Otherwise, analysis of the structure is carried out in order to identify large clefts that may act as active sites, or particular clusters of residues that could engender functionality

to the protein. The relationship between structure and function is examined more deeply in elsewhere in the encyclopedia.

X-ray crystallography thus holds great potential to assist the characterization of proteins identified by genomics projects. Of course, its more traditional role of elucidating proteins of known function will continue to be fundamental in the understanding of protein function, both in and out of the large-scale genomics projects. Methodology is constantly improving in all aspects of crystallography, from the protein expression level, to data collection facilities, to the validation of structures. It is hoped that, as our knowledge and experience increase, our ability to extract information from structures will also improve, enabling us to gain a full understanding of the proteins within us.

Reference

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H and Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181** (4610), 662–666.

NMR

C. Jeremy Craven
The University of Sheffield, Sheffield, UK

1. Introduction

Solution state Nuclear Magnetic Resonance (NMR) provides an alternative technique to X-ray crystallography (*see* Article 104, **X-ray crystallography**, Volume 6) for the determination of protein structures. NMR has so far proved much less amenable to automation than X-ray crystallography, but it exhibits a number of features that make it highly complementary to the diffraction technique for the determination of structures of small proteins in a high-throughput manner for structural proteomics projects.

2. Basic technique

The signal detected in an NMR experiment arises from the interaction between the nuclei of atoms in the molecule and a very strong applied magnetic field (Figure 1). There is a wide range of possible protein NMR experiments, whose full description is beyond the scope of this article. The rich information content of such spectra arises mainly from three phenomena. First, the frequencies of the signals observed depend upon the detailed electronic environment of the nuclei, which in turn is sensitive to the conformation of the protein molecule. Therefore, information can be obtained as to whether the protein is folded or unfolded, or whether a conformational change has occurred. Unfortunately, the dependence of the chemical shift upon conformation is so complex in proteins that it has not proved possible to obtain extensive information for structure *determination* from this property. (Chemical shift *changes* that occur upon binding of a ligand are, in contrast, particularly informative in probing interactions of proteins with macromolecular or small molecule ligands.) The second phenomenon is that the frequency of the NMR signal from a particular nucleus can be sensitive to the spin state (i.e., parallel or antiparallel to the magnetic field) of neighboring nuclei to which it is chemically bonded. This phenomenon, known as *scalar coupling*, is particularly useful during the procedure of assigning NMR signals to particular nuclei in the molecule, since it enables signals to be linked together into bonded networks that can be matched to the covalent structure of the molecule. Third, nuclei can influence each other owing to the interaction of the magnetic moment

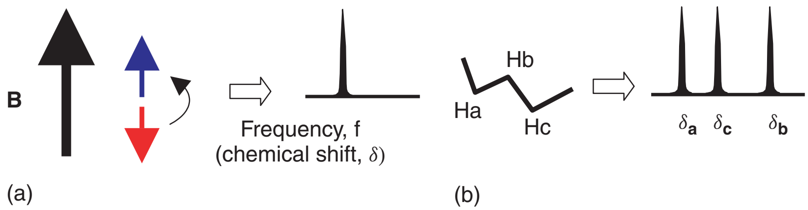


Figure 1 The basic phenomenon of NMR. (a) In a very strong applied magnetic field B, spin 1/2 nuclei (in proteins: ^{1}H, ^{15}N, ^{13}C) align parallel (blue spin) or antiparallel (red spin) to the applied field. The essence of the simplest one-dimensional NMR experiment is to perturb the distribution of spins between these two energy levels by applying a radiofrequency magnetic field. Radiofrequency electromagnetic radiation is emitted as the spins make transitions between the two energy levels and return to thermal equilibrium. The detected radiation can be represented as a peak at a particular frequency, or *chemical* shift, in a one-dimensional NMR spectrum. (b) The chemical shifts of the signals depend upon the precise strength of the magnetic field experienced by a nucleus, which depends, in turn, upon the extent to which it is shielded by its surrounding electrons. In a molecule containing three chemically distinct hydrogen atoms, three signals would generally be observed in the ^{1}H NMR spectrum

of one nucleus with the small magnetic field generated by the magnetic moment of another nucleus that is nearby in space. In solution state NMR, this affects the *relaxation* rates of nuclei, for instance, the rate at which nuclei return from an excited state antiparallel to the magnetic field to the ground state parallel to the magnetic field. This through-space (as opposed to through-bond) process also gives rise to the nuclear Overhauser effect (NOE), which provides the primary information for structure determination by NMR, since it indicates which nuclei are close in space due to the 3D fold of the protein. The NOE, which is fundamentally a relaxation phenomenon, is the change in intensity of an NMR peak due to the perturbation of the populations of the energy levels of a nearby nucleus.

3. Sample requirements

Protein molecules comprise such a large number of atoms that a simple, one-dimensional spectrum of the signals arising from all the hydrogen nuclei is too complex to interpret. The most efficient method to alleviate such an overlap is to incorporate the stable isotopes ^{15}N and/or ^{13}C into the protein to allow the use of multidimensional methods (see below). Such incorporation is straightforward to achieve in overexpressed proteins, by limiting the nitrogen or carbon sources to labeled compounds such as ^{15}N ammonium salts and ^{13}C glucose. Labeling with ^{15}N is sufficiently inexpensive and useful to be routine at the early stages of a project. Labeling with ^{13}C is an order of magnitude more expensive, and is generally only justified once a protein has been selected for structure determination.

A typical sample for structure determination by solution state NMR consists of a solution of 300–500 µL of 0.2–2 mM protein dissolved in water. This corresponds to ca. 1–20 mg of a 20-kDa protein. The protein must be soluble, robust, and pure enough to remain in such solution for a period of at least 1 week, and preferably

longer, although in cases of rapid degradation, multiple samples can be used. As for X-ray crystallography, the protein must be well folded, and if there exists any posttranslational modification, then such modification should be uniform within the sample.

4. Size limitations

There are two factors that limit the size of proteins that can be studied by NMR. The first is that the number of signals in an NMR spectrum is proportional to the size of the protein; overlap of these signals can make the spectrum too complex to interpret, even with isotope labeling. The second factor is that the rate of Brownian molecular tumbling in solution has a large effect on the sensitivity of NMR spectra. In general, larger proteins tumble more slowly than smaller proteins, which, for reasons beyond the scope of this article, leads to a reduction in sensitivity of NMR spectra. These factors dictate that full structure determination is relatively straightforward for well-behaved proteins less than about 180 a.a. (20 kDa).

The tumbling rate of proteins is sensitive not only to the molecular weight (and, to a lesser extent, the molecular shape) but also to the aggregation state. For instance, a protein that is dimeric will tumble with the characteristics of a protein of twice the molecular weight. In the case of physiological "true" dimers, this is unavoidable. The most problematic cases can be where nonspecific aggregation at the high concentrations required for NMR gives rise to spectra of a much poorer quality than expected for the monomeric molecular weight. It is not uncommon that proteins are simply insoluble at these high concentrations, or precipitate over a timescale of a few days.

5. Sample screening by ^{15}N HSQC

The detailed topic of multidimensional NMR is beyond the scope of this article; however, one particular spectrum, the ^{15}N HSQC spectrum, is worthy of detailed mention since it is the major tool for protein characterization at the early stages of many structural proteomics programs by both NMR and crystallography. An example of such a spectrum is shown in Figure 2.

In the screening stage of a structural proteomics program, this spectrum would indicate several very important things about the behavior of this protein. First, it shows that the protein is folded since the signals are well spread out, or *dispersed*, in the hydrogen dimension. Such dispersion arises from the fact that in a well-folded protein, the different amide groups along the protein backbone experience distinct environments. In an unfolded or molten globular protein, the environments will, on average, be much more similar to each other, and the signals would be observed to cluster much more tightly around the center of the range. Second, the correct number of signals are observed for the number of residues in the protein, indicating that structured parts of the protein are not undergoing slow conformational or intermolecular interactions that would either attenuate some signals or cause several peaks to be observed for one residue. The signals are also relatively sharp, which

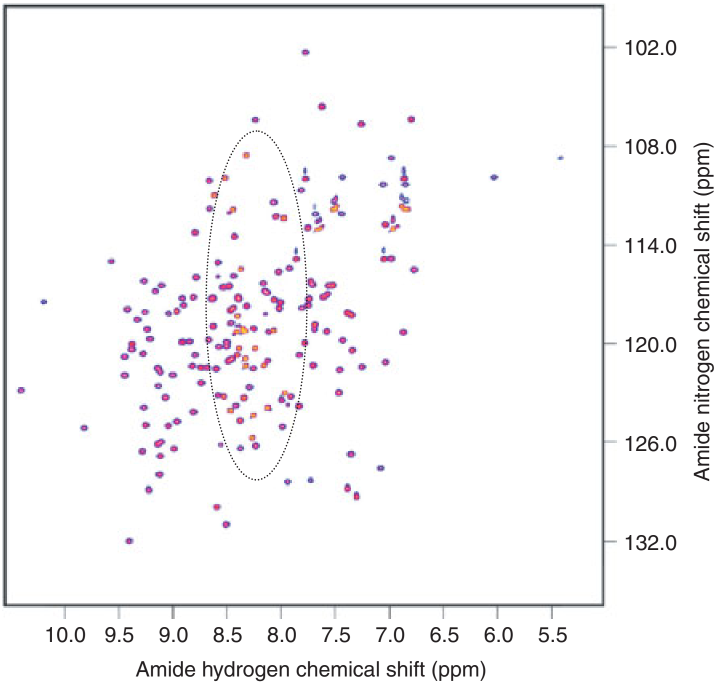


Figure 2 ^{15}N HSQC spectrum of the 19-kDa protein TCTP from *Schizosaccharomyces pombe*. A ^{15}N HSQC spectrum is a two-dimensional spectrum in which the frequencies of directly bonded ^{1}H–^{15}N pairs are simultaneously measured. One signal, or *crosspeak*, is observed for each NH group in the protein. Since the position of each crosspeak is determined by the frequency of the signal from both the ^{1}H and ^{15}N nuclei, the problem of overlap of the hydrogen frequencies is much reduced. The dotted line marks the area containing many intense and sharp peaks with ^{1}H shifts characteristic of unfolded proteins

indicates that the protein has suitable tumbling characteristics. Finally, the spectrum contains about 20 very sharp and intense signals clustered around the centre of the hydrogen dimension. Such signals are characteristic of a portion of the protein being unfolded. This is not a particular problem for an NMR structure determination, provided it does not have a severe effect on slowing the molecular tumbling. It would, however, often be problematic for crystallization of such a protein, in which case it would be necessary to identify the mobile residues involved and make constructs that eliminate these residues. Where alternative constructs are made, NMR provides an efficient method to test whether the construct indeed removes the mobile residues, and also whether the fold of the protein is retained. Note that HSQC screening is feasible for proteins up to approximately twice the size for which full structure determination is possible.

In cases where a poor HSQC spectrum is obtained, it is possible to screen alternative buffer and temperature conditions. Alternatively, the screening of orthologous proteins has been shown to increase the proportion of proteins for which a satisfactory HSQC spectrum can be obtained (Savchenko *et al*., 2003).

6. Structure determination by NMR

Once a ^{15}N HSQC spectrum of adequate quality has been obtained, the procedure for determining a 3D structural model of a protein is fairly straightforward, and can be obtained on average in about 2–6 months of a single person's effort. The first stage is to assign the different NMR signals observed to particular nuclei in the molecule, a process known as resonance assignment. This is followed by the derivation of distance restraints from NOESY spectra (Figure 3). Given enough such distance restraints, well-established computational techniques allow the derivation of a 3D model of the protein. Typically, the structure is calculated using molecular dynamics (Figure 4), in which extra energy terms are included to implement the distance restraints.

One particular problem in obtaining structural information from NOESY spectra is that overlap of chemical shifts can make it difficult to establish unambiguously which pair of atoms gives rise to a particular crosspeak. Therefore, the normal procedure is first to calculate a low-precision structure, using distance restraints only from crosspeaks that are readily assigned. Ambiguous crosspeaks can then be reexamined, and assignments incompatible with this preliminary model can be eliminated. By thus incorporating additional restraints, the precision of the structure can be iteratively improved. Some progress has been made in semiautomating this process by software packages such as ARIA; however, it remains the single greatest impediment to more rapid structure determination by NMR. While distance restraint data remains very much the primary source of NMR structural data, the structure calculation can be supplemented with torsional angle data from chemical shifts or scalar couplings, and with more global restraints from residual dipolar couplings or paramagnetic ion effects.

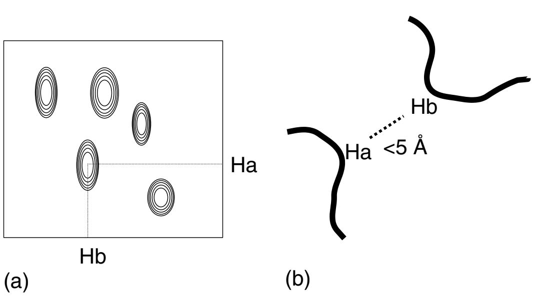


Figure 3 NOESY spectrum and distance restraints. The crosspeaks in a NOESY spectrum (a) occur for pairs of ^{1}H nuclei that are closer in space than about 5 Å. The simplest NOESY spectrum is a two-dimensional spectrum in which the position of crosspeaks is determined by the chemical shifts of the two nuclei involved. The observation of a crosspeak linking the frequencies of two nuclei Ha and Hb indicates that these two nuclei are close in space (b). If the two nuclei are far apart in sequence in the protein, this provides a powerful *distance restraint* from which the 3D structure can ultimately be calculated. Even with multidimensional techniques, such spectra are very complex and the extraction of the data from these spectra, therefore, remains a very challenging problem, which requires significant human input

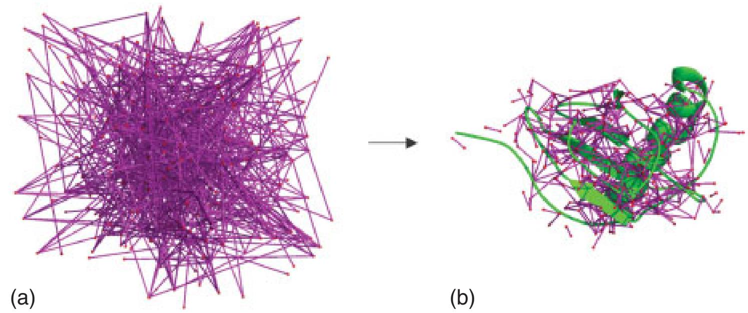


Figure 4 Structure determination by NMR. In one implementation of a molecular dynamics–based structure determination by NMR, the atoms of the protein are initially placed randomly (a). The distance restraints (represented by the magenta lines in (a)) can be visualized as taut "elastic bands" between the atoms. The atoms are then allowed to move under the influence of these restraints, and also subject to bond lengths and bond angles dictated by the covalent structure of the protein. The atoms move to satisfy the restraints to create a 3D model of the protein (b). By starting from different sets of random starting coordinates, an ensemble of structures can be determined, all of which satisfy the distance restraints, from which an estimate of the precision of the model can be made. For clarity, intraresidual restraints have been omitted, and only atoms to which restraints have been determined are shown individually. The protein shown is the 11-kDa cysteine proteinase inhibitor, human cystatin A

7. Complementarity to X-ray crystallography

In highly automated proteomics projects, one might expect X-ray crystallography to dominate owing to its much more automated structure determination methods. One of the advantages of NMR over X-ray crystallography is that, once purified recombinant protein is obtained, screening for suitability for further study can be achieved very rapidly by the ^{15}N HSQC method discussed above. Furthermore, this early screening is very strongly diagnostic: if the screening result is favorable, then a structure can be more or less guaranteed within a well-defined timescale. This contrasts with the much greater uncertainty in the timescale for the production of crystals that diffract well. Thus, it has been argued (Savchenko *et al*., 2003) that NMR may represent the most efficient strategy for structural proteomics of small proteins.

This conjecture appears to be supported by the distribution of structures that have been deposited so far (Figure 5). It, therefore, appears that the ability to rapidly screen and select targets for NMR enables it to remain a competitive technique for the structure determination of small proteins, even in the context of highly automated proteomics programs. In addition, NMR also provides an excellent route for screening small protein targets for *both* NMR and X-ray crystallography in order to optimize the constructs that proceed through to the structure determination stage, therefore greatly increasing the number of successfully studied small proteins by either technique.

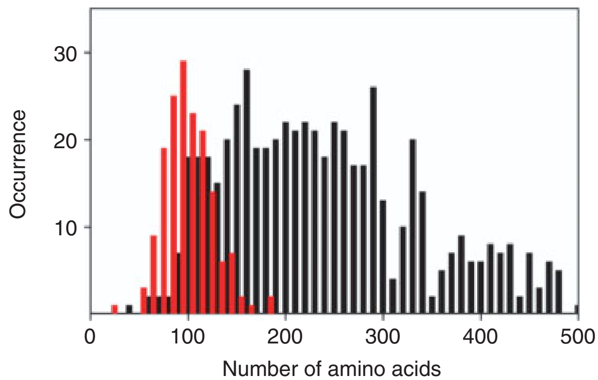


Figure 5 Distribution of protein structure determinations in structural proteomics programs, determined by NMR (red) and X-ray crystallography (black). The data was extracted from **targetdb.pbd.org**. These distributions are affected by the scale of resources dedicated to the two techniques, and also by the relative difficulty of crystallizing small proteins, especially if they are domains of larger proteins. It is markedly different from the distribution of structures deposited in the PDB as a whole, where, for proteins of size 50–120 a.a., there are approximately twice as many entries for structures determined by X-ray compared to those determined by NMR (**http://oca.ebi.ac.uk/oca-bin/ocamain**)

Further reading

Cavanagh J, Fairbrother WJ, Palmer AG and Skelton NJ (1996) *Protein NMR Spectroscopy: Principles and practice*, Academic Press: San Diego.

Neuhaus D and Williamson MP (2000) *The Nuclear Overhauser Effect in Structural and Conformational Analysis*, Second Edition, Wiley-VCH: New York.

Rattle H (1995) *An NMR Primer for Life Scientists*, Partnership Press: Hants.

Reid DG (Ed.) (1997) *Protein NMR Techniques*, Human Press: New Jersey.

Reference

Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, *et al*. (2003) Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. *Proteins: Structure, Function and Genetics*, **50**, 392–399.

Electron microscopy as a tool for 3D structure determination in molecular structural biology

Neil A. Ranson
University of Leeds, Leeds, UK

1. Introduction

In recent years, electron microscopy (EM) has been elevated from a tool for studying ultrastructure and the *morphology* of large protein assemblies, into a tool for studying protein *structure*. These advances have been made possible by the introduction of novel methodologies, and by dramatic improvements in instrumentation and computing resources available to researchers.

2. EM image formation

With the small apparent wavelength of high-energy electrons (~0.025 Å at 200 kV), a modern electron microscope is potentially an excellent tool for protein structure determination. However, many physical and optical considerations limit the resolution that can be obtained in practice.

First, electrons are focused in an EM as they move through a magnetic field, rather than a traditional lens. Such "magnetic lenses" are imperfect, and these imperfections mean that the focused beam must be passed through a small aperture; a small (~40 μm) hole in a metal foil, which excludes highly scattered electrons and prevents their contributing to the image. The use of a small aperture, just as in photography, limits the attainable resolution (to ~1 Å in this case). It also gives a large depth of field, meaning that all objects in the image are at approximately the same level of focus, and this has profound consequences for the interpretation of EM data that are discussed below. Secondly and more importantly, the EM image recorded is a combination of elastic scattering events (which generate information from the specimen), and inelastic scattering events (which damage the specimen and generate noise). Typical inelastic collisions between electrons and the specimen are actually less damaging than for X rays, but the scattering cross section for electrons is vastly greater. Radiation damage is therefore of primary importance in EM of biological specimens. Such damage is a fundamental limit on the type of biological

specimens that can be examined, as they must be thin enough to prevent multiple scattering events (typically less than 2000 Å). Radiation damage also limits the attainable resolution as limiting dose results in noisy images that are difficult to analyze.

Thin, biological specimens typically scatter electrons very weakly. This means that, somewhat counterintuitively, when the microscope is precisely in focus, very little can be seen as there is almost no contrast. Phase contrast is introduced through a combination of imperfections in the magnetic lenses used, and by the user defocusing the system. In such conditions however, the image observed is modulated by a contrast transfer function (CTF). The CTF is an oscillating function of increasing frequency and decreasing amplitude, which describes the transfer of information within the microscope as a function of resolution. An EM image therefore has alternating bands of positive and negative contrast as resolution increases, and the positions of these bands are dependent on the defocus at which the image was recorded. In practice, this has the effect of grossly distorting the image. Structural information can only be restored by a computational correction of the CTF in the observed, digitized images, and by combining information from images taken at different levels of defocus. This last point is critically important: as the CTF oscillates between positive and negative contrast; at the point where the contrast flips, there is no information at all. The resolutions at which these reversals occur are dependent on defocus, and so combining data at a range of defocus ensures that there is information at all resolutions. The CTF also describes the falloff in information as resolution increases, that is, there is very little signal at the high resolutions needed to reconstruct protein secondary structure. This phenomenon is fundamental to electron microscopy but its effects are ameliorated by the use of field emission gun (FEG) microscopes. For a review of this subject, see Hawkes and Valdrè, 1990 and Saibil, 2000.

3. Basic imaging methods

Isolated protein complexes have traditionally been imaged by using heavy metal staining on continuous carbon (or other) support films. Such an approach has compelling advantages as the stain coats (and to some extent penetrates) the biological molecule, and dramatically enhances both sample stability in the electron beam and the contrast that is attainable in the resulting images. However, a number of profound disadvantages exist. The preparation method presents the specimen with a series of extremely nonnative environments, typically resulting in a somewhat dehydrated specimen with an ~30% flattening against the support film. In addition, the nature of the stains used means that it is actually the stain shell, rather than the biomolecule itself, that is imaged. Finally, the granularity of such stains, and their lack of penetration into biomolecules, ultimately limits the meaningful information obtained to > ~20-Å resolution. Despite these limitations, the low-resolution information obtainable from staining methods has made, and continues to make, an enormous contribution to biology that cannot be underestimated (see for example Burgess *et al*., 2003).

A new EM method for studying biomolecules was pioneered in the late 1980s by Dubochet and colleagues at EMBL, in which specimens were stabilized by

low temperatures (Dubochet *et al*., 1988). Such "Cryo-EM" methods involve placing a solution of the biomolecule in question on to an EM grid bearing a holey carbon support film. Excess liquid is removed by blotting, leaving a thin aqueous film suspended in the holes. This is then cooled *extremely* rapidly by plunging into liquid ethane cooled to near–liquid nitrogen temperature. The enormous thermal conductivity of liquid ethane, and the thinness of the aqueous film (typically <2000 Å) allows vitrification of the solution into a glasslike state, and the biomolecule is trapped in a frozen-hydrated state. Such vitrified specimens have compelling advantages for structural biologists. First, the native structure of the specimen is preserved within the vitreous ice layer, even in the vacuum of the microscope column. Second, observing the specimen at such low temperatures significantly slows radiation damage, and while such damage still fundamentally limits the dose that may be used, usable images of unstained specimens can be recorded. Although such unstained images have extremely low contrast, the native structure of the specimen is preserved and internal structural details of the object recorded. The attainable resolution thus becomes limited only by the homogeneity of the specimen, and the quality of images recorded.

4. Data analysis

Two considerations dominate the analysis of any EM data with the aim of solving the three-dimensional structure of biological macromolecules. First, the large depth of field discussed earlier, together with the penetrating nature of high-energy electrons results in the electron-scattering potential of the object being projected onto the 2D plane of the image. The images obtained are therefore *projection images*, analogous to the familiar images from medical X rays. The 3D structure of an object can be obtained from a set of such projection images only if different views of the object, and their relative orientations, are available. Second, in order to limit radiation damage to acceptable levels, the dose used to record images is limited. This results in the observed images having an extremely poor signal-to-noise (S/N) ratio, and necessitates the use of averaging techniques to combine data from different images and improve S/N. The way in which these considerations are handled depends in large part upon the geometry of the specimen under study.

2D arrays of membrane proteins are particularly well suited to study by EM. Crystalline arrays by definition have a large number of individual molecules oriented by the crystal lattice, which satisfy the requirements for averaging, and mean that the biomolecules are identically oriented (or at least nearly so). In such studies however, different views of the crystalline material are achieved by tilting the specimen inside the microscope, and the maximum tilt angle that can be achieved is $\sim 70^\circ$. This results in a missing wedge of information in Fourier space, and corresponding nonisotropic resolution in the resulting model (with resolution generally poorer perpendicular to the plane of the membrane). Nevertheless, such methods have now resulted in density maps for a number of different membrane proteins in which the majority of the polypeptide chain can be unambiguously assigned. Indeed, the first atomic resolution structure of a membrane protein was determined by EM, in a study that exploited the fact that bacteriorhodopsin occurs in naturally crystalline patches of membrane (a number of studies, culminating

in Henderson *et al*., 1990). Subsequently, a number of artificially crystalline molecules have been determined to similar resolutions, including a plant light harvesting complex (Kuhlbrandt *et al*., 1994), and the cytoskeletal protein tubulin (Nogales *et al*., 1998).

Helical assemblies are also well suited to study using EM. In a helical assembly, a range of angular views of the asymmetric unit is displayed as the helical axis is traversed. Tilting is therefore not usually required for 3D reconstruction. However, order within the helix is generally lower (i.e., variable pitch and twist) than in 2D crystals, and the number of asymmetric units imaged is generally far lower. Despite this, recent work on an exceptionally well-ordered bacterial flagellum has yielded atomic resolution information using helical processing (Yonekura *et al*., 2003). Helical techniques are also extremely powerful when membrane proteins form (or can be made to form) tubular crystals. In such a case, the structure of the nicotinic acetylcholine receptor has been determined to ∼4.5 Å (Miyazawa *et al*., 2003).

It is perhaps in the study of true single particles using cryo-EM that the most visible progress has been made in recent years. The term *single particle* is used to cover any structures that are not part of a higher order assembly. Single particles range from highly symmetric structures such as icosahedral viruses (although some purists insist that icosahedral structures are not true single particles owing to their high symmetry), to objects with no intrinsic symmetry such as the ribosome. Single particle approaches generally rely on the intrinsic randomness of the frozen-hydrated specimen to give randomly oriented biomolecules. A full sampling of molecular views is therefore possible without specimen tilting. Many different approaches in analyzing such data are possible, but in essence they all share the common goals of determining the orientation of the images with respect to each other, and averaging of molecular views in the same orientation to improve S/N (for a comprehensive description, see Frank, 1996; Van Heel *et al*., 2000).

Single particles in solution, with their lack of externally imposed order, are inherently challenging structural projects, and to date no true single particle has been solved with sufficient resolution to unambiguously assign a polypeptide chain. A great many projects are currently in the 6–10 Å range. The most conspicuous progress has been seen in an explosive increase in our understanding of ribosome function (for the review, see Frank, 2003).

5. Synergy with other structural methods

The intermediate (6–12 Å) resolutions generally achieved with EM methods yield models of biomolecules that contain vast amounts of information, but are extremely difficult to interpret. The availability of atomic resolution coordinates of components of an EM map, or of related structures, can dramatically increase this interpretability. The fitting of atomic coordinates into intermediate resolution EM density maps is now routinely used to produce "pseudo-atomic" resolution models that greatly enhance the interpretability of EM data. For example, fitting of the crystal structure of the molecular chaperone GroEL in its unliganded state, into EM density for a GroEL:ATP complex, allowed a detailed description of ATP-induced conformational change (Ranson *et al*., 2001). This description was sufficiently

accurate to allow design of specific point mutations affecting the cooperativity of ATP binding.

6. Current state-of-the-art and future prospects

Spectacular progress has been made in recent years, and will continue into the future. We are, however, approaching a point in time where instrument quality, processing methods, and computing power will be sufficient to give resolutions that approach the homogeneity of the samples that can be prepared for microscopy. EM techniques determine solution structures, where the ensemble of conformations present is specimen-dependent. The attainable resolution will therefore vary from atomic resolution for exceptionally rigid or well-ordered specimens, to perhaps 50 Å or poorer for more dynamic or flexible ones. In addition, the relationship between resolution and the quantity of data required to attain that resolution is decidedly nonlinear: a 10-Å map might require 10^4 images, while a 6-Å map might need 10^5. Especially given the obvious synergy with X-ray crystallography and NMR, perhaps the most powerful application of EM is in studying conformational change in large assemblies, rather than as a tool for atomic resolution structure determination.

References

Burgess SA, Walker ML, Sakakibara H, Knight PJ and Oiwa K (2003) The structure of dynein-c by negative stain electron microscopy. *Nature*, **421**, 715–718.

Dubochet J, Adrian M, Chang JJ, Homo JC, Lepault J, McDowall AW and Schultz P (1988) Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, **21**, 129–228.

Frank J (1996) *Three-dimensional Electron Microscopy of Macromolecular Assemblies*, Academic Press.

Frank J (2003) Towards an understanding of the structural basis of translation. *Genome Biology*, **4**, 237.

Hawkes PW and Valdrè U (1990) *Biophysical Electron Microscopy*, Academic Press: London.

Henderson R, Baldwin JM, Ceska TA, Zemlin F, Beckmann E and Downing KH (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology*, **213**, 899–929.

Kuhlbrandt W, Wang DN and Fujiyoshi Y (1994) Atomic model of plant light-harvesting complex by electron crystallography. *Nature*, **367**, 614–621.

Miyazawa A, Fujiyoshi Y and Unwin N (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature*, **423**, 949–955.

Nogales E, Wolf SG and Downing KH (1998) Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature*, **391**, 199–203.

Ranson NA, Farr GW, Roseman AM, Gowen B, Fenton WA, Horwich AL and Saibil HR (2001) ATP bound states of GroEL captured by Cryo-Electron microscopy. *Cell*, **107**, 869–879.

Saibil HR (2000) Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica. Section D*, **56**, 1215–1222.

Van Heel M, Gowen B, Matadeen R, Orlova EV, Finn R, Pape T, Cohen D, Stark H, Schmidt R, Schatz M, *et al*. (2000) Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, **33**, 307–369.

Yonekura K, Maki-Yonekura S and Namba K (2003) Complete atomic model of the bacterial flagellum by electron cryo-microscopy. *Nature*, **424**, 643–650.

Introductory Review

Integrative approaches to biology in the twenty-first century

Marvin Cassman
Consultant, San Francisco, CA, USA

Systems or integrative biology has become wildly popular in just the past few years. (The two terms will be used interchangeably to refer to the quantitative analysis of biological networks, at whatever level of organization.) The reason is clear – the volume of experimental data generated by genomics and other high-throughput techniques pose both a problem and an opportunity. The problem is that the mass of data overwhelms the ability to intuitively understand it. The opportunity is that the data may be able to be manipulated in a manner that allows for an unprecedented understanding of the integrated behavior of biological systems. The mechanistic studies that have generated an amazing ability to decipher the molecular workings of individual biological molecules can now be linked to computational approaches, which will illuminate the functions of tightly coupled and highly regulated biological systems. At the end, the processes underlying health and disease will be understood in greater depth than has previously been possible.

Although there are frequent comments that there is no good definition of systems biology (Henry, 2003), this is somewhat of an overstatement. It is generally agreed that the availability of detailed information from the genome, proteome, metabolome, ad nauseome, has provided an opportunity to understand biological phenotype in a way that has not been previously possible. Many of the initial, and current, attempts to predict phenotype focus on the prevailing paradigm of the past 40 years – single gene defects. This extraordinarily powerful approach has been the major contributor to an understanding of the function of individual genes and proteins. It seems less likely that it can yield an understanding of complex biological behavior, from cellular activities such as motility to the operation and integration of organ systems. Indeed, the excitement surrounding systems biology derives from the underlying assumption that phenotype is governed by the behavior of networks, rather than simply the consequence of individual gene action. It is also assumed that to understand the behavior of complex (or even simple) networks, computational approaches will be required. It is worth a short digression to see how this variant fits into the spectrum of what is broadly referred to as “computational biology”.

Computational biology is not new and is not even one thing. Computational chemistry is what most people have meant when they refer to computational

biology. It is a long-established and well-understood area that involves computation on the molecular level, such as protein folding, molecular dynamics, and small-molecule docking. Another more recent form of computational biology is computational genomics, the offshoot of the genome programs, and the most common variant of bioinformatics. This is a discipline that is of more recent vintage than computational chemistry and is now very well appreciated. It involves storage, retrieval, and analysis of information from large data sets. Perhaps the oldest variant of computational biology is computational physiology, which can be traced back at least as far as the Hodgkin–Huxley equation in 1947. This form of computational biology involves the development of quantitative models that describe the behavior of physiological systems. The newest version is computational cell biology, which primarily involves the analysis of cellular networks through modeling and simulation, with targets such as signal transduction networks, cell motility, and cis-regulatory controls on the gene. (Let me say that I realize that there are other variants of computational biology that I have not included, such as epidemiology, population genetics, and computational neuroscience. My apologies to these and others.)

The heart of integrative/systems biology is the use of computational tools for modeling and simulation when applied to cellular and organ systems. Coupled to this are the tools that are required to construct the networks, which include the approaches of computational genomics and proteomics.

Modeling and simulation have a long and honored history in most of the sciences, but not in much of modern biology, and certainly not in modern cell biology. The reasons for this are complex, but at least in part, derive from moving from the concept of a cell as a fuzzy and ultimately unknowable unit of infinite complexity to the cell as defined by finite and understandable molecular elements governed by the laws of chemistry. In the process, the idea that cellular phenotypes were emergent properties that result from complex interactions was largely discarded. In its place came the purely mechanistic concept of the cell as little more than an assemblage of loosely linked molecular devices. On the whole, this was a change decidedly for the better. “Emergent properties”, after all, sounds much like a new-age mantra that should be associated with bland music and pyramids. More importantly, the rigorous application of chemical principles has allowed an understanding of biological specificity and its consequence, the ability to design specific interventions. However, the mass of information that has been generated in recent years has forced a reconsideration of the need to apply computational models to biological systems. Nowhere has this been more apparent than in the analysis of cellular regulatory networks. Decades of research have led to the kinds of connectivity diagrams shown in Figure 1. These networks predominate in living cells and particularly in eukaryotic cells, which are differentiated from their simpler brethren, not so much by the pathways of metabolism, which tend to be relatively similar, but by the increase in regulatory controls. Within eukaryotes, humans are overrepresented in gene functions related to signaling networks, both intra and extracellular. It is clear that the input–output relationships and the effects on phenotype of variation in their components, for networks such as that shown in Figure 1, cannot be understood intuitively. Computational modeling and subsequent experimental verification of in silico–predicted behavior are required to understand the behavior of such systems.

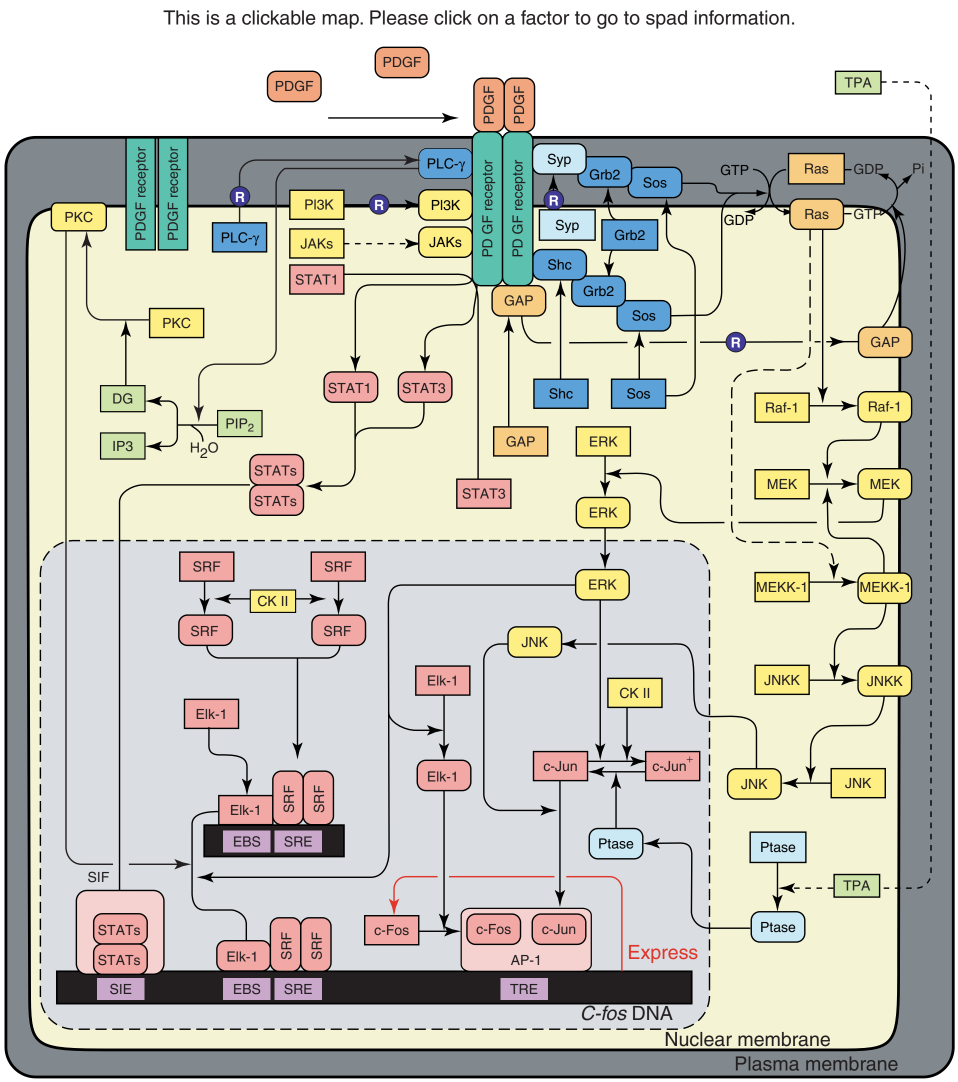


Figure 1 Signaling pathway mediated by PDGF (Reprinted with permission from Satoru Kuhara, Kyushu University, Japan)

How can we begin to deal with networks of such complexity, much less extend them into higher levels of organization such as organs and organ systems? As always, it is useful to begin with a simplifying hypothesis. (Although data-driven research may yield information, it still requires a hypothesis to generate knowledge.) This is an extension of the hypothesis stated earlier: Phenotype is governed by modular systems of complex networks whose output is highly resistant to change. This hypothesis brings into focus two important and somewhat controversial concepts – modularity and robustness. What can we say about these, and why are they important?

Biological modules have been defined in a number of ways. One is purely operational – the identification of groups of genes that tend to respond in a joint manner, for example, through temporally coordinated gene expression (Segal *et al*., 2003). Another is the identification of discrete groups of interconnected elements that are abstracted from the topology of the network (Wuchty *et al*., 2003). Yet a third functional definition is a set of genes and gene products that perform some task nearly autonomously (von Dassow *et al*., 2000). One example of the latter is the set of proteins involved in bacterial chemotaxis, which has been modeled by Leibler, Bray, and others (Barkai and Leibler, 1997; Morton-Firth *et al*., 1999; Yi *et al*., 2000).

The necessity for identification of modules is simple – the cell is too complex to be modeled as a whole. Of course, this is a problem for humans, but not necessarily one in which nature is obliged to accommodate us. Luckily, it appears that modular networks as well as network motifs are common in nature and may even be a natural consequence of evolution (Hartwell *et al*., 1999). However, a major problem remains – the identification of modules and their boundaries. In this effort, genomic, topological, and functional approaches will all be needed.

A second issue is that of robustness. One definition of robustness was given by Waddington (1942) as "canalization" – the constancy of phenotype in the face of genetic and environmental perturbation. On a molecular level, it can be defined as the insensitivity of network output, that is phenotype, to variation in system parameters. Evidence for robustness has been found in the analysis of a number of networks (Barkai and Leibler, 1997; von Dassow *et al*., 2000; Lee *et al*., 2003). Further, the evolution of complex networks has been suggested by Siegal and Bergman (2002) as the driving force for "canalization". Nevertheless, the question remains – robust in regard to what? For bacterial chemotaxis, it is in the regulation of a switch that rotates the bacterial flagellum in one direction or another (Barkai and Leibler, 1997). In *Drosophila* development, it is moving from one developmental state to another (von Dassow *et al*., 2000). However, not all phenotypic characteristics are equally insensitive to parameter change. For example, in bacterial chemotaxis, the rate constant for the flagellar switch is very sensitive to parameter variation (Barkai and Leibler, 1997). This raises the obvious question as to how and why some systems might evolve to a robust phenotype while others are appropriately more plastic and responsive to the environment (Marder and Prinz, 2002; Cassman, 2003).

Apart from regulatory networks, metabolic pathways are pervasive and highly conserved across organisms. The analysis of these networks is sometimes referred to as "metabolomics". Modeling of metabolic pathways has a long history, dating to the development of metabolic control analysis (MCA) in the 1970s (Kacser and Burns, 1973; Heinrich *et al*., 1997). More recently, flux-balance analysis (FBA), which assumes a steady state and a known stoichiometry of reaction, has been developed and applied on a genome-wide scale (Famili *et al*., 2003) and has been able to make in silico predictions that are consistent with some in vivo observations. Robustness can be observed in these approaches as well, since Edwards and Palsson have been able to show that the *E. coli* metabolic network is insensitive to large changes in a set of critical gene products (Edwards and Palsson, 2000). Modularity in metabolic networks has also been demonstrated by Barkai and her colleagues,

who have shown that in the yeast *S. cervisiae*, examination of gene coexpression showed a hierarchical organization of metabolic pathways into groups of varying expression coherence (Ihmels *et al*., 2004). A similar nested hierarchy of modules was also found from topological analysis (Ravasz *et al*., 2002), although there were significant differences between the two hierarchical trees.

As we move up in complexity, modeling of physiological functions such as action potentials in the heart, calcium flux, and the operation of neuronal circuits have been attempted at various levels of integration. Modeling the heart has linked cell studies with detailed anatomical models to produce some remarkable results (Kohl *et al*., 2000), including the mechanism of arrhythmogenic effects of sodium channel mutations (Noble, 2002). Several systems have been used to generate computational models of calcium flux (Sherman *et al*., 2002). Here again the various transport and cellular regulatory mechanisms have been treated as independent modules and then combined to establish a cellular process with specific phenotypic properties, in particular the oscillations that are characteristic of cellular calcium concentrations.

Robustness can be also be observed at these more complex levels of organization, albeit through a somewhat different mechanistic process than we have identified thus far. In neurons, conductances and activity patterns must remain stable in the face of ongoing processes of degradation and synthesis of ion channels. Models have shown that synapses do not have to be perfectly tuned to generate acceptable physiological outputs and that different initial synaptic strengths will result in similar network dynamics (Marder and Prinz, 2002).

At this point in the development of biological modeling and simulation, what is most interesting are the questions. What constitutes a module and the existence of and mechanism(s) underlying robustness are only two of these? What detailed information is minimally necessary to model cellular systems? Are there characteristic design elements in biological networks, and do the global properties of these networks fit some common structure, that is, scale-free (Barabasi and Albert, 1999)? What are the system characteristics that allow plasticity and which convey stability and how are these balanced? These and other questions will be addressed in the next few decades and their answers will enhance our knowledge and improve our ability to manipulate biological systems for the benefit of humanity.

References

Barabasi A-L and Albert R (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barkai N and Leibler S (1997) Robustness in simple biochemical networks. *Nature*, **387**, 913–917.

Cassman M (2003) Computational biology: Counting on the neuron. *Science*, **300**, 756–757.

Edwards JS and Palsson BO (2000) Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology Progress*, **16**, 927–939.

Famili I, Forster J, Nielson J and Palsson BO (2003) *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13134–13139.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Heinrich R, Rapoport SM and Rapoport TA (1997) Metabolic regulation and mathematical models. *Progress in Biophysics and Molecular Biology*, **32**, 1–82.

Henry CM (2003) Systems biology. *Chemical and Engineering News*, **81**, 45–55.

Ihmels J, Levy R and Barkai N (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nature Biotechnology*, **22**, 86–92.

Kacser H and Burns JA (1973) The control of flux. *Symposia of the Society for Experimental Biology*, **27**, 65–104.

Kohl P, Noble D, Winslow RL and Hunter P (2000) Computational modeling of biological systems: tools and visions. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences*, **358**, 576–610.

Lee E, Salic A, Kruger R, Heinrich R and Kirschner MW (2003) The roles of APC and Axin derived from experimental and theoretical analysis of the WNT pathway. *PLOS Biology*, **1**, 116–132.

Marder E and Prinz AA (2002) Modeling stability in neuron and network function: the role of activity in homeostasis. *Bioessays*, **24**, 1145–1154.

Morton-Firth CJ, Shimuzu TS and Bray D (1999) A free-energy-based stochastic simulation of the tar receptor complex. *Journal of Molecular Biology*, **286**, 1059–1074.

Noble D (2002) The rise of computational biology. *Nature Reviews*, **3**, 460–463.

Ravasz E, Somers AL, Mongru DA, Oltvai ZN and Barabasi A-L (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.

Sherman A, Li Y-X and Keizer J (2002) Whole cell models. In *Computational Cell Biology*, Fall C, Marland E, Wagner J and Tyson J (Eds.), Springer-Verlag: New York, pp. 101–139.

Siegal ML and Bergman A (2002) Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10528–10532.

von Dassow G, Meir E, Munro EM and Odell GM (2000) The segment polarity network is a robust developmental module. *Nature*, **406**, 188–192.

Waddington CH (1942) Canalization of development and the inheritance of acquired characteristics. *Nature*, **150**, 563–565.

Wuchty S, Oltvai ZN and Barabasi A-L (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, **35**, 176–179.

Yi TM, Huang Y, Simon MI and Doyle J (2000) Robust perfect adaptation through integral feedback control. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4649–4653.

Specialist Review

Functional networks in mammalian cells

Yuguang Xiong and Ravi Iyengar
Mount Sinai School of Medicine, Department of Pharmacology and Biological Chemistry, NY, USA

1. Introduction

The mammalian cell is a complex system composed of many types of components ranging from simple inorganic molecules to complex macromolecular assemblies that form organelles. The complexity of mammalian cell arises from the large number of molecular components contained in the cell, the complex interactions between these cellular components, and the anisotropic spatial distribution of the components within the cell. The identity of the mammalian cell is defined by its genetic makeup. Gene expression programs allow for the transcription and translation of specified genes such that the expression of key proteins defines the phenotypic behavior of the cell. However, both the genetic code and the protein products are subject to changes: genetic codes can be altered by mutation caused by normal physiological or pathological conditions and protein molecules can be modified during their synthesis, for example, posttranslation lipid modifications steps, or after their synthesis, for example, biochemical reactions with other molecules such as phosphorylation. As a consequence, the system is composed of a very large number of molecules with different functional properties. For example, the total number of protein molecules in a liver cell has been estimated to be about 7.9×10^9; so, if there are about 10 000 different gene products in a liver cell, each protein can have on average nearly one million molecules (Lodish *et al*., 2000). Understanding how these molecules are distributed, regulated, and how they interact with one another is a challenge.

Advances in biotechnology, such as rapid sequencing of entire genomes, and in information technology, such as large-scale web-based databases, have led to the discovery of a substantial number of previously unknown genes and their protein products in viruses, bacteria, archaea, and eukaryotes. Many large databases have been constructed to store and analyze genetic sequences and protein sequences and structures, such as the nucleotide database GenBank of National Center of Biotechnology Information (NCBI online, 2004), the protein database SwissProt at Geneva (SwissProt online, 2004), and Stanford Human Genome Center (SHGC online, 2004).

In addition to the large number of cellular components, interactions between the components occur through a variety of mechanisms. These interactions often directly represent the functions of components within the cell system. For example, gene transcription is controlled by different transcriptional regulators that initiate, terminate, inhibit, to regulate transcriptional processes by interacting with each other and with specific genetic sequences. The interactions between components integrate distinct cellular components into a central regulatory signal network as well as functional machines within the cell. Interactions between the regulatory network and the machines result in the observed systems behavior (Figure 1). A defining aspect of cellular systems is its regulatory topology based on component interactions. The regulatory topology of the cell represents the dynamic structure of the system and determines system behavior, therefore directly reflecting the functionality of cell. Because of its importance, regulatory topology continues to be the target of intense research. Such research usually focuses on identifying the regulatory subcelllular maps of specific or common cellular signaling processes, such as G protein-coupled signaling pathways (Neves *et al*., 2002). The microarray technology developed in recent years (Spellman *et al*., 1998; Iyer *et al*., 1999) has been successfully applied to the measurements of large-scale genetic regulatory networks (Davidson, 2002), and even protein interaction networks (Nielsen *et al*., 2003). Various computational approaches have been developed to analyze the structure of these complex biological networks, such as the clustering analysis to identify coexpressed genes in genetic regulatory networks (Eisen *et al*., 1998), the statistical analysis to study the connectivity of metabolic biochemical networks (Ravasz, 2002) and probabilistic models to infer regulatory cellular networks from disparate biological data (Friedman, 2004) and identification of regulatory motifs (Milo *et al*., 2002).

A significant challenge in understanding intracellular systems is how a complex cellular system containing thousands of components, in response to internal or external signals, performs the desired physiological functions accurately in spite of occasional component malfunction or noise from adjacent processes. In other words, what are the mechanisms by which individual cellular components are integrated into a coherent and functional system? Intuitively, a straightforward approach to this challenge is to define an exhaustive anatomy of the cellular system by listing every component and every interaction. This is a good starting point, but it is not sufficient. It is like trying to assemble a complex mechanical machine from a number of individual parts. Although we know exactly what function each part performs and which other parts it can be assembled with, we do not know how to put all the parts together to build a machine that is able to perform its designed functions accurately, effectively, and robustly. What we need in order to build a working machine from its parts and interactions list is the design principles of the machine. More explicitly, design principles describe how to build the machine with required functionalities from the functions of the individual parts and the interactions between those parts. For example, which parts should be assembled together? What is the state of each part to generate specific required functions? How is each part controlled? How do we protect key parts? How do we resist systems perturbations during function? What is the working range of machine within which each part functions in an error-free manner? How

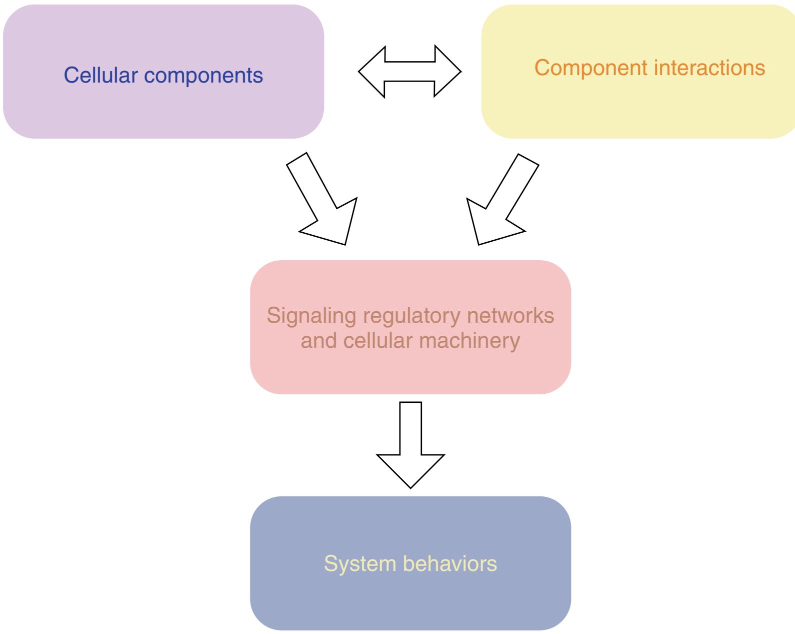


Figure 1 A schematic representation of a bottom-up approach to explain how systems level behavior is generated from components and interactions. The cell integrates cellular components and interactions into a functional systems by the means of regulatory signaling networks and cellular machines. Regulation of the functions of the cellular machines leads to the observed systems behavior

do we recover from error? A well-defined design diagram would provide answers to these questions. In understanding the design principles of cells, we can take advantage of natural selection processes. The cell has been going through billions of years of evolution during which only those cells with optimized systems and functions survived. To satisfy rigorous requirements for survival, it is necessary to have functionality coupled with accuracy and robustness. One way to approach this issue is to assume that survivors must follow similar design principles as human-made machines. Compared with engineering, systems biological research is a reverse process (*see* Article 110, **Reverse engineering gene regulatory networks**, Volume 6). Nature has designed a complex cell system, and the task of systems biologist is to study and understand how this system works and thus identify the overall the design principles. Because the cell is a complex but coherent functional system, the approaches to study this system have to be "systems aware", which means that we must analyze the cell from a systems perspective that can guide the design of experiments to explain the seemingly contradictory behavior of subsystems, search for hidden components, and predict unknown cellular behavior (*see* Article 107, **Integrative approaches to biology in the twenty-first century**, Volume 6). Such approaches are a major part of the emerging field of systems biology. This review covers two aspects of cell systems biology, dynamic machines, and design principles underlying the regulation of these machines.

2. Dynamic machines

The phenotypic behaviors of cells arise from the function of subcellular machines as the vesicle release machinery or the electrical response machinery of excitable cells that are located in a spatially restricted manner. Such machines exhibit two key features: tight spatially restricted control of these machines and dynamic stability. Local machine functions result from the dynamic control of cellular machines by the central signaling network to perform time- and space-dependent functions. Dynamic stability allows the machine to maintain its functionality within the context of the function of multiple cellular machines operating simultaneously.

It is essential that, within the cell, a variety of cellular processes are scheduled to perform required functions at appropriate times. Although the cell contains a large number of components, each cellular process uses only a fraction of total components, which means that only the components necessary for current process are active while others are quiescent or engaged in other functions. For signaling networks, different signaling pathways are activated at different times (Jordan *et al*., 2000), and for genetic networks, different gene expressions are turned on at different times (Davidson *et al*., 2002). These dynamic behaviors enable the cell to assemble and disassemble local machinery to perform the functions required for a specific process. This feature distinguishes cellular machinery from human-engineered systems whose structure cannot be changed on demand. There are two known mechanisms by which these dynamic features can be achieved. One mechanism is the transient formation of local cellular domains solely by the interactions between machinery components (e.g., the spatial patterns formed by signaling reactions) (Eldar *et al*., 2002). The other mechanism is the localization of machinery components at cellular compartments (e.g., plasma membrane, cytoplasm, and nucleus) (Weng *et al*., 1999) and cellular scaffolds (e.g., actin cytoskeleton) (Medalia *et al*., 2002).

Once assembled, the local machinery performs its designated function for the specific cellular process. Such functions result from the components and their interactions determined by the machine's regulatory topology. These local machineries can form higher-level cellular organelles such as the rough endoplasmic reticulum that contains both ribosomes and endoplasmic reticulum–bound posttranslational modifying enzymes required for the production of fully functional proteins. Therefore, the cell system can be always considered as a set of machines consisting of multiple components forming various levels of organization. The machines perform their functions by setting the components at preferred states through component interactions. These preferred states of components define the state of the whole machinery, and the state change of any components can affect the machinery state. For example, the fast growth of dendritic actin filaments is characterized by the high activities of machine components that regulate three processes: actin polymerization, actin depolymerization, and actin recycling. If the activity of any the components that regulates these processes changes, the overall growth rate will change (Pollard *et al*., 2000). Therefore, the dynamic behavior of cell system is governed by the state of the components that make up the cellular machines. Cells display both steady state and periodic dynamics. Steady dynamics is a status in which cellular machines operate at steady states, maintained by a constant level

of activity with the central regulatory signal networks. Steady state dynamics may be observed both in cells functioning at basal levels as well as in cells that have undergone activity-induced state changes such as neurons that display the long-term potentiation and long-term depression of synaptic responses (Sheng and Kim, 2002). Cellular machines can also operate at multiple steady states. For example, the mitogen-activated protein kinase (MAPK) signaling pathway that connects extracellular growth factor signals with nuclear transcriptional machinery exhibits bistable states dependent on the signaling stimulation by growth factors, and interestingly, this bistable behavior can be altered to monostable behavior by other signals (Bhalla *et al*., 2002). Periodic dynamics is perhaps one of the most important behaviors of biological systems such as the population model of predator and prey in ecology and the electric oscillation in neuronal and cardiac cells. Particularly, periodic dynamics exists in the biochemical signaling networks that regulate a variety of cellular processes such as glucose metabolism, calcium oscillation, pulsatile hormone signaling, and gene transcription (Goldbeter, 1996). A very common biological periodicity is the circadian phenomenon in which biological systems can adapt themselves to the 24-h periodic variations of environment (Dunlap, 1999). An early detailed molecular model for circadian rhythms was based on a set of five biochemical kinetic equations that described the oscillations of the period protein of Drosophila (Goldbeter, 1995). However, the periodic oscillations of cellular machinery can go out of control and become chaotic when oscillation dynamics is highly sensitive to the initial condition of the system (Goldbeter *et al*., 2001). Owing to the complex nature of periodic phenomena, computational approaches are valuable tools to analyze such systems (Goldbeter, 2002).

3. Design principles

With the large number of components and interactions, combined with the dynamics of local machinery, cellular system can exhibit many uncharacterized complex dynamic behaviors besides the steady and periodic dynamics described above. Since such complexity greatly complicates the in vivo observations of cellular dynamics, it becomes increasingly difficult to explain these observations through intuitive speculation or brute-force exhaustive mathematical modeling. Instead, systems biologists can try to deal with such complexity by applying the concepts of system control from engineering theory (Zhou *et al*., 1995).

3.1. Modular regulation

Despite its complexity, the cell is a biochemical machine constructed from molecules, so it must follow all the physiochemical rules of nature. In order to generate the required biological functions and maintain life, cellular machines need to manage the tremendously large number of components and their interactions in an efficient way. This efficiency requires cellular machines to organize their components into distinct modules, each of which is responsible for specific functions. To control these modules, a well-designed intermodular regulatory system is

present (Hartwell *et al*., 1999). Because modular regulation heavily determines the functionality of cellular machinery, many efforts, both experimentally and computationally, have been made to search for characteristic regulatory motifs that can explain various observations of the system. For example, autoregulatory motifs generate quick response to control input (Ren *et al*., 2000) and increase stabilility (McAdams and Arkin, 1997), chain regulatory motifs generate temporal sequential control (Simon *et al*., 2001), negative feedback motifs generate steady state dynamics (Buchman, 2002) and oscillatory dynamics (Goldbeter, 2002), positive feedback motifs generate bistability dynamics (Bhalla *et al*., 2002; Ferrell, 2002; Hasty *et al*., 2002), positive feedforward motifs generate quick response, large amplification, and high sensitivity to input signals (Goldbeter and Koshland, 1984; Lee *et al*., 2002; Shen-Orr *et al*., 2002), and multi-input motifs generate logical gate dynamics (Hasty *et al*., 2002). The design of these regulatory motifs for cellular machinery control share great similarities with engineering systems and highlight the physical origins of cellular systems.

3.2. Design of robust systems

Cells live in a world full of disturbances coming from environmental variation, intracellular and extracellular noise, and machinery malfunction. Systems within the cell must be able to maintain all the normal functionalities during these disturbances for the cell to survive, that is, the cell should display robust behavior. To achieve such robustness, cellular machines are designed with the following principles: (1) a relative intrinsic stability determined by the internal insensitivity of components themselves and their interactions to disturbance. Therefore, stability does not depend on machinery structure; (2) a dispersed regulatory structure that is determined by modular regulation. Both the physical and functional modularity of cellular machinery can prevent the malfunction or failure in one part of machinery from spreading to other parts and causing system-wide failures. Optimized modular regulatory structure can effectively stabilize the behavior of cellular machines against internal and external disturbances. For example, internal feedback regulation in the biochemical networks of bacterial chemotaxis makes the precise adaptation of bacteria to attractive chemical signals in environment very robust to the large variation of molecular concentrations and rate constants (Barkai and Leibler, 1997). Another example is the DNA robustness against damage by a feedback checking mechanism mediated by p53 protein. If DNA damage occurs, damage can be sensed by DNA-dependent protein kinases that can activate p53 protein. Active p53 protein then transactivates p21, which results in G1 arrest in cell cycle. This state is released when DNA is repaired. Thus, propagation of mutations due to DNA damage is limited. The design of cellular machines is a major mechanism for cellular machinery to attenuate intracellular noise to achieve various cellular functions (Rao *et al*., 2002); and (3) machinery redundancy, which introduces additional backups for important machinery components such that the failure of one component can be compensated by its substitution with another. Redundancy exists in both gene level and protein level. For example, at gene level, only about 256 out of 468 protein-coding genes are considered necessary and sufficient to sustain

the existence of *Mycoplasma genitalium* bacterium (Mushegian and Koonin, 1996) and, at the protein level, chemokine proteins are redundant in both their production and their function (Mantovani, 1999).

Unless individual cell systems are intrinsically stable in terms of their components and component interactions, system robustness is often achieved by including extra machines that can serve the same overall function. Thus, redundancies could be at the level of regulatory motifs or in actual machine itself. In either case, the overall goal is to offset the influence of disturbance on cellular systems and maintain functional stability within limits. When cell systems contain a large number of components and perform various functions, substantial increases in machinery structure is required to maintain the robustness of the large system. Thus, the maintenance of system robustness relies on an essential feature of cellular systems: system complexity. Cell systems are thought to be complex due to numerous cellular components, a large number of component interactions, and the emergent complex system behaviors (Weng *et al*., 1999). However, the complexity of a cell system can be concealed by its robust adaptation to a variety of internal and external disturbances until the emergence of unexpected changes and failures. This type of approach can be better understood by comparison with the design of complex engineered products such as the Boeing 777 airplane and the allometric scaling chart of optimal cruise speed (Csete and Doyle, 2002). The design principles of biological systems share great similarity with those of complex engineering systems in terms of two aspects of system design: the design of machinery function and the design of functional robustness. Both types of design can significantly increase the complexity of the system. This is illustrated by considering some examples. The design of complex biorhythmicity and its effects on the behavior of metabolic and genetic networks (Goldbeter *et al*., 2001), or the systems with extreme robustness, such as robust control of airplane flight (Yeh, 1998), reveal a paradoxical systems property. System robustness achieved by the increased complexity of system machinery comes at the cost of system fragility, which represents unanticipated system vulnerability to malfunctions and disturbances. To achieve robustness, extra regulation of machinery modules are added to the system. Robust metabolic networks usually have scale-free type of connections between network nodes (Ravasz *et al*., 2002). This feature makes metabolic networks more robust to random node failure than randomly connected networks, but more fragile to system-wide failures when hub nodes fail (Buchman, 2002). Thus, system robustness and fragility are two intrinsically linked system properties originating from system complexity. Well-designed systems have to make deliberate choices to balance between robustness and fragility to achieve an optimal function (Carlson and Doyle, 2002; Csete and Doyle, 2002).

3.3. Initiation of cell spreading: A case study in signaling network regulation of the actin-based cell motility machine

Motility is an important physiological phenomenon (*see* Article 114, **A complex systems approach to understand how cells control their shape and make cell fate decisions**, Volume 6). During development, the movement and migration of

cells is very important for the proper development of the embryo. The ability to move in response to extracellular signals is a critical property of hematopoietic cells and the mounting of immune responses. Cell movement involves the actin cytoskeletal machinery as well as the tubulin-based cytoskeleton. Sheetz and colleagues have recently developed quantitative approaches to imaging the spreading of mammalian fibroblasts when they come in contact with extracellular matrix proteins such as fibronectin (Dubin-Thaler *et al*., 2004). We have started to develop a computational process to model this process in terms of components and interactions.

A simplified connections map that can be used to represent the regulated prohypbreakcess is shown in Figure 2. This system is made up of a regulatory signaling network and the actin cytoskeletal machinery. The regulatory signaling network senses extracellular motility signals, for example, fibronectin in this case, by integrin receptors on cell membrane and transduces the signal to the interface layer between the signaling machinery and the actin machinery. The interface layer consists of several actin-binding proteins capable of regulating the process of actin polymerization and depolymerization. Each protein in the interface layer receives signals from the respective pathway of upstream signaling pathway and in response to these signals regulates the corresponding aspect of downstream actin machinery. Then, the actin cytoskeletal machinery reorganizes the structure of the actin cytoskeleton during polymerization and depolymerization reactions, which results in growth of actin filaments close to the plasma membrane. This growth of filament pushes the plasma membrane, generating force required for the cell to spread.

The dynamics of this motility machinery is well characterized by the multiphase model observed for the fibroblast cell spreading in the experimental setup. This model describes the motility behavior of the fibroblast cell before, upon, and after it contacts the fibronectin substrate on glass slide. Three phases are observed: (1) the basal phase in which spherical fibroblast cell demonstrates random local membrane protrusion before it contacts fibronectin substrate; (2) the transitional phase in which fibroblast cell explores the spatial distribution of fibronectin substrate by a few trial contacts and then decides whether or not to spread; and (3) the spreading phase in which spherical fibroblast cell spreads over fibronectin substrate to form a flat disk. This multiphase motility of fibroblast cell is realized by the cooperation of the multiple aspects of the dynamics of actin cytoskeletal machinery, each of which is regulated by the corresponding signaling pathway of upstream signaling regulatory machinery. This modular design ensures the effective control of complex dynamics of actin machinery by extracellular signals and intracellular signaling machinery. Moreover, this motility machinery is designed to be both sensitive to extracellular signals as well as have an intrinsically robust performance capability. This complex behavior can be seen in the multiple phases involved in fibroblast spreading (Dubin-Thaler *et al*., 2004). The time needed for fibroblast cell to explore extracellular fibronectin in the transitional phase is dependent on the distribution concentration of fibronectin on glass slide: the larger this concentration is, the lesser is the time needed in this phase. This sensitivity enables fibroblast cell to move toward more concentrated extracellular signals. Thus, the trigger to move is directly controlled by extracellular signals. However, when fibroblast cell enters the spreading phase,

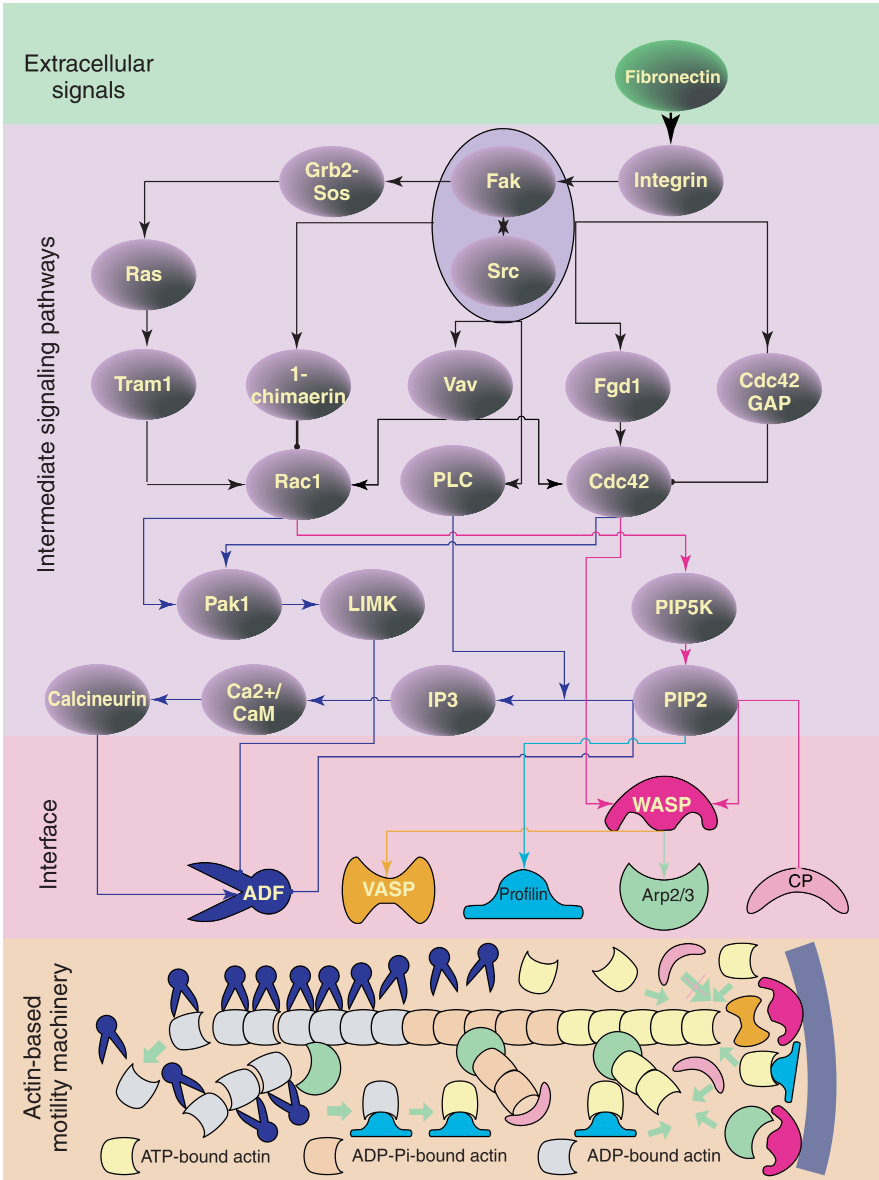


Figure 2 A cellular motility system consisting of the regulatory signaling network and the actin-based motility machine. The biochemical components of actin-based and integrin-regulated motility of fibroblast cell is shown as a connections map. The system can be thought to consist of four layers: (1) the extracellular signals, here the extracellular motility protein fibronectin that induce fibroblast cell motility is shown. There are other extracellular signals such as growth factors and chemotactic peptides that can affect this process; (2) the layer of intermediate signaling pathways that interact to form a network, that contains the integrin receptor, GTPases, tyrosine, and Serine-Threonine kinases; (3) the interface layer formed by several actin-binding proteins, which receive signals from the intermediate signaling pathways and in response to these signals regulate the downstream actin cytoskeletal machine; and (4) the layer of actin-based motility machinery, consisting of actin filaments and monomers that by polymerization, branching, and depolymerization generates cell motility. This is a simplified connection map in which only key components are shown

spreading rate is independent of fibronectin concentration. This behavior suggests that once spreading is initiated, the spreading of fibroblast cell mainly depends on intracellular signaling regulation and control of the actin motility machinery and is insensitive to extracellular signal. Thus, the system exhibits a robust functional response above a threshold level of an extracellular signal.

4. Conclusions

Although the cell is a complex system, its overall design favors functional modularity and robustness. Functional modularity is achieved by an array of cellular machines that can operate in a coordinated fashion. These machines are highly regulated and a central signaling network coordinates the regulation in response to extracellular and intracellular signals. In order to achieve overall robustness, the system has developed extensive redundancies at both the level of regulation as well as functional machines. The inclusion of these redundant subsystems results in a highly complex system. The high complexity of cellular regulatory networks and machinery unavoidably brings fragility into the system. Successful survival represents the ability to achieve balance between the robust and fragile properties of the system through functional modularity.

Acknowledgments

We thank Avi Mayaan for critically reading the manuscript. Research in the Iyengar laboratory is supported by NIH grants GM 54508 and CA 81050.

Further reading

Maayan A, Blitzer RD, and Iyengar R (2005) Towards predictive models of mammalian cells. *Annual Review of Biophysics and Biomolecular Structure* **34**, in press.

References

Barkai N and Leibler S (1997) Robustness in simple biochemical networks. *Nature*, **387**, 913–917.

Bhalla US, Ram PT and Iyengar R (2002) MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science*, **297**, 1018–1023.

Buchman TG (2002) The community of the self. *Nature*, **420**, 246–251.

Carlson JM and Doyle J (2002) Complexity and robustness. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2538–2545.

Csete ME and Doyle JC (2002) Reverse engineering of biological complexity. *Science*, **295**, 1664–1669.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, *et al*. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

Dubin-Thaler BJ, Giannone G, Dobereiner HG and Sheetz MP (2004) Nanometer analysis of cell spreading on matrix-coated surfaces reveals two distinct cell states and STEPs. *Biophysical Journal*, **86**, 1794–1806.

Dunlap JC (1999) Molecular bases for circadian clocks. *Cell*, **96**, 271–290.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Eldar A, Dorfman R, Weiss D, Ashe H, Shilo BZ and Barkai N (2002) Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. *Nature*, **419**, 304–308.

Ferrell JE Jr. (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, **14**, 140–148.

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Goldbeter A (1995) A model for circadian oscillations in the Drosophila period protein (PER). *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **261**, 319–324.

Goldbeter A (1996) *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*, Cambridge University Press: Cambridge, UK.

Goldbeter A (2002) Computational approaches to cellular rhythms. *Nature*, **420**, 238–245.

Goldbeter A, Gonze D, Houart G, Leloup JC, Halloy J and Dupont G (2001) From simple to complex oscillatory behavior in metabolic and genetic control networks. *Chaos*, **11**, 247–260.

Goldbeter A and Koshland DE Jr. (1984) Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. *Journal of Biological Chemistry*, **259**, 14441–14447.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Hasty J, McMillen D and Collins JJ (2002) Engineered gene circuits. *Nature*, **420**, 224–230.

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, *et al*. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Jordan JD, Landau EM and Iyengar R (2000) Signaling networks: the origins of cellular multitasking. *Cell*, **103**, 193–200.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Lodish H, Berk AS, Zipursky L, Matsudaira P, Baltimore D and Darnell J (2000) *Molecular Cell Biology*, Fourth Edition, W. H. Freeman and Company: New york.

Mantovani A (1999) The chemokine system: redundancy for robust outputs. *Immunology Today*, **20**, 254–257.

McAdams HH and Arkin A (1997) Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 814–819.

Medalia O, Weber I, Frangakis AS, Nicastro D, Gerisch G and Baumeister W (2002) Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science*, **298**, 1209–1213.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Mushegian AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10268–10273.

NCBI (2004) online: **http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide**

Neves SR, Ram PT and Iyengar R (2002) G protein pathways. *Science*, **296**, 1636–1639.

Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G and Sorger PK (2003) Profiling receptor tyrosine kinase activation by using Ab microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9330–9335.

Pollard TD, Blanchoin L and Mullins RD (2000) Molecular mechanisms controlling actin filament dynamics in nonmuscle cells. *Annual Review of Biophysics and Biomolecular Structure*, **29**, 545–576.

Rao CV, Wolf DM and Arkin AP (2002) Control, exploitation and tolerance of intracellular noise. *Nature*, **420**, 231–237.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Shen-Orr SS, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**, 64–68.

Sheng M and Kim MJ (2002) Postsynaptic signaling and plasticity mechanisms. *Science*, **298**, 776–780.

SHGC (2004) online: **http://www-shgc.stanford.edu**

Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, *et al*. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

SwissProt (2004) online: **http://www.expasy.org/cgi-bin/sprot-search-de**

Weng G, Bhalla US and Iyengar R (1999) Complexity in biological signaling systems. *Science*, **284**, 92–96.

Yeh YCB (1998) Design Consideration in Boeing 777 Fly-by-wire Computers, presented at *The Third IEEE High-Assurance Systems Engineering Symposium (HASE), at Washington, DC, 13-14 November, 1998*.

Zhou K, Doyle JC and Glover K (1995) *Robust and Optimal Control*, Prentice Hall: New Jersey.

Specialist Review

Analyzing and reconstructing gene regulatory networks

J. Jeremy Rice, Aaron Kershenbaum and Gustavo Stolovitzky

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

1. Introduction

The study of networks is becoming an increasingly important area of biological research. A fundamental goal in the postgenomic era is to understand cellular function in terms of complex behaviors emerging from simpler interactions at the biomolecular level. Networks offer a natural representation for aggregating biomolecular interactions for a wide range of tasks including protein–protein interaction maps, pathway diagrams and in silico based simulations. Current gene regulatory networks are reconstructed by combining data from many experiments and multiple labs using hypothesis-driven approaches, but there is a growing effort to infer biological networks from high-throughput and genome-wide data sets. This chapter discusses topological properties of gene regulatory networks, and the methods to reconstruct these networks from high-throughput data.

2. Principles in network design

Networks arise in diverse areas including telecommunications, electric grids, social interactions, and many areas of biology (for review, see Strogatz, 2001). The entire field of network analysis is beyond the scope of this work; however, some general network design principles can be considered when analyzing biological networks. We mention some the most pertinent examples here:

Shared resources: Sharing resources is often beneficial, especially in the presence of economies of scale.

Cascaded functions: Networks may be used to perform sequential sets of operations or processes that can be organized as modules that are small subnetworks.

Control and stability: Networks may exist to provide control mechanisms for an otherwise uncontrolled set of processes.

3. Defining network properties

Figure 1 shows several examples of networks. Figure 1(a) illustrates a specific topology known as the Erdös–Rényi (E-R) network in which edges are randomly placed between nodes with equal probability (Erdös and Rényi, 1959). In this example, the graph is mostly disconnected so that no path exists between most pairs of nodes. One of the earliest theorems in random graph theory from Erdös and Rényi relates the number of edges in an undirected graph to the connectedness. The theorem states that the probability of the network being connected rises from 0 to 1 over the interval $N\log N - cN < E < N\log N + cN$, where N is the number of nodes, E is the number of edges, and c is a constant. Hence, these networks make a relatively quick transition from disconnected to connected as the number of edges approaches $N\log N$. This example is a directed graph that is typical for a gene regulatory network in which nodes represent genes and edges represent genes influencing other genes. We can define an indegree (outdegree) as the number of edges going into (out of) each node. For example, the node with vertical crosshatches has an indegree of 1 and an outdegree of 1, whereas the node with horizontal crosshatches has an indegree of 0 and an outdegree of 1.

Several network properties are now defined. The network in Figure 1(a) is considered sparse because the number of edges is $O(N)$ rather than $O(N^2)$ (i.e., only 16 edges out of 484 possible connections). For a directed graph with N nodes, there are N^2 possible connections if a node can connect to any other node including itself. Sparseness has been observed for many biological networks (e.g., see Yeung *et al*., 2002); however, the E-R graph in Figure 1(a) differs from real gene regulatory networks in many other respects. The connectedness of a graph describes the degree to which paths exist between the nodes that comprise the network. Figure 1(b) shows several networks that are connected, but with different properties. Network 1, called a clique, is considered cohesive because all nodes connect directly with all others. Here undirected edges are shown, but the properties are similar for directed edges. Network 2 is also connected but is not as cohesive because only the center node connects directly with the others. Network 3 illustrates the scale-free property that is defined by the distribution of node degrees following an approximate power law relation (i.e., $P(k) \approx k^{-\gamma}$). Many biological networks are scale-free (Jeong *et al*., 2001; Ravasz *et al*., 2002) with few high-degree nodes connected to many lower-degree nodes. In contrast, an E-R graph has a nodal degree distribution that is Poisson (i.e., the probability of a node having degree k is $P(k) \approx \mathrm{e}^{-d} d^k/k!$, where d is the average node degree).

Figure 1(c) illustrates a network that is both scale-free and small-world. Small-world refers to the property that any two nodes are connected by a relatively small number of hops (transitions across edges), as reported from many biological networks (Ravasz *et al*., 2002). This property is quantified by the diameter which is the maximum distance between any pair of nodes in the network, where distance is defined here as the number of hops in the shortest path between a pair of nodes. In the example shown in Figure 1(c), the scale-free property is reflected by few nodes with high degrees called hubs connected to many nodes of low degree. The hubs are then joined by connections (dashed traces) so that the radius of the network remains small. Figure 1 shows other network architectures that do not show this

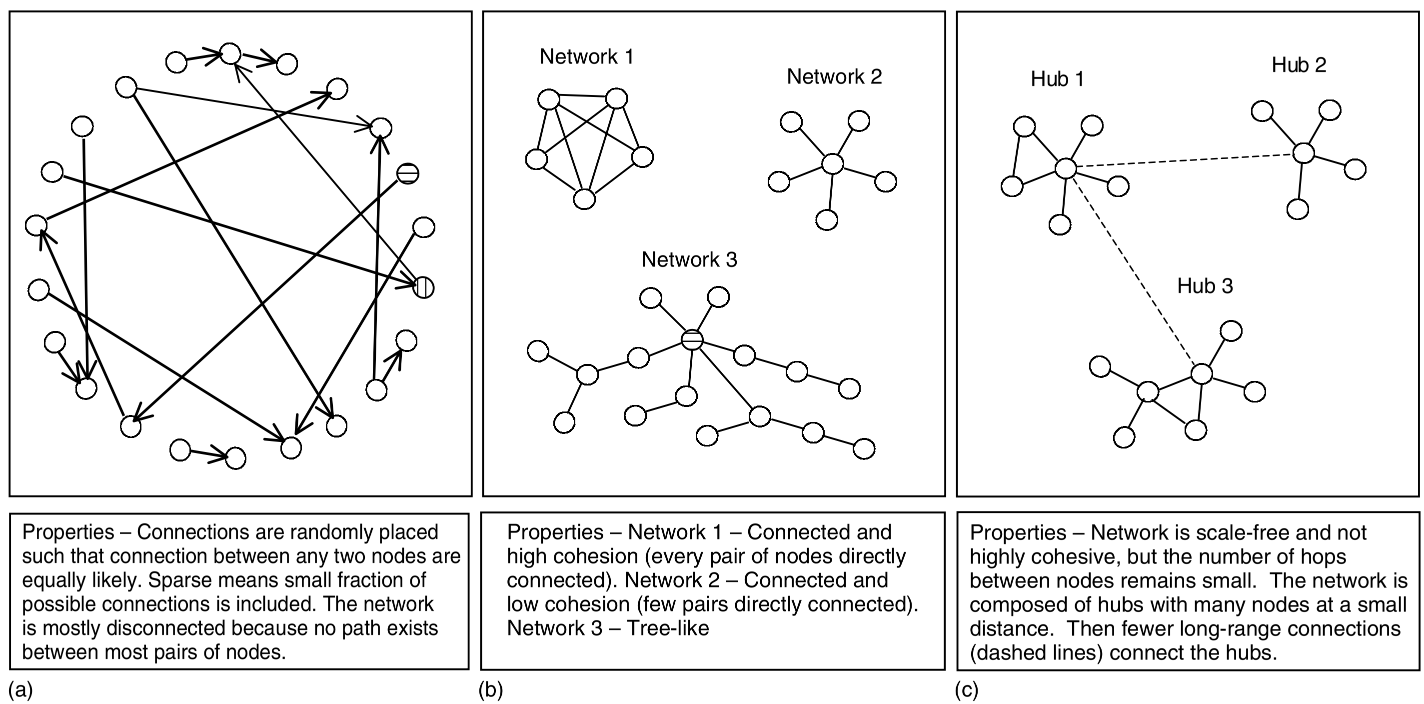


Figure 1 Sample networks illustrate network properties. (a) Erdös–Rényi (E-R) network; (b) three sample connected networks; and (c) scale-free and small-network is based on hubs with longer range links

scale-free property. For example, network 3 in Figure 1(b) is treelike with a root node. In these networks, the diameter grows as a function of the logarithm of the number of nodes.

4. Gene regulatory networks

Much of the previous discussion can be applied to a wide variety of networks. We now get more specific and consider gene regulatory networks in which connections represent the regulatory effect of one gene on another. For example, if gene A increases (or decreases) the expression of gene B, then an arrow with a positive (or negative) sign is placed from node A to node B. The actual mechanism is that gene A is transcribed and translated into a protein which in turn increases (or decreases) the rate of transcription of gene B. The expression of gene B can have further downstream actions on other genes or produce a protein that may have other effects (e.g., an enzyme could be produced to alter metabolic pathways (Reed *et al*., 2003) or repair DNA (Ronen *et al*., 2002)). Often, multiple genes can affect the target gene, and a gene may regulate itself, (*autoregulation*).

The transcriptional network from *Escherichia coli* is probably the most completely characterized gene regulatory network to date. Considerable knowledge of the regulatory mechanisms exists in the literature and has been condensed in databases such as Regulon DB (Salgado *et al*., 2000). In addition, considerable work has been done to analyze and reconstruct this network based on high-throughput data sources, as discussed later. By compiling the interactions in Regulon DB and other sources, networks similar to that shown in Figure 2(a) can be reconstructed (Shen-Orr *et al*., 2002). The network has 423 nodes and 578 connections. The representation does not show the autoregulation that exists in 59 of the 423 nodes. In addition, the regulatory network has three types of edges to represent activating (335), repressing (214), and both activating and repressing (29).

Several features of the network can be described using concepts already discussed. The network has a small number of nodes with high outdegree and a large number of nodes with only a single input or output connection to the rest of the network (see the starlike nodes in the center of Figure 2(a)). The node with the highest outdegree connects to 72 other nodes, and 15 other nodes have outdegree in the range of 10–26. The large number of nodes with high outdegree does not follow the Poisson distribution of E-R networks or the power law distributions of scale-free networks. Roughly, three-quarters of the nodes in the network (328 of 423) belong to a large connected component (here, a connection is assumed if any edge exists between a pair of nodes). The remaining nodes exist in components of 12 or fewer nodes. Figure 2(b) illustrates five distinct types of nodes in this network. External sources (red) and sinks (purple) are defined as having only one edge, either input or output, respectively. If one conceptualizes a flow of information from top to bottom, these nodes are starting and terminating points. Moving inward, external sources (yellow) and sinks (blue) connect to internal sources and sinks, respectively, that have two or more edges. The intermediary class (green) has relatively few nodes that connect only to internal sinks and sources (*see* Article 108, **Functional networks in mammalian cells**, Volume 6 and Article 113, **Metabolic dynamics in**

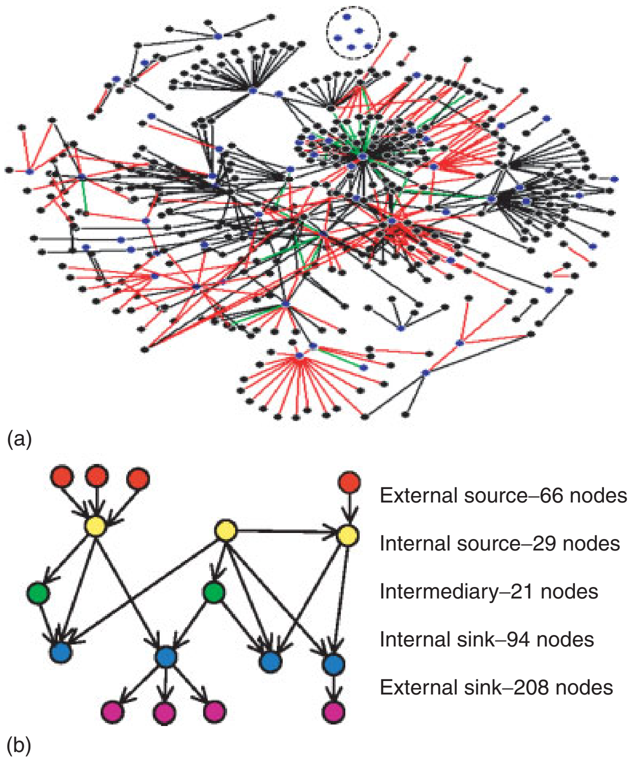


Figure 2 The transcriptional regulation network from *E. coli* (Adapted from Shen-Orr SS, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**, 64–68 by permission of Nature Publishing Group). (a) The network has 423 nodes and 578 edges and is displayed using Pajek (Batagelj and Mrvar, 2003). The regulatory network has three types of edges to represent activating (black), repressing (red), and both activating and repressing (green) connections. The 59 nodes with autoregulation are shown in blue, and the others are shown in black. (b) All but the five isolated nodes (in the dashed circle) can be placed in five classes of nodes, as labeled, based on the number and types of edges (see text for details)

cells viewed as multilayered, distributed, mass-energy-information networks, Volume 6.

5. Motif-based analysis

The connectivity of networks is only one static aspect of the interactions in the cellular environment. Yet, the complexity contained at this level of description is already considerable. Sufficiently large networks have interesting large-scale properties, such as power law degree distributions and relatively small diameter. Another level of analysis seeks to identify local properties of the network such as motifs. A motif is a recurring, possibly inexact, pattern in the network that may represent a reused module within the gene regulatory network. Moreover,

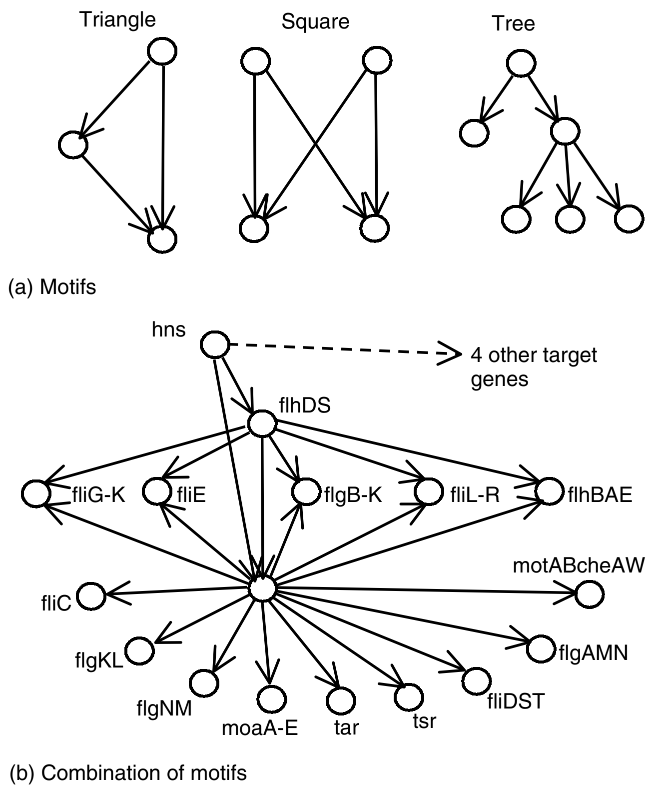


Figure 3 Samples of motifs discovered in the transcriptional regulation network from *E. coli* in Figure 2. Simple motifs in (a) can be combined in larger structures such as the example in (b) that corresponds to regulatory modules to assemble the flagella (see text for details)

the number of occurrences of motifs is usually compared to what is expected by chance, that is, the number discovered in randomized networks of similar properties (e.g., networks with randomized connections that preserve the in- and outdegree of each node). With this kind of analysis, motifs such as those shown in Figure 3(a) are found (Shen-Orr *et al*., 2002; Milo *et al*., 2002; Kershenbaum *et al*., 2003). The feed-forward triangle and the squares (also called bifan (Milo *et al*., 2002)) occurred at roughly two and three times the rate, respectively, as for the randomized network (Z-scores equal 4.9 and 7.5, respectively, Kershenbaum *et al*., 2003). The two examples are simple motifs that can be rigidly defined; however, motifs with more flexible structures can also be discovered. The descendent tree is a source node with several children, perhaps spread across levels (a two-level descendent tree is called a single input module in Shen-Orr *et al*., 2002).

One might also analyze networks as a composition of simple motifs. For example, the composite of motifs in Figure 3(b) is a composition of a descendent tree and several feed-forward triangles (Kershenbaum *et al*., 2003). The genes in this subnetwork are mostly associated with regulating the sequence and timing of proteins required to assemble the flagella (Kalir *et al*., 2001) (compare Figure 3(b)

with Figure 1 in the previous reference). Interestingly, the ancestor in this motif is a gene called H–NS, a master regulator in maintaining bacterial homeostasis under rapidly changing environments (Schroder and Wagner, 2002). Directly downstream of H–NS is flhDC, the master controller for the flagellar assembly process characterized in the previous work cited (Kalir *et al*., 2001). Note that the composite motif in Figure 3(b) was discovered on the basis of topology alone without regard to the name or function of the nodes. Hence, these results suggest that motif-based analyses can point to functional modules without full characterization of the nodes.

6. Biological interpretation of *E. coli* network topology

The layered network structure in Figure 2(b) suggests several properties. One can envision an "information flow" from a small number of sources (internal or external) to a much larger number of external sinks that can affect cellular processes. In the middle of the network, intermediary nodes may act as information integrators that subsequently affect larger numbers of downstream events. The layered networks produce short paths, perhaps suggesting that short temporal delays are favored over additional information integration that might be achieved with more levels of processing. Moreover, the negative autoregulation that is present in over 40% of *E. coli* transcription factors is thought to decrease rise-times in gene expression and hence may also decrease delay (Rosenfeld *et al*., 2002). Note that negative feedback loops at the level of gene regulation are not found, suggesting that additional stability as provided by this mechanism is not favored. Open loop control typically has better temporal responses and may be sufficient in this case. For example, the fast rate of division and high rate of error in DNA replication are thought to allow *E. coli* to rapidly "adjust parameters" in the metabolic pathways to make optimal use of the nutrients available in the current medium (Rosenfeld *et al*., 2002). Hence, this case suggests there is a regulation at the population level instead of at the level the individual gene network in this organism.

7. Reconstruction of pathways from high-throughput data

The *E. coli* network considered so far is constructed as compilations of data collected from the literature and other sources. Hence, the network represents a considerable part of the existing knowledge about the system and summarizes much of the underlying biology even though the functions of some genes may be better characterized than others. As mentioned earlier, however, *E. coli* is a special case in that it has been the object of a sustained research effort as one of the preferred model organisms for understanding prokaryotic biology. However, the knowledge base is sparser for most higher organisms. To address these limitations, considerable effort is being made to reconstruct networks based on high-throughput data and genome-wide data (*see* Article 110, **Reverse engineering gene regulatory networks**, Volume 6 and Article 118, **Data collection and analysis in systems biology**, Volume 6). However, high-throughput data is no panacea, and initial attempts to reconstruct large-scale, full kinetic models of the cell have faced

serious difficulties (Rice and Stolovitzky, 2004). From a theoretical perspective, the underdetermined nature of the problem (more unknowns than equations) implies that a unique solution is not generally possible because an infinite number of reconstructed systems are consistent with any given set of data. To deal with this nonuniqueness, the solution space is often limited by a priori, and often reasonable, assumptions such as linearity and sparseness, but even then the limitations of existing data render these approaches as mostly theoretical exercises except for reasonably small systems with high-quality data (Gardner *et al*., 2003; *see also* Article 110, **Reverse engineering gene regulatory networks**, Volume 6).

Another line of research seeks a reconstruction of the connectivity without full kinetic detail. We shall mention several efforts that are characteristic of the challenges facing the field (for a comprehensive review, see van Someren *et al*., 2002). A common approach is Bayesian inference methods as first used by Friedman *et al*. (2000) to analyze gene expression data. While these methods can handle noisy and incomplete data, the initial results showed that even three-node networks were hard to reconstruct in the yeast galactose metabolic pathway (Hartemink *et al*., 2001); however, much of the trouble may lie in the quality of the experimental data and not the method per se. Later work by the same group showed better results using synthetic gene networks where the researchers had better control of the data quality and quantity (Smith *et al*., 2003). In an elegant work (Gardner *et al*., 2003), researchers were able to reconstruct much of a nine-gene subnetwork in a DNA repair pathway in *E. coli* by controlled perturbations of a subset of the member genes. This work made the assumption of sparseness of the pathway connections and included a robust experimental design that kept low levels of noise in the real-time PCR measurement of transcript levels. In other recent studies (Segal *et al*., 2003; Troyanskaya *et al*., 2003), genome-wide yeast expression data and preliminary clustering were used to determine likely functional modules, that is, the sets of genes working together to perform a particular function. In addition, other data sources (candidate regulatory genes (Segal *et al*., 2003) and yeast two-hybrid data (Troyanskaya *et al*., 2003)) were combined to predict the functional modules. Hence, better results are found for understanding the regulation of sets of genes (versus single genes in initial attempts), more and perhaps better-quality data, and the use of complementary data types in addition to expression data alone.

8. Reconstruction of biological networks with pair-wise conditional correlation with a perturbed gene

We have proposed an alternative reconstruction method that seeks to determine the network connectivity only without generating a predictive model of the system (Rice *et al*., 2004). Starting with topologies based on the network in Figure 2(a), the network is endowed with dynamic behavior based on the work of Yeung *et al*. (2002). We propose an experimental design in which a single node is perturbed and the resulting effect on the rest of the network is measured. By computing the Pearson correlation of the expression profile of the perturbed gene to all the other genes in the network, the functionally connected genes can be inferred when the

correlation is above a set threshold. The process is repeated in a gene-by-gene fashion in order the reconstruct as large a network as needed. With this method, the network can be reconstructed with a high degree of accuracy that produces a reasonably low normalized fraction of false-positives and false-negatives.

The low reconstruction error rates of this and other similar studies are encouraging enough that full-scale automated reconstructions of gene regulatory networks may be possible in the foreseeable future with genome-wide data sets. False connections may be problematic for large network reconstruction (e.g., the *E. coli* 423 node network has $\sim 423^2$ possible connections so that a 1% false-positive error translates into more that 1000 false-positives, twice as many as the number of real connections). Therefore, reconstruction methods should usually be complemented with false-positive reducing approaches. Specifically, an optimal threshold to reject connections can be determined by modeling the distribution of correlation values of separate components representing connected versus disconnected nodes (Rice *et al*., 2004). Additional heuristics were developed to distinguish directly connected genes from indirectly connected genes that may also show high correlations (i.e., $X \rightarrow Z$ is distinguished from $X \rightarrow Y \rightarrow Z$).

While a large sparse network may make false-positives problematic, experimental noise may increase the rate of false-negatives. Other aspects of the *E. coli* gene regulatory network facilitated reconstruction. Specifically, the network hierarchical shape with large descendent trees that link relatively few sources to many sinks appears of be a topology that favors reconstruction. Specifically, the case of a gene being regulated by only one other (such as external sinks) is easier to reconstruct because the input and output show high correlation. In contrast, the infrequent structure of many nodes impinging on a single node is more difficult to reconstruct as the output is a function of many inputs and hence shows weaker correlation. Indeed, methods have shown similar difference in reconstruction accuracy for the case of single versus many inputs to a node (Smith *et al*., 2003). This latter study has shown difficulty with feedback pathways, but these are not seen to exist in the currently known *E. coli* network in Figure 2(a).

9. Toward a theory of biological systems

While allowing an unprecedented view into complex biological systems, the recent explosion of information has also opened a "Pandora's box". The systems biologist is like Newton under the tree who is pondering the apple in his hand. Fortunately, Newton connects the dots between the apple, the Earth, and his headache and develops calculus as the analytical tool to both understand and predict all motions, terrestrial and otherwise. A similar grand challenge for the systems biologist is to create a framework that allows the biologist to unify existing knowledge, to compare biological systems across species, and to make quantitative predictions of experimental manipulations. Network approaches are emerging as important ingredients in this quest. Networks are natural tools for compiling and mining high-throughput, genome-wide datasets. Networks are also tools to aggregate simpler interactions into large systems with complex behaviors. At the intersections of these tasks, large datasets can be used to infer network connectivity, pathways, and

potentially the kinetics between the interacting cellular players. Going the other direction, network-based tools such as motif discovery may help to decompose large systems into modules to more easily deduce functional relations. Clearly, networks will play important roles as the systems biologist tries to reign in the complexity and put the lid back on Pandora's box.

References

Batagelj V and Mrvar A (2003) In *Graph Drawing Software*, Vol. Jünger M and Mutzel P (Eds.) Springer: Berlin, **http://vlado.fmf.uni-lj.si/pub/networks/pajek/**.

Erdös P and Rényi A (1959) On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Friedman N, Linial M, Nachman I and Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Gardner TS, di Bernardo D, Lorenz D and Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Hartemink AJ, Gifford DK, Jaakkola TS and Young RA (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, **6**, 422–433.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG and Alon U (2001) Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, **292**, 2080–2083.

Kershenbaum A, Rice JJ and Stolovitzky G (2003) *Discovering Motifs in Biological Networks using Sub-Graph Isomorphism*. BMES 2003 Annual Fall Meeting, presented at *BMES 2003 Annual Fall Meeting*, Nashville.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Reed JL, Vo TD, Schilling CH and Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, **4**, R54.1465–R54.6914.

Rice JJ and Stolovitzky G (2004) Making the most of it: pathway reconstruction and integrative simulation using the data at hand. *Biosilico*, **2**(2), 70–77.

Rice JJ, Yu T and Stolovitzky G (2004) Reconstructing biological networks using conditional correlation analysis. *Bioinformatics*, Advanced Access published on October 14, 2004; doi:10.1093/bioinformatics/bti064.

Ronen M, Rosenberg R, Shraiman BI and Alon U (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10555–10560.

Rosenfeld N, Elowitz MB and Alon U (2002) Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, **323**, 785–793.

Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Blattner FR and Collado-Vides J (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Research*, **28**, 65–67.

Schroder O and Wagner R (2002) The bacterial regulatory protein H-NS–a versatile modulator of nucleic acid structures. *Biological Chemistry*, **383**, 945–960.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.

Shen-Orr SS, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**, 64–68.

Smith VA, Jarvis ED and Hartemink AJ (2003) Influence of network topology and data collection on network inference. *Pacific Symposium on Biocomputing*, **8**, 164–175.

Strogatz SH (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Troyanskaya OG, Dolinski K, Owen AB, Altman RB and Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8348–8353.

van Someren EP, Wessels LF, Backer E and Reinders MJ (2002) Genetic network modeling. *Pharmacogenomics*, **3**, 507–525.

Yeung MK, Tegner J and Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6163–6168.

Specialist Review

Reverse engineering gene regulatory networks

Michael J. Thompson, Michael E. Driscoll, Timothy S. Gardner and James J. Collins
Boston University, Boston, MA, USA

1. Introduction

Cellular function arises from the interaction of thousands of distinct chemical components, including DNA, RNAs, proteins, and small molecules. Determining the networks of interactions between such molecules has become a major interest of the biological community, with the mapping of chemical interactions underway in many model organisms. This effort involves the consolidation of data from focused biochemical, molecular biology, and genetics studies (Bhalla and Iyengar, 1999; Davidson *et al*., 2002), with data from high-throughput, systematic approaches. Here, we discuss approaches for building network models from biological data, a task referred to as *network inference* or *reverse engineering* (*see* Article 109, **Analyzing and reconstructing gene regulatory networks**, Volume 6 and Article 60, **Extracting networks from expression data**, Volume 7) (D'haeseleer *et al*., 2000; Banerjee and Zhang, 2002; Brazhnik *et al*., 2002; de Jong, 2002; Wyrick and Young, 2002; Li and Wang, 2003; Stark *et al*., 2003a,b; Friedman, 2004; Herrgård *et al*., 2004; Rice and Stolovitzky, 2004; Taverner *et al*., 2004).

Network inference approaches can be categorized (Table 1) by the level of physical detail in the model, the biological data used to construct the model, the inference method or underlying mathematical structure, and the type of biological insight desired from the model. There is no direct correspondence between the features in these different categories. For instance, different types of biological data and different model types can be used to infer a qualitative network model to describe cellular function. In the next sections, we will discuss the different levels of physical detail used in molecular network models and the types of biological data that are available for network inference. We will then describe four recent studies in detail (Beer and Tavazoie, 2004; Liao *et al*., 2003; Ronen *et al*., 2002; Gardner *et al*., 2003) to present different strategies for inferring gene regulatory networks. These four studies are notable in their emphasis on gaining biological insight, and each uses a different type of biological data to accomplish the inference task. We will also discuss the novel advances and limitations of each approach. In

Table 1 Distinguishing between network modeling approaches

Distinguishing characteristic	Examples
Level of description (Figure 1)	Topological, qualitative, and quantitative models
Data used to construct model	Physical and functional interactions sequence, physical interaction, and molecular abundance data; time-series and steady state perturbation experiments
Model type	Boolean, Bayesian, linear, and nonlinear models
Model purpose	Descriptive, interpretive, and predictive models

the final section, we will provide concluding remarks and suggest future directions for network inference studies.

2. Cellular network descriptions

There are different levels of detail (Rice and Stolovitzky, 2004) with which one can describe the structure of a molecular network (Table 1; Figure 1). Each level of description is appropriate for answering a different set of biological questions, and successful modeling involves a choice of detail appropriate for the question posed. In the first, and most simplified model (Figure 1a), the network is described as a set of homogenous components interacting through topological connections, indicating only which components interact. Models of this type have been used to compare the structure of molecular networks with other interaction networks, such as social networks made up of human components, and information networks consisting of website components. The potential biological insight from this work includes determining the universal principles responsible for the evolution, organization, and function of molecular networks (reviewed in Barabasi and Oltvai, 2004).

The second, and most predominant type of model (Figure 1b) includes additional details to describe the functional heterogeneity of the components and the nature of the interactions. These qualitative connections indicate properties such as the direction (causality) and sign (activating or inhibiting) of an interaction. This type of model is most prevalent because the function of a molecular network is often obvious based on its qualitative structure and biological intuition.

The third, and most detailed network model (Figure 1c) provides a quantitative description of the interactions between network components, indicating how each component behaves as a function of its inputs. This level of detail is required to predict functions that are not conceptually intuitive, such as determining the mode of action of a pharmaceutical compound or simulating dynamic cellular behavior. With increasing model detail, the task of inferring the structure of a network becomes more challenging and requires more biological data.

The level of detail of network models can also differ by the amount of physical realism described by the model (Table 1). A network connection can represent the interaction of two components that physically interact, as in the case of

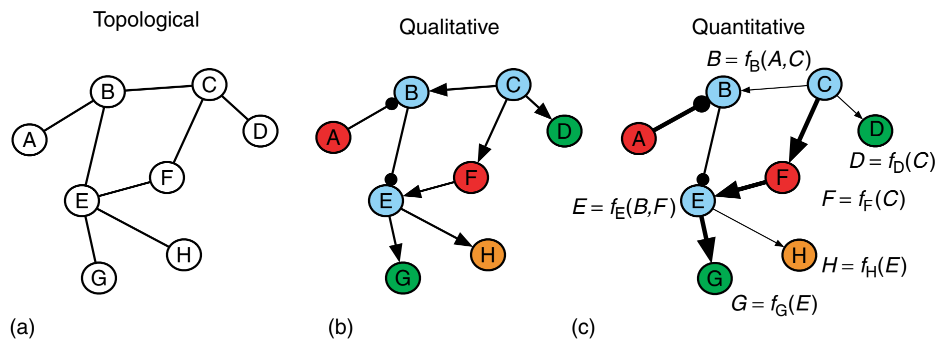


Figure 1 Molecular interaction network models can be described at different levels of detail. (a) Topological models describe graphs with homogenous molecular components and topological connections, indicating only which components interact. (b) Qualitative models include details of the functional heterogeneity of the molecular components (indicated by component color) and contain information on the direction (causality) and sign (activating or inhibiting) of interactions. Arrows begin at the causal factor and end at the component regulated by that factor. Arrow-tip terminated connections are activating; circle-tip connections are inhibiting. Component colors indicate different molecule functions; for instance, red components might be transcription factor proteins and blue components might be kinase proteins. (c) Quantitative models indicate how each component behaves as a function of the state of its inputs. For instance, component B will respond as a function of its inputs from components A and C. The strengths (indicated by line width of the arrow) of the interactions are captured in such a model

protein–protein interaction networks (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002; Giot *et al*., 2003; Li *et al*., 2004), or it can simply represent a functional relationship between two components. For example, a functional connection between two genes in a gene regulatory network means that a change in abundance of the product (mRNA or protein) of one gene affects the abundance of the product of the other gene. The physical interactions that mediate such a connection are unknown or ignored for the purpose of a simplified model (Brazhnik *et al*., 2002). This review focuses on the modeling of gene regulatory networks, so all of the connections described are functional in nature.

3. Inferring gene networks from biological data

The achievable level of model detail and corresponding network inference strategy both depend on the type of biological data that is available. For example, data on the absolute or relative abundance of molecular species can provide a quantitative measure of the response of molecular networks to stimuli. This can enable topological, qualitative, or quantitative network inference, and is especially important for qualitative and quantitative models. Technologies such as gene expression microarrays (Schena *et al*., 1995; Lockhart *et al*., 1996), which monitor mRNA abundance, have made the systematic collection of such molecular abundance data accessible to individual laboratories.

The most common approach to gaining biological insight with systematically collected molecular abundance data is to cluster genes into groups with similar patterns of expression over a set of conditions or a time course following a

stimulus (Eisen *et al*., 1998; Wen *et al*., 1998; Alon *et al*., 1999; Tamayo *et al*., 1999; Holter *et al*., 2000). Coexpression clustering has been reviewed elsewhere (*see* Article 90, **Microarrays: an overview**, Volume 4) (Brazma and Vilo, 2000; D'haeseleer *et al*., 2000; Gerstein and Jansen, 2000; Lockhart and Winzeler, 2000; Quackenbush, 2001). Groups of genes that share expression patterns have been shown to function in similar cellular processes, and their protein products are more likely to participate in physical interactions with one another (Eisen *et al*., 1998; Ge *et al*., 2001; Jansen *et al*., 2002). Observing correlated expression for a set of genes is useful for the construction of a gene regulatory network because the coexpressed genes often have a similar set of regulatory inputs. However, correlation alone is insufficient to uncover the regulatory inputs and construct a model of causal gene–gene interactions. Such relationships can be identified using additional data types and/or methods that control the causal factor in the experimental design.

Data suitable for inference of causal gene relationships include DNA sequence data (Tavazoie *et al*., 1999; Bussemaker *et al*., 2001; Pilpel *et al*., 2001; Wang *et al*., 2002; Segal *et al*., 2003c; Beer and Tavazoie, 2004; Haverty *et al*., 2004), annotated lists of regulators (Ihmels *et al*., 2002; Segal *et al*., 2003a; Qian *et al*., 2003; Haverty *et al*., 2004), physical protein–DNA binding data (Bar-Joseph *et al*., 2003; Liao *et al*., 2003; Gao *et al*., 2004; Kao *et al*., 2004), time-series experiments (Arkin and Ross, 1995; Arkin *et al*., 1997; Holter *et al*., 2001; Ronen *et al*., 2002; Perrin *et al*., 2003; Sontag *et al*., 2004), and targeted steady state perturbation experiments (Kyoda *et al*., 2000; Ideker *et al*., 2001; Pe'er *et al*., 2001; Wagner, 2001; Bruggeman *et al*., 2002; de la Fuente *et al*., 2002; Kholodenko *et al*., 2002; Wagner, 2002; Wang *et al*., 2002; Yeung *et al*., 2002; Tegner *et al*., 2003; Gardner *et al*., 2003; Sontag *et al*., 2004; Vlad *et al*., 2004). Methods that use this data to uncover the causal relationships include Bayesian inference, signal decomposition, and parameter estimation, and will be described below.

In the remainder of this section, we will focus on four examples of how molecular network inference has been used to gain biological insight. Each of the four studies uses a different type of biological data to accomplish the network inference task. Our emphasis will be on the novel advances and limitations of each approach, and the lessons in network modeling that can be drawn in each case.

4. Inference from sequence data

A recent study by Beer and Tavazoie (2004) used systematically collected mRNA abundance data in conjunction with DNA sequence data to study the combinatorial regulatory program underlying gene expression (Figure 2). As described above, genes that have correlated expression patterns are likely regulated by one or more common transcription factor proteins. The response of a gene to the simultaneous binding of more than one transcription factor may not be a simple linear superposition of the responses to the individual factors (Davidson *et al*., 2002; Zeitlinger *et al*., 2003). Beer and Tavazoie (2004) hypothesized that in addition to the combination of transcription factors regulating a particular gene (incoming network connections), the binding location and orientation of these factors relative to each other and to the start of protein translation would influence

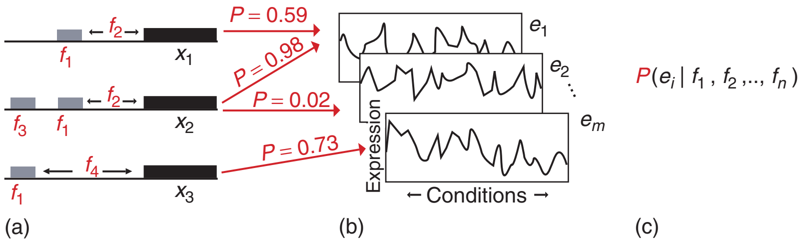


Figure 2 Approach taken by Beer and Tavazoie (2004) to determine quantitative connections between sequence determinants and gene expression. Red features are learned using the approach. (a) Sequence determinants of gene expression are enumerated as features, f_i, and include the presence of a motif, its orientation, and its distance from the translation start site. (b) Expression profiles, e_i, of genes under different conditions. The sequence determinants in (a) are mapped to corresponding gene expression profiles using a probabilistic function. (c) The conditional probability of a gene exhibiting a particular expression profile, e_i, based on sequence features, f_i, in its promoter. This probability is determined for all the expression profiles and sequence features using a Bayesian network

the expression of that gene. Information about the binding location and orientation of transcription factors is encoded in the DNA sequence, in which short patterns of nucleotides termed cis-regulatory motifs are recognized and bound by transcription factor proteins (Bulyk, 2003; Wasserman and Sandelin, 2004).

To accomplish their task, Beer and Tavazoie (2004) employed the Bayesian framework to describe the probabilistic dependencies between DNA sequence elements and gene expression profiles. First, they obtained data showing mRNA abundance changes in response to different environmental conditions, giving an expression profile for each gene. Next, they clustered the genes into sets of similar expression profiles, e_i (Figure 2b). They examined the promoter regions upstream of each gene, x_i, within a cluster of coexpressed genes for shared cis-regulatory motifs. Each motif was assigned a feature number, f_i, that could be used to indicate the presence (1) or absence (0) of that motif for a particular gene, x_i (Figure 2a). The motif orientation and distance from the translation start site were also assigned feature numbers. Finally, they described a Bayesian network (Pearl, 1988; Friedman *et al*., 2000; Pe'er *et al*., 2001; Perrin *et al*., 2003; Segal *et al*., 2003a,b,c; Friedman, 2004) mapping DNA sequence features ($f_1, f_2, \ldots, f_n$), to gene expression patterns (e_i) through the conditional probability, $P(e_i|f_1, f_2, \ldots, f_n)$, that a gene with a particular set of sequence features will participate in expression pattern i (Figure 2c). To train their model, Beer and Tavazoie (2004) searched through the space of sequence features to find a network (N) with the maximum probability of being correct given the observed data (D), using Baye's rule: $P(N|D) = P(N)P(D|N)/P(D)$. They withheld some data from the training set to test the predictive power of their model.

Beer and Tavazoie (2004) demonstrated their approach with gene expression data from the yeast *Saccharomyces cerevisiae*, responding to environmental stresses (Gasch *et al*., 2000) and a cell-cycle time course (Spellman *et al*., 1998). One expression profile they observed showed the stress-induced change in abundance of gene products involved in ribosomal RNA transcription and processing. They

identified two cis-regulatory motifs, termed PAC and RRPE (Sudarsanam *et al*., 2002), that are overrepresented in the promoter regions of these genes. Their Bayesian network model showed that the presence of the RRPE motif within 240 bp of the translation start site (ATG) of a gene indicates a 22% probability that the gene will exhibit the expression profile. Similarly, the presence of the PAC motif within 140 bp of the ATG indicates a 67% probability of the same. Importantly, the presence of both motifs within those respective locations is 100% indicative that the gene will exhibit the expression profile. Beer and Tavazoie (2004) describe this combinatorial program as AND logic, which strictly speaking means that the pattern is followed if both features are present and not followed if either individual feature or neither feature is present. Whether that same logical function would describe the response of a gene to the presence of activated versions of transcription factors that bind the PAC and RRPE motifs remains to be seen.

Beer and Tavazoie (2004) identified many other combinatorial sequence determinants of gene expression patterns in yeast, including some resembling OR and NOT logic functions. They demonstrated that those sequence features are predictive of the mRNA abundance changes of greater than 70% of the yeast genes responding to the observed environmental stress conditions. They also identified a program of combinatorial transcriptional regulation controlling the embryonic and larval development of the nematode *Caenorhabditis elegans*. They reasonably suggested that more and higher-quality expression data will enable the identification of many more regulatory programs.

The use of a Bayesian network approach provided some key benefits in this study. It allowed easy incorporation of heterogeneous determinants of gene expression, including not only sequence but also motif orientation and location. In addition, combinatorial gene regulation is commonly considered to involve a nonlinear response of gene expression to multiple simultaneous inputs. The Bayesian framework provides a probabilistic model with which these nonlinearities can be explored. Interestingly, it appears that regulation of genes by the PAC and RRPE motifs cannot be described using pure AND logic. For instance, genes containing only PAC or only RRPE can still exhibit the observed expression pattern (see above). Under these circumstances, which may be common in biological systems, it is likely that a linear model could also capture much of the important behavior. The additional benefit of using sequence data is that influential sequence determinants that are not associated with the binding of a particular protein, such as GC content affecting the physical properties of DNA, may influence transcription and could be learned using approaches similar to that of Beer and Tavazoie (2004).

Despite the excellent promise of this approach, some limitations should be considered in the future development of network inference strategies. For example, it is not yet clear how much of the information on the regulation of a gene is encoded in the DNA sequence in its promoter region. In higher organisms, the boundaries of promoter regions are less clear, and sequence determinants are commonly located tens of thousands of nucleotides upstream of a gene (Alberts, 2002). To infer a gene regulatory network model, one would also like to know the transcription factors that bind a given cis-regulatory motif. Coupling the approach of Beer and Tavazoie (2004) with genome-wide protein–DNA binding data determined using chromatin immunoprecipitation (Lee *et al*., 2002) is one

strategy to consider. Finally, as Beer and Tavazoie (2004) discuss, their statistical approach of averaging over genes that follow a common expression pattern may not enable the learning of very complex combinatorial regulation programs that have few or unique instances in a genome. However, the latter limitation may be partially overcome by using a strategy of comparing sequences of similar organisms (Bulyk, 2003; Cliften *et al*., 2003; Kellis *et al*., 2003; Wasserman and Sandelin, 2004).

5. Inference from physical interaction data

A recent study by Liao *et al*. (2003) demonstrated how mRNA abundance data and known qualitative network interactions from physical protein–DNA binding data (Ren *et al*., 2000; Lieb *et al*., 2001; Iyer *et al*., 2001; Lee *et al*., 2002; Kurdistani *et al*., 2002) can be used to determine quantitative gene regulatory network connections (Figure 3). Their approach also enabled the discovery of the activity profiles of a set of transcriptional regulators over a time course and a set of environmental conditions. The aim of their approach was to model the behavior of each of the regulated genes in the network as a linear combination of the activity profiles of the transcription factors that regulate that gene. The approach of Liao *et al*. (2003) is termed *network component analysis (NCA)*, and is suggested to yield more biologically relevant linear combinations than other signal decomposition approaches because of its use of prior information on regulatory relationships.

The network model includes a layer of transcription factors (Figure 3b) interacting with a larger set of regulated genes (Figure 3a). The NCA is provided

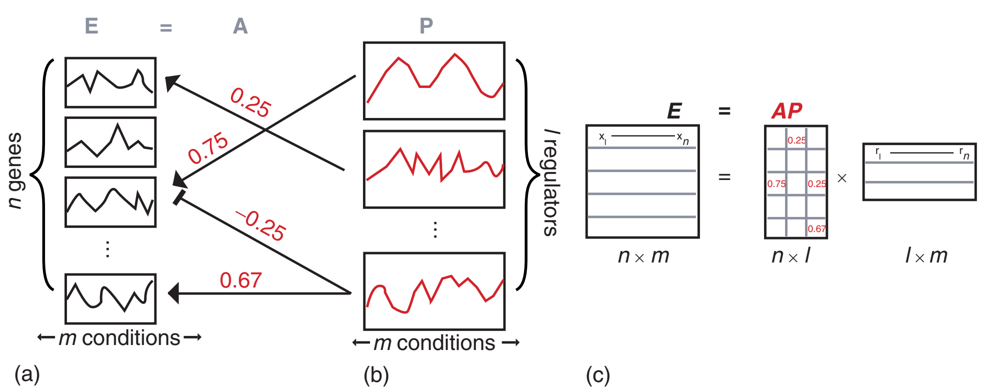


Figure 3 Approach taken by Liao *et al*. (2003) to determine quantitative gene regulatory connections and transcription factor activity (TFA) profiles using molecular abundance data and physical protein–DNA binding data. Red features are learned using the approach. (a) The algorithm is supplied with a matrix of experimental data, **E**, monitoring the abundance of n genes over m time points or environmental conditions. (b) The TFA profiles of the l transcription factors are contained in the **P** matrix. These TFA profiles are learned using the approach. The transcription factors direct regulatory connections, contained in matrix **A**, toward the genes in the network. The structure of these connections is known a priori, but the connection strengths and signs are learned using the approach. (c) The expression profiles, **E**, of the n regulated genes in the network are a linear combination of the l transcription factor activity profiles, **P**, weighted by the connectivity matrix, **A**

with a matrix, **E**, containing experimental data monitoring the abundance of each gene product in the network over a time course or set of conditions (Figure 3a). The approach by Liao *et al*. (2003) performs a linear decomposition of the expression data, **E**, into a connectivity matrix, **A**, and regulator profile matrix, **P** (Figure 3c). The connectivity matrix, **A**, describes the interactions directed from the transcription factors to the regulated genes. The **A** matrix is a pruned model of connectivity, representing a minimal nonoverlapping set of connections between transcription factors and regulated genes. Feedback connections are not allowed. The nonzero elements of matrix **A** are known from the physical protein–DNA binding studies, and the sign and magnitude (weights) of the elements (regulatory connections) are determined using NCA (Figure 3). The transcription factor activity (TFA) profile matrix, **P**, is determined using NCA, and describes the activity of the transcription factors (Figure 3b) over the same time course or conditions as the regulated genes. Decomposition of the expression data is achieved using a least-squares minimization, min $||\boldsymbol{E} - \boldsymbol{AP}||^2$, through an iterative procedure starting with the known network structure and random connectivity strengths.

Liao *et al*. (2003) demonstrated the NCA approach using mRNA abundance data taken over the time course of the yeast cell cycle (Spellman *et al*., 1998) and physical protein–DNA binding data obtained using systematic chromatin immunoprecipitation (Lee *et al*., 2002). They focused on 11 transcription factors known to regulate gene expression in a cell-cycle-dependent manner. To satisfy the criteria for NCA, the network of regulated genes was reduced to 441 genes under the control of these 11 factors and 22 other transcription factors, for a total of 33 regulators. Liao *et al*. (2003) used the NCA to determine the connectivity strengths and signs between these 33 transcription factors and the 441 genes, as well as the activity profiles of the transcription factors over the course of the cell cycle. Interestingly, they observed that the regulator activity profiles captured by the NCA do not necessarily correspond to the molecular abundance profiles of those regulators. They observed cases in which the activity profile of the regulator protein exhibited cyclical behavior, even though its corresponding mRNA abundance did not.

Even more compelling is a similar observation made by Liao and colleagues in a second study examining the response of *Escherichia coli* to carbon source transition (Kao *et al*., 2004). They decomposed the behavior of 100 genes responding to a shift from glucose to acetate growth media, into regulatory contributions by 16 transcription factors and the associated activity profiles of those regulators. One of the transcription factors they examined was CRP (combinatorial regulation programs), which requires the binding of the small molecule, cAMP, for its regulatory activity. Kao *et al*. (2004) showed that the activity profile (TFA) of the CRP-cAMP complex determined by the NCA algorithm matches the abundance profile of cAMP. With the NCA approach, it will be possible in principle to identify transcription factors whose abundance profile differs from their activity profile. These transcription factors would likely be regulated by means other than molecular abundance changes, including regulation by phosphorylation, localization, or complexation.

A limitation of this approach is the dependence on prior knowledge of the semiqualitative network connections. However, the construction of such network models is becoming more feasible with the use of DNA sequence data (described

above) and the determination of physical protein–DNA binding locations using chromatin immunoprecipitation.

6. Inference from time-series experiments

A recent study by Ronen *et al*. (2002) used time series of molecular abundance measurements to determine kinetic parameters (quantitative connections) for a network with previously known qualitative (direction and sign) connections (Figure 4). Their goal was to demonstrate the inference of a more complete descriptive model that could be used to predict the dynamics of the entire network on the basis of the observation of a single gene. The molecular network they examined was the DNA damage response and repair (SOS) pathway in *E. coli*. This network regulates the cellular response to DNA damage and involves more than 100 genes. Under normal conditions, the LexA transcriptional repressor ("R" in Figure 4b) blocks expression of other genes in the network. Under DNA damaging conditions, the RecA protein becomes activated and mediates LexA degradation, thereby relieving the repression of the SOS network genes.

Ronen *et al*. (2002) measured the temporal change in abundance (Figure 4a) of eight genes in the SOS network, including *lexA* and *recA*, following UV irradiation. They estimated the abundance of the proteins in the network using a fluorescent reporter protein placed under the control of the eight promoters. Their goal was to fit kinetic parameters for the production rate from each derepressed promoter (β_i), and the effective affinity of the LexA protein for each promoter (k_i), based on the time-dependent LexA repressor activity ($R(t)$; Figure 4b) and the observed promoter activity with respect to time ($X(t)$; Figure 4a). First, the LexA repressor activity ($R(t)$) was determined on the basis of singular value decomposition (SVD) from the different experiments. Next, the kinetic parameters were estimated using a Michaelis–Menten model for the regulatory dynamics (Figure 4c). This equation

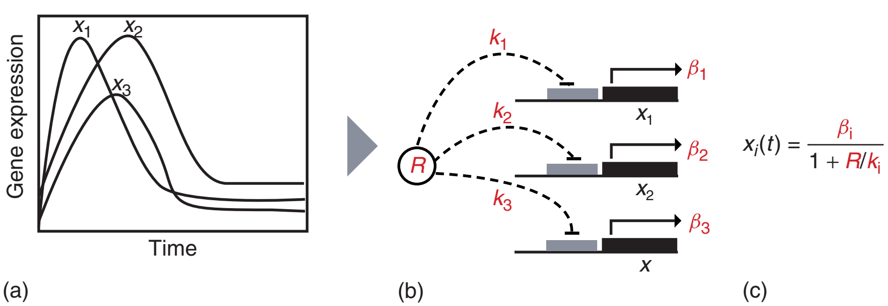


Figure 4 Approach taken by Ronen *et al*. (2002) to determine kinetic parameters (quantitative connections) for gene expression using time-series measurements of molecular abundance, and a prior model of the qualitative connections. Red features are learned using the approach. (a) Time-varying molecular abundances measured for every regulated species. (b) Regulatory interactions show that gene expression, x_i, is dependent on the repressor activity, R, the binding affinity for the repressor, k_i, and the expression level in the absence of repressor, β_i. (c) Michaelis–Menten model, showing the relationship between gene expression and the parameters in (b)

is specific to the system they investigated, and would take a different form if any of the genes were activated by LexA, or if LexA bound cooperatively to any of the promoters.

Ronen *et al*. (2002) were able to estimate kinetic parameters that reproduced the observed time-dependent upregulation of eight SOS network genes upon simulated LexA degradation. They discussed how the temporal order of the alleviation and reapplication of repression for the eight genes, as determined by the magnitude of the repressor binding affinity parameters, k_i, agrees with the order in which the genes function in the DNA repair process. Others have observed similar mechanistically significant temporal ordering of gene expression in different systems (Spellman *et al*., 1998; Laub *et al*., 2000; Kalir *et al*., 2001; Zaslaver *et al*., 2004), and this appears to be a key feature in transcriptional network dynamics. Ronen *et al*. (2002) also demonstrated how their kinetic model could be used to estimate the relative abundance of the active LexA regulator over the time course of the experiments, showing agreement with experiments that directly determined protein abundance (Sassanfar and Roberts, 1990). They also showed that the calculated error in estimated parameters could be used to anticipate additional regulation for a particular promoter that is not captured on the basis of LexA activity alone.

There are some limitations of this approach as a general solution for determining quantitative connections in gene regulatory networks. The strategy taken by Ronen *et al*. (2002) is specific for modeling the response of a group of genes to changes in activity of a single regulatory factor. The approach cannot describe the time-dependent response of a group of genes to multiple varying transcription inputs, or under circumstances in which there is feedback between the regulated genes and the regulatory factor. Elucidating the behavior in these more complicated regulatory system would require knowledge of the cis-regulatory logic describing multiple simultaneous inputs (Beer and Tavazoie, 2004). Another limitation of this approach is that it requires prior knowledge of the qualitative connections in the network, because the selection of an appropriate kinetic model will depend on the system structure. For the most part, the elucidation of such a qualitative network model remains a challenge. In principle, time-series data could be used to obtain the causality information required for inferring a network of qualitative connections with no prior structural information (Arkin and Ross, 1995; Arkin *et al*., 1997; Holter *et al*., 2001; Sontag *et al*., 2004). For example, Perrin *et al*. (2003) used the time-series data collected by Ronen *et al*. (2002) to infer a network of qualitative interactions between the eight genes in the SOS subnetwork.

7. Inference from perturbation experiments

A recent study by Gardner *et al*. (2003) demonstrated the use of molecular abundance data measured in response to targeted steady state perturbations to infer a gene network model with quantitative connections (Figure 5). The resulting network model describes the quantitative influence of each gene on the expression of every other gene in the network. One of the goals of the study was to use the resulting model to identify the major regulators in the network, which in their model need not

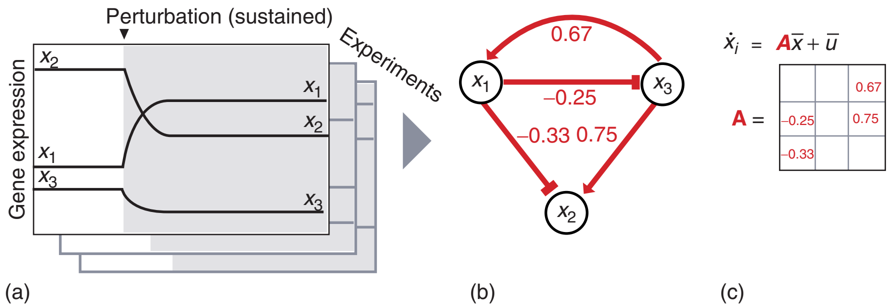


Figure 5 Approach taken by Gardner *et al*. (2003) to determine quantitative connections between genes on the basis of molecular abundance measurements following targeted steady state perturbations. No prior information on network structure was required. Red features are learned using the approach. (a) Steady state molecular abundance measured for every species following genetic perturbations. (b) Quantitative network model describes the influence of each gene on every other gene. The connectivity matrix, **A**, encodes these influence weights. (c) Linear model for the accumulation of each species, $\dot{x}_i$, based on the abundance of the other species, $\vec{x}$, the connectivity matrix, **A**, and the perturbations, $\vec{u}$. At steady state, the molecular abundances are not changing ($\dot{x}_i = 0$)

be transcription factor proteins. Another goal was to predict the targets of unknown perturbations (for instance, the molecular target of a pharmaceutical compound).

In the approach taken by Gardner *et al*. (2003), a cellular system near steady state is subjected to the specific perturbation (overexpression or downregulation) of a gene thought to have an influence on a network of interest. Once the system returns to steady state, the response of the system to the perturbation is measured by determining the molecular abundance of every species in the network (Figure 5a). Near a steady state, the behavior of the system can be approximated by a linear model (Figure 5c) describing the rate of accumulation, $\dot{x}_i$, of each species, i, in the network in terms of the abundance of every gene product in the network, $\vec{x}$, the quantitative connectivity matrix, **A**, and the perturbation, $\vec{u}$, made to the system. At steady state, the abundance of each species is not changing ($\dot{x}_i = 0$), so the equation in Figure 5(c) reduces to $A\vec{x} = -\vec{u}$. After making many perturbation-response measurements, the quantitative coefficients of A (Figure 5b), which represent the influence of each gene on every other gene, can be determined using linear regression.

Gardner *et al*. (2003) tested their approach, termed *network identification via multiple regression (NIR)*, on the SOS network in *E. coli* (described above). As a starting point, they applied the NIR method to a nine-gene subset at the core of the network. The authors used plasmid-borne copies of each gene to individually alter its abundance in nine separate experiments and measured the resulting changes in mRNA abundance for all nine components (Figure 5a). The NIR method was able to correctly identify 25 of the previously identified regulatory relationships between the nine genes, as well as 14 relationships that may be novel interactions or false positives. Moreover, the network model obtained by the NIR algorithm correctly identified the *recA* and *lexA* genes, the known principal regulators of the SOS response, as having the strongest influence (largest regulatory weights) on the

other genes in the network. Thus, the model can be used to identify which genes should be perturbed to elicit a maximal response from the network.

The network model obtained by the NIR algorithm was also used to identify the genes that mediate the network response to a particular stimulus. As illustrated in Figure 6, the network model correctly identifies the *recA* gene as the key mediator of the SOS network response to treatment with UV radiation, mitomycin C (MMC), and the quinolone antibiotic pefloxacin (each of which causes DNA damage). For novobiocin (Figure 6), a quinolone that does not cause DNA damage, *recA* is not predicted as the mediator of the expression response. The predictive power of the network model obtained using the NIR algorithm is due to its use of quantitative connections, and this level of model detail was specifically chosen to enable the identification of compound mode of action.

There are some limitations of the NIR approach for inferring molecular network structure and function. The predictive capability of the model was gained at the expense of its ability to describe many of the molecular interactions with mechanistic detail. For instance, while the NIR approach can include regulatory connections directed by nontranscription factors, the physical interactions that mediate those connections are often unknown, making it difficult to interpret the meaning of any particular connection. Other limitations of the approach include the challenge of delivering targeted (single gene) perturbations, and the difficulty in knowing which genes to perturb to most efficiently reconstruct the network model.

8. Concluding remarks and future directions

We have examined in detail four recent approaches to gene network inference (Beer and Tavazoie, 2004; Liao *et al*., 2003; Kao *et al*., 2004; Ronen *et al*., 2002; Gardner *et al*., 2003). These four approaches can be distinguished by several different model characteristics (Table 1). Despite differing in their underlying model structure, each of the four approaches provides a gene network model with some quantitative detail about the connections. For instance, the Bayesian network model of Beer and Tavazoie (2004) described the behavior of a cluster of genes as a probabilistic function of DNA sequence determinants, and the linear influence network model of Gardner *et al*. (2003) described the accumulation of one species as a weighted sum of the abundance of the other species in the network.

Although every gene regulatory network model has connections that represent functional relationships rather than physical interactions, the models can differ in their descriptive power. For instance, the approaches of Beer and Tavazoie (2004), Liao *et al*. (2003), and Ronen *et al*. (2002) describe connections between transcription factors (or the corresponding DNA sequence motifs to which they bind) and regulated genes. These interactions occur through well-understood mechanisms, and identifying such an interaction is therefore descriptive of the molecular behavior. The approach by Gardner *et al*. (2003), on the other hand, yields connections that are often mediated by several unknown physical interactions, making descriptive interpretation challenging. The resulting model, however, enables the prediction of behaviors resulting from the perturbation of nontranscription factors, including protein targets of pharmaceutical compounds.

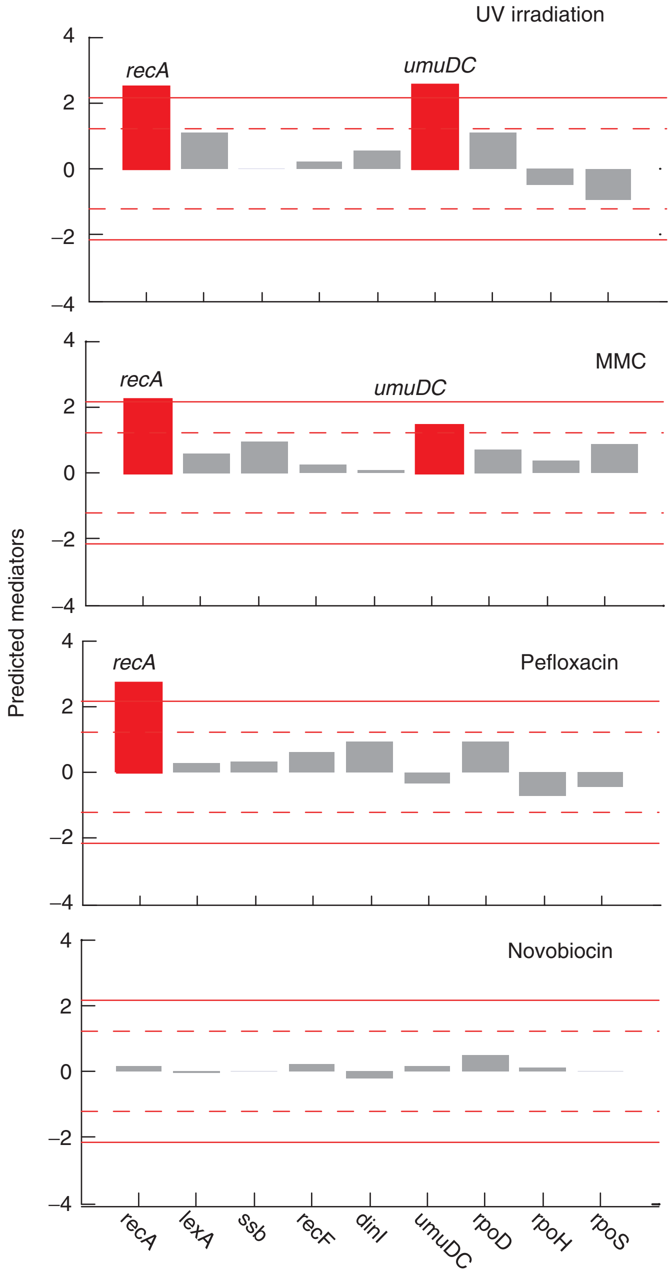


Figure 6 Prediction of genes mediating response to four different stimuli. The network model identified with the approach of Gardner *et al*. (2003) (Figure 5) was used to predict the mediators of expression responses following UV irradiation and treatment with three drug compounds. In the case of treatment with UV radiation, mitomycin C (MMC), and pefloxacin, all of which cause DNA damage, the *recA* gene is correctly predicted as the mediator of the expression response. For treatment with novobiocin, which does not damage DNA, *recA* is not predicted as the mediator of the expression response. Lines denote significance levels: $P = 0.3$ (dashed), $P = 0.1$ (solid)

The current trend in molecular network inference is in coupling different types of experimental data to improve the quality and completeness of the inferred model. The Bayesian formalism has been invoked in many studies as the most convenient approach for including previous knowledge and tolerating missing information (Beaumont and Rannala, 2004; Friedman, 2004). As Liao *et al*. (2003) and Kao *et al*. (2004) demonstrated, it is also possible to include prior information on network structure in linear models using a constraint-based approach (*see* Article 112, **Constraint-based modeling of metabolomic systems**, Volume 6). Another type of data that has not been discussed here are the synthetical lethal interactions, determined systematically in yeast by Tong *et al*. (2004). These interactions represent functional relationships between genes that are involved in similar or related cellular processes, and coupling these data with network inference approaches should prove fruitful. Finally, the use of physical protein–protein interaction data (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002; Giot *et al*., 2003; Li *et al*., 2004) and metabolic network structure information (Covert *et al*., 2001, 2004) to trace the physical interaction pathways responsible for the functional relationships between genes will help elucidate gene network structure and extend the value of gene regulatory networks as descriptive models of cellular function.

Research into approaches for network inference would benefit from the availability of large, standardized sets of molecular abundance data collected with the goal of network inference in mind. As we have described in this study, time-series experiments and targeted steady state perturbation experiments will be particularly useful. The availability of such data in several model organisms (such as *E. coli*, *S. cerevisiae*, and *C. elegans*) would significantly advance biological discovery.

Acknowledgments

We thank Boris Hayete and Mads Kærn for critical readings of the manuscript. This work was supported by the NIH, the NHLBI Proteomics Initiative, the DOE, and the NSF.

References

Alberts B (Ed.) (2002) *Molecular Biology of the Cell*. Fourth Edition, Garland Science: New York.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.

Arkin A and Ross J (1995) Statistical construction of chemical reaction mechanisms from measured time-series. *Journal of Physical Chemistry*, **99**, 970–979.

Arkin A, Shen P and Ross J (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**, 1275–1279.

Banerjee N and Zhang MQ (2002) Functional genomics as applied to mapping transcription regulatory networks. *Current Opinion in Microbiology*, **5**, 313–317.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon B, Fraenkel E, Jaakkola T, Young R, *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **21**, 1337–1342.

Barabasi AL and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, **5**, 101–113.

Beaumont MA and Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews. Genetics*, **5**, 251–261.

Beer MA and Tavazoie S (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Bhalla US and Iyengar R (1999) Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381–387.

Brazhnik P, de la Fuente A and Mendes P (2002) Gene networks: how to put the function in genomics. *Trends in Biotechnology*, **20**, 467–472.

Brazma A and Vilo J (2000) Gene expression data analysis. *FEBS Letters*, **480**, 17–24.

Bruggeman FJ, Westerhoff HV, Hoek JB and Kholodenko BN (2002) Modular response analysis of cellular regulatory networks. *Journal of Theoretical Biology*, **218**, 507–520.

Bulyk ML (2003) Computational prediction of transcription-factor binding site locations. *Genome Biology*, **5**, 201.

Bussemaker HJ, Li H and Siggia ED (2001) Regulatory element detection using correlation with expression. *Nature Genetics*, **27**, 167–171.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA and Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Covert MW, Knight EM, Reed JL, Herrgard MJ and Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E and Palsson BO (2001) Metabolic modeling of microbial strains in silico. *Trends in Biochemicaj Sciences*, **26**, 179–186.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, *et al*. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

de Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, **9**, 67–103.

de la Fuente A, Brazhnik P and Mendes P (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genetics*, **18**, 395–398.

D'haeseleer P, Liang S and Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Friedman N, Linial M, Nachman I and Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Gao F, Foat BC and Bussemaker HJ (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Gardner TS, di Bernardo D, Lorenz D and Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, **11**, 4241–4257.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ge H, Liu Z, Church GM and Vidal M (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics*, **29**, 482–486.

Gerstein M and Jansen R (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Current Opinion in Structural Biology*, **10**, 574–584.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.

Haverty PM, Hansen U and Weng Z (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Research*, **32**, 179–188.

Herrgård MJ, Covert MW and Palsson B (2004) Reconstruction of microbial transcriptional regulatory networks. *Current Opinion in Biotechnology*, **15**, 70–77.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Holter NS, Maritan A, Cieplak M, Fedoroff NV and Banavar JR (2001) Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 1693–1698.

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR and Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 8409–8414.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y and Barkai N (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**, 370–377.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M and Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Jansen R, Greenbaum D and Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Research*, **12**, 37–46, letter.

Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG and Alon U (2001) Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, **292**, 2080–2083.

Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V and Liao JC (2004) Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 641–646.

Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E, Westerhoff HV and Hoek JB (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12841–12846.

Kurdistani SK, Robyr D, Tavazoie S and Grunstein M (2002) Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nature Genetics*, **31**, 248–254.

Kyoda KM, Morohashi M, Onami S and Kitano H (2000) A gene network inference method from continuous-value gene expression data of wild-type and mutants. *Genome Informatics Series: Workshop on Genome Informatics*, **11**, 196–204.

Laub MT, McAdams HH, Feldblyum T, Fraser CM and Shapiro L (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Li H and Wang W (2003) Dissecting the transcription networks of a cell using computational genomics. *Current Opinion in Genetics & Development*, **13**, 611–616.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C and Roychowdhury V (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15522–15527.

Lieb JD, Liu X, Botstein D and Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics*, **28**, 327–334.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.

Lockhart DJ and Winzeler EA (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann: San Francisco, CA.

Pe'er D, Regev A, Elidan G and Friedman N (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**(Suppl 1), S215–S224.

Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J and D'Alché-Buc F (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**(Suppl 2), II138–II148.

Pilpel Y, Sudarsanam P and Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, **29**, 153–159.

Qian J, Lin J, Luscombe NM, Yu H and Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.

Quackenbush J (2001) Computational analysis of microarray data. *Nature Reviews. Genetics*, **2**, 418–427.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rice J and Stolovitzky G (2004) Making the most of it: pathway reconstruction and integrative simulation using the data at hand. *Drug Discovery Today: BioSilico*, **2**, 70–77.

Ronen M, Rosenberg R, Shraiman BI and Alon U (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10555–10560.

Sassanfar M and Roberts JW (1990) Nature of the SOS-inducing signal in Escherichia coli. The involvement of DNA replication. *Journal of Molecular Biology*, **212**, 79–96.

Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003a) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.

Segal E, Wang H and Koller D (2003b) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**(Suppl 1), I264–I272.

Segal E, Yelensky R and Koller D (2003c) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**(Suppl 1), I273–I282.

Sontag E, Kiyatkin A and Kholodenko BN (2004) Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, **20**(12), 1877–1886.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast

Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Stark J, Brewer D, Barenco M, Tomescu D, Callard R and Hubank M (2003a) Reconstructing gene networks: what are the limits? *Biochemical Society Transactions*, **31**, 1519–1525.

Stark J, Callard R and Hubank M (2003b) From the top down: towards a predictive biology of signalling networks. *Trends in Biotechnology*, **21**, 290–293.

Sudarsanam P, Pilpel Y and Church GM (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Research*, **12**, 1723–1731.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907–2912.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281–285.

Taverner NV, Smith JC and Wardle FC (2004) Identifying transcriptional targets. *Genome Biology*, **5**, 210.

Tegner J, Yeung MK, Hasty J and Collins JJ (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 5944–5949.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Vlad MO, Arkin A and Ross J (2004) Response experiments for nonlinear systems with application to reaction kinetics and genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7223–7228.

Wagner A (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps. *Bioinformatics*, **17**, 1183–1197.

Wagner A (2002) Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Research*, **12**, 309–315.

Wang W, Cherry JM, Botstein D and Li H (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16893–16898.

Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Reviews. Genetics*, **5**, 276–287.

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL and Somogyi R (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 334–339.

Wyrick JJ and Young RA (2002) Deciphering gene expression regulatory networks. *Current Opinion in Genetics & Development*, **12**, 130–136.

Yeung MKS, Tegner J and Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6163–6168.

Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG and Alon U (2004) Just-in-time transcription program in metabolic pathways. *Nature Genetics*, **36**, 486–491.

Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR and Young RA (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, **113**, 395–404.

Specialist Review

Functional inference from probabilistic protein interaction networks

Joel S. Bader
Johns Hopkins University, Baltimore, MD, USA

1. Introduction

Access to complete genome sequence reveals how few genes have well-characterized function. A minority of genes in human and important model organisms have been classified according to the widely used Gene Ontology terms (Ashburner *et al*., 2000; Harris *et al*., 2004). The need to annotate gene function has driven the development of technologies for high-throughput experimental characterization of genes and proteins, both to generate raw data and to serve as the foundation for wider computational inference based on genome sequence.

Protein-coding genes are the most numerous class of annotated functional units in all sequenced genomics. The most direct characterization of a protein's function is to understand its physical interactions with other biomolecules: interactions with metabolites for enzymes, and interactions with other proteins and DNA for proteins that function in regulatory, signal transduction, structural, or other nonmetabolic processes. As the organism complexity increases, from bacteria to unicellular eukaryotes to metazoans, the fraction of proteins with nonmetabolic function also increases. These proteins' functions are best described by placing a protein in the context of a larger pathway defined by discrete protein–protein interactions. Two technologies, two-hybrid and mass spectrometry, have answered the challenge of probing protein–protein interaction networks at the full-proteome level.

Two-hybrid methods use an engineered system in yeast to test pairwise interactions between proteins from yeast or a foreign species. Large-scale protein interaction maps based on the two-hybrid method were generated originally for yeast (Uetz *et al*., 2000; Ito *et al*., 2001). Full-genome screens have now been reported for fly (Giot *et al*., 2003), and a screen for metazoan-specific proteins has been reported for worm (Li *et al*., 2004). Full-genome coverage in this context refers to an attempt to use each predicted protein-coding gene as input to the assay. Technical difficulties with the assays and undersampling combine to limit coverage to 5–10% of the total number of possible interactions. Coverage estimates remain imprecise.

Mass spectrometry methods use engineered proteins with affinity tags to permit immunoprecipitation of entire complexes whose protein components are then

characterized by mass spectrometry. This method has generated large-scale data sets for yeast (Gavin *et al*., 2002; Ho *et al*., 2002). Technical challenges have limited its use for more complex organisms. Other methods, particularly protein chips (Zhu *et al*., 2001), may rise to prominence in generating large-scale interaction data.

Admitted shortcomings of high-throughput proteomic screens are the 90–95% false-negative rates mentioned above, and also a high false-discovery rate (the fraction of reported interactions that are technological artifacts or not biologically relevant). These may arise from spurious or nonspecific interactions in the two-hybrid system or from high-abundance contaminants in mass spectrometry data. Joint analysis of multiple data sets, together with benchmark data from small-scale experiments and well-characterized complexes, suggests that 50% or more of the high-throughput interactions may be spurious (Deane *et al*., 2002; von Mering *et al*., 2002).

Confidence metrics that permit discrimination of true-positive and false-positive interactions have been crucial to the analysis of high-throughput proteomic data. The experimental groups generating high-throughput data have certainly been aware of this issue. Groups working on yeast (Ito *et al*., 2001) and worm (Li *et al*., 2004) use decision trees to classify interactions qualitatively as a highest-confidence core set and one or more lower-confidence categories. The fly group (Giot *et al*., 2003) developed a more sophisticated logistic regression model to generate a numerical score representing the probability that an interaction is biologically relevant. Training sets for this method were generated by cross-species network comparisons to avoid a possible bias from human experts. A similar model was then applied retrospectively to characterize published yeast interaction data (Bader *et al*., 2004). Protein complexes were represented by including all possible within-complex pairs, and a previous finding that bait–hit interactions are higher confidence than hit–hit interactions was reproduced (Bader and Hogue, 2002). Heterogeneous predictors, such as mRNA coexpression (Kemmeren *et al*., 2002), can be incorporated into such models. On-line databases of protein interactions are incorporating confidence measures in their data models (Xenarios *et al*., 2002; Bader *et al*., 2003).

Related methods, roughly equivalent to extrapolating statistical confidence models from observed to unobserved protein pairs, can be used to predict novel interactions. As with the experimental data, these methods report a quantitative probability or likelihood for each predicted interaction. Numerous computationally predicted interaction sets are now available (Goldberg and Roth, 2003; Jansen *et al*., 2003; Schlitt *et al*., 2003; Troyanskaya *et al*., 2003; von Mering *et al*., 2003; Fraser and Marcotte, 2004; Lehner and Fraser 2004).

An appealing computational model for protein network data is a weighted graph, where vertices represent proteins and weighted edges represent confidence levels for observed or predicted physical interactions. This type of model can incorporate information about characterized protein function by defining protein classes and attaching functional class labels to the corresponding vertices. The problem of protein functional inference maps to the classification problem of inferring labels for unlabeled vertices in the weighted network.

This review traces the evolution of network-based algorithms for protein function inference. Algorithms fall into three general classes: nearest-neighbor algorithms, which infer function on the basis of fixed or adaptive thresholds of neighbors,

including tree-building methods; energy minimization methods, which extend nearest-neighbor approaches by iteratively using expected class assignments to transfer annotations over longer distances; and computationally intensive Gibbs sampling that calculated label probabilities by averaging over a Monte Carlo trajectory. Where possible, comparisons of algorithm performance are included to analyze the trade-off between computational cost and predictive power. All the applications discussed below are to inference based on yeast protein interactions.

For some algorithms, a benefit of computational expense may be reduced sensitivity to choice of parameters, with a corresponding reduction in generalization error from highly tuned parameter sets. Algorithms that in principle yield exact answers to model problems but converge slowly do not necessarily perform better than simpler, faster methods based on approximate statistical distributions.

2. Nearest-neighbor algorithms

A nearest-neighbor algorithm satisfies the expectations of guilt-by-association: if an unlabeled protein interacts with proteins labeled primarily for a single class, it probably belongs to the same class. Schwikowski *et al*. (2000) used a nearest-neighbor algorithm based on a network of 2358 interactions among 1548 yeast proteins using 42 overlapping cellular role categories from the Yeast Protein Database. Neighbors were defined as all proteins within 1 interaction link, and the most frequent annotation in the neighbor list was assigned to unlabeled proteins. The accuracy of this method was reported as 72% versus 12% for a randomized network, with new predictions for 364 uncharacterized proteins. The authors detected significant cross-talk, or interactions between proteins belonging to different classes. While some cross-talk may be due to false-positive interactions, a certain amount of cross-talk is inherent due to real, transient interactions between well-defined protein complexes with related, but not identical, function. The cross-talk inherent in biological networks limits the predictive accuracy of methods based purely on nearest neighbors.

Work by Hishigaki *et al*. (2001) extended the nearest-neighbor approach to a greater number of links outward from a central protein. Protein classes were taken from the Yeast Protein Database, with 22 classes related to subcellular localization, 41 classes for biological process, and 57 classes for molecular function. The predictive accuracy of their method was 73%, 64%, and 53% for these three ontologies, respectively. Nearest-neighbor inference was based on the observed class frequencies within a distance of d links of a central protein. Class frequencies in the local neighborhood were converted to a χ^2 statistic based on the frequencies in the entire network, and the class with the highest value of χ^2 was assigned to the central protein. Use of χ^2, rather than raw frequency, was suggested to improve classification accuracy for low-abundance classes. The optimal radius d was determined by best performance for proteins with known classifications. The optimal radius for biological process and subcellular component was 1 link, with a clear degradation of performance for larger neighborhoods. Molecular function showed a slight performance increase from 1 to 2 links, but it is not clear that this increase is statistically significant. Together, these results suggest that

nearest-neighbor methods that do not consider edge weights are limited to ~70% accuracy. This study also introduces a recurring theme that protein interactions are more closely aligned with subcellular component and biological process than with molecular function.

Other measures of distance have been used to define neighborhoods. The PROTDISTIN algorithm of Brun *et al*. uses the Czekanovski–Dice–Sorenson (CDS) measure to calculate pairwise distances d_{ij}^{CD} between proteins,

$$d_{ij}^{\mathrm{CDS}} = \frac{\left|N_i \cup N_j\right| - \left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right| + \left|N_i \cap N_j\right|} \tag{1}$$

where N_i denotes the set of proteins within 1 link of protein i (including i itself) and the absolute value notation indicates set cardinality (Brun *et al*., 2003). The BioNJ algorithm was used to generate a tree for 602 proteins based on the network distance (Gascuel, 1997). The PROTDISTIN algorithm then identifies maximal clades in which an absolute majority of classified proteins share a single class and transfers the class assignment to the other clade members. The distance threshold is clade-specific and hence adaptive. The method was applied to 2946 yeast protein interactions involving 2139 proteins from two-hybrid experiments, including proteins with at least three interactions and excluding highly connected proteins. Cross-talk was observed for interpenetrating clades, which arise from proteins belonging to multiple classes, and for smaller clades embedded in larger clades. The performance of PROTDISTIN was found to be superior to 1-link nearest-neighbor methods and more robust to noise in the input network data. A second tree-building method has been reported using the hypergeomtric distribution as a route to defining distances (Samanta and Liang, 2003). Following the calculation of the hypergeometric p-value for shared neighbors between each pair of proteins, the log of the p-value is used to drive distance-based clustering based on average linkage. While no direct comparison has been reported of these two tree-building methods, other work has shown that the hypergeometric p-value provides superior performance compared to other connectedness measures for inferring single links between proteins with shared interaction partners (Goldberg and Roth, 2003).

Bader's SEEDY algorithm extended the nearest-neighbor approach in two ways: first, edge weights improved the definition of a neighborhood; second, multiple sources permitted classification based on seed proteins of interest (Bader, 2003). The SEEDY method uses a breadth-first search on a weighted graph to identify a list of proteins ranked by their highest scoring path to a seed protein. Path weights are calculated by multiplying the confidence of each edge in the path, a number in the range 0–1. Algorithm performance was measured by selecting known protein complexes, using a subset of protein components as the seeds, and assessing the recall of other members by a receiver operating characteristic (ROC) curve. True-positives rates of 20–60% corresponded to false-discovery rates of 50–80%. Performance using the actual edge weights was demonstrated to improve upon performance with a permuted set of edge weights keeping the network topology fixed.

While the SEEDY method improves on earlier nearest-neighbor methods, it will eventually fail when high-confidence edges percolate the network. At the

percolation transition, virtually every protein will have an edge of confidence $\sim$1, and virtually every pair of proteins will be connected by a path of weight 1. The PRONET algorithm by Asthana *et al*. (2004) introduces probabilistic sampling into the neighborhood calculation. In the PRONET method, the weighted probability assigned to each edge defines a probability distribution on the space of all networks:

$$\Pr\left(\{e_{ij}\}\,\middle|\,\{p_{ij}\}\right) = \prod_{(ij)} p_{ij}^{e_{ij}} (1 - p_{ij})^{1-e_{ij}} \tag{2}$$

where $p_{ij} \in [0, 1]$ is the edge weight and $e_{ij} = 1/0$ denotes the presence/absence of an edge for each pair of proteins (ij). An ensemble of networks is sampled from this distribution, and pairs of proteins are ranked by frequency of co-occurrence in the same connected component. Trials with known complexes were used to optimize the edge weights and to determine an appropriate rank-based threshold for prediction. In practice, the probabilities $\{p_{ij}\}$ are rescaled by a single adjustable parameter $(1 - \alpha)$, where α is interpreted as an additional error rate for the probabilistic edges.

A further generalization of PRONET has been described by Huang *et al*. (2004) as a series of algorithms PROPATH-BIN, PROPATH-ALG, and PROPATH-EXP. The series of PROPATH algorithms use the same network probability space as PRONET. The PROPATH algorithms retain information that is discarded by PRONET, however: the number of links separating pairs of proteins that are located in the same connected component. The distances observed over the ensemble of sampled networks are combined into a single summary statistic. Note that the average distance is ill defined because proteins in separate connected components are at infinite distance. The PROPATH algorithms transform the distances prior to averaging: PROPATH-EXP takes the average of $\exp\left(-\lambda d_{ij}^{(k)}\right)$, and PROPATH-ALG takes the average of $\left(d_{ij}^{(k)}\right)^{-\lambda}$, where $d_{ij}^{(k)}$ is the shortest-path distance between protein pair (ij) in the kth sampled network, and λ is an adjustable parameter representing the inverse length scale of the network neighborhood. The PROPATH-BIN algorithm is the limit of PROPATH-EXP and PROPATH-ALG as $\lambda \to \infty$, which coincides with the PRONET algorithm.

Huang *et al*. performed a head-to-head test of the SEEDY, PRONET, and PROPATH algorithms to recover known protein complexes seeded by a subset of components. These tests led to two conclusions. First, for optimized parameters, all of the algorithms had roughly equivalent performance. This is significant because the SEEDY algorithm does not require sampling of probabilistic networks and is thus $\sim$100 $\times$ faster than the PRONET and PROPATH algorithms. Second, optimized parameters do not necessarily transfer across the methods. In particular, parameters optimized for SEEDY give strikingly poor performance with PRONET. The PROPATH-ALG and PROPATH-EXP methods showed the least sensitivity to parameterization. Results were not sensitive to the choice of the inverse length-scale λ, the choice of algebraic versus exponential decay, or the optimization of the adjustable error-rate parameter required by PRONET. The PROPATH-ALG and PROPATH-EXP algorithms have the same computational cost as PRONET, and

good performance over a range of parameters may provide increased protection against over-fitting and generalization error.

3. Energy minimization and generalized belief propagation

The 1-link nearest-neighbor method is isomorphic to the following energy minimization problem. Let $\{s_i\}$ represent the class assignments for proteins with known classes, and let $\{s_\alpha\}$ represent predicted classes for the remaining proteins. Let $e_{i\alpha} = 1/0$ represent the presence/absence of an interaction edge between proteins i and α. Define the potential energy U as

$$U\left(\{s_\alpha\}\,|\{s_i\}\right) = -\sum_{(i\alpha)} e_{i\alpha}\delta\left(s_i, s_\alpha\right) \tag{3}$$

where $\delta(s,s')$ is the discrete Kronecker delta function, 1 if $s = s'$ and 0 otherwise. The minimum energy is clearly obtained by selecting each s_α to maximize its agreement with its direct neighbors, which is identical to 1-link nearest-neighbor classification.

The energy function may be generalized by introducing terms corresponding all pairs of connected proteins, in particular, proteins without known classification:

$$U\left(\{s_\alpha\}\,|\{s_i\}\right) = -\sum_{(ij)} e_{ij}\delta\left(s_i, s_j\right) - \sum_{(i\alpha)} e_{i\alpha}\delta\left(s_i, s_\alpha\right) - \sum_{(\alpha\alpha')} e_{\alpha\alpha'}\delta\left(s_\alpha, s_{\alpha'}\right) \tag{4}$$

The sums are over distinct pairs of connected vertices, corresponding to a single contribution from each edge. The first sum has fixed value determined by interactions between proteins with prior classifications and is ignored. The second sum corresponds to the 1-link nearest-neighbor energy. The final sum reflects interactions between proteins with newly predicted classifications. This energy function corresponds to a frustrated spin system on a disordered lattice. Equivalently, it is termed a *Markov random field* because the state of each vertex depends only on the states of its immediate neighbors. Global energy minimization is intractable for such systems, and the lowest energy state may be degenerate. Nevertheless, methods such as simulated annealing (Kirkpatrick *et al*., 1983) that avoid the local minima identified by steepest descent methods can used to characterize low-energy states.

Vazquez *et al*. (2003) used simulated annealing with the energy function of equation (4) to predict protein function based on 424 classes Compared to the 1-link nearest-neighbor method, the energy minimization method has 5–10% greater classification accuracy and is more robust to random errors in the interaction network. In a subsequent comparison to the tree-building method, however, the energy minimization fared poorly: PROTDISTIN was able to predict accurate functions for 58% of recently characterized proteins, while energy minimization yielded accurate functions for only 31%. Although this comparison is preliminary, it seems unlikely that energy minimization as described by Vazquez *et al*. will outperform the much faster, deterministic tree-building approach. In related work, Karaoz *et al*. (2004) used local relaxation to quench to a local minimum in the

potential energy surface, with interaction edges corresponding to both protein interactions and associations inferred from mRNA co-expression.

Ignoring for the moment the distinction between classified and unclassified nodes, equation (4) becomes

$$U(\{s_i\}) = -\sum_{(ij)} e_{ij}\delta\left(s_i, s_j\right) - \sum_i h_i(s_i) \tag{5}$$

where $h_i(s_i)$ represents prior knowledge of the class s_i for protein i. An alternative to calculating the configuration $\{s_i\}$ that minimizes the potential energy is to calculate the probability of a configuration according to its Boltzmann weight,

$$\Pr(\{s_i\}) = \frac{\exp\left[-U(\{s_i\})\right]}{\sum_{\{s_i\}} \exp\left[-U(\{s_i\})\right]} \tag{6}$$

and then to obtain expectations for each label as the marginal sums,

$$\Pr(s_i) = \sum_{\{s'\}} \Pr\left(s_i, \{s'\}\right) \tag{7}$$

where $\{s'\}$ includes all labels other than s_i. These types of models gained prominence in image recognition (Geman and Geman, 1984). For acyclic networks, belief propagation provides an efficient shortcut to calculating the exact marginal distributions that is linear in the number of interactions. Belief propagation corresponds to calculating the label distribution that minimizes an approximation to the free energy that includes all 1-vertex and 2-vertex terms, termed the *Bethe approximation*, and can be considered loosely as a mean field theory for vertex pairs. These types of algorithms are not trustworthy for networks with loops because of slow or incorrect convergence. Kikuchi (1951) described higher-order free energy approximations based on larger clusters of vertices, which become exact as the cluster size becomes as large as the largest loop in a network. Yedidia *et al*. demonstrated the equivalence between message passing algorithms and free energy approximation, then showed how Kikuchi cluster expansions lead to generalized message passing algorithms that are readily applied to networks with loops (Yedidia *et al*., 2000; Yedidia *et al*., 2002). Mezard and Parisi (2001) developed similar approaches in the context of spin glasses.

Letovsky and Kasif (2003) recognized that instabilities with belief propagation can also be avoided by limiting the number of iterations to be smaller than the minimum loop size, which is three edges for an undirected graph. They implemented probability updates on the basis of a local Bayesian model for the classification of a protein based on its nearest neighbors, ran two iterations with probability updates for unclassified proteins, then set a threshold to fix the classification of a subset of unclassified proteins. The algorithm was then restarted and run exhaustively until no further labels changed for a network of 13,607 interactions between 4708 proteins. Each of 1951 GO terms was used in turn for classification. While no direct comparison was made with other methods for the

same network, the first probability propagation step should be roughly equivalent to a 2-link nearest-neighbor method. Furthermore, Letovsky and Kasif report that subsequent iterations resulted in ~50% additional unclassified proteins acquiring labels. Thus, this method is likely to have accuracy similar to nearest-neighbor methods but with significantly higher recall.

Leone and Pagnani (2005) instead implemented a full version of message passing at the Bethe approximation level. In careful comparisons on the same underlying network, they find that message passing far outperforms simulated annealing, which in turn outperforms nearest-neighbor methods. Leone and Pagnani incorporate a temperature parameter that is below the critical temperature, leading to stable predictions for class probabilities. Presumably, these probabilities represent Boltzmann-weighted sums over the same low-energy states that simulated annealing is designed to sample. The superior performance of message passing suggests that the simulated annealing method is inefficient in sampling the low-energy states. Thus, the computational efficiency of message passing in generating the exact probability distribution for an approximate free energy function provides an advantage over simulated annealing, which uses the exact free energy but is far slower to converge.

4. Gibbs samplers

An alternative to message passing, which rapidly calculates marginal probability distributions based on an approximate free energy function, is Gibbs sampling, which uses Markov chain Monte Carlo to generate a trajectory whose long-time average should yield exact marginal distributions. Convergence requires ergodicity; on the basis of the slow convergence of simulated annealing, ergodicity should not be assumed, even for thermal distributions.

Deng, Chen, Sun, and coworkers have used Gibbs sampling to perform functional inference based on protein interaction networks (Deng *et al*., 2003) and combined protein interaction/genetic interaction networks (Deng *et al*., 2004). Parameters for the conditional probabilities of the Gibbs sampler are estimated using label correlations for interacting proteins with known classes, which corresponds to a Markov random field and sets a natural temperature scale. While it would be interesting to know whether this temperature is above or below the critical temperature for the network, this aspect of the sampling is not discussed. Performance is superior to nearest-neighbor methods. Comparisons with simulated annealing and message passing methods, however, have not yet been made.

5. Summary and prospects

Performance criteria for functional inference algorithms include computational cost, statistical rigor, and predictive accuracy. Ideally, one would expect that all three criteria increase together, with a trade-off of higher computational cost for increasing rigor and accuracy. Perhaps surprisingly, this is not the case.

The fastest and least sophisticated algorithms, which transfer class predictions from nearest neighbors, indeed have the lowest accuracy. Recall is also limited because these algorithms do not transfer predictions across multiple unclassified proteins. Simple heuristics based on tree building, breadth first search with weighted edges, and sampling over probabilistic networks governed by edge weights, all provide substantial improvement. These algorithms can be interpreted as adapting the spatial range of neighbor sampling to account for network topology and confidence. The probabilistic sampling approaches have moderately increased computational cost, but are significantly less sensitive to precise optimization or tuning of edge weights. The PROPATH-ALG, PROPATH-EXP, and PROTDISTIN algorithms are likely the best of this class; comparisons of these algorithms have yet to be made.

More rigorous algorithms define a probability distribution for protein classes and attempt to predict classes based on this distribution. The form of the probability distribution is a Markov random field or spin lattice, in which the class of a protein depends only on the classes of its interaction partners. Despite the short-range potential energy function, connected series of links generate effective long-range correlations. Simulated annealing attempts to identify the ground state of an exact probability distribution that minimizes clashes between class labels of interacting proteins. It is computationally expensive, and actually performs worse than the much faster PROTDISTIN method. The poor performance is likely due to convergence and inadequate sampling of the relevant low energy states. Message passing belief propagation instead calculates exact marginal distributions for an approximate probability distribution. Message passing methods can be much faster to converge than simulated annealing, and also offer improved performance in reported comparisons. The statistical foundation of the free energy minimization problem underlying message passing offers a theoretical advantage over less rigorous nearest-neighbor methods.

Gibbs sampling methods use Markov chain Monte Carlo to calculate trajectories whose time averages for ergodic systems converge to the correct properties for the correct probability distribution function. Requiring lengthy trajectories and assuming ergodicity are the primary drawbacks of Gibbs sampling. The simulated annealing results on related potential surfaces suggest in particular that trajectories may not be ergodic.

Further improvement in protein functional inference should be realizable by combining features of the methods described. For example, the message passing methods have not incorporated edge weights for protein interactions. Edge weights fit naturally into the Markov random field/spin lattice formalism and have in fact been used to incorporate mRNA co-expression measurements into these models. Advances may also come from improved conceptual models. One limitation of all the models described here is the transitive assumption: if A interacts with B, and B interacts with C, then A and C are functionally linked. This assumption fails for transient connections between protein complexes (Han *et al*., 2004) and also for distinct complexes with a shared set of subunits. Explicit consideration of apparent cross-talk could simplify the protein complex prediction problem, thereby improving functional inference.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1), 25–29.

Asthana S, King OD, Gibbons FD and Roth FP (2004) Predicting protein complex membership using probabilistic network reliability. *Genome Research*, **14**(6), 1170–1175.

Bader GD, Betel D and Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, **31**(1), 248–250.

Bader GD and Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, **20**(10), 991–997.

Bader JS (2003) Greedily building protein networks with confidence. *Bioinformatics*, **19**(15), 1869–1874.

Bader JS, Chaudhuri A, Rothberg JM and Chant J (2004) Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, **22**(1), 78–85.

Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A and Jacq B (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, **5**(1), R6.

Deane CM, Salwinski L, Xenarios I and Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, **1**(5), 349–356.

Deng M, Tu Z, Sun F and Chen T (2004) Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, **20**(6), 895–902.

Deng M, Zhang K, Mehta S, Chen T and Sun F (2003) Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, **10**(6), 947–960.

Fraser AG and Marcotte EM (2004) A probabilistic view of gene function. *Nature Genetics*, **36**(6), 559–564.

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**(7), 685–695.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions in Pattern Analysis and Machine Learning*, **6**, 721–741.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A Protein Interaction Map of Drosophila melanogaster. *Science*, **302**(5651), 1727–1736.

Goldberg DS and Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(8), 4372–4376.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995), 88–93.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261.

Hishigaki H, Nakai K, Ono T, Tanigami A and Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**(6), 523–531.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**(6868), 180–183.

Huang H, Zhang LV, Roth FP and Bader JS (2004) Probabilistic paths in protein interaction networks. *In review*.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF and Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**(5644), 449–453.

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR and Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(9), 2888–2893.

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A and Holstege FC (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, **9**(5), 1133–1143.

Kikuchi R (1951) A theory of cooperative phenomena. *Physical Review*, **81**, 988–1003.

Kirkpatrick S, Gelatt CD and Vecchi MP (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Lehner B and Fraser AG (2004) A first-draft human protein-interaction map. *Genome Biology*, **5**(9), R63.

Leone M and Pagnani A (2005) Predicting protein functions with message passing algorithms. *Bioinformatics*. **21**, 239–247.

Letovsky S and Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19**(Suppl 1), i197–i204.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**(5657), 540–543.

Mezard M and Parisi G (2001) The Bethe lattice spin glass revisited. *The European Physical Journal B*, **20**, 217–233.

Samanta MP and Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(22), 12579–12583.

Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E and Brazma A (2003) From gene networks to gene function. *Genome Research*, **13**(12), 2568–2576.

Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**(12), 1257–1261.

Troyanskaya OG, Dolinski K, Owen AB, Altman RB and Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, **100**(14), 8348–8353.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Vazquez A, Flammini A, Maritan A and Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, **21**(6), 697–700.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, **31**(1), 258–261.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, **30**(1), 303–305.

Yedidia JS, Freeman WT and Weiss Y (2000) Generalized belief propagation. *MERL Technical Report*, **TR-2000-26**, 1–9.

Yedidia JS, Freeman WT and Weiss Y (2002) Understanding belief propagation and its generalizations. *MERL Technical Report*, **TR2001-022**, 1–35.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**(5537), 2101–2105.

Specialist Review

Constraint-based modeling of metabolomic systems

Daniel A. Beard
Medical College of Wisconsin, Milwaukee, WI, USA

Hong Qian
University of Washington, Seattle, WA, USA

1. Introduction

Since the time of Newton, a central dogma of scientific thinking has been that physical systems are best understood by representation in terms of the smallest possible subsystem (i.e., model) that captures the important mechanistic interactions. The influence of gravity in maintaining the earth's orbit about the sun is satisfactorily explained by analyzing the equations of motion representing a universe consisting of two massive bodies. Still, a complete mathematical analysis of the three-body problem remains out of reach, and most quantitative analyses of practical engineering problems, such as the analysis of the material properties of metal composites and the building of aircraft, are based on a combination of mechanistic models with intermediate-level empirical descriptions in quantitative terms. This approach is typical of engineering solutions to complex problems in physical sciences.

Living biological systems consist of not two, or even two hundred interacting components. Analysis, prediction, and rational manipulation of cellular function requires a mechanistic understanding of physical systems of unimaginable complexity. These are truly *complex systems*. Furthermore, each biological system is geared toward a unique function (or functions), a concept usually lacking in the physical sciences but central to engineering (Hopfield, 1994).

Two other disciplines that have tackled complexities similar to those encountered in Systems Biology are economics and ecology. It is arguable that mathematical thinking, language, and modeling are well-accepted components of these two fields. In addition, in both fields there is a clear distinction between the so-called micro- and macroscopic views. The microscopic view focuses on the small-scale mechanistic components of a system, while the macroscopic view treats the systems as a whole. As in biology, both economics and ecology have yet to establish a quantitative framework connecting the two viewpoints.

While it may seem natural to follow the path paved by Newton in constructing a differential equation–based representation for biological systems, we should not

treat uncritically the assumption that a useful representation of this form always exists. Popular mythology aside, Newton stood on no one's shoulders. Similarly, we do not require Systems Biology to stand on his alone.

Clearly, certain aspects of cellular physiology are successfully quantified via a traditional kinetic differential equation–based theory. A classical example is the electrophysiology of excitable cells, a field pioneered by Hodgkin and Huxley (1952). On the basis of their foundational description, models of ever-increasing complexity have been introduced, leading toward more complete computational models of cellular electrophysiology (Noble, 2004; *see also* Article 115, **Systems biology of the heart**, Volume 6). However, it is not immediately clear that differential equations will provide a universal framework for integrating these models with cellular mechanics, metabolism, genetics, and other phenomena. It seems more likely that no such universal mathematical framework will exist for Systems Biology.

In the past several decades, progress in cell biology has been driven mainly by the philosophies of and methodology from biochemistry and molecular genetics (Lazebnik, 2002). Biochemistry emphasizes individual reactions and their chemical nature: what are the reactants and products; which enzyme is involved; and what is the protein structure and enzyme mechanism? Molecular genetics, on the other hand, focuses on the function and/or malfunction of molecular components and on information processing in living cells. The traditional modeling approach in biochemistry is differential equation–based enzyme kinetics. Modeling studies of one or a few enzymatic reactions at a time have been successfully applied to in vitro kinetics. It remains to be demonstrated, however, that this approach can be scaled up to an in vivo system of hundreds of reactions and species with thousands of parameters (Teusink *et al*., 2000). More importantly, it is not clear that this is the ideal approach for integrating biochemistry with molecular genetics.

The constraint-based modeling (CBM) approach circumvents several of these difficulties. It facilitates integration of experimental data of disparate types and from disparate sources while increasing the accuracy in its prediction. It does not require a priori knowledge (or assumptions) regarding all of the mechanisms and parameters for a given system. However, when a priori knowledge exists, such as data on enzyme kinetics or measurements of in situ concentrations and fluxes, this knowledge can be introduced in the form of constraints, on equal footing with molecular genetic observations on the topology and information flow in a biochemical network. While the differential equation–based modeling provides predictions with "high information content and high false rate", the CBM approach, by design, provides a low level of false prediction, but can often give little information in its prediction. Thus, we believe that these two approaches will ultimately prove to be complementary.

It is the aim of this chapter to review this alternative, constraint-based approach to modeling biochemical systems (Reed and Palsson, 2003). We shall focus on modeling of cellular metabolic networks, provide a comprehensive description of the mathematical and theoretical basis for CBM of metabolic systems. The material has been organized with the objective of a convenient exposition, rather than as an

historical account. Wherever possible, we point to detailed reviews on specialized topics.

2. Biochemical variables and constraints on fluxes

For metabolic systems, the variables of interest are the concentrations of biochemical species (c_i), the fluxes of the biochemical reactions (J_j), and the activities of the associated enzymes (X_j). Furthermore, metabolic fluxes are responsible for maintaining the homeostatic state of the cell, a condition that translates into the assumption that the metabolic network functions in or near a steady state, that is, all of the concentrations are treated as constant in time.

2.1. The mass-balance constraint

In a metabolic steady state, the biochemical fluxes are balanced to maintain constant concentrations of all internal metabolic species. If the stoichiometry of a system made up of M species and N fluxes is known, then the stoichiometric numbers can be systematically tabulated in an $M \times N$ matrix, known as *the stoichiometric matrix* $\widetilde{\mathbf{S}}$ (Clarke, 1980). The S_{ij} entries of the stoichiometric matrix are determined by the stoichiometric numbers appearing in the reactions in the network. For example, if the jth reaction has the form:

$$S_{1j}^{L} A_1 + \cdots + S_{Mj}^{L} A_M \rightleftharpoons S_{1j}^{R} A_1 + \cdots + S_{Mj}^{R} A_M \tag{1}$$

where A_i represents the ith species, then the stoichiometric matrix has the form $S_{ij} = S_{ij}^{R} - S_{ij}^{L}$.

The fundamental law of conservation of mass dictates that the vector of steady state fluxes, $\mathbf{J}$, satisfies

$$\widetilde{\mathbf{S}}\mathbf{J} = \mathbf{b} \tag{2}$$

where $\mathbf{b}$ is the vector of boundary fluxes that transport material into and out of the system. If the values of the boundary fluxes are not known, equation (2) can be written as $\mathbf{SJ} = \mathbf{0}$ in which the boundary fluxes have been incorporated into the $M \times (N + N')$ matrix $\mathbf{S}$, where N' is the number of boundary fluxes (Qian *et al*., 2003). As an example, consider a simple network of three unimolecular reactions:

$$A \overset{J_1}{\rightleftharpoons} B, \quad B \overset{J_2}{\rightleftharpoons} C, \quad C \overset{J_3}{\rightleftharpoons} A \tag{3}$$

where all reactions are treated as reversible; left-to-right flux is the direction defined as positive. The network of equation (3) is represented by the stoichiometric matrix:

$$\widetilde{\mathbf{S}} = \begin{matrix} A \\ B \\ C \end{matrix} \begin{bmatrix} -1 & 0 & +1 \\ +1 & -1 & 0 \\ 0 & +1 & -1 \end{bmatrix} \tag{4}$$

Next, consider that species A is transported into the system at rate b_A and species B is transported out at rate b_B. Then the mass-balance equations $\widetilde{\mathbf{S}}\mathbf{J} = \mathbf{b}$ can be expressed:

$$\begin{bmatrix} -1 & 0 & +1 \\ +1 & -1 & 0 \\ 0 & +1 & -1 \end{bmatrix} \begin{bmatrix} J_1 \\ J_2 \\ J_3 \end{bmatrix} = \begin{bmatrix} -b_A \\ +b_B \\ 0 \end{bmatrix} \tag{5}$$

Algebraic analysis of this equation reveals that mass-balanced solutions exist if and only if $b_A = b_B$. Equation (5) can be simplified to $J_2 = J_3 = J_1 - b_A$. Thus, mass balance does not provide unique values for the internal reaction fluxes. In fact, for this example, solutions exist for

$$J_1 \in (-\infty, +\infty), \quad J_2 \in (-\infty, +\infty), \quad J_3 \in (-\infty, +\infty) \tag{6}$$

Equation (6) illustrates the fact that often the mass-balance constraint poses an underdetermined problem; typically, it is necessary to identify additional constraints and/or to formulate a model objective function to arrive at meaningful estimates for biochemical fluxes.

2.2. *Thermodynamic constraints*

In addition to the stoichiometric mass-balance constraint, constraints on reaction fluxes and species concentration arise from nonequilibrium steady state biochemical thermodynamics (Hill, 1977). For a biochemical reaction at a constant temperature T, such as

$$A + B \rightleftharpoons C + D \tag{7}$$

there exists a forward flux J_+ and a backward flux J_-, with net flux $J = J_+ - J_-$. The concentrations of the reactants and products in this reaction are related to the chemical potential of the reaction via (Qian, 2001)

$$\Delta\mu = \mu_C + \mu_D - \mu_A - \mu_B = k_B \mathrm{T} \ln(J_-/J_+) \tag{8}$$

where k_B is the Boltzmann constant. We see from equation (8) that J and $\Delta\mu$ always have opposite signs, and are both zero only when a reaction is in equilibrium.

For a system of reactions, let $\Delta\mu$ be the vector that contains potential differences for all the reactions. The magnitude of the product $-\mathbf{J}\Delta\mu$ is the amount of heat the reactions dissipate per unit time. The negative sign reflects the Second Law of Thermodynamics according to Lord Kelvin: One cannot convert 100% of heat energy into useful work.

The fundamental law of conservation of energy (First Law of Thermodynamics) dictates that the heat dissipated by the reactions is supplied by constant feeding

of the *open* biochemical system. We have shown that this statement can be mathematically expressed as

$$\boldsymbol{\mu}\mathbf{J} = \boldsymbol{\Delta\mu}\widetilde{\mathbf{S}}\mathbf{J} = \boldsymbol{\mu}\mathbf{b} \tag{9}$$

where the right-hand side of equation (9) is the amount of energy supplied to the system through boundary fluxes (Qian *et al.*, 2003). If the system is closed by removing the boundary fluxes, Equation (9) becomes

$$\boldsymbol{\Delta\mu}\mathbf{K} = \mathbf{0} \tag{10}$$

where the matrix $\mathbf{K}$ contains a basis for the right null space of $\widetilde{\mathbf{S}}$. This relation reflects *energy balance*, a generalization of Kirchhoff's loop law and the Tellegen's theorem in electrical circuits (Beard *et al.*, 2002).

For the example of equation (3), equation (10) is expressed as:

$$[\Delta\mu_1\ \Delta\mu_2\ \Delta\mu_3]\begin{bmatrix}1\\1\\1\end{bmatrix} = 0 \tag{11}$$

Here, the matrix $\widetilde{\mathbf{S}}$ of equation (4), has a one-dimensional null space, for which the vector $[1\ 1\ 1]^T$ is a basis. Equation (11) corresponds to summing the reaction potentials about the closed loop formed by the reactions in Equation (3).

The requirement that J and $\Delta\mu$ for each reaction in a network have opposite signs provides a stringent constraint on the vectors $\mathbf{J}$ and $\Delta\mu$. Each vector $\mathbf{J}$ that satisfies mass balance $\widetilde{\mathbf{S}}\mathbf{J} = \mathbf{b}$ must also satisfy the following condition in order to be thermodynamically feasible. The system of inequalities

$$\sum_i \mu_i \widetilde{S}_{ij} J_j \leq 0,\ \ (j = 1, 2, \ldots, N) \tag{12}$$

must have at least one nonzero solution for $\{\mu_i\}$ (Qian *et al.*, 2003; Beard *et al.*, 2004).

Under this constraint, the bounds on the fluxes of equation (5) are narrowed from those of equation (6) to:

$$J_1 \in (0, b_A), \quad J_2 \in (0, b_A), \quad J_3 \in (0, b_A) \tag{13}$$

Thus, in general, the thermodynamic constraint narrows the feasible flux space, but not necessarily to a unique solution. Knowledge of the boundary fluxes translates into constraints on the reaction directions. Thus, the feasible reaction directions are a function of an open system's (i.e., a cell's) interaction with its environment.

Prior to the introduction to the generalized constraint-based thermodynamic theory outlined above, applications of constraint-based modeling of metabolic systems relied on an empirical set of irreversibilities that were obtained from prior observations of reaction directions in physiological settings. By treating certain reactions as

implicitly unidirectional, biologically reasonable results can often be obtained without considering the system thermodynamics as outlined above. Since the system-level thermodynamic constraint is inherently nonlinear, in current and future application of constraint-based modeling, it may be practical to implement unidirectional (irreversibility) constraints on specific reactions in large-scale network models.

3. Biochemical concentrations and enzyme activities

Combining mass balance with the thermodynamic constraint, one can determine the feasible ranges for the biochemical concentrations and the levels of enzyme activities.

3.1. Feasible biochemical concentrations from potentials

Introducing the chemical potential and the energy balance equation provides a solid physical chemistry foundation for the CBM approach to metabolic systems analysis. Proper treatment of the network thermodynamics not only improves the accuracy of the predictions on the steady state fluxes, but can also be used to make predictions on the steady state concentrations of metabolites. To see this, we substitute the well-known relation between the chemical potential μ and the concentration c of a biochemical species, $\mu = \mu_o + k_B \mathrm{T} \ln c$ into the equation (12):

$$\sum_i \ln c_i \widetilde{S}_{ij} J_j \leq -\Delta\mu_j^o \tag{14}$$

where $\Delta\mu_j^o$ is the standard state chemical potential difference for reaction j, which may be obtained from a standard chemical reference source. There is also a concerted effort to establish a central database for this information at the National Institute of Standard and Technology (NIST). If a solution for μ_i exists for equation (12), then there exists a set of corresponding concentrations c_i. In fact, equation (14) provides a feasible space for the metabolites concentrations as a convex cone in the log-concentration space. If the set of feasible concentrations is empty, then the vector $\mathbf{J} = \{J_j\}$ is thermodynamically infeasible.

3.2. Biochemical conductance and enzyme activity

From the traditional biochemical kinetics standpoint, both steady state biochemical concentrations and reaction fluxes are predictable from known enzyme reactions with appropriate rate constants and initial conditions. In steady state, the fluxes are computable from the concentrations of the reactants and products. However, a realistic challenge we confront is that our current understanding of the reaction mechanisms and measurements of rate constants are significantly deficient. From the standpoint of CBM, however, the ratio between the J_k and $\Delta\mu_k$ of a particular reaction is analogous to the conductance, which can be shown to be proportional

to the enzyme activity of the corresponding reaction. We emphasize that the magnitudes of both J_k and $\Delta\mu_k$ are functions of the reaction networks topology. Therefore, each one alone will not be sufficiently informative of the level of enzyme activity (i.e., the level activity due to gene expression or posttranslational modification).

3.3. Conserved metabolic pools

In addition to the constraint on concentrations imposed by equation (14), a reaction network's stoichiometry imposes a set of constraints on certain conserved concentration pools (Alberty, 1991). These constraints follow from the equation for the kinetic evolution of the metabolite concentration vector:

$$\mathbf{P}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \mathbf{SJ} \tag{15}$$

where $\mathbf{P}$ is a diagonal matrix, with diagonal entries corresponding to the partition coefficients, or fractional intracellular spaces, associated with each metabolite in the system. In equation (15), columns corresponding to the boundary fluxes have been grouped into the matrix $\mathbf{S}$. Here, the vector $\mathbf{J}$ includes both internal reaction fluxes and boundary fluxes. The left null space of the matrix $\mathbf{P}^{-1}\mathbf{S}$ may be computed and bases for this space stored in a matrix $\mathbf{L}$, such that:

$$\mathbf{L}\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \mathbf{LP}^{-1}\mathbf{SJ} = \mathbf{0} \tag{16}$$

It follows from equation (16) the product $\mathbf{Lc}$ remains constant and defines a number of conserved pools of metabolic concentrations. For example, if we were to consider the glycolytic series as an isolated system, with no net flux of phosphate-containing metabolites into or out of the system, then as phosphate is shuttled among the various metabolites, the total amount of phosphate in the system is conserved.

3.4. Incorporating metabolic control analysis

One of the theoretical frameworks in quantitative analysis of metabolic networks is metabolic control analysis (Westerhoff and van Dam, 1987; Heinrich and Schuster, 1996). In metabolic control analysis, the enzyme elasticity coefficients provide empirical constraints between the metabolites concentrations and the reaction fluxes. These constraints can be considered in concert with the interdependencies in the $\mathbf{J}$ and $\mathbf{c}$ spaces that are imposed by the network stoichiometry. If the coefficients $\epsilon_k^i = (c_k/J_i)\partial J_i/\partial c_k$ are known, then these values bind the fluxes and concentrations to a hyperplane in the $\mathbf{J}$, $\mathbf{c}$ space. In addition, the constraint-based approach finds applications (Beard *et al*., 2003) in dynamic metabolic control

analysis, which derives flux control coefficients from linear relaxation kinetics in response to perturbations of a steady state (Reder, 1988; Liao and Delgado, 1993).

4. Optimization theory and objective functions

Constraint-based metabolic analysis is intimately tied to the mathematical field of optimization. Readers may find an accessible introduction to optimization theory, which represents a mature field of modern applied mathematics, in Strang's *Introduction to Applied Mathematics* (Strang, 1986). Optimization theory has been the major mathematical engine behind bioinformatics and genomic analysis (Waterman, 1995). Constraint-based approaches have also been very successful in biological structural modeling ranging from distance geometry calculations for protein structure prediction from NMR (*see* Article 105, **NMR**, Volume 6) to the recent structural determination of large macromolecular complexes (Alber *et al*., 2004). For the purposes of this chapter, it is sufficient to be familiar with the following basic concepts of optimization theory.

Mathematical optimization deals with determining values for a set of unknown variables $x_1, x_2, \ldots, x_n$, which best satisfy (optimize) some mathematical objective quantified by a scalar function of the unknown variables, $F(x_1, x_2, \ldots, x_n)$. The function F is termed the *objective function*; bounds on the variables, along with mathematical dependencies between them, are termed *constraints*. Constraint-based analysis of metabolic systems requires definition of the constraints acting on biochemical variables (fluxes, concentrations, enzyme activities) and determining appropriate objective functions useful in determining the behavior of metabolic systems.

Therefore, in CBM, the objective functions play a crucial role. A given objective function can be thought of as a mathematical formulation of a working hypothesis for the function of particular cell or cellular system. These objective functions should not be considered to be as theoretically sound as the physiochemical constraints; but they may be informative and biologically relevant. They can serve as concrete statements about biological functions and powerful tools for quantitative predictions, which must be checked against experimental measurements. One of the surprising discoveries in constraint-based modeling is how well certain simple objective functions have described biological function (see below).

That a cell functions precisely following some rule of optimality (such as optimal energetic efficiency for optimal cell growth rate) is, of course, highly suspect. There may be an evolutionary argument in favor of certain objective functions, but the ultimate justifications lie in the correctness of its predictions. In this sense, the constraint-based optimization approach provides a convenient way to efficiently generate quantitative predictions of biological hypotheses formulated in terms of objective functions. The value of this approach is in facilitating the systematic prediction–experimental verification–hypothesis modification cycle, ideally leading to new discoveries.

The aim of metabolic engineering is different from that of the traditional biological research (Bailey, 1991; Stephanopoulos, 1994). In metabolic engineering, one is more interested in the "capacity" and optimal behavior of a "biological

hardware" (Edwards and Palsson, 2000; van Dien and Lidstrom, 2002) rather than its natural function per se. For this reason, the metabolic engineering community has been a major proponent of the constraint-based optimization approach, while the cell biologists, whose traditional inclination follow reductionism, view the approach with certain healthy skepticism. Hopefully, with the introduction of thermodynamics, and the establishment of physiochemical basis of the CBM, the two communities will join force in furthering the research on Systems Biology of cells (Hartwell *et al*., 1999).

5. Applications of constraint-based modeling

For bacterial cells, growth rate (rate of biomass production) has been a widely used objective function. This objective is constructed as a net flux out of the cell of the components of biomass (amino acids, nucleotides, etc.) in their proper stoichiometric ratios, which translates into a linear function of the reaction fluxes. On the basis of this elegant paradigm, predictions from FBA of the fate of the *E. coli* MG1655 cell following deletions of specific genes for central metabolic enzymes have been remarkably accurate (Edwards and Palsson, 2000). When combined with energy balance analysis (EBA), it has been shown (Beard *et al*., 2002) that cells with nonessential genes deleted can redirect the metabolic fluxes under relatively constant enzyme activity levels, with few changes due to gene expression regulation and/or posttranslational regulation. The FBA/EBA combined approach predicts which enzyme must be upregulated, which must be down regulated, and which reactions must be reversed given perturbations to the genotype and/or cellular environment. Using this combined approach, a clear relation is established between the enzyme regulation and constraint-based analysis of metabolism.

Different objective functions can be used in studying other biological systems and problems. When addressing cellular metabolic pathway regulation, robustness has proven to be a useful concept. Robustness can be defined as minimal changes in metabolic fluxes, in steady state concentration, or in enzymes activities following perturbations. For example, minimization of flux adjustment has been used to model the metabolic response of *E. coli* JM101 with pyruvate kinase knockout (Segrè *et al*., 2002). In this study, it is assumed that the cell acts to maintain its wild-type flux pattern in response to the challenge imposed by a gene knockout. However, a minimal change in the flux pattern may require an unrealistic level of metabolic control. We have shown (Beard and Qian, 2005) that the objective of minimal changes in enzyme activities predicts the key regulatory sites in switching between glycogenic and gluconeogenic operating modes in hepatocytes. This approach facilitates inverse analyses, where the regulatory system is treated as a black box and control mechanisms are identified from measurements of the inputs and outputs to the system.

6. Summary

In summary, the constraint-based approach to the analysis of metabolic systems is based on a set of constraints that are imposed by the fundamental laws of mass

conservation and thermodynamics. These laws impose mathematical dependencies on the feasible fluxes, potentials, and concentrations in a given metabolic system. Biological hypotheses may be formulated as mathematical objectives that metabolic systems optimize under the imposed constraints. These hypotheses (e.g., optimal growth or optimally efficient control) are accepted, rejected, or revised on the basis of comparisons between the model predictions and experimental measurements.

Further reading

Schilling CH, Letscher D and Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal Theoretical Biology*, **203**, 229–248.

Schilling CH, Schuster S, Palsson BO and Heinrich R (1999) Metabolic pathway analysis: basic concepts and scientific applications in post-genomic era. *Biotechnology Progress*, **15**, 296–303.

Schuster S, Dandekar T and Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, **17**, 53–60.

Schuster S and Hilgetag C (1994) On elementary flux modes in biochemical reaction systems at steady-state. *Journal of Biological Systems*, **2**, 165–182.

References

Alber F, Eswar N and Sali A (2004) Structure determination of macromolecular complexes by experiment and computation. In *Practical Bioinformatics*, *Nucleic Acids and Molecular Biology Series Vol 15*, Bujnicki J (Ed.). Springer-Verlag: Berlin, 73–96.

Alberty RA (1991) Equilibrium compositions of solutions of biochemical species and heat of biochemical reactions. *Proceedings of the National Academy of Science United States of America*, **88**, 3268–3271.

Bailey JE (1991) Toward a science of metabolic engineering. *Science*, **252**, 1668-1674.

Beard DA and Qian H (2005) Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism. *American Journal of Physiology: Endocrinology and Metabolism*, **288**, E633–E644.

Beard DA, Liang SD and Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophysical Journal*, **83**, 79–86.

Beard DA, Qian H and Bassingthwaighte JB (2003) Stoichiometric foundation of large-scale biochemical system analysis. In *Modeling in Molecular Biology*, Ciobanu GM (Ed.), Springer: New York.

Beard DA, Babson E, Curtis E and Qian H (2004) Thermodynamic constraints for biochemical networks. *Journal of Theoretical Biology*, **228**, 327–333.

Clarke BL (1980) Stability of complex reaction networks. *Advances in Chemical Physics*, **43**, 1–215.

Edwards JS and Palsson BO (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capacities. *Proceedings of National Academy of Science United States of America*, **97**, 5528–5533.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Heinrich R and Schuster S (1996) *The Regulation of Cellular Systems*, Chapman & Hall: New York.

Hill TL (1977) *Free Energy Transduction in Biology: The Steady-State Kinetic and Thermodynamic Formalism*, Academic Press: New York.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**, 500–544.

Hopfield JJ (1994) Physics, computation, and why biology looks so different. *Journal of Theoretical Biology*, **171**, 53–60.

Lazebnik Y (2002) Can a biologist fix a radio?-Or, what I learned while studying apoptosis. *Cancer Cell*, **2**, 179–182.

Liao JC and Delgado J (1993) Advances in metabolic control analysis. *Biotechnology Progress*, **9**, 221–233.

Noble D (2004) Modeling the heart. *Physiology*, **19**, 191–197.

Qian H (2001) Nonequilibrium steady-state circulation and heat dissipation functional. *Physical Review E*, **64**, 022101.

Qian H, Beard DA and Liang SD (2003) Stoichiometric network theory for nonequilibrium biochemical systems. *European Journal of Biochemistry*, **270**, 415–421.

Reed JL and Palsson BO (2003) Thirteen years of building constraint-based in silico models of *Escherichia coli*. *Journal of Bacteriology*, **185**, 2692–2699.

Reder C (1988) Metabolic control theory: a structural approach. *Journal of Theoretical Biology*, **135**, 175–201.

Segrè D, Vitkup D and Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of National Academy of Science United States of America*, **99**, 15112–15117.

Stephanopoulos G (1994) Metabolic engineering. *Current Opinion in Biotechnology*, **5**, 196–200.

Strang G (1986) *Introduction to Applied Mathematics*, Wellesley-Cambridge Press: Wellesley.

Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, *et al*. (2000) Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, **267**, 5313–5329.

van Dien SJ and Lidstrom ME (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C_3 and C_4 metabolism. *Biotechnology and Bioengineering*, **78**, 296–312.

Waterman MS (1995) *Introduction to Computational Biology*, Chapman & Hall: New York.

Westerhoff HV and van Dam K (1987) *Thermodynamics and Control of Biological Free-Energy Transduction*, Elsevier: Amsterdam.

Specialist Review

Metabolic dynamics in cells viewed as multilayered, distributed, mass-energy-information networks

Miguel A. Aon and Sonia Cortassa
The Johns Hopkins University, Baltimore, MD, USA

1. Introduction

The biological sciences are in the midst of a transition between an *analytical* and an *integrative* period. During the analytical period, researchers have unraveled the main macromolecular components of cellular machinery including the cytoskeleton, entire routes of intermediary metabolism, the circuitry involved in the expression of genetic information, and signal transduction pathways. This has generated significant amounts of information, most of which is contained in databases. The advent of functional genomics, following the sequencing of several genomes including human, is now producing an increased understanding of the functions of gene products at several levels (named as genome, transcriptome, proteome, and metabolome) (Oliver, 2000). Now we are challenged to address the question how a cell works during its main life stages, that is, growth, division, differentiation, and death. We would like to know how the mass-energy transformation and -information (genetic and signaling)-carrying networks in the cell interact with each other to produce cellular responses under defined physiological conditions. This view of the cell as a multilayered, distributed, mass-energy-information network (Figure 1) emphasizes that information processes pervade all levels from genome to metabolome. Furthermore, it points out where the main gap in our knowledge stands and why our efforts should be directed to answer questions such as the way in which cellular phenotype dynamically “emerges” from the mass-energy-information carrying networks of the cell.

2. Mathematical models: criteria of reliability

Mathematical models are now playing a major role at integrating information across the molecular, cellular, tissue and organ levels, in order to improve our

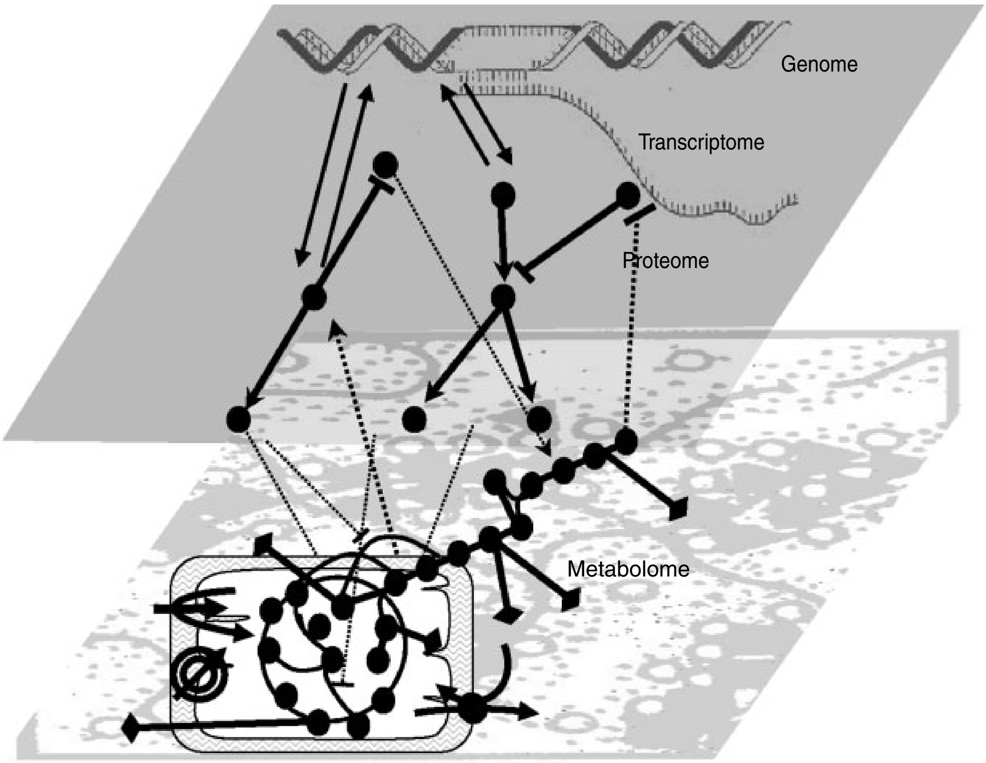


Figure 1 A view of cells as multilayered mass-energy-information networks of reactions. Metabolic reactions embedded in the cytoplasmic scaffolds are shown diagrammatically with each chemical species represented by a filled circle (bottom layer). Central catabolic pathways (glycolysis, TCA cycle) and mitochondria are sketched. As an example, mitochondria emphasize the interaction between subcellular compartments through transport reactions and exchange of metabolites or intracellular messengers that take part of information-carrying networks. The chemical reaction network for the synthesis of a microorganism is a complex one comprising around 1000 chemical reactions (Alberts *et al*., 1989). A typical mammalian cell synthesizes more than 10 000 different proteins, a major proportion of which are enzymes that carry out the mass-energy transformations represented in the bottom-layer networks. The information processing function of cells is carried out by other networks, that is, signaling pathways and the regulatory circuitry related with gene expression (genome, represented in the top layer), intertwined with the network depicted in the bottom layer. The information-carrying networks are shown in a different layer just for the sake of clarity and its presence on top does not imply hierarchy. On the contrary, as emphasized by arrows, cross-talk connections exist between all layers (activating, with arrowheads, or inhibitory, with a crossed line). Nevertheless, each network has its own type and mechanisms of interactions, for example, the nodes on the top layer are proteins or second messengers. The proteins (proteome) are continuously synthesized from mRNA (transcriptome) through transcription and translation mechanisms of gene expression, and exert feedback regulation (e.g., DNA-binding proteins) on its own, or the expression of other proteins (depicted as double arrows on top layer). Moreover, proteins taking part of the cytoskeleton either exert feedback regulation on its own expression (e.g., modulation of intracellular levels of tubulin through the stability of mRNAs) (Cleveland, 1988), or influence metabolic fluxes through epigenetic mechanisms (e.g., (de)polymerization, dynamic instability) (Reproduced from Cortassa *et al*. (2002) by permission of World Scientific Publishing)

understanding of biological function in general and specific processes operating in vivo in a defined context. Reliability of the integrative power of mathematical models is reflected in the following main properties: (1) whether or not there is a sound physico-(bio)chemical basis for the models; (2) the ability of the model to reproduce reported experimental evidence; (3) the ability of the model to provide insights into biophysical/biochemical processes (explanatory power); and (4) ability of the model to predict novel behaviors.

Figures 2 and 3 show several of the tests according to criteria (1–4) mentioned above. An integrative model of mitochondrial energetics, calcium dynamics, and reactive oxygen species (ROS) production and scavenging (Cortassa *et al*., 2003; Cortassa *et al*., 2004) serve to illustrate these four criteria:

1. *Sound physico-(bio)chemical basis*: Stimulation of respiration by ADP has been proposed as one of the mechanisms by which mitochondrial ATP production meets the cellular ATP demand (Chance and Williams, 1956; Harris and Das, 1991). The integrated model simulates the expected relationship between mitochondrial membrane potential, $\Delta\Psi_m$, and respiration upon a transition between states 4 (absence of ADP that leads to null ATP synthesis and low respiration) and 3 (presence of ADP leading to maximal ATP synthesis and high respiration).
2. *Ability to reproduce reported experimental evidence*: The integrated model of mitochondrial energetics is capable of simulating time-dependent behavior under two relevant physiological situations: (1) NADH and Ca^{2+} dynamics in mitochondria of heart trabeculae subjected to changes in workload (Brandes and Bers, 2002) (Figure 2a–c); (2) the dynamics of respiration following a shift between states 4 and 3 after a pulse of ADP (Boveris *et al*., 1999) (Figure 3c).

   Recent development of the integrated model of mitochondrial energetics and Ca^{2+} dynamics includes mitochondrial ROS production, cytoplasmic ROS scavenging, and ROS activation of inner membrane anion flux. This upgraded model is able to simulate the period and phase of mitochondrial oscillations in $\Delta\Psi_m$, NADH, and ROS (Cortassa *et al*., 2004) in agreement with experimental data obtained in heart cells under oxidative stress (Aon *et al*., 2003).
3. *Explanatory power*: The participation of various regulatory interactions was made clear in the analysis of the time-dependent behavior of Ca^{2+} and NADH in heart trabeculae during changes in workload. Model simulations mimicking the participation of just one of the regulatory mechanisms established that: (1) the recovery phase of NADH level after a sudden increase in pacing frequency is due to the stimulatory effect of Ca^{2+} on the tricarboxylic acid (TCA) cycle dehydrogenases and (2) the fast response of the NADH under(over)shoot following an increase (decrease) in pacing frequency is related to the stimulation of respiration by changes in mitochondrial ADP. Thus, the model results suggest a coparticipation of Ca^{2+} and ADP to explain mitochondrial NADH dynamics during an increase in workload (Figures 2 and 3).
4. *Predictive power*: The computational mitochondrial model has made several predictions, which have been verified experimentally. These predictions include: (1) modulation of the period of the mitochondrial oscillator described in heart cells by altering the concentration of ROS scavengers or the rate of oxidative

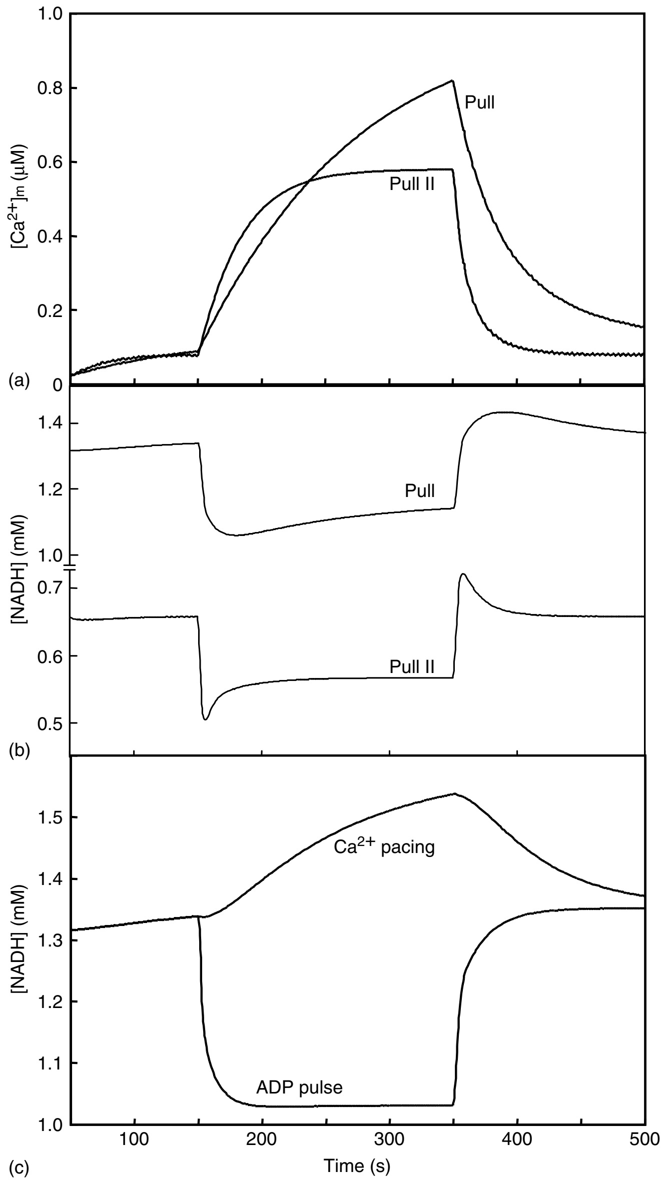


Figure 2 Time-dependent behavior of mitochondrial Ca^{2+} and NADH after changes in workload according to an integrated model of mitochondrial energetics. Model response of mitochondrial Ca^{2+} (a) to the pacing protocol applied at low frequency (0.25 Hz) for 100 s, and then a high frequency (2 Hz) for 200 s, followed by a return to the low frequency. (b) Time course of NADH levels during changes in pacing frequency, simulated by increasing the Ca^{2+} pulse frequency combined with an increase in ADPi. (c) The effects of a change in Ca^{2+} pulse frequency or ADPi applied separately. The simulations reproduce the experimental data reported by Brandes and Bers (2002) for rat cardiac trabeculae subjected to the same pacing protocol as in the simulations. "Pull" corresponds to a parametric condition in which processes downstream NADH control the respiratory flux (Reproduced from Cortassa *et al*. (2003) by permission of Biophysical Society)

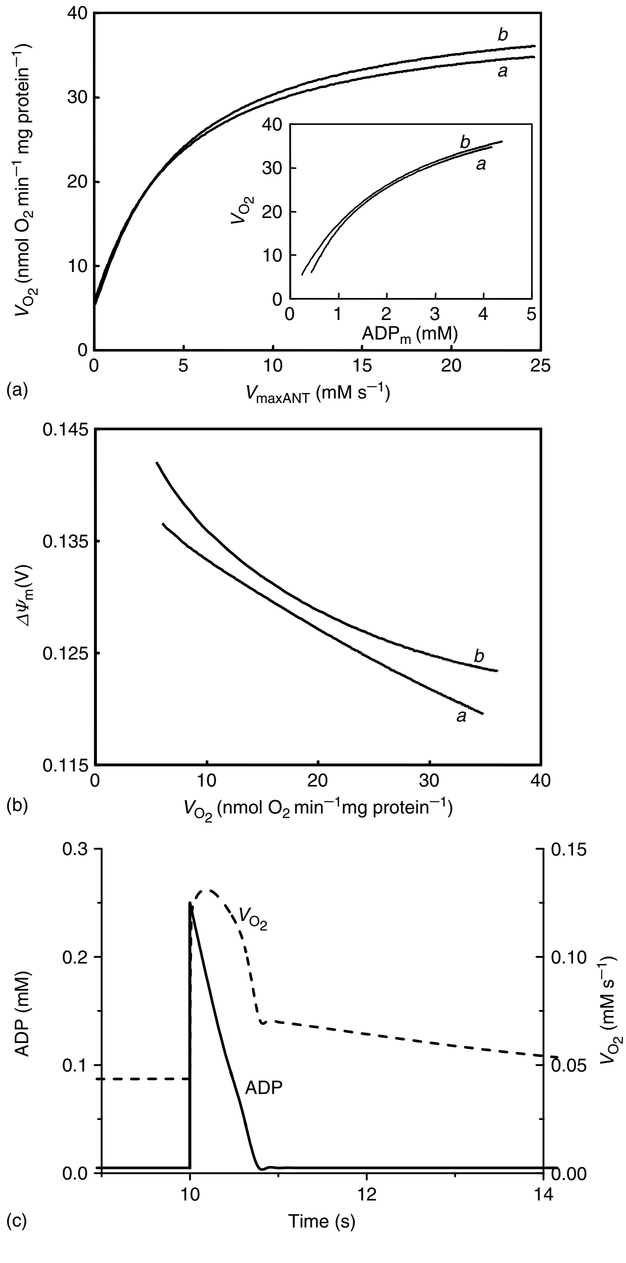


Figure 3 Relationship between membrane potential, $\Delta\Psi_m$, and respiratory flux, V_{O2}, upon changes in adenine nucleotide translocator (ANT) activity, and the transition from state 4 to state 3, obtained with an integrated model of mitochondrial energetics. The changes in V_{O2} and $\Delta\Psi_m$ are determined by the behavior of the model according to a steady state analysis performed at different ANT maximal activities (V_{maxANT}) that result in different ADP$_m$ concentrations (a, inset). The left portion of curves *a* and *b* in (a) and (b), corresponds to mitochondria in state 4 (i.e., low ADP) and the increase in V_{O2} results from the respiratory control exerted by augmenting ADP concentrations. (c) Model simulation of the transition state 4 → state 3 after a pulse of 0.25 mM ADP (Panels a and b are reproduced from Cortassa *et al*. (2003) by permission of Biophysical Society)

phosphorylation and (2) that the redox state of the glutathione pool oscillates (Cortassa *et al*., 2004). Another remarkable (yet untested) prediction of this model is that the mitochondrial oscillator of heart cells would be able to modulate its period over a wide range, from milliseconds to hours (Cortassa *et al*., 2004). This opens the possibility of a role of mitochondrial oscillations in intracellular signaling or timekeeping since bursts of ROS are a primary output of the mitochondrial oscillator. There is widespread evidence that many signaling and transcriptional activation cascades respond to ROS (Dröge, 2002; see the following text).

3. Mathematical models: the road to quantitation

Quantitation of cellular metabolism is one of the main facets that characterize the *integrative* period of biology. After most of the metabolic pathways were described and their constituent enzymes kinetically characterized, the road was paved for quantitation. Three main theoretical developments opened the way for a quantitative analysis of metabolism: kinetic modeling, metabolic control analysis (MCA), and metabolic flux analysis (MFA). Each of these theoretical developments was achieved during the second half of the past century. We will refer to the two former developments since MFA is reviewed elsewhere in this volume (*see* Article 112, **Constraint-based modeling of metabolomic systems**, Volume 6).

3.1. Kinetic modeling

This modeling technique is expanding and becoming a common tool for gaining mechanistic insight into important biological phenomena (Figures 2 and 3). Several efforts such as the insulin secretion in pancreatic β-cells (Magnus and Keizer, 1998), receptor-triggered cell signaling and cell communication during proliferation, migration, and cellular differentiation (Wiley *et al*., 2003), circadian rhythms (Goldbeter, 2002), or ultradian clocks (Lloyd *et al*., 2002) are worth mentioning. Many of these research efforts emphasize quantitative rather than qualitative data, and provide accurate tests for hypothesis-driven research (Winslow *et al*., 1999; Phair and Misteli, 2001; Tomita, 2001; Endy and Brent, 2001; Hasty *et al*., 2002; Slepchenko *et al*., 2002; Weiss *et al*., 2003).

Kinetic modeling of signaling networks is a timely topic for cell function (Bhalla, 2003; *see also* Article 108, **Functional networks in mammalian cells**, Volume 6 and Article 117, **EGFR network**, Volume 6). From the perspective of metabolic dynamics, the information-carrying networks often function transiently, under time frames fast enough for protein synthesis to be significant. Such is the case of cells subjected to sudden environmental changes where covalent modification by kinases activates production of allosteric effectors (AMP:ATP), or enzymes sense ratios of key allosteric metabolites (3PGA/Pi) producing an ultrasensitive amplification of entire metabolic pathways (Gomez-Casati *et al*., 2003; Hardie and Hawley, 2001; Vaseghi *et al*., 2001; Marsin *et al*., 2000; Marsin *et al*., 2002), or transcription factors whose critical cysteine residues are oxidized result in decrease of gene

promoter activity and subsequent gene expression (Morel and Barouki, 1999 and references therein).

In our view of cells as multilayered mass-energy-information networks of reactions (Figure 1), kinetic modeling is a basic tool for understanding their dynamics and control (Figures 2 and 3). This highlights the importance of being able to quantify the control and regulation of reaction networks in organelles and cells. Metabolic Control Analysis has provided these tools.

3.2. *Metabolic control analysis (MCA)*

Independently developed by Kacser and Burns (1973) and Heinrich and Rapoport (1974) elaborating on the work of Higgins (1965), this quantitative methodology has been extended and improved (see Fell, 1997; Cortassa *et al*., 2002, for reviews). Several aspects of control and regulation in metabolic networks have been quantified and conceptually refined, using MCA as a tool.

MCA addresses the question of what controls, and to what extent, the flux through a metabolic pathway at the steady state. Considering the mass-energy-information networks associated with mitochondrial energy production and regulation (Figure 4), we will illustrate how MCA is applied. In this particular case, relevant questions are: what controls the respiratory flux? and what controls relevant metabolite and ion concentrations such as ADP and Ca^{2+}? The latter question is relevant since ADP and Ca^{2+} play an important role as effectors in the regulation of several nodes in the network. More specifically, ADP for F_1F_0-ATP synthase and succinyl CoA lyase, and Ca^{2+} for the TCA cycle enzymes α-ketoglutarate (KGDH) and isocitrate (IDH) dehydrogenases. In this way, ADP and Ca^{2+} take part of the information-carrying networks as second messengers and allosteric effectors.

According to MCA, control is quantified through two different types of coefficients: the flux control coefficient and the elasticity coefficient. Both coefficients represent either systemic (flux control) or local (elasticity) properties of the network and its components, respectively. In mitochondria, the control exerted by each network component on the respiratory flux at the steady state is quantified by the flux control coefficient (Kacser and Burns, 1973; Heinrich and Rapoport, 1974):

$$C_{E_k}^{J_i} = \frac{\mathrm{E_k}}{\mathrm{J_i}} \frac{\partial J_i}{\partial E_k} \tag{1}$$

where the flux control coefficient, $C_{E_k}^{J_i}$, is the fractional change in respiratory flux (J_i) in response to a fractional change in a parameter of the system, typically an enzyme concentration, E_k. Similarly, a metabolite, M_i, concentration control coefficient, $C_{E_k}^{M_i}$, is defined:

$$C_{E_k}^{M_i} = \frac{\mathrm{E_k}}{M_\mathrm{i}} \frac{\partial M_i}{\partial E_k} \tag{2}$$

The link between the kinetic properties of an enzyme and its potential for flux control is given by the elasticity coefficient, ε, through the connectivity theorem (Fell, 1997; Cortassa *et al*., 2002). The elasticity coefficient ε_{ij} quantifies the effect of ionic species, or intermediate metabolite S (e.g., α-ketoglutarate) on the velocity

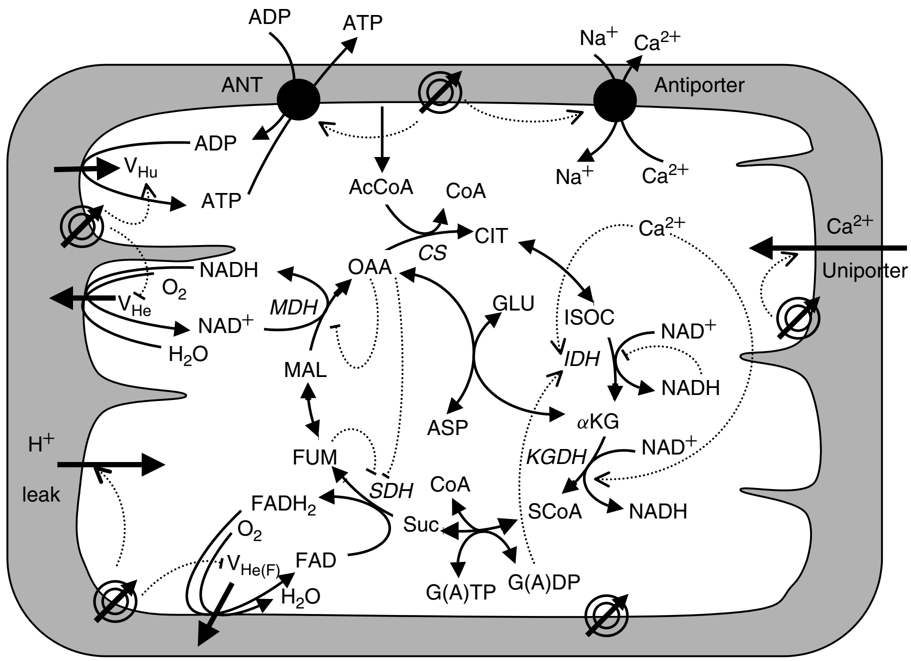


Figure 4 Schematic representation of electrophysiological and metabolic processes and their interactions, as described by an integrated model of mitochondrial energy metabolism (Cortassa *et al*., 2003). The model takes into account oxidative phosphorylation and matrix-based processes in mitochondria. The tricarboxylic (TCA) cycle of the mitochondrial matrix is fed by acetyl CoA (AcCoA). The TCA cycle oxidizes AcCoA to CO_2 and produces NADH and $FADH_2$, which provide the driving force for oxidative phosphorylation. NADH and $FADH_2$ are oxidized by the respiratory chain, and the concomitant pumping of protons across the mitochondrial inner membrane establishes an electrochemical gradient, or proton motive force ($\Delta\mu_H$). This proton motive force drives the phosphorylation of matrix ADP to ATP by the F_1F_0-ATPase (ATP synthase). The $\Delta\Psi_m$ of the inner membrane also governs the electrogenic transport of ions, including the cotransport of ATP and ADP by the adenine nucleotide translocator, Ca^{2+} influx via the Ca^{2+} uniporter and Ca^{2+} efflux via the Na^+/Ca^{2+} antiporter. The model also accounts for the explicit dependence of the TCA cycle enzymes isocitrate dehydrogenase (IDH) and α-ketoglutarate dehydrogenase (KGDH) on Ca^{2+}. Key to symbols: The concentric circles with an arrow across located at the inner mitochondrial membrane represent the $\Delta\Psi_m$. Dotted arrows indicate regulatory interactions either positive (arrow head) or negative (−) (Reproduced from Cortassa *et al*. (2003) by permission of Biophysical Society)

v_i of enzyme E_i (e.g., α-ketoglutarate dehydrogenase) is:

$$\varepsilon_S^{v_i} = \frac{\mathrm{S}}{v_i}\frac{\partial v_i}{\partial \mathrm{S}} \tag{3}$$

being the elasticity coefficient, $\varepsilon_S^{v_i}$, the fractional change in rate of the isolated enzyme, ∂v_i, in response to a fractional change, ∂S, in the amount of effector, S.

When MCA was applied to elucidate the control of mitochondrial respiration, two major conditions of mitochondrial physiology were found and defined as “push” or “pull” (Table 1, see also Cortassa *et al*., 2003). A “push” condition happens when the steps controlling the respiratory flux are upstream NADH (e.g., TCA cycle), whereas “pull” corresponds to a situation in which respiration is

Table 1 Control of the respiratory flux by TCA cycle enzymes and the rates of membrane-associated processes[a]

	Parametric set, flux control coefficient	
	Push	Pull
TCA cycle		
Citrate synthase (CS)	0.33	0.35
Isocitrate dehydrogenase (IDH)	22.8	5.3
α-ketoglutarate dehydrogenase (KGDH)	70.0	0.3
Malate dehydrogenase (MDH)	3.6	13.9
Membrane-associated processes		
Adenine nucleotide translocator (ANT)	3.8	38
Respiratory electron carriers	2.1	6.7
F_1F_0-ATPase	−0.2	2.7
Proton leak	0.1	0.3

[a]Control exerted by the rate of each process on mitochondrial respiration is quantified around the steady state corresponding to two set of parameters (Cortassa *et al*., 2003). The flux control coefficient (in %) indicates to which extent the respiratory flux changes upon a change on the activity of an enzyme in the pathway (see text equation 1). The flux control coefficient was calculated from the slope of the double logarithmic plot of V_{O2} as a function of the process rate under study.
In the cases of CS, IDH, KGDH, and MDH, the k_{cat} is varied around the steady-state level (<5% change) in order to compute the control coefficient. ρ^{res} and ρ^{F1} are varied (<10% change) to compute the control exerted by respiratory electron carriers and the F_1F_0-ATP synthase, respectively. The control exerted by the adenine nucleotide translocator and the proton leak is performed with parameters V_{maxANT} and g_H, respectively (Cortassa *et al*., 2003) (Reproduced from Cortassa *et al*. (2003) by permission of Biophysical Society).

controlled by processes downstream NADH (e.g., adenine nucleotide translocator, ATPase, respiration itself) (Table 1). Under “push” conditions, the Ca^{2+}-sensitive dehydrogenases (KGDH and IDH) of the TCA cycle share the control of respiration with 70% and 23%, respectively. For this situation, Ca^{2+} stimulates the rates of KGDH and IDH in the TCA cycle, and, in turn, the rates of respiration and F_1F_0-ATP synthase (Cortassa *et al*., 2003). This is an important point since it illustrates the difference between control and regulation, that is, respiration in mainly *controlled* by KGDH and IDH and *regulated* by Ca^{2+}. Flux control indicates the influence of the rate of a metabolic reaction or a biological process on a flux, whereas regulation refers to the modulation of an enzyme activity or pathway in response to the change in the level of a metabolite or ionic species. The difference is reflected in the way these concepts are quantitatively estimated. In our example, the regulatory effect of Ca^{2+} is quantified by the elasticity (equation 3) of the KGDH rate to Ca^{2+} (Figure 5), whereas control is quantified by the flux control coefficient (equation 1). A close inspection of Figure 5 shows that the initial reaction rate response of KGDH with respect to its substrate, α-ketoglutarate, exhibits substantial elasticity to Ca^{2+} within the concentration range detected in the mitochondrial matrix for the latter. This point also illustrates the importance of knowing the control exerted by different nodes in the network on Ca^{2+} or ADP concentration (quantified by the metabolite concentration coefficient, equation 2),

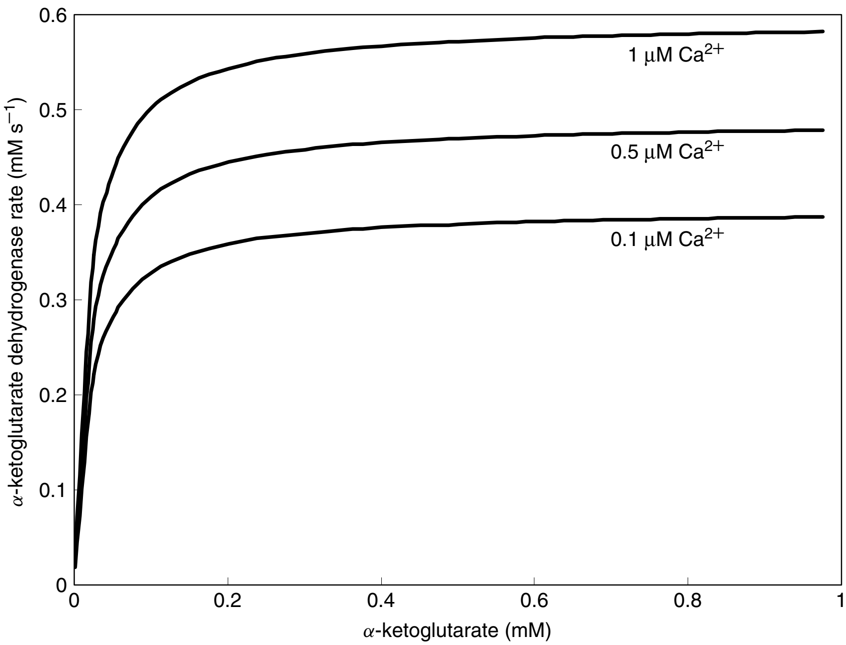


Figure 5 α-ketoglutarate dehydrogenase dependence on α-ketoglutarate and Ca^{2+}. The rate of KGDH is plotted against α-ketoglutarate concentration at various levels of Ca^{2+} as reported in Cortassa *et al*. (2003). The increase in V_{max} attained by the enzyme at saturating substrate levels indicates a positive elasticity of the enzymatic rate with respect to Ca^{2+} levels in the physiological range (Reproduced from Cortassa *et al*. (2003) by permission of Biophysical Society)

that is, their target enzymes will be responsive to the effector depending on its level.

The weight of the respiratory flux control shifts downstream of NADH ("pull") where the respiratory chain itself (i.e., through the concentration of electron carriers) and the adenine nucleotide translocator become the most rate-controlling steps (Table 1).

When targeting gene products differentially expressed in the disease state, the potentiality of MCA for detecting the main rate-controlling or regulatory steps in integrated metabolic systems becomes evident. For pathology, one would like to target the main determinants of a particular disease. In this sense, unique possibilities are open for the rational design of cells (Cortassa *et al*., 2002), or of effective therapeutic strategies (Cascante *et al*., 2002).

4. Proteomics and cell physiology: redundancy, control, and regulation

The changes in protein amounts and quality under defined conditions are main pieces of evidence provided by proteomics (Kovarova *et al*., 2002; Beranova-Giorgianni *et al*., 2002). This information can be very powerful when allied with

the theoretical concepts described before. As a matter of fact, many proteins have enzymatic activity. From the point of view of MCA, if the amount of an enzyme changes it may result in a change in the control of the flux. However, if the amount of an enzyme or protein remains unchanged it does not necessarily rule out its importance for regulation. Usually, regulatory enzymes in a metabolic network do not control the flux but exhibit high elasticity, or are responsive to an effector. Usually, these effectors are cofactors such as, for example, ATP, NAD(P)H, intracellular messengers (e.g., cAMP, Ca^{2+}), or a physiological condition (e.g., pH). If the control exerted by a protein on a metabolic network is known or can be known, then proteomics may highlight the importance of the role played by that protein in a well defined condition. Thus, proteomics combined with mathematical models will help elucidate the control aspects of metabolism.

Another important issue for cell physiology concerns the proteins that do not change or functionally redundant ones (i.e., different proteins may execute the same function). Crucial mechanisms such as energy provision in cells tend to be redundant conferring robustness to cells (Bailey, 1999). Bioenergetics is an excellent example of this. Considering the functional implications for cells under normal physiological conditions, proteins that do not change imply that they take part of vital mechanisms that tend to have little control on cell function, precisely because of that importance. In the same sense, a current misconception associated with adenine nucleotides is that they cannot be important because their concentrations hardly vary (Hardie and Hawley, 2001). Thus, cellular energetics and all its protein machinery constitute a compelling case illustrating the validity of the concept of redundancy (Koefoed *et al*., 2002).

5. Cytoplasmic organization: the missing link

The reliability of metabolic networks is firmly rooted in organization as well as redundancy. By the former, we mean supramolecular organization given by the cytoskeleton and protein–protein homologous or heterologous associations. Parallel pathways of identical function represent redundancy.

Since the beginning of the 1970s, compelling experimental evidence showed that enzymes are associated to other macromolecules in the cytoplasm of cells (Welch, 1977; Clegg, 1984; see Aon and Cortassa, 1997, for a review). These findings and the numerous reports derived from them progressively led to the idea that the physical chemistry of cellular cytoplasm is characterized by (supra)molecular organization and crowding (Aon *et al*., 2000; Ovadi and Saks, 2004). To this regard, at least two main ideas concerning the organization of metabolism in cells have emerged in the last decades: (1) metabolic channeling, in which reactions are facilitated by product/substrate transfer between closely associated enzymes (stable or transient enzymatic complexes) (Ovadi and Srere, 2000; Ovadi and Saks, 2004) and (2) cytoskeleton-driven modulation of enzymatic fluxes (Aon and Cortassa, 1997; Aon *et al*., 2000). Importantly, the kinetic properties of enzymes change during their association with macromolecular complexes composed of other enzymes or macromolecules taking part of the cytoskeleton or in cellular scaffolds.

In this sense, the cytoskeleton takes part of the information-transduction networks of cells (Aon and Cortassa, 2002).

The fractal organization exhibited by cytoskeletal proteins (Aon and Cortassa, 1994), and mitochondrial networks' physiology (Aon *et al*., 2004a) is a key ingredient for the scaling (power law) behavior exhibited by the connectivity of metabolites (Wagner and Fell, 2001; Bray, 2003) or mitochondrial networks at criticality (Aon *et al*., 2004b). The consequences for cellular (patho)physiology (O'Rourke, 2000) emerging from this "top-down" view must be further explored. Recent theoretical studies suggest that fractal organization allows enzyme reactions, following Michaelis–Menten or allosteric type kinetics, exhibits higher rates and sensitivity amplification in fractal media (for short times and at low substrate concentrations) than in Euclidean media (Aon *et al*., 2004a). These findings point out that owing to their fast kinetics, molecular mechanisms qualify for conferring fast and modulated response to cells facing environmental challenges (Koshland, 1987; Gomez-Casati *et al*., 2003). This feature is crucial when cells should adapt their energetics and signaling mechanisms following sudden changes in workload (Appaix *et al*., 2003) as in heart, or substrate supply (Vaseghi *et al*., 1999) or cell cycle dynamics (Müller *et al*., 2003) as in yeast cells.

6. Conclusions

6.1. The future of metabolic networks: a prospective

Jacob and Monod (1961) first proposed the metaphor of the genetic program as one entirely contained within the genomic DNA sequence. However, paraphrasing Fox Keller (2000), "If we wanted to keep the computer metaphor, we could describe the fertilized egg as a massively parallel and multilayered processor in which both programs (or networks) and data are distributed throughout the cell. (. . .) In this distributed program all the various DNA, RNA, and protein components function alternatively as instructions and data." This is in agreement with the view of cells as multilayered, distributed, mass-energy-information networks (Figure 1).

Many would agree that "cells are robust systems insensitive to many mutations, particularly those affecting 'core' activities" (Bailey, 1999). However, reported evidence also shows that several gene mutations affect more (sometimes seemingly unrelated) processes than the ones supposed to be targeted by the mutation. Conspicuous examples are given by cell-cycle genes affecting catabolite repression and, vice versa, catabolite repression genes influencing cell-cycle behavior (Mónaco *et al*., 1995; Aon and Cortassa, 1998); antisense inhibition of a plastidial isoform of adenylate kinase (unrelated, in principle, to the starch pathway) resulted in an increase in starch to 60% above that found in wild type (Regierer *et al*., 2002); or lack of correlation even between metabolites only separated by a single enzymatic step (Carrari *et al*., 2003). These lessons suggest caution when one annotates metabolic pathways (usually the mass-energy carrying ones) from genomic information as usually carried out with Metabolic Flux Analysis. This is due to the following reasons (see Cortassa *et al*., 2002, for several examples):

(1) some complex metabolic phenotypes are the outcome of multiple interacting genes and the environment; (2) the phenotypic consequences of gene inactivation depend on genetic background and pleiotropic effects; (3) pleiotropy may also arise through the strategic function of a gene product deeply nested in the mass-energy or information networks; (4) the missing information concerning regulatory interactions and signaling pathways that link all layers as shown in Figure 1.

The presumption that changes observed in vitro for enzymes or proteins do not necessarily happen under cellular conditions, points out that two main features representing the physicochemical condition of the cytoplasm ((supra)molecular organization and crowding) are not represented in the test tube. However, enzyme-mediated signal amplification has been shown to happen, or be modulated by cytoskeletal proteins or molecular crowding under in vitro or in situ conditions. These properties of individual enzymes' regulation when integrated in metabolic networks allow the emergence of new, unexpected, behaviors such as coherence and robustness (Aon and Cortassa, 2002).

Overall, the present state of knowledge dictates the necessity of a more profound understanding of the role and dynamics of information (signaling) networks as they relate to gene expression, under normal or pathophysiological conditions. Compelling experimental evidence points out the importance of accounting for cytoplasmic organization such as compartmentation, channeling, enzyme-cytoskeleton interaction, and crowding, in our mathematical models.

6.2. The evolution of mathematical models

There will be a survival of some of the mathematical models that are, and will be, proposed to do the task of integrating and making sense of the overflowing biological information. The survival of the fittest for mathematical models will be given by how closely interdependent have theory and experiment been during model construction provided the four main conditions for model reliability are fulfilled. In this way, we will be able to unfold the full potential of our models at descriptive, explanatory, and predictive levels, and through this, the organized complexity of nature.

While synthesizing biological information into a coherent scheme, there is a trade-off to be accomplished between *completeness* and *understandability*. We still have a long way to go to apprehend the fundamental role of signaling networks and cytoplasmic organization on cell function, a task for which mathematical modeling is unavoidable. We have greatly developed our ability to be more *complete* at integrating biological data. However, by doing so we may reach levels of trade-off between completeness and understandability that are beyond the reach of our present knowledge. We believe that there is much more room to accomplish the mentioned trade-off than any previous theoretical argument will suggest us to believe. However, at this point it is useful to remember that the *complete* simulation of a cell could involve a program whose description is *as large and complex as the cell itself*. As Barrow (1999) put it: "It is like having a full-scale map, as large as the territory it describes: extraordinarily accurate; but not so useful, and awfully tricky to fold up". The interpretation of the in silico behavior of such a cell may

well be beyond our understanding. Thus, at some stage of the *integrative* period of biology the important question to answer will be: do we have any tool(s) to keep the trade-off between completeness and understandability optimal enough and still retain the possibility of a significant and accessible knowledge of how a cell works? This is largely the problem that will be elucidated in the decade(s) to come.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson JD (1989) *Molecular Biology of the Cell*, Garland Publishing: New York.

Aon MA and Cortassa S (1994) On the fractal nature of cytoplasm. *FEBS Letters*, **344**, 1–4.

Aon MA and Cortassa S (1997) *Dynamic Biological Organization. Fundamentals as Applied to Cellular Systems*, Chapman & Hall: London.

Aon MA and Cortassa S (1998) Catabolite repression mutants of *Saccharomyces cerevisiae* show altered fermentative metabolism as well as cell cycle behaviour in glucose-limited chemostat cultures. *Biotechnology and Bioengineering*, **59**, 203–213.

Aon MA and Cortassa S (2002) Coherent and robust modulation of a metabolic network by cytoskeletal organization and dynamics. *Biophysical Chemistry*, **97**, 213–231.

Aon MA, Cortassa S, Gomez-Casati DF and Iglesias AA (2000) Effects of stress on cellular infrastructure and metabolic organization in plant cells. *International Review of Cytology*, **194**, 239–273.

Aon MA, Cortassa S, Marbán E and O'Rourke B (2003) Synchronized whole-cell oscillations in mitochondrial metabolism triggered by a local release of reactive oxygen species in cardiac myocytes. *The Journal of Biological Chemistry*, **278**, 44735–44744.

Aon MA, O'Rourke B and Cortassa S (2004a) The fractal architecture of cytoplasmic organization: scaling, kinetics and emergence in metabolic networks. *Molecular and Cellular Biochemistry*, **256/257**, 169–184.

Aon MA, Cortassa S and O'Rourke B (2004b) Percolation and criticality in a mitochondrial network. *Proceedings of the National Academy of Sciences United States of America*, **101**, 4447–4452.

Appaix F, Kuznetsov AV, Usson Y, Kay L, Andrienko T, Olivares J, Kaambre T, Sikk P, Margreiter R and Saks V (2003) Possible role of cytoskeleton in intracellular arrangement and regulation of mitochondria. *Experimental Physiology*, **88**, 175–190.

Bailey JE (1999) Lessons from metabolic engineering for functional genomics and drug discovery. *Nature Biotechnology*, **17**, 616–618.

Barrow JD (1999) *Impossibility. The Limits of science and the Science of Limits*, Vintage: London.

Beranova-Giorgianni S, Giorgianni F and Desiderio DM (2002) Analysis of the proteome in the human pituitary. *Proteomics*, **2**, 534–542.

Bhalla US (2003) Understanding complex signaling networks through models and metaphors. *Progress in Biophysics and Molecular Biology*, **81**, 45–65.

Boveris A, Costa LE, Cadenas E and Poderoso JJ (1999) Regulation of mitochondrial respiration by adenosine diphosphate, oxygen, and nitric oxide. *Methods in Enzymology*, **301**, 188–198.

Brandes R and Bers DM (2002) Simultaneous measurements of mitochondrial NADH and Ca^{2+} during increased work in intact rat heart trabeculae. *Biophysical Journal*, **83**, 587–604.

Bray D (2003) Molecular networks: the top-down view. *Science*, **301**, 1864–1865.

Carrari F, Urbanczyk-Wochniak E, Willmitzer L and Fernie AR (2003) Engineering central metabolism in crop species: learning the system. *Metabolic Engineering*, **5**, 191–200.

Cascante M, Boros LG, Comin-Anduix B, Atauri P, Centelles JJ and Lee PWN (2002) Metabolic control analysis in drug discovery and disease. *Nature Biotechnology*, **20**, 243–248.

Chance B and Williams GR (1956) The respiratory chain and oxidative phosphorylation. *Advances in Enzymology*, **17**, 65–134.

Clegg JS (1984) Properties and metabolism of the aqueous cytoplasm and its boundaries. *The American Journal of Physiology*, **246**, R133–R151.

Cleveland DW (1988) Autoregulated instability of tubulin mRNAs: a novel eukaryotic regulatory mechanism. *Trends in Biochemical Sciences*, **13**, 339–343.

Cortassa S, Aon MA, Lloyd D and Iglesias AA (2002) *An Introduction to Metabolic and Cellular Engineering*, World Scientific: Singapore.

Cortassa S, Aon MA, Marbán E, Winslow RL and O'Rourke B (2003) An integrated model of cardiac mitochondrial energy metabolism and calcium dynamics. *Biophysical Journal*, **84**, 2734–2755.

Cortassa S, Aon MA, Winslow RL and O'Rourke B (2004) A mitochondrial oscillator dependent on reactive oxygen species. *Biophysical Journal*, **87**, 2060–2073.

Dröge W (2002) Free radicals in the physiological control of cell function. *Physiological Reviews*, **82**, 47–95.

Endy D and Brent R (2001) Modelling cellular behaviour. *Nature*, **409**, 391–395.

Fell D (1997) *Understanding the Control of Metabolism*, Portland Press: London.

Fox Keller E (2000) *The Century of the Gene*, Harvard University Press: Cambridge.

Goldbeter A (2002) Computational approaches to cellular rhythms. *Nature*, **420**, 238–245.

Gomez-Casati DF, Cortassa S, Aon MA and Iglesias AA (2003) Ultrasensitive behavior in the synthesis of storage polysaccharides in cyanobacteria. *Planta*, **216**, 969–975.

Hardie GD and Hawley SA (2001) AMP-activated protein kinase: the energy charge hypothesis revisited. *BioEssays: News and Reviews in Molecular, Cellular and Development Biology*, **23**, 1112–1119.

Harris DA and Das AM (1991) Control of mitochondrial ATP synthesis in the heart. *The Biochemical Journal*, **280**, 561–573.

Hasty J, McMillen D and Collins JJ (2002) Engineered gene circuits. *Nature*, **420**, 224–230.

Heinrich R and Rapoport TA (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *European Journal of Biochemistry*, **42**, 89–95.

Higgins J (1965) Dynamics and control in cellular reactions. In *Control of Energy Metabolism*, Chance B, Estabrook RW and Williamson JR (Eds.), Academic Press: New York, pp. 13–46.

Jacob F and Monod J (1961) On the regulation of gene activity. *Cold Spring Harbor Symposia Quantitative Biology*, **26**, 193–211.

Kacser H and Burns JA (1973) The control of flux. *Symposia of the Society for Experimental Biology*, **27**, 65–104.

Koefoed S, Otten M, Koebmann B, Bruggeman F, Bakker B, Snoep J, Krab K, van Spanning R, van Verseveld H, Jensen P, *et al*. (2002) A turbo engine with automatic transmission? How to marry chemicomotion to the subtleties and robustness of life. *Biochimica et Biophysica Acta*, **1555**, 75–82.

Koshland DE (1987) Switches, thresholds and ultrasensitivity. *Trends in Biochemical Sciences*, **12**, 225–229.

Kovarova H, Halada P, Man P, Dzubak P and Hajduch M (2002) Application of proteomics in the search for novel proteins associated with the anti-cancer effect of the synthetic cyclin-dependent kinases inhibitor, bohemine. *Technology in Cancer Research and Treatment*, **1**, 247–256.

Lloyd D, Salgado EJ, Turner MP, Suller MTE and Murray DB (2002) Cycles of mitochondrial energization driven by an ultradian clock in a continuous culture of *Saccharomyces cerevisiae*. *Microbiology*, **148**, 3715–3724.

Magnus G and Keizer J (1998) Model of β-cell mitochondrial calcium handling and electrical activity. I. cytoplasmic variables. *The American Journal of Physiology*, **274**; *Cell Physiology*, **43**, C1158–C1173.

Marsin AS, Bertrand L, Rider MH, Deprez J, Beauloye C, Vincent MF, Van den Berghe G, Carling D and Hue L (2000) Phosphorylation and activation of heart PFK-2 by AMPK has a role in the stimulation of glycolysis during ischaemia. *Current Biology*, **10**, 1247–1255.

Marsin AS, Bouzin C, Bertrand L and Hue L (2002) The stimulation of glycolysis by hypoxia in activated monocytes is mediated by AMP-activated protein kinase and inducible 6-phosphofructo-2-kinase. *The Journal of Biological Chemistry*, **277**, 30778–30783.

Mónaco ME, Valdecantos P and Aon MA (1995) Carbon and energetic uncoupling associated with cell cycle arrest of cdc mutants of *Saccharomyces cerevisiae* may be linked to glucose-induced catabolite repression. *Experimental Cell Research*, **217**, 52–56.

Morel Y and Barouki R (1999) Repression of gene expression by oxidative stress. *The Biochemical Journal*, **342**, 481–496.

Müller D, Exler S, Aguilera-Vázquez L, Guerrero-Martín E and Reuss M (2003) Cyclic AMP mediates the cell cycle dynamics of energy metabolism in *Saccharomyces cerevisiae*. *Yeast*, **20**, 351–367.

Oliver S (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.

O'Rourke B (2000) Pathophysiological and protective roles of mitochondrial ion channels. *The Journal of Physiology*, **529**(Pt 1), 23–36.

Ovadi J and Saks V (2004) On the origin of intracellular compartmentation and organized metabolic systems. *Molecular and Cellular Biochemistry*, **256/257**, 5–12.

Ovadi J and Srere PA (2000) Macromolecular compartmentation and channeling. *International Review of Cytology*, **192**, 255–280.

Phair RE and Misteli T (2001) Kinetic modeling approaches to *in vivo* imaging. *Nature Molecular Cell Biology*, **2**, 898–907.

Regierer B, Fernie AR, Springer F, Perez-Melis A, Leisse A, Koehl K, Willmitzer L, Geigenberger P and Kossmann J (2002) Starch content and yield increase as a result of altering adenylate pools in transgenic plants. *Nature Biotechnology*, **20**, 1256–1260.

Slepchenko BM, Schaff JC, Carson JH and Loew LM (2002) Computational cell biology: spatiotemporal simulation of cellular events. *Annual Review of Biophysics and Biomolecular Structure*, **31**, 423–441.

Tomita M (2001) Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology*, **19**, 205–210.

Vaseghi S, Baumeister A, Rizzi M and Reuss M (1999) *In vivo* dynamics of the pentose phosphate pathway in *Saccharomyces cerevisiae*. *Metabolic Engineering*, **1**, 128–140.

Vaseghi S, Macherhammer F, Zibek S and Reuss M (2001) Signal transduction dynamics of the protein kinase-A/phosphofructokinase-2 system in *Saccharomyces cerevisiae*. *Metabolic Engineering*, **3**, 163–172.

Wagner A and Fell DA (2001) The small world inside large metabolic networks. *Proceedings of the Royal Society of London Series B*, **268**, 1803–1810.

Weiss JN, Qu Z and Garfinkel A (2003) Understanding biological complexity: lessons from the past. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, **17**, 1–6.

Welch GR (1977) On the role of organized multienzyme systems in cellular metabolism: a general synthesis. *Progress in Biophysics and Molecular Biology*, **32**, 103–191.

Wiley HS, Shvartsman SY and Lauffenburger DA (2003) Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends in Cell Biology*, **13**, 43–50.

Winslow RL, Rice J, Jafri S, Marbán E and O'Rourke B (1999) Mechanisms of altered excitation-contraction coupling in canine tachycardia-induced heart failure, II Model studies. *Circulation Research*, **84**, 571–586.

Specialist Review

A complex systems approach to understand how cells control their shape and make cell fate decisions

Donald E. Ingber and Sui Huang
Harvard Medical School and Children's Hospital, Boston, MA, USA

1. Introduction

Systems Biology emerged from our increasing awareness that the study of individual genes, proteins, and pathways cannot explain the characteristic behaviors of biological systems as a whole. In this chapter, we focus on the challenge of understanding how the myriad of molecular pathways in living cells are coordinated so as to give rise to switches between the different discrete system-level behaviors of cells in multicellular organisms, such as proliferation, differentiation into one of a finite number of specialized cell types, cell migration or apoptosis. We also discuss how this whole-cell behavior is linked to the physical shape of cells – a connection that is crucial for the formation of tissue patterns.

Systems biologists take various approaches to address the problem cellular regulation (Huang, 2004). Some see in a systems approach the comprehensive detailed characterization of all molecular processes using the burgeoning massively parallel, high-throughput analytical methods. Although such approaches are indispensable, their goal still remains gene-centered in nature (Bray, 2003). The notion here is that an exhaustive description at the highest possible resolution of mechanistic details of a system equals the understanding of that system. In contrast to this "bottom-up" approach, other investigators who typically come to Systems Biology from fields outside classical molecular biology often take a "top-down" approach. Their main goal is to extract generalizable (even "universal") principles that govern the behavior of complex biological systems without specifying the identity of individual molecular players. One central premise here is that higher-order systems-level features that result from multicomponent interactions (von Bertalanffy, 1969; Bar-Yam, 1997) cannot be understood by studying the properties of the individual components or groups of components (e.g., signaling modules) alone – be it qualitatively by traditional biochemistry, or quantitatively by computational modeling. A first step in this approach is to

recognize and characterize the sometimes abstract higher-order qualities or patterns of behavior at the whole-system level. It is noteworthy that this top-down approach of Systems Biology is distinct from classical organismal biology that similarly studies biological behavior at a higher level, in that now we have access to the molecular component parts, and hence seek to understand the nature of the relationship between them and the integrated, system-level features.

Thus, our group has raised the molecular analysis of how mammalian cells decide whether to grow, differentiate, migrate, or die to that of a systems-level, top-down approach. In this chapter, we review work addressing the following fundamental questions: (1) How do interactions between different molecular filaments within structural networks, such as the cytoskeleton, lead to the production of living cells and tissues with characteristic shapes and mechanical properties? (2) How do regulatory interactions among genes and proteins produce a coherent information processing machinery that enables cells to sense multiple simultaneous inputs and orchestrate a concerted response? and (3) How do changes in cell structural networks impact these information processing networks? In addressing these questions, we will describe two integrating concepts of how cells organize their structural and information networks, which may explain system-level (whole cell) properties from interactions among multiple subcomponents.

2. Cell hardware: tensegrity as a design principle

Current understanding of cell and tissue regulation is explained largely in terms of changes in the activity of individual molecules and their interactions. To ultimately understand how biochemistry is related to the physical nature of tissue patterns, we have focused on the cellular mechanism that accounts for the spatial differentials of cell behavior, for example, growth versus quiescence between neighboring cells. Such spatial heterogeneity of cell fates drives tissue patterns formation, such as the branching capillary networks of the vascular system. Our work has revealed the importance of cell adhesion to extracellular matrix (ECM), mechanical forces, and associated changes in cell shape for control of cell fates (Huang and Ingber, 1999). Local alterations in ECM mechanics and tension-dependent changes in cell shape determine the cell's fate with respect to proliferation, differentiation, movement, contractility, and apoptosis in response to soluble hormones and cytokines present in the tissue. For example, as cell spreading on the matrix is progressively inhibited by culturing cells on microfabricated adhesive islands of decreasing area, the cells will proliferate (when spread), shut off growth and differentiate (when spreading is partially inhibited), or undergo apoptosis (when fully retracted and round), even in the presence of attachment to the matrix and of saturating concentrations of growth factors. Similar observations were made in various cell types, including fibroblasts, capillary endothelial cells, vascular smooth muscle cells, and primary liver epithelial cells (Ingber and Folkman, 1989; Singhvi *et al*., 1994; Chen *et al*., 1997; Huang *et al*., 1998; Parker *et al*., 2002; Polte *et al*., 2004). These findings have changed the paradigm in cell-growth regulation and developmental control by placing molecular signaling cascades in the context of the structural and mechanical complexity of living tissues.

2.1. The tensegrity model

Given our finding that mechanical distortion of cells (i.e., global changes in cell shape) plays a central role in cell fate switching during development, a major focus in our laboratory has been to uncover the principles governing how cells assemble and rearrange macromolecules to control their shape, and how mechanical forces impact cellular biochemistry and gene expression via cell distortion.

Our approach to understand how multiple molecular filaments collectively control the mechanical stability of the cytoskeleton – the molecular framework that gives the cell its shape – is based on cellular tensegrity theory (Ingber, 1998; Ingber, 2003a). Tensegrity is a building principle that was first described by the architect R. Buckminster Fuller and first visualized by the sculptor Kenneth Snelson. Fuller defines tensegrity systems as structures that stabilize their shape by continuous tension or "tensional integrity" rather than by continuous compression (e.g., as used in a stone arch). Tensegrity systems are composed of a tensed network of structural members that are stabilized in space because they are balanced by a subset of other members that resist being compressed. It is this preexisting tensile stress (prestress) or isometric tension within a tensegrity network that holds its joints in a stable position, and governs how the whole system of interconnected members responds when mechanically stressed. Our bodies provide a simple example of a tensegrity structure: our bones act like struts to resist the pull of tensile muscles, tendons, and ligaments, and the shape stability (stiffness) of our bodies varies depending on the tone (prestress) in our muscles.

The cellular tensegrity model proposes that the whole cell is a structural hierarchy of molecules that is organized as a prestressed tensegrity structure (Ingber, 2003a). In cells, tensional forces are borne by cytoskeletal microfilaments and intermediate filaments, and these forces are balanced by interconnected structural elements that resist compression, most notably internal microtubule struts and ECM adhesions. However, individual filaments can have dual functions and hence bear either tension or compression in different structural contexts or at different size scales (e.g., contractile microfilaments generate tension, whereas actin microfilament bundles that are rigidified by cross-links bear compression in filopodia). The tensional prestress that stabilizes the whole cell is generated actively by the contractile actomyosin apparatus. Additional passive contributions to this prestress come from cell distension through adhesions to the ECM and other cells, osmotic forces, and forces exerted by filament polymerization. Furthermore, intermediate filaments that interconnect at many points along microtubules, microfilaments, and the nuclear surface as well as the highly elastic, cortical cytoskeletal (actin-spectrin-ankyrin) network directly beneath the plasma membrane provide additional stiffness. The entire integrated cytoskeleton is then permeated by a viscous cytosol and enclosed by a differentially permeable surface membrane.

Importantly a theoretical formulation of the cellular tensegrity model starting from first mechanistic principles was developed by Dimitrije Stamenovic working with our group and by others (Stamenovic and Ingber, 2002; Ingber, 2003a). In this quantitative model of the cell as a mechanical system, actin microfilaments, and intermediate filaments carry the prestress that is balanced internally by microtubules and externally by focal adhesions to the ECM substrate. Work on variously shaped

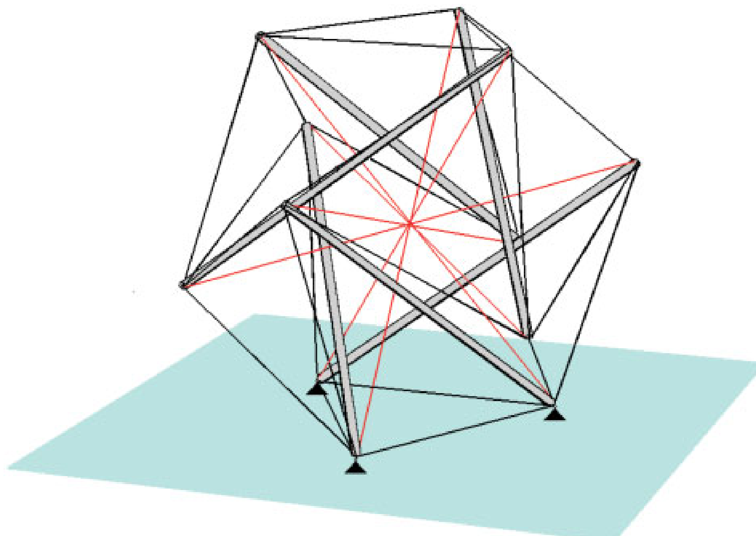


Figure 1 A spherical tensegrity structure used to generate a computational tensegrity model of the cell. The thin tendons represent microfilaments (black lines) and radial intermediate filaments (red lines) in the cytoskeleton; the thick gray struts indicate microtubules. Anchoring points to the substrate (blue) are indicated by the black triangles

models revealed that even the simplest prestressed tensegrity structure embodies the key mechanical properties of all members of this tensegrity class. Thus, for simplicity, we used a symmetrical cell model in which the tensed filaments are represented by 24 cables and the microtubules by 6 struts organized as shown in the structure in Figure 1. The cytoskeleton and substrate together were assumed to form a self-equilibrated, stable mechanical system; the prestress carried by the cables was balanced by the compression of the struts.

A microstructural analysis of this model using the principle of virtual work led to two a priori predictions relating to emergent mechanical behaviors of the entire, integrated, cell-wide, cytoskeletal system: (1) the stiffness of the model (or cell) will increase as the prestress (*preexisting* tensile stress) is raised and (2) at any given prestress, stiffness will increase linearly with increasing stretching force (*applied* stress). The former is consistent with what we know about how muscle tone alters the stiffness of our bodies and it closely matches data from experiments with living cells (Wang *et al*., 2002; Stamenovic *et al*., 2002a, 2003). The latter meshes nicely with the mechanical measurements of stick-and-string tensegrity models, cultured cells, and whole living tissues, although it can be explained by other models also. This mathematical approach strongly supports the idea that architecture and prestress in the cytoskeleton are key to a cell's emergent shape robustness.

Largely through the work of Stamenovic and coworkers, this oversimplified micromechanical model continues to be progressively modified and strengthened over time. A more recent formulation of the model includes, for example, semiflexible struts analogous to microtubules, rather than rigid compression struts,

and incorporates values for critical features of the individual cytoskeletal filaments (e.g., volume fraction, bending stiffness, and cable stiffness) from the literature (Coughlin and Stamenovic, 1997; Stamenovic and Coughlin, 1999). This more refined model is qualitatively and quantitatively superior to the one that contains rigid struts. Another formulation of the tensegrity model that includes intermediate filaments as tension cables that link the cytoskeletal lattice and surface membrane to the cell center (Wang and Stamenovic, 2000) was able to explain mechanical properties of cells lacking the intermediate filament vimentin. All these tensegrity models yield elastic moduli (stiffness) quantitatively similar to those of cultured adherent cells (Stamenovic and Coughlin, 1999, 2000) when measured through cell surface receptors that link to the internal cytoskeleton (Stamenovic and Coughlin, 2003). Although the current formulation of the tensegrity theory relies on a highly simplified architecture (6 struts and 24 cables), it nevertheless effectively predicts many *static* mechanical behaviors of living mammalian cells. Most critically, the a priori prediction of the tensegrity model that cell stiffness will increase in proportion with the prestress has been confirmed in various experimental studies (Wang *et al*., 2002; Stamenovic *et al*., 2002a, 2003).

Interestingly, recent studies have revealed that the *dynamic* mechanical behavior of mammalian cells depends on generic system properties, as indicated by a spectrum of time constants when the cells are stressed over a wide range of force application frequencies (Fabry *et al*., 2001). This work suggests that these dynamic behaviors reflect an emergent property of the cell or cytoskeleton that is observable only at some higher system level of molecular interaction. It is not consistent with the notion of a single type of cytoskeletal filament or molecular interaction (e.g., actin cross-linking) being responsible for cell dynamic behavior. It is also not consistent with standard ad hoc models of cell mechanics that assume the elastic and frictional behaviors of the cell originate from two distinct compartments (the elastic cortex and the viscous cytoplasm). Importantly, computer simulations suggest that dynamic mechanical behaviors exhibited by living cells, including the dependence of both their elastic and frictional moduli on prestress (Wang *et al*., 2001; Stamenovic *et al*., 2002a) may also be explained by their use of tensegrity (Canadas *et al*., 2002; Sultan *et al*., 2004). In other words, tensegrity could provide a common structural basis for both the elastic and viscous behaviors of living cells.

In summary, the work on tensegrity revealed that great insight into cellular behavior due to molecular interactions could be gained by viewing the system as a whole and working from the "top-down" starting with system-level mechanical properties. Specifically, we found that when trying to understand collective mechanical behavior within supramolecular assemblies, higher-order architecture and physical forces also must be considered. This work also revealed how robust behaviors, such as persistence, mechanical adaptability, and shape stability, can be generated using "sloppy" parts (e.g., flexible molecular filaments): since the system properties are "emergent", they do not put very high constraints on the precision of individual component parts and their rules of assembly (Carlson and Doyle, 2002). Thus, tensegrity may represent the "hardware" behind living systems.

3. Attractors in information networks

But what about the software? This leads us to the problem of how structural networks affect information processing networks at the level of the whole cell where tensegrity and the cytoskeleton seem to exert their effects on cell shape and on signal integration. Experiments show that while individual cells in the tissue may receive multiple simultaneous inputs, they are able to rapidly integrate these signals so as to produce just one of a few possible outputs, or cell fate programs (e.g., proliferation, quiescence, differentiation, apoptosis). Furthermore, as described above, cells are able to respond to mechanical cues (e.g., cell shape distortion) with the same behavioral repertoire as they do in response to signals carried by specific molecular messengers, such as hormones or growth factors. The finding that mechanical or geometrical factors that are devoid of the molecular information carried by soluble ligands that bind to specific cell receptors can regulate the exact same biochemical machinery (e.g., cell cycle G1/S transition or apoptosis) as do the specific soluble factors raises a fundamental question: how can a gradual change in a physical control parameter over a broad continuum, such as cell shape (distortion from round to spread), be translated into one of a few discrete and mutually exclusive cellular programs?

Signal transduction, the biochemical process of converting an external molecular signal into a cellular effect, has historically been viewed in terms of "signaling pathways" that lead from a specific input to a particular outcome (*see* Article 108, **Functional networks in mammalian cells**, Volume 6). However, information processing in cells appears to be "distributed": Different stimuli can elicit the same behavior, whereas the same stimulus can generate different responses, depending on the cellular context, including the activity of other regulatory proteins and the physical state of the cell (Huang and Ingber, 2000). The apparent lack of hard-wired, point-to-point, function-specific pathways and the high prevalence of cross-talk between canonical signaling cascades are best documented in signal transduction of cytokines (Kerr *et al*., 2003). DNA microarray analysis now confirms that engaged cytokine receptors induce thousands of genes that can overlap to a great extent (Fambrough *et al*., 1999; Iyer *et al*., 1999; Tamayo *et al*., 1999). Yet, despite this apparent lack of pathway specificity, receptor ligands reliably produce a concerted biological effect, such as proliferation, differentiation, or apoptosis. Thus, equating a molecular signaling cascade with a logical chain of causation may be too simplistic to explain how cells process environmental signals and make discrete cell fate decisions.

It has become evident that signal transduction and gene regulatory pathways form a single, large, connected genome-scale network that spans virtually across the cell's entire regulatory machinery (Marcotte, 2001; Bray, 2003; Huang, 2004). How then can discrete cell phenotypes, such as cell fates, emerge from such genome-wide molecular networks? More generally, a long-standing question in physics has been: Can such large, apparently irregular, that is, "complex" networks produce global, coherent patterns of gene and protein activation – the prerequisite for expression of discrete behaviors at the whole-cell level? While the detailed "wiring diagram" of such molecular information processing networks of the mammalian genome remains largely unknown (*see* Article 110, **Reverse engineering gene regulatory**

networks, Volume 6), early computer models of the dynamics of statistical ensembles of large, generic networks of interacting genes (Kauffman, 1993) have revealed that, for a subset of types of network architectures, "ordered" rather than "chaotic" behavior can arise, in that only a few stable states spontaneously emerge as a result of the global constraints imposed by the regulatory interactions (Kauffman, 1993; Fox and Hill, 2001; Aldana and Cluzel, 2003; Huang, 2004). Importantly, this subset of network architectures that exhibit ordered behavior encompasses those found in nature in lower organisms where gene or protein network topology has been elucidated (Huang, 2004). In such networks with ordered behavior, most states of the network (defined by the configuration of activity of all molecules of the network) are unstable because the respective configuration of gene or protein activation is incompatible with their regulatory relationships: for instance, mutually inhibitory genes cannot be active at the same time. As a consequence, most network states are "attracted" to the few stable network states, known as the "attractor" states, which are stable with respect to the activity status of all the genes of the network, and hence are said to be "high-dimensional". Such attractor states define the activity configuration of genes across the network that complies with the regulatory interactions (see Figure 2 for details). The central idea then is that a cell fate, or more generally, any distinct, stable phenotypic state, is a high-dimensional attractor state in terms of network dynamics (Kauffman, 1993; Huang, 2002). This is essentially an extension to the genomic scale of the old idea that cell differentiation is due to a bistable switch in a gene circuit containing a self-activating gene or protein (Delbrück, 1949; Thomas, 1998)

Since attractors in gene regulatory networks are akin to energy potential wells in some abstract sense, one can think of all the possible states of a network's dynamic behavior, the so-called state space, as constituting a "potential landscape". Then a marble that represents a network state at a given time point will roll down into the valleys or pits to the bottom that would represent the stable attractor states (Figure 2). The movement (change of network state) is driven by the execution of the regulatory interactions. Upon small perturbations (change in activity status of few individual genes), the marble will generally remain within its local neighborhood and roll back down into the same valley; an attractor state is therefore robust to small perturbations in many directions (=with respect to many genes). In response to a larger perturbation, typically involving large sets of genes, however, the marble could traverse a mountainous peak in the landscape and irrevocably be committed to rolling down the other side of the hill (driven by the regulatory interactions in the network) until it reaches another stable attractor in a neighboring valley. In the case of the cell, this corresponds to a cell fate switch that would cause it to take on a different, stable gene activation configuration and thus, exhibit a different stable phenotype.

The "attractor landscape" schematically captures the network's entire behavioral repertoire. Its structure is determined by the genomic gene and protein regulatory network hard-wired in the genome sequence that encodes protein structures and promoters motifs, and hence determines which molecules interact with which one and in what modality. The attractor landscape may be what Waddington had in mind when he proposed in the 1940s, not knowing about gene networks, the metaphor of an "epigenetic landscape" (Figure 3) to explain the discreteness and stability of

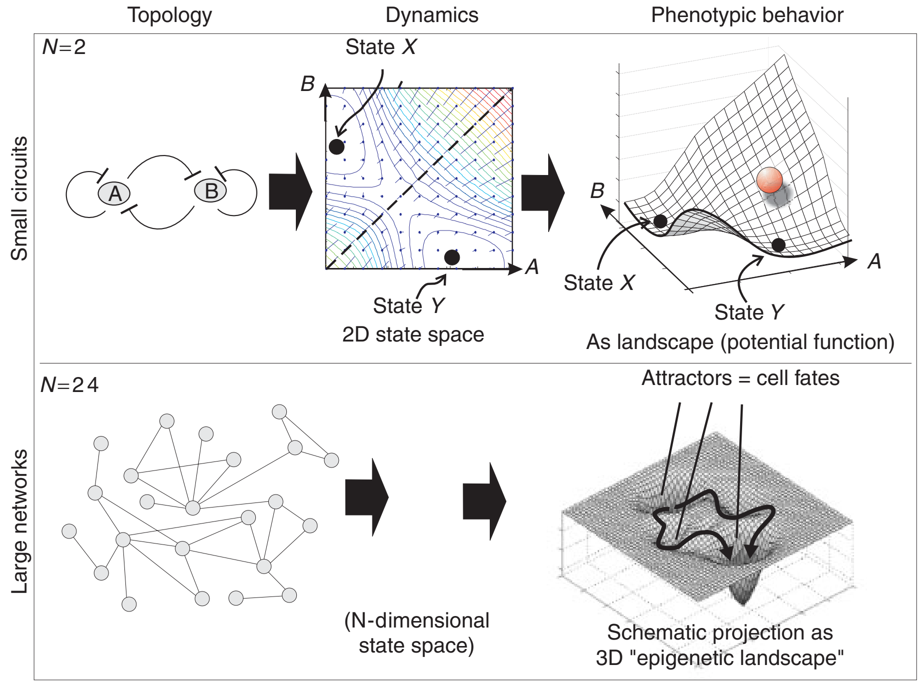


Figure 2 From network topology to dynamics, illustrated for small circuits and large network. Top panel: Small circuit (signaling module) toy model consisting of two proteins, A and B, which are mutually inhibitory (via sigmoidial functions), and promote their own (first order) decay. For a wide range of kinetic parameters, this circuit topology gives rise to bistability. Middle panel shows a phase plane (two-dimensional state space) that represents the dynamics as a vector field. Each dot represents a state $S(t) = [x_A(t), x_B(t)]$ of the circuit at time t, defined by the expression or activity values $x(t)$ of the two nodes, A and B (see text). The ticks emanating from each dot points to the direction in state space in which that respective state will move in the next time unit to comply with the interactions (i.e., systems equations) of the circuit. Contour lines help visualize the landscape character that arises from the dynamics at each point. There are two stable states, X and Y (black dots), to which the ticks point. The dashed diagonal represents the separatrix, the border line between the two basins of attraction. The right panel illustrates the structure of the state space after numerical integration of the vector field from an arbitrary starting point in x and y directions. Although the vector field in this case is not integrable, that is, not a gradient field, the graph shows the landscape character and the bistability with states X and Y as "potential wells" (attractors) and unstable regions in-between. Bottom panel: Analogous illustration of topology and dynamics, but for a larger network ($N = 24$). Now the state space is N-dimensional, containing all the network states, defined by $S(t) = [\, x_1(t), x_2(t), \ldots, x_N(t)\,]$ and cannot be drawn; however, again it can be schematically thought of as some sort of a landscape in which the pits represent (high-dimensional) attractor states (right panel). In this projection, every point in the xy-plane represents a (high-dimensional) network state $S(t)$ arranged so as to generate the landscape, where the vertical axis would represent some "state change potential" value reflecting the "dissimilarity" of the state S_{xy} at position (x,y) to that of the attractor state that it will eventually reach. Each attractor state corresponds to a distinct phenotypic cell state, or a cell fate. The landscape can be thought of as the formal and molecular equivalent of Waddington's "epigenetic landscape" (see Figure 3)

cell fates in development (Waddington, 1956). The attractor landscape metaphor naturally explains systems features of cell fate switching behavior discussed earlier: the existence of discrete specialized cellular programs (cell fates), the all-or-none switch between them, and their mutual exclusivity. It also makes plausible why

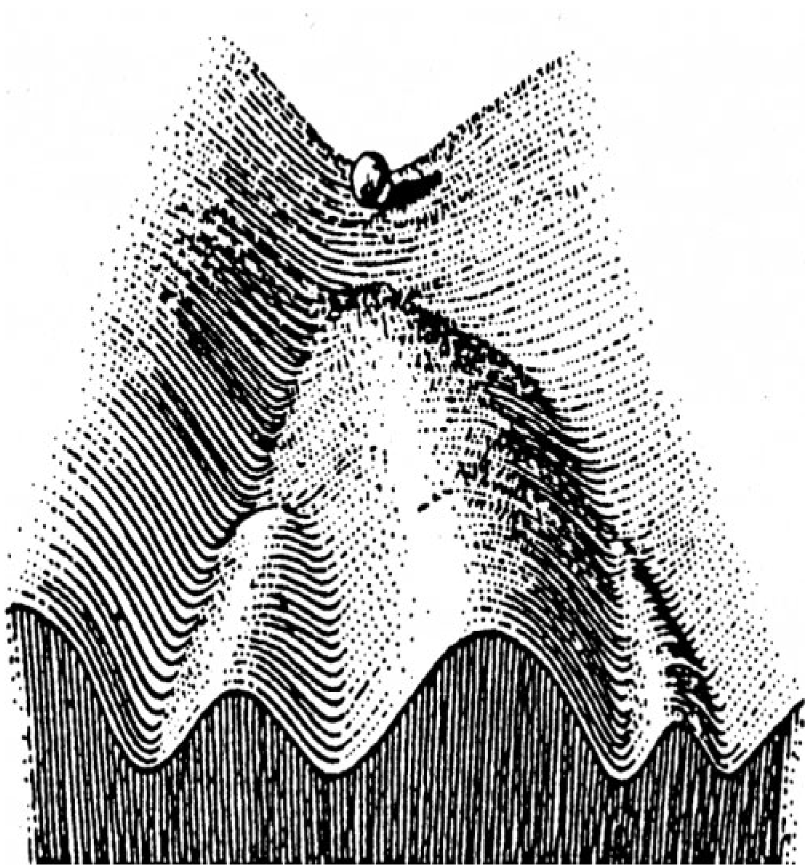


Figure 3 "Epigenetic landscape". A metaphor proposed by Waddington in the 1940s to explain discreteness and stability of cell fates during embryonic development (Reproduced by permission from C. Waddington, (1957), *The Strategy of the Genes*. © Allen and Unwin)

many unrelated signals, including those devoid of molecular specificity can lead to the same cell fate.

Recent experimental work from our laboratory using dynamic gene expression profiling of human promyelocytic HL60 cells that undergo differentiation into neutrophils in response to two distinct stimuli, either a specific (retinoic acid) or a nonspecific (dimethylsulfoxide) agent, provides direct support for the attractor hypothesis. A necessary signature of a transition into a high-dimensional attractor is the convergence of genome-scale gene expression profiles from different perturbed network states as the cell assumes the new phenotypic state (Figure 2, bottom right). Indeed, the mature neutrophil differentiated state in HL60 cells was approached from two different directions in the epigenetic landscape: starting from identical network states, defined by a gene expression profile over ~3000 genes (correlation coefficient $r = 1$), in response to these two distinct agents, the cells visited globally distinct states (profiles) ($r = 0.2$) before being "attracted" to the common one specific for the neutrophils ($r = 0.9$). In terms of individual genes, this

means that differentiation can take alternative paths, that is, genes are turned on and off differently in these two alternative neutrophil differentiation processes. This observation defies the traditional notion of unique, specific, and dedicated "differentiation pathways".

An important corollary of this alternative paradigm of cellular signaling is that, due to the inherent stability of a high-dimensional attractor state, a change in a large set of genes is necessary for a transition from one stable state to another (e.g., switch from proliferation to differentiation) (Huang, 2002). This could explain the pleiotropy and promiscuity in signal transduction pathways that may have been wired by evolution precisely to trigger transitions between different attractor states.

Another fundamental consequence of the concept of a high-dimensional attractor landscape hard-wired in the genome is that it provides a deterministic guiding structure, so to speak a stage on which the inherently noisy molecular processes of regulated gene expression take place (Paulsson, 2004). The stability of high-dimensional attractors ensures that the stochastic fluctuations in the levels of expression and activation of multiple interacting proteins will not affect cell phenotype. Thus, a stable phenotype (e.g., a quiescence state) of an apparently homogeneous cell population may actually represent a "swarm" of points rather than a single point at the bottom of an attractor. Accordingly, signal transduction then may be viewed as a process that orchestrates these random fluctuations in the activity of thousands of individual genes so that their joint probability of activation may produce an "outlier" cell in the swarm that would jump over a crest and roll down to another attractor (Figure 4). Since the attractor requires that

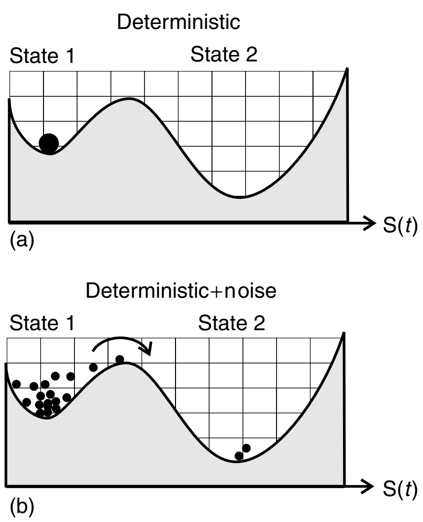


Figure 4 Effect of noise on a network with two attractor states. Schematic illustration of how the attractor landscape controls and exploits molecular noise. Every dot corresponds to a network state, defined by its position in the horizontal axis. (a) A situation without noise. (b) Noise is added to the same system. Noise can "kick" the system out of one state and force it to transition to another stable state. Because of the attractors, a specific transition (from state 1 to state 2) can occur following a nonspecific (noisy) perturbation

multiple genes be regulated to produce an attractor switch, a combinatorial scheme in which the fluctuations in the individual genes are biased by the external signal would allow the fine-tuning of the "macroscopic" cell fate transition probability. In fact, the stochastic nature of cell lineage commitment in stem cells as well as many forms of cell fate regulation appears to be due to the combined effect of deterministic, instructive external regulators and intrinsic probabilistic, digital events (Enver *et al*., 1998; Hume, 2000; Lahav *et al*., 2004).

4. Integration of structural and information processing networks

If system-level behavior of a whole cell is a transition between robust, high-dimensional attractor states of the regulatory network, then the requirement for specific details of input signals can be relaxed. This provides the evolvability necessary for cells to evolve the capability to harness environmental cues that lack molecular specificity, such as physical forces and mechanical distortion, to serve as signals that can cause a particular cell fate transition and the associated highly concerted changes in gene expression.

The precise molecular mechanism by which a physical change in a cell is transduced into a biochemical signal is not fully understood. However, what is now well documented is that the actin cytoskeleton acts as a scaffold for the assembly of many signaling pathway protein complexes (Geiger and Bershadsky, 2001) and that a change in the actin network architecture (or prestress) can affect the activity of large arrays of signaling pathways. In fact, many of the biochemical events that mediate cell metabolism and signal transduction require solid-state biochemistry: cellular enzymes and substrates that mediate these biochemical reactions are immobilized on insoluble molecular scaffolds within the cytoskeleton and nuclear matrix (Ingber, 2003a). Mechanical forces are transmitted across cell surface adhesion receptors, such as integrins, that mechanically couple ECM scaffolds to the same internal cytoskeletal scaffolds (Wang *et al*., 1993) and are thought to be converted into chemical and electrical signals through stress-dependent distortion of molecules that associate with load-bearing elements of the cytoskeleton (Ingber, 2003a). In this manner, physical distortion of the structural network in cells may affect a large set of signaling pathways (Alenghat and Ingber, 2002). Yet, because of the existence of stable attractors in the information-processing network, physical perturbations and associated pleiotropic biochemical signals would be channeled down a valley in the epigenetic landscape, and hence automatically orchestrated to trigger a specific cellular program. This is consistent with the widely seen ability of cell shape distortion to influence cell fates as discussed earlier.

The integration of the global mechanical state of a cell, as predicted by the tensegrity model, with the regulation of cell fates is key to development and tissue homeostasis because it governs the spatial organization of the expression of distinct cellular phenotype so as to comply with the macromechanical demands (stretching or compression of tissue) within an organism. A prosaic example for this sculpting and molding of tissues guided by physical reality is wound healing. When

cells are lost from an epithelium due to injury, cells surrounding the wound will spread, move to fill the space, and divide until the initial state of cell compaction (compression) is restored; then proliferation shuts off. Similarly, during embryonic development cells, in regions of tissues that experience higher levels of deformation also often exhibit higher proliferation rates (reviewed in Huang and Ingber, 1999; Ingber, 2003b), thus balancing cell proliferation with macroscopic need.

5. Conclusion

In this chapter, we presented two complex system aspects of cells that elude explanation by the molecular characterization of individual components or pathways: the cell *as a physical entity* with mechanical properties and the cell's *global information processing behavior* that mediates cell fate decisions. We showed how the concept of "network" in the broadest sense facilitates the bridging of the gap between molecules and system-level behaviors. It is important to note that cytoskeletal and regulatory networks belong to entirely different classes of objects: the former is a concrete physical scaffold made of individual proteins, while the latter is an abstract wiring diagram between symbols representing molecular species. Nevertheless, both network concepts help us understand how a robust, system-level feature emerges from rule-governed interactions between molecules. Tensegrity structures explain the mechanical properties of a whole cell, and attractor states explain cell fate switching. Such a conceptual understanding does, of course, not obviate the need for our continuing efforts to achieve a comprehensive characterization of all the molecular constituents and processes in the cell. But it emphasizes the importance of complementing the pathway-centered approach that currently dominates Systems Biology with top-down approaches that embrace the actual system properties. Only in this manner will it be possible for us to understand how and why the whole is more than the sum of its parts once we have catalogued all the molecular components.

Acknowledgments

This work was supported by the National Health Institutes and the Air Force Office for Scientific Research.

Further reading

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

References

Aldana M and Cluzel P (2003) A natural class of robust networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8710–8714.

Alenghat FJ and Ingber DE (2002) Mechanotransduction: all signals point to cytoskeleton, matrix, and integrins. *Science's STKE: Signal Transduction Knowledge Environment*, **2002**, PE6.

Bar-Yam Y (1997) *Dynamics in Complex Systems (Studies in Nonlinearity)*, Perseus Publishing: Boulder.

Bray D (2003) Molecular networks: the top-down view. *Science*, **301**, 1864–1865.

Canadas P, Laurent VM, Oddou C, Isabey D and Wendling S (2002) A cellular tensegrity model to analyse the structural viscoelasticity of the cytoskeleton. *Journal of Theoretical Biology*, **218**, 155–173.

Carlson JM and Doyle J (2002) Complexity and robustness. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(Suppl 1), 2538–2545.

Chen CS, Mrksich M, Huang S, Whitesides GM and Ingber DE (1997) Geometric control of cell life and death. *Science*, **276**, 1425–1428.

Coughlin MF and Stamenovic D (1997) A tensegrity structure with buckling compression elements: application to cell mechanics. *ASME Journal of Applied Mechanics*, **64**, 480–486.

Delbrück M (1949) Discussion. In *Unités Biologiques Douées de Continuité Génétique Colloques Internationaux du*, Centre National de la Recherche Scientifique: Paris.

Enver T, Heyworth CM and Dexter TM (1998) Do stem cells play dice? *Blood*, **92**, 348–351.

Fabry B, Maksym GN, Butler JP, Glogauer M, Navajas D and Fredberg JF (2001) Scaling the microrheology of living cells. *Physical Review Letters*, **87**, 148102.

Fambrough D, McClure K, Kazlauskas A and Lander ES (1999) Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent sets of genes. *Cell*, **97**, 727–741.

Fox JJ and Hill CC (2001) From topology to dynamics in biochemical networks. *Chaos*, **11**, 809–815.

Geiger B and Bershadsky A (2001) Assembly and mechanosensory function of focal contacts. *Current Opinion in Cell Biology*, **13**, 584–592.

Huang S (2002) Regulation of cellular states in mammalian cells from a genome-wide view. *Gene Regulation and Metabolism: Post-Genomic Computational Approach*, MIT Press: Cambridge, pp. 181–220.

Huang S (2004) Back to the biology in systems biology: what can we learn from biomolecular networks. *Briefings in Functional Genomics & Proteomics*, **2**, 279–297.

Huang S, Chen CS and Ingber DE (1998) Control of cyclin D1, p27(Kip1), and cell cycle progression in human capillary endothelial cells by cell shape and cytoskeletal tension. *Molecular Biology of the Cell*, **9**, 3179–3193.

Huang S and Ingber DE (1999) The structural and mechanical complexity of cell-growth control. *Nature Cell Biology*, **1**, E131–E138.

Huang S and Ingber DE (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Experimental Cell Research*, **261**, 91–103.

Hume DA (2000) Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. *Blood*, **96**, 2323–2328.

Ingber DE (1998) The architecture of life. *Scientific American*, **278**, 48–57.

Ingber DE (2003a) Tensegrity I. Cell structure and hierarchical systems biology. *Journal of Cell Science*, **116**, 1157–1173.

Ingber DE (2003b) Tensegrity II. How structural networks influence cellular information processing networks. *Journal of Cell Science*, **116**, 1397–1408.

Ingber DE and Folkman J (1989) Mechanochemical switching between growth and differentiation during fibroblast growth factor-stimulated angiogenesis in vitro: role of extracellular matrix. *The Journal of Cell Biology*, **109**, 317–330.

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, *et al*. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Kauffman SA (1993) *The Origins of Order*, Oxford University Press: New York.

Kerr IM, Costa-Pereira AP, Lillemeier BF and Strobl B (2003) Of JAKs, STATs, blind watchmakers, jeeps and trains. *FEBS Letters*, **546**, 1–5.

Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, Elowitz MB and Alon U (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nature Genetics*, **36**, 147–150.

Marcotte EM (2001) The path not taken. *Nature Biotechnology*, **19**, 626–627.

Parker KK, Brock AL, Brangwynne C, Mannix RJ, Wang N, Ostuni E, Geisse NA, Adams JC, Whitesides GM and Ingber DE (2002) Directional control of lamellipodia extension by constraining cell shape and orienting cell tractional forces. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, **16**, 1195–1204.

Paulsson J (2004) Summing up the noise in gene networks. *Nature*, **427**, 415–418.

Polte TR, Eichler GS, Wang N and Ingber DE (2004) Extracellular matrix controls myosin light chain phosphorylation and cell contractility through modulation of cell shape and cytoskeletal prestress. *American Journal of Physiology. Cell Physiology*, **286**, C518–C528.

Singhvi R, Kumar A, Lopez GP, Stephanopoulos GN, Wang DI, Whitesides GM and Ingber DE (1994) Engineering cell shape and function. *Science*, **264**, 696–698.

Stamenovic D and Coughlin MF (1999) The role of prestress and architecture of the cytoskeleton and deformability of cytoskeletal filaments in mechanics of adherent cells: a quantitative analysis. *Journal of Theoretical Biology*, **201**, 63–74.

Stamenovic D and Coughlin MF (2000) A quantitative model of cellular elasticity based on tensegrity. *ASME Journal of Biomechanical Engineering*, **122**, 39–43.

Stamenovic D and Coughlin MF (2003) A prestressed cable network model of the adherent cell cytoskeleton. *Biophysical Journal*, **84**, 1328–1336.

Stamenovic D and Ingber DE (2002) Models of cytoskeletal mechanics and adherent cells. *Biomechanics and Modeling in Mechanobiology*, **1**, 95–108.

Stamenovic D, Liang Z, Chen J and Wang N (2002a) Effect of the cytoskeletal prestress on the mechanical impedance of cultured airway smooth muscle cells. *Journal of Applied Physiology*, **92**, 1443–1450.

Stamenovic D, Mijailovich SM, Tolic-Norrelykke IM and Wang N (2003) Experimental tests of the cellular tensegrity hypothesis. *Biorheology*, **40**, 221–225.

Sultan C, Stamenovic D and Ingber DE (2004) A computational tensegrity model predicts dynamic rheological behaviors in living cells. *Annals of Biomedical Engineering*, **32**, 520–530.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907–2912.

Thomas R (1998) Laws for the dynamics of regulatory networks. *The International Journal of Developmental Biology*, **42**, 479–485.

von Bertalanffy L (1969) *General System Theory*, George Brazillet: New York.

Waddington CH (1956) *Principles of Embryology*, Allen & Unwin Ltd: London.

Wang N, Butler JP and Ingber DE (1993) Mechanotransduction across the cell surface and through the cytoskeleton. *Science*, **260**, 1124–1127.

Wang N, Naruse K, Stamenovic D, Fredberg J, Mijailovic SM, Maksym G, Polte T and Ingber DE (2001) Mechanical behavior in living cells consistent with the tensegrity model. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7765–7770.

Wang N and Stamenovic D (2000) Contribution of intermediate filaments to cell stiffness, stiffening, and growth. *American Journal of Physiology. Cell physiology*, **279**, C188–C194.

Wang N, Tolic-Norrelykke IM, Chen J, Mijailovich SM, Butler JP, Fredberg JJ and Stamenovic D (2002) Cell prestress. I. Stiffness and prestress are closely associated in adherent contractile cells. *American Journal of Physiology. Cell physiology*, **282**, C606–C616.

Specialist Review

Systems biology of the heart

Denis Noble
University Laboratory of Physiology, Oxford, UK

1. Introduction

Systems biology of the heart has focused on understanding cellular and multicellular function in terms of the interactions between proteins forming channels, transporters, receptors, contractile elements, sequestration structures, and buffers. It is a good example of the need for reduction and integration to complement each other. Without identifying the elements, their coding genes, molecular structure, and their reactions there would be no basis on which to build models; equally, without modeling the interactions of these components there would be no understanding of overall function. The first successful example of this approach came when Hodgkin and Huxley (1952) published their equations for the squid giant axon. This was the first model to use mathematical reconstruction of ion channel transport and gating, rather than abstract equations and it correctly predicted the shape of the action potential, the impedance changes and the conduction velocity. It was a brilliant demonstration of the need for biological modeling to respect the details of experimental results.

Theoretical biology has only rarely been as successful as this. The squid axon is a relatively simple nervous mechanism. Most nerve cells are much more complex, and so are cardiac cells, with the added complexity of electromechanical coupling and mechanoelectric feedback (Noble *et al*., 1998). Modeling the heart has therefore required many more rounds of iteration between experimental and modeling work.

Even in the physical sciences, modeling has usually been an iterative process of interaction between mathematics and experimentation, involving successive approximations toward predictive capability. The iterative process generates insights on the way, even when models are wrong or incomplete. In fact, the mistakes can be as illuminating as the successes (Noble, 2002b). Models of complex biological systems should be judged as much by the insights they have given, as by their predictive power.

2. Potassium channels: energy conservation and the fragility of repolarization

The first integrative questions to be resolved were fundamental to cardiac electrophysiology. Compared to nerve, where the total duration is only a millisecond

or two, the cardiac action potential is exceedingly long. Hundreds of milliseconds are required for repolarization to occur. And when the resting potential has been restored, it is not necessarily stable. In pacemaker regions, like the SA node, the AV node, and the Purkinje fibers, it immediately starts to depolarize again to generate spontaneous rhythm.

What is responsible for these differences? The answer lies primarily in the nature of the potassium channels in cardiac cells. The first experimental results (Hutter and Noble, 1960; Carmeliet, 1961) showed the existence of two kinds: the inward rectifier, i_{K1}, now known to be coded by Kir2.x, and delayed rectifier i_K. The delayed rectifier was later shown to be generated by multiple components (Noble and Tsien, 1969; Sanguinetti and Jurkiewicz, 1990; Zeng *et al*., 1995), including i_{Kr}, coded by hERG and MinK, and i_{Ks}. We now know that there are many other K^+ channels whose expression levels vary with spatial location (Antzelevitch *et al*., 2001). These include i_{to}, which shapes the early phases of the action potential.

The discovery of the inward rectifier was a surprise. On depolarization, it closes so that the conductance changes in precisely the opposite direction to that of the potassium channel in squid giant axon, and it was natural to ask the question whether this could account for the long-lasting nature of the cardiac action potential. The model incorporating this mechanism (Noble, 1962) showed unambiguously that it could. In fact, it is a major energy-saving device since, by reducing the membrane permeability, very small inward currents (generated by sodium and calcium ions) suffice to maintain a long plateau of depolarization. The energy cost of a long action potential generated in this way is an order of magnitude less than it would otherwise be. The 1962 model also correctly attributed repolarization to i_K, and identified its decay as an important contribution to pacemaker activity.

Evolution involves a play-off between competing factors: energy conservation and safety being two of them. Engineers are familiar with this concept in the context of constructing aircraft, bridges, spacecraft, elevators, and a host of other technologies. Evolution approaches these competing factors in similar ways (Diamond, 1993). Energy conservation may have been one reason for the evolution of channels like i_{K1}, but in the play-off there was a very severe price. What this model and all later models showed is that repolarization becomes a much more fragile process. Very small current changes can disturb and even prevent it. Moreover, the channel mainly responsible, i_{Kr}, is one of the most promiscuous receptors known to pharmacology. This channel protein or the G-coupled proteins controlling it, interact with a large range of compounds, including drugs as diverse as antibiotics, antihistaminics, anticancer drugs, antihypertension drugs, and so on, the list is almost endless, many of which interfere with repolarization in a way that generates potentially fatal arrhythmias. A major challenge for systems modeling now is whether it can help overcome these problems (Muzikant and Penland, 2000).

While the 1962 model successfully represented the roles of i_{K1} and i_K, it completely lacked calcium fluxes, and its representation of pacemaker activity was far too simple. There are many other mechanisms involved, including the hyperpolarizing-activated current i_f (DiFrancesco and Noble, 1985; DiFrancesco, 1993) and various sodium and calcium channels (Irisawa *et al*., 1993). Modeling of pacemaker activity is currently a hot topic, with the roles of a number of ion transporters being reassessed (Rigg and Terrar, 1996; Cho *et al*., 2003) and

variations in expression levels incorporated (Zhang *et al*., 2000). A major new focus of this work is the mouse, which is of great importance in correlating genetic information with electrophysiology (Liu *et al*., 2003; Demolombe *et al*., 2004) and for interpreting the effects of genetic modifications (Noble, 2002d; Papadatos *et al*., 2004). Without a systems biological approach, it would be impossible to understand a process as multifactorial as cardiac rhythm generation. This is true not only of normal pacemaker activity but also of abnormal arrhythmic mechanisms.

3. Calcium balance: calcium channels, sodium–calcium exchange, sarcoplasmic reticulum

Harald Reuter first discovered calcium channels in the heart (Reuter, 1967), and discovered the sodium–calcium exchanger (Reuter and Seitz, 1969). Later, Alex Fabiato first demonstrated calcium-induced release of calcium from the sarcoplasmic reticulum (Fabiato, 1983).

The next major role for systems biological modeling was to integrate these three processes into the cellular web of interactions. Calcium currents were incorporated into a Purkinje fiber model by McAllister *et al*. (1975), and into the first ventricular cell model by Beeler and Reuter (1977), but the first integrative models that addressed the key questions of calcium balance were those developed with DiFrancesco (DiFrancesco and Noble, 1985) and Hilgemann (Hilgemann and Noble, 1987). These became the generic models from which all the modern models of excitation-contraction coupling derive (Luo and Rudy, 1994a,b; Jafri *et al*., 1998; Noble *et al*., 1998; Winslow *et al*., 1999; Winslow *et al*., 2000). The Hilgemann–Noble model addressed a number of important questions concerning calcium balance:

1. How quickly is calcium balance achieved? Net calcium efflux is established as soon as 20 ms after the beginning of the action potential (Hilgemann, 1986), which was considered to be surprisingly soon. In the model, this was achieved by calcium activation of efflux via the Na–Ca exchanger, thus revealing the time course of one of the important functions of this transporter in the heart.
2. Since the exchanger is electrogenic, where was the current that this would generate and did it correspond to the quantity of calcium that the exchanger needed to pump? Mitchell *et al*. (1984) provided the first experimental evidence that the action potential plateau is maintained by sodium–calcium exchange current. The Hilgemann–Noble model showed that this is precisely what one would expect, both qualitatively and quantitatively. Subsequent experimental and modeling work has fully confirmed this conclusion (Egan *et al*., 1989; LeGuennec and Noble, 1994; Eisner *et al*., 2000; Bers, 2001).
3. Could a model of the sarcoplasmic reticulum SR that reproduced the major features of Fabiato's experiments be incorporated? The model followed as much of the Fabiato data as possible, but while broadly consistent with the Fabiato work it could not be based on that alone. Fabiato's experiments were heroic but they were done on skinned fibers, which removes many of the relevant

mechanisms. It is an important function of systems biological models to reveal when experimental data needs extending.

4. Were the quantities of calcium, free and bound, at each stage of the cycle consistent with the properties of the cytosol buffers? The great majority of the cytosol calcium is bound so that, although large calcium fluxes are involved, the free calcium transients are much smaller, as they are experimentally.

The major deficiency of this model was that it could not account for graded release of calcium from the SR, or for the phenomenon of calcium sparks. Much more complex models, incorporating finer detail of the excitation–contraction process, including the communication between L-type calcium channels and the SR calcium release channels, are required to achieve this (Jafri *et al*., 1998; Rice *et al*., 1999; Winslow *et al*., 2000; Greenstein and Winslow, 2002; Hinch, 2004).

4. Modeling disease states

An important criterion for success in systems biology is the ability to account for disease states. In this section, I will show that modeling of the heart has begun to do this.

4.1. Genetic mutations

Connecting genetics to physiological function at cellular and other levels is an exciting new development in which systems biological simulations are helping to unravel complex behavior resulting from simple changes at the genetic level. Clancy and Rudy (1999) have revealed the arrhythmogenic effects of a mutation that affects sodium channel inactivation and is associated with a congenital form of the long-QT syndrome, LQT3. Long-QT refers to prolongation of the electrocardiogram interval between excitation of the ventricles (Q) and their repolarization (T), a phenomenon that can indicate life-threatening arrhythmias. The simulations show that the mutant channel generates a persistent inward sodium current during the action potential plateau, and at long cycle lengths this generates an early after-depolarization so delaying repolarization and prolonging the QT interval. Noble and Noble (2000) have shown similar consequences of shifting the voltage-dependence of sodium channel inactivation. This mimics part of the experimental effect of a missense mutation underlying the Brugada syndrome, an inherited condition in which sudden fatal ventricular fibrillation can occur in people who show no other manifestation of heart disease.

4.2. Congestive heart failure

The ability to track multiple changes in gene expression levels using cDNA arrays, real-time PCR, and other methods has revolutionized the range and quantity of data available on disease states. A major problem, however, is to distinguish those

changes that are primary from secondary changes attributable to remodeling in response to the disease state. Systems biology can help in this distinction since it can identify those changes that explain the disease effect, such as cardiac arrhythmia. An excellent example of this approach is the modeling of congestive heart failure. Experimental data shows that i_{to} and i_{K1} are downregulated, leading to a prolongation of the action potential, while the SR calcium pump, SERCA, is also downregulated leading to a reduction and slowing of the intracellular calcium transient. Sodium–calcium exchange is upregulated, which may be a secondary response to the other changes in calcium handling. Introducing these changes into cardiac cells models succeeds in reproducing failure of repolarization and multiple reexcitations of the kind that generate major arrhythmia (Winslow *et al*., 1999) and in showing that the major factor involved is the change in calcium balance.

4.3. Ischemia

Ischemia is a multifactorial process, starting with interruption of the blood supply, leading to rundown of metabolites like ATP involved in driving energy-using mechanisms, with consequent changes in ionic concentrations. These processes have been partially modeled (Ch'en *et al*., 1998). Both the metabolic and the ionic changes in turn modulate the activity of channels and transporters. Thus, some potassium channels, $i_{K,ATP}$ are activated by the fall in ATP (Noma, 1983), so shortening the action potential. Extracellular accumulation of potassium ions also contributes to this shortening through changes in i_{K1} and i_{Kr}. Meanwhile, the development of sodium and calcium overload may generate oscillatory changes in intracellular calcium that may initiate ectopic beating. These are the acute changes. Long-term changes create scar tissue that can contribute to determining the pathways for reentrant arrhythmia. Unraveling the ways in which all these processes interact to generate life-threatening arrhythmia is a problem that requires modeling. The interactions are too complex to be understood without quantitative study. However, at present, all the model studies are partial ones, and they have mostly been carried out at the cellular level, though Rudy and his colleagues (Shaw and Rudy, 1997; Wang and Rudy, 2000) have done important studies on one-dimensional multicellular models and an ectopic focus has been simulated in 2D and 3D models (Winslow *et al*., 1993). Computations at the whole organ level have yet to appear. These will be important since reentrant arrhythmia and fibrillation are properties of the whole ventricle.

5. Linking levels: putting cell models into tissue and organ models

Life-threatening arrhythmias depend on molecular and cellular mechanisms but they are fatal because of their actions at the level of the whole organ. Anatomically detailed models of the ventricles, including fiber orientations and sheet structure (Crampin *et al*., 2004), have been used to incorporate the cellular models in an attempt to reconstruct the electrical and mechanical behavior of the whole organ.

These models have been used to reconstruct the spread of the activation wave front, generated largely by the rapidly activated sodium channels and to reconstruct reentrant waves and their breakdown into the multiple wavelets of fibrillation (Bernus *et al*., 2003; Ten Tusscher and Panfilov, 2003; Panfilov and Kerkhof, 2004). These processes are heavily influenced by cardiac ultrastructure, with preferential conduction along the fiber-sheet axes, and the result corresponds well with that obtained from multielectrode recording from dog hearts in situ. Accurate reconstruction of the depolarization wavefront promises to provide reconstruction of the early phases of the ECG to complement work already done on the late phases (Antzelevitch *et al*., 2001) and as the sinus node, atrium and conducting system are incorporated into the whole heart model, we can look forward to the first example of reconstruction of a complete physiological process from the level of protein function right up to routine clinical observation.

The whole ventricular model has been incorporated into a virtual torso (Bradley *et al*., 1997), including the electrical conducting properties of the different tissues, to extend the external field computations to reconstruction of multiple-lead chest and limb recording. Incorporation of biophysically detailed cell models into whole organ models (Noble, 2002a,c; Trayanova *et al*., 2002; Crampin *et al*., 2004) is still at an early stage of development, but it is essential to attempts to understand heart arrhythmias. So also is the extension of modeling to human cells (Nygren *et al*., 1998; Ten Tusscher *et al*., 2003).

Acknowledgments

Denis Noble is a British Heart Foundation Professor. His laboratory is also supported by the MRC, BBSRC, EPSRC, and the Wellcome Trust.

References

Antzelevitch C, Nesterenko VV, Muzikant AL, Rice JJ, Chien G and Colatsky T (2001) Influence of transmural gradients on the electrophysiology and pharmacology of ventricular myocardium. Cellular basis for the Brugada and long-QT syndromes. *Philosophical Transactions of the Royal Society, Series A*, **359**, 1201–1216.

Beeler GW and Reuter H (1977) Reconstruction of the action potential of ventricular myocardial fibres. *Journal of Physiology*, **268**, 177–210.

Bernus O, Verschelde H and Panfilov A (2003) Reentry in an anatomical model of the human ventricles. *International Journal of Bifurcation and Chaos*, **13**, 3693–3702.

Bers D (2001) *Excitation-contraction Coupling and Cardiac Contractile Force*, Kluwer: Dordrecht.

Bradley CP, Pullan AJ and Hunter PJ (1997) Geometric modeling of the human torso using cubic Hermite elements. *Annals of Biomedical Engineering*, **25**, 96–111.

Carmeliet EE (1961) Chloride ions and the membrane potential of Purkinje fibres. *Journal of Physiology*, **156**, 375–388.

Ch'en FC, Vaughan-Jones RD, Clarke K and Noble D (1998) Modelling myocardial ischaemia and reperfusion. *Progress in Biophysics and Molecular Biology*, **69**, 515–537.

Cho H-S, Takano M and Noma A (2003) The electrophysiological properties of spontaneously beating pacemaker cells isolated from mouse sinoatrial node. *Journal of Physiology*, **550**, 169–180.

Clancy CE and Rudy Y (1999) Linking a genetic defect to its cellular phenotype in a cardiac arrhythmia. *Nature*, **400**, 566–569.

Crampin EJ, Halstead M, Hunter PJ, Nielsen P, Noble D, Smith N and Tawhai M (2004) Computational physiology and the physiome project. *Experimental Physiology*, **89**, 1–26.

Demolombe S, Liu J, Marionneau C, Escande D and Lei M (2004) Expression of ion channel genes in mouse heart. *Journal of Physiology Glasgow Proceedings*, **557**, C8.

Diamond JM (1993) Evolutionary physiology. In *The Logic of Life*, Boyd CAR and Noble D (Eds.), OUP: Oxford, pp. 89–111.

DiFrancesco D (1993) Pacemaker mechanisms in cardiac tissue. *Annual Reviews of Physiology*, **55**, 455–472.

DiFrancesco D and Noble D (1985) A model of cardiac electrical activity incorporating ionic pumps and concentration changes. *Philosophical Transactions of the Royal Society*, **B 307**, 353–398.

Egan T, Noble D, Noble SJ, Powell T, Spindler AJ and Twist VW (1989) Sodium-calcium exchange during the action potential in guinea-pig ventricular cells. *Journal of Physiology*, **411**, 639–661.

Eisner DA, Choi HS, Diaz ME and O'Neill SC (2000) Integrative analysis of calcium cycling in cardiac muscle. *Circulation Research*, **87**, 1087–1094.

Fabiato A (1983) Calcium induced release of calcium from the sarcoplasmic reticulum. *American Journal of Physiology*, **245**, C1–C14.

Greenstein JL and Winslow RL (2002) An integrative model of the cardiac ventricular myocyte incorporating local control of Ca2+ release. *Biophysical Journal*, **83**, 2918–2945.

Hilgemann DW (1986) Extracellular calcium transients and action potential configuration changes related to post-stimulatory potentiation in rabbit atrium. *Journal of General Physiology*, **87**, 675–706.

Hilgemann DW and Noble D (1987) Excitation-contraction coupling and extracellular calcium transients in rabbit atrium: Reconstruction of basic cellular mechanisms. *Proceedings of the Royal Society B*, **230**, 163–205.

Hinch R (2004) A mathematical analysis of the generation and termination of calcium sparks. *Biophysical Journal*, **86**, 1–15.

Hodgkin AL and Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**, 500–544.

Hutter OF and Noble D (1960) Rectifying properties of heart muscle. *Nature*, **188**, 495.

Irisawa H, Brown HF and Giles WR (1993) Cardiac pacemaking in the sinoatrial node. *Physiological Reviews*, **73**, 197–227.

Jafri S, Rice JJ and Winslow RL (1998) Cardiac Ca2+ dynamics: The roles of Ryanodine receptor adaptation and sarcoplasmic reticulum load. *Biophysical Journal*, **74**, 1149–1168.

LeGuennec JY and Noble D (1994) The effects of rapid perturbation of external sodium concentration at different moments of the action potential in guinea-pig ventricular myocytes. *Journal of Physiology*, **478**, 493–504.

Liu J, Pritchard C, Billeter R, Underhill P, Lei M and Noble D (2003) Profiling of ion channel gene expression in mouse heart. *Journal of Physiology Manchester Proceedings*, **99**, 6210–6215.

Luo C and Rudy Y (1994a) A dynamic model of the cardiac ventricular action potential - simulations of ionic currents and concentration changes. *Circulation Research*, **74**, 1071–1097.

Luo C and Rudy Y (1994b) A dynamic model of the cardiac ventricular action potential: II. Afterdepolarizations, triggered activity and potentiation. *Circulation Research*, **74**, 1097–1113.

McAllister RE, Noble D and Tsien RW (1975) Reconstruction of the electrical activity of cardiac Purkinje fibres. *Journal of Physiology*, **251**, 1–59.

Mitchell MR, Powell T, Terrar DA and Twist VA (1984) The effects of ryanodine, EGTA and low-sodium on action potentials in rat and guinea-pig ventricular myocytes: Evidence for two inward currents during the plateau. *British Journal of Pharmacology*, **81**, 543–550.

Muzikant AL and Penland RC (2000) Models for profiling the potential QT prolongation risk of drugs. *Current opinion in Drug Discovery and Development*, **5**, 127–135.

Noble D (1962) A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pacemaker potentials. *Journal of Physiology*, **160**, 317–352.

Noble D (2002a) Modelling the heart: From genes to cells to the whole organ. *Science*, **295**, 1678–1682.

Noble D (2002b) Modelling the heart: Insights, failures and progress. *BioEssays*, **24**, 1155–1163.

Noble D (2002c) The rise of computational biology. *Nature Reviews Molecular Cell Biology*, **3**, 460–463.

Noble D (2002d) Unravelling the genetics and mechanisms of cardiac arrhythmia. *Proceedings of the National Academy of Sciences, United States of America*, **99**, 5755–5756.

Noble D and Tsien RW (1969) Outward membrane currents activated in the plateau range of potentials in cardiac Purkinje fibres. *Journal of Physiology*, **200**, 205–231.

Noble D, Varghese A, Kohl P and Noble PJ (1998) Improved guinea-pig ventricular cell model incorporating a diadic space, iKr & iKs, and length- & tension-dependent processes. *Canadian Journal of Cardiology*, **14**, 123–134.

Noble PJ and Noble D (2000) Reconstruction of the cellular mechanisms of cardiac arrhythmias triggered by early after-depolarizations. *Japanese Journal of Electrocardiology*, **20**(Suppl 3), 15–19.

Noma A (1983) ATP-regulated K+ channels in cardiac muscle. *Nature*, **305**, 147–148.

Nygren A, Fiset C, Firek L, Clark JW, Lindblad DS, Clark RB and Giles WR (1998) A mathematical model of an adult human atrial cell: The role of K+ currents in repolarization. *Circulation Research*, **82**, 63–81.

Panfilov A and Kerkhof P (2004) Quantifying ventricular fibrillation: In silico research and clinical implications. *IEEE Transactions on Bio-medical Engineering* **51**, 195–196.

Papadatos GA, Wallerstein P, Head C, Ratcliff R, Brady P, Benndorf K, Saumarez R, Trezise A, Huang C-H, Vandenberg JL, *et al*. (2004) Slowed conduction and ventricular tachycardia following targeted disruption of the cardiac sodium channel, Scn5a. *Proceedings of the National Academy of Sciences*, **286**, H1573–H1589.

Reuter H (1967) The dependence of slow inward current in Purkinje fibres on the extracellular calcium concentration. *Journal of Physiology*, **192**, 479–492.

Reuter H and Seitz N (1969) The dependence of calcium efflux from cardiac muscle on temperature and external ion composition. *Journal of Physiology*, **195**, 451–470.

Rice JJ, Jafri MS and Winslow RL (1999) Modeling gain and gradedness of Ca2+ release in the functional unit of the cardiac diadic space. *Biophysical Journal*, **77**, 1871–1884.

Rigg L and Terrar DA (1996) Possible role of calcium release from the sarcoplasmic reticulum in pacemaking in guinea-pig sino-atrial node. *Experimental Physiology*, **81**, 877–880.

Sanguinetti MC and Jurkiewicz NK (1990) Two components of cardiac delayed rectifier K^+ current. Differential sensitivity to block by class III antiarrhythmic agents. *Journal of General Physiology*, **96**, 195–215.

Shaw RM and Rudy Y (1997) Electrophysiological effects of acute myocardial ischemia: A mechanistics investigation of action potential conduction and conduction failure. *Circulation Research*, **80**, 124–138.

Ten Tusscher KH and Panfilov A (2003) Influence of nonexcitable cells on spiral breakup in twodimensional and three-dimensional excitable media. *Physical Review. E*, **68**, 062902.

Ten Tusscher KHWJ, Noble D, Noble PJ and Panfilov AV (2003) A model of the human ventricular myocyte. *American Journal of Physiology*, **12**, 962–972.

Trayanova N, Eason J and Aguel F (2002) Computer simulations of cardiac defibrillation: A look inside the heart. *Computing and Visualization in Science*, in press.

Wang Y and Rudy Y (2000) Action potential propagation in inhomogeneous cardiac tissue: Safety factor considerations and ionic mechanism. *American Journal of Physiology*, **278**, H1019–H1029.

Winslow R, Varghese A, Noble D, Adlakha C and Hoythya A (1993) Generation and propagation of triggered activity induced by spatially localised Na-K pump inhibition in atrial network models. *Proceedings of the Royal Society B*, **254**, 55–61.

Winslow RL, Greenstein JL, Tomaselli GF and O'Rourke B (1999) Computational models of the failing myocyte: Relating altered gene expression to cellular function. *Philosophical Transactions of the Royal Society, Series A*, **359**, 1187–1200.

Winslow RL, Scollan DF, Holmes A, Yung CK, Zhang J and Jafri MS (2000) Electrophysiological modeling of cardiac ventricular function: From cell to organ. *Annual Review Biomedical Engineering*, **2**, 119–155.

Zeng J, Laurita KR, Rosenbaum DS and Rudy Y (1995) Two components of the delayed rectifier K+ current in ventricular myocytes of the guinea pig type: Theoretical formulation and their role in repolarization. *Circulation Research*, **77**, 1–13.

Zhang H, Holden AV, Kodama I, Honjo H, Lei M, Varghese T and Boyett MR (2000) Mathematical models of electrical activity of central and peripheral sinoatrial node cells of the rabbit heart. *American Journal of Physiology*, **279**, H397–H421.

Specialist Review

Integrative modeling of the pancreatic β-cell

Arthur Sherman
National Institutes of Health, Bethesda, MD, USA

Richard Bertram
Florida State University, Tallahassee, FL, USA

1. Introduction

Mammalian metabolism is well adapted to switch between carbohydrate and lipid fuels depending on availability. One of the key hormones involved is insulin, which promotes glucose oxidation and fat storage when carbohydrate is abundant. In the developed world, unprecedented affluence has resulted in chronic excess of both nutrients relative to energy expenditure requirements, leading to an epidemic of obesity, hypertension, hyperlipidemia, and heart disease. An integral part of this metabolic syndrome is resistance to the effects of insulin, possibly as a compensation for overnutrition (Unger, 2003). In response, the β-cells of the pancreatic islets of Langerhans oversecrete insulin to maintain plasma glucose within normal ranges. In many cases (currently about 6% of the population of the United States), the β-cells fail to compensate, leading to hyperglycemia (Type II diabetes), further morbidity, and premature death. For these reasons, there has been much interest in understanding how β-cells regulate plasma glucose. This control mechanism centers on complex patterns of oscillation in membrane potential and cytosolic calcium, which are difficult to understand without the aid of mathematical models. Those models are the focus of this review.

2. Schematic model

The β-cell models build on the Hodgkin–Huxley paradigm (Hodgkin and Huxley, 1952), in which the plasma membrane is represented as an RC circuit, with the ion channels providing the resistance and the lipid bilayer providing the capacitance. The great success of such models in explaining the diverse electrical properties of neurons, muscle, and endocrine cells stems from the ability to reduce the electrical apparatus to a simple physical model, which can be represented by a small system of ordinary differential equations. The availability of dynamic readouts (voltage

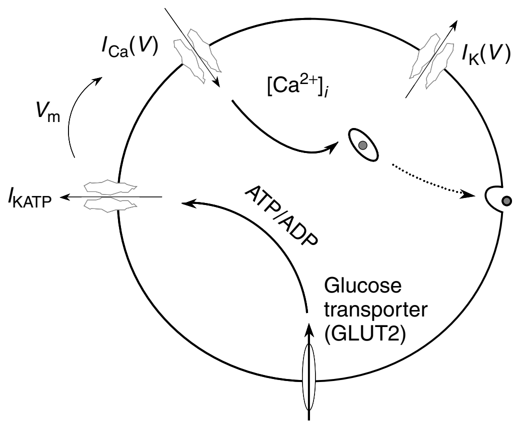


Figure 1 Consensus schematic of β-cell. Glucose triggers depolarization and Ca^{2+} entry, which signals exocytosis of insulin granules

and calcium) with high time resolution has also been crucial. The validity of isolating the electrical module was strikingly demonstrated by transfecting voltage-dependent Na^+ and K^+ channels into nonexcitable Chinese hamster ovary cells, which endowed them with neuron-like action potentials (Hsu *et al*., 1993).

The minimal parts list for the β-cell is shown in Figure 1. The voltage-dependent Ca^{2+} channel conducts Ca^{2+} ions into the cell, which raises the transmembrane voltage, V, whereas the K^+ channel gates efflux of K^+ and restores V to a low level. The temporal interaction of the two channels is sufficient to explain the repetitive spiking observed in β-cells. The Ca^{2+} influx also provides the primary chemical signal to trigger exocytosis of insulin-containing granules. These components are typical of many endocrine cells, but the β-cell needs an additional metabolic module to connect the level of electrical activity to plasma glucose concentration. Metabolism of glucose raises the ratio of ATP to ADP, which closes a third channel, the K(ATP) channel. Thus, in the absence of glucose, β-cells are electrically silent, but they generate Ca^{2+}-dependent action potentials when glucose is elevated.

This is the consensus schematic model for β-cells and is sufficient for understanding many aspects of diabetes pathology and therapy. For example, one of the classes of drugs used to treat Type II diabetes, the sulfonylureas, block K(ATP) channels independent of glucose metabolism and augment insulin secretion (Ashcroft and Rorsman, 1989). Also, recently it has been shown that either a surplus or a deficit of K(ATP) activity can lead to diabetes (Seino and Miki, 2003). In the former case, K(ATP) channels that fail to close in response to glucose result in inadequate secretion of insulin. In the latter case, K(ATP) channels that are always closed independent of glucose result in childhood hyperinsulinism. Some patients who survive this devastating hypoglycemia into adulthood become diabetic, possibly because their β-cells fail from chronic depolarization and exposure to high $[Ca^{2+}]$ (Glaser, 2003).

In both mice (Ashcroft and Rorsman, 1989) and humans (Martin and Soria, 1996), insulin secretion is controlled by oscillations of calcium, which are driven by *bursts* of action potentials with periods ranging from tens of seconds to several

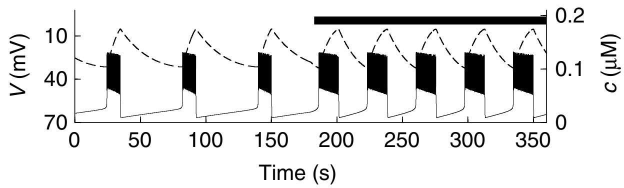


Figure 2 Simulation of glucose step with Chay–Keizer-like model (equations 1–5). Each plateau in voltage (V; solid) is a train of brief spikes. Cytosolic calcium (c; dashed) rises slowly to terminate the bursts and recovers slowly in the silent phase. The black bar indicates period of elevated glucose

minutes (see Figure 2 for an illustration using the model). We want to understand how bursting arises, how it is modulated by glucose and other signals, and how such a broad dynamic range is achieved. Three illustrative models are presented below in a didactic sequence of increasing complexity and explanatory power. Bear in mind that the actual historical development was more complex and subject to alternate interpretations.

3. The Chay–Keizer model

The first widely adopted mathematical model for bursting in β-cells was developed by Chay and Keizer (1983), on the basis of the hypothesis of Atwater and Rojas that the slow dynamics of intracellular free Ca^{2+} are responsible for packaging impulses into bursts (Atwater *et al*., 1980); that is, the sustained high voltage of the *active phase* of each burst would slowly increase cytosolic $[Ca^{2+}]$ (c), which would gradually activate a Ca^{2+}-activated K^+ (K(Ca)) channel until a critical level was reached that would shut the spiking off. In the absence of spiking, c would slowly recover, giving a long *silent phase* (Figure 2).

Chay and Keizer also sought to explain how an increase in plasma glucose modulates the bursts by prolonging the active phases and shortening the silent phases, that is, increasing the *plateau fraction*. They suggested that increased glucose would increase activity of the plasma membrane Ca^{2+}-ATPase (PMCA), slowing the rise of c and accelerating its fall. In keeping with the modular approach, the machinery of metabolism enters only through its effect on the pump rate (black bar, Figure 2).

The equations that correspond to the above scheme and that were used to produce Figure 2 are in outline:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = -\frac{I_{\mathrm{Ca}}(V) + I_{\mathrm{K}}(V,n) + I_{\mathrm{K(Ca)}}(V,c) + I_{\mathrm{K(ATP)}}(V)}{C_{\mathrm{m}}} \tag{1}$$

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{n_\infty(V) - n}{\tau_n} \tag{2}$$

$$\frac{\mathrm{d}c}{\mathrm{d}t} = f_{\mathrm{cyt}} J_{\mathrm{mem}} \tag{3}$$

(The full equations and parameter values are posted at **http://mrb.niddk.nih.gov/sherman**.) The electric circuit is represented by equations (1) and (2) for the membrane potential (V) and the activation level (open probability) of the voltage-dependent K^+ current (n), respectively. Equation 1 is Kirchhoff's law, which balances capacitive current with the four ionic currents, the voltage-dependent Ca^{2+} (I_{Ca}), voltage-dependent K^+ (I_K), Ca^{2+}-dependent K^+ ($I_{K(Ca)}$), and ATP-sensitive K^+ ($I_{K(ATP)}$) currents. Note that in order to obtain spikes, the RC circuit must be elaborated by assuming that the resistive elements (i.e., the ion channels) change their conducting state in response to V. These changes have been measured (Ashcroft and Rorsman, 1989) and incorporated semiquantitatively into the model. For the present, the conductance of $I_{K(ATP)}$ is held constant, unlike the conductances of the other currents, which vary as V or c change in time. Later, we will introduce dependence on glucose through the ATP production rate.

The slow dynamics of c are modeled by simple mass balance, assuming one well-mixed compartment. Influx through Ca^{2+} channels increases c, while efflux through the PMCA decreases it:

$$J_{mem} = -\alpha I_{Ca}(V) - k_{PMCA}c \tag{4}$$

(Note that I_{Ca} is negative by convention, so $-\alpha I_{Ca}$ is positive.) The net flux of Ca^{2+}, J_{mem}, is multiplied by f_{cyt}, the ratio of free to total (free + buffered) Ca^{2+}. f_{cyt} is small here (0.00025), which makes c slow. Calcium responds to the membrane potential through I_{Ca}, and in turn influences the membrane potential through $I_{K(Ca)} = g_{K(Ca)}\omega(V - V_K)$, where the activation variable ω is:

$$\omega = \frac{c^2}{c^2 + K_D^2} \tag{5}$$

A noteworthy feature of the Chay–Keizer model is that the voltage plateau is not explicitly included in the model. It is an emergent property arising from the $V - n$ interactions. This contrasts with an earlier model (Matthews and O'Connor, 1979), which postulated a special plateau current. This made for a more complex and cumbersome model than Chay–Keizer and may have contributed to its later neglect, in spite of its anticipation of many aspects of Chay–Keizer and other later models.

4. The endoplasmic reticulum: A second Ca^{2+} compartment

In addition to accounting for two major known properties of β-cells, the Chay–Keizer model made a prediction that cytosolic Ca^{2+} would oscillate, with a slow rise and fall. When Ca^{2+} was later imaged in β-cells (Valdeolmillos *et al.*, 1989), it was indeed seen to rise and fall, which was not surprising, but the kinetics were not as expected: c rose rapidly at the beginning of the active phase to a plateau and sometimes even fell slightly toward the end of the active phase. The K(Ca) channel hypothesis fell into disfavor, but has been revived, in part because

of the discovery of a candidate K(Ca) channel in β-cells (Goforth *et al*., 2002; Göpel *et al*., 1999), and in part because of an improved model due to Chay (1996), which included a second, internal Ca^{2+} compartment, representing the endoplasmic reticulum (ER).

There are several reasons for including the ER. As a storehouse for Ca^{2+}, it can influence the Ca^{2+} dynamics of both excitable and nonexcitable cells (Berridge, 1997). In islets in particular, releasing Ca^{2+} from the ER by either blockade of the ER Ca^{2+} pump (SERCA) or activation of IP_3 receptors dramatically affects electrical activity, Ca^{2+} oscillations, and secretion (Bertram *et al*., 1995b; Goforth *et al*., 2002; Worley *et al*., 1994).

The ER compartment is incorporated into the model by adding an ER-to-cytosol flux term (J_{ER}) to equation (3) and by adding a differential equation for the free ER Ca^{2+} concentration (c_{ER}). All forms of efflux are subsumed in a leakage pathway (J_{leak}), and influx occurs through SERCA pumps (J_{SERCA}):

$$J_{ER} = J_{leak} - J_{SERCA} \tag{6}$$

We neglect the known nonlinearities of these processes and write $J_{leak} = p_{leak}(c_{ER} - c)$ and $J_{SERCA} = k_{SERCA}c$. The effect of IP_3 can be represented simply by an increase in p_{leak}. The equations for c and c_{ER} are

$$\frac{dc}{dt} = f_{cyt}(J_{mem} + J_{ER}) \tag{7}$$

$$\frac{dc_{ER}}{dt} = -f_{ER}\left(\frac{V_{cyt}}{V_{ER}}\right)J_{ER} \tag{8}$$

where f_{ER} is the ER Ca^{2+} buffering factor, and V_{cyt} and V_{ER} are the cytosolic and ER volumes.

The behavior of the enhanced model is shown in Figure 3. The rapid jump in c at the beginning of each active phase reflects the intrinsic kinetics of cytosolic Ca^{2+}, which have been accelerated compared to Figure 2 by reducing the cytosolic buffer capacity (i.e., increasing f_{cyt} to the more realistic value of 0.01). This rise in c does not alone provide adequate inhibition to terminate the active phase, but as the ER slowly fills, c is pushed up to the necessary critical level. When spiking ceases, c jumps down rapidly and then slowly decays as the ER gives back the Ca^{2+} it took in during the active phase. Thus, like a buffer, the ER slows down c, but because the ER is slow (whereas buffers are fast), c displays two distinct kinetic phases, in agreement with experiments.

The model parameters in Figure 3 are chosen to give an oscillation period comparable to that typically observed in islets (10–60 s). However, if $g_{K(Ca)}$ is increased, the initial jump in c becomes sufficient to terminate an active phase without assistance from c_{ER}, resulting in a period of only a few seconds. (Similar fast bursting can also be obtained by partially emptying the ER. See Figure 4.) Conversely, reducing $g_{K(Ca)}$ (and possibly adjusting k_{PMCA}), produces periods of several minutes, which are also observed in experiments. This broad range of frequencies is possible because the time scale of oscillations is determined by

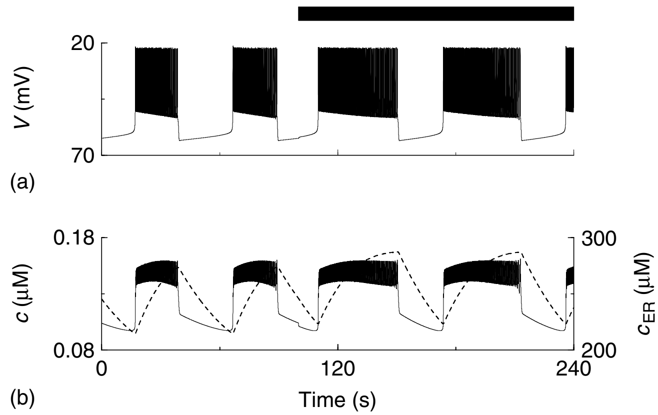


Figure 3 Chay–Keizer model with a slow ER compartment (c_{ER}, dashed). c now has fast and slow components, and increasing PMCA rate now increases both plateau fraction and mean c

mixing the relatively fast c and the relatively slow c_{ER} components. We call this "phantom bursting" because the time scale is not determined by any single process (Bertram and Sherman, 2004).

A comparison of Figures 2 and 3 shows that the ER enhances the model in another way. Whereas both the original Chay–Keizer model and the ER model explain how glucose can raise plateau fraction, in the one-compartment Chay–Keizer model, this does not lead to an increase in the mean level of cytosolic [Ca^{2+}]. In the ER model, in contrast, the approximately square shape of c means that an increase in plateau fraction produces an increase in the average level of c, without any change in the minimum and maximum values, as more time is spent at the higher level relative to the lower level. Note that this occurs in spite of the fact

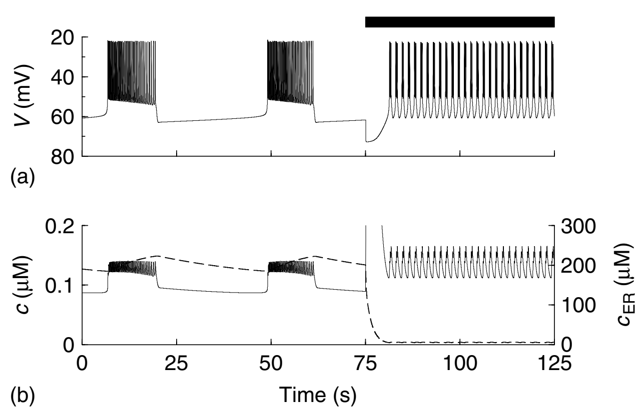


Figure 4 Response of model with ER to acetylcholine (ACh), modeled as a step increase in p_{leak}, throughout period denoted by black bar. Release of Ca^{2+} from ER (b; dashed) gives a transient sharp peak of c (solid; truncated for clarity). Plateau of high c is maintained by an ACh-activated inward current, which drives depolarized, fast bursting (a)

that the rate of pumping Ca^{2+} out of the cell is increased, illustrating that steady state intuitions may not apply when there are oscillations.

As alluded to earlier, inclusion of the ER also allows the model to account for the effects of acetylcholine (ACh) application (Figure 4). In vivo, ACh is released from parasympathetic nerve terminals in the pancreas to stimulate the islets in response to food ingestion (Woods and Porte, 1974). This auxiliary control mechanism raises insulin in anticipation of the coming glucose load and helps limit the rise in plasma glucose. At the cell level, ACh stimulates the production of IP_3, which activates IP_3 receptors on the ER. Here we again collapse an external regulatory module into a change in a parameter of the core electrical/calcium module: the efflux permeability p_{leak} is stepped up 45-fold. This is sufficient to produce the increase in burst frequency seen experimentally. However, in order to capture the rise in V and c that are also seen, we need to add a small depolarizing current ($I_{\mathrm{ACh}} = g_{\mathrm{ACh}}(V - V_{\mathrm{ACh}})$, where $g_{\mathrm{ACh}} = 5$ pS and $V_{\mathrm{ACh}} = 0$) to the V equation. Such a current has now been found (Mears and Zimliki, 2004; Rolland *et al*., 2002). The depolarized bursting and elevation of the cytosolic Ca^{2+} concentration are critical because they contribute to the increase in insulin secretion.

5. Oscillations in glucose metabolism

We now consider the glucose dependence of the K(ATP) channel. If the K(ATP) conductance is assumed to decrease with external glucose concentration, then plateau fraction is modulated similarly to the changes obtained in Figures 1 and 2 by varying PMCA pump rate (not shown). Beyond this, there is some evidence that ATP and ADP concentrations in β-cells vary over time (Ainscow and Rutter, 2002) in a Ca^{2+}-dependent manner (Kennedy *et al*., 2002). This was proposed by Keizer and Magnus (1989), who assumed that the rate of ATP production decreases when cytosolic Ca^{2+} inhibits the mitochondrial membrane potential. Phenomenologically, this can be represented by the following equation for ADP concentration:

$$\frac{\mathrm{d}[ADP]}{\mathrm{d}t} = k_{\mathrm{a}}\left[ATP - ADP\exp\left(r\left(1 - c/r_1\right)\right)\right] \tag{9}$$

We assume that the total nucleotide concentration, ADP + ATP, is constant, so no differential equation is needed for ATP. The effect of ATP to block the K(ATP) conductance is antagonized by ADP by competitive inhibition:

$$I_{\mathrm{K(ATP)}} = g_{\mathrm{K(ATP)}}\frac{1 + ADP/K_1}{1 + ADP/K_1 + ATP/K_2}(V - V_{\mathrm{K}}) \tag{10}$$

The net result is that a rise in c results in a slow (minutes) rise in ADP and an increase in $I_{\mathrm{K(ATP)}}$. An increase in glucose can now be represented by an increase in the parameter r in equation 9 in combination with changes in pump rates. We illustrate this in Figure 5 for the situation where glucose is raised from a subthreshold level to a level that permits bursting. Low glucose is simulated by reducing r and SERCA pump rate, k_{SERCA}, which drains the ER and elevates ADP. Starting in that condition, we assume that when glucose is raised, k_{SERCA} increases rapidly,

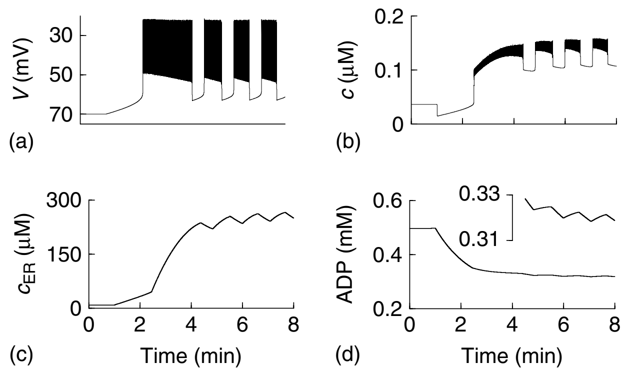


Figure 5 Model with ER and calcium feedback on nucleotide dynamics. Glucose is stepped at $T = 1$ min, initiating a triphasic transient (see text). In the steady state, ADP oscillates as well as cytosolic and ER calcium

producing a transient drop in c (Figure (5)b). Over several minutes, ADP declines, reducing K(ATP) conductance, and V gradually increases. When the voltage threshold for activation of I_{Ca} is reached, V jumps up sharply and spiking begins. The ER, however, is still depleted and acts as a sink for Ca^{2+}. This limits activation of the K(Ca) conductance, resulting in a prolonged period of spiking. After several more minutes, the ER is filled, and bursting can begin. ADP oscillates during the bursting phase (inset, Figure (5)d) and combines with c to drive the bursting. Thus, with inclusion of dynamics for ADP, the model is able to account for the complex triphasic transient of electrical and cytosolic Ca^{2+} activity observed experimentally (Bertram *et al*., 1995b; Goforth *et al*., 2002; Worley *et al*., 1994).

The incorporation of two very slow processes, c_{ER} and ADP, makes the model more robust and endows it with some redundancy. This may help account for the observations that oscillations can persist when the ER is disabled by blocking SERCA (Fridlyand *et al*., 2003) (K(ATP) can compensate for the loss of K(Ca)) or when K(ATP) channels are blocked by sulfonylureas (Rosario *et al*., 1993) (K(Ca) can compensate for the loss of K(ATP)). (See also Bertram and Sherman, 2004.) A third slow process recently proposed by Fridlyand *et al*. (2003), Na^+ accumulation that activates the Na^+ pump, might also be helpful to explain this complex phenomenology.

6. Summary and prospects

We have traced in outline the evolution of the Chay–Keizer family of models over 20 years. The first models explained bursting of membrane potential and modulation of plateau fraction by glucose. They also predicted oscillations of cytosolic Ca^{2+}, which were later seen, but with a different kinetic character. This led to a second wave of models with more accurate Ca^{2+} kinetics, and predictions of oscillations in ER Ca^{2+}. The latter have not yet been demonstrated experimentally, and indeed their existence has been challenged on the basis of ER measurements in

permeabilized cells (Tengholm *et al*., 2001). Measurements of ER Ca^{2+} in intact cells undergoing cytosolic Ca^{2+} oscillations are needed to test this prediction.

The current generation of models can account for the full range of observed oscillation periods, spanning 2 orders of magnitude, from seconds to minutes. They also account for modulation of electrical activity (and, implicitly, insulin secretion) by cholinergic stimuli as well as glucose (Figure 4), and the complex transients seen when glucose is first elevated (Figure 5). The initial prolonged spiking phase may be related to the early large peak of insulin secretion, which is lost in Type II diabetes (Kahn, 2001). However, a serious treatment of secretion will require consideration of the dynamics of vesicle exocytosis and recycling (Rorsman and Renstrom, 2003) in response to both local (submembrane) and bulk cytosolic Ca^{2+} and also to non-Ca^{2+} signals.

Future models will need to address other modulators, such as glucagon-like peptide, which potentiates secretion by elevating cAMP, and epinephrine, which blocks secretion and slows down oscillations. The former pathway is a drug target of great current interest because, unlike K(ATP) channel blockade, it is only effective when glucose is elevated, eliminating the danger of therapy-induced hypoglycemia. Recent data indicate that insulin may have both stimulatory and inhibitory effects on its own secretion (Khan *et al*., 2001). This is an unsettled area that merits further investigation both experimentally and theoretically.

We have included metabolic oscillations based on Ca^{2+}-dependent effects in the mitochondria (Figure 5). However, there is also some evidence for glycolytic oscillations (Tornheim, 1997). One possible role of glycolytic oscillations is in explaining the packaging of the bursts themselves into higher-order patterns, or "bursts of bursts" (Wierschem and Bertram, 2004), and other roles are currently under investigation.

For brevity, we have described here only models for a single cell, representing an electrically synchronized islet. For an overview of models that explicitly treat the nonobvious effects of electrical coupling, see Sherman (1996).

Thanks to judicious pruning of unnecessary detail, the current models are similar in complexity to the original Chay–Keizer model, in spite of their greater explanatory power. The use of simple models has paid dividends in terms of insight and mathematical analysis that goes beyond mere simulation. Of particular note is the establishment of a general classification scheme for burst mechanisms (Bertram *et al*., 1995a). This has not only enhanced our understanding of the β-cell models but also situated them in a general perspective of bursting models for neurons and other secretory cells.

The β-cell has provided an interesting case study of the interaction of physiology and modeling with genomic methods. For example, the K(ATP) channel was identified by physiological experiments, and its possible roles, as either a passive modulator or active participant in oscillations, have been clarified by modeling. In parallel, sequencing has suggested how its structure underlies its function, and genetic scans have uncovered connections to disease (hyperinsulinism and Type II diabetes). Yet, we do not yet know why, according to limited data, in rats glucose appears not to modulate plateau fraction, but rather gross depolarization (Antunes *et al*., 2000), similarly to ACh in mice (Figure 4). The islet α-cells in both species also possess K(ATP) channels, but their response is just the reverse

of β-cells – electrical activity and secretion are inhibited by glucose. How such different behaviors can arise in cell types that differ in rather subtle and quantitative ways poses a challenge for future studies integrating all the techniques at our disposal.

Related articles

See Article 115, **Systems biology of the heart**, Volume 6

Acknowledgments

This work was supported by NSF grant DMS-0311856 to R. Bertram.

References

Ainscow EK and Rutter GA (2002) Glucose-stimulated oscillations in free cytosolic ATP concentration imaged in single islet β-cells. *Diabetes*, **51**, S162–S170.

Antunes CM, Salgado AP, Rosario LM and Santos RM (2000) Differential patterns of glucose-induced electrical activity and intracellular calcium responses in single mouse and rat pancreatic islets. *Diabetes*, **49**(12), 2028–2038.

Ashcroft FM and Rorsman P (1989) Electrophysiology of the pancreatic β-cell. *Progress in Biophysics and Molecular Biology*, **54**, 87–143.

Atwater I, Dawson CM, Scott A, Eddlestone G and Rojas E (1980) The nature of the oscillatory behavior in electrical activity for the pancreatic β-cell. In *Biochemistry and Biophysics of the Pancreatic β-cell*, G, Thieme (Ed.), Verlag: New York, pp. 100–107.

Berridge MJ (1997) Elementary and global aspects of calcium signalling. *The Journal of Physiology*, **499**, 291–306.

Bertram R, Butte M, Kiemel T and Sherman A (1995a) Topological and phenomenological classification of bursting oscillations. *Bulletin of Mathematical Biology*, **57**, 413–439.

Bertram R and Sherman A (2004) A calcium-based phantom bursting model for pancreatic islets. *Bulletin of Mathematical Biology*, **66**(5), 1313–1344.

Bertram R, Smolen P, Sherman A, Mears D, Atwater I, Martin F and Soria B (1995b) A role for calcium release-activated current (CRAC) in cholinergic modulation of electrical activity in pancreatic β-cells. *Biophysical Journal*, **68**, 2323–2332.

Chay TR (1996) Electrical bursting and luminal calcium oscillation in excitable cell models. *Biological Cybernetics*, **75**, 419–431.

Chay TR and Keizer J (1983) Minimal model for membrane oscillations in the pancreatic β-cell. *Biophysical Journal*, **42**, 181–190.

Fridlyand LE, Tamarina N and Philipson LH (2003) Modeling of Ca^{2+} flux in pancreatic beta-cells: role of the plasma membrane and intracellular stores. *American Journal of Physiology. Endocrinology and Metabolism*, **285**(1), E138–E154.

Glaser B (2003) Dominant SUR1 mutation causing autosomal dominant type 2 diabetes. *Lancet*, **361**(9354), 272–273.

Goforth PB, Bertram R, Khan FA, Zhang M, Sherman A and Satin LS (2002) Calcium-activated K^+ channels of mouse β-cells are controlled by both store and cytoplasmic Ca^{2+}: experimental and theoretical studies. *The Journal of General Physiology*, **114**, 759–769.

Göpel SO, Kanno T, Barg S, Eliasson L, Galvanovskis J, Renström E and Rorsman P (1999) Activation of Ca^{2+}-dependent K^+ channels contributes to rhythmic firing of action potentials in mouse pancreatic β cells. *The Journal of General Physiology*, **114**, 759–769.

Hodgkin AL and Huxley AF (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**(4), 500–544.

Hsu H, Huang E, Yang XC, Karschin A, Labarca C, Figl A, Ho B, Davidson N and Lester HA (1993). Slow and incomplete inactivations of voltage-gated channels dominate encoding in synthetic neurons. *Biophysical Journal*, **65**(3), 1196–1206.

Kahn SE (2001) Clinical review 135: The importance of beta-cell failure in the development and progression of type 2 diabetes. *The Journal of Clinical Endocrinology and Metabolism*, **86**(8), 4047–4058.

Keizer J and Magnus G (1989) The ATP-sensitive potassium channel and bursting in the pancreatic beta cell. *Biophysical Journal*, **56**, 229–242.

Kennedy RT, Kauri LM, Dahlgren GM and Jung S-K (2002) Metabolic oscillations in β-cells. *Diabetes*, **51**, S152–S161.

Khan FA, Goforth PB, Zhang M and Satin LS (2001) Insulin activates ATP-sensitive K(+) channels in pancreatic beta-cells through a phosphatidylinositol 3-kinase-dependent pathway. *Diabetes*, **50**(9), 2192–2198.

Martin F and Soria B (1996) Glucose-induced [Ca^{2+}]i oscillations in single human pancreatic islets. *Cell Calcium*, **20**(5), 409–414.

Matthews EK and O'Connor MD (1979) Dynamic oscillations in the membrane potential of pancreatic islet cells. *The Journal of Experimental Biology*, **81**, 75–91.

Mears D and Zimliki CL (2004) Muscarinic agonists activate Ca^{2+} store-operated and -independent ionic currents in insulin-secreting HIT-T15 cells and mouse pancreatic beta-cells. *The Journal of Membrane Biology*, **197**(1), 59–70.

Rolland JF, Henquin JC and Gilon P (2002) G protein-independent activation of an inward Na(+) current by muscarinic receptors in mouse pancreatic beta-cells. *The Journal of Biological Chemistry*, **277**(41), 38373–38380.

Rorsman P and Renstrom E (2003). Insulin granule dynamics in pancreatic beta cells. *Diabetologia*, **46**(7), 1029–1045.

Rosario LM, Barbosa RM, Antunes CM, Silva AM, Abrunhosa AJ and Santos RM (1993) Bursting electrical activity in pancreatic beta-cells: evidence that the channel underlying the burst is sensitive to Ca^{2+} influx through L-type Ca^{2+} channels. *Pflugers Archiv: European Journal of Physiology*, **424**(5–6), 439–447.

Seino S and Miki T (2003). Physiological and pathophysiological roles of ATP-sensitive K^+ channels. *Progress in Biophysics and Molecular Biology*, **81**(2), 133–176.

Sherman A (1996) Contributions of modeling to understanding stimulus-secretion coupling in pancreatic β-cells. *The American Journal of Physiology*, **271**, E362–E372.

Tengholm A, Hellman B and Gylfe E (2001) The endoplasmic reticulum is a glucose-modulated high-affinity sink for Ca^{2+} in mouse pancreatic beta-cells. *The Journal of Physiology*, **530**(Pt 3), 533–540.

Tornheim K (1997) Are metabolic oscillations responsible for normal oscillatory insulin secretion? *Diabetes*, **46**, 1375–1380.

Unger and RH (2003). Lipid overload and overflow: metabolic trauma and the metabolic syndrome. *Trends in Endocrinology and Metabolism: TEM*, **14**(8), 398–403.

Valdeolmillos M, Santos RM, Contreras D, Soria B and Rosario LM (1989) Glucose-induced oscillations of intracellular Ca^{2+} concentration resembling electrical activity in single mouse islets of Langerhans. *FEBS Letters*, **259**, 19–23.

Wierschem K and Bertram R (2004) Complex bursting in pancreatic islets: a potential glycolytic mechanism. *Journal of Theoretical Biology*, **228**(4), 513–521.

Woods SC and Porte DJ (1974) Neural control of the endocrine pancreas. *Physiological Reviews*, **54**, 596–619.

Worley JF, McIntyre MS, Spencer B, Mertz RJ, Roe MW and Dukes ID (1994) Endoplasmic reticulum calcium store regulates membrane potential in mouse islet β-cells. *The Journal of Biological Chemistry*, **269**, 14359–14362.

Short Specialist Review

EGFR network

Stanislav Y. Shvartsman
Princeton University, Princeton, NJ, USA

In 1962, a peptide purified from the salivary gland of a mouse was shown to accelerate incisor eruption and eyelid opening in newborn mice. The peptide was named the epidermal growth factor (EGF) (Cohen, 1962). Soon thereafter, EGF and members of the family of peptide growth factors had been identified in countless physiological and pathological contexts. EGF binds to a cell surface receptor (EGFR, epidermal growth factor receptor), inducing its dimerization and phosphorylation of several tyrosines residues within its cytoplasmic tail. The phosphorylated tyrosines provide binding sites for cytoplasmic proteins; this couples the activated receptor to the signal transduction cascades and, eventually, to gene expression and processes such as cell division, differentiation, or migration. Abnormal EGFR signaling due to overactive receptors or overexpressed ligands leads to developmental defects and is also associated with many types of cancers. This much was understood about the EGF system when the 1986 Nobel Prize in Medicine was awarded to Stanley Cohen and Rita Levi-Montalcini for their discovery of EGF and NGF (nerve growth factor).

Today, the EGFR is the subject of some 30 000 research papers. Many individual molecules mediating the EGFR-induced responses have been identified and are drug targets in oncology and other areas of medicine (Yarden and Sliwkowski, 2001). Genomics and proteomics approaches will soon make it possible to follow all genes and protein/protein interactions affected by EGFR/ligand binding on the cell surface. The structural details of EGFR interaction with its ligands are becoming progressively understood at the atomistic level and can be followed in real time with modern imaging tools (Burgess *et al*., 2003). However, neither the contribution of EGFR to tissue morphogenesis in development nor the exact role of deregulated EGFR signaling in disease processes is understood at this time. For example, the mechanistic explanation of the EGF-induced eyelid opening is only beginning to emerge and the attempts to use EGFR expression as prognostic marker in cancer have met with only limited success (Arteaga, 2003).

Quantitative models are necessary in order to deal with the immense complexity associated with the large number of signaling components and multiple feedback loops in the EGFR network (Wiley *et al*., 2003). Synthesis of the existing information into quantitative models is becoming possible due to the highly conserved nature of EGFR signaling and the ability to test the constructed models in well-developed experimental systems. Only a few dozen out of

~30 000 EGFR-related Medline entries are dedicated to modeling and computational analysis. The existing models can be classified into three groups. Models of binding and trafficking describe the levels and cellular location of free and ligand-bound receptors. Explicit models of signal transduction describe the biochemical transformations induced by extracellular ligands at the level of a single cell. Finally, molecular modeling has provided insight into the structural biology of the EGFR system and has guided the design of receptor tyrosine kinase (RTK) inhibitors.

Quantitative models of ligand–receptor binding and trafficking have been indispensable in dissecting the individual steps in the cycle of receptor (and ligand) endocytosis (Wiley, 2003). Using these models to fit data obtained for the kinetics of ligand-induced receptor internalization and degradation, it has been possible to extract the rate constants for different steps within the endocytic cycle. The extracted constants have explained the experimentally observed differences in the biological effects of different EGFR ligands (notably EGF and TGF-α). At the same time, the pharmacological effects of drugs can now be interpreted and predicted in terms of their effects on different steps in receptor-mediated endocytosis. Importantly, the models developed for the EGFR have been successfully used in the quantitative analysis of other ligand–receptor systems.

Several computational models have been used to interpret and plan EGFR signal transduction experiments in cell culture assays. These models accounted for ligand–receptor binding, receptor dimerization, endocytosis, and signal transduction through the PLCγ and MAPK cascades (Schoeberl *et al*., 2002; Kim *et al*., 1999; Haugh *et al*., 2000). The models were used to describe the signals induced by a steplike change in the concentration of exogenous ligand. The predictive capability of these models was demonstrated in two separate cell culture experiments, even though the parameters used for the computational models had been assembled from measurements in multiple cell types and from several in vitro studies.

Almost exclusively, current models of EGFR signaling focus on events at the level of a single cell (or at the molecular level). However, many of the most important in vivo effects of EGFR signaling involve interactions between multiple cells. There are no existing models that describe EGFR signaling at this level, and there is relatively little quantitative understanding that can be translated into directed manipulation of tissue regulation and development. Systematic analysis of EGFR-mediated cell communication requires computational models that simultaneously account for intracellular events *and* intercellular communication by locally produced EGFR ligands. Only with approaches that consider all of these elements simultaneously can predictions be made that are of value on the organismal scale. Given this complexity, integrated models are nontrivial to test experimentally. Appropriate experimental paradigms that can be used to directly test predictions of complex models are critically needed. Cultured epithelial layers and model organisms of developmental genetics, such as the fruit flies, show great promise for achieving this goal (Shvartsman *et al*., 2002; Vermeer *et al*., 2003; Casci and Freeman, 1999).

Even though much of the quantitative studies of EGFR network have assumed that it acts as an independent signaling module, it is highly unlikely that EGFR acts alone in mediating a specific cell or tissue response. Fortunately, an increasing amount of information on how the EGFR system interacts with other pathways,

such as TGFβ and cytokine receptor pathways, promises to help us understand how cells integrate signals from multiple pathways to produce a final response.

References

Arteaga CL (2003) ErbB-targeted therapeutic approaches in human cancer. *Experimental Cell Research*, **284**(1), 122–130.

Burgess AW, Cho HS, Eigenbrot C, Ferguson KM, Garrett TP, Leahy DJ, Lemmon MA, Sliwkowski MX, Ward CW and Yokoyama S (2003) An open-and-shut case? Recent insights into the activation of EGF/ErbB receptors. *Molecular Cell*, **12**(3), 541–552.

Casci T and Freeman M (1999) Control of EGF receptor signalling: Lessons from fruitflies. *Cancer and Metastasis Reviews*, **18**, 181–201.

Cohen S (1962) Isolation of a mouse sugmaxillary gland protein accelerating incisor eruption and eyelid opening in new-born animal. *The Journal of Biological Chemistry*, **237**(5), 1555.

Haugh JM, Wells A and Lauffenburger DA (2000) Mathematical modeling of epidermal growth factor receptor signaling through the phospholipase C pathway: mechanistic insights and predictions for molecular interventions. *Biotechnology and Bioengineering*, **70**(2), 225–238.

Kim HG, Kassis J, Souto JC, Turner T and Wells A (1999) EGF receptor signaling in prostate morphogenesis and tumorigenesis. *Histology and Histopathology*, **14**(4), 1175–1182.

Schoeberl B, Eichler-Jonsson C, Gilles ED and Muller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, **20**(4), 370–375.

Shvartsman SY, Muratov CB and Lauffenburger DA (2002) Modeling and computational analysis of EGF receptor-mediated cell communication in Drosophila oogenesis. *Development*, **129**(11), 2577–2589.

Vermeer PD, Einwalter LA, Moninger TO, Rokhlina T, Kern JA, Zabner J and Welsh MJ (2003) Segregation of receptor and ligand regulates activation of epithelial growth factor receptor. *Nature*, **422**(6929), 322–326.

Wiley HS (2003) Trafficking of the ErbB receptors and its influence on signaling. *Experimental Cell Research*, **284**(1), 78–88.

Wiley HS, Shvartsman SY and Lauffenburger DA (2003) Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends in Cell Biology*, **13**(1), 43–50.

Yarden Y and Sliwkowski MX (2001) Untangling the ErbB signalling network. *Nature Reviews. Molecular Cell Biology*, **2**(2), 127–137.

Data collection and analysis in systems biology

Trey Ideker
University of California, San Diego, CA, USA

1. Introduction

An exciting trend in molecular biology involves the use of systematic genomic, proteomic, and metabolomic technologies to construct large-scale models of biological systems. These endeavors, collectively known as *systems biology* (Ideker *et al*., 2001a; Kitano, 2002), establish a paradigm by which to systematically interrogate, model, and iteratively refine our knowledge of the cell. While the field of systems science has existed for some time (Ashby, 1958; Bertalanffy, 1973), systems approaches have recently generated a great deal of excitement in biology due to a host of new experimental technologies that are high throughput, quantitative, and large scale. Owing to the time and expense associated with large-scale measurements as well as the enormous amounts of data produced, principled strategies will be indispensable for using these data to construct biological models.

2. Large-scale experimental methods

The first technology to revolutionize modern biology was automated DNA sequencing (Hood *et al*., 1987), which was instrumental in defining the list of 30 000 genes in the human genome (Lander *et al*., 2001; Venter *et al*., 2001). More recently, DNA microarrays have enabled simultaneous measurement of all gene states to reveal which genes are expressed (i.e., turned on vs. off) in a particular cell type or biological condition (Slonim, 2002). Other molecular states, such as changes in protein levels (Gygi *et al*., 1999), phosphorylation states (Zhou *et al*., 2001), and metabolite concentrations (Griffin *et al*., 2001), can be quantified with mass spectrometry, nuclear magnetic resonance, and other advanced technologies. A final cellular state measurement that is gaining in importance is the genomic phenotyping experiment (Begley *et al*., 2002), also called *parallel phenotypic analysis* (Deutschbauer *et al*., 2002). In this type of experiment, a library of

gene knockouts is screened to identify which genes are essential for a particular phenotype. In single-celled organisms, the phenotype associated with each gene is typically the growth rate, but it can really be any measure of the phenotypic consequences of perturbing a gene.

Of the approaches for characterizing cellular states, measurements made by DNA microarrays are currently the most comprehensive (every mRNA species is detected); high throughput (a single technician can assay multiple conditions per week); well characterized (experimental error is appreciable, but understood); and cost-effective (whole-genome microarrays are purchased commercially for US $50 to $1000, depending on the organism). However, continued advances in protein labeling and separation technology are making measurement of protein abundance and phosphorylation state almost as feasible, with the primary barrier being the expense and expertise required to set up and manage a mass spectrometry facility. Measurement of metabolite concentrations, an endeavor otherwise known as *metabonomics* (Nicholson *et al*., 2002), is currently limited not by detection (thousands of peaks, each representing a different molecular species, are found in a typical NMR spectrum) but by identification (matching each peak with a chemical structure is difficult). Clearly, measuring changes in cellular state at the protein and metabolic levels will be crucial if we are to gain insight into not only regulatory pathways but also those pertaining to the cell's signaling and metabolic circuitry.

Equally as exciting, another set of recent technological advances have enabled us to characterize DNA and protein interaction *networks.* Several methods are available for measuring protein–protein interactions at large scale – two of the most popular being the yeast two-hybrid system (Uetz *et al*., 2000b; Fields and Song, 1989) and protein co-immunoprecipitation (coIP) followed by mass spectrometry (Gavin *et al*., 2002; Ho *et al*., 2002a). Protein–DNA interactions, which commonly occur between transcription factors and their DNA binding sites, constitute another interaction type that can now be measured at high throughput using the technique of Chromatin ImmunoPrecipitation followed by promoter microarray chip analysis (ChIP-chip) (Iyer *et al*., 2001; Ren *et al*., 2000). Large protein–protein or protein–DNA interaction data sets are now available for a variety of species including *Saccharomyces cerevisiae* (Uetz *et al*., 2000a; Lee *et al*., 2002; Ito *et al*., 2001; Ho *et al*., 2002b; Gavin *et al*., 2002), *Helicobacter pylori* (Rain *et al*., 2001), *Drosophila melanogaster* (Giot *et al*., 2003), and *Caenorhabditis elegans* (Walhout *et al*., 2000; Li *et al*., 2004). Additional types of molecular interactions, such as those between proteins and small molecules (carbohydrates, lipids, drugs, hormones, and other metabolites), are difficult to measure at large scale, although protein array technology (MacBeath and Schreiber, 2000; Zhu *et al*., 2001; Haab *et al*., 2001) might enable high-throughput measurement of protein–small molecule interactions in the near future. A current drawback of high-throughput interaction measurements is a potentially high error rate (Deane *et al*., 2002). An emerging approach for addressing this problem is to construct models that integrate several complementary data sets together (e.g., two-hybrid interactions with coIP data or gene expression profiles) to reinforce the common signal (Bar-Joseph *et al*., 2003; Hanisch *et al*., 2002; Ideker *et al*., 2002; Jansen *et al*., 2002; Yeger-Lotem and Margalit, 2003; Ge *et al*., 2001).

3. Modeling and systems analysis

The enormous amount of data arising from high-throughput biology provides a more complete picture of cellular function than ever before. However, new data sets are being generated at a rate that far outpaces our ability to analyze and interpret the results – a disparity that has thus far limited the impact of these data on basic biomedical research. Eliminating this disparity therefore presents a number of grand challenges to computational researchers: how to best associate high-level information about proteins and protein interactions with functional roles; how to enrich the true biological signal in noisy data; and, most importantly, how to organize global measurements at different levels into full-fledged models of cellular signaling and regulatory machinery.

Systems biology attempts to address these goals by integrating the various levels of global measurements together and with a mathematical model of a biological system or pathway of interest. Although these model-driven approaches may differ in the particulars of implementation, all follow a fundamental framework involving the following four distinct steps (Figure 1):

1. *Define the System Components.* Discover all of the genes in the genome as well as the particular molecules and molecular interactions that constitute the pathway of interest. If possible, define an initial model of how these molecular components and interactions relate to govern pathway function.
2. *Perturb the System.* Perturb each pathway component through a series of genetic or environmental manipulations. Detect and quantify the corresponding global cellular response to each perturbation, using genomic, proteomic, and/or metabolomic technologies.

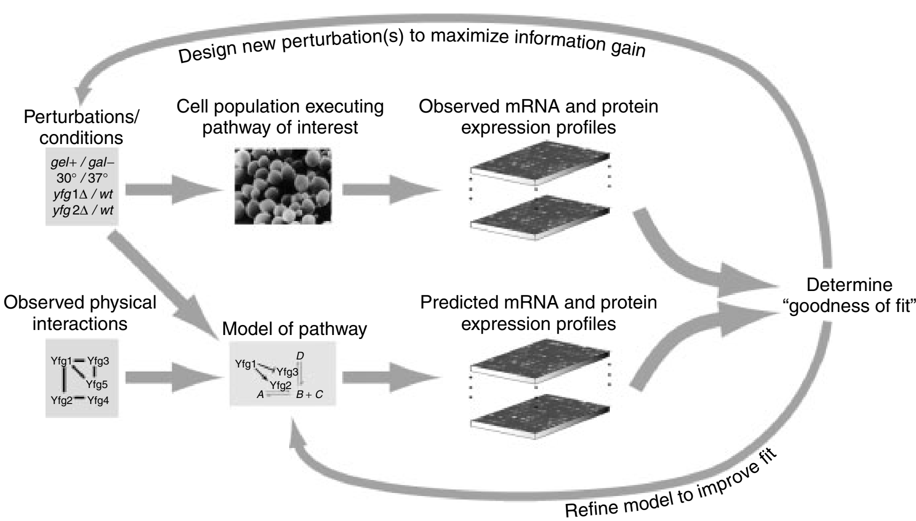


Figure 1 A systems approach to biology. (Reprinted, with permission, from the Annual Review of Genomics and Human Genetics, Volume 2 © 2001 by Annual Reviews **www.annualreviews.org**)

3. *Model Reconciliation.* Integrate the observed responses with the current pathway model and with the global networks of protein–protein interactions, protein–DNA interactions, and biochemical reactions.
4. *Model Verification/Expansion.* Formulate new hypotheses to explain observations that are not predicted by the model. Design additional perturbation experiments to test these and iteratively repeat steps (2), (3), and (4).

Steps (1) and (2) are focused on biological *discovery* through construction of a library of potential components, interactions, and system responses. Steps (3) and (4) are driven by a set of hypotheses encoded by the computational models. Systems approaches following this general framework have been used most actively to interrogate pathways in model organisms, including yeast (Forster *et al*., 2003; King *et al*., 2004; Ideker *et al*., 2001b; Bar-Joseph *et al*., 2003), *Escherichia coli* (Gardner *et al*., 2003), *Halobacterium* (Baliga *et al*., 2002), and sea urchin (Davidson *et al*., 2002). These works provide a roadmap of how to model cellular processes through large-scale measurement and integration of biological data.

Central to this systems approach are computer-aided models for understanding and interrogating complex cellular processes. These models promise to revolutionize biology and medicine by providing a comprehensive blueprint of normal and diseased cell functions and by allowing researchers to simulate the effects of drugs on cells long before they are tested in humans. Several strategies have been developed for integrating gene expression profiles with other large-scale data to formulate models of regulatory networks, including Bayesian learning, neural nets, process algebra, and systems of differential equations [for a review, see van Someren (van Someren *et al*., 2002)]. Computer-aided approaches for experimental design are equally as important – that is, designing perturbations to most effectively and efficiently verify and expand the models in step (4) above (Tong and Koller, 2001; King *et al*., 2004; Ideker *et al*., 2000).

With the recent appearance of large networks of protein–protein and protein–DNA interactions as a new type of measurement, systems researchers are trying to identify the specific network structures and network "design principles" that have been most favored by evolution. These efforts have shown that biological networks encode modular functional units (Rives and Galitski, 2003; Ravasz *et al*., 2002) that have likely evolved to be robust to perturbation (Jeong *et al*., 2001). Moreover, these modules often contain recognizable configurations such as the feedback and feed-forward loops that are also prevalent in electronic circuitry and other types of man-made systems (Milo *et al*., 2002). Several other groups have proposed methods for constructing regulatory models of the cell, using molecular interaction networks as the central framework (Bar-Joseph *et al*., 2003; Kelley *et al*., 2003; Hanisch *et al*., 2002; Ideker *et al*., 2002; Jansen *et al*., 2002; Yeger-Lotem and Margalit, 2003; Ge *et al*., 2001). The key idea is that, by identifying which parts of the molecular interaction network correlate most strongly with other biological evidences such as gene expression profiles or genomic phenotypes, it will be possible to organize the network into circuit modules representing the repertoire of distinct functional processes in the cell. Several approaches are also available for identification of metabolic pathways or protein complexes that have

been conserved over evolution (Kelley *et al*., 2003; Forst and Schulten, 2001; Dandekar *et al*., 1999). Evolutionarily conserved pathways allow interpretation of the network of a poorly understood organism based on its similarity to that of a well-known species. These tools also have application to infectious disease, for example, by targeting drugs to pathways that are present in a pathogenic organism but absent from its human host.

4. Current and future challenges

Systems biology now faces several major challenges. First, given a first-generation toolbox of modeling and systems approaches, an immediate next step is to leverage these tools to elucidate disease pathways. However, although systems approaches have been successfully applied to species such as *S. cerevisiae* (Gavin *et al*., 2002; Ho *et al*., 2002a; Habeler *et al*., 2002; Gygi *et al*., 1999; Haab *et al*., 2001; Kumar *et al*., 2002; Lee *et al*., 2002; Tong *et al*., 2001; Uetz *et al*., 2000b), similar studies in medically relevant organisms such as human, mouse, and rat have remained largely out of reach. This disparity is due to the increased complexity of mammalian systems; technical and ethical problems in subjecting them to perturbation; and a relative lack of experimental data on protein–protein, protein–DNA, or other molecular interactions for human. Encouragingly, ongoing data production efforts in human cell lines (e.g., the Alliance for Cell Signaling; **http://www.afcs.org**) and emerging technologies such as RNAi gene knockdowns (Dykxhoorn *et al*., 2003) promise to address many of these difficulties in the near future. In the meantime, yeast remains an attractive alternative for studying basic biological pathways that influence pathogenesis and genetic disorders.

A second challenge is that almost all previous systems biology studies have been strongly reliant on a preexisting model of the pathway of interest. For instance, in our case study of the galactose-utilization pathway (Ideker *et al*., 2001b), the initial components and interactions of the model were drawn directly from review articles and the primary literature. An initial literature-based model was also indispensable for the now classic explorations of bacterial chemotaxis (Barkai and Leibler, 1997) and infection by phage lambda (Arkin *et al*., 1998; McAdams and Shapiro, 1995). These studies significantly expanded our biological knowledge, but if systems approaches continue to be successful, they will quickly exhaust the few well-studied biological systems that are available. Therefore, systems biology will only become a sustainable paradigm if it can generate new models in the absence of extensive prior knowledge.

If these challenges can be met, it is likely that systems and modeling approaches will have substantial payoffs in basic medicine as well as the pharmaceutical industry. In this regard, it is revealing that outside of biotechnology, many sectors of manufacturing already depend heavily on computer simulation and modeling for product development. Using computer-aided design (CAD) tools, digital circuit manufacturers explore the wiring of transistors and other components on the silicon wafer, just as automotive engineers estimate how many miles per gallon is to be expected from the next-generation sedan long before it is built on the assembly line. Biology will undoubtedly also benefit from these “classical” engineering

approaches. Given that more than six out of every seven drugs that undergo human testing ultimately fail because of unanticipated side effects, systems modeling may act as a much-needed filter between high-throughput screening for drug candidates and the time-consuming and costly follow up of human trials.

5. Perspective

The field of systems biology still faces many challenges but also holds much promise. By increasing our basic repertoire of experimental strategies and modeling approaches, systems biology provides the starting point for advances in many facets of biotechnology, not least of which is an enhanced ability to appropriately target therapeutics in diseased cells. Thus, we can move one step closer to the day when systems modeling techniques will have widespread influence on basic biological research and replace high-throughput screening as a de facto standard in drug development.

References

Arkin A, Ross J and McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, **149**, 1633–1648.

Ashby R (1958) General Systems Theory as a New Discipline. *General Systems Yearbook*, **3**, 1–6.

Baliga NS, Pan M, Goo YA, Yi EC, Goodlett DR, Dimitrov K, Shannon P, Aebersold R, Ng WV and Hood L (2002) Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14913–14918.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **21**, 1337–1342.

Barkai N and Leibler S (1997) Robustness in simple biochemical networks. *Nature*, **387**, 913–917.

Begley TJ, Rosenbach AS, Ideker T and Samson LD (2002) Damage Recovery Pathways in Saccharomyces cerevisiae Revealed by Genomic Phenotyping and Interactome Mapping. *Molecular Cancer Research*, **1**, 103–112.

Bertalanffy Lv (1973) *General Systems Theory: Foundations, Development, Applications*, Penguin: Harmondsworth.

Dandekar T, Schuster S, Snel B, Huynen M and Bork P (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *The Biochemical journal*, **343**, 115–124.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, *et al*. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

Deane CM, Salwinski L, Xenarios I and Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular and Cellular Proteomics*, **1**, 349–356.

Deutschbauer AM, Williams RM, Chu AM and Davis RW (2002) Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomycescerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15530–15535.

Dykxhoorn DM, Novina CD and Sharp PA (2003) Killing the messenger: short RNAs that silence gene expression. *Nature Reviews. Molecular Cell Biology*, **4**, 457–467.

Fields S and Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Forst CV and Schulten K (2001) Phylogenetic analysis of metabolic pathways. *Journal of Molecular Evolution*, **52**, 471–489.

Forster J, Famili I, Fu P, Palsson BO and Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research*, **13**, 244–253.

Gardner TS, di Bernardo D, Lorenz D and Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ge H, Liu Z, Church GM and Vidal M (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics*, **29**, 482–486.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.

Griffin JL, Mann CJ, Scott J, Shoulders CC and Nicholson JK (2001) Choline containing metabolites during cell transfection: an insight into magnetic resonance spectroscopy detectable changes. *FEBS Letters*, **509**, 263–266.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Haab BB, Dunham MJ and Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, **2**.

Habeler G, Natter K, Thallinger GG, Crawford ME, Kohlwein SD and Trajanoski Z (2002) YPL.db: the Yeast Protein Localization database. *Nucleic Acids Research*, **30**, 80–83.

Hanisch D, Zien A, Zimmer R and Lengauer T (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**(Suppl 1), S145–S154.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002a) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*. (2002b) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Hood LE, Hunkapiller MW and Smith LM (1987) Automated DNA sequencing and analysis of the human genome. *Genomics*, **1**, 201–212.

Ideker T, Galitski T and Hood L (2001a) A new approach to decoding life: Systems Biology. *Annual Review of Genomics and Human Genetics*, **2**, 343–372.

Ideker T, Ozier O, Schwikowski B and Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**(Suppl 1), S233–S240.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Bumgarner R, Aebersold R and Hood L (2001b) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ideker TE, Thorsson V and Karp RM (2000) Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design. *Pacific Symposium on Biocomputing*, 305–316.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M and Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Jansen R, Greenbaum D and Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Research*, **12**, 37–46.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR and Ideker T (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11394–11399.

King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, Kell DB and Oliver SG (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, **427**, 247–252.

Kitano H (2002) Computational systems biology. *Nature*, **420**, 206–210.

Kumar A, Cheung KH, Tosches N, Masiar P, Liu Y, Miller P and Snyder M (2002) The TRIPLES database: a community resource for yeast molecular biology. *Nucleic Acids Research*, **30**, 73–75.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al*. (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

MacBeath G and Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

McAdams HH and Shapiro L (1995) Circuit simulation of genetic networks. *Science*, **269**, 650–656.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Nicholson JK, Connelly J, Lindon JC and Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews. Drug Discovery*, **1**, 153–161.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of Helicobacter pylori. *Nature*, **409**, 211–215.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rives AW and Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1128–1133.

Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, **32** (Suppl 1), 502–508.

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Tong, S and Koller, D (2001) Active Learning for Structure in Bayesian Networks. *Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, Washington, 863–869.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000a) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000b) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae [see comments]. *Nature*, **403**, 623–627.

van Someren EP, Wessels LF, Backer E and Reinders MJ (2002) Genetic network modeling. *Pharmacogenomics*, **3**, 507–525.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N and Vidal M (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.

Yeger-Lotem E and Margalit H (2003) Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation. *Nucleic Acids Research*, **31**, 6053–6061.

Zhou H, Watts JD and Aebersold R (2001) A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnology*, **19**, 375–378.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.

Introductory Review

Contig mapping and analysis

Asim Siddiqui and Steven Jones
British Columbia Cancer Agency, Vancouver, BC, Canada

The creation of ordered sets of overlapping clones or "contigs" has historically been the goal of chromosome walks in gene hunting and more recently in providing tiling paths of clones for whole genome sequencing. Various methods have been used to establish clone overlaps, including simple cross hybridization. In radiation hybrid mapping, the overlaps between the propagated large genomic fragments are detected through shared Sequence Tags Sites (STSs) detected through PCR amplification (Cox *et al*., 1990). Currently, the most common way to establish clone overlaps is through shared restriction fragments.

Restriction enzyme fingerprint mapping has undergone a number of iterations in its development. However, the basic concept remains the same. A set of clone fragments is derived by cutting the clone with a restriction endonuclease that matches a very short (4 or 6 bp) nucleotide sequence. The fragments from each clone are then sized using an electrophoresis technique. Clones that overlap will share a number of similarly sized fragments. The earliest methods used radioactive labeling and acrylamide gels (Nathans, 1979). Using an essentially similar method, Coulson *et al*. (1986) generated a preliminary version of the *Caernorhabditis elegans* physical map by using Cosmid clones libraries. The contiguation of this map was later substantially improved through the addition of YAC clones to the contigs (Coulson *et al*., 1988). The larger insert size of YAC clones allowed them to bridge gaps in the coverage of the Cosmid clone map. The relatively small size of cosmids, approximately 30–40 kb, remained a limitation in the ability of such fingerprinting techniques to map large plant and mammalian genomes until the development of Bacterial Artificial Chromsomes (BACs) with a much larger insert size of between 180 and 200 kb (*see* Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3). Other improvements, such as the use of agarose instead of acrylamide gels also made the approach amenable to high-throughput data generation and has been used to create physical maps for a number of mammalian genomes including human (McPherson *et al*., 2001), mouse (Gregory *et al*., 2002), and rat (Krzywinski *et al*., 2004). In this chapter, we will discuss the methods and approaches used in contiguating clones on the basis of restriction fragment fingerprint information.

A genomic library consists of a set of overlapping clones representing coverage of the whole genome. By identifying common overlap regions between clones, clones that have adjacent positions in the complete genome may be found and

joined together in a common contig. The level of redundancy in the genome coverage of a library is often quoted as "X coverage". For example, in a 4X genomic library, any section of DNA is represented on average in four different clones – see Figure 1. The average coverage can be calculated by simply multiplying the average insert size of the library by the number of clones in the library and dividing by an estimate of the genome size. Since short clone overlaps are difficult to reliably detect, the joining of the clones is more likely to be reliable with a high redundancy of clone coverage, which will also reduce the number of gaps in clone coverage.

To generate a restriction enzyme fingerprint map, the clone insert is cut by a restriction endonuclease that recognizes a short, specific nucleotide sequence. This breaks the clone into a series of fragments, the largest of which are around 30 kb in length. By separating the clone fragments using an electrophoresis gel, it is possible to size each of the fragments. This record of the sizes of fragments represents the fingerprint of the clone. Owing to limitations of gel electrophoresis, fragments that are less than 600 bases in length are not recovered. By fingerprinting all of the clones using the same restriction enzyme, clones with similar fingerprints, that is, those containing similarly sized fragments, can be identified and they will likely overlap. The probability that two clones overlap increases as a function of the number of bands their fingerprints have in common. The Sulston score (Sulston *et al*., 1988) defines that probability that *im* or more bands matching between two

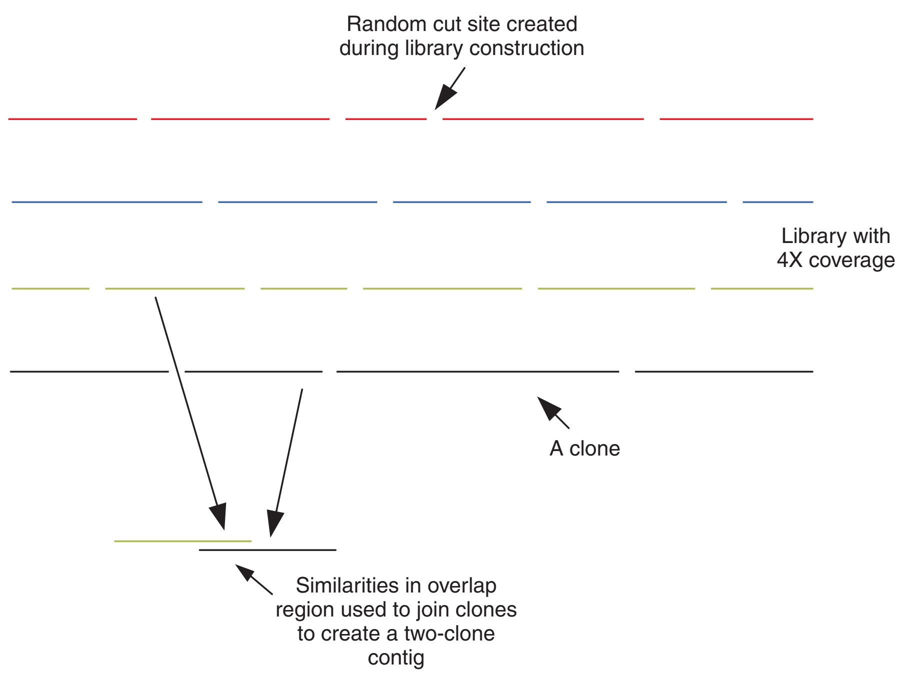


Figure 1 A representation of clones in a genomic library with 4X coverage. Each colored line represents a different copy of the original genomic DNA. Each fragment in each line represents an individual clone. Overlaps between clones can be used to identify the two clones as neighbors

fingerprints do so by chance in the following way:

$$S = \sum_{n=im}^{n=nbandl} \left[\frac{nbandl\,!}{(nbandl\ - n)!n!} * (1 - p)^n\, p^{(nbandl - n)} \right] \tag{1}$$

where *nbandl* is the number of bands in the clone with the fewer bands, *nbandh* is the number of bands in the clone with more bands and

$$p = \left(1 - \frac{2 * tolerance}{gellength}\right)^{nbandh} \tag{2}$$

The *tolerance* is a measure of the experimental error in determining the position of the band (and hence the size of the fragment). The equation assumes that all bands are equally likely. Hence, the probability of a single band in the gel with fewer bands matching one band in the gel with more bands is given by

$$\frac{2 * tolerance}{gellength} \tag{3}$$

which leads to equation (2), which is the probability that a single band in the fingerprint with fewer bands matches no bands in the other fingerprint. Since the number of matches between the fingerprints will follow a binomial distribution, the extension of equation (2) for all bands leads to the Sulston score, equation (1).

The Sulston score forms the basis for the FPC (FingerPrinted Contigs) software (Soderlund *et al*., 1997). Figure 2 shows a screenshot from the FPC software. This software is used to manage the process of creation of clone contigs. FPC performs an automated fast assembly of the clones producing a set of ordered and "buried" clones. A buried clone is one that contains no identifiable unique fragments and whose entire sequence may be wholly represented by another clone. In this case, the clone is "buried" in the clone that contains it. Clones that contain at least one unique fragment cannot be buried and are referred to as *canonical clones* (Figure 3).

The FPC algorithm works by first calculating the Sulston score for each pair of clones, burying clones that are an exact or a close match to another clone, and then building a consensus band map (CB map). The CB map is initially built using a hybrid greedy/stochastic algorithm to identify adjacent clones, which is then greedily extended.

Unfortunately, both the experimental methods and computational methods for identifying and sizing clone fragments are subject to experimental error. Therefore, it is not possible to create a perfectly ordered CB map and the CB map created by FPC must be refined by manual methods. A common problem is when a band is missed (undercalled) or a band is identified when none is present (overcalled). This problem is particularly evident when bands are very close together such that they cannot be resolved as separate bands on the gel, but may also be due to faint bands. A human map analyst can use FPC to review gel images and reorder clones by correcting for overcalled or undercalled bands to create the most likely clone order. This is a manually intensive task and is the bottleneck in the production of high



Figure 2 A screenshot from the FPC software. The clone names (e.g., H005D09) run across the page near the top of the figure and the fragment mobilities (related to fragment size) are on the left-hand side of the figure. The scanned gel images for each clone run down the page. On either side of the scanned image are horizontal bars that show where the band-calling software has identified a band. The figure shows examples of faint bands that have not been called by the software (undercall) and thick bands where the number of fragments composing the band is unclear and may result in too many bands being identified (overcall)

quality maps. Correcting for these artifacts in an automated manner is an active area of research and requires either improving the data quality (Fuhrmann *et al*., 2003) or improving the ordering and burying algorithm (Flibotte *et al*., 2004). Bandleader analyses the pixel values of a scanned gel image to identify bands. By assuming a Gaussian signal for a band, it is able to improve the resolution of image using standard signal processing techniques. CORAL uses a machine-learning algorithm to improve the burying and ordering of an FPC map.

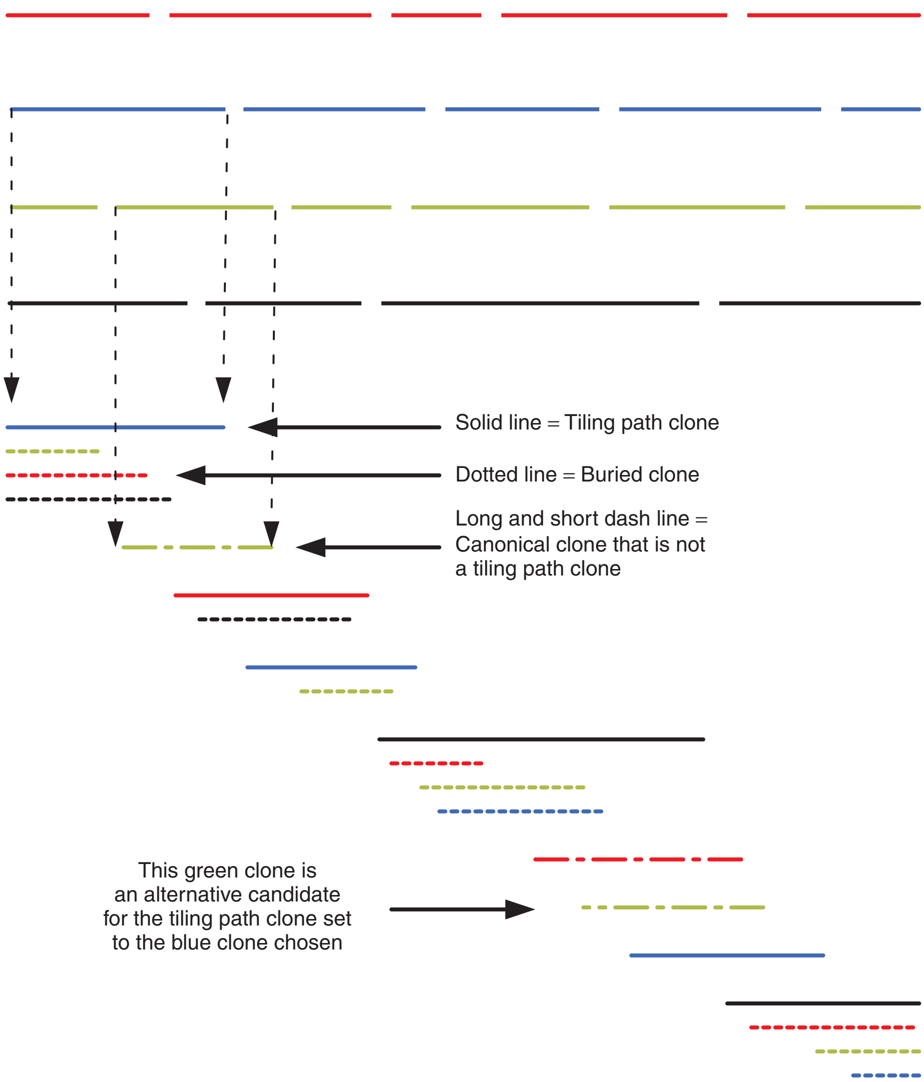


Figure 3 The different clones in Figure 1 have now been joined together to form a map contig. Dotted clones have their sequence fully represented by another clone are termed buried clones. Clones that cannot be buried (solid line and alternating dash/dot) are called canonical clones. From the canonical clones, a minimal set of tiling path clones have been selected (solid line) that represent all of the genomic DNA covered by this map contig with little or no redundancy. Note there are alternative clones that could have been selected when choosing clones for the tiling path

Another common problem is the creation of chimeric clusters that contain clones from a different part of the genome. These clones have been placed together because they have high Sulston scores by chance. FPC provides a framework and the necessary software tools for the map analyst to manipulate the clone order, change buried clones, and tease apart chimeric clusters.

Further difficulties are caused by genomes that contain a large number of repeat regions, such as plant genomes. The repeat regions tend to be compressed in fingerprint-based physical maps since if the genome contains the enzyme site, a repeat region will cause the overrepresentation of a band in the fingerprint, while if there is an absence of an enzyme site, this will lead to a large fragment that is not cut by the enzyme (Chen *et al*., 2002). There have been some attempts to improve the ordering and burying provided by the raw Sulston score. Soderlund and coworkers incorporated marker data into the original FPC software (Soderlund *et al*., 2000). The markers (for example, based on STS data) were used in conjunction with the fingerprint data to enhance the score. If two clones share markers, the cutoff of the Sulston score for the clones to be related is lowered by an order of magnitude for each marker they share. A further improvement to FPC has been the creation of a parallel version that distributes the calculation of the Sulston score for all pairs of clones across the nodes of a cluster (Ness *et al*., 2002). This reduces the time to create an FPC map making it feasible to optimize build parameters by running multiple FPC builds.

Software available for viewing an FPC map includes WebFPC (available from **http://www.genome.arizona.edu/software/fpc**) and Internet Contig Explorer (Fjell *et al*., 2003). Further tools allow virtual fingerprints to be created for sequenced clones for addition to the map and the extraction of a minimal tiling set from the map (Engler *et al*., 2003).

As described above, the final set of clones will consist of a number of buried clones and an ordered number of overlapping, canonical clones. Depending on the depth of sampling of the genomic library, several clones may represent any one region of the genome. By identifying a minimum tiling path for the clones (see Figure 3), a set of canonical clones that cover the entire genome with minimal overlap may be extracted. The minimum tiling path may be used as a starting point for sequencing the genome, providing a set of clones that have the same coverage of the genome, but with lower degree of redundancy than the original library, thereby reducing the sequencing cost. Clone tiling paths are also useful in the generation of comparative genomic hybridization (CGH) arrays (*see* Article 23, **Comparative genomic hybridization**, Volume 1) and for fluorescent in-situ hybridization (FISH) experiments (*see* Article 22, **FISH**, Volume 1).

More data can be provided by end sequencing of the clones. In this process, a sequence read is generated from each end of the clone. These reads may be used to help align the physical map with an assembled genome sequence. By aligning the end reads against the sequence, for example by using BLAST sequence alignment program, the clone's position in the genome sequence may be defined (Engler *et al*., 2003). By tying the map assembly to the sequence assembly, it is possible to uncover misassemblies in both the sequence and the map and to identify clone resources that can be used to fill sequence gaps.

This verification can be performed at several levels. Large-scale rearrangements will result in the clone order as determined by the fingerprints to be different from that specified by sequence assembly. The insert size of the clone (i.e., the length of DNA sequence that was inserted into the BAC or YAC) can be determined by summing the sizes of fragments in the fingerprint. This insert size will be an underestimate since small fragments (less than 600 bp) cannot be measured.

The insert size can be calculated for each clone individually or alternatively an average insert size can be calculated for the library (insert sizes for high-quality genomic libraries follow a pseudonormal distribution). Since reads from each end of the clone were taken, the sequence assembly can be used to calculate the insert size and this may be compared to the known insert size or average for discrepancies. The Arachne software is capable of incorporating this information (Batzoglou *et al*., 2002).

Resolution at a higher level may be performed by digesting with the assembled sequence in silico to create virtual fingerprints (i.e., finding cut sites in the assemble sequence that match the recognition site of the enzyme used to create the real fingerprints). The in silico fingerprints from the sequence assembly may be compared to the real fingerprints to identify discrepancies. The resolution of this type of analysis may be extended even further by creating additional fingerprints of the clones cut with different enzymes. See Article 8, **Genome maps and their use in sequence assembly**, Volume 7 for more details on how map data may be used to improve sequence assemblies.

There are a number of efforts ongoing to apply fluorescent labeling in conjunction with capillary electrophoresis using a DNA analyzer to automatically size fragments. The latest methods extend this technique by performing multiple digests using different enzymes and labeling the ends of the fragments to identify fragments cut by the same enzyme (Luo *et al*., 2003) or to sequencing 1–4 bases at the 5′ end of the fragment (Ding *et al*., 2001). Although a DNA analyzer presents an upper limit on the size of fragments that can be measured (about 600 bp), the measurements are much more accurate than those obtained by gel electrophoresis. Hence, the value of the tolerance used in equation (2) is smaller, resulting in a stricter Sulston score probability. This factor and the additional information that these methods provide about each fragment greatly improve the ability to identify related clones. These methods represent the next generation in fingerprinting technology and will allow for substantial improvements in automated methods of contig mapping and analysis.

With the improvement in whole genome shotgun assembly techniques, the generation of a physical map is no longer a necessary precursor to genome sequencing. However, the physical map remains a valuable resource for aiding the assembly of a genome. The correct assembly of a genome sequence from sequence data alone remains a difficult task and the physical map can act as a framework for the assembly. Furthermore, physical maps can be used to generate a set of clones that cover the entire genome with less redundancy than the original genomic library (known as a set of tiling path clones). Since the generation of a clone fingerprint remains relatively cheap, reducing the redundancy of the clones that are to be sequenced can reduce the cost of the sequencing operation.

Related articles

Article 18, **Fingerprint mapping**, Volume 3; Article 19, **Restriction fragment fingerprinting software**, Volume 3; Article 8, **Genome maps and their use in sequence assembly**, Volume 7

Acknowledgments

The authors wish to thank Jacquie Shein for helpful comments and Heesun Shin for providing the FPC figure. Asim Siddiqui and Steven Jones gratefully acknowledge the support of the BC Cancer Foundation. Steven Jones is a scholar of the Michael Smith Foundation for Health Research

References

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP and Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research*, **12**, 177–189.

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, YuY, Sun S, *et al*. (2002) An integrated physical and genetic map of the rice genome. *The Plant Cell*, **14**, 537–545.

Coulson A, Waterston R, Kiff J, Sulston J and Kohara Y (1988) Genome linking with yeast artificial chromosomes. *Nature*, **335**, 184–186.

Coulson AR, Sulston J, Brenner S and Karn J (1986) Towards a physical map of the genome of the nematode Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7821–7825.

Cox DR, Burmeister M, Price ER, Kim S and Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, **250**, 245–250.

Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, Lam JY, Kim YM and Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics*, **74**, 142–154.

Engler FW, Hatfield J, Nelson W and Soderlund CA (2003) Locating sequence on FPC maps and selecting a minimal tiling path. *Genome Research*, **13**, 2152–2163.

Fjell CD, Bosdet I, Schein JE, Jones SJ and Marra MA (2003) Internet Contig Explorer (iCE)–a tool for visualizing clone fingerprint maps. *Genome Research*, **13**, 1244–1249.

Flibotte S, Chiu R, Fjell C, Krzywinski M, Schein JE, Shin H and Marra MA (2004) Automated ordering of fingerprinted clones. *Bioinformatics*, **20**(8), 1264–1271.

Fuhrmann D, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, Chinwalla A, Hillier LW, Waterston RH, McPherson JD, *et al*. (2003) Software for automated analysis of DNA fingerprinting gels. *Genome Research*, **13**, 940–953.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al*. (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.

Krzywinski M, Wallis J, Gösele C, Bosdet I, Chiu R, Graves T, Hummel O, Layman D, Mathewson C, Wye N, *et al*. (2004) Integrated and Sequence-Ordered BAC and YAC-based Physical Maps for the Rat Genome. *Genome Research*, **14**(4), 766–779.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD and Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.

Nathans D (1979) Restriction endonucleases, simian virus 40, and the new genetics. *Science*, **206**, 903–909.

Ness SR, Terpstra W, Krzywinski M, Marra MA and Jones SJ (2002) Assembly of fingerprint contigs: parallelized FPC. *Bioinformatics*, **18**, 484–485.

Soderlund C, Humphray S, Dunham A and French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Research*, **10**, 1772–1787.

Soderlund C, Longden I and Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences*, **13**, 523–535.

Sulston J, Mallett F, Staden R, Durbin R, Horsnell T and Coulson A (1988) Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biosciences*, **4**, 125–132.

Specialist Review

Algorithmic challenges in mammalian whole-genome assembly

Serafim Batzoglou
Stanford University, Stanford, CA, USA

1. Introduction

The basic methodology used today for sequencing a large genome is double-barreled shotgun sequencing. Shotgun sequencing, introduced by Sanger and colleagues early on (Sanger *et al*., 1977), involves obtaining a redundant representation of a genomic segment with sequenced reads and then assembling the reads into contigs on the basis of sequence overlap. *Double-barreled* shotgun sequencing enhances this method by introducing the notion of paired-end reads that are obtained by sequencing both ends of a fragment of known size, and thus provides useful long-range linking information (Edwards *et al*., 1990).

It is relatively straightforward to cover an arbitrarily long genome with sequenced reads, so that few gaps remain (e.g., by the Lander–Waterman model (1988), a 10-fold redundancy coverage of the genome with 500-long reads yields less than 1 gap in every 10^6 bp). Sequencing a genome is simple then, if defined as obtaining a representation of the majority of nucleotides in a genome. What makes sequencing challenging is the subsequent computational problem of correctly assembling the reads into the original sequence.

2. Repeats, sequencing errors, and their relationship

Assembly would be easy were it not for repeats. If genomes had the properties of random strings on four letters {A, C, G, T}, then it would be unlikely for two reads to share a significant overlap unless they represented overlapping segments in the genome – a *true overlap*. Unfortunately, nearly half a mammalian genome's length consists of repeats, and true overlaps can be confused with *repeat-induced overlaps*. The main challenge of assembly is to distinguish between true overlaps, where reads should be merged, and repeat-induced overlaps, where merging should be (usually) avoided to prevent *misassemblies* – assembled sequence that does not occur in the genome. Repeats come in many varieties. The number of copies ranges

from 2, such as in gene duplications, to about 10^6 in the case of ALU repeats in the human genome. Sequence similarity between copies can be almost 100%, such as in recent duplications, or very low, such as in ancient transposon lineages that have become inactive. Our ability to distinguish between true and repeat-induced overlaps is intimately tied to the sequencing error rate. In essence, if more than a proportion r of sequencing errors is unlikely, then any overlap with more than $2r$ mismatches and gaps is likely to be repeat-induced. Therefore, increasing read accuracy reduces the repeat content of a genome for the purposes of assembly.

3. Two main ways to deal with repeats: clustering the reads and pairing the reads

There are broadly two methods for assembling long genomic regions in the presence of repeats. The first method is *clone-based sequencing*, whereby the reads are being clustered by performing shotgun sequencing on medium-sized segments such as Bacterial Artificial Chromosome (BAC) clones of length ∼200 kb. With this method, the effective repeat content is drastically cut because most segments that are repeated across the genome are unique within such short regions. Moreover, the number of reads to be assembled is reduced by a factor of 10^4, and therefore the computational assembly problem becomes much simpler. The second method is *double-barreled whole-genome shotgun (WGS) sequencing*, whereby paired reads are obtained from both ends of inserts of various sizes (Edwards *et al*., 1990). Paired reads can resolve repeats by jumping across them: when taken from plasmid inserts longer than 5000 bp, they can resolve LINE repeats of that length; paired reads from longer inserts such as cosmids and BACs can resolve duplications, help build longer scaffolds, and verify the large-scale accuracy of an assembly.

4. Clone-based sequencing

Clone-based sequencing was used to sequence several genomes including those of the yeast *Saccharomyces cerevisiae* (Oliver *et al*., 1992; Dujon, 1996), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998), *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), and humans (International Human Genome Sequencing Consortium, 2001). Most applications of clone-based sequencing were performed under the “map first, sequence second”, or *physical mapping* approach. Physical mapping involves constructing a complete physical map of a large set of clones covering the genome with redundancy, and then selecting a minimal tiling subset of those clones to be fully sequenced. After sequencing and assembly, errors in the original map, gaps, and misassemblies can be spotted, and if necessary, corrected with additional targeted sequencing. This approach was successfully used to generate physical maps for sequencing the *S. cerevisiae* and *C. elegans* genomes (Coulson *et al*., 1986; Olson *et al*., 1986). The human genome was also sequenced by the public consortium using this approach, although the first draft large-scale assembly was still a challenging

computational task: ordering and orienting the assembled contigs into a global genome assembly required integration of various sources of information such as sequence overlap, mRNA, and EST fragments, and end-sequencing of BACs. All these pieces of evidence were put together by a specialized system called *GigAssembler* (Kent and Haussler, 2001). Physical mapping is by no means the only possible way to perform clone-based sequencing. Other methods are possible but less explored, such as the walking approach (Venter *et al*., 1996; Batzoglou *et al*., 1999; Roach *et al*., 1999; Roach *et al*., 2000), or light shotgun sequencing of completely unmapped random clones, whereby a collection of clones covering the genome with redundancy C_1 are sequenced to redundancy C_2 for a total sequencing depth C_1C_2. As clustering directly addresses the repeat challenge by effectively eliminating most repeats, assembly under any of the above schemes is, in principle, easier than under double-barreled sequencing.

5. Whole-genome shotgun sequencing

Shotgun sequencing was initially applied by Sanger to the genome of bacteriophage λ (Sanger *et al*., 1980, 1982). Over the years, shotgun sequencing has been applied to genomic segments of increasing length. In the 1980s, it was applied to plasmids, chloroplasts (Ohyama *et al*., 1986), viruses (Goebel *et al*., 1990), and cosmid clones of length roughly 40 kb, which were considered to be the limit of this approach. In the 1990s, mitochondria (Oda *et al*., 1992), BACs of length 200 kb carrying genomic DNA from human or other organisms, as well as the entire 1800 kb genome of the bacterium *H. influenzae* (Fleischmann *et al*., 1995) were sequenced. Compared to clone-based shotgun sequencing, WGS poses a much harder computational challenge, especially in large, repeat-rich genomes. When a WGS strategy was proposed for sequencing a mammalian genome such as human (Weber and Myers, 1997), scientists debated whether such as strategy would succeed (Green, 1997). Today, we know that this is possible, as demonstrated by WGS assemblies of several complex genomes such as *Drosophila melanogaster* (Myers *et al*., 2000), human (Venter *et al*., 2001), and mouse (Mouse Genome Sequencing Consortium, 2002; Mural *et al*., 2002). These successes have brought computer science to the forefront of genomics: WGS assembly became possible through advances in algorithms and software systems rather than in either computer processor speed or sequencing accuracy.

6. Overview of WGS assembly

The prevailing methodology for WGS assembly is the three-stage *overlap-layout-consensus* strategy (Peltola *et al*., 1984; Kececioglu, 1991; Huang, 1992; Kececioglu and Myers, 1995). During the *overlap* phase, all read-to-read sequence overlaps are detected. During *layout*, subsets of the overlaps are selected in order to merge reads into longer scaffolds of ordered and oriented reads. Those scaffolds are converted into sequence through a multiple alignment during the *consensus* phase. One of the most successful systems employing this three-phase strategy is

PHRAP (Green, 1994), which has been the standard for BAC-size assembly, but is unable to handle whole mammalian-size genomes. Current WGS assemblers include the Celera assembler (Myers *et al*., 2000), Euler (Pevzner *et al*., 2001), Arachne (Batzoglou *et al*., 2002; Jaffe *et al*., 2003), Phusion (Mullikin *et al*., 2003), Atlas (Havlak *et al*., 2004), and others. Most of these systems, at least implicitly, use the overlap-layout-consensus strategy.

In the following paragraphs, we will briefly describe each of the assembly phases, focusing on the challenges they pose and on the algorithms that current WGS assemblers employ to address these challenges. We will be using the following terminology: a *contig* is a multiple alignment of reads, which can be converted into contiguous genomic sequence during the consensus phase. A *supercontig* is a structure of ordered and oriented contigs, usually derived by linking contigs according to paired reads. Some papers refer to *ultracontigs* as higher structures of supercontigs that are derived by various sources of linking information, such as paired reads from large inserts, mapping of BAC clones, or even cross-species comparison. The general term for assembled structures is the *scaffold*, a term usually interchangeable with supercontig. As the layout phase is the most complex, a useful conceptual (and implementation) division of this phase is in two subphases: (1) *contig layout*, during which most assembly systems strive to identify repeat boundaries and connect the reads into *contigs* that are as long as possible without crossing those boundaries, as that would induce misassemblies and (2) supercontig assembly, perhaps the hardest phase of all, where contigs are chained into longer structures by use of paired reads crossing contig boundaries, and subsequently, gaps are filled and misassemblies are identified and remedied.

7. The overlap phase

In the *overlap* phase, read-to-read sequence overlaps are detected. Note that since read orientation is not known, all overlaps between reads in forward or reverse complement direction are detected. The nature of the data requires several preprocessing tasks to take place prior to overlap detection, such as trimming of reads to remove bases with quality scores that are too low and filtering of contamination from sequencing/cloning vector or mitochondrial sequences. Next, in the case of large-scale WGS, a significant algorithmic challenge is to find the overlaps efficiently. The strategies that have been employed rely on detecting exact word matches between the reads. Celera, Phusion, and Arachne, all perform a global step of indexing all exact fixed-length word occurrences in the reads and then aligning all pairs of reads that share a word. Prior to alignment, filtering of very frequent words is necessary to avoid a huge number of pairwise read matches. Phusion and Arachne mask all highly repetitive k-mers, while Celera masks repeats, a step that is inherently more conservative and will discard additional spurious matches, but is also more complicated. Alignment is done with fast procedures that rely on the fact that only almost exact sequence identity between two reads can induce an overlap. Phusion and Arachne's original indexing strategies were more time-efficient, but required more memory, than Celera's strategy. Arachne performs this step in several passes, trading off speed for memory.

Upon detecting potential read overlaps, most assemblers filter them so as to remove the ones that are very likely to be repeat-induced. Arachne and Euler in particular, use a sequencing error-correction strategy that greatly simplifies subsequent assembly. In Arachne, the main idea is that a sequencing error reveals itself as a singleton base aligned to several other reads that all agree to a different base – in contrast to a repeat polymorphism, which should be shared by several reads. Euler's error correction is based on finding and modifying low-frequency words in the reads, and is generally more aggressive, resulting in a drastic reduction in the number of errors by as much as a factor of 50. After error correction, Arachne aggressively discards read-to-read overlaps that disagree in high-quality bases, because such overlaps are overwhelmingly likely to be repeat-induced. Most assemblers at this stage also discard likely *chimeric* reads, that is, reads that are formed by two distant segments of the genome being glued together. The signature of such reads is that they overlap other reads to the left and right, but no overlap spans the chimeric joint.

8. Contig layout

The prevailing way of representing all read overlaps is the *overlap graph* (Myers, 1995), where reads are nodes, and overlaps are labeled edges that contain information on the location and (optionally) quality of overlap. Only full overlaps between reads x and y are stored in the graph: partial overlaps that end somewhere within the two sequences are clearly repeat-induced and not stored in the graph. Repeat boundaries can be detected in an overlap graph by finding a read x that extends to two reads y and z to the left (or right), where y and z do overlap one another. The Celera assembler and Arachne use this overlap graph to merge reads into contigs up to the boundaries of repeats, thus joining all local interval subgraphs into contigs. To assemble contigs, Arachne employs a further heuristic: paired pairs of reads, defined as two plasmids of similar insert size with sequence overlaps occurring at both ends, are identified and merged, except under a heuristic condition that detects and avoids the instance of long repeats. Paired pairs help to cross high-fidelity repeats of length slightly longer than a read length. Phusion instead clusters the reads and then passes the clusters to local assembly calls to PHRAP. One advantage of this approach is the use of PHRAP for low-level assembly, because PHRAP has been highly tuned to work well with real data. Most assemblers use several additional heuristics during contig layout, in order to extend contigs as far as possible. For example, sequencing errors near the end of one read are identified as such by noticing that the read does not overlap any other read on that side. Contig breaks due to such sequencing errors are then crossed. Also, contig breaks due to repeats that are shorter than one read length are resolved by finding reads that span the different repeat copies. Precisely, a repeat can be resolved if each copy, except for at most one, is spanned by a read.

The resulting contigs constructed by Celera and Arachne are either unique, representing sequence that occurs once in the genome, or repetitive, representing a high-fidelity repeat sequence whose reads were merged together. In Arachne, the repeat has to exhibit nearly 100% fidelity. Contigs are classified as unique or

repetitive by measuring the density of reads within a contig: repetitive contigs are likely to be twice or more times as dense as unique contigs (Myers *et al*., 2000).

9. Supercontig layout

Once contigs are constructed and classified as unique or repetitive, the Celera and Arachne assemblers link them into larger structures, called *scaffolds* or *supercontigs*. One read-pair link can be an error, but two or more mutually confirming links that connect a pair of contigs are very unlikely by chance, and – in general – can be safely used to join the contigs into supercontigs. Most assemblers first join the unique contigs while excluding likely repeat ones to produce supercontigs that represent correctly ordered and oriented unique sequence with gaps. Then, gaps are filled by recruiting the repeat contigs. This last stage of repeat resolution is rather complicated, and involves a combination of several heuristic approaches. The Celera assembler fills in repeats in a series of increasingly aggressive steps by recruiting first repetitive contigs that contain two or more read-pair links to the flanking unique contigs, then those that contain one link and some sequence overlap, and finally the rest conditionally on some heuristics. Arachne attempts to walk through the gap between two unique contigs by a shortest-path search that requires anchors spaced at almost every 2 kb, where an anchor is a repetitive contig with read-pair links to the flanking unique contigs. This heuristic, in practice, results in a reasonably accurate repeat reconstruction, except in the case of high-fidelity tandem repeats, which are often collapsed and show up as deletions with respect to the true sequence.

10. Derivation of consensus sequence

Once a layout has been obtained, deriving consensus sequence is a relatively straightforward task. Starting from the leftmost read of each contig, consensus sequence is derived from the multiple alignment of reads. Conservative quality scores are given to each consensus base according to the quality and agreement of bases in the multiple alignment column. The Celera assembler will additionally report positions that appear polymorphic and give an estimate of the likelihood that the polymorphism is real versus repeat-induced.

11. Discussion on existing WGS assemblers

An elegant formulation of sequence reconstruction as an Eüler Path Problem on a de Bruijn graph (Pevzner, 1989), originally proposed in the context of sequencing by hybridization, has led to the development of the Eüler assembler (Pevzner *et al*., 2001). According to that approach, a graph is constructed whose nodes are the k-long words in reads and whose edges connect pairs of k-long words that share a $(k-1)$-long overlap. Read overlaps are thus implicitly found in an efficient manner

similar to Arachne's and Phusion's approach. Error correction is done directly on the k-long words by modifying words that are too infrequent. Assembly then reduces to finding the correct *Eülerian path* – there are usually a large number of Eülerian Paths. An Eülerian Path is a path that uses every edge exactly once. To achieve that, a series of graph simplifications are taken on the basis of the reads and the read-pair links. This approach has the advantage of clean formulation and demonstrably strong performance in resolving repeats – or identifying all assembly ambiguities due to repeats. One challenge is to develop an implementation that is efficient enough for whole mammalian genome assembly. The de Bruijn graph provides the same branching structure (repeat representation) as Myers's original overlap graph but at k-mer resolution rather than fragment resolution and provides more precise repeat boundaries at the cost of extra memory usage.

Several assemblers exist for large-scale application. The Celera assembler, Arachne, and Jazz effectively use the overlap-layout-consensus paradigm combined with read-pair information. Phusion follows a different approach, motivated by the fact that PHRAP is a well-tried and effective approach for small-scale assembly: Phusion first clusters the reads efficiently and then sends several local assembly problems to PHRAP. The assemblies are then merged and corrected according to read-pair information. Arachne and Phusion were both used for assembling the Mouse genome, with comparable performance (Jaffe *et al*., 2003; Mullikin *et al*., 2003). Atlas is a suite of programs for assembling a genome with an approach that combines clone-based and WGS data, and was used to assemble the rat genome (Rat Genome Sequencing Project Consortium, 2004).

12. Sequencing and assembly in the future

A few years ago, WGS assembly was considered impossible for large and repeat-rich genomes. Today, it is routinely applied, thanks to the development of computational methods led by the Celera assembly team and by public efforts. The success of WGS assembly contributed to the interest in bioinformatics methods within the biology community. Although the assembly research pertinent to current sequencing technology seems to be reaching diminishing returns, additional open problems are coming from the development of new and revolutionary sequencing technologies that promise to deliver the "$1000 genome" within a decade (**http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html**). Computationally, such technologies can be largely divided in two categories: (1) those that produce short, accurate reads cheaply, such as Pyrosequencing (Ronaghi *et al*., 1998; Ronaghi, 2001) or other technologies (Mitra and Church, 1999; Mitra *et al*., 2003; Braslavsky *et al*., 2003) and (2) those that produce very long reads, but potentially with very high error rates (see Pennisi, 2002). Short reads should be harder to assemble, even if the mate-pair information is available. The reason is, again, effective repeat content: fewer exact repeats will be completely spanned by a read length. However, if reads are clustered without sacrificing the method's speed and low cost (e.g., by following a random clone-based sequencing approach that does not involve physical mapping), assembly should be possible. Long reads will present interesting alignment and large-scale indexing

problems. In general, once the technological hurdles are overcome and a concrete computational problem is formulated, we can be confident that computational biologists will solve it and help ultimately make the $1000 genome a reality.

Acknowledgments

The author would like to thank Granger Sutton for helpful comments on the first draft of this manuscript. The author thanks NSF, NIH, and the Alphred P. Sloan Foundation for support.

References

Batzoglou S, Jaffe D, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP and Lander ES (2002) ARACHNE: A whole genome shotgun assembler. *Genome Research*, **12**, 177–189.

Batzoglou S, Mesirov JP, Berger B and Lander ES (1999) Sequencing a genome by walking with clone-ends: a mathematical analysis. *Genome Research*, **9**, 1163–1174.

Braslavsky I, Hebert B, Kartalov E and Quake S (2003) Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3960–3964.

Coulson A, Sulsto J, Brenner S and Karn J (1986) Towards a physical map of the genome of the nematode Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7821–7825.

Dujon B (1996) The yeast genome project: What did we learn? *Trends Genetics*, **12**, 263–270.

Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT and Ansorge W (1990) Automated DNA sequencing of the human HPTR locus. *Genomics*, **6**, 593–608.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–511.

Goebel SJ, Johnson GP, Perkus ME, Davis SW, Winslow JP and Paoletti E (1990) The complete DNA sequence of Vaccinia virus. *Virology*, **179**, 247–266.

Green P (1994) PHRAP documentation. **http://www.phrap.org**.

Green P (1997) Against a whole-genome shotgun. *Genome Research*, **7**, 410–417.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM and Gibbs RA (2004) The Atlas Genome Assembly System. *Genome Research*, **14**, 721–732.

Huang X (1992) A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, **14**, 18–25.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody M and Lander ES (2003) Whole-Genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research*, **13**, 91–96.

Kececioglu J (1991) *Exact and Approximation Algorithms for DNA Sequence Reconstruction*. PhD thesis, Department of Computer Science, University of Arizona, Tuscon.

Kececioglu J and Myers E (1995) Exact and approximate algorithms for the sequence reconstruction problem. *Algorithmica*, **13**, 7–51.

Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, **11**, 1541–1548.

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Mitra R and Church G (1999) In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, **27**, 1–6.

Mitra R, Schendure J, Olejnik J and Church G (2003) Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, **320**, 55–65.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Mullikin JC and Ning Z (2003) The Phusion assembler. *Genome Research*, **13**, 81–90.

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, *et al*. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, **296**, 1661–1671.

Myers EW (1995) Towards simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, **2**, 275–290.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KHJ, Remington KA, *et al*. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Oda K, Katsuyuki Y, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, *et al*. (1992) Gene organization deduced from the complete sequence of Liverwort Marchantia polymorpha mitochondrial DNA. *Journal of Molecular Biology*, **223**, 1–7.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, *et al*. (1986) Chloroplast gene organization deduced from complete sequence of livewort Marchantia polymorpha chloroplast DNA. *Nature*, **322**, 572–574.

Oliver SG (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R and Frank T (1986) Random-clone strategy for genome restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 7826–7830.

Peltola H, Soderlund H and Ukkonen E (1984) SEQUAID: a DNA sequence assembly program based on a mathematical model. *Nucleic Acids Research*, **12**, 307–321.

Pennisi E (2002) Gene researchers hunt bargains, fixer-uppers. *Science*, **298**, 735–736.

Pevzner PA (1989) L-tuple DNA sequencing: computer analysis. *Journal of Biomolecular Structural Dynamics*, **7**, 63–73.

Pevzner PA, Tang H and Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9748–9753.

Rat Genome Sequencing Project Consortium (RGSPC) (2004) Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Roach JC, Siegel AF, Trask B, van den Engh G and Hood L (1999) Gaps in the Human Genome Project. *Nature*, **401**, 843–845.

Roach JC, Thorsson V and Siegel AF (2000) Parking strategies for genome sequencing. *Genome Research*, **10**, 1020–1030.

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **11**, 3–11.

Ronaghi M, Uhlen M and Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–365.

Sanger F, Coulson AR, Barrell BG, Smith AJH, and Roe BA (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology*, **143**, 161–178.

Sanger F, Coulson AR, Hong GF, Hill DF and Peterson GB (1982) Nucleotide sequence of bacteriophage_DNA. *Journal of Molecular Biology*, **162**, 729–773.

Sanger F, Nicklen S and Coulsen AR (1977) DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

The Arabidopsis Genome Initiative (AGI) (2000) Analysis of the genome of the flowering plant Arabidopsis thaliana. *Nature*, **408**(6814), 796–815.

The C. elegans Sequencing Consortium (1998) Genome sequencing of the nematode C. elegans: a platform for investigating biology. *Science*, **282**(5396), 2012–2018.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Venter JC, Smith HO and Hood L (1996) A new strategy for genome sequencing. *Nature*, **381**, 364–366.

Weber JL and Myers EW (1997) Human whole-genome shotgun sequencing. *Genome Research*, **7**, 401–409.

Specialist Review

Genome signals and assembly

Marek Kimmel
Rice University, Houston, TX, USA

1. Introduction

Sequencing of entire genomes of various organisms has become one of the basic tools of biology. However, quality of genome assembly depends to a large extent on the structure of genomic sequences, notably, signals such as repeats, polymorphisms, and nucleotide asymmetry as well as structural motifs such as protein motifs (*see* Article 28, **Computational motif discovery**, Volume 7), gene promoters, enhancers and suppressors (*see* Article 19, **Promoter prediction**, Volume 7 and Article 24, **Exonic splicing enhancers and exonic splicing silencers**, Volume 7), transcription factor binding sites, exon/intron splice junctions, regions of homology between sequences (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7), and protein docking sites. Probabilistic and statistical properties of the structural motifs are covered elsewhere in this volume (*see* Article 6, **Statistical signals**, Volume 7). We focus here on global features of genome assembly. The importance of these has been recognized in the book by Waterman (1995), in seminal papers of Lander and Waterman (1988) and Li and Waterman (2003), as well as in the book by Ewens and Grant (2001). Genome assembly can be modeled as a binomial/Poisson stochastic process. From the properties of this process, it is possible to derive the statistics of contig size and in this way to determine the coverage needed to achieve an assembly of desired quality. Analogously, probing the fragments with l-mers, it is possible to estimate the size of the genomic sequence, even when repeats are involved. This is accomplished by reconstructing the repeat structure, using an expectation-maximization algorithm. In the framework of this theory, it is also possible to estimate the total gap length and the stringency ratio.

In this review, we provide an account of these topics as well as of some other more specialized sequence signals such as polymorphisms, nucleotide asymmetry, and annotation by words. We also consider some real-life and made-up examples.

2. Statistics of genome assembly – contigs and anchors

The following account is mainly based on Ewens and Grant (2001), Chapter 5. If biological details are omitted, shotgun genome sequencing can be reduced to

assembly of a sequence of total length G, from N fragments (also called *reads*) of equal length L. Fragments are randomly selected from sequence G, and the assembly is feasible if there exists enough overlap between the fragments. To ensure this, the coverage $a = NL/G$ should be large enough and certainly greater than 1. Depending on the strategy used, G may represent the entire genome (WGS, whole genome shotgun) or a subset of the genome. In Figure 1, we depict an example, this one with incomplete coverage.

A model for the random fragments is the binomial/Poisson stochastic point process, in which the coordinates of the left ends of the fragments are independent random variables uniformly distributed over G. Neglecting boundary effects, we obtain that the probability that the number of fragments with left ends in the interval $(x, x-h)$ is equal to k and has distribution *binomial* $(N, h/G)$ or approximately *Poisson* (Nh/G). A union of overlapping fragments is named a *contig*. In order to obtain a good quality of assembly, it is necessary that the contigs cover as much of G as possible. Mean number of contigs is equal to

$$\begin{aligned} E(\text{ \#contigs }) &= N \times \Pr(\text{fragment is rightmost in a contig}) \\ &= N \times \Pr(\text{fragment does not include the left end of any} \\ &\qquad \text{other fragment}) \\ &= N \times \exp\frac{-NL}{G} = \frac{aG}{L} \times \exp(-a) \end{aligned} \tag{1}$$

Figure 2 depicts $E(\text{ \#contigs })$ as a function of coverage a. $E(\text{ \#contigs })$ first increases but then decreases back, since smaller contigs coalesce with increasing coverage. A single (length G) contig is expected when coverage reaches $a \approx 8$. Expected contig size can be expressed as

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}(\text{\#fragments} - 1)\ \mathrm{E}(\text{inter} - \text{epoch distance}) + L \\ &= \frac{1-\exp(-a)}{\exp(-a)} \int_0^L x\lambda \frac{\exp(-\lambda x)}{1-\exp(-\lambda L)} \mathrm{d}x + L = L\ \frac{\exp(a)-1}{a} \end{aligned} \tag{2}$$

$E(S)$ increases as smaller contigs coalesce. For the data as in Figure 2, $E(S) = G$ is expected when coverage reaches $a \approx 8$.

We might require that the contigs be anchored, that is, each of them include at least one anchor, which is a genomic site of known location (an element of a

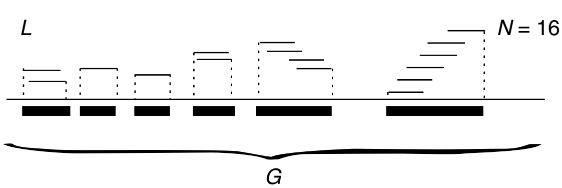


Figure 1 Schematic depiction of 6 contigs assembled from $N = 16$ fragments, of length L each, of genomic sequence of length G

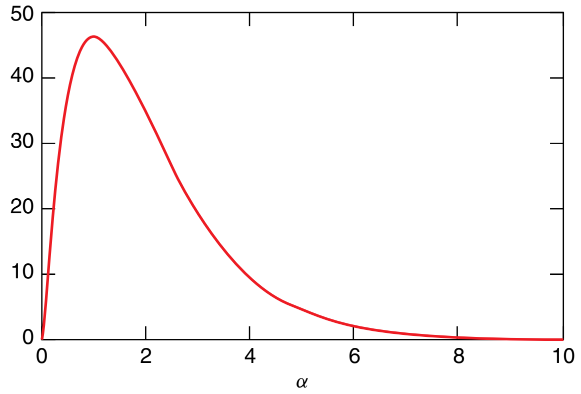


Figure 2 Expected number of contigs as a function of coverage a. Parameter values, $L = 800$, $G = 100\,000$

gene map). Let us define the coverage with M anchors as equal to $b = ML/G$ and suppose that the anchors follow the binomial/Poisson process. Then, we have

$$\text{E(\# anchored contig)} = Nb\,\frac{\exp(-a) - \exp(-b)}{b - a} \tag{3}$$

which is reduced to the nonanchored case as $b \to \infty$, but usually is smaller.

3. Gap statistics

In a recent paper, Wendl and Yang (2004) discuss prediction of the size of gaps in WGS projects. According to the Lander and Waterman (1988) theory, the expected number of gaps is equal to

$$E = N \exp[-\alpha(N - 1)] \tag{4}$$

where $\alpha = (L - T)/G$ is the effective fractional read (fragment) length, with T being the minimum overlap required. Maximum gaps number consequently is equal to

$$E_{\max} = \frac{\exp(\alpha - 1)}{\alpha} \tag{5}$$

Lander and Waterman (1988) theory also implies that stringency σ, that is, the proportion of existing gaps to the expected maximum, can be expressed by the effective coverage $\delta = N(L - T)/G$, as $\sigma = E/E_{\max} = \delta \exp(1 - \delta)$. Wendl and Yang (2004) proceed to demonstrate that these expressions underestimate stringency (as it was claimed earlier on) and provide semiempirical estimates of stringency, given coverage, $\sigma_{\text{emp}} = 1.187 \exp(-0.334\delta)$ for eukaryotes and $\sigma_{\text{emp}} = 0.701 \exp(-0.211\delta)$ for prokaryotes. These estimates were obtained by applying linear regression to log-linearized data from 20 WGS assemblies. Empirical stringencies

were computed as the quotients of the numbers of gaps in the assembly and a theoretical maximum, while the coverages were based on the numbers of reads in an assembly (further details in Wendl and Yang, 2004).

4. Repetitive elements and estimation of genome length

Repetitive elements are among the most frequent dynamic features of the genome. Many of them display complicated evolutionary dynamics (for some models, see e.g., Kimmel and Axelrod, 2002). Here, mostly based on Li and Waterman (2003), we review the influence of repetitive elements on estimation of genome length. In the shotgun sequencing method, the genomic sequence is assembled from overlapping fragments. Uncertainties of the assembly (see further on) lead to the need to analyze the repeat structure of the sequence to be able to estimate the true size of the genome or its fragment sequenced. Indeed, let us consider a given DNA sequence. If we can choose l such that almost all l-mers appear only once in the sequence, if they appear at all, then, by counting how many different l-mers there are in the fragments, we can have a good estimate of the length of the sequence. This is accurate, assuming that the sequence is random with letters generated independently. However, there may be many repeats in the sequence (Primrose, 1998), such that many l-mers appear in it twice or more, even if l is quite large. In this case, the number of different l-mers in the fragments should be much smaller than the number of nucleotides in the sequence. However, if we know how many times each l-mer is repeated, and what fraction of the sequence it accounts for, we can estimate the sequence length. That is, we may have to estimate the repeat structure and coverage in order to estimate the sequence length.

Recall that, for an l-mer appearing in a sequence, instead of the number of this l-mer's occurrences in this sequence, we obtain the number of this l-mer's occurrence in N fragments. Because the fragments are randomly chosen from clone libraries, their positions in the original sequence are random, as is the number of occurrences of any l-mer in the sequence. Let us consider the distributions of these random variables first. Assume we know the coverage a of the genome by these N fragments. Also assume that an l-mer in the sequence cannot appear twice or more in any fragment, that is, any two copies of an l-mer are at least L base pairs apart in the original sequence. For a given l-mer, for instance, w, that appears in the sequence $n(w)$ times, how many times will it appear in N fragments?

Assume $x_i(w)$ is the number of fragments that cover the ith copy of w, in which i is from 1 to $n(w)$. Then, w appears $x_{n(w)}(w)$ times in the fragments. Note that $x_i(w)$, $i = 1, \ldots, n(w)$ are independent identically distributed (i.i.d.) and approximately Poisson. Assume the parameter of the Poisson process is λ

$$\lambda = \frac{N}{G - L + 1} \tag{6}$$

(Lander and Waterman, 1988). The distribution of $x_1(w)$ is Poisson with parameter $\lambda\,(L - l + 1)$. Owing to the additivity of Poisson process, the distribution of $x(w)$

is Poisson with parameter $n(w)\ \lambda\ (L - l + 1)$. For any given l-mer w in the sequence, we do not have a vector of $\{x_i(w), i = 1, \ldots, n(w)\}$, but $x(w)$, the sum of the elements in the vector. For estimation, we can use all observations from those l-mers, that appear approximately $n = n(w)$ times in the original sequence, as samples of w (family n). Recall that the number of occurrences of those l-mers may not be independent, although they have the same distribution. Assume w_1, $w_2, \ldots, w_m$ are those family n l-mers. We have

$$\frac{x(w_1) + \cdots + x(w_m)}{m}$$

approach $n(w)\ \lambda\ (L - l + 1)$ when m is large.

Assume there are k families of l-mers in the original sequence, the number of occurrences of any l-mer in the fragments from the ith family is a Poisson random variable with parameter $a_i\, a$, in which $a = \lambda\ (L - l + 1)$ is the coverage and a_i is an unknown integer. Different l-mers in the ith family account for $\alpha_i \times 100\%$ of all different mers in the original sequence. Mathematically, we have a set of samples from a mixture of Poisson distributions, and some of the samples are dependent. We estimate a_i, α_i, and a for $i = 1, 2, \ldots, k$. For this mixed-Poisson proportion problem, we get the following expressions (Lange, 1998)

$$a_i a = \frac{\sum_w n(w)\,\Pr(w \in \text{family } i|n(w))}{\sum_w \Pr(w \in \text{family } i|n(w))}$$

$$\alpha_i = \frac{\sum_w \Pr(w \in \text{family } i|n(w))}{\sum_{j=1}^{k} \sum_w \Pr(w \in \text{family } i|n(w))}$$

$$\Pr(w \in \text{family } i|n(w)) = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j \left(\frac{a_j}{a_i}\right)^{n(w)} \exp[(a_i - a_j)a]} \tag{7}$$

We have no data for w with $n(w) = 0$. That is, we do not know group 0 if we assume that there are group i-mers in the sequence, which appear i times in the fragments, for $i = 0, \ldots, groupNum$. But, we can use the following formulas to estimate group 0.

$$group(0) = \frac{\sum_{i=1}^{k} \alpha_i \exp(-a_i a) \sum_{i=1}^{groupNum} group(i)}{\sum_{i=1}^{k} \alpha_i (1 - \exp(-a_i a))} \tag{8}$$

According to Li and Waterman (2003), the length of the sequence is estimated by the second or third expression below.

$$\text{percentage of the tuples from the } i\text{th family} = \frac{\alpha_i a_i}{\sum_{j=1}^{k} \alpha_j a_j}$$

$$\text{length of the sequence} = \sum_{j=1}^{k} \alpha_j a_j \sum_{j=1}^{groupNum} group(j) \tag{9}$$

or

$$\text{length of the sequence} = \frac{N(L-l+1)}{\min\{a_j a : j = 1, \ldots, groupNum\}} \tag{10}$$

The expressions developed can be used to create a recursive algorithm of the Expectation-Maximization (EM) type. This first entails setting a large number for k and a proper number for l and calculation of $group(j)$, for $j = 1, 2, \ldots, groupNum$, based on data. Then, we set initial values for a_{i}, a, and α_{i}, respectively for $i = 1, 2, \ldots, k$. Subsequently, we compute $group(0)$ and new values for a_{i}, a, and α_{i}, respectively for $i = 1, 2, \ldots, k$, based on the expression above. We repeat this step until convergence is reached. Finally, we calculate the percentage of each family of substrings and the length of the sequence.

The proper choice of l is critical for the functioning of the method. As stated above, we should let l be large enough so that many l-mers in the original sequence are unique l-mers. That is, when the DNA is G base pairs (bp) long, l should satisfy $4^l > G$ if the sequence is generated from a uniform i.i.d. mechanism. On the other hand, we cannot let l be too large. For instance, if we let $l = L$, then there are N l-mers in all, and each l-mer appears once in the fragments in general. Then, our estimate of a is 1, which is incorrect. Moreover, the larger l is, the fewer the number of samples, and the less accurately we can estimate a_i, a, and α_{i}, for $i = 1, 2, \ldots, k$. Therefore, l must be large, but not too large. Further details can be found in Li and Waterman (2003).

Statistical properties of this algorithm do not seem to be well developed. Estimation of $group(0)$ is based on the Poisson assumption. Figure 2 of Mullikin and Ning (2003) shows a frequency graph of the 17-mer word occurrence, based on mouse $a = 7.5$ assembly. It shows a large excess of single-occurrence words above the model curve, indicating that these words are bad data resulting from sequencing errors.

5. Examples of estimation of repeat structure and genome length

Li and Waterman (2003) provide examples of how their method functions. Their Example 3 assumes $G = 80\,000$, $L = 500$, and $a = 3$. There are two families

of repeats. One is 6-kb long with 2 copies, whereas the other is 1-kb long with 12 copies. Repeats appear in tandem in the original sequence. In more than 95% of the simulations, they observed that the estimated genome length is 73 104 bp; the estimated coverage is 3.283; the unique sequence accounts for 82.7% of the genome. There is only one family of repeats, which has 12 copies and accounts for 17.3% of the sequence. This example illustrates difficulties inherent when short genomes and low coverages are considered. Further details are discussed in Li and Waterman (2003).

6. Polymorphisms

The following is mostly based on Fasulo *et al*. (2002). Polymorphisms are differences in variants of DNA sequences present in different individuals and on paternal and maternal chromosomes in a single individual. Most algorithms for large-scale DNA assembly make the simplifying assumption that the input data is derived from a single homogeneous source. Frequency of polymorphisms (single nucleotide polymorphisms (SNP), variable-length microsatellites or block insertions and deletions (indels)) depend on the evolutionary history of the species. For example, indels are twice as frequent in the human genome than in the *Drosophila* genome. Discrepancies due to polymorphisms will cause false negatives in the overlap relations among fragments and may result in confusing the polymorphisms with evolutionary divergence in repeat regions of the genome. The Celera Assembler (Myers *et al*., 1995) includes the A-statistic, which can help determine if the region containing the bubble seems repetitive. A sequence of overlapping fragments (for simplicity assumed to have equal lengths) of a genomic sequence can be depicted using a directed graph, which is linear for an unambiguous complete assembly. Polymorphisms will be represented as bubble-like structures (for examples, see Fasulo *et al*., 2002). Bubbles can be resolved in two ways: (1) A single path through the bubble can be designed, resulting in a consensus sequence. (2) Multiple alternative paths may be accepted and represented in the final assembly. An algorithm for bubble resolution (smoothing), which is a subroutine of the Celera Assembler, is described in Fasulo *et al*. (2002). The final validation of bubble smoothing is resequencing of the putative polymorphisms.

7. Bias in AT versus GC composition

It is known that high AT contents of the genome degrades the quality of assembly. Wendl and Yang (2004) provide a semiempirical relationship for stringency in prokaryotes $\sigma_{\text{emp}} = 1.129\exp(-4.93\gamma)$, where γ is the GC content, with correlation of around 76%. Analogous relationship in eukaryotes is not so well defined.

8. Annotation by words and comparison of genome assemblies

It is possible to annotate large genomic sequences with exact word matches. Healy *et al*. (2003) developed a tool for rapidly determining the number of exact matches

of any word within large, internally repetitive genomes or sets of genomes. This is achieved by creating a data structure that can be efficiently queried and that can also entirely reside in random access memory (RAM). The solution depends upon the Burrows–Wheeler transform, a method used to create a reversible permutation of a string of text that tends to be more compressible than the original text (Burrows and Wheeler, 1994). First, given a genome **G** of length G, we create a new string **G**$ of length $G + 1$ by appending a "$" to the end of that genome. We then generate all $G + 1$ "suffixes" of **G**$, that is, substrings that start at every position and proceed to the end. We next associate with each suffix the letter preceding it. In the case of the suffix that starts at the first position, we associate the new $ character and assume that "$" has the lowest lexicographical value in the genome alphabet. The string of antecedent letters, in the lexicographical order of their suffixes, is the Burrows–Wheeler transform of the G$ string. For example, if the genome were simply "CAT", the G$ string would be "CAT$". Then the suffixes of the genome in sorted order would be: "$", "AT$", "CAT$", and "T$". The Burrows–Wheeler transform of this particular G$ would be the letters that precede each of these suffixes taken in the same order, specifically "TC$A". The list of pointers taken in the order of the sorted suffixes would be (3,1,0,2). This list of pointers is the suffix array for the string "CAT$". The suffix array for the human genome approximately constitutes 12 Gb (3 billion, 4-byte integers) of RAM. However, the B–W string alone is sufficient to determine word counts and it can be compressed to about 1 Gb of RAM. Furthermore, for querying, all but a negligibly small portion of the compressed form can remain so throughout execution (Healy *et al*., 2003).

Using the above tools, any region of the genome can be annotated with its constituent mer frequencies. Healy *et al*. (2003) have depicted annotations of a 5-kb region of chromosome 19 four-mer lengths, 15, 18, 21, and 24 bases. For each coordinate and each word length, they determined the count of the succeeding word of the given length, in both the sense and antisense directions. One of the most striking features of this terrain is the presence of narrow spikes in 15-mer counts. This is a virtually universal property of all regions of the human sequence examined, including coding exons and it is statistically significant (based on a comparison with random genomes). Hypothetically, these spikes might result from the accidental coincidence of 15-mers in ordinary sequence with 15-mers present in high-copy-number repeats.

Another application of the method was to monitor successive human genome assemblies. Healy *et al*. (2003) compared December 2001 and June 2002 assemblies and found that, unexpectedly, 1.2% of their probe words vanished from the June 2002 assembly (i.e., all of their constituent 21-mers went from copy number one in their original assembly, to copy number zero in a subsequent assembly). Systematic studies revealed both losses and gains of single-copy words between assemblies. Although there may be technical reasons explaining the dropout of some of these fragments, such as difficulty in assembly or poor-quality sequence, it is also likely that, due to insertion/deletion and order-of-sequence polymorphisms in humans, no fixed linear rendition of the genome is feasible. A detailed discussion is provided in Healy *et al*. (2003).

9. Examples

The following examples illustrate the complex nature of global-level organization of human chromosomes and demonstrate that the theory developed above is important for the applications. Recently, edited sequences of human chromosome 14 and of the male-specific region of chromosome Y (MSY) were published (Heilig *et al*., 2003; Skaletsky *et al*., 2003). In chromosome 14, an ancient duplication involving 70% of chromosome 14 and a portion of chromosome 2 has been reported. This event is, however, only visible at the protein level and predates the mouse – human separation. Heilig *et al*. (2003) found that 1.6% of chromosome 14 consists of interchromosomal segmental duplications contained in fragments of 1 kb or more that show at least 90% sequence identity. On the basis of further references in Heilig *et al*. (2003), comparable value, based on a different comparison procedure, was reported earlier, and confirms that chromosome 14 has the lowest content of interchromosome segmental duplications in the human genome. In a similar analysis that excluded repetitive DNA, it was found that internal duplications account for 1.1% of chromosome 14 and are clustered in four segments. The largest includes an 800-kb region adjacent to the centromere, which is also part of the segmental duplication shared with chromosome 22.

The male-specific region of the Y chromosome, the MSY, differentiates the sexes and comprises 95% of the chromosome's length. Skaletsky *et al*. (2003) determined, among others the most pronounced structural features of the ampliconic regions of Y are eight massive palindromes. In all eight palindromes, the arms are highly symmetrical, with arm-to-arm nucleotide identities of 99.94–99.997%. The palindromes are long, their arms ranging from 9 kb to 1.45 Mb in length. They are imperfect in that each contains a unique, nonduplicated spacer, 2–170 kb in length, at its centre. Palindrome P1 is particularly spectacular, having a span of 2.9 Mb, an arm-to-arm identity of 99.97%, and bearing two secondary palindromes (P1.1 and P1.2, each with a span of 24 kb) within its arms. The eight palindromes collectively comprise 5.7 Mb, or one-quarter of the MSY euchromatin.

In addition to palindromes and inverted repeats, the ampliconic regions of Yq and Yp contain a variety of long tandem arrays. Prominent among these are the newly identified NORF (no long open reading frame) clusters, which in aggregate account for about 622 kb on Yp and Yq, and the previously reported TSPY (testis specific protein Y) gene clusters, which comprise about 700 kb of Yp. The NORF arrays are based on a repeat unit of 2.48 kb. Numerous further structural features of MSY and their evolutionary explanations, are discussed in Skaletsky *et al*. (2003).

10. Concluding remarks

As more and more genomes become sequenced, structural features such as repeated regions, short repeats (mini- and microsatellites), palindromes, and polymorphisms are emerging (for the human Y chromosome example, see Skaletsky *et al*., 2003). Current understanding of genome assembly has been dominated by practical

considerations. In particular, sequencing is time consuming and constrained by cost-efficiency. However, shotgun sequencing also is a stochastic procedure and therefore even edited genome assemblies may contain inaccuracies (Healy *et al*., 2003). On the other hand, stability of a genome as a whole is not a priori known and may be less perfect than expected (Healy *et al*., 2003). These observations underscore the need for methods, such as Li and Waterman (2003) algorithm, to estimate structural features of the genome on the basis of reads available. More generally, further mathematical work seems to be needed, transcending the existing theory (see, e.g., Lander and Waterman, 1988), to understand the uncertainties of genome assembly.

References

Burrows M and Wheeler DJ (1994) A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation: Palo Alto.

Ewens WJ and Grant GR (2001) *Statistical Methods in Bioinformatics*, Springer: New York.

Fasulo D, Halpern A, Dew I and Mobarry C (2002) Efficiently detecting polymorphisms during the fragment assembly process. *Bioinformatics*, **18**(Suppl. 1), S294–S302.

Healy J, Thomas EE, Schwartz JT and Wigler M (2003) Annotating large genomes with exact word matches. *Genome Research*, **13**, 2306–2315.

Heilig R, Eckenberg R, Petit JL, Fonknechten N, Da Silva C, Cattolico L, Levy M, Barbe V, de Berardinis V, Ureta-Vidal A, *et al*. (2003) The DNA sequence and analysis of human chromosome 14. *Nature*, **421**, 601–607.

Kimmel M and Axelrod DE (2002) *Branching Processes in Biology*, Springer: New York.

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Lange K (1998) *Numerical Analysis for Statisticians*, Springer: New York.

Li X and Waterman MS (2003) Estimating the repeat structure and length of DNA sequences using l-mers. *Genome Research*, **13**, 1916–1922.

Mullikin JC and Ning Z (2003) The Phusion assembler. *Genome Research*, **13**, 81–90.

Myers E (1995) Toward simplifying and accurately formulating fragment assembly. *Journal of Computation Biology*, **2**, 275–290.

Primrose SB (1998) *Principles of Genome Analysis*, Blackwell Science Ltd.: Oxford.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, *et al*. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, **423**, 825–837.

Waterman MS (1995) *Introduction to Computational Biology*, Chapman and Hall/CRC: Boca Raton.

Wendl MC and Yang S-P (2004) Gap statistics for the whole genome shotgun DNA sequencing projects. *Bioinformatics*, **20**, 1527–1534.

Microbial sequence assembly

George Weinstock
Baylor College of Medicine, Houston, TX, US

The first microbial genome projects began in the 1990s and focused either on important model systems (e.g., *Escherichia coli*, *Saccharomyces cerevisiae*) or on important pathogens (e.g., *Mycobacterium tuberculosis*). The prevailing view for these early microbial projects was that assembling the complete genome sequence must be piecemeal from large insert clones such as cosmids that were ordered into a tiling path along the chromosome (*see* Article 8, **Genome maps and their use in sequence assembly**, Volume 7). Each large insert clone would be sequenced by shotgun sequencing and its sequence assembled on the basis of sequence overlaps between the individual random shotgun reads. The assembled large insert clones would then be stitched together to form the complete genome.

The alternative assembly method to this clone-by-clone (CBC) approach was whole-genome shotgun (WGS) assembly, where the entire genome would be fragmented and cloned and the ends of each clone would be sequenced to produce a large amount of random sequence reads. These would then be assembled en masse to produce the final assembly. However, in the early years of microbial genome projects, the CBC approach was broadly adopted out of a concern that repeated sequences in the genome would pose an insurmountable problem for the WGS approach. The repeated sequences could not be unambiguously placed on the basis of sequence overlaps, and would therefore create gaps in the genome assembly.

The concern for the repeated sequence problem was dramatically shattered with the assembly of the *Haemophilus influenzae* genome using the WGS method at The Institute for Genomic Research in 1995 (Fleischmann *et al*., 1995). This was the first large-scale (for microbes) WGS project to be completed and set the stage for virtually every subsequent microbial genome project. The software developed for this project, the TIGR assembler (Pop and Kosack, 2004), had many ideas that are used in latter day genome assembly programs, even though the assembly software has grown in sophistication and now handles genomes over 1000 times the size of a bacterium.

The general approach used in most WGS genome projects nowadays is to first create a series of shotgun genomic libraries representing clones with different insert sizes (usually small, 3 kb, medium, 10 kb, and large, 40 kb). Both ends of each clone are sequenced until the total number of bases produced is about eight times the number of bases in the genome. The reads are then assembled on the basis of sequence overlaps between reads to produce contigs: blocks of contiguous sequence

representing a path of overlapping reads. Contigs end, however, resulting in a gap. The gaps may be due to regions that are missing from the clone libraries, or regions containing repeated sequences that cannot unambiguously be placed in the genome and so are skipped by the assembly software.

The contigs that result from assembling the WGS reads are ordered and oriented with respect to each other using read pair (sometimes called mate pair) information. Each clone in the shotgun libraries was sequenced from both ends, resulting in a pair of reads separated by a distance in the genome equal to the insert size of the clone. This information is used to link contigs together when the reads in a pair fall in different contigs. Contigs that are linked by read pairs are often called scaffolds. Thus, most assembly algorithms deal with the automated construction of contigs and scaffolds.

At this stage of the assembly, there are two kinds of gaps remaining to be filled: gaps between contigs within a scaffold (captured gaps since clones exist that span them) and gaps between scaffolds (uncaptured gaps). In the former case, the polymerase chain reaction is used to generate a product spanning the gap, since primers can be designed and properly oriented on the basis of the sequence of the contigs flanking the scaffold. The latter case poses the greatest challenge since one does not know which contigs flank uncaptured gaps. The principal solution to sequencing these regions is primer walking, where a primer is made from the end of a contig adjacent to a gap and used to sequence from the genome into the gap region. This new sequence either connects the contig to scaffold or, if not, it is used as the target for the design of a new primer and another walk is made. The process is repeated until the gap is spanned.

The gap-filling steps just described tend to be more labor-intensive and costly than the WGS sequencing and automated assembly stages. In fact, the major advantage of the WGS approach is that there is no need for up-front efforts to map large insert clones or develop a tiling path. As a result, a WGS genome project may stop after automated assembly and before the gap-filling stage. The product of such a project is called a "draft" genome sequence. In the contigs, the sequence is of generally high quality and sufficient for gene predictions or design of microarrays, which are often the main deliverables from a genome sequence. The gaps that remain may principally be due to repeated sequences that are often not critical in understanding the phenotype of an organism. Draft sequencing is particularly popular for genomes where gap filling is difficult due to the large size of a genome or other properties. However, a "finished" genome sequence, one with gaps filled and other defects removed, is still superior for detailed analysis of genes and regulatory sequences, where single nucleotide differences can be significant.

References

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, **269**(5223), 496–512.

Pop M and Kosack D (2004) Using the TIGR assembler in shotgun sequencing projects. *Methods in Molecular Biology*, **255**, 279–294.

Short Specialist Review

Comparative analysis for mapping and sequence assembly

Aleksandar Milosavljevic
Baylor College of Medicine, Houston, TX, USA

1. Comparative analysis for mapping and sequence assembly

Comparisons of mammalian genomic sequences reveal extensive similarity at both the chromosome and basepair levels (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004). The increasing number of assembled reference sequences produced by ongoing genome sequencing projects thus provides information that is potentially useful for the mapping and assembly of related genomes.

Detection of unique orthologous sequence fragments, which are sometimes referred to as "orthologous anchors" or "syntenic anchors", is a key step in many comparative genomic methods (Mural *et al*., 2002). The density of anchor sequences and the conservation of their relative order, orientation, and distances is a key indicator of utility of comparative information for mapping and assembly of one organism using the assembled sequence of a related organism as a template.

2. Detection of orthologous anchors

Orthologous anchors may be identified by comparing two assembled genomes, one assembled genome against an extensive collection of contigs from another genome, or by comparing contigs from two partially assembled genomes. (The word "contig" here refers to either to a group of contiguous assembled sequencing reads or to a single read.)

On one hand, the comprehensive nature of all-against-all sequence comparisons of two genome-scale databases is computationally demanding. On the other hand, it allows identification of truly orthologous anchors with high specificity via the application of a reciprocal best match filter (Mural *et al*., 2002).

The most widely used comparison programs are optimized for fast and sensitive comparison of short queries against a single large database (Altschul *et al*., 1997; Pearson, 2000). In contrast, for maximum utility of a "reciprocal best match" filter, anchoring ideally requires simultaneous comparison of genome-sized databases. A new generation of comparison programs employ various types of genome-scale

indices that are designed to fit within computer memory (RAM), which provides the speed required for genome comparison (Ning *et al*., 2001; Kent, 2002; Ma *et al*., 2002; Kalafus *et al*., 2004). One of these programs was specifically designed for the anchoring task (Kalafus *et al*., 2004).

3. Comparative mapping of reads and contigs

Anchoring of contigs from one species onto an assembled genome of the other related species provides hypothetical order, orientation, and distance information. This information can be independently confirmed by a variety of methods. For example, mate-pair reads obtained from both ends of clone inserts of known size may bridge the gap between two anchored contigs, thus confirming their putative distance and orientation. Anchored contigs may also be bridged computationally from the read overlap graph by constructing a tiling path of overlapping reads. Alternatively, the bridging may be performed by targeted experiments, such as PCR and primer walking.

There are two general contig mapping scenarios. The first scenario involves one assembled genome and the contig set from another, partially assembled genome (Pletcher *et al*., 2001). The second scenario involves two genomes assembled at the contig level only, each providing only fragmentary information for bootstrapping each other's assembly. The algorithmic challenge inherent in the latter method has recently been addressed (Veeramachaneni *et al*., 2003).

4. Comparative mapping of clones

A variety of methods exists for mapping cloned genomic fragments, such as Bacterial Artificial Chromosomes (BACs), of one species onto an already assembled genome of another. The information may be either used to select clones for targeted comparative sequencing or to infer their hypothetical order, overlap and distance, or for both purposes.

One method for comparative physical mapping starts with a comparison of already assembled genomes to identify regions generally conserved across species. The conserved regions are then used to design hybridization probes for identification of homologous clones from third species. Key assumption is that the sequence conservation between two species is a good indicator of its conservation in the third, related species. The method has been applied to the construction of orthologous clone–contig maps in multiple species (Thomas *et al*., 2002).

A more direct probe design requires whole-genome shotgun reads from the mapped species. Reads anchored onto an assembled genomic sequence of a related species can be used as intraspecies hybridization probes to identify clones that map onto the anchoring sites. A large collection of mouse BACs has been mapped onto the human genome by this method (Thomas *et al*., 2000).

An even more direct comparative clone mapping method involves sequencing of clone ends and their anchoring onto the related genome. A clone is mapped by this method if the distance between the two anchoring sites approximates

clone insert size. Recent large-scale applications of this method involved BAC end sequencing (BES) of large collections of chimpanzee (Fujiyama *et al*., 2002) and bovine (Larkin *et al*., 2003) BACs and their mapping onto the assembled human genome.

Finally, recently proposed Pooled Genomic Indexing (PGI) method (Milosavljevic *et al*., 2005; Csuros and Milosavljevic, 2004) maps BACs of one species onto the genome of another by shotgun sequencing of BAC pools. Pools are designed so that two pools have at most one BAC in common. Consequently, if reads originating from two pools anchor within the same BAC-sized genomic location of a related genome, the BAC that is present in both pools is mapped onto that location. PGI has the potential to achieve significant reduction in cost and efficiency of comparative mapping. In a first genome-scale application of PGI, a library of rhesus macaque BACs is being mapped onto the human genome (Milosavljevic *et al*., 2005 #716).

5. Comparative sequence assembly

Assembled genomic sequence of one species can be used as a template to guide the assembly of another, related species. In contrast to the independent assembly, where only read overlap information is utilized in the assembly process, comparative assembly maximizes both read overlap and similarity of the newly assembled sequence to a reference template (Milosavljevic, 1999). An information-theory argument shows that the comparative information enables better detection of conserved regions than the comparison of independently assembled genomes (Milosavljevic, 1995).

Comparison of independently assembled genomic sequences frequently results in differences due to either evolutionary rearrangements or to assembly errors. Detection of the assembly errors leads to an improved assembly (Pletcher *et al*., 2001; Rat Genome Sequencing Project Consortium, 2004). To avoid assembly errors from occurring in the first place, comparative information may be employed earlier in the assembly process. Ideally, the comparative information is employed so as to simultaneously improve both the extent and quality of assembly.

Selection of appropriate comparative assembly strategy depends on specific circumstances. For example, if the genome in question has been covered by shotgun reads, comparative information may be employed in order to localize the assembly process. Specifically, some of the reads are first directly anchored onto their orthologous locations in the reference genome. The anchored reads are then used as “baits” to “fish out” some of the remaining nonanchored reads using intraspecies read overlap information, thus increasing the number of localized reads. Finally, the localized reads are locally assembled. Localized assembly process improves the extent and quality of assembly at low read coverage, particularly in the presence of repetitive elements.

Human genomic sequence has been used as a reference for the initial assembly of genomes of the mouse and dog (Abbott, 2000; Kirkness *et al*., 2003). A published NIH report anticipates that primate genome sequencing projects will be greatly aided by the availability of the finished human genome (NIH-NCRR, 2001). A

comprehensive pipeline for comparative assembly of bacterial strains has been recently developed (Pop, 2004 #740).

6. Comparative assembly of transcribed sequences

Expressed sequence tags (ESTs) obtained from sequencing cDNA clones can be assembled into transcript sequences using an assembled genomic sequence as a reference. In addition to their inherent biological significance, assembled transcript sequences have utility for the design of gene expression probes. Specifically, human EST fragments can be grouped using human genomic sequence and then locally assembled for the purpose of designing oligonucleotide probes for the analysis of gene expression.

7. Significant trends in sequencing technology

In contrast to the current applications of comparative sequence assembly, which are driven by the availability of assembled genomes, early applications were driven by necessity: short sequence fragments detected by hybridization technology could not be assembled independently and comparative assembly provided the solution (Milosavljevic, 1995; Milosavljevic, 1999). With the advent of DNA chip technologies, comparative sequencing using hybridization probes is now expanding to the genome scale (Frazer *et al*., 2001; Frazer *et al*., 2003).

New sequencing technologies have the potential to increase the throughput and decrease the cost of sequencing by orders of magnitude. These trends may lead to the expansion of the applications of sequencing in the same way that the increased performance and decreased cost of microprocessors expanded the applications of computing. One of the obstacles for wide adoption of the new technologies, such as pyrosequencing (Ronaghi, 2001), single molecule array-based sequencing (*see* Article 7, **Single molecule array-based sequencing**, Volume 3), massively parallel genome sequencing, and Real-time DNA Sequencing (*see* Article 8, **Real-time DNA sequencing**, Volume 3), is higher sequencing error rate and significantly shorter read length compared to the standard Sanger method. Such sequence fragments tend to be too short for independent assembly, particularly in view of the repeat-rich nature of mammalian genomes (*see* Article 2, **Algorithmic challenges in mammalian whole-genome assembly**, Volume 7) but they are long enough for the sort of anchored comparative mapping and assembly described above.

8. Summary

Comparative information has the potential to both decrease the cost and accelerate mapping and sequencing projects by reducing experimental effort. Various comparative mapping and sequencing methods have already been put in practice. Future demand for comparative mapping and assembly is likely to be driven

by two trends: the increasing availability of reference genomes as potential templates for assembly and advances in sequencing technology. If current trends continue, comparative assembly of individual human genomes and even diagnostic sequencing of individual tumor samples will become a routine practice.

A number of challenges will have to be overcome along the way to efficient gigabase-scale sequence assembly. One is the development of computational means for fast and accurate sequence anchoring based on all-against-all comparisons of exponentially increasing sequence databases. Another significant challenge is the development of comparative mapping and assembly methods and their systematic characterization and validation.

References

Abbott A (2000) Drive for more genomes threatens mouse sequence. *Nature*, **405**(6787), 602–603.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Csuros M and Milosavljevic A (2004) Pooled Genomic Indexing (PGI): Analysis and design of experiments. *Journal of Computational Biology*, **11**(5), 1001–1021.

Frazer KA, Chen X, Hinds DA, Pant PV and Patil N (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Research*, **13**(3), 341–346.

Frazer KA, Sheehan JB, Stokowski RP, Chen X, Hosseini R, Cheng JF, Fodor SP, Cox DR and Patil N (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Research*, **11**(10), 1651–1659.

Fujiyama A, Watanabe H, Toyoda A, Taylor TD, Itoh T, Tsai SF, Park HS, Yaspo ML, Lehrach H, Chen Z, *et al*. (2002) Construction and analysis of a human-chimpanzee comparative clone map. *Science*, **295**(5552), 131–134.

Kalafus KJ, Jackson AR and Milosavljevic A (2004) Pash: Efficient genome-scale sequence anchoring by positional hashing. *Genome Research*, **14**(4), 672–678.

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, *et al*. (2003) The dog genome: Survey sequencing and comparative analysis. *Science*, **301**(5641), 1898–1903.

Larkin DM, Everts-Van Der Wind A, Rebeiz M, Schweitzer PA, Bachman S, Green C, Wright CL, Campos EJ, Benson LD, Edwards J, *et al*. (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Research*, **13**(8), 1966–1972.

Ma B, Tromp J and Li M (2002) PatternHunter: Faster and more sensitive homology search. *Bioinformatics*, **18**(3), 440–445.

Milosavljevic A (1995) DNA sequence recognition by hybridization to short oligomers. *Journal of Computational Biology*, **2**(2), 355–370.

Milosavljevic A (1999) DNA Sequence Similarity Recognition by Hybridization to Short Oligomers, U.S. Patent 6001562.

Milosavljevic A, Harris RA, Sodergren EJ, Jackson AR, Kalafus KJ, Hodgson A, Cree A, Dai W, Csuros M, Zhu B, *et al*. (2005) Pooled genomic indexing of rhesus macaque. *Genome Res*, **15**(2), p. 292–301.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, *et al*. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, **296**(5573), 1661–1671.

NIH-NCRR (2001) *Recommendations for Future Efforts in Primate Genomics*. National Institutes of Health, National Center for Research Resources Report, submitted to NCRR February 14, 2001, endorsed by the National Advisory Research Resources Council May 17, 2001.

Ning Z, Cox AJ and Mullikin JC (2001) SSAHA: A fast search method for large DNA databases. *Genome Research*, **11**(10), 1725–1729.

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, **132**, 185–219.

Pletcher MT, Wiltshire T, Cabin DE, Villanueva M and Reeves RH (2001) Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics*, **74**(1), 45–54.

Pop M, Phillippy A, Delcher AL and Salzberg SL (2004) Comparative genome assembly. *Brief Bioinform*, **5**(3), 237–248.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **11**(1), 3–11.

Thomas JW, Prasad AB, Summers TJ, Lee-Lin SQ, Maduro VV, Idol JR, Ryan JF, Thomas PJ, McDowell JC and Green ED (2002) Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Research*, **12**(8), 1277–1285.

Thomas JW, Summers TJ, Lee-Lin SQ, Maduro VV, Idol JR, Mastrian SD, Ryan JF, Jamison DC, and Green ED (2000) Comparative genome mapping in the sequence-based era: Early experience with human chromosome 7. *Genome Research*, **10**(5), 624–633.

Veeramachaneni V, Berman P and Miller W (2003) Aligning two fragmented sequences. *Discrete Applied Mathematics*, **127**(1), 119–143.

Short Specialist Review

Statistical signals

Andrew J. Gentles and Samuel Karlin
Stanford University, Stanford, CA, USA

1. Introduction

Detection of statistical signals in protein and DNA sequences is often a powerful method for uncovering biological significance and function. Examples of sequence features that can be identified through their statistical properties include protein motifs (*see* Article 28, **Computational motif discovery**, Volume 7), gene promoters, enhancers and suppressors (*see* Article 19, **Promoter prediction**, Volume 7 and Article 24, **Exonic splicing enhancers and exonic splicing silencers**, Volume 7), transcription binding sites, exon/intron splice junctions, regions of homology between sequences (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7), and protein docking sites. Each generally possesses some unique statistical property, or signal, that distinguishes it from other sequence segments. Signals may be interesting in their own right but are often combined in algorithms such as multiple alignments, homology detection, and gene finding (*see* Article 13, **Prokaryotic gene identification *in silico***, Volume 7, Article 14, **Eukaryotic gene finding**, Volume 7, Article 20, **Operon finding in bacteria**, Volume 7, Article 21, **Gene structure prediction in plant genomes**, Volume 7, Article 25, **Gene finding using multiple related species: a classification approach**, Volume 7, and Article 27, **Gene structure prediction by genomic sequence alignment**, Volume 7).

2. Maximal aggregate scoring segments

Central to identification of statistically significant signals, either in individual sequences or when comparing a group of sequences, is the general scoring scheme (Karlin and Altschul, 1990). Given a score defined on individual letters of a sequence (e.g., the amino acid substitution matrices BLOSUM62 and PAM120, commonly used in sequence alignment; for recent developments, see Yu *et al*., 2003), the general problem is to determine the statistical significance of a sequence region exhibiting a high aggregate (additive) score. Let S be a sequence composed of letters from an alphabet $A = \{a_1, a_2, \ldots, a_r\}$ that are independently randomly

distributed with probabilities $\{p_1, p_2, \ldots, p_r\}$. For each letter in S, one can associate a score s_i (e.g., the score for a matching letter in another sequence). In general, we are interested in finding the segment of the sequence that has the maximum aggregate score. Two important conditions are applied. First, we require that at least one score be positive. Second, we require the expected score per letter, $E = \sum p_i s_i$, be negative. If E were positive, the maximal segment would tend to be the whole sequence. The maximal aggregate scoring segment M_n in a sequence of length occurs with probability given by

$$\text{Prob}\left\{M_n \geq (\ln n)/\lambda^* + x\right\} \approx 1 - \exp\left\{-K^* e^{-\lambda^* x}\right\} \tag{1}$$

where λ^* is the unique positive solution of $\sum_{i=1}^{r} p_i e^{\lambda^* s_i} = 1$, and K^* is determined from a rapidly convergent series (see Karlin and Altschul, 1990). This approach can be extended to allow for neighbor dependencies (e.g., Markov models), as well as random and vector scores. To use the maximal scoring formulas, one chooses a significance level p (e.g., $p = 0.001$) and solves for $x = x(p)$ so that a segment score exceeding $(\ln n)/\lambda^* + x(p)$ is significant at the p level. The homology search algorithm, BLAST applies the ideas of maximal aggregate score to find matching sequence blocks between query and database sequences (Altschul *et al*., 1997).

For example, consider the case of aligning two sequences where gaps are not allowed (Dembo *et al*., 1994). The first sequence is composed of letters a_i with probability p_i, and the second sequence of letters a_j with probability p'_j, so that the pairing of a_i with a_j occurs with probability $p_i p'_j$. Assuming that the average pair score $\sum_{i,j} p_i p'_j s_{ij}$ is negative, then λ^* is the unique positive solution of $\sum_{i,j} p_i p'_j \exp\{\lambda^* s_{ij}\} = 1$. If the sequence lengths m and n grow at roughly similar rates, then equation (1) holds for the maximal scoring segmental alignment with n replaced by mn:

$$\text{Prob}\left\{M > \frac{\ln nm}{\lambda^*} + x\right\} \approx K^* \exp\left\{-\lambda^* x\right\}. \tag{2}$$

In other words, the alignment of two sequences has an unusually high (statistically significant at the $p = 0.01$ level) score if M exceeds $(\ln nm)/\lambda^* + x_0$, with x_0 determined so that $K^*\exp\{-\lambda^* x_0\} = 0.01$. The number of "separate" high-scoring segments is close to a Poisson distribution with parameter $K^*\exp\{-\lambda^* x_0\}$. This distribution can be used to estimate whether the number of high-scoring segments over a whole protein is unusually high. The theory readily generalizes to the case of multiple sequence alignments (Dembo *et al*., 1994).

The above results can be used in discriminating among possible scoring schemes. Suppose that $\{p_i\}$ are the overall letter frequencies in some random reference sequence. Now, let $\{q_i\}$ be desirable *target frequencies* that correspond to known representatives of some type of protein sequence that we wish to identify. Then, in a high-scoring segment, $q_i \approx p_i \exp\{\lambda^* s_i\}$, and appropriate scores are naturally defined as $s_i = \ln(q_i/p_i)$, With these log ratio scores, the mean score for a sample letter is automatically negative. The PAM and BLOSUM matrices are examples of natural log odds ratio score matrices.

3. Significant segment pair alignments (SSPA)

Most measures of similarity between protein sequences are local. One way to score global similarity between two protein sequences A and B is as follows: determine all high-scoring sequence segment pairs (HSSPs) significant at the $p = 0.01$ level. A *consistent matching array* is defined as a combination of HSSPs (optimizing overlaps) into a single "gapped" alignment. The significant segment pair alignment (SSPA) score for the sequence A and B is the maximal value with respect to all sets of consistently matching arrays, obtained by summing the segment scores. Normalization is then performed so that the SSPA score is

$$\sigma(A, B) = 100 \times \frac{\text{maximal total score w.r.t. consistent matching arrays}}{\max(\text{self score of } A, \text{self score of } B)} \qquad (3)$$

or, alternatively, normalization can be done with respect to the minimum of the two self-scores. This normalization allows comparison of proteins of different sizes and quality. In either case, $0 \leq \sigma(A,B) \leq 100$. For sequence pairs with at least one significant SSPA match, additional matching segments are identified at a lower probability threshold (typically $p = 0.50$) and appended to the alignment. The secondary lower threshold helps to fill in gaps between significant HSSPs. The SSPA method can be incorporated into a multiple alignment algorithm that employs dynamic programming to accommodate insertions/deletions and variable-length unaligned segments (Brocchieri and Karlin, 1998).

4. Analysis of marker arrays: *r*-scan analysis

A frequent problem in sequence analysis is to determine the statistical significance of *spacings* between specific markers distributed throughout a sequence. For example, one might seek to determine whether direct repeats or dyads (a sequence fragment followed by its reverse complement) are significantly clumped (many neighbors with short spacings) or over-dispersed. The method of *r-scans* provides a rigorous framework for such determinations (reviewed in Reinert *et al*., 2000). Let the r-scan length $R_i^{(r)}$ be the distance between marker i and marker $i + r$, that is the cumulative length of r distances between markers. Extremal statistics for the $\{R_i^{(r)}\}$ can be determined from their limiting distributions. The distribution of a marker is determined by comparing the distribution of $\{R_i^{(r)}\}$, calculated for a random sequence to those actually observed. Varying r permits detection of clustering or dispersion at different scales. Let the minimum and maximum r-scans be $m_r^* = \min_i R_i^{(r)}$ and $M_r^* = \max_i R_i^{(r)}$. The theoretical probabilities for a marker array of n points obey the asymptotic relations

$$\text{Prob}\left\{m_r^* \geq \frac{x}{n^{(1+1/r)}}\right\} \approx \exp\frac{-x^r}{r!} \text{ and}$$
$$\text{Prob}\left\{M_r^* \leq \frac{\ln n + (r-1)\ln(\ln n) + x}{n}\right\} \approx \exp\frac{-e^{-x}}{(r-1)!}. \qquad (4)$$

These equations provide thresholds for whether minimum and maximum observed spacings deviate significantly from random expectation. For example, setting the probability on the left-hand side to a required significance (typically 0.01) yields the formula $\exp(-x_b^r/r!) = 0.01$, which is solved for x_b, and the threshold value $b_r^* = x_b/n^{(1+1/r)}$ is then determined. For a sequence of length L, observed r-scan lengths of $m_r^* \leq b_r^* L$ define r-scan clusters. Similarly, significantly even spacing is indicated by $m_r^* > a_r^* L$, where a_r is determined from equations (4) by setting the probability to 0.99. Analogous expressions can be obtained for the maximal spacing M_r*, and markers are considered to be distributed randomly if the r-scan lengths are in the ranges $(a_r^* L,\ b_r^* L)$ and $(A_r^* L,\ B_r^* L)$. r-scan statistics can also identify whether counts in sliding window plots are significant. For a window of size w, let $N(t)$ be the number of markers in $(0,t)$. The critical value r^* that corresponds to a significance value p (e.g., $p = 0.01$) for the marker counts in the window can be calculated from

$$\text{Prob}\left\{\max_{w\leq t\leq 1}\,[N(t) - N(t-w)] > r\right\} = \text{Prob}\,\{m(r) \leq w\} \approx 1 - \exp\frac{-x^r}{r!} = p \qquad (5)$$

where $m(r)$ is the length of the minimum r-scan. Defining n by $x = wn^{1+1/r}$ gives $n(nw)^r/r! + \ln(1-p) = 0$, which can be solved numerically to give r^* for values of n, w, and p. This sliding window method correctly predicts the region of the lytic origin of replication for human cytomegalovirus (Masse *et al*., 1992).

5. Frequent words (oligonucleotides and peptides)

Another topic of interest is determination of sequence words that occur statistically more, or less, frequently than would be expected from random chance. For a sequence of length L, comprising letters drawn from an alphabet of size A, there is a natural word length s defined by the inequality $A^{s-1} \leq L < A^s$ and a natural copy number r determined by $(r-1)/r < (\log L)/\log A^s \leq r/(r+1)$. For sufficiently large L, the number of words of length s that occur more than r times is approximately Poisson distributed. This model, based on approximations to Poisson distributions concerned with generalized occupancy problems of balls-in-urns, works well in the case where the letter frequencies are roughly equal for the sequence (Karlin and Leung, 1991). In the case of a biased sequence, for example, an A + T rich genome such as *Haemophilus influenzae*, the method can be generalized to the case of words w that have separate frequencies p_w. The word size s is as defined previously, but the copy number threshold, r_w, for a particular word now satisfies $p_w L \exp(-p_w L)/r_w! \leq 1/L$. In this formalism, the lower the expected frequency of the word, the lower the cutoff for it to be considered frequent. If the inequality is saturated, at most, one frequent word is expected in a random sequence of the same length. For independent letters, $p_w = f_1 f_2 \ldots f_s$, where f_i is the frequency of occurrence of the ith letter in the sequence. For a Markov model, with transition probabilities $f_{i,j}$ between the ith and jth letters, $p_w = f_1 f_{1,2} \ldots f_{s-1,s}$. For the *H. influenzae* genome, $L \approx 1.8$ million bp, which, with $A = 4$ for a nucleotide alphabet, gives a natural frequent word length of $s = 9$ bp. Applying the copy number threshold reveals

that, amongst others, the sequence AAGTGCGGT and its inverted complement qualify as frequent words. These 9-mers can be recognized as the DNA-uptake signal sequences of *H. influenzae*, which are highly enriched in the genome of this bacterium and which facilitate the binding and uptake of *H. influenzae's* own DNA or DNA from closely related species. In *Escherichia coli*, REP elements and Chi sites qualify as frequent words, along with pentapeptides corresponding to the ATP-binding domain. For *Synechocystis*, the palindrome GGCGATCGCC is highly frequent and significantly evenly distributed. In eukaryotes, amino acid frequent words include words characteristic of zinc fingers (CGKAF, CEECG), chymotrypsin proteases (LTAAH, GDSGGP), serine/threonine and tyrosine kinases (ADFGL, FGQGT), and homeobox proteins (FQNRR, HFNRY).

6. Repeat structures

The patterns of significantly long repeats, frequency, length, and genomic environment may provide insights on evolution of genomes. All large repeats of the *H. influenzae* genome can be determined, allowing for error blocks (Leung *et al*., 1991). There are 150 occurrences in 51 repeat families; all involve at least 75% identities:

Sizes (bp)	25–29	30–39	40–59	60–99	100–139	140–179	180–399	400–699	700–1699	=3000
Occurrences	14	18	17	14	19	17	19	19	10	3

Spacings between repeats:

	tandem	2 bp–1 kb	1–10 kb	10–200 kb	=200 kb
Number of matches	18	8	7	28	33

Number of occurrences in distinct repeat families:

Family size	2	3	4	5	6	13	14
Count of families	38	5	3	2	1	1	1

Some atypical examples of direct repeats in *H. influenzae* include the following: (1) A ~3.2-kb direct repeat (97% identity). The first occurrence at about 1371–1380 kb in the published sequence contains two major insertions (0.9 and 4.2 kb long). The second uninterrupted occurrence is at 1416–1420 kb. (2) A 600-bp tandem-perfect direct repeat (one base insertion) spans the 3′ part of gene *guaB* and the 5′ part of gene *guaA*. (3) Three copies of ~500 bp with 97% identical sequences, all intergenic, around positions 173, 723, and 1344 kb.

Further reading

Dembo A and Karlin S (1992) Poisson approximations for *r*-scan processes. *Annals of Applied Probability*, **2**, 329–357.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Brocchieri L and Karlin S (1998) A symmetric-iterated multiple alignment of protein sequences. *Journal of Molecular Biology*, **276**, 249–264.

Dembo A, Karlin S and Zeitouni O (1994) Critical phenomena for sequence matching with scoring; Limit distribution of maximal non-aligned two-sequence segmental score. *Annals of Probability*, **22**, 1993–2039.

Karlin S and Altschul S (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA*, **87**, 2264–2268.

Karlin S and Leung M-Y (1991) Some limit theorems on the distributional patterns of balls in urns. *Annals of Applied Probability*, **1**, 513–538.

Leung M-Y, Blaisdell BE, Burge C and Karlin S (1991) An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *Journal of Molecular Biology*, **221**, 1367–1378.

Masse MJO, Karlin S, Schachtel G and Mocarski ES (1992) Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proceedings of the National Academy of Sciences of the USA*, **89**, 5246–5250.

Reinert G, Schbath S and Waterman MS (2000) Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7**, 1–46.

Yu Y-K, Wootton JC and Altschul SF (2003) The compositional adjustment of amino acid substitution matrices. *Proceedings of the National Academy of Sciences of the USA*, **100**, 15688–15693.

Short Specialist Review

Errors in sequence assembly and corrections

Martti T. Tammi

Karolinska Institutet, Stockholm, Sweden

National University of Singapore, Singapore

Björn Andersson

Karolinska Institutet, Stockholm, Sweden

1. Introduction

The major source of all the challenges is the limitation in sequencing technologies, which today allows us to routinely sequence only about 500 to 800 bases of contiguous DNA sequence. To overcome the limitation of short contiguous sequences, Frederick Sanger devised the shotgun sequencing technique and in 1982 demonstrated its potential by sequencing the genome of the bacteriophage lambda (Sanger *et al.*, 1982). The process of shotgun sequencing begins with physically shearing the original DNA molecules into small pieces as randomly as possible. The fragments are subsequently inserted into cloning vectors and amplified by growing them in *Escherichia coli*. The ends of the inserts are sequenced and sequenced reads assembled by specialized computer software, shotgun fragment assembly programs (Havlak *et al.*, 2004; Huang *et al.*, 2003; Huson *et al.*, 2001; Jaffe *et al.*, 2003; Kent and Haussler, 2001; Mullikin and Ning, 2003; Pevzner *et al.*, 2001; Pop *et al.*, 2004; Tammi *et al.*, 2003a; **www.phrap.org**.).

Despite continuous technology improvements during the three decades after Sanger devised the dideoxy DNA sequencing method (Sanger *et al.*, 1977), the average sequence length in routine production has only increased from about 200–300 contiguous bases to 500–800 bases. Although the average length has more than doubled, the increase is still too small to make a significant impact on the efficiency the fragment assembly. However, the throughput has increased enormously thanks to fluorescent detection of the base sequence (Smith *et al.*, 1986) and automatic sequencers.

Although the sequence assembly problem may appear simple, it is known to be NP-complete. In addition, sequence assembly is complicated by a number of

factors, but the three most important ones are sequencing or base-calling errors, genomic repeats, and polymorphism (*see* Article 11, **Algorithms for sequence errors**, Volume 7 and Article 2, **Algorithmic challenges in mammalian whole-genome assembly**, Volume 7).

2. Sequencing errors

Sequenced fragments contain base-calling errors, which can be incorrectly determined bases, insertions, and deletions. The number of errors tends to increase toward the ends of the reads. This is unfortunate, since most of the overlaps between reads are determined using the very ends of the sequenced fragments.

3. Repeats

In shotgun sequencing, the original idea of putting the puzzle of sequences together by using sequence similarity often fails in the case of repeats. Genomic sequence contains many kinds of repeats of varying lengths. Repeat copies may be identical or almost identical, only differing by a few bases. They can be dispersed all over the genome, and/or repeated in tandem, and contain any number of copies. Repeated regions are difficult to separate and cause assembly programs to assemble fragments originating from different copies together, resulting in erroneous assemblies. The combination of sequencing errors and repeated sequences pose the greatest challenge in shotgun fragment assembly. It is probably close to impossible to correctly assemble identical repeat copies. However, it may be argued that no information is lost, but identical repeat copies may cause large artificial genomic rearrangements. Repeats that are shorter than the average read length cause fewer problems than longer copies.

4. Polymorphism

In whole-genome shogun sequencing, polymorphism complicates the assembly of nearly identical repeats and in some cases also nonrepetitive regions, but it is not a problem in a “clone-by-clone” approach, since only one variant of each genomic region is sampled (*see* Article 12, **Polymorphism and sequence assembly**, Volume 7).

5. Incomplete coverage

The fraction $\frac{NL}{G}$ represents the amount of oversampling of the genomic sequence G, where N is the number of sampled reads and L is the average length of the reads. This is also called *coverage*. Some assembly programs use statistical methods, searching the best overlaps and may easily assign false overlaps if no better overlap is present due to lack of coverage. There are genomic regions that are impossible to sequence due to biological reasons. However, the output from an assembly program

may consist of several contigs also due to an erroneous assembly, or due to the nature of sampling process itself, all resulting in gaps in coverage. The coverage has a certain probability to be zero, depending on the amount of sampling. When reads are sequenced from random fragments, not all genomics positions are equally sampled. The Lander and Waterman expression (Lander and Waterman, 1988) can be used to calculate the average number of gaps, contigs, and their average length in a shotgun project; if a target sequence G is redundantly sampled, at an average coverage $\overline{c}$ and assuming that the sheared fragments are uniformly distributed along the genomic sequence, the coverage at a give base b is a Poisson random variable with mean $\overline{c}$:

$$P(\text{base } b \text{ convered by } k \text{ fragments}) = \frac{e^{-\overline{c}} \cdot \overline{c}^k}{k!} \tag{1}$$

For example, the fraction of G that is covered by at least one fragment is $1 - e^{-\overline{c}}$. The Lander and Waterman expressions can also be used to compute the fraction of reads involved in overlaps, detected by an assembly program, but which ones of the particular gaps are due to an erroneous assembly of reads is of course hard to estimate. However, some methods have used a high coverage as an indicator of many copies of repeats assembled together (Myers *et al*., 2000).

An additional complicating factor is the unknown orientation of the shotgun fragments. It is not known from which DNA strand each sequenced fragment originates, and this increases the complexity of the assembly task. Hence, a read may be present on one strand or it is the reverse complement sequence on the other strand. In a sequencing project containing one million sequenced reads, a complementary set of reads must be generated resulting in two million reads. The complete set is required for the evaluation of all overlaps. This and possible contamination and unremoved vector sequences may cause severe errors in assemblies.

A number of methods have been invented in order to address these problems. One powerful way is the assignment of error probabilities to base-calls (Ewing and Green, 1998). The knowledge of the quality of the base-calls allows efficient trimming of the reads and statistical evaluation of candidate overlaps (**www.phrap.org**). A detailed analysis of highly similar repeat copies may be performed (Tammi *et al*., 2002) as well as error correction of the sequenced reads (Tammi *et al*., 2003b; Pevzner *et al*., 2001). Since clone inserts are usually sequenced from both ends and the insert lengths are known, this information can be used to position the read pairs within the assembly (Edwards and Caskey, 1990; Myers *et al*., 2000). It is likely that the combination of the information given by sequenced pairs and statistical methods for analysis of highly similar repeats yields the most powerful approach in the struggle against misassemblies caused by repeats and sequencing errors. It is, however, clear that current sequence assembly software is only capable of producing a draft sequence, and much labor-intensive manual finishing is required to arrive at a good quality, reliable, finished sequence.

The human genome was sequenced both by a public (Lander *et al*., 2001) and a private initiative (Venter *et al*., 2001). The public initiative used the hierarchical or "clone-by-clone" approach, a method that involves extensive mapping. The

mapping step was avoided by the whole-genome shotgun (WGS) approach, which was used by Celera in the privately funded effort. A comparison of these approaches by She *et al*. (2004) showed that the WGS approach runs into problems on the repeated parts of the genome. Apart from the highly repetitive telomere and centromere regions that are not targeted by either initiative, the WGS approach was unable to adequately resolve larger than 15 kb where the difference between copies was 3% or less. About 4% of the sequence was lost or erroneously assembled because of collapsed repeated regions. This leads to significant reduction of the actual genome length and the loss of many biologically important regions including genes. It is likely that a combination of the sequencing strategies is the most advantageous one.

Related articles

Article 3, **Hierarchical, ordered mapped large insert clone shotgun sequencing**, Volume 3, Article 24, **The Human Genome Project**, Volume 3 and Article 25, **Genome assembly**, Volume 3

References

Edwards A and Caskey CT (1990) Closure strategies for random DNA sequencing methods. *A Companion to Methods in Enzymology*, **3**, 41–47.

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**(3), 186–194.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM and Gibbs RA (2004) The Atlas genome assembly system. *Genome Research*, **14**(4), 721–732.

Huang X, Wang J, Aluru S, Yang SP and Hillier L (2003) PCAP: a whole-genome assembly program. *Genome Research*, **13**(9), 2164–2170.

Huson DH, Reinert K, Kravitz SA, Remington KA, Delcher AL, Dew IM, Flanigan M, Halpern AL, Lai Z, Mobarry CM, *et al*. (2001) Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics*, **17**(Suppl 1), S132–S139.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC and Lander ES (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research*, **13**(1), 91–96.

Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, **11**(9), 1541–1548.

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3), 231–239.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Mullikin JC and Ning Z (2003) The phusion assembler. *Genome Research*, **13**(1), 81–90.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mabarry CM, Reinert KH, Remington KA, *et al*. (2000) A whole-genome assembly of Drosophila. *Science*, **287**(5461), 2196–2204.

Pevzner PA, Tang H and Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(17), 9748–9753.

Pop M, Phillippy A, Delcher AL and Salzberg SL (2004) Comparative genome assembly. *Briefings in Bioinformatics*, **5**(3), 237–248.

Sanger F, Coulson AR, Hong GF, Hill DF and Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. *Journal of Molecular Biology*, **162**(4), 729–773.

Sanger F, Nicklen S and Coulson A (1977) DNA sequencing with chainterminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL and Eichler EE (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**(7011), 927–930.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB and Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**(6071), 674–679.

Tammi MT, Arner E and Andersson B (2003a) TRAP: Tandem Repeat Assembly Program produces improved shotgun assemblies of repetitive sequences. *Computer Methods and Programs in Biomedicine*, **70**(1), 47–59.

Tammi MT, Arner E, Kindlund E and Andersson B (2003b) Correcting errors in shotgun sequences. *Nucleic Acids Research*, **31**(15), 4663–4672.

Tammi MT, Arner E, Britton T and Andersson B (2002) Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. *Bioinformatics*, **18**(3), 379–388.

Venter JC, *et al*. (2001) The sequence of the human genome. *Science*, **291**(5507), 1304–1351. **www.phrap.org**.

Basic Techniques and Approaches

Genome maps and their use in sequence assembly

Paul H. Dear
MRC Laboratory of Molecular Biology, Cambridge, UK

1. Introduction

Genomes range in size from around a million base pairs to many thousands of millions, and yet a typical sequencing reaction yields less than a thousand base pairs of contiguous sequence information. From these tiny fragments of data, the complete genome sequence must be reconstructed as accurately and as completely as possible if genes and other features are to be reliably identified. For the smallest and simplest genomes, the so-called "shotgun" approach is sometimes sufficient: the genome is fragmented and subcloned, and many randomly selected clones are sequenced. As the data accumulates, overlapping stretches of sequence are identified computationally to build contiguous sequences ("contigs") that grow and merge as the shotgun sequencing project progresses. When enough sequence data has been obtained, the hope is that a single contig can be assembled representing each chromosome.

For more complex genomes, however, shotgun sequencing alone is insufficient. Some parts of the genome may prove refractory to cloning, leaving gaps between a large number of unlinked (and therefore unordered) contigs. Other parts may contain complex repeats – large tracts of sequence that occur at multiple points in the genome – making unambiguous assembly impossible.

Genome maps address this problem by defining the relative positions of selected sequence features over distances of a few kilobases or more, complementing the short-range information provided by the shotgun data. They thus act like the index of a book, which gives the locations of keywords to help the reader navigate the detailed text. This article will outline the principal methods of genome mapping and will then consider how these approaches are integrated into the sequencing process to produce finished genome sequences.

2. Methods of genome mapping

A number of methods have evolved to produce maps of genomes, but here we will focus only on those methods that give sufficient detail to assist the assembly of sequence data. This largely excludes genetic mapping and fluorescence in situ

hybridization, both of which yield coarse maps that are useful mainly for validating the final sequence assembly on the largest scale.

2.1. Physical mapping

Physical (or clone-based) mapping is conceptually similar to shotgun sequence assembly but takes place on a much larger scale (*see* Article 13, **YAC-STS content mapping**, Volume 3 and Article 18, **Fingerprint mapping**, Volume 3). The aim is to find a series (or "tiling path") of cloned fragments that represent consecutive overlapping segments of the genome, but the clones that are used are one or two orders of magnitude larger than the small-insert clones used for shotgun sequencing. Typically, the genome is first cloned in BACs (bacterial artificial chromosomes), with insert sizes of 50–200 kb, and each clone in the library is analyzed. This analysis may involve tabulating the sizes of restriction fragments produced when each clone is digested with a chosen restriction enzyme ("fingerprinting") or testing each of the clones by PCR for the presence of many different specific sequences (known as "sequence tagged sites" or STSs, perhaps chosen from the early data from a shotgun sequencing project). Two clones are then inferred to overlap if they are found to have several restriction-fragment sizes in common or if they both contain the same STS (Figure 1).

Physical mapping suffers from the same limitations as the finer-scale shotgun sequence assembly: regions of the genome that cannot be cloned in BACs leave gaps in the physical map, while very large repetitive tracts (larger than the size of a single BAC clone) can lead to ambiguities in the construction of the tiling path. These problems can be partly overcome in some cases by the use of yeast artificial chromosomes (YACs) as the cloning vector, as these will accept larger inserts and will propagate some sequences that are unstable in bacterial cloning systems. However, YACs are technically more difficult to work with and are also more susceptible to a range of cloning artifacts. In spite of these difficulties, physical mapping is widely used in genome sequencing, often in conjunction with other approaches.

2.2. Optical mapping

Optical mapping (Samad *et al*., 1995; Zhou *et al*., 2003) is an elegant method, though it has so far been applied only to a rather narrow range of genomes. Genomic DNA is spread across a glass slide in such a way that the long molecules are stretched out in one direction and are loosely bound to the surface. The DNA is stained with a fluorescent dye and then treated with a restriction enzyme that cleaves the DNA at each occurrence of the enzyme's recognition sequence. When examined through a fluorescence microscope, the long DNA molecules appear as dotted lines, broken at each restriction site, and the sizes of the restriction fragments can be measured directly. By compiling many such images, a fairly precise "restriction map" is generated, showing the location of all the restriction sites in the genome. The pattern of restriction sites in the contigs produced by a

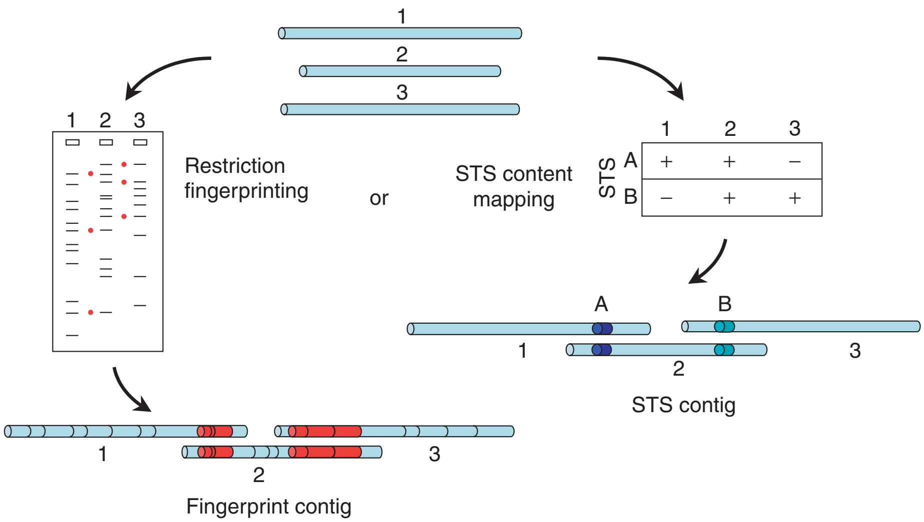


Figure 1 Physical mapping. Large-insert clones (top) can be characterized by digesting them with a restriction enzyme and measuring the sizes of the resulting fragments on a gel (left). Clones that have several fragment sizes in common (indicated by red dots) can be assumed to overlap, with the shared fragments arising from the regions of overlap (bottom left; common restriction fragments shown in red). Alternatively (right), each clone can be tested for the presence of many different STS markers (in this case, two markers A and B). Clones that carry the same STS markers can be inferred to overlap, the shared markers lying in the region of overlap (bottom; markers indicated by colored segments)

shotgun sequencing project can then be compared with the genome-wide restriction map to find the precise location of each contig in the genome (Figure 2).

The biggest advantage of optical mapping is its independence from cloning and the associated artifacts, though the data it produces (locations of restriction sites) is not as rich as that produced by other methods, complicating its direct use in large genomes.

2.3. *Radiation hybrid mapping*

Radiation hybrid (RH) mapping is conceptually similar to genetic linkage mapping (*see* Article 15, **Linkage mapping**, Volume 3) and gives information on the order and spacing of selected sequences (STSs) in the genome (Cox *et al*., 1990; *see also* Article 14, **The construction and use of radiation hybrid maps in genomic research**, Volume 3). Living cells of the species to be mapped (the "donor") are irradiated to break their chromosomes at random locations and are then fused with unirradiated cells of another species (the "host"). The result is a population of hybrid cells, each containing the host chromosomes along with a few random fragments of the donor genome. A library of many such cells is then tested, by PCR, to determine which donor-derived STSs are present in each hybrid. If two STSs are close to one another in the donor genome, then they will often remain on the same chromosome fragment after irradiation and hence will often be found together (or

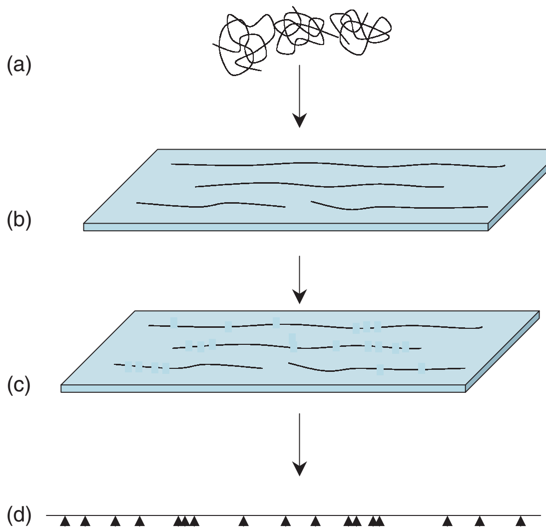


Figure 2 Optical mapping. Large fragments of genomic DNA (a) are spread and aligned on a glass surface (b) and stained for visualization. A restriction enzyme cuts the molecules at its recognition sites, and the cuts can be seen using a fluorescence microscope (c). Computerized measurement and analysis of many such images covering overlapping parts of the genome allows the precise pattern (d) of restriction sites (indicated by arrowheads) in the genome to be determined

"cosegregate") in the same hybrid cell. Conversely, STSs lying far apart in the donor genome will reside on different fragments after the irradiation and hence will segregate independently amongst the hybrids. By analyzing cosegregation frequencies, therefore, distances between STS markers can be estimated and a map constructed. The basic principle is similar to that used in HAPPY mapping (see below), as illustrated in Figure 3.

RH mapping is a powerful tool used to making maps of large genomes since it can be used to map widely spaced markers. However, it suffers from some technical limitations, including the difficulty of making hybrid cells from many donor species, the complicating presence of the host genome in the hybrids, and artifacts caused by biological factors influencing the retention of donor fragments in the hybrid cells.

2.4. *HAPPY mapping*

HAPPY mapping (Dear and Cook, 1993; *see also* Article 22, **The Happy mapping approach**, Volume 3) is analogous to RH mapping, but is entirely an in vitro process (Figure 3). Again, it begins by randomly breaking genomic DNA of the species to be mapped, either by mechanical means or by radiation. However, instead of segregating the fragments into hybrid cells, they are segregated simply by diluting and dispensing them into a series of samples, each containing only a few random fragments of the genome. Each sample is then screened by PCR to determine

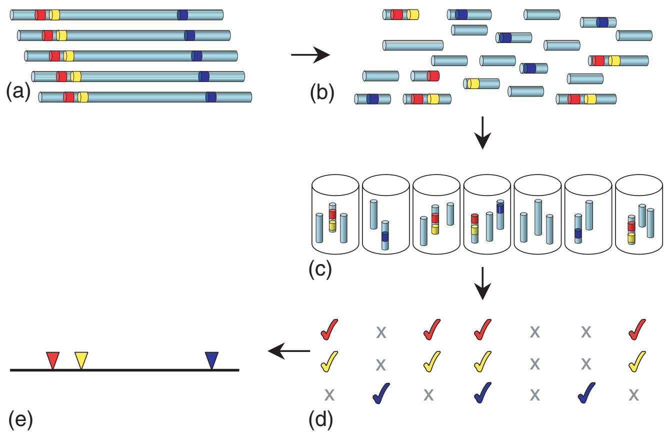


Figure 3 HAPPY mapping. Genomic DNA (a; colored segments represent STS markers) is broken into random fragments (b), which are greatly diluted and dispensed into a series of samples (c). Each sample is tested by PCR to determine the markers it contains (d). Closely linked markers (red and yellow) will tend to occur together (cosegregate) amongst the samples. By analyzing the cosegregation of many such markers, their order and spacing along the chromosome can be calculated to produce a map (e). The process of radiation hybrid mapping is similar, except that the DNA fragments are propagated in hybrid cells rather than as in vitro samples

the specific sequences (STSs) it contains. Much as in RH mapping, cosegregation frequencies can be used to deduce the order and spacing of the markers.

The main limitation of HAPPY mapping is the technical challenge of preparing and analyzing minuscule DNA samples, although this is now relatively straightforward. Its advantages arise from its in vitro nature and the consequent control that this gives. The method works equally well on all genomes and does not suffer from biological artifacts induced by specific sequences. Moreover, by choosing how finely to break the DNA at the outset, the level of detail in the maps can be precisely controlled, from relatively coarse long-range maps suited to large genomes (Dear *et al*., 1998) to detailed maps with accuracies of a few kilobases (Bankier *et al*., 2003).

3. Mapping as part of a sequencing program

Very few genome sequencing programs use a single methodology in its pure form, but we can outline the elements of two very different, simplified strategies as examples.

In the “top-down” approach, a physical map consisting of overlapping large-insert clones (typically BACs) is first constructed and carefully checked. Some years ago, physical mapping was conducted exclusively using the fingerprinting approach. Nowadays, it is often done by STS content mapping, and, in many cases, a proportion of the STSs would have been previously mapped by RH or other methods to provide an additional layer of positional information and to guard against errors in the physical map. Once the physical map is complete, a set of

mapped clones is chosen that covers the genome with the minimum of overlap. Each of the BAC clones in this "minimal tiling path" is then purified, subcloned as small fragments, and sequenced as an individual shotgun project. The genome is therefore sequenced segment by segment in an orderly manner.

The "bottom-up" approach, conversely, starts with extensive whole-genome shotgun sequencing and assembling of the data into sequence contigs. The contigs typically grow and merge as more data are accumulated, until a point of diminishing returns is reached where new sequence data largely duplicates those that are already obtained. At this point, the genome consists of many contigs, separated either by difficult-to-clone regions or by repeats that impede further unambiguous assembly. Mapping strategies are then used to find the arrangement of these contigs in the genome. For example, short sequences chosen from one end of each contig can be HAPPY mapped, showing how the sequence contigs are arranged in the genome. This is often sufficient to resolve repeat-induced gaps; clone-gaps can be closed, for example, by PCR between the sequences now known to flank the gap.

In reality, however, most genome projects use a mixed strategy to bring the sequence to completion. For example, the publicly funded human genome project began as a "top-down" approach, starting with detailed BAC physical maps, reinforced by RH and other marker-based maps. In response to a commercially funded pure shotgun approach, it then adopted a more shotgun-dependent strategy to allow rapid production of an unfinished draft sequence. The public consortium has since largely finished the sequence through a return to more methodical map-led approaches, whereas the pure shotgun strategy of Celera left the genome in over 100 000 unlinked contigs, as expected for a sequence of this size and complexity.

Most large-genome sequencing projects now rely heavily on a combination of shotgun sequencing and physical mapping approaches and also on "map as you go" strategies in which initial sequence data is used to identify those BAC clones that overlap with the existing contigs and those that can be sequenced to extend the contigs in a stepwise manner.

Conversely, the sequencing of the genome of the amoeba *Dictyostelium discoideum* began as a modified bottom-up approach, with shotgun sequencing not of the whole genome but of individual purified chromosomes. HAPPY mapping was then used to locate each sequence contig precisely in the genome, allowing the gaps in the sequence to be closed methodically using a variety of approaches. This type of approach is widely favored for smaller genomes since the initial shotgun phase yields valuable sequence data, and the subsequent mapping (by whichever means) can be directed toward closing the remaining gaps.

References

Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M, Vogel C, Teichmann SA, Ivens A and Dear PH (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. *Genomics*, **1**, 1787–1799.

Cox DR, Burmeister M, Price ER, Kim S and Myers RM (1990) Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, **250**, 245–250.

Dear PH, Bankier AT and Piper MB (1998) A high-resolution metric HAPPY map of human chromosome 14. *Genomics*, **48**, 232–241.

Dear PH and Cook PR (1993) HAPPY mapping - linkage mapping using a physical analog of meiosis. *Nucleic Acids Research*, **21**, 13–20.

Samad A, Huff EF, Cai W and Schwartz DC (1995) Optical mapping: A novel, single-molecule approach to genomic analysis. *Genome Research*, **5**, 1–4.

Zhou S, Kvikstad E, Kile A, Severin J, Forrest D, Runnheim R, Churas C, Hickman JW, Mackenzie C, Choudhary M, *et al*. (2003) Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides* strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome Research*, **13**, 2142–2151.

Basic Techniques and Approaches

Repeatfinding

Richard Sucgang
Baylor College of Medicine, Houston, TX, USA

Genomes in general, and eukaryotic genomes in particular, are rife with segments of repetitive sequence. Repetition is the most obvious pattern found in genetic information, and is usually indicative of a biologically significant motif or landmark in that particular biomolecule. Repetitive segments of DNA appear to be necessary for the structural function of centromeres and telomeres (Louis, 2002; Schueler *et al*., 2001). Gene duplications also lead to repetitive motifs, such as those found in biologically important regions such as ribosomal RNA genes (Heilig *et al*., 2003) or the human Y chromosome (Repping *et al*., 2002), reflecting the evolutionary importance of these areas. Consequently, some of the earliest analytical tools developed involve detection, identification, and classification of repeat regions in protein and DNA sequences. With the availability of whole genomes, such repeat finding algorithms are an essential aspect of genome level analysis and annotation. In fact, repetitive regions often confound the assembly stage of genome sequencing, so the successful assembly of whole genomes is itself dependent on identifying and masking repeats.

Although identifying repeats in biological sequences would, on its face, appear to be a mathematically simple problem of searching for repeated substrings within a larger string, several factors confound the issue. Biological repeat segments are not necessarily exact repeats, and the periodicity of the pattern seldom matches an exact formula. Despite these fuzzy boundaries, biological repeats retain a physiological significance, and therefore, the algorithms written have accommodated these inexact specifications. In addition, comparisons of repeat motifs within and among whole genomes demand logarithmically increasing amounts of memory and processing power; most recent implementations of repeat finding software utilize more sophisticated searching algorithms such as suffix-tree mapping to address these issues (Delcher *et al*., 1999).

In general, repeat analysis software packages have focused on one or more of the following tasks: detection of repeat patterns (which can be subtle), de novo identification of basic repeat units (Bao and Eddy, 2002), clustering and classification of known repeat units (Volfovsky *et al*., 2001), and curation of such repeats. In this section, we will take a survey of the different software implementations for repeat finding.

Detection of repeated segments in strings of characters is a difficult problem in computation, particularly in biology where the basic unit of repetition can vary not

only in sequence but also in size and displacement. The human eye, however, is adept at picking out patterns laid out in a two-dimensional figure; thus, one of the simplest and yet most effective means of detecting repeated patterns in sequences is the dot-matrix plot (Gibbs and McIntyre, 1970). At the simplest implementation of this plot, two sequences that are being compared are laid out along the two axes of a two-dimensional grid. For every character in common between the two strings, one dot is placed at the intersection of the two coordinates. While this approach can be used to visualize the commonalities between any two strings, when the same string is laid on both axes, repeated motifs become evident to the human eye (Figure 1). A central diagonal is always present as the sequence is compared against itself, direct repeats manifest as parallel lines to the central diagonal, and inverted repeats are present as perpendicular lines. The distance from the diagonal indicates displacement, and palindromic repeats cross the central diagonal. This approach has proved to be so effective in detecting repeat motifs that despite increasing sophistication in implementation, the basic algorithm has remained essentially unchanged.

The earliest implementation as a software package was described by Pustell and Kafatos (1982), and implementations are found in various software packages ranging from commercial to open source; one such example is dottup, a component of the EMBOSS package (Rice *et al*., 2000). Since a dot-matrix plot relies on the pattern recognition skills of the human eye, good interactivity and flexibility

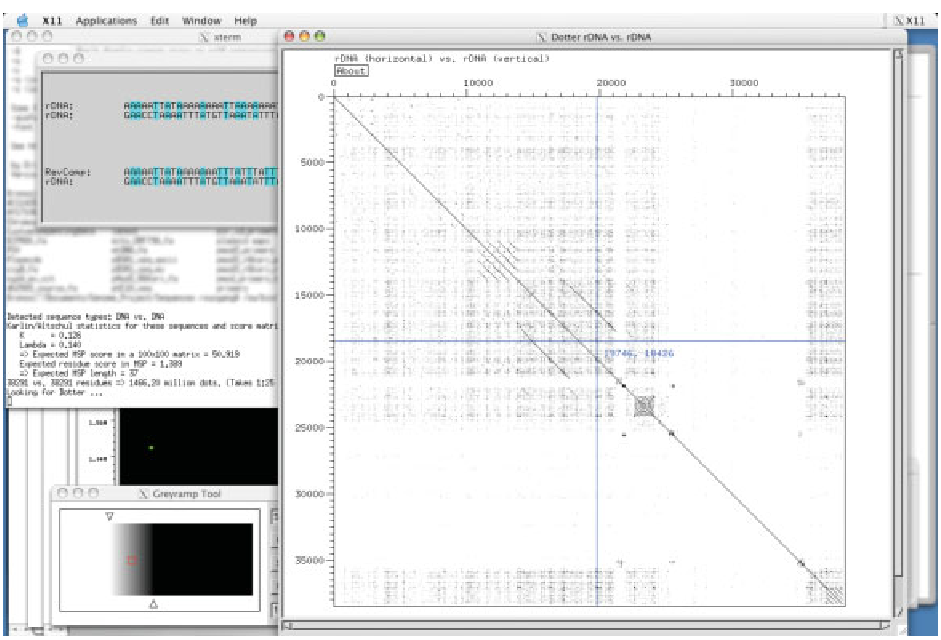


Figure 1 The dotter interface. Loaded here is one arm of the *Dictyostelium discoideum* extrachromosomal rDNA element, lines parallel to the central diagonal indicate direct repeats in the sequence. Clusters of shorter repeats are seen as multiple parallel lines. Note the greyramp tool at the lower left corner that can be used to set the threshold of the image to improve detection of shorter repeat motifs

with the manipulation of the dot-matrix image is key to a good implementation. While lines can denote significant repeat structures, nonspecific or low complexity matches will be displayed as clouds of dots that can occlude the visibility of significant sequence features. Perhaps the most sophisticated interface for exploring dot-matrix plots currently available is in the dotter package (Sonnhammer and Durbin, 1995). Relying on the X11 graphical user layer in Unix, the dotter package plots the images as a gray-scale plot, and provides a gray ramping tool that permits on the fly alterations to the visualization threshold. Dynamic zooming into segments, and the ability to export to high-resolution PostScript data completes the package.

While dotter theoretically presents no real limit to the amount of data that it can process, realistically, as the input data increases (such as that of a chromosome-sized piece of DNA), the dotter package slows down dramatically. Much of the performance bottleneck is in the matching algorithm used to align segments to the genome. Word-search-based algorithms that rely on preindexing a table of all possible word combinations in a sequence are far faster in traversing large sequences, and is the method implemented by the software package lbdot (Huang and Zhang, 2004). Although the interface is less polished for exploring the resulting plot, it is able to process large sequences at a fraction of the time and memory required by dotter.

Inherent to the problem of detecting repeats is the need to align and search whole genomes efficiently. The REPuter package efficiently searches for repeats and palindromes in genome scale sequences. Initially implemented just to detect the largest exact matches for the substrings, it has since been improved to permit the presence of degenerate repeats with allowable mismatches (Kurtz *et al*., 2001). It is available in limited form for Web-based interaction, but a licensed stand-alone version provides increased flexibility. An approach to aligning whole genomes that uses the suffix-tree algorithm to detect repeats is the MUMmer package from The Institute of Genome Research (TIGR) (Kurtz *et al*., 2004). Now at version 3, the MUMmer package is uniquely flexible in not only aligning any genome (or fragment thereof) with any other, but can base the alignment on the protein coding potential of the genome, detecting duplicated genes that may have diverged sufficiently on the nucleotide level to avoid detection with a direct dot-matrix plot (Figure 2). The MUMmer package is fast enough that an on-line version is available for arbitrary alignments between the different completed microbial genomes archived at the Comprehensive Microbial Resource at TIGR (Peterson *et al*., 2001). More recently, a new Repeat Analysis Program has become available that improves on these approaches by indexing gapped words, increasing the sensitivity of the word-searching method (Campagna *et al*., 2005).

Eukaryotic genomes carry not only a load of simple and tandem repeats but also a number of interspersed repeats, particularly those deriving from transposable elements and their by-products. Although simple repeats are easy to identify, repeated de novo identification of more complex repeats through word-finding can be impractical for every new genome that needs to be analyzed, or at the very least inefficient for genomes that are closely related to a previously analyzed one. In these circumstances, repeat finding can be accelerated by the use of a preexisting library of curated known repeat sequences.

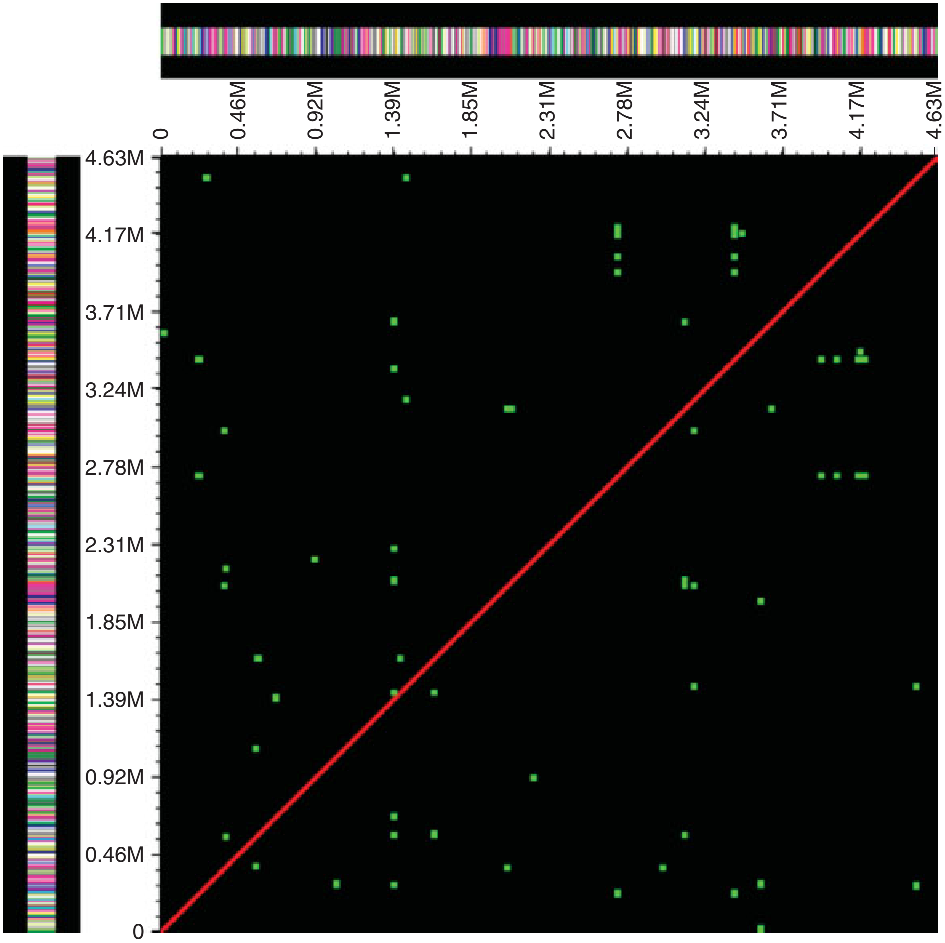


Figure 2 MUMmer output. This illustrates MUMmer displaying the *E. coli* genome against itself, highlighting the presence of short inverted repeats. For this display, the minimum sensitivity is 100 bp. A red dot indicates an alignment in the same direction, and a green dot for the opposite direction. The multicolored bars on each side show the gene density for the chromosome, each colored bar representing one gene

Although repeated sequence motifs can be indicative of biologically important regions of genomes, the sheer abundance of interspersed complex repetitive sequences can obscure annotation and analysis efforts, particularly if a particular element encodes potential coding sequences that can interfere with ab initio gene prediction software. Thus, masking out the repeats is an equally important subsequent step to repeat detection. The most common software package for this purpose is RepeatMasker, now an open source package housed by the Institute for Systems Biology (Smit *et al*., 1996–2004). At its core, RepeatMasker relies on a database of known repeat sequences, and the matching algorithm cross_match (Gordon *et al*., 1998). However, at the highest sensitivity settings, cross_match can be very slow with large sequences, leading to the development of

Table 1 Applications and websites described in this section

Application	URL
dotter	**http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html**
EMBOSS	**http://emboss.sourceforge.net/**
lbdot	**http://www.lynnon.com/dotplot/files.html**
MUMmer	**http://www.tigr.org/software/mummer/**
RepeatMasker/MaskerAid	**http://www.repeatmasker.org**
RepeatFinder	**ftp://ftp.tigr.org/pub/software/repeatFinder/**
Repbase/CENSOR	**http://www.girinst.org**
REPuter	**http://genomes.de**
Repeat Sequence Database	**http://rsdb.csie.ncu.edu.tw/**
CMR	**http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl**

the accelerator package MaskerAid (Bedell *et al*., 2000), which replaced the matching algorithm with the significantly faster WU-BLAST algorithm. The current implementation of RepeatMasker integrates the option to use the WU-BLAST method.

Databases of prototypic sequences from repetitive elements are key to the operation of such masking software. Although the on-line resource RepBase was originally crafted to curate prototypic sequences from the human genome, it has since expanded to accommodate most known eukaryotic repeats, and releases them in database form for use by programs such as RepeatMasker (Jurka *et al*., 1992). Managed by the nonprofit Genetic Information Research Institute, RepBase also electronically publishes a specialized electronic journal dedicated to studying eukaryotic repeats called RepBase Reports. The RepBase team also implements an alternative masking program called CENSOR (Jurka *et al*., 1996). Housed in Taiwan, the Repeat Sequence Database is another such database, originally designed to explore the relationship between repetitive sequence data and transcriptional activation sites (Horng *et al*., 2003).

As with most software implementations, the packages described in this section continue to evolve, and will no doubt be supplemented and supplanted by new software in the future. Table 1 summarizes the different URLs where the software is available, and all are either open source or available for free for academic use at the time of this writing.

References

Bao Z and Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, **12**, 1269–1276.

Bedell JA, Korf I and Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.

Campagna D, Romualdi C, Vitulo N, Del Favero M, Lexa M, Cannata N and Valle G (2005) RAP: a new computer program for de novo identification of repeated sequences in whole genomes. *Bioinformatics*, **21**(5), 582–588.

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O and Salzberg SL (1999) Alignment of whole genomes. *Nucleic Acids Research*, **27**, 2369–2376.

Gibbs AJ and McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European Journal of Biochemistry*, **16**, 1–11.

Gordon D, Abajian C and Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Heilig R, Eckenberg R, Petit JL, Fonknechten N, Da Silva C, Cattolico L, Levy M, Barbe V, de Berardinis V, Ureta-Vidal A, *et al*. (2003) The DNA sequence and analysis of human chromosome 14. *Nature*, **421**, 601–607.

Horng JT, Lin FM, Lin JH, Huang HD and Liu BJ (2003) Database of repetitive elements in complete genomes and data mining using transcription factor binding sites. *IEEE Transactions on Information Technology in Biomedicine*, **7**, 93–100.

Huang Y and Zhang L (2004) Rapid and sensitive dot-matrix methods for genome analysis. *Bioinformatics*, **20**, 460–466.

Jurka J, Klonowski P, Dagman V and Pelton P (1996) CENSOR–a program for identification and elimination of repetitive elements from DNA sequences. *Computers & Chemistry*, **20**, 119–121.

Jurka J, Walichiewicz J and Milosavljevic A (1992) Prototypic sequences for human repetitive DNA. *Journal of Molecular Evolution*, **35**, 286–291.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J and Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, **29**, 4633–4642.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biology*, **5**, R12.

Louis EJ (2002) Are Drosophila telomeres an exception or the rule? *Genome Biology*, **3**, REVIEWS0007.

Peterson JD, Umayam LA, Dickinson T, Hickey EK and White O (2001) The comprehensive microbial resource. *Nucleic Acids Research*, **29**, 123–125.

Pustell J and Kafatos FC (1982) A high speed, high capacity homology matrix: zooming through SV40 and polyoma. *Nucleic Acids Research*, **10**, 4765–4782.

Repping S, Skaletsky H, Lange J, Silber S, Van Der Veen F, Oates RD, Page DC and Rozen S (2002) Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *American Journal of Human Genetics*, **71**, 906–922.

Rice P, Longden I and Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K and Willard HF (2001) Genomic and genetic definition of a functional human centromere. *Science*, **294**, 109–115.

Smit A, Hubley R and Green P (1996–2004) RepeatMasker-Open 3.0.

Sonnhammer EL and Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.

Volfovsky N, Haas BJ and Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biology*, **2**, RESEARCH0027.

Basic Techniques and Approaches

Graphs and metrics

Andreas Dress
Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstrasse, Leipzig, Germany

1. Generalities

Graphs and metrics are structures that allow to specify and represent mutual relationship between pairs of objects from a given collection of objects under consideration. The most abstract form of representation consists in just a list specifying all pairs of objects that are considered to be "related" (relative to some preconceived concept of relatedness). Additional information regarding a "quantifiable degree of relatedness" can be specified by providing, in general for every pair x, y of objects, a nonnegative real number $S(x, y)$ measuring their *similarity* – or, conversely, a number $D(x, y)$ measuring their *dissimilarity* (or *distance*).

2. Basic definitions

1. In consequence, one defines a (simple undirected) *graph* G to be a pair (V, E) consisting of a set $V = V_G$ called the *vertex set* of G (representing the objects) and a subset $E = E_G$ of the set $\binom{V}{2}$ of 2-subsets of V called its *edge set*. The *degree* $d_G(v)$ of a vertex $v \in V$ is the number of elements in its *G-neighborhood* $N_G(v) := \{u \in V : \{v, u\} \in E\}$. If V is a finite set of cardinality $|V| =: n = n(G)$ – in which case G is called a finite graph – one has $0 \leq d_G(v) < n$ for every $v \in V$, $2|E| = \sum_{v \in V} d_G(v) = \sum_{i=0}^{n-1} i\, n_i$ with $n_i = n_i(G) := |V_i(G)|$, the cardinality of the set $V_i(G) := \{v \in V : d_G(v) = i\}$ of vertices of degree i, and $|V| = \sum_i n_i$ (which, by the way, implies that every finite graph contains at least two vertices of the same degree because $n_i \leq 1$ for all $i = 0, 1, \ldots, n-1$ would imply $n_i = 1$ for all $i = 0, 1, \ldots, n-1$, while $n_i = 0$ holds for all $i > n - (n_0 + 1)$). A *clique* of G is a subset C of V with $\binom{C}{2} \subseteq E$, an *independent* set of G is a subset I of V with $\binom{I}{2} \cap E = \emptyset$, a *G-path* of *length* ℓ from a vertex v to a vertex u is a sequence $v_0, v_1, \ldots, v_\ell$ of $\ell + 1$ vertices on V with $v_0 = v, v_\ell = u$, and $\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{\ell-1}, v_\ell\} \in E$. A *connected component* of G is a minimal nonempty subset C of V such that $u \in C$ holds for every $u \in V$ for which a G-path from u to some vertex $v \in C$ exists (or, equivalently, a maximal subset C of V such that a G-path from u to v exists for any two vertices $u, v \in G$). The vertex sets of the connected components of G form a *partition* of V, denoted

by π_G. G is connected if $|\pi_G| = 1$ (or, equivalently, $\pi_G = \{V\}$) holds. A *G-cycle* is a path $x_0, x_1, \ldots, x_n$ with $x_n = x_0$ and $|\{x_1, \ldots, x_n\}| = n > 2$. G is called a *tree* if it is connected and does not contain any G-cycle. One has $|V| \leq |E| + 1$ for any finite connected graph $G = (V, E)$, with equality holding iff (=if and only if) G is a tree, implying that $2\sum_{i>0} n_i = 2|V| = 2 + 2|E| = 2 + \sum_{i>0} i\, n_i$ and, therefore, $n_1 = 2 + \sum_{i>2}(i-2)n_i > 1$ and $2 + \sum_{i>1}(i-1)n_i = |V|$ hold for every finite tree with at least two vertices. More generally, the *cycle number* $C_1(G); := |E| + |\pi_G| - |V|$ is a non-negative integer for every finite graph G. For every connected graph $G = (V, E)$, there exists a *spanning tree*, that is, a tree with vertex set V whose edge set is a subset of E. Every finite tree (V, E) with $|V| \geq 2$ contains at least two *leaves*, that is, vertices of degree 1, and – in case $|V| \geq 5$ and $n_2 = O$ – at least two disjoint *cherries*, that is, pairs of distinct leaves u, v with $N_G(u) = N_G(v)$.

2. Given any set X, a map D from $X \times X := \{(x, y): x, y \in X\}$ into the set $\mathbb{R} \cup \{+\infty\}$ is called *symmetric* if $D(x, y) = D(y, x)$, a *similarity* if $D(x, x) \geq D(x, y)$, and a *dissimilarity* if $D(x, x) = 0 \leq D(x, y)$ holds for all $x, y \in X$. A dissimilarity is called a *metric* if $D(x, y) \leq D(x, z) + D(y, z)$ holds for all $x, y, z \in X$. The *diameter* $D(C)$ of a subset C of X relative to a dissimilarity D is defined by $D(C) := \sup(D(x, y) : x, y \in C)$. The D-rank $\mathrm{rk}_x^D(y)$ of an element $y \in X$ relative to some element $x \in X$ is defined by $\mathrm{rk}_x^D(y) := |\{z \in X : D(x, z) \leq D(x, y)\}|$ and the *D-rank* of a subset C of X by $\mathrm{rk}^D(C) := \max(\mathrm{rk}_x^D(y) : x, y \in C)$. One has $\mathrm{exc}^D(C) := \mathrm{rk}^D(C) - |C| \geq 0$ for every $C \subseteq X$, and $|C_1 \cup C_2| \leq \max\left(\mathrm{rk}^D(C_1), \mathrm{rk}^D(C_2)\right) \leq \max(|C_1|, |C_2|) + \max(\mathrm{exc}^D(C_1), \mathrm{exc}^D(C_2))$ for all $C_1, C_2 \subseteq X$ with $C_1 \cap C_2 \neq \emptyset$. In particular, one has $C_1 \cap C_2 = \emptyset$, $C_1 \subseteq C_2$, or $C_2 \subseteq C_1$ for all $C_1, C_2 \subseteq X$ with $\mathrm{exc}^D(C_1) = \mathrm{exc}^D(C_2) = 0$. Note that $\mathrm{rk}_x^{D_1}(y) = \mathrm{rk}_x^{D_2}(y)$ holds for all $x, y \in X$ for any two *homologous* dissimilarities D_1 and D_2, that is, with $D_1(x, z) \leq D_1(x, y) \Leftrightarrow D_2(x, z) \leq D(x, y)$ for all x, y, z in X.

Given n sets $X_1, \ldots, X_n$ with metrics $D_i : X_i \times X_i \to \mathbb{R}_{\geq 0}$ $(i = 1, \ldots, n)$, the product space $X_1 \times \cdots \times X_n$ consisting of all sequences $(x_1, \ldots, x_n)$ with $x_i \in X_i$ for all $i = 1 \ldots, n$, is a metric space relative to the metric $D := D_1 \times \cdots \times D_n$ defined by $D((x_1, \ldots, x_n), (x\prime_1, \ldots, x\prime_n)) := \sum_{i=1}^{n} D_i(x_i, x\prime_i)$ for any two sequences $(x_1, \ldots, x_n)$, $(x'_1, \ldots, x'_n)$ in $X_1 \times \cdots \times X_n$.

3. From graphs to metrics

Given a graph $G = (V, E)$ together with a *weighting scheme* $w : E \to \mathbb{R}_{>0}$, there exists a unique largest metric $D = D_{(G,w)} : V \times V \to \mathbb{R}_{\geq 0} U\{\infty\}$ with $D(u, v) \leq w(\{u, v\}$ for all $u, v \in V$ with $\{u, v\} \in E$. In case $w(e) = 1$ holds for all $e \in E$, one also writes D_G instead of $D_{(G,w)}$. Clearly, one has (1) $D_{(G,w)}(x, y) = +\infty$ for some $x, y \in V$ (and every weighting scheme w) iff no G-path from x to y exists, (2) the diameter $D_G(C)$ of a subset C of V equals 1 iff C is a clique, (3) $\{u, v\} \in E$ for any two elements $u, v \in V$ iff $D_G(u, v) = 1$ holds and, provided G is finite, also (4) $D_{(G,w)}(x, y) + D_{(G,w)}(u, v) \leq \max(D_{(G,w)}(x, u) + D_{(G,w)}(y, v), D_{(G,w)}(x, v) + D_{(G,w)}(y, u))$ for all $x, y, u, v \in V$ for some (or all) weighting scheme(s) w iff G is

a tree in which case a map $f : V_1 \to \mathbb{R}$ is of the form $f = f_v : V_1 \to \mathbb{R} : x \mapsto D_{(G,w)}(x, v)$ for some vertex $v \in V$ of degree $k \geq 3$ iff $f(x) = \max(D_{(G,w)}(x, y) - f(y){:}y \in V_1)$ holds for all $x \in V_1$ and a clique of cardinality k exists in the graph $G_f{:=}(V_1, E_F)$ with $E_f := \{\{x, y\} \in \binom{V_1}{2} : f(x) + f(y) = D_{(G,w)}(x, y)\}$; so, we can recover V (and then also E and w) from the restriction of $D_{(G,w)}$ to V_1 in case $n_2 = 0$ holds.

4. From metrics to graphs

Given a dissimilarity $D : X \times X :\to \mathbb{R}_{\geq 0}$ defined on a finite set X and a number $r \in \mathbb{R}$, put $E(r) := E_D(r) := \{\{x, y\} \in \binom{X}{2} : D(x, y) \leq r\}$ and $G(r) = G_D(r){:=} (X, E(r))$. As $r_1 \leq r_2$ implies $E(r_1){\subseteq}E(r_2)$, every connected component of $G(r_2)$ is a disjoint union of connected components of $G(r_1)$ in this case, implying that the graph $\mathcal{C}(\mathcal{D})$ with vertex set $C(D) := \bigcup_{r\geq 0} \pi_{G(r)}$ and edge set all subsets $\{U_1, U_2\}$ of $C(D)$ with $|\{U \in C(D){:}U_1 \subsetneq U \subseteq U_2\}| = 2$ is a tree. Similarly, the graph $\mathcal{R}(\mathcal{D})$ with vertex set $A(D){:=}\{C{\subseteq}X{:}\mathrm{exc}^D(C) = 0\}$ and edge set all subsets $\{C_1, C_2\}$ of $A(D)$ with $|\{C \in A(D){:}C_1 \subsetneq C \subseteq C_2\}| = 2$ is also a tree. If X is a set of species and D a measure of their "genetic dissimilarity", these trees can be viewed as putative phylogenetic trees for X.

Of interest is also the graph with vertex set X and edge set $E(D) := \{\{x, y\} \in v\binom{X}{2} : D(x, y) < D(x, z) + D(z, y)$ for all $z \in X - \{x, y\}\}$. If D is a metric with $D(x, y) > 0$ for any two distinct points $x, y \in X$, then $D = D_{(G,w)}$ holds for the weighting scheme $w : E(D) \to \mathbb{R} : \{x, y\} \mapsto D(x, y)$, and $E(D){\subseteq}E$ holds for the edge set E of any graph $G = (X, E)$ for which a weighting scheme w with $D = D_{(G,w)}$ exists.

5. An example: graphs in sequence assembly

In *whole-shotgun sequencing*, sequence assembly can be performed by studying the *overlap graph* of a given collection of subsequences identified, for example, by PCR procedures: Given a finite set A of "letters", a string $\alpha := a_1a_2 \ldots a_N$ of length N, a positive integer $k < N$, and a collection $\mathcal{C}$ of subsequences of length k of the form $a_{i+1}a_{i+2} \ldots a_{i+k}$ of α, the problem is to reconstruct α from the subsequences in $\mathcal{C}$. While it is obvious that this is not always possible (e.g., the sequences *abcabab* and *ababcab* have the same set of subsequences of length 3 – so this is not even necessarily possible if $\mathcal{C}$ consists of *all* subsequences of length k *and* if all of these are distinct), one can still try to find at least one, or even all, strings α that would be *compatible* with the given collection $\mathcal{C}$ of k-strings. To this end, one can proceed as follows: One considers the associated *overlap graph* $G = G(\mathcal{C})$ whose vertices are the strings in $\mathcal{C}$ and whose edges are the pairs $a_1 \ldots a_k$ and $b_1 \ldots b_k$ of strings in $\mathcal{C}$ for which $a_2 \ldots a_k = b_1 \ldots {}_{k-1}$ holds (note that these edges come with a *preferred* direction, that is, that from $a_1 \ldots a_k$ to $b_1 \ldots b_k$). The task is to find a shortest path in that graph that respects directions and visits every vertex at least once. This is a typical problem in the important

area of *graph algorithms*, which, though unfortunately much too extensive to be covered in this article, still relies strongly on the concepts collected above.

6. Discussion

The above concepts and facts provide a highly flexible framework for dealing with complex data (in particular, discrete data with large degrees of freedom) such as collections of genetically related sequences, interaction patterns of biomolecules, and/or biochip data. Genomics, proteomics, toponomics, and metabolomics all rely on comparative data analysis and, hence, on the analysis of relationships encoded by coincidence or, at least, (dis)similarity. References could easily fill many pages. However, endowed with a solid understanding of the concepts and constructions presented above, most of their applications within these contexts should be sufficiently transparent or, at least, accessible.

Algorithms for sequence errors

Björn Andersson
Karolinska Institutet, Stockholm, Sweden

Martti T. Tammi
Karolinska Institutet, Stockholm, Sweden
National University of Singapore, Singapore

The current state-of-the-art sequencing relies on the Sanger dideoxy chain termination method (Sanger *et al*., 1977). The method is based on a DNA synthesis process controlled by adding labeled dideoxy terminators, ddNTPs, into a polymerase reaction. DNA polymerases copy single-stranded DNA templates by incorporating nucleotides at the 3′ end of a primer annealed to the template DNA. The polymerase grows a new DNA strand by adding a nucleotide to the 3′-hydroxyl group at the growing end. The Sanger method exploits the fact that DNA polymerases are able to incorporate nucleotide analogs, such as dideoxynucleotides, ddNTPs, and fluorescently labeled terminators. ddNTPs are nucleotides that lack the 3′-hydroxyl group, while the label of the terminators block the 3′-position. By adding terminators, along with ordinary dNTPs, to the reaction, the elongation process is terminated once a terminator is incorporated. Since each terminator carries a different fluorescently colored label, we will be able to tell at which nucleotide the elongation was terminated. Several rounds of elongation will generate fragments of all sizes. These fragments are subsequently size-ordered by electrophoresis, and the nucleotide sequence can be read by detecting fluorescence after the labels are excited by a laser beam. This analysis is routinely performed in automatic sequencers, which usually utilize capillary electrophoresis.

Automatic sequencers produce so-called electropherograms or chromatograms (Figure 1), which are analyzed by specialized base-calling software (Ewing *et al*., 1998; Andrade and Manolakos, 2003; ABI, 1996) to determine the nucleotide sequence of the original template. The peak height is related to the number of fragments, which have the same length and end with the same nucleotide. The x-axis in a chromatogram is the time when the subpopulation has arrived to the detector of the sequencer. Depending on the terminators used, a specific mobility shift correction may be necessary, due to differences in migration caused by the different labels.

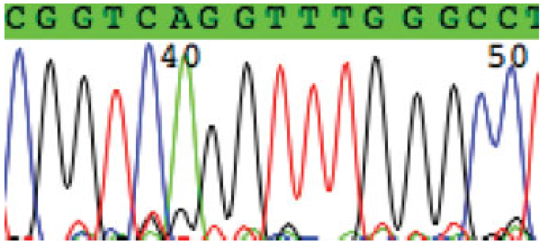


Figure 1 A section of a chromatogram produced by an automatic DNA sequencer. There are four different traces: Red for T, black for G, green for A, and blue for C. The x-axis is the time line when a subpopulation passes the detection point. The peak height is related to the number of fragments passing the detection point

Ideally, peaks in chromatograms for all four traces, one for each base, would be Gaussian-shaped and distributed evenly with no background noise. In real life, many factors perturb the signal detection, which make the base-calling procedure challenging and occasionally error prone. Some of the most important factors are:

1. In the beginning of the chromatogram, large and noisy peaks appear owing to unincorporated dyes, usually roughly the first 10 to 30 bases (Figure 2).
2. Toward the end of the chromatogram, peaks will be less evenly spaced and lumpy caused by diffusion of the fragments in the gel and by the loss of resolution due to the decrease in the difference of the relative mass of the fragments (Figure 2).

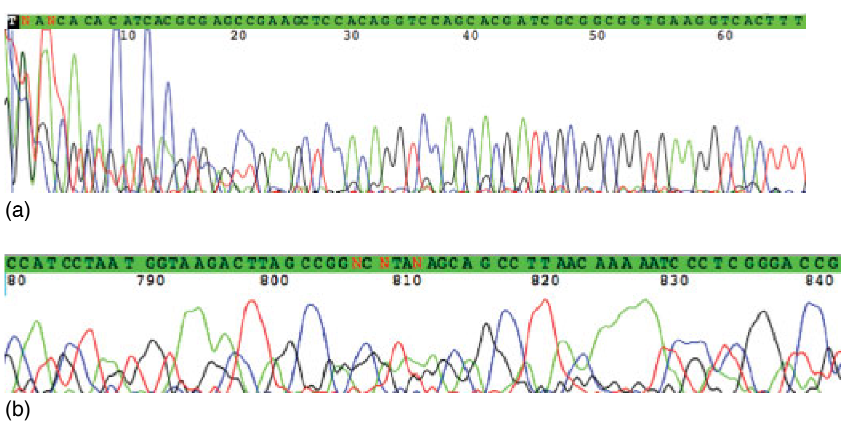


Figure 2 In the beginning of the chromatogram, large and noisy peaks appear because of unincorporated dyes. The first 66 positions are shown in (a). The peak height gets smaller toward the end of the chromatogram due to decreasing quantity of ddNTPs. The peaks may also be lumpy and unevenly spaced, which is caused by diffusion and the loss of resolution. Toward the end, the relative difference of the mass of the fragments decreases. The last 780–841 positions are shown in (b)

3. Toward the end of the chromatogram, the peak height gets smaller due to decreasing quantity of terminators relative to dNTPs.
4. The peaks may be compressed as a result of the formation of secondary structures in DNA fragments, such as hairpins. This is particularly common in GC-rich sequences.
5. Primer-dimer formation may cause false stops.
6. Uneven quantity of fragments may in some cases cause problems (ABI, 1996).

Thus, a base may be erroneously called, uncalled, or missed. These errors result in substitutions, deletions, and insertions of nucleotides respectively. Several attempts have been made to improve the quality of base calling (Ewing *et al*., 1998; Andrade and Manolakos, 2003; Davies *et al*., 1999; Brady *et al*., 2000; Giddings *et al*., 1998). It is, however, unlikely that the sequence data will ever become error free. Thus, it is necessary to try to improve the data quality using downstream methods. The latter include redundant sequencing, where the same region is sequenced several times, estimation of error probabilities, and error-correction methods.

In the shotgun sequencing method, the same genomic region is routinely sampled 5 to 10 times. This will increase the reliability of the data, since each base is sequenced several times and a so-called consensus sequence may be computed from the multiple alignments of the redundant sequences (Figure 3).

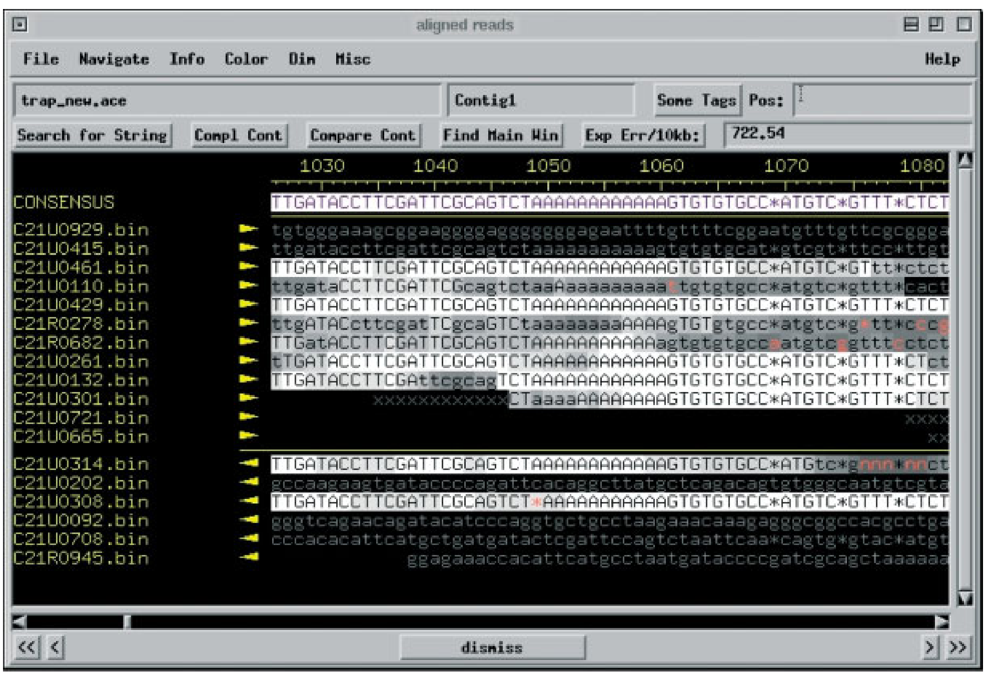


Figure 3 In a multiple alignment, a consensus sequence may be computed. A section of an alignment is shown in the Consed editor (Gordon *et al*., 1998). The numbers at the top are contig positions, followed by the consensus sequence and the reads. On the left are the read names, and arrows show the direction of the reads. The line in the middle separates the two DNA strands

Estimation of error probabilities of called bases is a powerful method to cope with sequencing errors. Many early base-calling algorithms estimated confidence values (Giddings *et al*., 1993; Giddings *et al*., 1998; Golden *et al*., 1993). Lawrence and Solovyev (1994) performed an extensive study and defined a large number of chromatogram parameters and analyzed them to determine which are most effective in distinguishing accurate base-calls. The base-calling program Phred (Ewing *et al*., 1998) is widely used today because of its ability to estimate the probability of an error in a base-call. Phred quality values have become the de facto standard within the sequencing community (Richerich, 1998). The quality values are logarithmic and range from 0 to 99. They are defined as follows:

$$q = -10.1\text{g}P \tag{1}$$

where q is the quality of the base-call and P is the estimated error probability of the call. The range of the Phred quality values is wide. For example, a quality value 10 means that in average one out of 10 calls is wrong and a value of 60 means that in average one out of a million base-calls is in error. In a typical sequence read, the quality is usually lowest toward the end of the reads as described above.

By characterizing signal traces in chromatograms and assigning parameter values to base-calls in sequences from an already known nucleotide sequence, it is possible to compare the output from a base-calling program to a benchmark data set. The relation to the frequency of base-calling errors can now be related to the parameter values calculated for each base-call. This can be used to estimate error probabilities in an unknown sequence by comparing parameter values of the known error frequencies. Phred uses a lookup table containing 2011 lines to set quality values. However, Phred does not provide any probabilities of missing base-calls or probabilities for any of the three other uncalled bases.

A widely used fragment assembly program, Phrap (**http://www.phrap.org**, 2005) utilizes Phred quality values. Phrap does not perform error correction, but uses a statistical framework to determine read overlaps. The quality values are also used in the computation of the consensus sequence, in most cases resulting in a more reliable sequence. Phrap assigns error probabilities to the consensus sequence itself, which helps to pinpoint problematic areas in assemblies.

The key problem in fragment assembly is the correct assembly of reads sampling repeated regions. The original genome sequence is reconstructed from short, redundantly sampled reads by using the criteria of similarity in the sequence. Similar reads are aligned together into multiple alignments, which form contiguous sequences, so-called contigs. In general, genomic sequences contain many repeats that may be identical or almost identical, and it is extremely difficult to distinguish the differences between reads originating from different repeated regions from sequencing errors. This is especially true when the difference between repeat copies is below 3%. Note that the error rate in sequencing varies and may be up to 11%. However, reads are normally trimmed at the ends to reduce errors, and an average remaining sequencing error in trimmed reads is somewhere around 1–2%. Error correction would reduce this problem.

Currently, there are three software tools that perform error correction on shotgun sequences: MisEd (Tammi *et al*., 2003), EULER (Pevzner *et al*., 2001), and

AutoEditor (Gajer *et al*., 2004), each of which use a different approach to the problem.

MisEd creates multiple alignments of shotgun reads. In this approach, the errors are corrected by first detecting differences between repeated reads using a statistical approach described by Tammi *et al*. (2002). If the multiple alignments are properly constructed, all the remaining differences must be sequencing errors. MisEd corrects these simply by converting the remaining differences to agree with the consensus base. In alignments that solely consist of nonrepeated or unique reads, all observed differences should be errors, since these reads sample the same genomic region. In the shotgun sequencing approach, the same genomic regions are generally sampled 5 to 10 times in average, and it is fairly easy to assemble reads originating from nonrepeated regions of the genome.

Patterns of differences between reads can be observed by creating multiple alignments of all similar reads in a shotgun project. An example of such an alignment is shown in the Figure 4, in which reads from different, very similar genomic repeats are aligned. Differences between reads caused by sequencing errors appear random, but a systematic appearance of differences between reads originating from separate repeat copies can be observed. In fact, the appearance of sequencing errors can be approximated by the Poisson process. This phenomenon is utilized in the MisEd program, which computes the probability of a random appearance of such patterns. If the estimated probability of observing a pattern of differences by random sequencing errors is very low, below a certain threshold, the differences are considered to be true differences. These differences are labeled as Defined Nucleotide Positions (DNPs). Errors are corrected by converting all, but DNPs into the same base as the consensus base in the alignment.

EULER or EULER assembler is a shotgun assembly program, which contains an error-correction step. In this approach, the reads are chopped into small pieces, roughly 40-bases long, and so-called spectral alignments using these short fragments are constructed (Figure 5). These alignments of short fragments display a similar pattern of differences between fragments as multiple alignments constructed of whole-shotgun reads. Sequence errors appear sporadically and real differences between repeats appear more systematically, that is, more frequently in a column of the alignment. The count of short fragments aligned on each read is analyzed

```
     ATGTAGGCATCTCATGACACGATCGACGGT
     ATGTAGGCATCTCATGACACGATCGACGGT
R1   ATGTAGGCATCTCATGATACGATCGACGGG
     ATGTAGGCATCTCATGACACGATCGACGGT
     ATGTAAGCATCTGATGACACGATCGATGAT
     ATGTAAGCATCTGATGACACGATCGATGAT
R2   ATGAAAGCATCTAATGACACGATCGATGAT
     ATGTAAGCATCTGATGACACAATCGATGAT
     ATGTAAGCATCTGATGACACGATCGATGAT
     ATGTAAGCATCTGATGACGCGATCGACTAT
R3   ATGTAAGCATCTGACGACGCGGTCGACGAT
     ATTGAAGCATCTCATGACGCGATCAACGAT
```

Figure 4 An example of an alignment of reads originating from repeat copies located in different genomic regions. The systematic appearance of differences can be observed for the real differences between repeat copies. The appearance of sequencing errors is sporadic. The repeat copies are marked as R1, R2, and R3

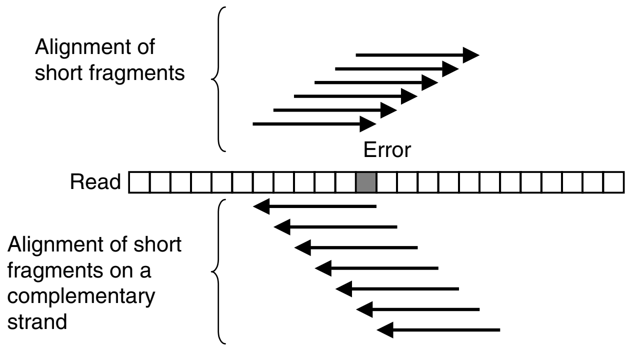


Figure 5 A schematic illustration of a spectral alignment of short fragments of reads. An error in a read affects the number of short read fragments, which can be aligned exactly on a read and its complementary strand

and erroneous positions determined by the number of aligned short fragments (Figure 5). The EULER program does not rely on any statistical framework. Certain parameters are used to set the lower threshold of the number of differences accepted as sequencing errors. All bases below this limit are considered to be caused by sequencing errors and converted to the consensus base. The parameters are user definable.

AutoEditor combines an approach of realigning shotgun reads to the consensus sequence and reexamination of the chromatogram data produced by sequencers. By recalling suspicious base-calls, AutoEditor is effectively able to correct errors. However, surprisingly, this program does not yield better results than the two previously described methods, even though chromatogram data is reexamined.

Sequencing errors are the root of the challenges in genome sequencing, particularly in combination with repeated genomic regions. They cause many errors in genome assemblies, and finishing of the projects requires a lot of manual labor. Perhaps we will see many automated finishing tools utilizing some kind of error correction/analysis methods in the near future.

Related articles

Article 25, **Genome assembly**, Volume 3, Article 2, **Algorithmic challenges in mammalian whole-genome assembly**, Volume 7 and Article 7, **Errors in sequence assembly and corrections**, Volume 7

References

ABI (1996) *ABI PRISM, DNA Sequencing Analysis Software, User's Manual*, PE Applied Biosystems: Foster City.

Andrade L and Manolakos ES (2003) Signal background estimation and baseline correction algorithms for accurate DNA sequencing. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, **35**(3), 229–243.

Brady D, Kocic M, Miller A and Karger B (2000) Maximum likelihood Base-Calling for DNA sequencing. *IEEE Transactions on Bio-medical Engineering*, **47**(9), 1271–1280.

Davies SW, Eizenman M and Pasuphaty S (1999) Optimal structurefor automating processing of DNA sequences. *IEEE Transactions on Bio-medical Engineering*, **46**(9), 1044–1056.

Ewing B, Hillier L, Wendl M and Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Gajer P, Schatz M and Salzberg SL (2004) Automated correction of genome sequence errors. *Nucleic Acids Research*, **32**(2), 562–569.

Giddings M Jr, Haker RB and Smith LM (1993) An adaptive, object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Research*, **21**, 4530–4540.

Giddings MC, Severin J, Westphall M, Wu J and Smith LM (1998) A software system for data analysis in automated DNA sequencing. *Genome Research*, **8**, 644–665.

Golden J, Torgersen D and Tibbets C (1993) Pattern recognition for automated DNA sequencing, i: on-line signal conditioning and feature extraction for basecalling. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, Hunter L, Searls DB and Shavlik JW (Eds.), AAAI Press: Menlo Park.

Gordon D, Abajian C and Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**(3), 195–202.

http://www.phrap.org. (2005)

Lawrence C and Solovyev V (1994) Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acids Research*, **22**, 1272–1280.

Pevzner PA, Tang H and Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(17), 9748–9753.

Richerich P (1998) Estimation of errors in “raw” DNA sequence: a validation study. *Genome Research*, **8**, 251–259.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Tammi MT, Arner E, Britton T and Andersson B (2002) Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. *Bioinformatics*, **18**(3), 379–388.

Tammi MT, Arner E, Kindlund E and Andersson B (2003) Correcting errors in shotgun sequences. *Nucleic Acids Research*, **31**(15), 4663–4672.

Basic Techniques and Approaches

Polymorphism and sequence assembly

Brinda K. Rana, Douglas W. Smith and Nicholas J. Schork
University of California, San Diego, CA, USA

1. Introduction

Worldwide research efforts to characterize the human genome, such as the Human Genome Project (*see* Article 24, **The Human Genome Project**, Volume 3), the ENCODE project, and the International HapMap Project, along with advances in DNA sequencing technologies, have produced an enormous amount of DNA sequence information. This information is becoming available in public databases and will allow researchers to identify and characterize naturally occurring variations in the human DNA sequence across individuals. Such genetic variation, that is, differences in DNA sequence among a group of individuals or between populations, occurring with frequency $>1\%$ is known as genetic *polymorphism*. Sources of polymorphisms within DNA sequence include variable numbers of tandem repeats (VNTRs) such as microsatellite repeats and short tandem repeats (STRs), small insertions and deletions of sequence (in/dels), gene copy number variation (Sebat *et al*., 2004; Fredman *et al*., 2004), and single-nucleotide polymorphisms (SNPs) (Kwok *et al*., 1994). SNPs are the most abundant form of genetic polymorphisms found in the human population and have become a focal point in the study of the genetic basis of multifactorial diseases and traits (*see* Article 57, **Genetics of complex diseases: lessons from type 2 diabetes**, Volume 2), interindividual differences in response to therapeutic drugs ("pharmacogenomics"), and human evolutionary and population history (*see* Article 71, **SNPs and human history**, Volume 4). This article will outline sequencing-based molecular and bioinformatics approaches for the identification of SNPs in the human population. Although approaches discussed can be extended to study polymorphism in other organisms and/or other types of polymorphism, the sequencing-based identification and scoring of repeats and in/dels can be more challenging than for SNPs.

As noted, SNPs are the most simple and abundant form of a genetic polymorphism and occur along a stretch of DNA sequence when one of the four nucleotides that contain the base adenine(A), cytosine(C), thymine(T), or guanine(G) is replaced by one of the remaining three nucleotides. A SNP that occurs between the purines (A, G) or pyrimidines (C,T) is known as a transition and constitutes two-thirds

of all SNPs (Iida *et al*., 2001a,b,c; Freudenberg-Hua *et al*., 2003); a transversion occurs between a purine and a pyrimidine.

By 2001, data arising from the public Human Genome Project identified over one million SNPs averaging one SNP every 1331 bp in comparison of two human chromosomes drawn from an ethnically diverse sample of individuals (The International SNP Map Working Group, 2001). A combination of polymorphisms that are closely linked on a single chromosome and that are inherited together on a single maternal or paternal chromosome is termed a *haplotype*. SNPs have become instrumental in defining haplotype “blocks” that reflect stretches of genome encompassing multiple SNPs with alleles that appear to be inherited together in the human genome (The International HapMap Consortium, 2003; **http://www.hapmap.org**).

2. In silico SNP discovery

Numerous SNPs were initially detected by comparing sequences from independent sources of mRNA and expressed sequence tags (ESTs) that had either been deposited in the public domain (dbEST; Boguski *et al*., 1993) or generated by large sequencing projects (Marth *et al*., 1999; Irizarry *et al*., 2000; Useche *et al*., 2001). Over 48 000 SNPs were identified through the use of purely computational and bioinformatic tools. This “in silico” mining of SNPs consist of simple comparisons of extant DNA sequence obtained from different sources, individuals, or investigators. These SNPs have been integrated into public databases such as dbSNP at the National Center for Biotechnology Information (NCBI). Owing to the potential errors in sequence information, either introduced by Taq polymerase during PCR or computational errors associated with base calling algorithms necessary for interpreting the results of sequence assays, in silico–defined SNPs should be validated through direct sequencing to confirm their existence in the human population. NCBI’s dbSNP catalogs over 900 000 SNPs (Sherry *et al*., 2001) as well as small in/dels, and retroposable element insertions; other SNP databases contain SNPs specific to particular ethnic populations (e.g., JSNP), while others catalog SNPs involved in diseases (e.g., HGVSNP) (Table 1). SNP databases are a good resource for investigators to verify SNPs that they have identified in their own sequencing studies, for choosing loci for further sequencing studies, and

Table 1 SNP databases

Database	URL
NCBI dbSNP	**http://www.ncbi.nlm.nih.gov/SNP/**
The SNP Consortium LTD	**http://snp.cshl.org/**
The International HapMap Project	**http://www.hapmap.org**
Japanese SNP database	**http://snp.ims.u-tokyo.ac.jp/**
SNPer at CHIP Bioinformatics	**http://snpper.chip.org**
Utah’s SNP database	**http://www.genome.utah.edu/genesnps/**
Whitehead Institute	**http://www-genome.wi.mit.edu/snp/human/**
Karolinska’s HGVBase	**http://hgvbase.cgb.ki.se/**
NHLBI Programs for Genomic Applications	**http://pga.lbl.gov/PGA/PGA_inventory.html**
Seattle SNPs	**http://pga.gs.washington.edu/**

for selecting SNPs to genotype for association studies. However, the investigator should be aware that SNPs contained in these databases have been obtained through various techniques, including in silico approaches, and a number of these SNPs remain unvalidated while others may have been missed because of the technique used or the population studied. Recent studies have investigated the quality of SNP databases (Jiang *et al*., 2003; Reich *et al*., 2003; Mitchell *et al*., 2004).

3. Targeted polymorphism discovery at candidate loci

Computational mining of SNPs from available DNA sequence databases has led to the identification of only a fraction of the SNPs' existing in the population; rare SNPs having low allele frequencies, or SNPs that are unique to a specific population ("population specific SNPs") are less likely to be found by these approaches (Kruglyak and Nickerson, 2001). Direct sequencing of a population of individuals followed by a comparison of the resulting sequences through multiple sequence alignment procedures is the most reliable of methods for identifying SNPs and other polymorphisms.

Fluorescence-based sequencing has become an important tool for such polymorphism discovery studies (see Griffiths *et al*., 1996, for a description of available sequencing methods). Figure 1 summarizes the steps in sequence-based polymorphism discovery following the selection of samples for sequencing: Polymerase chain reaction (PCR) amplification of locus in samples; sequencing of PCR products from forward and reverse directions; base calling, contig assembly, and polymorphism detection by sequence analysis software; and finally visualizing aligned sequence chromatograms for polymorphism identification and heterozygous base calling.

3.1. Samples

The choice of DNA samples depends on the ethnic population of interest, trait, or disease being studied, and the power desired to detect a polymorphism of a certain frequency. For example, sequencing 50 individuals (100 chromosomes) will enable the identification of SNPs that are at least 1% in that population with high (>80%) confidence. The frequencies of a number of SNPs are known to vary in different populations. Thus, limiting studies to a single population will reduce the chance of identifying SNPs common in other populations that are frequent or rare in the studied population. Therefore, to identify disease-associated variants, sequencing DNA from a pool of affected individuals is ideal, and ethnic background should be considered (*see* Article 75, **Avoiding stratification in association studies**, Volume 4).

However, to maximize the probability of finding a variant in the general human population, ethnic diversity panels is the best strategy. To facilitate the discovery of genetic variants in the entire human population, the National Human Genome Research Institute (NHGRI) of NIH, in conjunction with the Centers for Disease Control and Prevention, the National Institute of Environmental Health Sciences, and individual investigators, has assembled a DNA Polymorphism Discovery Resource consisting of DNA samples from 450 unrelated individuals from the United States, with ancestry from major regions of the world (Collins *et al*., 1998).

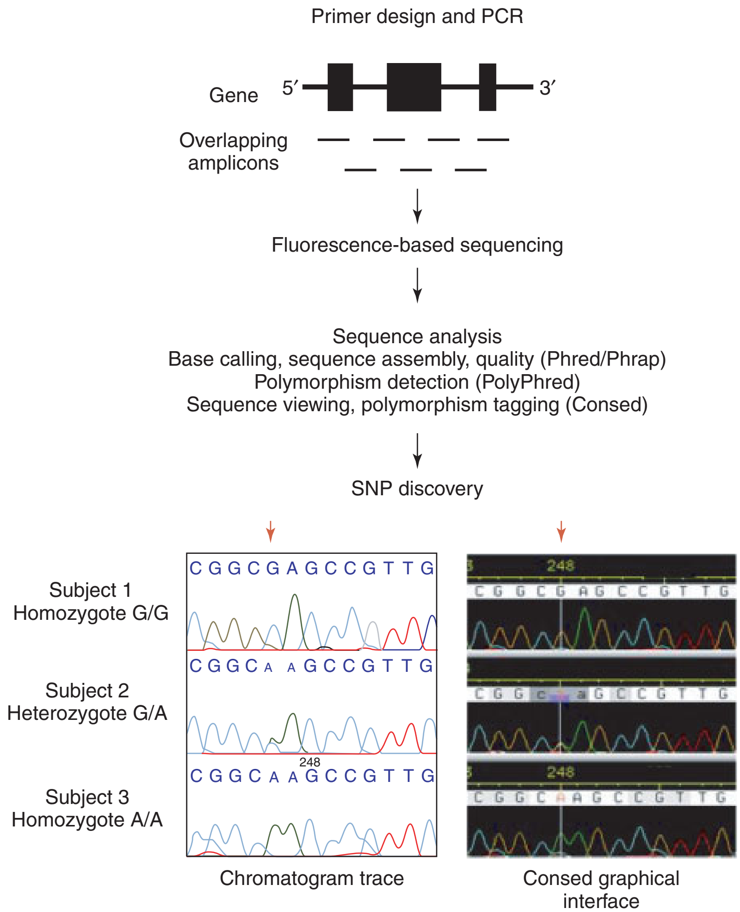


Figure 1 Targeted polymorphism discovery at candidate loci. PCR primers are designed to generate overlapping fragments (amplicons) of DNA to accommodate sequencing the entire gene of interest. Fluorescence-based sequencing generates chromatograms that are analyzed through the software package of choice to facilitate polymorphism discovery. Consed tags an SNP at position 248. A red arrow above both chromatogram views indicates the location of the discovered SNP. Subjects 1 and 3 are homozygous for nucleotide G and A, respectively. Subject 2 is determined to be heterozygote based on overlapping G and A traces that have approximately one-half the amplitude of either homozygote peak

These samples, as well as the NIH Diversity Panel, are available from the Coriell Institute for Medical Research (**http://coriell.umdnj.edu**) in collaboration with the National Institute of General Medical Sciences (**http://locus.umdnj.edu/nigms/**).

3.2. *PCR amplification of DNA samples and sequencing*

Forward and reverse sets of oligonucleotide primers for PCR amplification of a desired region of the genome are designed on the basis of gDNA sequence obtained, for example, from GenBank (NCBI). Messenger RNA sequence does not contain

intronic sequence found in gDNA, so care should be taken when designing PCR primers on the basis of these edited sequences for gDNA templates. Depending on the sequencing assay and instrument used, PCR primers should be designed to amplify approximately 600–1000-bp fragments. A popular software for optimizing primer design is Primer3 written at the Whitehead Institute (**http://www-genome.wi.mit.edu/cgi-gin/primer/primer3_ www.cgi**); PCR-Overlap© (**http://droog.gs.washington.edu/PCR-Overlap.html**; ©1998-1999 University of Washington, Rieder MJ and Nickerson DA) and The PCR Suite (**www.eur.nl/gff/kgen/primer/**; van Baren and Heutink, 2004) are free software packages that extend Primer3 by facilitating the design of PCR primers to generate overlapping fragments of DNA for loci that are too large to be accommodated by a single PCR fragment. Overlapping at least 100 bp on each end of the PCR fragment will reduce the chance of missing SNPs located in the primer region or on the ends of the PCR-amplified fragments (amplicons), which generally do not sequence well by fluorescence-based methods. Following the PCR amplification reaction, amplicons are purified and prepared for sequencing. The set of forward and reverse PCR primers for each amplicon can be used as sequencing primers to sequence from both ends of the amplicon. Alternatively, sequencing primers can be designed on the basis of other sequences contained in the amplicon.

3.3. Sequence analysis

Accurate and efficient sequence analysis for fluorescence-based sequencing requires software for base calling, sequence assembly, polymorphism detection, and sequence visualization. Two examples of reliable software packages for sequence analysis are MacVector (Rastogi, 2000) and the Phred/Phrap/Consed suite of sequence analysis software, the latter of which we present here in more detail. Phred is used to automate base calling including accuracy assessments of base calls (Ewing *et al*., 1998; Ewing and Green, 1998); Phrap assembles sequence fragments of multiple amplicons on the basis of a representative sequence of the locus; and Consed enables the visualization of sequence alignments as well as the original fluorescent graphical data (“chromatogram traces”) from which the base calls were made (Gordon *et al*., 1998). Polymorphisms called by sequence analysis software should also be reverified manually using a graphical interface such as Consed.

3.4. Heterozygous base calling

Since humans are diploid (have two copies of each chromosome – one inherited from the mother and one inherited from the father), a DNA sequence from gDNA template represents both copies of the chromosome. An individual is *heterozygous* at a polymorphic site when two different alleles (two different bases for a SNP) exist at that site. It can be difficult to accurately identify heterozygous sites because of the variability in fluorescence signals and inconsistency of base calling at these sites. The detection of heterozygous sites when comparing sequence traces of homozygotes with heterozygotes is based on the observation of a significant drop in peak height (to approximately one-half) along with the presence of a second,

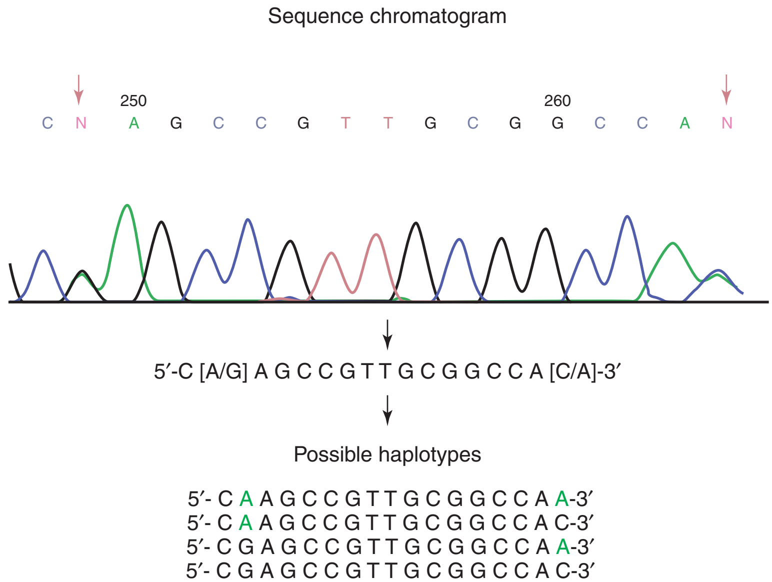


Figure 2 Complexity of haplotype determination in diploid organisms. A segment of continuous sequence from a human subject has two SNP sites, position 249 G/A and position 264 A/C. There are four possible combinations of two SNPs representing four possible haplotypes. All four haplotypes or only a subset may be represented in the human population. Further analysis is needed to determine the true haplotypes of this subject

overlapping peak in heterozygotes. PolyPhred is one of the most reliable software packages available for automated detection of heterozygous sites (Nickerson *et al*., 1997). Chromatogram traces from multiple individuals may be compared, for example, with Consed, to help assess difficult heterozygous calls.

3.5. *Verifying rare genetic variants*

The advantage of polymorphism discovery studies on a large number of samples is being able to identify rare variants in the population. However, genetic variants observed only once should be verified through reamplification from original gDNA template followed by sequencing to rule out the possibility of Taq polymerase induced or other sequence analysis errors.

3.6. *SNP haplotype assembly*

Because of the predominant diploid character of the human genome (see above), SNP identification in fragmented sequences introduces an algorithmic problem for assembly of multiple SNP genotypes into true haplotypes. If there are two or more SNPs in a fragment of genomic sequence, it is impossible to determine the haplotype just from that sequence (Figure 2). The most reliable method for delineating haplotypes is to clone a DNA fragment containing multiple SNP sites

Table 2 Computer programs for inferring and assigning haplotypes from genotype data

Haplotype programs	URL
EH: Infers haplotype frequencies on the basis of Expectation Maximum (EM) algorithm	**ftp://bioinformatics.weizmann.ac.il/pub/software/linkage_and_mapping/rockefeller/eh/**
Arlequin: Infers haplotypes using EM algorithm	**http://anthropologie.unige.ch/arlequin/**
HAPLOTYPER: Bayesian estimation of haplotypes	**http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm**
HFES: Computes the maximum likelihood estimate of up to five-locus haplotypes.	**http://www.cib.nig.ac.jp/dda/timanish/hfes.html**
SNPHAP: Infers haplotypes using EM algorithm	**http://www-gene.cimr.cam.ac.uk/clayton/software/**
SNPEM: Assesses differences in haplotype frequency between case and control subjects	**http://ww.pgx.ucsd.edu/PRL/snpem.txt**
Rhode's Program: Integrates nuclear family information into the EM algorithm to infer haplotypes	Available on request: rohde@mdc-berlin.de
Cooper and Wilson: Infers haplotypes from genotypes	**http://www.maths.abdn.ac.uk/~ijw/downloads/download.htm**
GENEHUNTER: Infers haplotypes from pedigrees	**http://linkage.rockefeller.edu/soft/gh/**
STATA PACKAGES: Chooses "haplotype tagging" SNPs	**http://www-gene.cimr.cam.ac.uk/clayton/software/**
Simwalk2: Infers haplotypes from any size pedigrees	**http://cmag.cit.nih.gov/lserver/SIMWALK2.html**
Haplo: Infers haplotypes by conditional probabilities	**http://watson.hgen.pitt.edu/register/docs/haplo.html**
Zaplo: Estimates haplotype frequencies of SNPs assuming no recombination	**http://watson.hgen.pitt.edu/register/docs/zaplo.html**

into a bacterial plasmid vector. After transformation of cloned vector into bacteria, a single colony representing the sequence from one chromosome can be chosen for sequencing. (see Griffiths *et al*., 1996, pp. 424–432). The sequence of the fragment representing the other chromosome can be accurately inferred on the basis of the sequenced chromosome fragment. When multiple individuals are studied, haplotyping each individual using the cloning method is time consuming and costly. In such cases, algorithms that statistically infer haplotypes may be a more efficient approach (Table 2). When the number of individuals is large or family data exists, such statistical inference can be reasonably accurate (Salem *et al*., 2005).

4. Conclusion

Polymorphism discovery relies heavily on sequencing-based methods of detection as presented here, however, other nonsequencing-based alternatives have been used. For example, the denaturing high-performance liquid chromatography (dHPLC) method, which utilizes the different mobility of homo- versus heteroduplexes of PCR amplicons during liquid chromatography under partial denaturing conditions, has successfully been employed for SNP discovery (Han *et al*., 2004). Such nonsequence-based genotype detection methods must be followed up with sequencing to determine the actual sequence change of the variant. Once SNPs and other polymorphisms have been discovered, one of the alternate high- or moderate-throughput strategies for genotyping large numbers of individuals can be employed as these will be more efficient than direct sequencing when the investigator is interested in typing subjects at single polymorphic sites (reviewed in Kwok, 2001; Gut, 2001).

Further reading

Sequencing Strategies and Maps:

Griffiths AJF, Wessler SR, Lewontin RC, Gelbart WM, Suzuki DT, Miller JH, Fixsen W and Lavett DK (2004) *An Introduction to Genetic Analysis*, *Eighth Edition*, WH Freeman and Company: New York.

SNP Detection and Genotyping Methods:

Kwok PY and Chen X (2003) Detection of single nucleotide polymorphisms. *Current Issues in Molecular Biology*, **5**, 43–60.

Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, **2**(12), 930–942. Review

Haplotype Maps using SNPs for Disease Association Studies:

Belmont JW and Gibbs RA (2004) Genome-wide linkage disequilibrium and haplotype maps. *American Journal of Pharmacogenomics*, **4**, 253–262.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA and Dudbridge F (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**, 233–237.

References

Boguski MS, Lowe TM and Tolstoshev CM (1993) dbEST–database for expressed sequence tags. *Nature Genetics*, **4**(4), 332–333.

Collins FS, Brooks LD and Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, **8**, 1229–1231.

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl MC and Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT and Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, **36**(8), 861–866.

Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P and Nothen MM (2003) Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. *Genome Research*, **13**, 2271–2276.

Gordon D, Abajian C and Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Griffiths AJF, Wessler SR, Lewontin RC, Gelbart WM, Suzuki DT, Miller JH, Fixsen W and Lavett DK (1996) DNA sequence determination. *An Introduction to Genetic Analysis*, *Sixth Edition*, WH Freeman and Company: New York, pp. 444–448.

Gut IG (2001) Automation in genotyping of single nucleotide polymorphisms. *Human Mutation*, **17**, 475–492.

Han W, Yip SP, Wang J and Yap MK (2004) Using denaturing HPLC for SNP discovery and genotyping, and establishing the linkage disequilibrium pattern for the all-trans-retinol dehydrogenase (RDH8) gene. *Journal of Human Genetics*, **49**(1), 16–23.

Iida A, Saito S, Sekine A, Kitamura Y, Kondo K, Mishima C, Osawa S, Harigae S and Nakamura Y (2001a) High-density single-nucleotide polymorphism (SNP) map of the 150-kb region corresponding to the human ATP-binding cassette transporter A1 (ABCA1) gene. *Journal of Human Genetics*, **46**, 522–528.

Iida A, Saito S, Sekine A, Kitamoto T, Kitamura Y, Mishima C, Osawa S, Kondo K, Harigae S and Nakamura Y (2001b) Catalog of 434 single-nucleotide polymorphisms (SNPs) in genes of the alcohol dehydrogenase, glutathione S-transferase, and nicotinamide adenine dinucleotide, reduced (NADH) ubiquinone oxidoreductase families. *Journal of Human Genetics*, **46**, 385–407.

Iida A, Sekine A, Saito S, Kitamura Y, Kitamoto T, Osawa S, Mishima C and Nakamura Y (2001c) Catalog of 320 single nucleotide polymorphisms (SNPs) in 20 quinone oxidoreductase and sulfotransferase genes. *Journal of Human Genetics*, **46**, 225–240.

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W and Lee CJ (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genetics*, **26**, 233–236.

Jiang R, Duan J, Windemuth A, Stephens JC, Judson R and Xu C (2003) Genome-wide evaluation of the public SNP databases. *Pharmacogenomics*, **4**, 779–789.

Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nature Genetics*, **27**, 234–236.

Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics Human Genetics*, **2**, 235–258.

Kwok PY, Carlson C, Yager TD, Ankener W and Nickerson DA (1994) Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics*, **23**(1), 138–144.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY and Gish WR (1999) A general approach to single nucleotide polymorphism discovery. *Nature Genetics*, **23**, 452–456.

Mitchell AA, Zwick ME, Chakravarti A and Cutler DJ (2004) Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics*, **20**, 1022–1032.

Nickerson DA, Tobe VO and Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.

Rastogi PA (2000) MacVector. Integrated sequence analysis for the Macintosh. *Methods in Molecular Biology*, **132**, 47–69.

Reich DE, Gabriel SB and Altshuler D (2003) Quality and completeness of SNP databases. *Nature Genetics*, **33**(4), 457–458.

Salem RM, Wessel J and Schork NJ (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics*, **2**(1), 39–66.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**(5683), 525–528.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308–311.

The International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796.

The International SNP Map Working Group (2001) Variation is the spice of life. *Nature* **409**, 928–933.

Useche FJ, Gao G, Harafey M and Rafalski A (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Informatics Series Workshop on Genome Informatics*, **12**, 194–203.

van Baren MJ and Heutink P (2004) The PCR suite. *Bioinformatics*, **20**, 591–593.

Introductory Review

Prokaryotic gene identification *in silico*

Mark Borodovsky and Rajeev Azad
Georgia Institute of Technology, Atlanta, GA, USA

Statistical methods for the identification of protein-coding regions in prokaryotic genomes have been the main tools for gene annotation since the start of genomic era. The pioneer papers describing gene-recognition algorithms appeared in press in the 1980s (Fickett, 1982; Staden, 1984; Gribskov *et al*., 1984; Almagor, 1985; Claverie and Bougueleret, 1986; Borodovsky *et al*., 1986a, b, c; Fichant and Gautier, 1987). These methods were using the so-called intrinsic information, statistical patterns in nucleotide ordering, characteristic for protein-coding and noncoding regions observed in a single DNA sequence. Methods utilizing the so-called extrinsic information embodied in common or similar sequences conserved in evolution and identified by comparison of DNA or proteins of different species were developed later (Gish and States, 1993; Robison *et al*., 1994). With the fast accumulation of DNA and protein sequences, use of extrinsic information has gained more attention and methods combining both intrinsic and extrinsic information have appeared (Borodovsky *et al*., 1994; Frishman *et al*., 1998).

Development of *ab initio* gene recognition algorithms that use intrinsic information generally follows three steps. At the first step, statistical measures (e.g., G + C composition) that may help distinguish sequence types (e.g., protein-coding vs. noncoding) are determined. At the second step, the statistical models are built for sequences of all types. At the third step, the models are integrated into a statistical pattern recognition algorithm.

In 1992, in a comprehensive review, Fickett and Tung analyzed more than 20 types of statistical measures of protein-coding regions and singled out as the most informative ones the frame-dependent (positional) hexamer frequencies (Fickett and Tung, 1992).

Introduced in 1986, inhomogeneous Markov chain models of a protein-coding region (Borodovsky *et al*., 1986b) have been a rigorous mathematical framework for using positional oligonucleotide frequencies. This model combined with a homogeneous Markov model of noncoding sequence in a Bayesian algorithm possesses a significant predictive power, as was demonstrated by the GeneMark gene finding program (Borodovsky *et al*., 1986c; Borodovsky and McIninch, 1993).

The GeneMark algorithm identifies protein-coding regions in both DNA strands by analyzing only one of them. To accomplish this, a Markov model of a DNA sequence complementary to a gene on the opposite strand (“shadow” model)

is added to the already-mentioned models of "direct" protein-coding sequence and noncoding sequence. The algorithm outline is as follows. Given the first-order Markov chain model, the probability that a DNA sequence segment, $S = \{s_1, s_2, \ldots, s_n\}$ (n is a multiple of 3), comes from a protein-coding region with s_1 being the first nucleotide of a codon (event E_1 or frame 1 setting) is given by the formula $P(S|E_1) = P^1(s_1) * P^1(s_2|s_1) * P^2(s_3|s_2) * P^3(s_4|s_3) * \cdots * P^2(s_n|s_{n-1})$ Here $P^i(s_1)$ is the probability for s_1 to be situated in frame i and $P^i(s_k|s_{k-1})$ is the transition probability of s_k to succeed s_{k-1} situated in frame i. Similar expressions for $P(S|E_i)$, $i = 2$ to 7 could be written for the remaining six events, which correspond to sequence S coming from the coding region in frame 2 and frame 3 settings (E_i, $i = 2, 3$); from a region complementary to a gene in frames 4, 5, and 6 (E_i, $i = 4, 5, 6$), here the shadow model is used; and from a noncoding region (E_7), here the noncoding model is used (and no frame setting is needed). Thus, the determined probabilities $P(S|E_i)$ are used to compute an *a posteriori* probability for each E_i to be associated with observed sequence S, $P(E_i|S) = \frac{P(S|E_i)*P(E_i)}{\sum_i P(S|E_i)*P(E_i)}$. Here $P(E_i)$ is the *a priori* probability of event E_i. It is possible to consider the GeneMark algorithm as an approximation to the forward–backward algorithm, known in the theory of hidden Markov models (HMM) (Rabiner, 1989), which computes the *a posteriori* probability of a hidden (functional) state underlying a given observed state (a nucleotide) in a particular sequence position (Azad and Borodovsky, 2004a). This interpretation could be unexpected as the forward–backward algorithm is used frequently at a training step for an HMM-based algorithm. Here we certainly deal with the prediction step. Still, we may consider the value of $P(S|E_i)$ computed by GeneMark for a short DNA segment S (within a sliding window) to be an approximation to the value of an *a posteriori* probability for the hidden state E_i for a nucleotide position in the middle of segment S computed by a rigorous forward–backward HMM algorithm, given the entire model of states and transitions.

Given the notion that genes in prokaryotic genomes are continuous open reading frames (ORFs), though by far not all ORFs are genes, the GeneMark program computes a scoring statistics for each ORF in the genomic sequence. This score is the average value of the *a posteriori* probabilities of the events E_i (for one and the same i defined by the ORF) for the short subsequences of the ORF in question. If this value is larger than a certain threshold, then the ORF is predicted as a gene.

It can be shown that Markov models of higher orders perform better for gene recognition than the first-order ones. The difficulty though may arise with the high-order model parameter estimation (model training). If the length of DNA sequence available for training is limited, an attempt to derive parameters of Markov models of high order may end up in overfitting of at least a fraction of the model parameters to the training data. However, the attempt to recover the informative parameters of the high-order model and combine them with informative parameters of lower-order models may result in better predictive ability (Jelinek *et al*., 1980). Such combined (interpolated) models have been used in the Glimmer algorithm (Salzberg *et al*., 1998; Delcher *et al*., 1999). Their parameters are defined by recursive equation, $P^{IMM}(b|c_k) = \lambda(c_k) * P(b|c_k) + (1 - \lambda(c_k)) * P^{IMM}(b|c_{k-1})$. Here $P^{IMM}(b|c_k)$ is the interpolated transition probability of a nucleotide b to succeed k-mer c_k,

$P(b|c_k)$ is the transition probability as defined in a regular Markov chain model, and $\lambda(c_k)$ is the interpolation parameter corresponding to k-mer c_k, $0 \leq \lambda(c_k) \leq 1$. It is reasonable to expect that if the frequency of c_k in the training sequence is sufficiently high, then $\lambda(c_k)$ should be close to 1; for low frequency c_k, $\lambda(c_k)$ should be close to 0. However, accurate estimation of $\lambda(c_k)$ is a critical issue and several approaches have been proposed to address it (Salzberg *et al*., 1998; Potamianos and Jelinek, 1998; Azad and Borodovsky, 2004b).

The Bayesian framework was further generalized by using HMM, which have been used in several prokaryotic as well as eukaryotic gene prediction algorithms developed in the last decade (Krogh *et al*., 1994; Yada and Hirosawa, 1996; Lukashin and Borodovsky, 1998; Yada *et al*., 2001; Larsen and Krogh, 2003; Burge and Karlin, 1997; Krogh, 2000; *see also* Article 14, **Eukaryotic gene finding**, Volume 7, Article 21, **Gene structure prediction in plant genomes**, Volume 7, and Article 26, **Dynamic programming for gene finders**, Volume 7). In an HMM, the gene recognition problem is reformulated as a problem of finding the most likely sequence $H = \{H_1, \ldots, H_n\}$, where H_i stands for a hidden state, for example protein-coding or noncoding, underlying a given DNA sequence $S = \{s_1, \ldots, s_n\}$. Hidden (functional) and observed (nucleotide) states are connected in an HMM by probabilistic (Markov chain type) dependencies. If HMM structure and parameters are defined, then for a given sequence S, the Viterbi algorithm solves the problem stated above and finds the most likely state sequence $H*$ by maximizing the conditional probability $P(H|S)$ with respect to H.

The first HMM-based gene prediction algorithm, ECOPARSE, was described by Krogh *et al*. (1994). ECOPARSE and the later developed GeneHacker (Yada and Hirosawa, 1996) are based on a classic HMM, with a hidden state emitting one nucleotide triplet (codon) from the protein-coding state and emitting one nucleotide from noncoding state. In the GeneMark.hmm program (Lukashin and Borodovsky, 1998), a hidden state is assumed to emit a whole sequence of observed states (nucleotides), with length specified by a probability distribution. The homogeneous and inhomogeneous Markov models (earlier utilized in GeneMark) are used as submodels of the main HMM to describe these emitted nucleotide sequences. The HMM architecture of GeneMark.hmm also included hidden states for rare “atypical” (e.g., horizontally transferred) genes, along with states for abundant native “typical” genes. Note that GeneMark.hmm and GeneMark possess complementary properties as the former implements the generalized Viterbi algorithm dealing with the sequence as a whole and the latter approximates the forward–backward algorithm, revealing local features of gene organization. Particularly, the *a posteriori* probability graphs generated by GeneMark in six reading frames help in detecting frameshifts (local irregularities) in the sequence that mainly occur because of sequencing errors.

Extrinsic information could be used in gene finding algorithms in several different ways. Early extrinsic algorithms used straightforward similarity search on protein level, that is, BLASTX, to identify ORFs whose protein products score statistically significant hits to known proteins. More recent algorithms combine extrinsic and intrinsic information. For instance, parameters of the gene models in the ORPHEUS program (Frishman *et al*., 1998) are estimated from ORF sequences whose protein products show significant similarity to known proteins.

As an indicator of the genetic code presence, ORPHEUS is using the statistic $\Omega = R(a_1a_2 \ldots a_n) - \max\{R(b_1b_2 \ldots b_n), R(c_1c_2 \ldots c_n)\}$, where $R(a_1a_2 \ldots a_n)$, is the normalized value (in units of standard deviation) of the quantity $\sum_{i=1}^{n} \log f(a_i)$. Here $f\ (x_\mathrm{i})$ is the frequency of x_i codon and $(a_1 \ldots a_n)$, $(b_1 \ldots b_n)$, $(c_1 \ldots c_n)$ stand for the sequences of codons in three possible frames. Similar validation of a training set by using extrinsic data is implemented in the recent algorithm EasyGene (Larsen and Krogh, 2003). EasyGene scores sequence S of each ORF by computing a statistics $W = \log\left(\frac{P(S|M)}{P(S|N)}\right)$ where $P(S|M)$ and $P(S|N)$ are the probabilities of S being generated by HMM of a gene, M, and HMM of noncoding region, N, respectively. A rather unique feature of EasyGene is the assessment of the statistical significance of the computed score. The CRITICA algorithm (Badger and Olsen, 1999) uses both extrinsic and intrinsic approaches to assign scores reflecting protein-coding propensity to a DNA sequence fragment in each of the six possible frames. The score is a sum of two components: a comparative (extrinsic) score based on the relative identities of nucleotides and corresponding potential amino acids and noncomparative (intrinsic) score based on dicodon bias in real protein-coding frames. Shibuya and Rigoutsos (2002) developed Bio-Dictionary, a database of patterns called "seqlets" extracted from protein sequence databases. If the number of seqlets matching to the ORF protein product exceeds a predetermined threshold, the ORF is predicted as a gene. This approach was implemented in a gene prediction algorithm, called Bio-Dictionary Gene Finder (BDGF) (Shibuya and Rigoutsos, 2002). Finally, one of the most extrinsic information oriented programs GeneHacker Plus (Yada *et al*., 2001), an improved version of GeneHacker (Yada and Hirosawa, 1996), has an ability to include the results of similarity searches as restrictions in the Viterbi algorithm.

The exact determination of gene start position is an important part of prokaryotic gene identification task. The algorithms mentioned above, ECOPARSE, EasyGene, GeneHackerPlus, GeneMark.hmm, and Glimmer use ribosome binding site (RBS) models to improve accuracy of gene start prediction. A two-component RBS model is included in the HMM architecture of GeneMark.hmm and its parameters along with other model parameters could be derived by GeneMarkS (Besemer *et al*., 2001), the unsupervised training program implementing the expectation-maximization procedure. This procedure iterates predictions and model derivation to the point where the gene predictions made with updated models coincide with gene prediction in the previous step (i.e., till the convergence is reached). The convergence is facilitated by use of heuristic pseudocounts at the initial step of the procedure (Besemer and Borodovsky, 1999). Note that several of the above-mentioned algorithms have extensions for deriving models by unsupervised learning procedure, which are useful if one has to start analysis of a new genome. Assessment of the performance of gene finding algorithms has been far from simple issue due to the lack of sufficient number of experimentally validated genes. Even an ability of a program to reproduce correctly most of the genes annotated in GenBank will not provide conclusive evidence of the algorithm's high predictive power unless the test on experimentally validated genes would support this statement. Two parameters are usually used to measure the prediction accuracy: the percentage of genes in the test set that are correctly identified by the

algorithm (Sensitivity, Sn) and the percentage of predictions that match the genes in the test set (Specificity, Sp).

All above-mentioned programs have shown high performance with Sn and Sp values higher than 90% in majority of the tests that were done with regard to GenBank annotated sequences. Still there is a need to carefully design test sets for various prokaryotic species and assess the performance if one is interested in this issue in depth.

Currently, the largest number of experimentally validated genes is known for *Escherichia coli*. When 195 *E. coli* genes with their starts determined by N-terminal protein sequencing were used for accuracy tests, both GeneMarkS and EasyGene correctly identified above 93% of the gene starts (Larsen and Krogh, 2003). Similar results were observed when the same algorithms were tested on a larger set comprising 850 *E. coli* genes with validated start positions (Rudd, 2000) (see **http://bmb.med.miami.edu/EcoGene/EcoWeb/**).

The rather short history of development of the prokaryotic gene identification algorithms presents a picture of significant accomplishments. However, the perfectionist goal to identify by computer all protein-coding genes in a given genome without false-positives remains a challenge. The main part of this challenge is the flawless detection of short genes. Identification of atypical or laterally transferred genes is a closely related problem as genes from this group are frequently escaping detection. Further improvement of gene start prediction is another issue that remains to be addressed. However, continuation of strong efforts toward integrating intrinsic and extrinsic approaches should bring us closer to a point when the prokaryotic gene identification problem would be declared completely solved.

References

Almagor H (1985) Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach. *Journal of Theoretical Biology*, **117**, 127–136.

Azad RK and Borodovsky M (2004a) Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Briefings in Bioinformatics*. **5**, 118–130

Azad, RK and Borodovsky, M (2004b) Effects of choice of DNA sequence model structure on gene identification accuracy. *Bioinformatics*, **20**, 993–1005

Badger JH and Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, **16**, 512–524.

Besemer J and Borodovsky M (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Research*, **27**, 3911–3920.

Besemer J, Lomsadze A and Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, **29**, 2607–2618.

Borodovsky M and McIninch J (1993) GeneMark: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, **17**, 123–133.

Borodovsky M, Rudd KE and Koonin EV (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Research*, **22**, 4756–4767.

Borodovsky M, Sprizhitsky YuA, Golovanov EI and Alexandrov AA (1986a) Statistical patterns in the primary structures of functional regions of the genome in *Escherichia coli*: I. Oligonucleotide Frequencies Analysis. *Moleculars Biology*, **20**, 826–833.

Borodovsky M, Sprizhitsky YuA, Golovanov EI and Alexandrov AA (1986b) Statistical patterns in the primary structures of functional regions of the genome in *Escherichia coli*: II. Nonuniform Markov models. *Molecular Biology*, **20**, 833–840.

Borodovsky M, Sprizhitsky YuA, Golovanov EI and Alexandrov AA (1986c) Statistical patterns in the primary structures of functional regions of the genome in *Escherichia coli*: III. Computer recognition of coding regions. *Molecular Biology*, **20**, 1144–1150.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Claverie JM and Bougueleret L (1986) Heuristic informational analysis of sequences. *Nucleic Acids Research*, **14**, 179–196.

Delcher AL, Harmon D, Kasif S, White O and Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**, 4636–4641.

Fichant G and Gautier C (1987) Statistical method for predicting protein−coding regions in nucleic acid sequences. *Computer Applications in the Biosciences*, **3**, 287–295.

Fickett JW (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, **10**, 5303–5318.

Fickett JW and Tung CS (1992) Assessment of protein coding measures. *Nucleic Acids Research*, **20**, 6441–6450.

Frishman D, Mironov A, Mewes H-W and Gelfand M (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Research*, **26**, 2941–2947.

Gish W and States DJ (1993) Identification of protein coding regions by database similarity search. *Nature Genetics*, **3**, 266–272.

Gribskov M, Devereux J and Burgess RR (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Research*, **12**, 539–549.

Jelinek F and Mercer, RL (1980) Interpolated estimation of Markov source parameters from sparse data. In *Pattern Recognition in Practice*, Gelsema ES and Kanal LN (Eds.), North-Holland Publishing Company: Amsterdam, pp. 381–397.

Krogh A (2000) Using database matches with HMMGene for automated gene detection in Drosophila. *Genome Research*, **10**, 523–528.

Krogh A, Mian IS and Haussler D (1994) A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Research*, **22**, 4768–4778.

Larsen TS and Krogh A (2003) EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.

Lukashin AV and Borodovsky M (1998) GeneMark.hmm: new solutions for gene Finding. *Nucleic Acids Research*, **26**, 1107–1115.

Potamianos G and Jelinek F (1998) A study of N−gram and decision tree letter language modeling methods. *Speech Communication*, **24**, 171–192.

Rabiner L (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

Robison K, Gilbert W and Church GM (1994) Large scale bacterial gene discovery by similarity search. *Nature Genetics*, **7**, 205–214.

Rudd KE (2000) EcoGene: a genome sequence database for *Escherichia coli* K−12. *Nucleic Acids Research*, **28**, 60–64.

Salzberg SL, Delcher AL, Kasif S and White O (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, **26**, 544–548.

Shibuya T and Rigoutsos I (2002) Dictionary−driven prokaryotic gene finding. *Nucleic Acids Research*, **30**, 2710–2725.

Staden R (1984) Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research*, **12**, 551–567.

Yada T and Hirosawa M (1996) Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model. *DNA Research*, **3**, 355–361.

Yada T, Totoki Y, Takagi T and Nakai K (2001) A novel bacterial gene- finding system with improved accuracy in locating start codons. *DNA Research*, **8**, 97–106.

Eukaryotic gene finding

Roderic Guigó
Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, and Centre de Regulació Genòmica, Barcelona, Spain

1. Introduction

Gene prediction is the process of inferring the sequence of the functional products encoded in genomic DNA sequences. In this chapter, we will review computational methods to predict protein-coding genes. These methods are usually limited to the prediction of the coding fraction of the genes and they ignore for the most part the prediction of the terminal 5′ and 3′ untranslated regions.

In the eukaryotic cell, the pathway leading from the DNA sequence of the genome to the amino acid sequence of the proteins proceeds in a number of steps. A gene sequence is first transcribed into a pre-mRNA sequence, this transcript is subsequently processed (i.e., capped, spliced, and polyadenylated), and the mature mRNA transcript is then transported from the nucleus into the cytoplasm for translation into a functional protein. A number of sequence motifs in the primary DNA sequence or in the intermediate RNA sequences are recognized along this pathway: promoter elements that indicate the site for the initiation of the transcription, splice sites that define the boundaries of the introns removed during splicing, and the initiation and termination codons that define the segment of the mature mRNA translated into protein. A natural approach is for computational gene-prediction methods to attempt to model these sequence motifs, and the mechanism by means of which they are recognized. However, known sequence motifs involved in gene specification carry not enough information to precisely elucidate eukaryotic gene structures, and current computational methods need to rely on additional information such as sequence compositional bias and similarity to known coding sequences. With the flowering of eukaryotic sequencing, genome comparisons have also recently emerged as a powerful approach for eukaryotic gene prediction.

In summary, gene-finding programs predict genes in genomic sequences by a combination of one or usually more of the following approaches:

1. analysis of sequence signals potentially involved in gene specification
2. analysis of regions showing sequence compositional bias correlated with coding function
3. comparison with known coding sequences (proteins or cDNAs)
4. comparison with anonymous genomic sequences.

The so-called ab initio or de novo gene predictors combine approaches 1 and 2. We will term here methods relying mostly in approach 3 as "homology-based" gene predictors, and those mostly based on approach 4 as "comparative" gene predictors. There is not, however, a widely accepted terminology, and, for instance, "comparative" gene predictors are often included within the "de novo" category to separate them from those methods relying on expression data (proteins or cDNAs). Moreover, as programs become increasingly complex, the frontiers between the different categories become very fuzzy, with systems existing nowadays that employ all of the strategies enumerated above.

In this article, we will first review methods to identify sequence signals and to capture the sequence composition bias characteristic of protein-coding regions. We will then describe how these methods are integrated in "ab initio" gene finders, and how these are extended into homology-based and comparative predictors.

2. Prediction of signals

Splice sites (donor and acceptor), translation initiation, and termination codons are the basic sequence signals involved in the definition of coding exons. Typically, these signals are modeled as position weight matrices (PWMs, see Figure 1). The coefficients in these matrices are the observed probabilities in known functional sites of each nucleotide at each position along the signal. Assuming independence between positions, PWMs can be used to compute along a genomic interval the probability of the sequence at a given position assuming that it corresponds to a functional site. Those positions in which this probability is much larger than the background probability of the sequence are typically predicted as potential functional sites. However, because of the degeneracy of the signals, usually many more sites are predicted to be functional than actually are.

To improve signal prediction, methods have been developed that relax the assumption of independence between adjacent positions. In the so-called first-order Markov or weight array model (Zhang and Marr, 1993), the probability of a nucleotide at a given position is conditioned on the nucleotide at the preceding position. Higher-order Markov models – in which the probability at one position is conditioned to a number of preceding positions – have also been considered (see, for instance, Salzberg *et al*., 1998). Dependencies, however, may not be limited to adjacent positions along the signal. Nonlocal dependencies are implicit in the maximal dependence decomposition (MDD) method (Burge and Karlin, 1997), which has been applied to the prediction of donor sites. More recently, methods based on Bayesian networks, and other probabilistic frameworks, that attempt to systematically recover all relevant dependencies between positions in sequence signals, have been developed for the prediction of splice sites (see, for instance, Castelo and Guigó, 2004). Machine-learning techniques, which capture nonlocal dependencies, have been applied to the prediction of sequence signals as well. The NetGene program (Brunak *et al*., 1991) is an early example of the application of neural networks (*see* Article 98, **Hidden Markov models and neural networks**, Volume 8) to splice site prediction. See Article 16, **Searching for genes and biologically related**

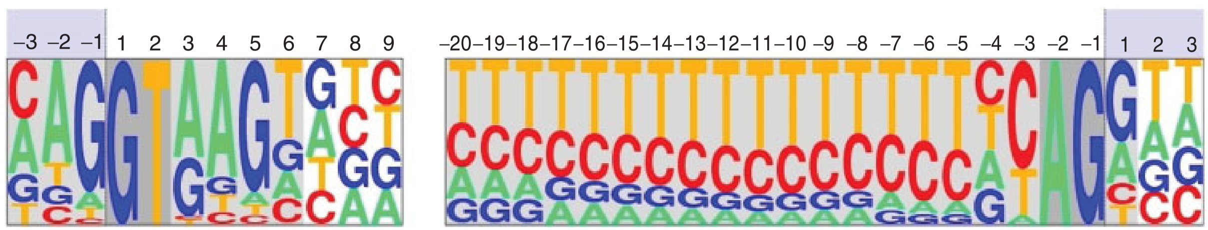


Figure 1 Sequence logo displaying the position weight matrices (PWMs) corresponding to the canonical donor and acceptor sites of U2 eukaryotic introns. The height of each letter at each position along the site is proportional to the frequency of the corresponding nucleotide as observed in real splice sites. The deviation from random composition can be used to develop methods to predict splice sites. Sequence logo obtained with the program pictogram (Burge *et al*., **http://genes.mit.edu/pictogram.html**)

signals in DNA sequences, Volume 7 and Article 28, **Computational motif discovery**, Volume 7 for a more in-depth review of methods to predict sequence signals.

Despite the variety of efforts, prediction of splice sites (and of other signal sequences) remains an elusive problem. In this regard, recent research suggests that signals other than those at the splice sites play a role in the definition of the intron boundaries. Computational analysis have revealed enhancer sequences that promote splicing of introns with constitutively weak sites, as well as repressor sequences that would prevent splicing of "pseudoexons" with otherwise strong splice signals. See Article 24, **Exonic splicing enhancers and exonic splicing silencers**, Volume 7 for a more in-depth review of exonic splicing enhancers.

3. Prediction of coding regions

Protein-coding regions exhibit characteristic DNA sequence composition bias, which is absent for noncoding regions (Figure 2). The bias is a consequence of (1) the uneven usage of the amino acids in real proteins, (2) the uneven usage of synonymous codons, and (3) dependencies between adjacent amino acid residues in proteins. To discriminate protein coding from noncoding regions, a number of content measures can be computed to detect this bias (see Guigó, 1999). Such content measures – also known as coding statistics – can be defined as functions that compute a real number related to the likelihood that a given DNA sequence codes for a protein (or a fragment of a protein). Since the early eighties, a great number of coding statistics have been published in the literature. Hexamer frequencies usually in the form of fifth-order Markov Models (Borodovsky and McIninch, 1993) appear to offer the maximum discriminative power, and are at the core of most popular gene finders today. Combination of a number of coding statistics, on the other hand, seem to improve performance, as in the pioneer GRAIL program (Uberbacher and Mural, 1991).

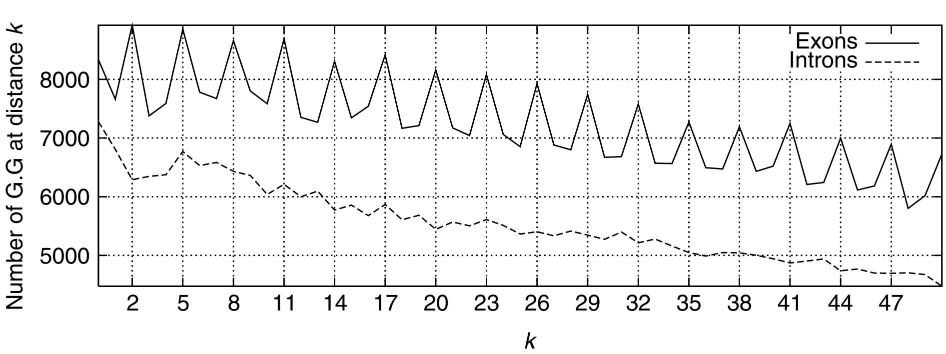


Figure 2 The absolute frequency of the pair G ... G with k (from 0 to 50) nucleotides between the two A's in the 200 first base pairs accumulated in a set of 1753 human exons and introns. A clear period-3 pattern appears in coding regions, which is absent in noncoding regions. This periodic pattern is a reflection of the characteristic codon usage that can be used to distinguish coding from noncoding regions (see Guigó, 1999)

4. "Ab intitio" gene prediction

Most existing computational gene-prediction programs combine at least prediction of sequence signals and prediction of coding regions. Typically, these programs carry out the following tasks: (1) identification of splice sites and other exon-defining sequence signals, (2) prediction of candidate exons based on the predicted signals, (3) scoring of the exons as a function, at least, of the score of the defining signals and the coding statistics computed on the exon sequence, and (4) assembling of a subset of the predicted exons into an "optimal" gene structure – usually the gene structure that maximizes a given scoring function. If this is a linear function of the exon scores, a computational technique known as *dynamic programming* (*see* Article 26, **Dynamic programming for gene finders**, Volume 7) is able to find the solution very efficiently. The programs GENEID (Guigó *et al*., 1992) and FGENESH (Solovyev *et al*., 1995) use this approach explicitly. Under the assumption of linearity in the scoring schema, hidden Markov models (HMMs, *see* Article 98, **Hidden Markov models and neural networks**, Volume 8) offer a robust framework where the optimal structure can also be obtained very efficiently, with the advantage that it has a precise probabilistic interpretation. Since the development of GENSCAN (Burge and Karlin, 1997), a program that has been widely used in the annotation of vertebrate genomes, HMMs have been very popular for gene finding. Pioneer HMM gene finders include GENIE (Kulp *et al*., 1996) and HMM-gene (Krogh, 1997).

5. Homology-based gene prediction

When the query genomic sequence encodes genes with known protein or expressed sequence tag (EST) homologs, this information can be used to improve the predictions. One approach is to incorporate in the exon-scoring schema of an underlying "ab initio" program the quality of the alignment between a putative exon and known coding sequences. GENOMESCAN (Yeh *et al*., 2001), an extension of GENSCAN, among others, incorporates protein similarity, while GRAIL-EXP, an extension of GRAIL, is an early example incorporating similarity to ESTs.

When the homology is very strong (or the genomic sequence encodes a known gene or a gene with ESTs), a more sophisticated approach involves aligning the genomic query directly against the known protein or cDNA target. These alignments, often referred to as "spliced alignments" (*see* Article 15, **Spliced alignment**, Volume 7), are usually computationally expensive and cannot be used for database searches, requiring the target sequence to be known in advance. PROCRUSTES (Gelfand *et al*., 1996) and GENEWISE (Birney and Durbin, 1997) are a few examples of this approach.

6. Comparative gene prediction

Comparative (or dual) genome gene predictors rely on the fact that functional regions in the genome sequence – protein-coding genes in particular – are more

conserved during evolution than nonfunctional ones. Over the last few years, a number of programs have been developed that exploit sequence conservation between two genomes. A wide variety of strategies has been explored, ranging from complex extensions of the classical dynamic programming algorithm for sequence alignment that allow for heterogeneous models of sequence conservation to extensions of single-genome predictors allowing for the integration of sequence conservation with another genome in the exon-scoring schema. Comparative gene predictors outperform their single genome "ab initio" counterparts, and three such predictors, SGP-2 (Parra *et al*., 2003), SLAM (Alexandersson *et al*., 2003), and TWINSCAN (Korf *et al*., 2001) have been used in the comparative annotation of the human and rodent genomes (mouse, rat papers). SLAM is a pair-HMM-based method (*see* Article 17, **Pair hidden Markov models**, Volume 7), in which gene prediction and sequence alignment are obtained simultaneously. TWINSCAN is an extension of GENSCAN, while SGP-2 is an extension of GENEID. The probability scores that these two programs assign to each potential exon are modified by the presence and quality of genome alignments.

As the number of genomes sequenced at different evolutionary distances increases, methods to predict genes on the basis of the analysis of multiple species – and not of two species only – are being investigated. In the so-called phylogenetic hidden Markov models (phylo-HMMs) or evolutionary hidden Markov models (EHMM), a gene-prediction HMM is combined with a set of evolutionary models, thus taking into account that the rate (and type) of evolutionary events differs in protein-coding and noncoding regions. The recent application of phylo-HMMs to gene prediction has produced encouraging results (for instance, Siepel and Haussler, 2004). See Article 27, **Gene structure prediction by genomic sequence alignment**, Volume 7 for a more in-depth review of comparative gene-prediction methods.

7. Gene prediction in complete genomes

While a number of gene-prediction programs are efficient enough that they can be easily run on complete eukaryotic genomes, integrated pipelines that execute a number of analysis programs and weigh the different sources of evidence provide the most accurate gene annotations on complex genomes. Probably, the best-known eukaryotic system is ENSEMBL (**http://www.ensembl.org**, *see* Article 97, **Ensembl and UniProt (Swiss-Prot)**, Volume 8). In the ENSEMBL pipeline, protein-coding database searches and predictions by GENSCAN are used to identify targets for genomic to protein alignments by GENEWISE. The annotation of the *Tetraodon nigroviridis* genome used a different pipeline, based on the GAZE system. GAZE (Howe *et al*., 2002) is a generic framework for the integration of arbitrary gene-prediction data (from sequence signals to complete gene predictions) using dynamic programming. A more in-depth review of gene-finding pipelines can be found elsewhere in this encyclopedia.

Integrated pipelines, complemented with comparative gene-prediction programs, appear to provide accurate enough first-pass annotation of complex eukaryotic genomes when good collections of cDNA or EST sequences exist. Coupled with

recent advances in pseudogene detection methods, they have reached a point where computational predictions can be used as a hypothesis to drive experimental annotation via systematic RT-PCR.

When the collections of expression data are poor or nonexistent, however, computational predictions are less reliable. For instance, experimental cloning of genes initially predicted in *Caenorhabditis elegans* revealed at least 50% of incorrect exonic structures (Reboul *et al*., 2003). The problem is compounded in those cases for which no phylogenetically related genome has been sequenced. The lack of expression data makes it very difficult to estimate the parameters of the gene-prediction programs, and the common practice of using a gene predictor trained on a different, phylogenetically distant species may lead to poor predictions (Korf, 2004). Since a precise assessment of the accuracy of the gene predictions is important, substantial effort has been made during the last decade to develop consistent accuracy metrics and standard datasets. See Guigó and Wiehe (2003) for a review of efforts to evaluate the accuracy of gene-prediction programs.

In summary, while substantial progress has occurred in the field during the last 15 years, computational gene prediction is still an open problem whose solution will require, not only more sophisticated computational methodologies, but a better understanding of the biological mechanisms by which the sequence signals involved in gene specification are processed and recognized in the eukaryotic cell.

Related articles

See Article 13, **Prokaryotic gene identification *in silico***, Volume 7; Article 15, **Spliced alignment**, Volume 7; Article 16, **Searching for genes and biologically related signals in DNA sequences**, Volume 7; Article 17, **Pair hidden Markov models**, Volume 7; Article 18, **Information theory as a model of genomic sequences**, Volume 7; Article 19, **Promoter prediction**, Volume 7; Article 20, **Operon finding in bacteria**, Volume 7; Article 21, **Gene structure prediction in plant genomes**, Volume 7; Article 22, **Eukaryotic regulatory sequences**, Volume 7; Article 23, **Alternative splicing in humans**, Volume 7; Article 24, **Exonic splicing enhancers and exonic splicing silencers**, Volume 7; Article 25, **Gene finding using multiple related species: a classification approach**, Volume 7; Article 26, **Dynamic programming for gene finders**, Volume 7; Article 27, **Gene structure prediction by genomic sequence alignment**, Volume 7; Article 28, **Computational motif discovery**, Volume 7

Further reading

Brent MR and Guigó R (2004) Recent advances in gene structure prediction. *Current Opinion in Structural Biology*, **14**, 264–272.

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, **3**, 698–709.

References

Alexandersson M, Cawley S and Patcher L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, **13**, 496–502.

Birney E and Durbin R (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Vol. 5, AAAI Press: pp. 56–64.

Borodovsky M and McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry*, **17**, 123–133.

Brunak S, Engelbrecht J and Knudsen S (1991) Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Research*, **18**, 4797–4801.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Castelo R and Guigó R (2004) Splice site identification by idlBNS. *Bioinformatics*, **20**, I69–I76.

Gelfand MS, Mironov AA and Pevzner PA (1996) Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 9061–9066.

Guigó R (1999) DNA composition, codon usage and exon prediction. In *Genetic Databases*, Bishop M (Ed.), Academic Press: pp. 53–80.

Guigó R, Drake N, Knudsen S and Smith T (1992) Prediction of gene structure. *Journal of Molecular Biology*, **226**, 141–157.

Guigó R and Wiehe T (2003) Gene prediction accuracy in large DNA sequences. In *Frontiers in Computational Genomics*, Galperin MY and Koonin EV (Eds.), Caister Academic Press: pp. 1–33.

Howe KL, Chothia T and Durbin R (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Research*, **12**, 1418–1427.

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.

Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Vol. 5, AAAI Press: pp. 179–186.

Kulp D, Haussler D, Reese MG and Eeckman FH (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, Vol. 4, AAAI Press: pp. 134–142.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW and Guigó R (2003) Comparative gene prediction in human and mouse. *Genome Research*, **13**, 108–117.

Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, *et al*. (2003) C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics*, **34**, 35–41.

Salzberg SL, Delcher AL, Kasif S and White O (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, **26**, 544–548.

Siepel A and Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, **11**, 413–428.

Solovyev VV, Salamov AA and Lawrence CB (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, Vol. 3, AAAI Press: pp. 367–375.

Uberbacher EC and Mural RJ (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 11261–11265.

Yeh R, Lim LP and Burge C (2001) Computational inference of the homologous gene structures in the human genome. *Genome Research*, **11**, 803–816.

Zhang MQ and Marr TG (1993) A weight array method for splicing signal analysis. *Computer Applications in the Biosciences: CABIOS*, **9**, 499–509.

Specialist Review

Spliced alignment

Xiaoqiu Huang
Iowa State University, Ames, IA, USA

1. Introduction

Finding genes in genomic DNA sequences is a fundamental step in genome analysis. Spliced alignment is an effective method for finding multiexon genes for which a similar cDNA or protein sequence is available. A cDNA sequence is either of full length or of partial length such as an expressed sequence tag (EST). The coordinates of exons that are similar to a cDNA or protein sequence are located on the basis of sequence similarity and splice site strength. This approach is very accurate if exons are highly similar to a cDNA or protein sequence.

In this chapter, we first present a mathematical formulation of the spliced alignment problem such that a mathematically optimal solution is likely to be exactly correct in exon coordinates. Both sequence similarity and splice site strength are included in the formulation. Then, we describe a dynamic programming algorithm for computing an optimal spliced alignment of a genomic sequence and a cDNA sequence. We briefly discuss spliced alignment methods that explore protein sequence similarity. Next, we present a fast method for dealing with a large set of genomic sequences and a set of cDNA sequences. We also describe methods used in existing spliced alignment programs and comment on their strengths. Finally, we offer suggestions for future improvements in spliced alignment.

2. An alignment model

A similarity relationship between a genomic sequence A and a cDNA sequence B is represented by an alignment of the two sequences, an ordered list of pairs of letters of A and B. The alignment consists of substitutions, deletion gaps, and insertion gaps. A substitution pairs a letter of A with a letter of B. A substitution is a match if the two letters are identical and a mismatch otherwise. A deletion gap is a gap in which letters of A correspond to no letter of B, and an insertion gap is a gap in which letters of B correspond to no letter of A. The length of a gap is the number of letters involved. Deletion and insertion gaps are defined with regard to transformation of sequence A into sequence B. An alignment of A and B shows a way to transform A into B, where a letter of A is replaced by a letter of B in every

```
TACTGGCAGCTA CGTACACTACCGTGCTACTACGGCCAGCTAGCTAGATCCGCTGTCAGTC

+++++++|||||-|||||--||||++++++++++++++++|||||| ||||+++++++++++

       AGCTATCGTAC  TACC                CTAGCTTGATC
```

Figure 1 An alignment of a genomic DNA sequence and a cDNA sequence. A match is indicated by a vertical bar, a mismatch is indicated by a blank, each intron deletion gap is indicated by a number of plus symbols, and other gaps are indicated by a number of dash symbols. This alignment contains 24 matches, 1 mismatch, an exon deletion gap of length 2, an insertion gap of length 1, and three intron deletion gaps of lengths 7, 16, and 11

substitution, the letters of A in every deletion gap are deleted, and the letters of B in every insertion gap are inserted.

There are two types of deletions gaps: exon deletion gaps and intron deletion gaps. Exon deletion gaps are introduced to represent minor differences between exon regions of A and regions of B, whereas intron deletion gaps are intended to represent major differences involving introns and other regions of A that do not correspond to any part of B. An alignment of a genomic DNA sequence and a cDNA sequence is shown in Figure 1.

The similarity of an alignment is measured by a numerical score. Let $\sigma(a,b)$ be the score of a substitution involving letters a and b. Let numbers q and r be gap-open and gap-extension penalties respectively. The numbers q and r are nonnegative. The score of an insertion or exon deletion gap of length h is $-(q + h \times r)$. Let k be a minimum cutoff on the lengths of intron deletion gaps. Let $d = q + k \times r$ be the penalty for a gap of length k. Let $do(i)$ denote the prediction score of a candidate donor site starting at position i of A, where position $i - 1$ belongs to a candidate exon region and position i belongs to a candidate intron region. Let $ac(i)$ denote the prediction score of a candidate acceptor site ending at position i of A, where position i belongs to a candidate intron region and position $i + 1$ belongs to a candidate exon region. The score of an intron deletion gap involving a region from positions i_s to i_e of A is $do(i_s) - d + ac(i_e)$. The similarity score of an alignment is just the sum of scores of each substitution and each gap in the alignment.

Values for the parameters σ, k, q, and r are specified by the user. A letter-independent substitution table is usually used for comparison of DNA and cDNA sequences. For example, each match is given a score of 10 and each mismatch a score of −20. Possible values for q and r are 40 and 2 respectively. For candidate donor sites with the GT dinucleotide and candidate acceptor sites with the AG dinucleotide, which are called canonical splice sites, their prediction scores are obtained by using a statistical method (Zhang and Marr, 1993; Salzberg, 1997; Brendel and Kleffe, 1998) and scaling resulting scores to a range from $-c \times r$ to $c \times r$, where c is a positive integer constant. For the remaining candidate donor and acceptor sites, their prediction scores are $-c \times r$. For example, a value of 5 is used for both c and k, each canonical candidate site is given a score of 10, and each noncanonical candidate site is given a score of −10. The score of the alignment in Figure 1 is −36 under the parameter values given above.

An optimal spliced alignment of two sequences A and B is an alignment of A and B with the maximum score. In the optimal alignment, exon regions of A are likely to be aligned with regions of B, and long introns and intergenic regions of A are likely to be treated as intron deletion gaps because intron deletions gaps are given a constant penalty. Exon–intron boundaries in A are likely to be exactly shown on the optimal alignment because the prediction scores of donor and acceptor sites are used in the definition of the optimal alignment.

3. A dynamic programming algorithm

Let $A = a_1\ a_2\ \ldots\ \ a_m$ be a genomic DNA sequence of length m and $B = b_1\ b_2\ \ldots\ \ b_n$ be a cDNA sequence of length n. A dynamic programming algorithm is developed to compute an optimal spliced alignment of A and B. Let $A_i = a_1\ a_2\ \ldots\ \ a_i$ and $B_j = b_1\ b_2\ \ldots\ \ b_j$ be the initial segments of lengths i and j of A and B. In the algorithm, a matrix S is introduced: $S(i,j)$ is the maximum score of all alignments of A_i and B_j. Thus, $S(m,n)$ is the score of an optimal spliced alignment of A and B. To compute the matrix S efficiently for $i > 0$ and $j > 0$, partition the set of all alignments of A_i and B_j into four groups. Group 1 consists of alignments of A_i and B_j ending with a substitution. Group 2 consists of alignments of A_i and B_j ending with an exon deletion gap. Group 3 consists of alignments of A_i and B_j ending with an insertion gap. Group 4 consists of alignments of A_i and B_j ending with an intron deletion gap.

Three additional matrices ED, ID, and I are introduced for groups 2 through 4. Define $ED(i,j)$ (ED for exon deletion) to be the maximum score of alignments in group 2. Define $I(i,j)$ (I for insertion) to be the maximum score of alignments in group 3. Define $ID(i,j) + ac(i)$ (ID for intron deletion) to be the maximum score of alignments in group 4. Note that the score of each alignment in group 4 includes the term $ac(i)$ as part of the score of an intron deletion gap ending at position i. A partial score of an alignment in group 4 is the score of the alignment minus $ac(i)$. Thus, $ID(i,j)$ is the maximum partial score of alignments in group 4. The introduction of the partial score leads to a simple formula for computing $ID(i,j)$.

On the basis of the definitions of the matrices given above, the recurrences for computing the matrices S, ED, I, and ID can be developed. The correctness proof for the recurrences is omitted. The complete recurrences for the matrices are given below.

$$
\begin{aligned}
S(0,0) &= 0 \\
S(i,0) &= \max\{ED(i,0), ID(i,0) + ac(i)\} \text{ for } i > 0 \\
S(0,j) &= I(0,j) \text{ for } j > 0 \\
S(i,j) &= \max\{S(i-1,j-1) + \sigma(a_i,b_j), ED(i,j), I(i,j), \\
&\qquad ID(i,j) + ac(i)\} \text{ for } i > 0 \text{ and } j > 0 \qquad (1)
\end{aligned}
$$

$$ED(0, j) = S(0, j) - q \text{ for } j \geq 0$$
$$ED(i, 0) = ED(i-1, 0) - r \text{ for } i > 0$$
$$ED(i, j) = \max\{ED(i-1, j) - r, S(i-1, j) - q - r\}$$
$$\text{for } i > 0 \text{ and } j > 0 \qquad (2)$$

$$I(i, 0) = S(i, 0) - q \text{ for } i \geq 0$$
$$I(0, j) = I(0, j-1) - r \text{ for } j > 0$$
$$I(i, j) = \max\{I(i, j-1) - r, S(i, j-1) - q - r\}$$
$$\text{for } i > 0 \text{ and } j > 0 \qquad (3)$$

$$ID(0, j) = S(0, j) + do(1) - d \text{ for } j \geq 0$$
$$ID(i, 0) = ID(i-1, 0) \text{ for } i > 0$$
$$ID(i, j) = \max\{ID(i-1, j), S(i-1, j) + do(i) - d\}$$
$$\text{for } i > 0 \text{ and } j > 0 \qquad (4)$$

The matrices can be computed in order of rows or columns. The value $S(m,n)$ is the score of an optimal spliced alignment of A and B. If only the score $S(m,n)$ is needed, then three linear arrays and a few scalars are sufficient to carry out the computation. This algorithm is an extension to standard global alignment algorithms (Needleman and Wunsch, 1970; Sellers, 1974; Wagner and Fischer, 1974; Waterman *et al*., 1976; Gotoh, 1982).

An optimal alignment is found by a traceback procedure through the matrices S, ED, I, and ID. The aligned pairs of an optimal alignment are generated in a reverse order with the last aligned pair produced first. The traceback procedure is based on the following observations. Let $R(A_m, B_n)$ denote an optimal alignment of A and B. If $S(m, n) = S(m-1, n-1) + \sigma(a_m, b_n)$, then the last aligned pair of $R(A_m, B_n)$ is a substitution pair. Otherwise, if $S(m, n) = ED(m, n)$, then thelast aligned pair is an exon deletion pair. Otherwise, if $S(m, n) = I(m, n)$, then the last aligned pair is an insertion pair. Otherwise, we have $S(m, n) = ID(m, n)$, and the last aligned pair is an intron deletion pair. Next, the procedure is repeated to produce the aligned pairs of the portion of $R(A_m, B_n)$ before the last pair.

Specifically, the traceback procedure consists of four cases, one for each matrix. Initially, the current entry is (m,n) in S. First, consider the case in which the current entry is (i,j) in S. If $i = 0$ and $j = 0$, then the traceback procedure terminates. Otherwise, if $i > 0$ and $S(i, j) = ED(i, j)$, then the current entry becomes (i,j) in ED. Otherwise, if $j > 0$ and $S(i, j) = I(i, j)$, then the current entry becomes (i,j) in I. Otherwise, if $i > 0$ and $S(i, j) = ID(i, j)$, then the current entry becomes (i,j) in ID. Otherwise, the current entry becomes $(i-1, j-1)$ in S and a substitution pair (a_i, b_j) is reported.

Then, consider the case in which the current entry is (i,j) in ED. An exon deletion pair $(a_i, -)$ is reported. If $i > 1$ and $ED(i, j) = ED(i-1, j) - r$, then the current entry becomes $(i-1,j)$ in ED. Otherwise, the current entry becomes $(i-1,j)$ in S.

Next, consider the case in which the current entry is (i,j) in I. An insertion pair $(-, b_j)$ is reported. If $j > 1$ and $I(i, j) = I(i, j-1) - r$, then the current entry becomes $(i,j-1)$ in I. Otherwise, the current entry becomes $(i,j-1)$ in S.

Finally, consider the case in which the current entry is (i,j) in ID. An intron deletion pair $(a_i, -)$ is reported. If $i > 1$ and $ID(i, j) = ID(i-1, j)$, then the current entry becomes $(i-1,j)$ in ID. Otherwise, the current entry becomes $(i-1,j)$ in S.

The traceback procedure requires that the complete matrices be saved or additional information be saved to indicate how the value at each matrix entry is generated, which takes computer memory proportional to the product $m \times n$. For example, on a genomic sequence of 100 000 bp and a cDNA sequence of 1000 bp, the algorithm requires about 100–400 MB of memory. The time requirement of the algorithm is also proportional to the product $m \times n$. For example, the algorithm requires less than 1 min on an ordinary computer to produce an optimal alignment of a genomic sequence of 100 000 bp and a cDNA sequence of 1000 bp. The space requirement of this algorithm is much more limiting than the time requirement.

An optimal alignment of a genomic sequence and a cDNA sequence can be computed in computer memory proportional to the sum of the sequence lengths by a divide-conquer technique (Hirschberg, 1975; Myers and Miller, 1988). The main idea of this technique is to determine a middle pair of positions on an optimal alignment in linear space. Then, the portions of the optimal alignment before and after the middle pair of positions are constructed recursively. This space-efficient technique is used in several programs for computing spliced alignments (Mott, 1997; Huang *et al*., 1997; Usuka *et al*., 2000).

We would like to point out that a dynamic programming algorithm could be developed for a version of the spliced alignment problem based on protein sequence similarity. The algorithm is able to produce, in presence of frameshifts, a strongest similarity relationship between codons of a genomic sequence and amino acids of a protein sequence. There are a few existing spliced alignment programs based on protein sequence similarity (Gelfand *et al*., 1996; Huang and Zhang, 1996; Usuka and Brendel, 2000).

4. Fast comparison of genomic and cDNA sequences

A general strategy for comparing a large set of genomic sequences with a set of cDNA sequences proceeds in two steps. In step 1, short word matches between genomic and cDNA sequences are quickly located through a lookup table (Pearson and Lipman, 1988; Altschul *et al*., 1990). Each word match is extended into a high-scoring segment pair (HSP), an ungapped alignment. Then, HSPs from the same pair of genomic and cDNA sequences are combined into chains (Wilbur and Lipman, 1983; Chao and Miller, 1995; Huang, 1996), where a chain of HSPs is an approximation to an alignment, with each HSP being an ungapped portion of the

alignment. In step 2, for each high-scoring chain between a region of a genomic sequence and a cDNA sequence, a small area of the dynamic programming matrix is located to cover all the HSPs in the chain. Then, a largest-scoring alignment of the genomic region and the cDNA sequence is produced by restricting the dynamic programming computation to the area.

Variations of this general strategy are used in the following programs: est2genome (Mott, 1997), DDS/GAP2 (Huang *et al*., 1997), sim4 (Florea *et al*., 1998), GeneSeqer (Usuka *et al*., 2000), Spidey (Wheelan *et al*., 2001), and BLAT (Kent, 2002).

The est2genome program works in three steps. In step 1, a genomic sequence is compared with a set of cDNA sequences with BLAST (Altschul *et al*., 1990). For each cDNA sequence similar to the genomic sequence, the start and end positions of an optimal local alignment are computed by the Smith–Waterman algorithm (Smith and Waterman, 1981). Then, the corresponding regions of the genomic and cDNA sequences are extracted and an optimal alignment of the two regions is computed by a dynamic programming algorithm in linear space. The GT-AG dinucleotides are used to locate splice sites. The est2genome program is good at computing cross-species alignments.

The DDS program identifies regions of the genomic sequence that are similar to a cDNA sequence in the database, on the basis of step 1 of the general strategy described above. The GAP2 program computes an alignment for each pair of genomic region and cDNA sequence by using a linear-space dynamic programming algorithm based on the GT-AG dinucleotides. The DDS/GAP2 program is suitable for computing cross-species alignments.

The sim4 program works in four steps. In step 1, perfect word matches of length 12 between genomic and cDNA sequences are extended into HSPs. In step 2, HSPs are combined into chains. Each exon core is represented by a set of close HSPs. In step 3, the boundaries of exon regions in the genomic and cDNA sequences are determined by a speedy method based on sequence similarity and the GT-AG dinucleotides. In step 4, for each pair of exon regions of the genomic and cDNA sequences, an alignment of the exon regions is produced by a fast method (Chao *et al*., 1997). A main strength of sim4 is its ability to perform cross-species comparisons at a great speed.

In the GeneSeqer program, pairs of similar genomic regions and cDNA sequences are found by extending word matches of length 12 into HSPs and combining HSPs into chains. Then, for each pair of similar genomic region and cDNA sequence, an optimal alignment is computed by scoring for both sequence similarity and potential splice site strength. The strengths of splice sites are scored by using a statistical method (Brendel and Kleffe, 1998) instead of the GT-AG dinucleotides. A main strength of GeneSeqer is its ability to find short exons on the basis of their splice site strength.

The Spidey program finds exon–intron boundaries in a genomic sequence by repeatedly performing fast searches with BLAST. Initially, HSPs between the genomic sequence and cDNA sequences are computed by BLAST at a high stringency ($e = 10^{-6}$). Then, regions of the genomic sequence are located such that all HSPs in a region are nonoverlapping and linearly consistent. Next, for each genomic region, HSPs between the genomic region and the cDNA sequence are computed

by BLAST at a lower stringency ($e = 10^{-3}$). For each gap between HSPs, additional HSPs are computed by BLAST at a low stringency ($e = 1$). Close HSPs are merged into alignments such that each alignment corresponds to one exon. Finally, potential splice sites around the ends of the alignments are scored and selected. A main strength of Spidey is that its performance is not affected by wide variations in intron size.

In the BLAT program, a lookup table is constructed for all nonoverlapping words of length k in the set of genomic sequences, where k is an integer between 8 and 16. Then, the set of cDNA sequences is scanned to locate perfect or near perfect word matches between genomic and cDNA sequences. If there are a sufficient number of word matches between a genomic sequence and a cDNA sequence, then those word matches are extended into HSPs. Close HSPs are linked together into an alignment by using a fast method based on sequence similarity and the GT-AG dinucleotides. BLAT is also able to speed up the comparison by using many processors. A main strength of BLAT is its ability to handle huge sets of genomic and cDNA sequences from the same species.

Four existing programs, DDS/GAP2, est2genome, sim4, and GeneSeqer, were used and compared by Haas *et al*. (2002) in a rigorous reannotation effort to map 5016 full-length cDNA sequences from *Arabidopsis* to its genome. The four programs produced identical splice sites for 4918 of the 5016 cDNA sequences; the programs disagreed on less than 2% of the cDNA sequences. However, the programs are different in their strengths. DDS/GAP2 is better at matching the entire cDNA to the genome than any of the other programs. GeneSeqer is excellent at finding short exons of 3–25 bp. SIM4 is much faster than any of the other programs.

5. Future developments

We suggest three issues for future studies: identification of short exons, construction of a weight array matrix for noncanonical splice sites, and comparison of cross-species sequences. Short exons are difficult to find because short exon matches are much less significant than long exon matches in assessment of exons. This issue may be addressed through development of a special method to put more weight on splice site strength. As more noncanonical splice sites are known, it may be possible to construct a weight array matrix for noncanonical splice sites. If the matrix is useful for finding noncanonical splice sites, then the matrix can be incorporated in existing alignment algorithms. The full matrix versions of dynamic programming algorithms are sensitive in computing cross-species alignments but are very slow for large-scale applications. It may be possible to improve the efficiency of those algorithms without a significant loss in sensitivity.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Brendel V and Kleffe J (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. *Nucleic Acids Research*, **26**, 4748–4757.

Chao K-M and Miller W (1995) Linear-space algorithms that build local alignments from fragments. *Algorithmica*, **13**, 106–134.

Chao K-M, Zhang J, Ostell J and Miller W (1997) A tool for aligning very similar sequences. *Computer Applications in the Biosciences*, **13**, 75–80.

Florea L, Hartzell G, Zhang Z, Rubin GM and Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, **8**, 967–974.

Gotoh O (1982) An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**, 705–708.

Gelfand MS, Mironov AA and Pevzner PA (1996) Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 9061–9066.

Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O and Salzberg SL (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biology*, **3**, 1–12.

Hirschberg DS (1975) A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, **18**, 341–343.

Huang X (1996) Fast comparison of a DNA sequence with a protein sequence database. *Microbial & Comparative Genomics*, **1**, 281–291.

Huang X, Adams MD, Zhou H and Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. *Genomics*, **46**, 37–45.

Huang X and Zhang J (1996) Methods for comparing a DNA sequence with a protein sequence. *Computer Applications in the Biosciences*, **12**, 497–506.

Kent WJ (2002) BLAT – The BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Mott R (1997) EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, **13**, 477–478.

Myers EW and Miller W (1988) Optimal alignments in linear space. *Computer Applications in the Biosciences*, **4**, 11–17.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Pearson WR and Lipman D (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Salzberg S (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Computer Applications in the Biosciences*, **13**, 365–376.

Sellers PH (1974) On the theory and computation of evolutionary distances. *SIAM Journal on Applied Mathematics*, **26**, 787–793.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Usuka J and Brendel V (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring. *Journal of Molecular Biology*, **297**, 1075–1085.

Usuka J, Zhu W and Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

Wagner RA and Fischer MJ (1974) The string-to-string correction problem. *Journal of the ACM*, **21**, 168–173.

Waterman MS, Smith TF and Beyer WA (1976) Some biological sequence metrics. *Advances in Mathematics*, **20**, 367–387.

Wheelan SJ, Church DM and Ostell JM (2001) Spidey: A tool for mRNA-to-genomic alignments. *Genome Research*, **11**, 1952–1957.

Wilbur WJ and Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*, **80**, 726–730.

Zhang MQ and Marr T (1993) A weight array method for splicing signal analysis. *Computer Applications in the Biosciences*, **9**, 499–509.

Specialist Review

Searching for genes and biologically related signals in DNA sequences

Mihaela Pertea
The Institute for Genomic Research, Rockville, MD, USA

1. Introduction

Complex software systems integrating sophisticated machine learning techniques have been developed to elucidate the structures and functions of genes, but accurate gene annotation is still difficult to achieve, primarily due to the complexity inherent in biological systems. If the biological signals surrounding coding exons could be detected perfectly, then the protein-coding regions of all genes could be identified simply by removing the introns, concatenating all the exons, and reading off the protein sequence from start to stop. Not surprisingly, most successful gene-finding algorithms integrate signal sensors that are intended to identify the usually short sequence patterns in the genomic DNA that correspond directly to regions on mRNA or pre-mRNA that have an important role in splicing or translation. Unfortunately, DNA sequence signals have low information content and are very hard to distinguish from nonfunctional patterns, making signal detection a challenging classification problem.

This review provides mainly background information about signal prediction in the gene-finding framework from a computational perspective, without giving too many biological details (for which the interested reader is invited to look at the cited papers). It starts by presenting the most frequently predicted biological signals and ends with a general overview of the most widely used methods in signal detection.

2. Commonly predicted biological signals

The signals that are usually modeled by gene-finding packages are transcription promoter elements, translational start and stop codons, splice sites, and polyadenylation sites, but computational methods have been designed for many other signals, including branch points, ribosome-binding sites, transcription terminators, exon enhancers, and various transcription factor binding sites. In this section, we present some of the biological signals that are most often addressed by signal recognition programs. These signals are shown in Figure 1 in rapport to the structure of a gene.

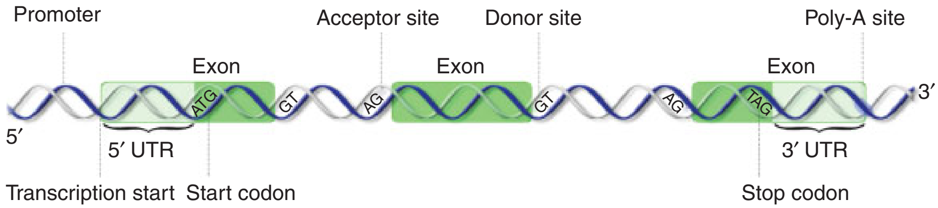


Figure 1 Important biological signals usually modeled by gene finders

2.1. Splice sites

Splice sites are one class of signals that determine the boundaries between exons and introns. The biological rules that govern splice site selection are not completely understood (Reed, 2000). It is well known, though, that recognition of the splicing signals in pre-mRNA is carried out by a complex of proteins and small nuclear RNAs known collectively as the spliceosome, which splices out the introns and produces the mature mRNA transcript. The components of the spliceosome are, in general, highly conserved among eukaryotes (Mount and Salz, 2000), but the splice site consensus sequences are short and degenerate. Usually, the 5′ boundary or donor site of introns in most eukaryotes contains the dinucleotide GT (GU in pre-mRNA), while the 3′ boundary or acceptor site contains the dinucleotide AG. In addition to these dimers, a pyrimidine-rich region precedes the AG at the acceptor site, a shorter consensus follows the GT at the donor site, and a very weak consensus sequence appears at the branch point, approximately 30 nucleotides (nt) upstream from the acceptor site.

A number of computational methods have been developed to identify splice sites. Some are stand-alone splice site finders (see for instance Hebsgaard *et al*., 1996; Reese *et al*., 1997; Brendel and Kleffe, 1998; Pertea *et al*., 2001 for a few examples), but often gene finders identify splice sites as a subroutine, and their performance is greatly influenced by the quality of these splice sensors.

2.2. Translational start and stop codons

Since translational start and stop codons, together with the splice sites, determine the complete coding portion of a gene, their recognition is crucial to the success of computational gene-finding strategies. Unfortunately, patterns in the vicinity of the translational start and stop sites are even less distinctive than those around the splice sites. While the translational stop is almost always identified by the first in-frame stop codon, computational recognition of the start codon proves to be a difficult task in gene finding. Shine-Dalgarno (Shine and Dalgarno, 1974) and Kozak sequences (Kozak, 2001) describe consensus motifs in the context of the translation initiation both in bacteria and in eukaryotes. However, these consensus patterns have very limited power in discriminating the true start codons from other potential starts in a DNA sequence. The lack of strong identifying signals means

that start signal recognition usually has very high false-positive rates, even higher than in the case of splice site detection (Salzberg, 1997; Pertea and Salzberg, 2002).

2.3. Promoters and transcription start sites

Finding the DNA site to which RNA polymerase will bind and initiate transcription is such a complicated task that most gene finders do not even attempt to integrate it into the prediction model. A gene recognition program will usually predict only the protein coding region of the gene, from start to stop codon, leaving the untranslated exons and exon segments to be discovered by other methods. These methods frequently focus on the discovery of the transcription start site (TSS) of a gene, by trying to identify the promoter region, which is usually located adjacent to the TSS.

The promoter region typically extends 150–300 bp upstream from the transcription start site (in prokaryotes), and contains binding sites for RNA polymerase and a number of proteins that regulate the rate of transcription of the adjacent gene. Most bacterial promoters contain the consensus sequences TATAAT and TTGACA at around 10 bp and 35 bp respectively upstream of the TSS (Galas *et al*., 1985). Much less is known about eukaryote promoters. Most computational methods try to recognize the Goldberg–Hogness or TATA box that is centered around 25 bp upstream of the TSS and has the consensus sequence TATAAAA. A CAAT box and a GC-rich element have also been identified in several promoters (Bucher, 1990). Although many promoter and TSS prediction programs have been developed, their performance is less than satisfactory, due to the large numbers of false-positives that are generated (see for instance Fickett and Hatzigeorgiou, 1997; Ohler and Niemann, 2001 for an assessment of recent promoter recognition programs).

2.4. Transcription terminators and polyadenylation sites

At the end of the coding sequences, a signal called the transcription terminator causes the RNA polymerase to slow down and fall off the nascent transcript. It is composed of a sequence that is rich in the bases G and C, and can form a hairpin loop. Prediction of transcription terminators has been achieved with some success in bacterial genomes. Specificity in the 89–98% range was obtained by TransTerm (Ermolaeva *et al*., 2000), which identifies terminators by searching for a hairpin structure in proximity of a downstream U-rich region.

As transcription is nearly completed, a polyadenylation signal where the mRNA is cut is encountered. A polyadenylic acid sequence of varying length is then added posttranscriptionally to the primary transcript as part of the nuclear processing of RNA. This poly-A tail is thought to help transport the mRNA out of the nucleus, and may stabilize the mRNA against degradation in the cytoplasm. Prediction of the location of the poly-A signal is included in many gene-recognition programs, since careful determination of the gene's ends plays an important role in determining the accuracy of the programs. When the end of a gene is not correctly predicted, it tends to be either truncated or fused with the next gene by the gene finder. Poly-A signal prediction has been also considered by stand-alone programs (Salamov

and Solovyev, 1997; Tabaska and Zhang, 1999) that can predict the well-known AAUAAA signal as well as other poly-A sites.

3. Computational methods for signal prediction

In this section, we describe several of the most popular methods for signal detection in DNA sequences, but many others have appeared over the years. These methods range from very simple ones, such as computation of consensus sequences and position weight matrices, to more complex computational methods such as neural networks, Markov models, and decision trees. In general, the best current methods combine many different types of evidence (Salzberg *et al*., 1998), using nucleotides in a relatively wide window around the core sequence pattern. Because all of these methods are essentially statistical, they have become more effective recently, due not to changes in the algorithms but rather to significant increases in the amount of genomic sequence data available.

3.1. Consensus analysis

Consensus sequence methods use a multiple alignment of functionally related sequences to identify the consensus, defined as a composite subsequence created using the most common base at each position in the multiple alignment. New sites are identified simply by matching them to the consensus sequence. The match can be exact or it can be within a neighborhood consisting of all subsequences within a fixed distance k from the consensus sequence, where the most common distance used between two sequences of the same length is the Hamming distance. These techniques have been used to characterize a variety of different biological signals; for example, the consensus (A,C)AG/GU(A,G)AGU with mismatches was described at the 5′ end of introns back in 1990 (Senapathy *et al*., 1990). Different programs implement this technique either by itself or combined with other methods to predict functional sites (e.g., SpliceView (Rogozin and Milanesi, 1997) or SplicePredictor (Brendel and Kleffe, 1998)).

3.2. Weight matrices

The weight matrix model (WMM) summarizes the consensus sequence information by specifying the probabilities of each of the four nucleotides in each position of the DNA sequence (Stormo, 1990; Gelfand, 1995). Values in the weight matrix are determined by aligning known sequences for a given signal type, and computing the fraction of each residue at each position. Formally, if the signal pattern is ℓ nucleotides in length, then the probability that the sequence $X = x_1x_2 \ldots x_\ell$ is a signal of that type is given by $p(X) = \prod_{i=1}^{\ell} p^{(i)}(x_i)$, where $p^{(i)}(x_i)$ is the probability of generating nucleotide x_i at position i of the signal. Fickett (1996) describes a common variation of the above technique, the position weight matrix, which uses the relative frequency of each base in each position by

replacing $p^{(i)}(x_i)$ with $p^{(i)}(x_i)/p(x_i)$, where $p(x_i)$ is the probability of generating nucleotide x_i independent of its position in the signal. An important limitation of the weight matrices is the underlying simplifying assumption of independence between positions. A modified form of the WMM known as "inhomogeneous Markov model" or "weight array model" (IMM/WAM) has been introduced to capture nonindependence between positions (Zhang and Marr, 1993). Most commonly used is the first-order WAM in which the distribution at position i depends on the residue at position $i - 1$, but higher-order WAMs have been also described. While such higher-order models might succeed in reducing the high false-positive rates, they require the estimation of a very large number of parameters, thus an increased amount of training data, which is often not available. To reduce the problem of sample-size limitation, Burge and Karlin (1997) implement a generalization of the WAM, called "windowed weight array model" (WWAM), to detect splice site signals. In this method, the transition probabilities at a given position i are estimated by averaging the conditional probabilities observed at position i with those observed in a neighboring window of positions centered at position i. For instance, a second-order WWAM will estimate the transition probabilities for positions i from the data at positions $i-2, i-1, i, i+1$, and $i+2$, the underlying assumption here being that nearby positions have similar triplet biases.

3.3. Decision trees

A decision tree is a supervised learning method that learns to classify objects from a set of examples (where an "object" can be a DNA sequence or partial sequence). It takes the form of a tree structure of nodes and edges in which the root and the internal nodes test on one of the objects' attributes. Each edge extending from an internal node of the tree represents one of the possible alternatives or courses of action available at that point. Depending on the outcome of the test, different paths in the tree are followed down to the tree's leaves, which carry the names of classes into which the objects are placed. Most decision tree learning systems use an information-theoretic approach that minimizes the number of tests to classify an object. A typical approach is to use an information-based heuristic that selects the attribute providing the highest information gain (Quinlan, 1986), that is, the attribute that minimizes the information needed in the resulting subtrees to classify the elements. Formally, the information gain uses $I = \sum (p_i \ln p_i)$, where p_i is the probability that an object is in class i, as evaluation function for measuring the information of the initial set and the subsets produced after splitting. There have been many other approaches based on different evaluation functions, including the Gini index, chi-square tests, and other metrics (Murthy *et al*., 1994).

Decision tree–based classification methods have applications in a wide variety of domains, including DNA sequence analysis. SpliceProximalCheck (Thanaraj and Robinson, 2000), for example, is a program that uses a decision tree to discriminate true splice sites from false ones that may be located physically close to real ones.

An interesting application of decision trees was introduced by (Burge and Karlin (1997) to model complex dependencies in signals. Their approach, called maximal dependence decomposition (MDD), tries to capture the most significant adjacent

and nonadjacent dependencies in a signal using an iterative subdivision of the training data. MDD trees are constructed from a set D of N aligned DNA sequences of length k, extracted from a set of signal sites. For each of the k positions, let the most frequent base at that position be the consensus base. The variable C_i will be 1 if the nucleotide at position i matches the consensus at position i, 0 otherwise. Next, compute the χ^2 statistics between the variables C_i and X_j (which identifies the nucleotide at position j), for each i, j pair with $i \neq j$. If strong dependencies are detected, then proceed as in Burge and Karlin (1997) by partitioning the data into two subsets based on those positions. Recursive application of this procedure builds a small tree. Each leaf of this tree contains a subset of the signal sites used for training.

3.4. Markov models

Markov models have been in use for decades as a way of modeling sequences of events, and they translate very naturally to DNA sequence data. To score a sequence using a Markov chain, a gene finder needs to compute a set of probabilities from training data. These probabilities take the form $P(b_i|b_{i-1}, b_{i-2}, \ldots)$, where b_i indicates the base in position i of the DNA sequence. Models that use Markov chains rely on the *Markov assumption*, which states that any event x is dependent only on a fixed number of prior events. Clearly, this assumption is not true for all DNA sequences, and the extent to which the Markov assumption is false defines a limitation on the model. Markov chains combined with the MDD method are implemented by the splice site predictor program GeneSplicer (Pertea *et al*., 2001).

Many gene finders such as Genscan (Burge and Karlin, 1997) or GeneMark. HMM (Lukashin and Borodovsky, 1998) model signals (e.g., promoters, poly-A sites, and different enhancers) by using hidden Markov models (HMMs) or their extended form, generalized hidden Markov models (GHMMs). HMMs are Markov chains in which the states are not directly observable, and only the output of a state is observable. HMMs are generally far more complex than Markov chains, because the model need not be a chain; rather, it can contain elaborate loops and branching structures. The output from (or input to) each state is a symbol (or a string of symbols in the case of GHMMs) from a finite set of symbols and is chosen randomly according to some probability distribution over the alphabet of the output symbols. The parameters of the HMM are estimated using an algorithm called the Expectation Maximization (E-M) algorithm, and the most probable interpretation of a DNA sequence is then found with the Viterbi algorithm (Eddy, 1996).

3.5. Neural networks

Neural networks (NNs) are a powerful tool for pattern classification. Some NNs are models of biological neural networks while others are not, but historically, much of the inspiration for the field of NNs came from the desire to produce artificial systems capable of sophisticated computations similar to those that the human brain

routinely performs. There is no room here to review the extensive and complex literature on neural nets, but generally speaking, NNs are networks of many simple processors, each possibly having a small amount of local memory. These networks are usually connected in a layered, feed-forward architecture and have some sort of training rule whereby the weights of connections are adjusted on the basis of data.

Many types of biological signals have been predicted using NNs, including the following: Promoter 2.0 (Knudsen, 1999) and NNPP (Reese, 2001) for promoter prediction, NNSPLICE (Reese *et al*., 1997) and NetGene2 (Hebsgaard *et al*., 1996) for splice site recognition, and NetStart (Pedersen and Nielsen, 1997) for start codon determination.

3.6. Linear and quadratic discriminant analysis

In linear discriminate analysis (LDA), the examples to be classified are represented as points in an n-dimensional space defined by features, each of which represents a measurement of some property of the input. The main purpose of LDA is to predict class membership on the basis of a linear combination of the scores of individual features. This can be represented as a multidimensional hyper-plane (a decision surface) whose coefficients are determined by minimizing the prediction error using positive and negative training data sets. In the field of signal detection, LDA is implemented, for instance, in PromH (Solovyev and Shahmuradov, 2003), a promoter prediction program that uses linear discriminant functions that take into account conserved features and nucleotide sequences of promoter regions in pairs of orthologous genes. Quadratic discriminant analysis (QDA) generalizes LDA by finding an optimal quadratic surface instead. The promoter prediction program CorePromoter (Zhang, 1998) uses QDA computed from 5-tuples in the input DNA sequence to predict promoters.

4. Conclusions and future challenges

The accuracy of gene-recognition programs is relatively high (around 90% sensitivity and specificity) when computed merely on the basis of the number of correctly predicted coding nucleotides, but decreases significantly when considering the number of exons or genes predicted correctly. Determining the exact boundaries of a gene's structural elements is highly influenced by the ability to predict correctly the signals surrounding the DNA coding regions.

Prediction of very small exons or shifted open-reading frames could benefit from better recognition of splicing regulatory patterns. Computational determination of alternative splicing of genes can also benefit from a better understanding of the rules that govern splice site selection (Black, 2000). Alternative splicing appears to be most common in vertebrates (e.g., at least 35% of the genes in the human genome have been shown to be alternatively spliced (Croft *et al*., 2000)), but other metazoan organisms and plants also experience a significant number of alternatively spliced genes (Brown and Simpson, 1998), a phenomenon that raises serious challenges for

the current software programs for gene recognition. A dominant role in the regulation of alternative splicing is played by exon splicing enhancers (ESEs), which activate nearby splice sites and promote inclusion of exons in which they reside. ESEs have only recently begun to be included in computational models (Fairbrother *et al*., 2002; Cartegni *et al*., 2003).

Signals upstream and downstream of genes also have an important role in determining the exact locations of genes. Using currently available methods, it is easier to find the 3′ end of a gene than the 5′ end. More effort needs to be devoted to improving detection of the patterns preceding genes. Numerous TSS and promoter databases (Zhu and Zhang, 1999; Hershberg *et al*., 2001; Shahmuradov *et al*., 2003; Schmid *et al*., 2004; Suzuki *et al*., 2004) have been constructed (though their reliability is highly variable), which can help with the determination of the 5′ end of the genes in various organisms. As more complete genomic sequences and extensive experimental data become available, systematic computational and experimental approaches can be further exploited to increase our ability to find genes and regulatory signals in all organisms.

Acknowledgments

This work was supported in part by the US National Institutes of Health under Grant R01-LM06845.

References

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**(3), 367–370.

Brendel V and Kleffe J (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Research*, **26**(20), 4748–4757.

Brown JWS and Simpson CG (1998) Splice site selection in plant pre-mRNA splicing. *Annual Review of Plant Physiology and Plant Molecular Biology*, **49**, 77–95.

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, **212**(4), 563–578.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**(1), 78–94.

Cartegni L, Wang J, Zhu Z, Zhang MQ and Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, **31**(13), 3568–3571.

Croft L, Schandorff S, Clark F, Burrage K, Arctander P and Mattick JS (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genetics*, **24**(4), 340–341.

Eddy SR (1996) Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3), 361–365.

Ermolaeva MD, Khalak HG, White O, Smith HO and Salzberg SL (2000) Prediction of transcription terminators in bacterial genomes. *Journal of Molecular Biology*, **301**(1), 27–33.

Fairbrother WG, Yeh RF, Sharp PA and Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**(5583), 1007–1013.

Fickett JW (1996) The gene identification problem – an overview for developers. *Computers and Chemistry*, **20**(1), 103–118.

Fickett JW and Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Research*, **7**(9), 861–878.

Galas DJ, Eggert M and Waterman MS (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *Journal of Molecular Biology*, **186**, 117–128.

Gelfand MS (1995) Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, **2**(1), 87–115.

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P and Brunak S (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Research*, **24**(17), 3439–3452.

Hershberg R, Bejerano G, Santos-Zavaleta A and Margalit H (2001) PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Research*, **29**(1), 277.

Knudsen S (1999) Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics*, **15**, 356–361.

Kozak M (2001) A progress report on translational control in eukaryotes. *Science's STKE*, **1**(71), PE1.

Lukashin AV and Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research*, **26**(4), 1107–1115.

Mount SM and Salz HK (2000) Pre-messenger RNA processing factors in the *Drosophila* genome. *The Journal of Cell Biology*, **150**, F37–F44.

Murthy SK, Kasif S and Salzberg S (1994) A System for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, **2**(1), 1–32.

Ohler U and Niemann H (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics*, **17**(2), 56–60.

Pedersen AG and Nielsen H (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, **5**, 226–233.

Pertea M, Lin X and Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research*, **29**(5), 1185–1190.

Pertea M and Salzberg SL (2002) A method to improve the performance of translation start site detection and its application for gene finding. *Lecture Notes in Computer Science*, **2452**, 210–219.

Quinlan JR (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

Reed R (2000) Mechanisms of fidelity in pre-mRNA splicing. *Current Opinion in Cell Biology*, **12**, 340–345.

Reese MG (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Computers & Chemistry*, **26**(1), 51–56.

Reese MG, Eeckman FH, Kulp D and Haussler D (1997) Improved splice site detection in Genie. *Journal of Computational Biology*, **4**(3), 311–323.

Rogozin IB and Milanesi L (1997) Analysis of donor splice signals in different organisms. *Journal of Molecular Evolution*, **45**, 50–59.

Salamov AA and Solovyev VV (1997) Recognition of 3′-processing sites of human mRNA precursors. *Computer Applications in the Biosciences*, **13**(1), 23–28.

Salzberg SL, Delcher AL, Fasman K and Henderson J (1998) A decision tree system for finding genes in DNA. *Journal of Computational Biology*, **5**(4), 667–680.

Salzberg SL (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Computer Applications in the Biosciences*, **13**(4), 365–376.

Schmid CD, Praz V, Delorenzi M, Périer R and Bucher P (2004) The Eukaryotic promoter database EPD: the impact of in silico primer extension. *Nucleic Acids Research*, **32**, D82–D85.

Senapathy P, Shapiro MB and Harris NL (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods in Enzymology*, **183**, 252–278.

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM and Solovyev VV (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Research*, **31**(1), 114–117.

Shine J and Dalgarno L (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences of the United States of America*, **71**(4), 1342–1346.

Solovyev VV and Shahmuradov IA (2003) PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Research*, **31**(13), 3540–3545.

Stormo GD (1990) Consensus patterns in DNA. *Methods in Enzymology*, **183**, 211–221.

Suzuki Y, Yamashita R, Sugano S and Nakai K (2004) DBTSS, DataBase of transcriptional start sites: progress report 2004. *Nucleic Acids Research*, **32**, D78–D81.

Tabaska JE and Zhang MQ (1999) Detection of polyadenylation signals in human DNA sequences. *Gene*, **231**(1–2), 77–86.

Thanaraj TA and Robinson AJ (2000) Prediction of exact boundaries of exons. *Briefings in Bioinformatics*, **1**(4), 343–356.

Zhang MQ (1998) Identification of human gene core promoters in silico. *Genome Research*, **8**(3), 319–326.

Zhang MQ and Marr TG (1993) A weight array method for splicing signal analysis. *Computational Applied Bioscience*, **9**(5), 499–509.

Zhu J and Zhang MQ (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**(7–8), 607–611.

Specialist Review

Pair hidden Markov models

Marina Alexandersson, Nicolas Bray and Lior Pachter
University of California Berkeley, Berkeley, CA, USA

1. Introduction

Many of the early contributions of computer science to biological sequence analysis consisted of the development of algorithms for pairwise sequence alignment, most notably the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), which was later extended and refined by Smith, Waterman (Waterman and Smith, 1981), and others. It was only 20 years later, when probabilistic models began to be developed for related problems in sequence analysis that it was first suggested that some of the classical algorithms might have a probabilistic interpretation (Bucher and Hofmann, 1996). This observation has led to numerous developments: more rigorous approaches to parameter estimation (Liu *et al*., 1995), novel annotation systems based on combinations of alignment and gene finding models (Alexandersson *et al*., 2003), and Bayesian sampling methods for alignments (Liu *et al*., 1995) to name a few.

The probabilistic interpretation of alignment is based on modeling sequences using *pair hidden Markov models* (PHMMs). Our goal in this chapter is to explain these models, and their connection to hidden Markov models (HMMs) and phylogenetic models of evolution. These models are now widely used in computational biology, and a complete understanding of their foundations is essential to their proper application and for the development of novel more powerful models that will capture the complexities of sequence evolution. The chapter is organized as follows: we begin by explaining evolutionary models for single bases. This is the beginning of a rich theory, and although we avoid exploring the topic in depth, we point out the crucial nonidentifiability property for two taxa models. This property plays a key role in pairwise sequence alignment. Next, we describe HMMs, which capture positional dependence in biological sequences, and are therefore relevant for PHMMs. We then turn to a complete description of PHMMs, which are instances of *Bayesian multinets*. These are graphical models in which the structure of the graph can depend on the assignments to nodes. This distinction places PHMMs in a separate category from standard HMMs. In fact, the term PHMM suggests that PHMMs are close relatives of HMMs, which is misleading; PHMMs are significantly more complex than HMMs. For biologists, the relevance of PHMMs stems from the fact that they can be used to infer alignments of

sequences. The standard alignment algorithms, such as the Needleman–Wunsch algorithm and its variants, can be interpreted as Viterbi algorithms for PHMMs. This is explained in Section 4 in which we also show that PHMMs can be used to sample alignments efficiently. We conclude by describing biologically relevant extensions of PHMMs.

2. What is an alignment?

PHMMs are mainly used to find alignments between pairs of sequences, so we begin by explaining exactly what we mean by *pairwise alignment*.

Definition 1. A pairwise alignment between two sequences σ^1 and σ^2 from a finite alphabet is a sequence on the alphabet H, I, D that has the property that $\#H + \#I = |\sigma^1|$ and $\#H + \#D = |\sigma^2|$. By $\#H$, $\#I$, $\#D$ we mean the number of occurrences of those characters in the alignment string.

An alignment captures the relationship between two sequences that have evolved from a common ancestor via a series of mutations of bases, represented by H that stands for homology, I (insertions), or D (deletions). The alignment string records the events that led to the present day sequences. Alternatively, alignment can be viewed as an *edit* of σ^1 to produce σ^2. That is, we allow characters from σ^1 to be changed or deleted, or new characters to be inserted. An alignment can be thought of, equivalently, as a record of the edits from σ^1 to σ^2. We make this statement precise in the following section.

The main problem in sequence alignment is to identify good alignments that explain the evolutionary history on the basis of parameters for mutation, insertion, and deletion. Mutation mechanisms are accounted for in PHMMs with the use of evolutionary models. An underlying HMM structure is used to model local sequence characteristics, or insertion/deletion structures. The challenging computational task is to find good alignments among the set of all possible alignments. In particular, if $\mathcal{A}_{n,m}$ is the set of all alignments for two strings of length n and m, the number of elements of $\mathcal{A}_{n,m}$ is the Delannoy number

$$D(n, m) = \sum_i \binom{n}{i} \binom{n + m - i}{m - i} \tag{1}$$

This number is exponential in the lengths of the sequences, leading to a nontrivial computational problem in identifying good alignments.

3. Models

3.1. Evolutionary models for two bases

PHMMs capture point mutations, as well as insertions and deletions in biological sequences. We begin by carefully analyzing single base evolutionary models that

are at the core of PHMMs. We restrict ourselves to DNA. Single base mutation is modeled by continuous time Markov chains:

Definition 2. A *rate matrix* (or *Q-matrix*) is a 4×4 matrix $Q = (q_{ij})$ (with rows and columns indexed by the nucleotides) satisfying the properties

$$\begin{aligned} q_{ij} &\geq 0 \quad \text{for} \quad i \neq j \\ \sum_{j \in \{A,C,G,T\}} q_{ij} &= 0 \quad \text{for all} \quad i \in \{A, C, G, T\} \\ q_{ii} &< 0 \quad \text{for all} \quad i \in \{A, C, G, T\} \end{aligned} \tag{2}$$

Rate matrices capture the notion of *instantaneous rate of mutation* and allow us to calculate the *substitution matrices* $P(t)$, giving the probability that a given base will mutate from one nucleotide to another over a given period of time. The entry of $P(t)$ in row i and column j equals the probability that the substitution $i \to \cdots \to j$ occurs in a time interval of length t. We use the notation $P(i, j;t)$ to denote such an entry. Given a rate matrix Q, these probabilities will be governed by a matrix of differential equations whose solution is provided by the following theorem:

Theorem 3. *Let Q be any rate matrix and $P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i t^i$. Then*

1. $P(s + t) = P(s)P(t)$,
2. *$P(t)$ is the unique solution to $P'(t) = P(t) \cdot Q$, $P(0) = \mathbf{1}$ for $t \geq 0$,*
3. *$P(t)$ is the unique solution to $P'(t) = Q \cdot P(t)$, $P(0) = \mathbf{1}$ for $t \geq 0$.*

Furthermore, a matrix Q is a rate matrix if and only if the matrix $P(t) = e^{Qt}$ is a stochastic matrix (nonnegative with row sums equal to one) for every t.

By fixing the entries in a rate matrix and specifying some time t^*, we obtain a matrix $P(t^*)$ that completely specifies the probabilities for observing the four nucleotides after time t^*, starting with each of the nucleotides A, C, G, or T. One key feature of the evolutionary models that we will consider is *reversibility*. In the case in which the background base frequencies are uniform, this property is equivalent to requiring that the substitution matrices are symmetric.

Substitution matrices can be combined along edges of a tree to form a *phylogenetic model*. We consider only the simplest case consisting of an ancestral base r, which after a speciation event mutates into a base a that we observe after time t_1 and a separate base b that we observe after time t_2. If $P(i, j; t_1, t_2) = Pr(a = i, b = j; t_1, t_2)$ is the probability of observing i and j at a and b after times t_1 and t_2 respectively, then we can write the probability

$$P(i, j; t_1, t_2) = \sum_k Pr(r = k)P(i, k; t_1)P(j, k; t_2) = \frac{1}{4} \sum_k P(i, k; t_1)P(j, k; t_2) \tag{3}$$

where, again, we have assumed a uniform distribution at r. If we include the assumption that $P(t)$ is symmetric for all t, then this completely specifies a statistical model that we call the *general reversible model for two taxa*. The key property of reversible models that we will need for PHMMs is the nonidentifiability of the two taxa general reversible model:

Proposition 4. If $t_1 + t_2 = t_3 + t_4$ then $P(i, j; t_1, t_2) = P(i, j; t_3, t_4)$.

Proof: Since $P(t)$ is symmetric for all t, $P(i, j; t) = P(j, i; t)$ for all i and j. Now as above, we have that

$$P(i, j; t_1, t_2) = \frac{1}{4}\sum_k P(i, k; t_1)P(j, k; t_2) = \frac{1}{4}\sum_k P(i, k; t_1)P(k, j; t_2) \qquad (4)$$

However, the right-hand side of equation (4) is just $\frac{1}{4}$ times the (i, j)th entry of the matrix $P(t_1)P(t_2) = P(t_1 + t_2)$. Similarly, we will have that $P(i, j; t_3, t_4)$ will be $\frac{1}{4}$ times the (i, j)th entry of the matrix $P(t_3 + t_4)$. The result now follows from the assumption that $t_1 + t_2 = t_3 + t_4$.

This proposition shows that we arrive at the same distribution on pairs of nucleotides if we set one of our branch lengths to 0, in which case the model describes not two bases evolving from a common ancestor but rather of one base changing directly into the other. This fact lies at the heart of our ability to refer to alignment as the editing of one sequence into the other, rather than as a description of the evolutionary history from a common ancestor.

3.2. *Phylogenetic models for two sequences*

In order to describe the evolution of entire sequences, rather than single bases, we make use of *HMMs*. In this section, we use them to describe an evolutionary model for sequences, which allows for point mutation and dependence between neighboring nucleotides, without allowing for insertion or deletion, a complexity that is added in the next section.

Definition 5. A discrete hidden Markov model consists of n observed states $Y_1, \ldots, Y_n$ taking on l possible values, and n hidden states $X_1, \ldots, X_n$ taking on k possible values. The HMM can be characterized by the following conditional independence statements for $i = 1, \ldots, n$:

$$p(X_i \mid X_1, X_2, \ldots, X_{i-1}) = p(X_i \mid X_{i-1}),$$
$$p(Y_i \mid X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) = p(Y_i \mid X_i).$$

We consider the model with uniform initial distribution, where all probabilities $p(X_i|X_{i-1})$ are given by the same $k \times k$ matrix and all probabilities $p(Y_i|X_i)$ are given by the same $l \times k$ matrix.

The relationship of HMMs to the modeling of pairs of sequences that evolved from a common ancestral sequence is as follows: the random variables $Y_1, \ldots,$

Y_n take on $l = 16$ possible values (consisting of all pairs of nucleotides). The hidden random variables $X_1, \ldots, X_n$ take on $k = 4$ possible states corresponding to the four nucleotides A, C, G, T. The hidden values play the role of the ancestral sequence. Fixing the conditional transition and output probabilities in the HMM results in a probability distribution being defined on pairs of sequences. If the output probabilities are constrained to a reversible model, it is clear from the previous discussion that we have described an evolutionary model for sequences that allows for dependence between neighboring nucleotides in the ancestral sequence.

3.3. *Modeling insertion and deletion*

We are now ready to introduce PHMMs, which add insertions and deletions into the model in Section 3.2. Graphically, this model is represented in Figure 1(a). The boxes around the nodes, called *plates*, are used to indicate that the nodes may not be present. The presence of a node is determined by *class nodes*, small circles in the plates that in our case take on the values 0 (node missing) or 1 (node present). The hidden states take on the values H, I_1, I_2 and D_1, D_2. In the H state, an ancestral base is generated, and from it two observed bases. In the D states, an ancestral base is generated but only one observed base. In the I states, no ancestral base is generated, and one observed base is generated.

Alternatively (and equivalently, because of the nonidentifiability discussed in Section 3.1), we can describe a PHMM with the model in figure 1b. In this case, there are three hidden states: H, I and D. In the I state a base is generated in sequence 1, and in the D state a base is generated in sequence 2. In the H state, two bases are generated from a joint distribution (note that the model as drawn technically requires four H states, one for each base, but this is usually omitted in favor of simply writing the joint probability distribution on the observed nodes directly). The model then consists of a distribution on the initial state of the model, $\{\pi_i\}_{i=1}^{k}$, transition probabilities $s_{ij} = Pr(X_{l+1} = j|X_l = i)$, and an output distribution from each state j, $b_j(x, y) = Pr(x, y|X_l = j)$, where x, y range over all possible pairings of bases, allowing for the null base. Note that the fact that the output distribution now includes "null characters" allows for the instances where one or the other observed nodes is not present in the model. For example, $b_I(A, \emptyset)$ is the probability of an A being observed at the first node when the second node is not present in the model.

Given two sequences σ^1 and σ^2 of lengths N and M respectively, this model allows us to assign a probability to any alignment h in $\mathcal{A}_{N,M}$. The alignment h is a sequence of Hs, Is, and Ds and this will be our sequence of hidden states. Adopting the notation of Pachter and Sturmfels (2005), we define $h[i]$ to be $\#H + \#I$ in the prefix $h_1 h_2 \ldots h_i$ of h, and $h\langle i\rangle$ to be $\#H + \#D$ in $h_1 h_2 \ldots h_i$. Thus, $h[i]$ is our position in σ^1 at step i of the alignment and $h\langle i\rangle$ is our position in σ^2. Then the probability of the alignment h according to our model will be

$$\pi_{h_1} b_{h_1}(\sigma^1_{h[1]}, \sigma^2_{h\langle 1\rangle}) \prod_{i=2}^{L} s_{h_{i-1}h_i} b_{h_i}(\sigma^1_{h[i]}, \sigma^2_{h\langle i\rangle}) \tag{5}$$

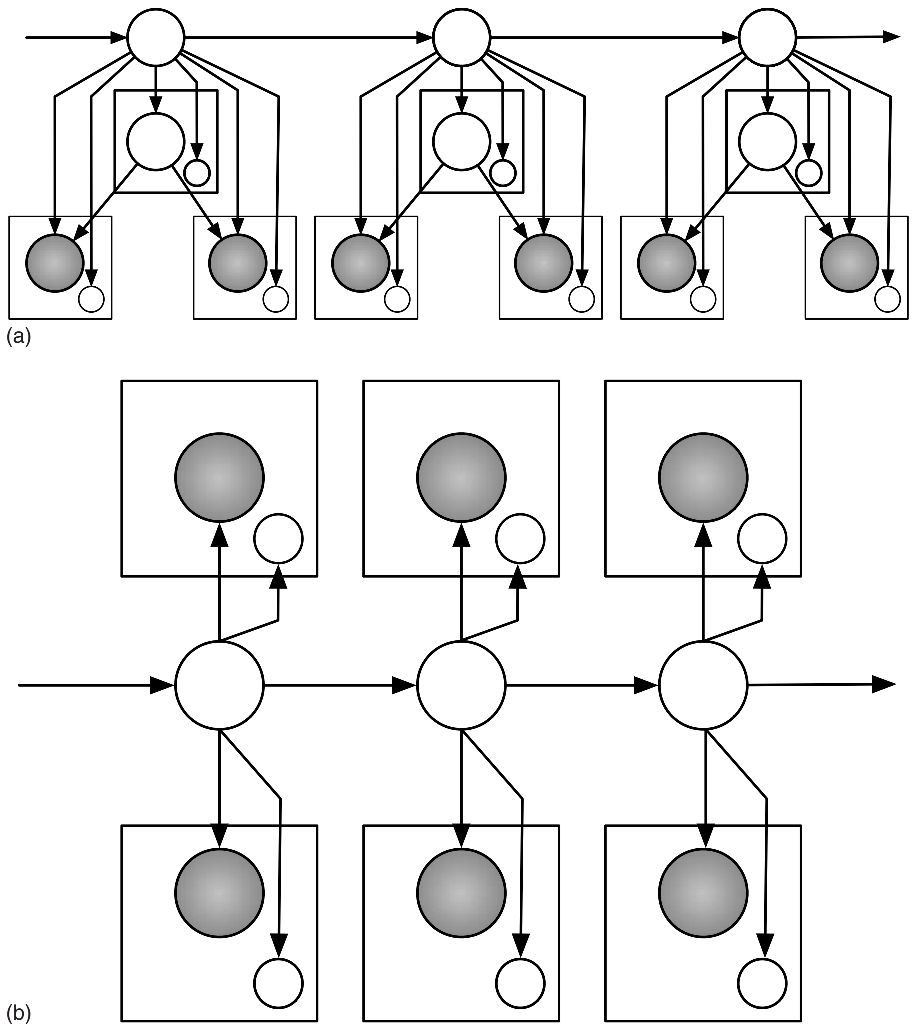


Figure 1 The pair hidden Markov model

From this we can calculate the probability of two sequences:

$$P(\sigma^1, \sigma^2) = \sum_{h \in \mathcal{A}_{N,M}} \pi_{h_1} b_{h_1}(\sigma^1_{h[1]}, \sigma^2_{h\langle 1\rangle}) \prod_{i=2}^{L} s_{h_{i-1}h_i} b_{h_i}(\sigma^1_{h[i]}, \sigma^2_{h\langle i\rangle}) \tag{6}$$

Despite the name, PHMMs are nontrivial modifications of standard HMM models, in that the structure of the models is not fixed. In the graphical model literature, these types of models are referred to as Bayesian multinets (Friedman *et al*., 1997).

4. Algorithms

4.1. Finding a best alignment

The model described in the previous section, for some fixed parameters and fixed observed sequences, provides a probability distribution over alignments. Thus, for a given pair of sequences, one obvious question is: what is the most likely alignment of the two sequences according to the model? Stated formally, the question is to find the *maximum a posteriori* (MAP) alignment. The naive method of answering this question would be to enumerate all possible alignments, calculate the probability of each, and select the one with highest probability. This quickly becomes infeasible as the number of possible alignments of two sequences grows exponentially in their length. However, by exploiting the structure of the model, we can find the optimal alignment in time proportional to the product of the lengths of the sequences using the Viterbi algorithm.

The Viterbi algorithm for PHMMs is related to the Viterbi algorithm for HMMs, which is itself a specific instance of the generalized distributive law for inference in graphical models (Aji and McEliece, 2000). It can also be viewed as an instance of the dynamic programming approach, a general method for exploiting recursive structure in optimization problems. As is typical of dynamic programming methods, the Viterbi algorithm solves what would appear to be a much more difficult problem: instead of finding the optimal alignment for the two given sequences, the optimal alignment is calculated for every pair of prefixes of the sequences. For a sequence σ, let $\sigma\langle i\rangle$ denote the prefix $\sigma_1\sigma_2 \ldots \sigma_i$ and denote the two sequences being compared by σ^1 and σ^2. The key observation of the Viterbi algorithm is that in an alignment of $\sigma^1\langle i\rangle$ and $\sigma^2\langle j\rangle$, the final hidden state must be either H, I, or D and so the alignment decomposes naturally into smaller alignments. This is used to recursively compute a matrix δ where $\delta(i, j, s)$ is the probability of the optimal alignment of $\sigma^1\langle i\rangle$ and $\sigma^2\langle j\rangle$ ending in hidden state s. Having computed this matrix, it is then possible to compute the optimal alignment of the entire sequences.

Formally, let $X_1X_2 \ldots X_L$ be the underlying hidden Markov process assuming values in the state space $S = \{H, I, D\}$, (where $L \le N + M$ is the length of the alignment). The process begins in a state determined by the initial distribution $\{\pi_i\}_{i=1}^{k}$ and evolves through the state space according to the transition probabilities $s_{ij} = Pr(X_{l+1} = j|X_l = i)$. The observed sequences σ^1 and σ^2 are sequences of random steps of the hidden process using some output distribution $b_j(\sigma_n^1, \sigma_m^2) = Pr(\sigma_n^1, \sigma_m^2|X_l = j)$. Moreover, in order for the model to represent a distribution over all possible pairs of sequences and their alignments, we need to introduce two silent states, *Begin* and *End*. We initialize the Viterbi algorithm by

$$\begin{aligned} \delta(1, 1, H) &= \pi_H b_H(\sigma_1^1, \sigma_1^2), \\ \delta(1, 0, I) &= \pi_I b_I(\sigma_1^1, \emptyset), \\ \delta(0, 1, D) &= \pi_D b_D(\emptyset, \sigma_1^2) \end{aligned} \tag{7}$$

and the recursion relations become, for $1 \leq n \leq N$ and $1 \leq m \leq M$,

$$\delta(n, m, H) = \max_j \{\delta(n-1, m-1, j) a_{jH} b_H(\sigma_n^1, \sigma_m^2)\}$$
$$\delta(n, m, I) = \max_j \{\delta(n-1, m, j) a_{jI} b_I(\sigma_n^1, \emptyset)\}$$
$$\delta(n, m, D) = \max_j \{\delta(n, m-1, j) a_{jD} b_D(\emptyset, \sigma^2, m)\} \quad (8)$$

where $\emptyset$ represents a gap and $\delta(n, m, j) = 0$ if $n < 0$ or $m < 0$. The recursion is terminated in the *End* state by $\delta(N+1, M+1, i) = \max_j \{\delta(N, M, j) a_{ji}\}$, and the probability of the most likely state sequence is simply $\max_i \delta(N+1, M+1, i)$, call the maximizing state i^*. The optimal hidden state sequence, given the data, is retrieved in the traceback by starting in i^* and recursively backtracking through $\text{argmax}_i \delta(n-k, m-l, i)$, where $(k, l) \in \{(1, 1), (1, 0), (0, 1)\}$.

Needleman and Wunsch were the first ones to describe an alignment algorithm for two biological sequences using the dynamic programming approach (Needleman and Wunsch, 1970), and although their paper is focused on a very specific problem, it contains the essence of many alignment methods in use today. It is easy to see that by taking logarithms of the parameters we have described, the PHMM Viterbi algorithm becomes a Needleman–Wunsch algorithm with negative scores. If alignments are normalized, some of the scores will be positive (Bucher and Hofmann, 1996). It is also true, that for any set of Needleman and Wunsch scores, a PHMM can be constructed for which the Viterbi algorithm will yield the same alignment as the Needleman–Wunsch algorithm (even for extensions such as affine gap penalties).

There have been countless improvements to the basic algorithm; we point the reader to textbooks such as Durbin *et al*. (1998) and Gusfield (1997) for detailed expositions. One point of note is that there is a linear-space divide and conquer algorithm for pairwise sequence alignment that is crucial for practical applications where memory is a limiting factor in computing large sequence alignments (Hirschberg, 1975).

4.2. Sampling alignments

One of the advantages of using PHMMs for alignment is that operations such as sampling of alignments can be performed in a statistically meaningful way, for example, by sampling from the posterior distribution on the alignments given the observations. This is not only useful for exploring alternatives to the MAP alignment discussed in the previous section but also for utilizing Bayesian methods for parameter estimation. Fortunately, there is an efficient way to sample alignments, which is also memory efficient (Cawley and Pachter, 2003). The algorithm is not well known, and we therefore include it here for completeness.

Consider the alignment of two sequences of lengths N and M, with a simple PHMM that only assigns probabilities for insertions or deletions, which we call w, and has substitution matrices s_{ij} (for observing nucleotides i, j). Let a_{ij} denote the sum of the probabilities of all alignments in which position i in one sequence is

aligned to position j in the other (i.e., the so-called forward variables). Observe that the a_{ij} can be computed using a standard Viterbi-type algorithm with *max* replaced by *sum*. The recursion is

$$a_{ij} = wa_{i-1,j} + wa_{i,j-1} + s_{ij}a_{i-1,j-1} \tag{9}$$

where $i, j \geq 1$ and $a_{0j} = a_{j0} = wj$. Let $v_j = (a_{1j}, a_{2j}, \ldots, a_{Nj})$ be the jth column vector $(0 \leq j \leq M)$. The notation $v_j(i)$ is used to denote a_{ij}. We fix a constant r, and write v_{r+1} in terms of v_r. First, consider the lower-triangular $(N+1) \times (N+1)$ matrix W where the ijth entry of W is $w^{|i-j|+1}$, $i \geq j$ and $w_{ij} = 0$ otherwise. Similarly, let S_r be the $(N+1) \times (N+1)$ matrix with 1 on the diagonal, and $S_r[ij] = s_{jr}$ for $i = j + 1$ and $S_r[ij] = 0$ otherwise. Observe that recursion (1) can be rewritten as

$$v_r = WS_r v_{r-1} \tag{10}$$

In order to sample efficiently, we want to compute v_{r-1} in terms of v_r:

$$v_{r-1} = S_r^{-1}W^{-1}v_r \tag{11}$$

The matrix W is invertible iff $w \neq 0$, which is always the case since $w = e^g$ for some real number g. All the matrices S_r are invertible. Thus, we can solve for v_{r-1} given v_r. Matrix inversion and multiplication are expensive, however, in our case, the equations yield an effective recursive procedure for computing a_{ir} because of the special structure of the matrices. It is easily seen that W^{-1} is a banded matrix with nonzero entries only on the diagonal and off-diagonal, and S_r^{-1} is a lower-triangular matrix. By multiplying them, we obtain the following set of equations for computing v_{r-1} from v_r:

$$\begin{aligned}
v_{r-1}(0) &= \frac{1}{w}v_r(0) \\
v_{r-1}(1) &= \frac{v_r(1)}{w} + v_r(0) - \frac{s_{1r}}{w}v_{r-1}(0) \\
&\vdots \\
v_{r-1}(k) &= \frac{v_r(k)}{w} + v_r(k-1) - \frac{s_{kr}}{w}v_{r-1}(k-1) \\
&\vdots \\
v_{r-1}(N) &= \frac{v_r(N)}{w} + v_r(N-1) - \frac{s_{Nr}}{w}v_{r-1}(N-1)
\end{aligned} \tag{12}$$

Thus, the vector v_{r-1} can be computed in time $O(N)$ from the vector v_r by computing the $v_{r-1}(k)$ in the order $v_{r-1}(0), v_{r-1}(1), \ldots, v_{r-1}(N)$. Although it is possible to derive the recursions (3)–(8) directly from equation (1), we have included the matrix derivation since it is useful in generalizing the backward computation of forward variables to other problems.

The sampling algorithm for randomly selecting an alignment path according to its weight is therefore to first compute v_M using the forward algorithm variation of the Needleman–Wunsch algorithm (*max* replaced by *sum*). This is done using only $O(N)$ memory by keeping only v_{r-1} in memory for the computation of v_r. The samples can then be computed simultaneously by computing $v_{M-1}, \ldots, v_0$ one column at a time and always discarding columns after they have been used (to keep the computations at $O(N)$ memory). The running time of the algorithm for generating k samples is therefore $O(NM + k(N + M))$. The memory requirement, like the Hirschberg algorithm (Hirschberg, 1975), is only $O(N + M)$.

Although we have presented the algorithm for a simple PHMMs, the same method with minor modifications works for more general models.

5. Extensions

5.1. Generalized pair hidden Markov models and alignment

Traditionally, the problems of gene finding and alignment have been treated separately. However, genes and other functional elements in the genome tend to be conserved between related organisms to a higher extent than nonfunctional regions and so strong alignments are an indicator of possible function for the sequences in the aligned regions. Conversely, inferences about sequence function help construct biologically meaningful alignments. Thus, combining the two holds promise to improve both. At a theoretical level, combining gene prediction with alignment is quite simple: the hidden states of the pair HMM are modified to reflect not only the evolutionary history but also the functionality of the bases, and the output states are modified depending on the functional type as well. So, for example, a hidden state might correspond to matching two codons or an output state could correspond to the specific patterns found in splice sites. However, there are serious practical difficulties with this approach. In a PHMM, the number of steps spent in any given hidden state will be geometrically distributed and so, in the case of gene finding, one would expect the lengths of the predicted exons, unlike the lengths of real exons, to have a geometric distribution. This requires using what is referred to as a generalized PHMM. The development of generalized PHMMs was undertaken in Pachter *et al*. (2002). The theory was implemented in a software called SLAM (Alexandersson *et al*., 2003), which aligns and identifies complete exon/intron structures in two related DNA sequences.

5.2. PhyloHMMs and multiple sequence alignment

The hidden Markov model described in Section 3.2. for sequence evolution without gaps can be generalized to multiple sequences related via a phylogeny (McAuliffe, 2004; Siepel and Haussler, 2003). These *phyloHMMs* can also include generalized states, although this introduces complexities when one allows dependencies between neighboring nucleotides in ancestral sequences. In particular, it may no longer be possible to perform efficient MAP inference. In a different direction, multiple sequence alignment by allowing insertions and deletions in ancestral sequences

(without generalized states) leads to interesting models studied by Mitchison (1999), Bruno and Holmes (2001), and others. An interesting future direction for research is the combination of generalized ancestral states with models allowing for insertion and deletion. In other words, it should be interesting to explore generalized phylogenetic Markov Bayesian multinets.

References

Aji SM and McEliece RJ (2000) The generalized distributive law. *IEEE Transactions on Information Theory*, **46**(2), 325–343.

Alexandersson M, Cawley S and Pachter L (2003) SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, **13**, 496–502.

Bucher P and Hofmann K (1996) A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In *ISMB-4*, States D, Agarwal P, Gaasterland T, Hunter L and Smith R (Eds.), AAAI Press: Menlo Park, 44–51.

Cawley S and Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19**(Suppl 2), 36–41.

Durbin R, Eddy S, Korgh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.

Friedman N, Geiger D and Goldszmit M (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

Gusfield D (1997) *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press.

Hirschberg DS (1975) A linear space algorithm for computing maximal common subsequences. *Communication of the Association for Computing Machinery*, **18**(6), 341–343.

Holmes I and Bruno WJ (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**(9), 803–820.

Liu JS, Lawrence CE and Neuwald A (1995) Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *Journal of American Statistical Association*, **90**, 1156–70.

McAuliffe JD, Pachter L and Jordan MI (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, **20**, 1850–1860.

Mitchison GJ (1999) A probabilistic treatment of phylogeny and sequence alignment. *Journal of Molecular Evolution* **49**(1), 11–22.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal Molecular Biology*, **48**, 443–445.

Pachter L, Alexandersson M and Cawley S (2002) Applications of generalized pair hidden Markov models to alignment and gene finding problems. *Journal of Computational Biology*, **9**(2), 389–400.

Pachter L and Sturmfels B (2005) *Algebraic Statistics for Computational Biology*, Cambridge University Press.

Siepel A and Haussler D (2003) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, **11**, 413–428.

Smith TF and Waterman MS (1981) Identification of common molecular sequences. *Journal Molecular Biology*, **147**, 195–197.

Specialist Review

Information theory as a model of genomic sequences

Chengpeng Bi and Peter K. Rogan
University of Missouri, Kansas City, MO, USA

1. Theory

Shannon and Weaver (1949) developed their theory of information in order to understand the transmission of electronic signals and model the communication system. Gatlin (1972) first described its extension to biology. Information theory is an obvious tool to use in looking for patterns in DNA and protein sequences (Schneider, 1995). Information theory has been applied to the analysis of DNA and protein sequences in several ways: (1) by analyzing sequence complexity from the Shannon–Weaver indices of smaller DNA windows contained in a long sequence; (2) by comparing homologous sites in a set of aligned sequences by means of their information content; and (3) by examining the pattern of information content of a sequence divided into successively longer words (symbols) consisting of a single base, base pairs, base triplets, and so forth.

Some of the most useful applications of molecular information theory have come from studies of binding sites (typically protein-recognition sites) in DNA or RNA recognized by the same macromolecule, which typically contain similar but nonidentical sequences. Because average information measures the choices made by the system, the theory can comprehensively model the range of sequence variation present in nucleic sequences that are recognized by individual proteins or multisubunit complexes.

Treating a discrete information source (i.e., telegraphy or DNA sequences) as a Markov process, Shannon defined entropy (H) to measure how much information is generated by such a process. The information source generates a series of symbols belonging to an alphabet with size J (e.g., 26 English letters or 4 nucleotides). If symbols are generated according to a known probability distribution p, the entropy function $H(p_1, p_2, \ldots, p_J)$ can be evaluated. The units of H are in bits, where one bit is the amount of information necessary to select one of two possible states or choices. In this section, we describe several important concepts regarding the use of entropy in genomic sequence analysis.

1.1. Entropy

Entropy is a measure of the average uncertainty of symbols or outcomes. Given a random variable X with a set of possible symbols or outcomes $A_X =$

$\{a_1, a_2, \ldots, a_J\}$, having probabilities $\{p_1, p_2, \ldots, p_J\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{x \in A_X} P(x) = 1$, the Shannon *entropy* of X is defined by

$$H(X) = \sum_{x \in A_X} P(x) \log_2 \frac{1}{P(x)} \tag{1}$$

Two important properties of the entropy function are: (1) $H(X) \geq 0$ with equality for one x, $P(x) = 1$ and (2) Entropy is maximized if $P(x)$ follows the uniform distribution. Here the *uncertainty* or *surprisal*, $h(x)$, of an outcome (x) is defined by

$$h(x) = \log_2 \frac{1}{P(x)} \text{ (bits)} \tag{2}$$

For example, given a DNA sequence, we say each position corresponds to a random variable X with values $A_X = \{\text{A, C, G, T}\}$, having probabilities $\{p_a, p_c, p_g, p_t\}$, with $P(x = \text{A}) = p_a$, $P(x = \text{C}) = p_c$, and so forth. Suppose the probability distribution $P(x)$ at a position of DNA sequence is $P(x = \text{A}) = 1/2$; $P(x = \text{C}) = 1/4$; $P(x = \text{G}) = 1/8$; $P(x = \text{T}) = 1/8$.

The uncertainties (surprisals) in this case are $h(\text{A}) = 1, h(\text{C}) = 2, h(\text{G}) = h(\text{T}) = 3$ (bits). The entropy is the average of the uncertainties $H(X) = \text{E}[h(x)] = 1/2(1) + 1/4(2) + 1/8(3) + 1/8(3) = 1.75$ bits. In a study of genomic DNA sequences, Schmitt and Herzel (1997) found that genomic DNA sequences are closer to completely random sequences than to written text, suggesting that higher-order interdependencies between neighboring or adjacent sequence positions make little contributions to the block entropy.

The entropy (average uncertainty), H, is 2 bits if each of the four bases is equally probable (uniform distribution) before the site is decoded. The information content (IC) is a measure of a reduction in average uncertainty after the binding site is decoded. $IC(X) = H_{\text{before}} - H_{\text{after}} = \log_2|A_x| - H(X)$, provided the background probability distribution $P(\text{before})$ is uniform (Schneider, 1997a). If the background distribution is not uniform, the Kullback–Leibler distance (relative entropy) can be used (Stormo, 2000). The information content calculation needs to be corrected for the fact that a finite number of sequences were used to estimate the information content of the ideal binding site, resulting in the corrected IC, R_{sequence} (Schneider *et al*., 1986). This measures the decrease in uncertainty before versus after the macromolecule is bound to a set of target sequences. Positions within a binding site with high information are conserved between binding sites, whereas low-information content positions exhibit greater variability. The R_{sequence} values obtained precisely describe how different the sequences are from all possible sequences in the genome of the organism, in a manner that clearly delineates the conserved features of the site.

1.2. *Relative entropy*

For two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet, the *relative entropy* (also known as the Kullback–Leibler divergence or

KL-distance) is defined by

$$D_{KL}(P||Q) = \sum_{x \in A_X} P(x) \log \frac{P(x)}{Q(x)} \quad (3)$$

Note that the relative entropy is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$; and although it is sometimes called the KL-distance, it is not strictly a distance (Koski, 2001; Lin, 1991). Relative entropy is an important statistic for finding unusual motifs/patterns in genomic sequences (Durbin *et al*., 1998; Lawrence *et al*., 1993; Bailey and Elkan, 1994; Hertz and Stormo, 1999; Liu *et al*., 2002).

1.3. R_{sequence} versus $R_{\text{frequency}}$

The fact that proteins can find their required binding sites among a huge excess of nonsites (Lin and Riggs, 1975; von Hippel, 1979) indicates that more information is required to identify an infrequent site than a common binding site in the same genome. The amount of information required for these sites to be distinguished from all sites in the genome, $R_{\text{frequency}}$, is derived independently from the size and frequency of sites in the genome. $R_{\text{frequency}}$, like R_{sequence}, is expressed in bits per site. R_{sequence} cannot be less than the information needed to find sites in the genome. With few exceptions, it has been found that R_{sequence} and $R_{\text{frequency}}$ are similar (Schneider *et al*., 1986). This empirical relationship is strongly constrained by the fact that all DNA-binding proteins operating on the genome are encoded in the genome itself (Kim *et al*., 2003).

1.4. Molecular machines

Molecular machines are characterized by stable interactions between distinct components, for example, the binding of a recognizer protein to a specific genomic sequence. The behavior of a molecular machine can be described with information theory. The properties of molecular machine theory may be depicted on multiple levels: on one level, sequence logos, which describe interactions between the molecules (see Figure 1), are equivalent to transmission of information by the recognizer as a set of binary decisions; on another level, the information capacity of the machine, which represents the maximum number of binary decisions (or bits) that can be made for the amount of energy dissipated by the binding event; and finally, the relationship between information content and the energy cost of performing molecular operations (Schneider, 1991; Schneider, 1994). The molecular machine capacity is derived from Shannon's channel capacity (Shannon, 1949). The error rate of the machine can be specified to be as low as necessary to ensure the survival of the organism, so long as the molecular machine capacity is not exceeded. Entropy decreases as the machine makes choices, which corresponds to an increase in information.

The second law of thermodynamics can be expressed by the equation $\mathrm{d}S \geq \mathrm{d}Q/T$. The equation states that for a given increment of heat $\mathrm{d}Q$ entering a volume at some

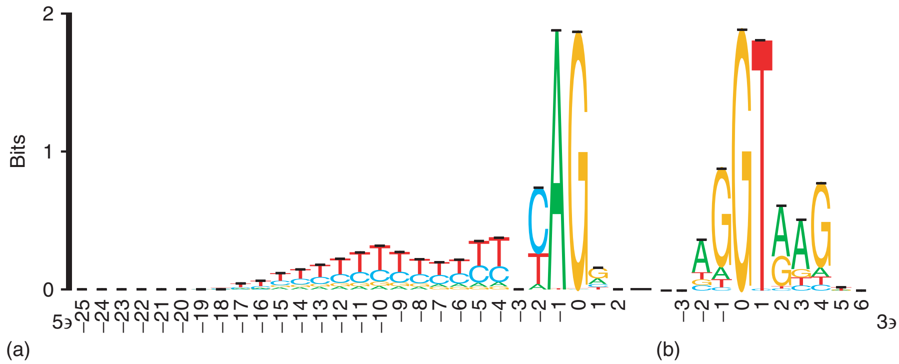


Figure 1 Examples of sequence logos. Models of human (a) 108 079 acceptor and (b) 111 772 donor splice sites derived from both strands of the human genome reference sequence (April, 2003) are shown. A sequence logo visually represents the sequence conservation among a common set of recognition sites, with the height of each nucleotide stack corresponding to the average information content at that position. The height of each nucleotide is proportional to its frequency. Sampling error is indicated by error bar at the top of each stack. The zero coordinate represents the intronic position immediately adjacent to the splice junction. The average information contents (R_{sequence}) of the acceptor and donor sites are respectively, 9.8118 ± 0.0001 bits/site and 8.12140 ± 0.00009 bits/site

temperature T, the entropy will increase dS at least by dQ/T. If we relate entropy to Shannon's uncertainty, we can rewrite the second law in the following form:

$$\varepsilon_{\min} = \kappa_{\text{B}} T \ln(2) \leq \frac{-q}{IC} \text{ (joules per bit)} \tag{4}$$

where κ_{B} is Boltzman constant and q is the heat. This equation states that there is a minimum amount of heat energy that must be dissipated (negative q) by a molecular machine in order for it to gain $IC = 1$ bit of information.

1.5. *Individual information*

The information contained in a set of binding sites is an average of the individual contributions of each of the sites (Shannon, 1948; Pierce, 1980; Sloane and Wyner, 1993; Schneider, 1995). The information content to each individual binding-site sequence can be determined by a weight matrix so that the average of these values over the entire set of sites is the average information content (Schneider, 1997a).

The individual information weight matrix is

$$R_{iw}(b, l) = 2.0 - (-\log_2(f(b, l) + e(n(l)))) \text{ (bits per base)} \tag{5}$$

in which $f(b,l)$ is the frequency of each base b at position l in the binding-site sequences; $e(n(l))$ is a correction of $f(b,l)$ for the finite sample size (n sequences at position l) (Schneider *et al*., 1986). The jth sequence of a set of binding sites is represented by a matrix $s(b,l,j)$, which contains 1's in cells from base b at position l of a binding site and zeros at all other matrix locations. The individual

information of a binding site sequence is the dot product between the sequence and the weight matrix:

$$R_i(j) = \sum_{l} \sum_{b=a}^{t} s(b, j, j) R_{iw}(b, l) \text{ (bits per site)} \tag{6}$$

2. Applications

2.1. *Displaying sequence conservation*

Sequence logos, which display information about both consensus and nonconsensus nucleotides, are visual representations of the information found in a binding site (an example is shown in Figure 1). This is the information that the decoder (i.e., a binding protein) uses to evaluate potential sites in order to recognize actual sites. The calculation of sequence logos uses the assumption that each site is evaluated independently, that is, that there is no correlation between a change in nucleotide at one position with another position, which is reasonable for most genomic sequences (Schmitt and Herzel, 1997). An advantage of the information approach is that the sequence conservation can be interpreted quantitatively. R_{sequence}, which is the total area under the sequence logo and measures the average information in a set of binding site sequences, is related to the specific binding interaction between the recognizer and the site. R_{sequence} is an additive measure of sequence conservation; thus, it is feasible to quantitatively compare the relative contributions of different portions within the same binding site.

Structural features of the protein–DNA complex can be inferred from sequence logos. When positions with high information content are separated by a single helical turn (10.4 bp), this suggests that the protein makes contacts across the same face of the double helix. Sequence conservation in the major groove can range anywhere between 0 and 2 bits depending on the strength of contacts involved, and usually correlates with the highest information content positions (Papp *et al*., 1993). Minor groove contacts of B-form DNA allow both orientations of each kind of base pair so that rotations about the dyad axis cannot easily be distinguished; hence, a single bit is the information content in native B-form DNA (Schneider, 1996). Higher levels of conservation for bases within the minor groove indicate that these positions are accessed protein distortion of the helix, that is, bending accompanied by base-pair opening and flipping (Schneider, 2001).

2.2. *Visualizing individual binding-site information*

Because sequence logos display the average information content in a set of binding sites, they may not accurately convey protein–DNA interactions with individual DNA sequences, especially at highly variable positions within a binding site. The walker method (Schneider, 1997b) graphically depicts the nucleotide conservation of a known or suspected site compared to other valid binding sites defined by the individual information weight matrix (Schneider, 1997a). Walkers apply to a

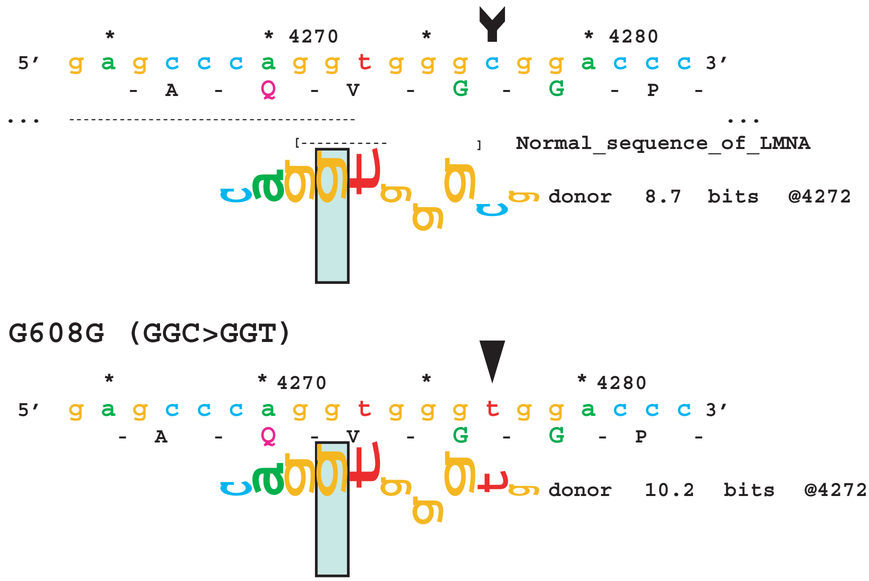


Figure 2 Examples of sequence walkers. A synonymous C > T substitution at codon 608 activates a cryptic donor splice site in exon 11, the *LMNA* gene in patients with Hutchinson–Gifford progeria (Eriksson *et al*., 2003). The walker, shown below the sequence, indicates a preexisting 8.7-bit cryptic site that is strengthened by the mutation to 10.2 bits (≥ 2.8-fold). The height and orientation of each nucleotide correspond to contribution that nucleotide makes to the overall information content in the binding site. The green rectangle indicates the location of valid splice site ($R_i > 0$) and delineates the scale displayed; the lower and upper limits shown are, respectively, −4 bits and +2 bits. Sequence coordinates are from GenBank accession L12401 (4277 C > T)

single sequence (rather than a set of binding sites); only a single letter is visualized at each position of the binding site (Figure 2). The height of the letter, which is in units of bits, represents the contribution of that nucleotide at each position in the binding site by the information weight matrix, $R_{iw}(b,l)$. Evaluation of the R_i value at each position in a genomic DNA sequence is equivalent to moving the walker along that sequence. Walkers are displayed for sequences with positive R_i values, since these are more likely to be valid binding sites (see equation 4 and discussion below). Sequence walkers facilitate visualization and interpretation of the structures and strengths of binding sites in complex genetic intervals and can be used to understand the effects of sequence changes (see below), and engineer overlapping or novel binding sites.

2.3. *Mutation and polymorphism analysis*

Because the relationship between information and energy can be used to predict the effects of natural sequence variation at these sites, phenotypes can be predicted from corresponding changes in the individual information contents (R_i, in bits) of the natural and variant DNA binding sites (Rogan *et al*., 1998; *see also* Article 68, **Normal DNA sequence variations in humans**, Volume 4). For

splice site variants, mutations have lower R_i values than the corresponding natural sites, with null alleles having values at or below zero bits (equation 4; Kodolitsch *et al*., 1999). The decreased R_i values of mutated splice sites indicate that such sites are either not recognized or are bound with lower affinity, usually resulting in an untranslatable mRNA. Decreases in R_i are more moderate for partially functional (or leaky) mutations that reduce but do not abolish splice site recognition and have been associated with milder phenotypes (Rogan *et al*., 1998). The minimum change in binding affinity for leaky mutations is $\geq 2^{\Delta Ri}$ lower fold than cognate wild-type sites. Mutations that activate cryptic splicing may decrease the R_i value of the natural site, increase the strength of the cryptic site, or concomitantly affect the strengths of both types of sites (see Figure 2). Nondeleterious changes do not alter the R_i value of splice sites significantly (Rogan and Schneider, 1995). Increases in R_i indicate stronger interactions between protein and cognate binding sites.

2.4. *Information evolution*

How do genetic systems gain information in a binding site of genomic DNA sequence by evolutionary processes? Schneider (2000) proposed an answer to this question. Given a binding site for an artificial protein, his simulation begins with zero information and, as in naturally occurring genetic systems, the information measured in the fully evolved binding sites (R_{sequence}) is close to that needed to locate the sites in the genome ($R_{\text{frequency}}$).

2.5. *Model refinement*

Information models based on small numbers of proven binding sites may fail to detect valid binding sites and tend to predict R_i inaccurately. Iterative selection of functional binding sites has been used to optimize (Lund *et al*., 2000) and to introduce bias (Shultzaberger and Schneider, 1999) into the frequencies of each nucleotide in computing the information theory–based weight matrices of binding sites. Significant differences between information weight matrices have been determined from their respective evolutionary distance metrics (e.g., see Shultzaberger *et al*., 2001). The effects of model refinements can be monitored by comparing the genome scan results for pairs of successive information weight matrices based on additional binding sites (Gadiraju *et al*., 2003). Other potential applications include the determination of the locations of overlapping binding sites recognized by different proteins and comparisons of binding sites detected with information models of orthologous proteins from different species.

2.6. *Regulatory binding sites*

Information theory–based models have been used in searching for regulatory sites in genomic DNA or RNA sequences of prokaryotes (Hengen *et al*., 1997) and eukaryotes (*see* Article 22, **Eukaryotic regulatory sequences**, Volume 7). The binding sites in prokaryotes include the ribosome binding sites (Shultzaberger *et al*.,

2001), T7 promoters, plasmid replication initiator protein-binding sites (Papp *et al*., 1993) and binding sites for repressors and polymerases (Schneider *et al*., 1986), and cyclic AMP receptor protein (Stormo and Fields, 1989) in *Escherichia coli*.

A bipartite pattern is an independent functional unit on the upstream or downstream side of a regulated gene that is recognized by a protein-binding complex such as a heterodimer. A model built to simulate a bipartite pattern in genomic sequences has left and right motif submodels, plus an associated gap penalty function, $g(d)$ defined as $-\log(n(d)/n)$, where $n(d)$ is the number of sites with gap size d. Shannon's entropy can be used to evaluate such sites by calculating the total information content, IC, given as

$$IC = IC(left|d) + IC(right|d) - g(d) \tag{7}$$

$$IC(m|d) = \sum_{l=1}^{J_m} \left(\log_2 |A_X| - H_{ml}(X)\right), \quad m \in \{left, \quad right\} \tag{8}$$

$$H_{ml}(X) = \sum_{x \in A_X} P_{ml}(x) \log_2 \frac{1}{P_{ml}(x)}, \quad A_X = \{A, C, G, T\} \tag{9}$$

Here J_{m} is the width of motif m and $P_{ml}(x)$ is the probability of x at position l given motif m. The left and right motifs are not allowed to overlap and the gap size (d) is set to a limited range $[d_{\min}, d_{\max}]$ on the basis of empirical observations. The goal is to maximize the total information content that can be reduced to minimize the total Shannon's entropy. We used a Monte Carlo strategy to greedily search the multiple alignment space and find an optimal solution to the bipartite pattern search problem (Bi *et al*., 2004).

We developed a new method for the bipartite cis-regulatory pattern discovery based on minimizing entropy, and applied the method to a set of known PXR/RXRα binding sites in the human genome. The PXR/RXRα heterodimer binding controls the expression of coregulated genes such as *CYP3A4*, which is involved in detoxification of drugs and xenobiotics (*see* Article 70, **Pharmacogenetics and the future of medicine**, Volume 4). This work is an extension of Shultzaberger *et al*. (2001). Using the assumption that two proteins (i.e., PXR and RXRα) cooperatively bind to the bipartite site with constrained spacers, we built models for different motif widths and validated them on the basis of the relative binding strength of a series of test sequences. The results supported our hypothesis that PXR and RXRα transcription factors cooperatively bind to two adjacent motifs with variable spacing (Bi *et al*., 2004).

2.7. Genome-wide analyses

Information weight matrices of binding sites can be developed directly from validated sets of binding sites extracted from genome sequences provided that

the locations of sequence features are accurately annotated. As this is not always the case, we built a genome-wide human splice junction database by initially extracting the coordinates and sequences of donor and acceptor splice regions from both strands of the human genome reference sequence (Rogan *et al*., 2003). After redundant sites were eliminated, the splice site database consisted of 170 144 acceptor and 170 450 donor sites. The information weight matrix was recomputed after each of iteration and scanning of the resultant set of sites. Successive models iteratively utilized the modified matrix by excluding sites with negative R_i values. After eight cycles of refinement, the final models were then defined by 108 079 unique acceptor sites and 111 772 donor sites (sequence logos of model sites are shown in Figure 1).

The average information contents of the acceptor and donor sites are respectively, 9.8 bits/site and 8.1 bits/site. These values are similar to those previously reported by Schneider and Stephens (1990), that is, 9.35 bits for acceptors and 7.92 bits for donors, which were based on about 65-fold fewer splice sites. Individual splice site strengths in the genome have an approximately Gaussian distribution.

3. Prospects for information theory–based analyses of genomic sequences

As the functions for regulatory and expressed nucleic acid sequences are elucidated, it is becoming evident that multiple protein components catalyze these processes. Modeling such molecular machines by information theory will require the development of procedures that account for cooperative and interdependent binding events between two or more recognizers. Frameworks for building multipartite information models will therefore have to incorporate corrections for overlapping sites and mutual information.

There are opportunities to enhance currently available genomic applications by scaling currently available software for information theory analyses (Gadiraju *et al*., 2003) to investigate genome annotation. For example, changes in IC contents due to mutation may be of assistance in prioritizing single nucleotide polymorphisms for functional analyses. It is also possible that comparative genomic analyses of binding sites with orthologous DNA recognition domains from multiple species may reveal possible identities of functionally analogous regulatory sequences in these systems.

Medical genetic applications of information theory–based binding-site models are a promising avenue to improve diagnosis and prognosis of disease-causing mutations (*see* Article 80, **Genetic testing and genotype–phenotype correlations**, Volume 2). Accurate models will be required for use of information theory in a clinical setting. To calibrate individual information measures of protein-nucleic acid binding with the thermodynamic properties of these complexes will require more comprehensive models, that is, based on larger numbers of binding sites spanning a wide range of binding affinities.

Related articles

See Article 80, **Genetic testing and genotype–phenotype correlations**, Volume 2; Article 68, **Normal DNA sequence variations in humans**, Volume 4; Article 19, **Promoter prediction**, Volume 7; Article 22, **Eukaryotic regulatory sequences**, Volume 7; Article 28, **Computational motif discovery**, Volume 7

Acknowledgments

This work was supported by NIEHS (Grant ES10855). We thank Dr. Thomas Schneider for valuable suggestions and comments.

Further reading

Chen X, Li M, Ma B and Tromp J (2002) DNACompress: fast and effective DNA sequence compression. *Bioinformatics*, **18**, 1696–1698.

Cover TM and Thomas JA (1991) *Elements of Information Theory*, John Wiley & Sons: New York.

Milosavljevic A and Jurka J (1993) Discovering simple DNA sequences by the algorithmic significant method. *Computer Applications in the Biosciences: CABIOS*, **9**, 407–411.

Schneider TD and Mastronarde DN (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics*, **71**, 259–268.

Stephens RM and Schneider TD (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology*, **228**, 1124–1136.

References

Bailey TL and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, **2**, 28–36.

Bi C-P, Vyhidal CA, Leeder JS and Rogan PK (2004) A minimization entropy based bipartite algorithm with application to PXR/RXRα binding sites. *Proceedings of the RECOMB2004 Annual Symposium, San Diego*, 453–454.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press: Cambridge.

Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robins CM, Moses TY, Berglund P, *et al*. (2003) Recurrent de novo point mutations in lamin A cause Hutchinson-Gifford progeria syndrome. *Nature*, **423**, 293–298.

Gadiraju S, Vyhlidal CA, Leeder JS and Rogan PK (2003) Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics*, **4**, 38.

Gatlin LL (1972) *Information Theory and the Living System*, Columbia University Press: New York.

Hengen PN, Bartram SL, Stewart LE and Schneider TD (1997) Information analysis of Fis binding sites. *Nucleic Acids Research*, **25**, 4994–5002.

Hertz GZ and Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Kim JT, Martinez T and Polani D (2003) Bioinformatic principals underlying the information content of transcription factor binding sites. *Journal of Theoretical Biology*, **220**, 529–544.

Kodolitsch Yv, Pyeritz RE and Rogan PK (1999) Splice site mutations in atherosclerosis candidate genes: relating individual information to phenotype. *Circulation*, **100**, 693–699.

Koski T (2001) *Hidden Markov Models for Bioinformatics*, Kluwer Academic Publishers.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lin J (1991) Divergence measure based on the Shannon entropy. *IEEE Transactions on Information Theory*, **37**, 145–151.

Lin S and Riggs AD (1975) The general affinity of lac repressor for E. coli DNA: implications for gene regulation in procaryotes and eucaryotes. *Cell*, **4**, 107–111.

Liu XS, Brutlag DL and Liu JS (2002) An algorithm, for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, **20**, 835–839.

Lund M, Tange TO, Dyhr-Mikkelsen H, Hansen J and Kjems J (2000) Characterization of human RNA splice signals by iterative functional selection of splice sites. *RNA*, **6**, 528–544.

Papp PP, Chattoraj DK and Schneider TD (1993) Information analysis of sequences that bind the replication initiator RepA. *Journal of Molecular Biology*, **233**, 219–230.

Pierce JR (1980) *An Introduction to Information Theory: Symbols, Signals and Noise*, Second Edition, Dover Publications: New York.

Rogan PK, Faux BM and Schneider TD (1998) Information analysis of human splice site mutations. *Human Mutation*, **12**, 153–171.

Rogan PK and Schneider TD (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Human Mutation*, **6**, 74–76.

Rogan PK, Svojanovsky S and Leeder JS (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics*, **13**, 207–218.

Schmitt AO and Herzel H (1997) Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, **188**, 369–377.

Schneider TD (1991) Theory of molecular machines. *Journal of Theoretical Biology*, **148**, 83–137.

Schneider TD (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, **5**, 1–18.

Schneider TD (1995) Information Theory Primer. **http://www.lecb.ncifcrf.gov/~toms/paper/primer/**.

Schneider TD (1996) Reading of DNA sequence logos: prediction of major groove binding by information theory. *Methods in Enzymology*, **274**, 445–455.

Schneider TD (1997a) Information content of individual genetic sequences. *Journal of Theoretical Biology*, **189**, 427–441.

Schneider TD (1997b) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Research*, **25**, 4408–4415.

Schneider TD (2000) Evolution of biological information. *Nucleic Acids Research*, **28**, 2794–2799.

Schneider TD (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Research*, **29**, 4881–4891.

Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**, 6097–6100.

Schneider TD, Stormo GD, Gold L and Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, **188**, 415–431.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–432.

Shannon CE (1949) Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, **37**, 10–21.

Shannon CE and Weaver W (1949) *The Mathematical Theory of Communication*, University of Illinois Press: Urbana.

Shultzaberger RK, Bucheimer RE, Rudd KE and Schneider TD (2001) Anatomy of *Escherichia coli* ribosome binding sites. *Journal of Molecular Biology*, **313**, 215–228.

Shultzaberger RK and Schneider TD (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Research*, **27**, 882–887.

Sloane NJA and Wyner AD (1993) *Claude Elwood Shannon: Collected Papers*, IEEE Press: Piscataway.

Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo GD and Fields DS (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 5699–5703.

von Hippel PH (1979). In *Biological Regulation and Development*, Vol. 1, Goldberger RF (Ed.), Plenum Press: New York, pp. 279–347.

Short Specialist Review

Promoter prediction

Vladimir B. Bajic
Institute for Infocomm Research, Singapore

Thomas Werner
Genomatix, Munich, Germany

1. Biological problem and importance for practice and science

Complex processes in cells of living organisms depend on synchronous actions of different groups of genes. Coordination of gene expression is achieved to a large extent by different transcriptional control mechanisms characteristic for each gene and controlling timing, rate, and level of its transcription. Promoters represent genomic regions containing many such regulatory signals. Polymerase II promoters are located at the beginning of the genomic region whose transcription they control. The boundaries of promoters are not very clear, but most important transcriptional signals known today are generally located within the segment of [−2000, +500] relative to the transcription start site (TSS), which often already includes proximal enhancers. The +1 position corresponds to the position of the first transcribed nucleotide.

Accurate promoter determination can help to (1) identify new genes; (2) complement annotation of known genes and identify their 5′ boundary; (3) predict alternative transcripts initiating from additional promoters; (4) localize the most important control region for gene activation; (5) annotate transcriptional regulatory patterns; and (6) determine organizational promoter models of different gene groups, which ultimately can be used to determine the cause–consequence relationships characteristic of different gene activation pathways and gene regulatory networks. The bottom line is that determining the location of promoters can ultimately contribute toward the understanding of the molecular control of activity of the respective genes.

2. Current possibilities for locating promoters

There are two basic categories of techniques allowing for large-scale identification of promoters: experimental and computational methods. The most promising experimental techniques are based on oligo-capping (Maruyama and Sugano, 1994), where the cap structure of mRNAs is selectively replaced by an oligonucleotide,

which subsequently serves as sequencing primer to determine the 5′ sequence of the transcript. More than 83 000 human and murine transcripts have been characterized that way to date. Other experimental techniques include generation of EST sequences and mRNA fragments, and require back-mapping to original DNA genomic sequences. Unfortunately, none of these methods is perfect and even the best (oligo-capping) produces about 20 to 30% false TSS data.

Computational detection of promoters has also advanced considerably recently and is reaching the level where the amount of correct promoter prediction is severalfold or higher than the amount of false-positive promoter prediction in large-scale genomic searches (Scherf *et al*., 2000; Bajic and Seah, 2003).

3. Current computational solutions

We will make comments about promoter prediction programs (PPPs) for eukaryotic promoters developed or significantly modified in 1999 and after. Two reviews (Fickett and Hatzigeorgiou, 1997; Prestridge, 2000) discussed earlier PPP solutions. We will comment on CONPRO (Liu and States, 2002), CpGProD (Ponger and Mouchiroud, 2002), CpG-Promoter (Ioshikhes and Zhang, 2000), Dragon Promoter Finder (DPF) (Bajic *et al*., 2003), Dragon Gene Start Finder (DGSF) (Bajic and Seah, 2003), Eponine (Down and Hubbard, 2002), First Exon Finder (FirstEF) (Davuluri *et al*., 2001), McPromoter (Ohler *et al*., 2001), NNPP2.1 (Reese, 2001), Promoter2.0 (Knudsen, 1999), PromoterInspector (Scherf *et al*., 2000), PromH (Solovyev and Shahmuradov, 2003), and the system of Hannenhalli and Levy (SHL) (Hannenhalli and Levy, 2001).

4. Biological signals

There are several types of biological signals utilized to enhance computational promoter predictions.

The most significant are CpG-islands (Bird *et al*., 1986), relatively short stretches of DNA characterized by high amount of CpG dinucleotides containing high amounts of unmethylated C nucleotides. Unmethylated C nucleotides facilitate factor binding to genomic DNA during transcriptional activation and initiation. CpG-islands are frequently found close to the gene starts and thus close to promoters. CpG-islands are characteristic for a large proportion of vertebrate genomes (Cross and Bird, 1995). PPPs that use the concept of CpG-island are CpGProD, CpG-Promoter, DGSF, FirstEF, and SHL.

Another significant characteristic, at least of the human genome, is an elevated GC content around the TSS, across the first exon, and around the first splice donor site as opposed to other parts of the genome (Louie *et al*., 2003). These properties are only partially related to CpG-islands. Moreover, there is a strong bias in the nucleotide content in 5′ region of genes (Louie *et al*., 2003; Majewski and Ott, 2002), including CpG dinucleotide and GGG trinucleotide. PPPs that utilize different GC content of promoters are DPF, DGSF, Eponine, FirstEF, and together with other features also PromoterInspector.

The third characteristic of promoters that is more universal across different taxa are sets of specific combinations of promoter elements (PEs) (transcription factor binding sites and promoter boxes, and their combinatorial and positional distribution patterns) for individual promoters and promoter groups of coregulated genes (e.g., Fessele *et al*., 2002). PPPs that use different combinations of promoter elements are McPromoter, NNPP2.1, and Promoter2.0.

The fourth characteristic of promoter regions versus nonpromoter regions can be found in different densities of potential PEs in these regions. Independent studies of authors suggest a number of overrepresented PE patterns, such as E2F, ETF, Elk-1, and ZF5, in human promoters as opposed to nonpromoters. However, as such overrepresentation is method/data dependent in most cases, this may impact performance of PPPs that utilize this promoter characteristic.

In addition, there are some other physicochemical properties that distinguish promoters from nonpromoters, such as DNA bendability, propeller twist, and so on (Ohler *et al*., 2001). A PPP that uses these properties in combination with other features is McPromoter.

PPPs employ numerous types of biological information, and currently, the most efficient systems utilize some of the key biological characteristics of promoter regions or their combinations.

5. Implemented technology

PPP implementations include position weight matrices and their higher-order derivatives (in DPF, DGSF, FirstEF, Eponine); various artificial intelligence approaches such as artificial neural networks (DPF, DGSF, McPromoter, NNPP2.1, Promoter2.0, PromoterInspector), Interpolated Markov Models (McPromoter), and relevance vector machine (Eponine); some standard statistical techniques such as linear (SHL, PromH), and quadratic discriminant analysis (FirstEF, CpG-Promoter); as well as comparison with orthologous promoter sequences (PromH). Some solutions use combinations of different separate solutions (CONPRO).

6. Information provided to end users

The existing PPPs provide different information to the end user. Prediction of promoters can be strand-specific (DPF, DGSF, Eponine, FirstEF, CONPRO, McPromoter, NNPP2.1, Promoter2.0, PromH) or non-strand-specific (CpG-Promoter, CpGProD, SHL, PromoterInspector). In the first case, the direction of the body of the transcript is predicted, while in the latter case only the genomic locations of its 5′ end is predicted, with no direction for the transcript. Some programs attempt to predict the actual TSS locations (DPF, DGSF, Eponine, FirstEF, CONPRO, McPromoter, NNPP2.1, Promoter2.0, PromH), while others predict a region expected to contain a TSS (CpG-Promoter, CpGProD, DGSF, SHL, PromoterInspector). In addition to these basic data, some PPPs also provide additional information such as potential TSBSs in the region of interest (DPF, DGSF, PromH), information on

other genomic signals, such as first donor (FirstEF), CpG-islands (FirstEF, CpG-ProD, CpG-Promoter), genomic location corresponding to the translation initiation codon (DPF), and so on.

7. Current performance

We tested several PPPs on three whole human chromosomes (4, 21, and 22). Details of the experiment were as presented in Bajic and Seah (2003), with the only difference being that we merged predictions of individual systems if they were no more than 1000 nucleotides apart (except for PromoterInspector). The performances in percentage for PPPs are given in the form (sensitivity, positive predictive value): DPF (for threshold 0.5) (80.70, 22.21), DGSF (65.04, 77.27), Eponine (41.04, 85.20), FirstEF (78.43, 48.08), McPromoter (for threshold −0.005) (55.65, 70.95), NNPP2.1 (for threshold 0.99) (60.52, 7.40), and Promoter2.0 (60.52, 4.61). PromoterInspector achieved (42, 52) and (11, 96) on the complete human and mouse genomes, respectively, measured relative to the known transcripts (not gene numbers) and including a large amount of alternative transcripts usually not considered.

8. Future solutions: the next generation

Since the large-scale localization of TSSs in mammalian genomes has reached maturity and levels of ~80% of correct promoter predictions falling into region [−500, +500] relative to real TSS (Bajic and Seah, 2003), there are two immediate next targets for future developments of promoter prediction. One goal would be the more position-accurate TSS prediction, hopefully in the range of [−20, +20] relative to the real TSSs, and the ability to accurately distinguish alternative TSSs. The second goal, naturally, is to increase the sensitivity of predictions while preserving the specificity reached so far. This will be very important in the light of a huge amount of alternative transcripts representing an important body of biologically relevant information.

Another issue will be the extension of high-accuracy predictors to nonmammalian species, including other vertebrates, invertebrates, and plants. These extensions may not necessarily be simple, as, for example, in *Fugu rubripes* the GC properties of promoters and nonpromoters are very different compared to the human genome. Many of the nonmammalian species do not have high proportion of unmethylated CpG-islands characterizing gene starts, while typical promoter elements such as TATA and CCAAT boxes may appear in completely different proportions than in mammalian genomes. Furthermore, the GC content of different genomes will influence our ability to effectively search for TSSs computationally.

Related articles

Article 8, **The role of gene regulation in evolution**, Volume 1; Article 32, **DNA methylation in epigenetics, development, and imprinting**, Volume 1; Article 27,

Noncoding RNAs in mammals, Volume 3; Article 28, **The distribution of genes in human genome**, Volume 3; Article 29, **Pseudogenes**, Volume 3; Article 31, **Overlapping genes in the human genome**, Volume 3; Article 33, **Transcriptional promoters**, Volume 3; Article 34, **Human microRNAs**, Volume 3; Article 9, **Repeatfinding**, Volume 7; Article 14, **Eukaryotic gene finding**, Volume 7; Article 16, **Searching for genes and biologically related signals in DNA sequences**, Volume 7; Article 17, **Pair hidden Markov models**, Volume 7; Article 18, **Information theory as a model of genomic sequences**, Volume 7; Article 22, **Eukaryotic regulatory sequences**, Volume 7; Article 28, **Computational motif discovery**, Volume 7; Article 98, **Hidden Markov models and neural networks**, Volume 8; Article 106, **Gibbs sampling and bioinformatics**, Volume 8; Article 110, **Support vector machine software**, Volume 8

References

Bajic VB and Seah SH (2003) Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Research*, **13**, 1923–1929.

Bajic VB, Seah SH, Chong A, Krishnan SPT, Koh JLY and Brusic V (2003) Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates. *Journal of Molecular Graphics & Modeling*, **21**, 323–332.

Bird AP, Taggart MH, Nichollas RD and Higgs DR (1986) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO Journal*, **6**, 999–1004.

Cross SH and Bird AP (1995) CpG islands and genes. *Current Opinion in Genetics & Development*, **5**, 309–314.

Davuluri RV, Grosse I and Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. *Nature Genetics*, **29**, 412–417.

Down TA and Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research*, **12**, 458–461.

Fessele S, Maier H, Zischek C, Nelson PJ and Werner T (2002) Regulatory context is a crucial part of gene function. *Trends in Genetics*, **18**, 60–63.

Fickett JW and Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Research*, **7**, 861–878.

Hannenhalli S and Levy S (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.

Ioshikhes IP and Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. *Nature Genetics*, **26**, 61–63.

Knudsen S (1999) Promoter2.0: for the recognition of Pol II promoter sequences. *Bioinformatics*, **15**, 356–361.

Liu R and States DJ (2002) Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Research*, **12**, 462–469.

Louie E, Ott J and Majewski J (2003) Nucleotide frequency variation across human genes. *Genome Research*, **13**, 2594–2601.

Majewski J and Ott J (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Research*, **12**, 1827–1836.

Maruyama K and Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.

Ohler U, Niemann H, Liao G and Rubin GM (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17**, S199–S206.

Ponger L and Mouchiroud D (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, **18**, 631–633.

Prestridge DS (2000) Computer software for eukaryotic promoter analysis. Review. *Methods in Molecular Biology*, **130**, 265–295.

Reese MG (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Computers & Chemistry*, **26**, 51–56.

Scherf M, Klingenhoff A and Werner T (2000) Highly specific localisation of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *Journal of Molecular Biology*, **297**, 599–606.

Solovyev VV and Shahmuradov IA (2003) PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Research*, **31**, 3540–3545.

Short Specialist Review

Operon finding in bacteria

Maria D. Ermolaeva
The Institute for Genomic Research, Rockville, MD, USA

Bacterial genes are often organized into multigene transcriptional units (TUs), a series of genes that are transcribed together into one messenger RNA (mRNA) molecule. A TU starts with a promoter, which initiates transcription, and ends with a terminator, which terminates transcription (Figure 1a). The expression of the genes in a TU is controlled by one set of regulatory gene(s), which is often located nearby. The term “operon” was originally defined to include both the TU and the associated regulatory genes; however, today, the term is often used to refer only to a set of cotranscribed genes. This is especially common in computational genomics, where it is often necessary to simplify the model of an operon in order to make computational predictions. Indeed, the model shown in Figure 1(a) is simplified in a number of other aspects, too. Additional promoters and terminators may be located between the cotranscribed genes, and genes may be cotranscribed in some conditions and transcribed separately in other conditions. These details usually cannot be captured by operon-prediction methods and, therefore, they are not discussed here. Genes that are transcribed separately from other genes can be considered as one-gene TUs.

Finding TUs computationally is a difficult task; the predictions are much less accurate than the gene predictions themselves. The apparent solution would be to find the stretches of genes located between a promoter and terminator; however, this requires accurate knowledge of promoters and terminators, which are usually at least as difficult to predict as operons. The only exception here is rho-independent terminators, which have a distinctive structure and can, for some species of bacteria, be predicted accurately.

While it is difficult to predict TUs’ boundaries, it is possible to make a rough estimate of their number in a genome. This estimate can be done using only gene predictions, which are usually more than 95% accurate in bacteria. Here, we have to introduce an additional term, “directon” (Salgado *et al*., 2000), which is a set of genes located consecutively, one after another, on the same DNA strand. All genes of a TU always belong to the same directon because genes must be on the same strand in order to be transcribed together. A directon may include one or more TUs and single genes (one-gene TUs). Figure 1(a) shows a directon that consists of one TU – the flanking TUs have the opposite orientation; that is, they are located on the opposite DNA strand. Figure 1(b) shows another possible situation, where the two neighboring TUs have the same orientation and belong to the same operon. If the orientation of all TUs were completely independent of their flanking TUs,

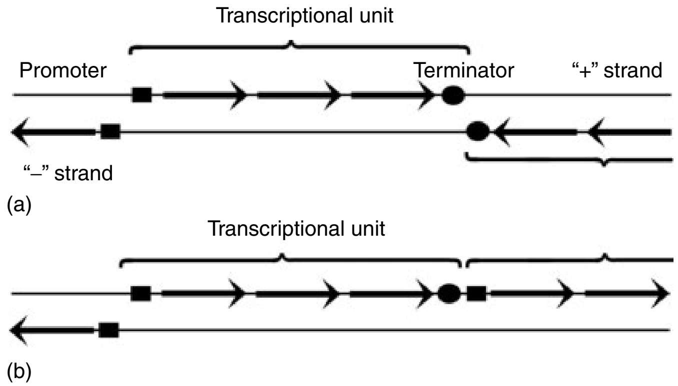


Figure 1 A scheme of a transcription unit. (a) The neighboring transcription units are located on the opposite DNA strand. (b) At least two consecutive transcription units are located on the same DNA strand, making it difficult to locate a boundary between them

then the average directon would have two TUs (Ermolaeva *et al*., 2001). Therefore, the size of an average TU would be half the average directon size, which can be easily calculated from the gene predictions. (Note that single genes are counted as one-gene TUs.) Most bacterial and archaeal genomes have some strand bias, with more genes located on the leading strand of replication than on the lagging strand. In this case, an average TU size t can be calculated as

$$t = \frac{d}{1+b} \tag{1}$$

where d is the average directon size and b is the strand bias, calculated as the number of genes on the leading strand divided by the number of genes on the lagging strand. We prove this as follows. First, the probability that a randomly chosen TU would be located on the leading strand is

$$p = \frac{b}{1+b} \tag{2}$$

The probability that a randomly chosen directon has exactly n TUs is the probability that the first n TUs are located on the leading strand and the next TU is on the lagging strand plus the probability that the first n TUs are located on the lagging strand and the next TU is on the leading strand:

$$P(n) = p^n \cdot (1-p) + (1-p)^n \cdot p \tag{3}$$

The average number of TUs in a directon is the sum of all possible directon sizes multiplied by their corresponding likelihoods:

$$t = \sum_{n=1}^{\infty} P(n) \cdot n \tag{4}$$

Combining equations (2), (3), and (4) and reducing the infinite summation yields equation (1).

The next logical step is to find the TU boundaries, that is, to predict which of the neighboring genes are transcribed together and which pairs of the neighboring genes belong to different TUs. For genes located on opposite DNA strands, the answer is obvious: they belong to the different TUs because they cannot be transcribed together. Consecutive genes that are located on the same DNA strand may belong to the same or different TUs.

TU and operon boundaries can, in some cases, be found using the knowledge of metabolic pathways and gene functions (Zheng *et al*., 2002). Genes that belong to the same metabolic pathway are often regulated together, and are located in the same operon. This method, however, can only be applied to well-studied genes with known metabolic pathways. Another method that relies on experimental data is described in Sabatti *et al*. (2002). Genes within a TU are transcribed together and, therefore, they have similar levels of expression. Microarray experiments allow us to measure the expression of genes in different conditions and to find genes with correlated expression levels. Although correlated expression does not directly imply that genes belong to the same operon, when coupled with those genes' adjacency information on the genome it can provide strong evidential support. The main shortfall of this method is the difficulty, using current technology, of obtaining accurate and reproducible gene expression measurements, which means that a large number of experiments are required to detect the correlation.

The two methods described above require extensive experimental data to support any conclusions about TUs. In addition, there are a few "purely computational" approaches that only rely on the DNA sequence data and gene predictions. Two such methods are described below.

The first method is based on calculating intergenic distances; that is, the distances between neighboring genes (Salgado *et al*., 2000). As illustrated in Figure 1(b), two neighboring genes must be separated by a terminator and a promoter, if they belong to different TUs. Thus, the distance between such genes is usually longer than the intergenic distances within TUs. Thus, genes that are located close to each other are likely to belong to the same TU, while genes that are separated by hundreds of nucleotides are (in prokaryotes) likely to have a TU boundary between them. This amazingly simple method has an impressively high (compared with the other methods) accuracy of 82% for the *Escherichia coli* genome. This method can be also used for other bacterial and archaeal genomes (Moreno-Hagelsieb and Collado-Vides, 2002), but its accuracy is not always clear. Different genomes have different distributions of intergenic lengths, either due to fundamental differences in their biology or, possibly, to differences in methods of computational gene prediction. Figure 2 shows such distributions for three bacterial genomes: *E. coli*, *Thermotoga maritima*, and *Synechocystis*. In the figure, *E. coli* has a distinctive maximum for short intergenic distances, most of which are located within TUs. Similarly, *T. maritima* has an even higher percentage of short distances, but *Synechocystis* appears to have few short intergenic distances.

It is not clear whether the genomes with longer average intergenic distances have fewer operons or whether the distances between genes within operons are longer. Figure 3 shows that there is a weak correlation (Pearson correlation coefficient 0.26)

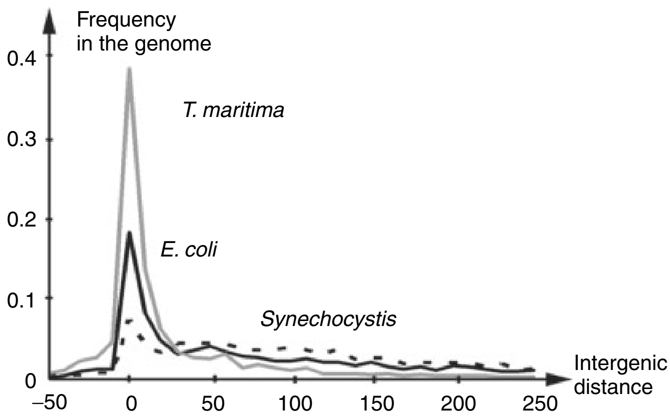


Figure 2 Frequency of intergenic distances (in 10 bp bins) in three bacterial genomes. Only distances between genes located on the same DNA strand were considered

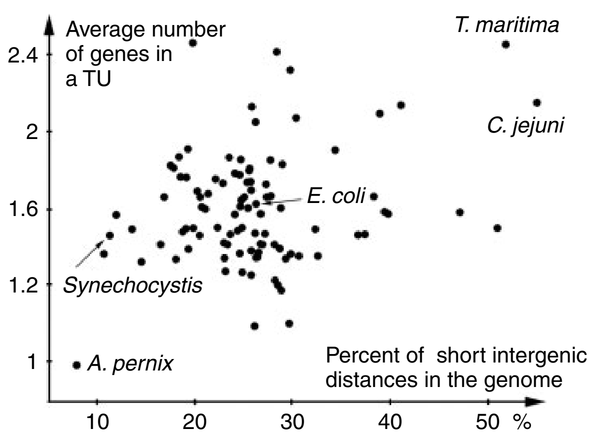


Figure 3 Each dot represents one sequenced bacterial or archaeal genome. Axis x shows percent of short intergenic distances (from −5 to 15 bp) in the genome (only distances between genes located on the same DNA strand were considered) and axis y shows the average number of genes in a transcriptional unit, calculated using equation (1)

between the percent of short distances in a genome and the average number of genes in a TU (calculated using equation (1)). We should mention here that intergenic distances may vary significantly due to the accuracy of placement of gene start sites (i.e., the initial ATG start codon), whose prediction varies dramatically depending on the gene prediction method used. This bias, however, does not appear to be the sole reason for the weak correlation: using subsets of genes whose start sites have greater accuracy (known from sequence homology with genes in other species and ribosome binding sites predictions) has only a small effect on the overall picture.

The second, more computationally sophisticated operon-finding technique is based on finding conserved gene clusters (Bansal, 1999; Overbeek *et al*., 1999; Huynen *et al*., 2000; Snel *et al*., 2000; Ermolaeva *et al*., 2001; Mering *et al*.,

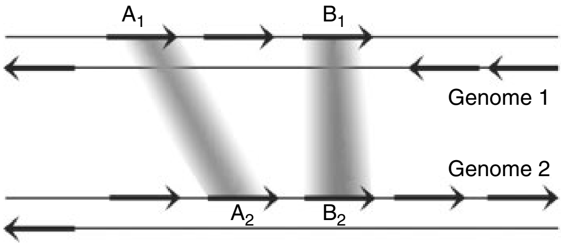


Figure 4 Genes A and B are located nearby, within one directon in both genomes, forming a conserved gene cluster. A_1 is ortholog of A_2, and B_1 and B_2 are also orthologs

2003). Bacterial genomes often reshuffle their genes, changing gene order and orientation. Even such evolutionary related genomes as *E. coli*, *Haemophilus influenzae*, and *Vibrio cholerae* have completely different gene orders when viewed at a genome scale. Genes that belong to the same operon – and that are regulated by the same mechanisms – are under greater selective pressure to remain together, even as other genes are shuffled. Therefore, when conserved gene clusters are observed in evolutionarily distant genomes, one can make the inference that these clusters are likely to represent TUs. Figure 4 shows a scheme of a conserved gene cluster that consists of two genes.

The specificity of these methods depends on the extent of gene cluster conservation, and may be as high as 98% or higher (Ermolaeva *et al*., 2001) if the gene cluster is shared by at least a few evolutionary distant genomes. Such methods, however, can locate only a portion of all operons, because many operons are present in only one or two currently sequenced genomes. Fortunately, the predictive power of this comparative genomics strategy will improve steadily over time, as the sequences of more bacterial and archaeal genomes become available.

Acknowledgments

The author would like to thank Steven L. Salzberg for helpful discussions and comments and National Science Foundation (grant DBI-0234704) for financial support.

References

Bansal AK (1999) An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*, **15**, 900–908.

Ermolaeva MD, White O and Salzberg SL (2001) Prediction of operons in microbial genomes. *Nucleic Acids Research*, **29**, 1216–1221.

Huynen M, Snel B, Lathe W and Bork P (2000) Exploitation of gene context. *Current Opinion in Structural Biology*, **10**, 366–370.

Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, **31**, 258–261.

Moreno-Hagelsieb G and Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**(Suppl 1), S329–S336.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–2901.

Sabatti C, Rohlin L, Oh MK and Liao JC (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, **30**, 2886–2893.

Salgado H, Moreno-Hagelsieb G, Smith TF and Collado-Vides J (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6652–6657.

Snel B, Lehmann G, Bork P and Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research*, **28**, 3442–3444.

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ and Kasif S (2002) Computational identification of operons in microbial genomes. *Genome Research*, **12**, 1221–1230.

Short Specialist Review

Gene structure prediction in plant genomes

Volker Brendel
Iowa State University, Ames, IA, USA

1. Introduction

In eukaryotes, the presence of intervening sequences (introns) within most genes makes the problem of computational gene structure prediction distinct from (and harder than) the same problem in prokaryotes. However, even among eukaryotes, the problem is varied beyond the basic need for species-specific training of algorithm parameters. For example, introns are rare in yeast genes, and, where found, they typically bear highly conserved signature patterns for splicing signals that make identification easy. The purpose of this short review is to discuss approaches to gene identification in plants. There are two reasons why this task is somewhat different from the same task in animals. The first reason is biological. Introns in plant genes have a similar length distribution as exons (Brendel *et al*., 1998). In particular, there appear to be none of the very long introns that confound gene prediction in vertebrates. The second reason is pragmatic. Expressed Sequence Tag (EST) sequencing and whole-genome as well as genome-survey sequencing is in progress for dozens of plant species, all of which appear closely enough related that gene identification in any given species can very efficiently leverage annotation and sequence information from all the other species. In practice, a combination of ab initio gene prediction, spliced alignment, and expert annotation will, in most cases, produce highly reliable protein-coding gene structures. Identification of noncoding RNA genes such as *snoRNA* and *miRNA* genes might prove a much harder problem (Chen *et al*., 2003; Reinhart *et al*., 2002; Llave *et al*., 2002; Marker *et al*., 2002).

1.1. Ab initio algorithms for plant gene finding

Most of the experience with computational gene structure prediction in plants was derived from annotation and reannotation of the Arabidopsis genome. The initial annotation (Arabidopsis Genome Initiative, 2000) was achieved with specially trained gene-finding programs such as GENSCAN (Burge and Karlin, 1997), GeneMark.hmm (Lukashin and Borodovsky, 1998), and GlimmerA (Salzberg *et al*., 1999). These programs derive their predictions “ab initio”: the input consists only of the genomic DNA to be annotated, and gene structures are derived on the

basis of models for exons, introns, and transcription and splicing signals (reviewed in Mathé *et al*., 2002). The underlying methods, also called "intrinsic methods", are similar between the various programs; however, combinations of the programs have been shown to perform better than any single program (Murakami and Takagi, 1998; Pavy *et al*., 1999). The performance of these methods has been thoroughly tested (Pavy *et al*., 1999). At best, exon level sensitivity and specificity were found around the 80% mark, and fewer than half of the predicted gene structures were completely correct. While unsatisfactory in terms of current understanding of gene structure and transcriptional and posttranscriptional control, in practice, such approximate annotations are highly useful because they immediately provide a glimpse of the gene space. Biologists can often find their genes of interest by a simple BLAST search (Altschul *et al*., 1997) against the annotated genes and then refine the respective gene models by more detailed analysis. On a large scale, reannotation efforts (Wortman *et al*., 2003) rely on approaches discussed in the next two sections.

1.2. Spliced alignment with cDNAs and ESTs

Full-length cDNA sequencing provides the most direct way for gene structure identification, because by definition the cDNA residues derive from exons in the genomic sequence. The problem of threading a cDNA sequence back into its corresponding genomic DNA is known as *spliced alignment*. The alignment task is straightforward in the absence of significant differences between the cDNA and the genomic DNA (some level of mismatch would simply result from sequencing errors or differences between DNA sources). A number of programs are available to this end, including dds/gap2 (Huang *et al*., 1997), sim4 (Florea *et al*., 1998), GeneSeqer (Usuka *et al*., 2000), and BLAT (Kent, 2002). In practice, large-scale EST-sequencing projects often provide the first view of a plant's gene space. If not accompanied by a genome-sequencing project, the ESTs are typically clustered and assembled on the basis of sequence similarity to derive nonredundant sets of putative transcripts ("unigenes"; e.g., Quackenbush *et al*., 2001; Dong *et al*., 2004). Otherwise, inherent problems of EST clustering (limited power to detect overlap and to distinguish duplicated genes, chimeric clones) can be largely eliminated by spliced alignment of ESTs onto the genome and subsequent assembly of these spliced alignments to complete gene structures (Zhu *et al*., 2003; Haas *et al*., 2003). In many cases, incorporation of splice site prediction into the derivation of optimal spliced alignments allows the use of nonnative ESTs for gene structure prediction (Brendel *et al*., 2004). This approach appears very promising, given the large combined EST resources across all plant species.

1.3. Spliced alignment with proteins

Although gene order is not necessarily conserved between even closely related plant species (Tarchini *et al*., 2000; Fu and Dooner, 2002), the proteomes of different plants are estimated to be more than 90% conserved (Bennetzen, 2000).

This suggests a successful complementary strategy to gene identification based on spliced threading of protein sequences into genomic DNA thought to encode a homologous protein. The implicit conservation of the reading frame across predicted coding exons makes the algorithms for such alignment tasks more complex than in the cDNA/EST spliced alignment case. Programs such as AAT (Huang *et al*., 1997) and GeneSeqer (Usuka and Brendel, 2000) can be used, provided close-enough homologs can be identified as probes. In practice, good results can be achieved with a combination of cDNA/EST and protein spliced alignments (Schlueter *et al*., 2003). This derives from the fact that EST coverage is mostly from the 5′- and 3′-ends of a gene, regions that are the most difficult to predict by protein spliced alignment because of natural variation in N- and C-termini of proteins, whereas the EST-sparse internal exons are best predicted with homologous proteins (Usuka and Brendel, 2000). The limitations of this approach are, first, that this method obviously does not apply to genes that are specific to a given species; and, second, that poor choices of the protein probe will lead to unreliable predictions. The latter problem could be exacerbated if the probes themselves are erroneously predicted proteins, thus potentially propagating annotation errors (Gilks *et al*., 2002).

1.4. Combined approaches and user-contributed annotations

The problem of automated computational gene structure prediction remains challenging. In particular, it is difficult to weight different sources of prediction and evidence to derive a consistent prediction. A human expert can often easily enough distinguish solid and plausible evidence from the spurious or mistaken. Examples of this include an unmasked poly-A tail in an EST sequence matched erroneously to an adenine-rich segment of genomic DNA in a spliced alignment, a missed U12-type intron, or a chance alignment of a low-complexity protein region. But large-scale applications of computational gene structure prediction will be hard-pressed to find program parameter settings that work for the special cases. It is to be hoped that wide community participation in expert-based gene structure annotation and editing will result in highly reliable annotations for the fully sequenced plant genomes, providing a foundation for further development of computational approaches (allowing model training on error-free data) and of homology-based annotation of subsequently sequenced genomes.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *The Plant Cell*, **12**, 1021–1029.

Brendel V, Carle-Urioste JC and Walbot V (1998) Intron recognition in plants. In *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants*, Bailey-Serres J and Gallie DR (Eds.), American Society of Plant Physiologists: Rockville, pp. 20–28.

Brendel V, Xing L and Zhu W (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, **20**, 1157–1169.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Chen CL, Liang D, Zhou H, Zhuo M, Chen YQ and Qu LH (2003) The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Research*, **31**, 2601–2613.

Dong Q, Schlueter SD and Brendel V (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research*, **32**, D354–D359.

Florea L, Hartzell G, Zhang Z, Rubin GM and Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, **8**, 967–974.

Fu H and Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9573–9578.

Gilks WR, Audit B, De Angelis D, Tsoka S and Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Haas B, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al* (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.

Huang X, Adams MD, Zhou H and Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. *Genomics*, **46**, 37–45.

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Llave C, Kasschau KD, Rector MA and Carrington JC (2002) Endogenous and silencing-associated small RNAs in plants. *The Plant Cell*, **14**, 1605–1619.

Lukashin AV and Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research*, **26**, 1107–1115.

Marker C, Zemann A, Terhorst T, Kiefmann M, Kastenmayer JP, Green P, Bachellerie JP, Brosius J and Hüttenhofer A (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Current Biology*, **12**, 2002–2013.

Mathé C, Sagot MF, Schiex T and Rouzé P (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, **30**, 4103–4117.

Murakami K and Takagi T (1998) Gene recognition by combination of several gene-finding programs. *Bioinformatics*, **14**, 665–675.

Pavy N, Rombauts S, Déhais P, Mathé C, Ramana DV, Leroy P and Rouzé P (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887–899.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B and Bartel DP (2002) MicroRNAs in plants. *Genes and Development*, **16**, 1616–1626. Erratum in: *Genes and Development*, **16**, 2313.

Salzberg SL, Pertea M, Delcher AL, Gardner MJ and Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24–31.

Schlueter SD, Dong Q and Brendel V (2003) GeneSeqer@PlantGDB – gene structure prediction in plant genomes. *Nucleic Acids Research*, **31**, 3597–3600.

Tarchini R, Biddle P, Wineland R, Tingey S and Rafalski A (2000) The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. *The Plant Cell*, **12**, 381–391.

Usuka J and Brendel V (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *Journal of Molecular Biology*, **297**, 1075–1085.

Usuka J, Zhu W and Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

Wortman J, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, *et al* (2003) Annotation of the *Arabidopsis* genome. *Plant Physiology*, **132**, 461–468.

Zhu W, Schlueter SD and Brendel V (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiology*, **132**, 469–484.

Eukaryotic regulatory sequences

Mathias Krull
BIOBASE GmbH, Wolfenbüttel, Germany

Edgar Wingender
University of Göttingen, Göttingen, Germany
BIOBASE GmbH, Wolfenbüttel, Germany

1. Introduction

In metazoan organisms, cells have to express specific sets of genes to maintain cellular housekeeping as well as the cell's specific identity and to respond to external stimuli that induce developmental differentiation, growth, and survival. To achieve the regulatory specificity, transcription of eukaryotic genes is controlled by a complex modular machinery of interacting proteins, which is not yet completely understood. The proteins, termed *transcription factors*, mediate their effects by targeting specific sites on the DNA nucleotide sequence, the regulatory sequences. The location and distance to the regulated gene differs with the type of the control element. Elements that constitute the core promoter and bind the RNA polymerase II (Pol II) and the general transcription factors are situated in the close vicinity of the transcription start site (TSS) (Butler and Kadonaga, 2002), whereas enhancers can be as distant as 50 kb up- or downstream from the TSS.

The DNA sites a transcription factor binds to do not have a unique sequence as the sites for most restriction enzymes (Stormo, 2000). The sequence patterns are degenerate so that a factor recognizes a family of sequences with variable affinity. In general, the sequences are short (5–15 bp) and occur frequently in the genome by chance (Bailey and Noble, 2003). This makes any in silico identification of "real" functional sites difficult and implies that additional mechanisms are involved in specific transcriptional regulation. According to the commonly accepted models, these are mainly the combination of sites to cis-element clusters (Frith *et al*., 2001; Bailey and Noble, 2003) and the accessibility of the clusters for transcription factors that is regulated by DNA methylation and chromatin structure (Wagner, 1999).

We give a short overview on the best characterized types of regulatory sequences and modules and on computational methods that have been applied for their genome-wide prediction, a knowledge that would bring enormous benefits for uncovering the network of transcription regulation and identifying gene targets for therapeutic intervention (Frith *et al*., 2001). Currently, prediction of promoters (*see*

Article 19, **Promoter prediction**, Volume 7) is far more successful in comparison with enhancer prediction, which is not yet possible.

2. Regulatory sequences and clusters

The core promoter of a eukaryotic gene is defined as the minimal part of contiguous DNA that is recognized by the preinitiation complex (PIC) and is sufficient for transcription initiation. The PIC consists of Pol II and the general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (Fickett and Hatzigeorgiou, 1997) that are multisubunit complexes themselves. The most prominent regulatory site is the TATA box that is typically located 25–30 bp upstream of the TSS. It binds TBP (a part of TFIID), but as all other sites it is not universal and is present only in a subset of core promoters (Butler and Kadonaga, 2002). The initiator (Inr) encompasses the TSS and is recognized by several transcription factors. The downstream core promoter element (DPE) is frequent in TATA-less promoters and is located at an invariant distance to the Inr 28–32 bp upstream of the TSS, whereas the TFIIB recognition element (BRE) is present immediately upstream of some TATA boxes (Butler and Kadonaga, 2002). CpG islands are stretches of GC-rich DNA (size of 0.5 to 2 kb) that contain multiple GC-box motifs as target sites of Sp1 and related transcription factors. They can typically be found in TATA and DPE-free core promoters of housekeeping genes and initiate transcription from multiple weak start sites. Thus, core promoters are different combinations of control elements and can as such be recognized by specific enhancers (Butler and Kadonaga, 2002).

The proximal promoter is the region ranging from −250 to +250 bp in relation to the TSS (Butler and Kadonaga, 2002). Together with the core promoter, it constitutes the promoter. It contains multiple binding sites for a subset out of the 2000 transcription factors estimated for humans. Alone, some of the sites are nonfunctional, but rather need the synergistic/antagonistic combination with one or more additional sites and are called *composite elements* (Kel-Margoulis *et al*., 2002). Single sites and composite elements tend to be clustered as higher order cis-regulatory modules, which may be a key for efficient prediction (Bailey and Noble, 2003). This model is supported by the fact that transcription factors often form multimeric functional complexes.

Enhancers are long distance (up to +/− 50 kb) regulatory elements with a length of 50–1500 bp that harbor multiple protein-binding sites targeted by a multiprotein complex, the enhancesome. The enhancesome can augment gene transcription via interaction with the promoter-bound multiprotein complex (Blackwood and Kadonaga, 1998). When the effect is repressive, the enhancer is called a *silencer*.

To prevent that enhancers regulate other unrelated genes, their influence is blocked by special regulatory sequences, termed *insulators*. They also protect a gene against the inactivating impairment by adjacent heterochromatin (Burgess-Beusse *et al*., 2002).

Also acting as boundary elements are the scaffold/matrix-attached regions (S/MARs), which are the anchor points for chromatin loops at scaffold/matrix proteins. Amongst other proteins, they also bind transcription factors and are involved in gene expression regulation (Schübeler *et al*., 1996).

Another distal regulatory sequence is the locus control region (LCR) that has influence on several genes of a locus, for example, the globin locus. For exerting its effects, it recruits a holocomplex out of chromatin-modifying enzymes, coactivators, and transcription factors that can make promoters accessible and competent for subsequent transcriptional activation by other control elements (Levings and Bungert, 2002; Blackwood and Kadonaga, 1998).

3. Prediction

A transcription factor tolerates significant variation in the sequences of its DNA binding sites. Different formats have established that try to reflect this pattern of binding specificity: consensus sequences, positional weight matrices (PWM), and profile hidden Markov models (HMM). A consensus sequence indicates the most preferred base at each position, also including the IUPAC alphabet for ambiguous bases. Its utility for site prediction has limitations due to the ambiguities in the pattern that give rise to a high number of false-positive matches on one side (Stormo, 2000) and to its rigidity causing a high rate of false-negatives on the other (Quandt *et al*., 1995). A PWM assigns a weight to all bases at each position of a site and returns a site score that is the sum of these weights. Ideally, the score gives an estimation of the free energy of the protein–DNA binding (Fickett and Hatzigeorgiou, 1997). A prerequisite, and often a bottleneck, for matrix construction is the availability of an adequate number of experimentally proven binding sites for a particular factor. Collections of known binding sites can be obtained from databases such as TRANSFAC (Matys, 2003).

HMMs are generative models for describing the probability distribution over a family of sequences. They are suited to model substitutions, insertions, and deletions very well, with the limitation that dependencies between particular positions in the sequence cannot be included (Eddy, 1998).

PWMs and profile HMMs can identify in vitro target sequences accurately but cannot predict sites with in vivo function alone. Besides knowledge about chromatin structure and protein–protein interactions, the identification of regulatory clusters and the construction of predictive models can help. Several algorithms for predicting regulatory clusters have been proposed (e.g., Bailey and Noble, 2003; Frith *et al*., 2001; Wagner, 1999). They can be grouped in three classes, two of which are generative in that they rely upon a rule set of a cluster model. They use a sliding window approach and HMMs, respectively. The third approach is discriminative and models the difference between regulatory and nonregulatory sequences (Bailey and Noble, 2003).

Phylogenetic footprinting for the identification of conserved regions in orthologous gene sequences may give supporting evidence for predicted transcription factor binding sites (Lenhard *et al*., 2003). Sets of coregulated genes identified by gene expression arrays are important sources for the systematic analysis of regulatory sequences by pattern matching and approaches as explained above, but also for pattern identification algorithms that may find new regulatory elements (*see* Article 28, **Computational motif discovery**, Volume 7).

Related articles

Article 29, **Pseudogenes**, Volume 3; Article 33, **Transcriptional promoters**, Volume 3; Article 19, **Promoter prediction**, Volume 7; Article 28, **Computational motif discovery**, Volume 7

References

Bailey TL and Noble WS (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19**, ii16–ii25.

Blackwood EM and Kadonaga JT (1998) Going the distance: a current view of enhancer action. *Science*, **281**, 60–63.

Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, Recillas-Targa F, Simpson M, West A and Felsenfeld G (2002) The insulation of genes from external enhancers and silencing chromatin. *Proceedings of the National Academy of Science of the United States of America*, **99**, 16433–16437.

Butler JEF and Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, **16**, 2583–2592.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fickett JW and Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Research*, **7**, 861–878.

Frith MC, Hansen U and Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV and Wingender E (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Research*, **30**, 332–334.

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N and Wasserman WW (2003) Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology*, **2**, 13. **http://jbiol.com/content/2/2/13**.

Levings PP and Bungert J (2002) The human beta-globin locus control region. *European Journal of Biochemistry*, **269**, 1589–1599.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, *et al*. (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**, 374–378.

Quandt K, Frech K, Karas H, Wingender E and Werner T (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research*, **23**, 4878–4884.

Schübeler D, Mielke C, Maass K and Bode J (1996) Scaffold/matrix-attached regions act upon transcription in a context-dependent manner. *Biochemistry*, **35**, 11160–11169.

Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Wagner A (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Short Specialist Review

Alternative splicing in humans

John G. Conboy and Inna Dubchak
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

1. Introduction

In the human genome, the protein coding sequence of most genes is organized into discrete blocks of sequence, the exons, which are interrupted by noncoding sequences called introns. Exons and introns are transcribed together as long nuclear pre-mRNAs that must be spliced so as to excise the introns and ligate the exons, forming mature mRNAs that are exported to the cytoplasm for translation. For small genes, this process of RNA splicing is often hard-wired so that the same array of exons is always joined to create a single unique mRNA. Many large eukaryotic genes, in contrast, exhibit alternative competing pathways for splicing, generating multiple mRNA isoforms that differ with respect to inclusion or exclusion of selected exons or parts of exons. Moreover, because alternative exons can encode unique peptides or introduce new translation initiation or termination signals, any aspect of protein structure and function may be affected, having important physiological impacts on virtually every major cellular process. In recognition of the critical role of splicing as a key mediator of gene expression, its discovery was celebrated with the 1993 Nobel Prize award in Physiology or Medicine to Richard Roberts and Phillip Sharp.

Others have reviewed the structure and function of the eukaryotic spliceosomal machinery and the biochemical mechanisms involved in catalysis of the splicing reactions (Black, 2003). Here, we will focus briefly on the importance of alternative splicing in the human genome. It is now known that a majority of human genes are alternatively spliced with respect to one or more exons (Modrek and Lee, 2002), and that this phenomenon allows a relatively small genome to encode an enormously complex proteome. A number of efforts are under way to catalog and classify human alternative splicing events on a genome-wide basis, and to compare alternative splicing among genomes of other species (e.g., Lee *et al*., 2003; Nurtdinov *et al*., 2003; Pospisil *et al*., 2004; Thanaraj *et al*., 2004). Some splicing events are poorly conserved phylogenetically and may represent experiments of nature, while others are very highly conserved through evolution and likely regulate important cellular functions. Many conserved splicing events are tightly regulated in cell type– and differentiation stage–specific patterns, so that individual exons or sets of exons can be switched on or off in response to appropriate environmental signals. Moreover, regulation of alternative splicing is functionally coupled to

other RNA processing events including transcription initiation and polyadenylation (Hirose and Manley, 2000; Maniatis and Reed, 2002). A major challenge in the splicing field remains identification and characterization of the regulatory networks that control alternative splicing switches and thereby play important roles in defining the specialized properties of differentiated cells.

2. Cis-acting regulatory sequences in the genome

The decision to activate splicing of an alternative exon requires mechanism(s) for recruiting spliceosomal machinery to the appropriate splice sites, a daunting task given the observed degeneracy of splice site consensus sequences and the abundance of cryptic splice sites in the genome. The cellular splicing machinery meets this challenge by employing not only the splice sites themselves but also accessory regulatory elements in the exon and/or its flanking introns, for the process of "exon definition" (Robberson *et al*., 1990). Analogous to regulatory elements that stimulate or repress transcription, positively acting enhancer elements and negatively acting silencer elements can exert dramatic effects on the control of pre-mRNA splicing by promoting or inhibiting spliceosome recruitment. While exons are generally enriched in enhancers and introns contain abundant silencers (Cartegni *et al*., 2002; Fairbrother and Chasin, 2000; Fairbrother *et al*., 2002), tissue-specific alternative exons often contain both types of elements as an integral part of the regulatory mechanism. As a result, exon sequences are evolutionarily constrained by the need to ensure their own splicing via retention of appropriate enhancers and silencers.

3. Trans-acting splicing factors

A major paradigm in pre-mRNA splicing regulation is the concept that splicing is controlled by mutually antagonistic splicing factor proteins that bind to enhancer and silencer elements in pre-mRNA, in order to control recruitment of the spliceosomes. Typically, these proteins fall into two classes, the heterogeneous nuclear ribonucleoprotein (hnRNP) proteins, and the SR proteins characterized by domains rich in arginine and serine. Splicing activator proteins typically act as spliceosome recruiting proteins by binding to regulatory site(s) in or near an alternative exon (via a consensus RNA binding motif), and to a component(s) of the spliceosome (via a protein interaction domain). Interestingly, recruitment itself can involve any one of several different adaptors interacting with distinct spliceosomal proteins, a flexibility that may enable a cell to switch on specific sets of alternative exons in response to specific environmental signals. Equally important to splicing regulation are splicing silencer proteins that can interfere with spliceosome recruitment by blocking activity of the positive regulators. This interplay between splicing activators and inhibitors, termed "dynamic antagonism" (Charlet *et al*., 2002), plays a critical role in regulating cell type–specific splicing.

4. Alternative splicing and normal development

Implementation of alternative splicing switches during differentiation and development requires signaling mechanisms to transduce extracellular stimuli into changes in expression or activity of splicing activators and inhibitors. Alternative splicing switches can be signaled by insulin, stress hormones, or antigen treatment (Stamm, 2002), and may ultimately be mediated by tissue-restricted expression of a splicing regulatory protein, or by physiological changes in expression of a widely expressed factor (e.g., Hou *et al*., 2002; Jensen *et al*., 2000; Ladd *et al*., 2001). However, at present these mechanisms are poorly understood.

5. Alternative splicing and human disease

This subject is clearly presented in a recent review (Faustino and Cooper, 2003). Mutations that alter normal splice sites or create aberrant splice sites have long been known to underlie many human diseases. More recently it has been appreciated that subtle exon mutations and distant intron mutations can also cause disease by disrupting enhancer or silencer elements. Alternatively, disturbances in the splicing machinery itself, whether engineered in mouse models or occurring naturally in human disease states, have been shown to affect splicing patterns of many genes and cause global disturbances of cellular metabolism. Quantitative shifts in the relative concentrations of enhancer and silencer proteins could alter splice site selection of key growth regulatory genes, DNA repair functions, or apoptosis genes, and contribute to cancer progression. Many individual reports of altered splicing in cancer, and two broad bioinformatic surveys of cancer-associated splicing defects, have been published recently (Wang *et al*., 2003; Xu and Lee, 2003).

6. Future perspectives

This is an exciting time in the alternative splicing field, as new approaches offer great promise for a better understanding of the extent of alternative splicing in the human genome and better knowledge of the mechanisms by which alternative splicing is regulated. Advances in genome-wide sequence analysis and microarray technologies offer opportunities to deduce global patterns of alternative splicing in normal tissues (Johnson *et al*., 2003). This will provide a more detailed and perhaps more accurate view of splicing than what is available today from studies that are often derived from transformed cell lines or tumors that may exhibit aberrant splicing compared to normal differentiated cells. Although computational analysis of splicing regulatory elements has already begun (Brudno *et al*., 2001; Cartegni *et al*., 2003b; Fairbrother *et al*., 2002), improved methods should increase our ability to mine this sequence data for splicing regulatory signals. Together with new biochemical and genetic methods for identifying the targets of splicing regulator proteins, these advances should ultimately lead to a map of the splicing regulatory networks that control alternative splicing. Finally, the next several years should

see significant progress in understanding splicing aberrations in human disease and, more importantly, in the development of exciting new potential therapies for the correction of such splicing defects (Cartegni and Krainer, 2003a; Sazani and Kole, 2003).

References

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, **72**, 291–336.

Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I and Conboy JG (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Research*, **29**, 2338–2348.

Cartegni L, Chew SL and Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, **3**, 285–298.

Cartegni L and Krainer AR (2003a) Correction of disease-associated exon skipping by synthetic exon-specific activators. *Nature Structural Biology*, **10**, 120–125.

Cartegni L, Wang J, Zhu Z, Zhang MQ and Krainer AR (2003b) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, **31**, 3568–3571.

Charlet BN, Logan P, Singh G and Cooper TA (2002) Dynamic antagonism between ETR-3 and PTB regulates cell type-specific alternative splicing. *Molecular Cell*, **9**, 649–658.

Fairbrother WG and Chasin LA (2000) Human genomic sequences that inhibit splicing. *Molecular Cell Biology*, **20**, 6816–6825.

Fairbrother WG, Yeh RF, Sharp PA and Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

Faustino NA and Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes & Development*, **17**, 419–437.

Hirose Y and Manley JL (2000) RNA polymerase II and the integration of nuclear events. *Genes & Development*, **14**, 1415–1429.

Hou VC, Lersch R, Gee SL, Ponthier JL, Lo AJ, Wu M, Turck CW, Koury M, Krainer AR, Mayeda A, *et al*. (2002) Decrease in hnRNP A/B expression during erythropoiesis mediates a pre-mRNA splicing switch. *The EMBO Journal*, **21**, 6195–6204.

Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY and Darnell RB (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, **25**, 359–371.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R and Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.

Ladd AN, Charlet N and Cooper TA (2001) The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Molecular Cell Biology*, **21**, 1285–1296.

Lee C, Atanelov L, Modrek B and Xing Y (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Research*, **31**, 101–105.

Maniatis T and Reed R (2002) An extensive network of coupling among gene expression machines. *Nature*, **416**, 499–506.

Modrek B and Lee C (2002) A genomic view of alternative splicing. *Nature Genetics*, **30**, 13–19.

Nurtdinov RN, Artamonova II, Mironov AA and Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, **12**, 1313–1320.

Pospisil H, Herrmann A, Bortfeldt RH and Reich JG (2004) EASED: extended alternatively spliced EST database. *Nucleic Acids Research*, **32**, Database issue, D70–D74.

Robberson BL, Cote GJ and Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Molecular Cell Biology*, **10**, 84–94.

Sazani P and Kole R (2003) Modulation of alternative splicing by antisense oligonucleotides. *Progress in Molecular and Subcellular Biology*, **31**, 217–239.

Stamm S (2002) Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Human Molecular Genetics*, **11**, 2409–2416.

Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Le Texier V and Muilu J (2004) ASD: the alternative splicing database. *Nucleic Acids Research*, **32**, Database issue, D64–D69.

Wang H, Lo S, Yang H, Gere S, Hu Y, Buetow KH and Lee MP (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research*, **63**, 655–657.

Xu Q and Lee C (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, **31**, 5635–5643.

Short Specialist Review

Exonic splicing enhancers and exonic splicing silencers

Stephen M. Mount
University of Maryland, College Park, MD, USA

1. Signals affecting the splicing of messenger RNA precursors

RNA splicing (*see* Article 30, **Alternative splicing: conservation and function**, Volume 3) is the process by which some sections of a primary RNA transcript (the introns) are removed, and those sections that are retained (the exons) are joined together. Splicing is carried out by the spliceosome, a large macromolecular machine consisting of five spliceosomal RNAs (U1, U2, U4, U5, and U6) and perhaps as many as 100 proteins. The assembly of a spliceosome on the pair of splice sites flanking an intron involves the recognition of splicing signals in the messenger RNA precursor (pre-mRNA) by splicing factors. Any element in the DNA sequence of a gene that helps to specify the accurate splicing of its primary RNA transcript to generate the mature RNA product is a splicing signal. In the case of genes encoding proteins, splicing signals are at the splice sites themselves and in the exons and introns flanking splice sites. Splicing signals that promote splicing and lie within exons are known as exonic splicing enhancers (ESEs), while signals within exons that repress splicing are known as exonic splicing silencers (ESSs).

5′ splice sites (those at the 5′ boundary of an intron) are generally similar to each other at the last three exon nucleotides and the first seven intron nucleotides, so that the sequence spanning the splice site resembles MAG|**GT**RAGTA (where M indicates A or C and R indicates A or G; the Ts here would be U in the RNA sequence). The underlined GT dinucleotide is invariant, or nearly so. This consensus is complementary to the 5′ end of the U1 small nuclear RNA, which is a component of the U1 snRNP (small nuclear ribonucleoprotein). Although recognition of 5′ splice sites during the initial stages of splicing is primarily accomplished by the U1 snRNP, it does not act in isolation. Not all sites bound by the U1 snRNP are ultimately used as 5′ splice sites. Factors bound to other sites on the pre-mRNA, some of which are discussed below, can promote binding between U1 snRNP and the 5′ splice site or can facilitate progression along the pathway toward splicing. Furthermore, the 5′ splice site is later "examined" by additional factors in the course of splicing. One such factor is U6 snRNA, which displaces U1 prior to the first catalytic step, and remains associated with the 5′ splice site throughout the

remainder of the splicing reaction. In general, initial selection of the 5′ splice site by the U1 snRNP is influenced by other factors, and must be followed by appropriate interactions between the 5′ splice site and other components of the spliceosome.

The 3′ splice site usually occurs immediately 3′ of the trinucleotide C**AG** or U**AG** (and occasionally AAG, but almost never GAG). In addition, there is some conservation of the first nucleotide of the exon adjacent to the 3′ splice site, bringing the number of conserved 3′ splice site nucleotides to four (Y**AG**|G, where Y is C or U and the underlined AG dinucleotide is invariant, or nearly so). How is it possible for the 3′ splice site to be specified by so few nucleotides? The answer to this question lies in the sequence immediately upstream of the 3′ splice site. Additional signals here include the branch site, which is the site where the 5′ end of the intron is joined to the intron via an unusual 2′5′ phosphodiester bond in the first step of splicing. The branch site is typically 16 to 45 nucleotides upstream of the 3′ splice site. The branch site is first recognized by a protein known as SF1 and later associates with U2 snRNA. The region between the branch site and the 3′ splice site is bound by the large subunit of U2AF (U2 auxiliary factor), a dimeric protein that later recruits the U2 snRNP to the branchpoint immediately upstream. The small subunit of U2AF binds to the 3′ splice site itself. As with the U1 snRNP, recruitment of U2AF by other factors plays an important role in the selection of splice sites. All of these factors are highly conserved among eukaryotes, but there are considerable differences among species with regard to the extent to which different components of the multipart 3′ splice site signal are conserved.

2. Exonic splicing enhancers

Auxiliary signals that are distinct from the core splice site signals described above also influence the outcome of pre-mRNA splicing. Prominent among these are ESEs, sequences in flanking exons that are required for splicing to occur, either in vivo or in vitro. Although such splicing enhancers have been identified in both exons and introns, ESEs are generally better characterized, and are probably more common. ESEs activate nearby splice sites (both 5′ and 3′ splice sites) and promote the inclusion (vs. skipping) of exons in which they reside. Initially, ESEs were recognized as purine-rich motifs containing repeated GAR (GAA or GAG) trinucleotides. However, many other sequences have now been shown to have enhancer activity (see Zheng, 2004 for review). A small number of well-defined enhancer-dependent splicing events (notably the IgM M2 exon and the female-specific exon of the *Drosophila melanogaster* doublesex gene) have been used by researchers in the field to define and characterize numerous splicing enhancers, and this approach is validated by the general observation that an ESE in one assay is typically active in others.

Many ESEs are bound and activated by one or more of several related splicing factors known as SR proteins (Graveley, 2000; Cartegni *et al*., 2002). SR proteins contain either one or two RNA-binding domains and “RS” domains that are characterized by numerous arginine-serine dipeptide repeats. SR proteins are not only essential for splicing but also for each of the first three recognizable steps of spliceosome assembly. In vitro, any one of the several SR proteins can restore splicing to a splicing extract lacking SR proteins. Thus, the essential functions of

individual SR proteins in splicing are at least partially redundant. However, there is considerable specificity to the activation of splicing by SR proteins through ESEs. Individual SR proteins differ with respect to the sequence-specificity of their RNA-binding domains, and with respect to their ability to recognize and activate different ESE sequences (e.g., Liu *et al*., 1998).

The relationship between sequence-specific binding by SR proteins and the activation of splicing by ESEs is complex and incompletely understood. Both restoration of splicing and activation of some enhancer-dependent splicing events by an SR protein lacking the RS domain have been reported (Zhu and Krainer, 2000). Conversely, recruitment of the RS domain of SR proteins to an RNA by means of an unrelated RNA-binding domain is sufficient to activate enhancer-dependent splicing events, implying that the role of ESEs is to recruit such an activation domain (Graveley and Maniatis, 1998). Although only a few dozen splicing events have been shown to be enhancer dependent, the existence of ESEs within constitutively spliced exons (e.g., Schaal and Maniatis, 1999) suggests that ESEs are ubiquitous, redundant, and required for all splicing events. This possibility is supported by the observation that RNA-binding activity is required in order for an SR protein to support splicing in vitro. Recent evidence that SR proteins function in mRNA export (Huang *et al*., 2004) and nonsense-mediated decay (Zhang and Krainer, 2004) suggests that ESEs may play a role in these processes as well.

3. Exonic splicing silencers

Sequences in exons that act to suppress local splicing events are known as exonic splicing silencers or ESSs. Although ESSs are less well characterized than ESEs, they may be just as important in overall splice site selection, particularly in suppressing the inclusion of pseudoexons (regions within long introns that might otherwise be incorporated into mRNA as an exon (Sun and Chasin, 2000)). In many cases, ESSs are known to be bound by hnRNP proteins (particularly hnRNP A, hnRNP H, or hnRNP I, which is also known as PTB, pyrimidine-tract binding protein), but it appears likely that the basis of ESS activity is more variable than ESE activity. In theory, an ESS (acting through whatever protein or proteins recognize it) can act either by blocking the action of one or more ESEs or by directly blocking the activation of nearby splice sites.

4. Identification of ESEs and ESSs

It appears likely that many sequences can act as splicing enhancers or silencers, and it is estimated that as many as 15–20% of random sequences 20-nt long contain a splicing enhancer (Blencowe, 2000). Originally, ESEs were identified when mutational analysis of genes revealed that sequences within exons were required for proper splicing. Significantly more sequences were found when high-throughput selections for short sequences that might act as enhancers were carried out, either in vitro (e.g., Liu *et al*., 1998) or in vivo (e.g., Coulter *et al*., 1997). The most definitive assays of this type have been carried out with extracts dependent

upon a single purified SR protein (Liu *et al*., 1998). Because these ESE assays depend on recovering the ESE as part of a spliced RNA, experimental identification of ESSs, which are not part of the spliced RNA, has lagged behind, but was recently accomplished by Wang *et al*. (2004) using a cell-sorting assay.

Methods for the computational identification of ESEs and ESSs are generally based on some measure of oligonucleotide frequency. Indeed, content-based measures for detection of coding sequence used in many genefinders (*see* Article 14, **Eukaryotic gene finding**, Volume 7 and Article 16, **Searching for genes and biologically related signals in DNA sequences**, Volume 7) (Zhang, 2002) are likely to reflect frame-independent preferences (Antezana and Kreitman, 1999), including ESEs and (the absence of) ESSs. Gibbs sampler techniques have been used to identify ESE motifs within experimental data (Liu *et al*., 1998) and are the basis of the SEE-ESE web server (**http://www.tigr.org/software/SeeEse/eseF.html**). A bivariate statistical analysis that requires both overrepresentation in constitutive exons relative to introns and overrepresentation in exons that lack strong splicing signals (RESCUE-ESE, Fairbrother *et al*., 2002) was successful in identifying hexamers whose activity was later verified experimentally and through patterns of conservation (Fairbrother *et al*., 2004a). Direct use of conservation at synonymous sites has proven useful for predicting experimentally verified ESEs in Arabidopsis (Pertea, Salzberg and Mount, in preparation). Zhang and Chasin (2004) identified ESS motifs by comparing the frequency of 8-mers in internal noncoding exons versus unspliced pseudoexons, and similar motifs were identified by the high-throughput experimental screens of Wang *et al*. (2004).

5. ESEs and ESSs in genetic disease

Web servers are available for the identification of potential animal ESEs based on in vitro selection (Cartegni *et al*., 2003) or RESCUE-ESE (Fairbrother *et al*., 2004b) and for Arabidopsis ESEs (SEE-ESE, Pertea, Salzberg and Mount, in preparation). In part due to these web servers, there is an increased recognition in recent years that mutations in ESEs can be responsible for genetic disease. As a result, there has also been an explosive growth (since 2000) in the number of papers that describe mutations that inactivate an ESE.

Related articles

Article 74, **Molecular dysmorphology**, Volume 2; Article 80, **Genetic testing and genotype–phenotype correlations**, Volume 2

Further reading

Fu X-D (2004) Towards a splicing code. *Cell*, **119**, 736–738.

Ladd AN and Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology*, **3**, 0008.1–0008.16.

Pagani F and Baralle FE (2004) Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews. Genetics*, **5**, 389–396.

References

Antezana MA and Kreitman M (1999) The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *Journal of Molecular Evolution*, **49**, 36–43.

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences*, **25**, 309–320.

Cartegni L, Chew SL and Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews. Genetics*, **3**, 285–298.

Cartegni L, Wang J, Zhu Z, Zhang MQ and Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, **31**, 3568–3571.

Coulter LR, Landree MA and Cooper TA (1997) Identification of a new class of exonic splicing enhancers by in vivo selection. *Molecular and Cellular Biology*, **17**, 2143–2150.

Fairbrother WG, Holste D, Burge CB and Sharp PA (2004a) Nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biology*, **2**, E268.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA and Burge CB (2004b) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Research*, **32**, W187–W190.

Fairbrother WG, Yeh R-F, Sharp PPA and Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

Graveley B (2000) Sorting out the complexity of SR protein function. *RNA*, **6**, 1197–1211.

Graveley B and Maniatis T (1998) RS domains of SR proteins can function as activators of pre-mRNA splicing. *Molecules and Cells*, **1**, 1765–1771.

Huang Y, Yario TA and Steitz JA (2004) A molecular link between SR protein dephosphorylation and mRNA export. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9666–9670.

Liu HX, Zhang MQ and Krainer AR (1998) Identification and functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes & Development*, **12**, 1998–2012.

Schaal TD and Maniatis T (1999) Multiple distinct enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Molecular and Cellular Biology*, **19**, 1705–1719.

Sun H and Chasin LA (2000) Multiple splicing defects in an intronic false exon. *Molecular and Cellular Biology*, **20**, 6414–6425.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M and Burge C (2004) Systematic identification and analysis of exonic splicing enhancers. *Cell*, **119**, 831–845.

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. *Nature Reviews. Genetics*, **3**, 698–709.

Zhang XH-F and Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Development*, **18**, 1241–1250.

Zhang Z and Krainer AR (2004) Involvement of SR proteins in mRNA surveillance. *Molecules and Cells*, **16**, 597–607.

Zheng Z-M (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *Journal of Biomedical Science*, **11**, 278–284.

Zhu J and Krainer AR (2000) Pre-mRNA splicing in the absence of an SR protein RS domain. *Genes & Development*, **14**, 3166–3178.

Short Specialist Review

Gene finding using multiple related species: a classification approach

Manolis Kellis

MIT Broad Institute for Biomedical Research, Cambridge, MA, USA

1. Introduction

Ideally, we should be able to systematically discover all the functional genes in a newly sequenced genome from its sequence alone. Computational discovery methods rely both on the direct signals used by the cell to guide transcription, splicing, and translation, and also on indirect signals such as evolutionary conservation. In this paper, we summarize the principles of a classification-based approach for systematic gene identification, on the basis of comparative sequence information from multiple, closely related species. We first frame gene identification as a classification problem of distinguishing real genes from spurious gene predictions. We then present the Reading Frame Conservation (RFC) test, a new computational method implementing such a classification approach, on the basis of the patterns of nucleotide changes in the alignment of orthologous regions. We finally summarize our results of applying this method to reannotate the yeast genome, and the challenges of using related methods to discover all functional genes in the human genome.

2. Defining real genes

What is a real gene? This question is relatively easy to answer for those genes that are abundantly expressed, encode well-characterized proteins, and whose disruption affects a specific function in the organism. Beyond these, the distinction is much more subtle between functional genes and spurious gene predictions.

Experimentally, the definition of a functional gene comes largely as accumulating evidence of its usage, including gene expression (Velculescu, 1997), protein fragments (Rezaul *et al*., 2005), biochemical function (Jackman *et al*., 2003), protein interactions (Bai and Elledge, 1997), or the effect of its disruption (McAlister and Holland, 1982). It is important to note, however, that absence of experimental evidence does not imply that a gene annotation is spurious: a real

gene may be missing experimental evidence because it is not used in the particular conditions surveyed. Conversely, any individual report of gene usage could be due to experimental noise, cross-hybridization, or chance transcription events due to a basal level of intergenic transcription. Thus, experimental evidence alone is insufficient to distinguish between real genes and spurious gene predictions.

Computationally, the processes of transcription, splicing, and translation can be thought of as a series of decisions taken by the cell, based on signals in the genome. These signals include distant enhancer elements, regulatory elements surrounding the transcription start site, splicing enhancer and repressor signals in the message, and translation signals in the sequence or structure of mature mRNAs (Fairbrother *et al*., 2002; Wang, 2004; Thanaraj and Robinson, 2000; Yeo and Burge, 2004). The subset of these signals that we currently understand is insufficient to specify the set of known genes. Thus, in addition to the *direct* signals used by the cell, gene identification methods (Burge and Karlin, 1997; Majoros *et al*., 2004; Kulp *et al*., 1996; Krogh, 1997; Henderson *et al*., 1997; Stanke and Waack, 2003; Salzberg *et al*., 1998) routinely identify genes using additional *indirect* signals (reviewed in Fickett and Tung, 1992), which are not available to the cell, although they are generally good indicators of protein-coding genes. These include the frequency of each codon in protein-coding regions, the overall length of the translated protein product, and importantly, the evolutionary conservation of protein sequences across related species.

Evolutionary conservation is perhaps the strongest indicator that a predicted gene is functional. A gene that confers even a slight evolutionary advantage can be conserved across millions of years, regardless of the rarity of its usage. Hence, even if experimental methods fail to detect the usage of a gene in a given set of experimental conditions, evolutionary methods are able to detect *indirect* evidence of its usage, by observing the pressure to preserve its function over millions of years.

3. Gene finding as a classification problem

The challenge of using evolutionary conservation to define genes lies in the ability to reject spurious genes. Although it is generally well understood that genes conserved across large evolutionary distances are functional, lack of conservation is generally attributed to evolutionary divergence, thus providing no evidence toward either accepting or rejecting a gene prediction.

Yet, the comparison of related genomes contains information much richer than simply the presence or absence of a protein in a genome. By working with closely related genomes, we can define regions of conserved gene order, or *synteny blocks*, which span several genes, and potentially entire chromosomes. Defining conserved synteny blocks allows us to construct global alignments, spanning both well-conserved regions and regions of low conservation. In particular, these give us access to full nucleotide alignments for all predicted genes and for all intergenic regions within orthologous segments.

With the availability of orthologous alignments for both protein-coding and noncoding regions, we can study their distinct properties and build a classifier between the two types of region. The simplest and most commonly used such classifier

observes the overall level of nucleotide conservation in the alignment, and selects regions of high conservation that are likely to be functional. Other classifiers may observe more subtle signals, such as the number of amino acid changes per nucleotide substitution (Hurst, 2002), the frequency of insertions and deletions, the periodicity of mutations, and so on. By working with properties of the nucleotide alignment, rather than protein alignment, we are able to apply it uniformly to evaluate both genes and intergenic regions in the same test.

One particular conservation property unique to protein-coding segments is the pressure to preserve the reading frame of translation. Since protein sequences are translated every three nucleotides, the length of insertions and deletions (indels) is largely constrained to remain a multiple of three, thus preserving the frame of translation. Within coding regions, indels that disrupt the frame of translation are excluded, or compensated with nearby indels that restore the reading frame. In noncoding regions, the length of indels does not have this constraint, and short spacing changes are tolerated.

To evaluate this property quantitatively, we developed the RFC test. This test evaluates the pressure to preserve the reading frame in a fully aligned interval, by measuring the portion of nucleotides in this interval for which the reading frame has been locally conserved (Kellis *et al*., 2004). The RFC test provides a classifier between coding and noncoding regions that is completely independent of the traditional signals used in gene identification. It does not rely on start, stop, or splicing signals, nor does it rely on the conservation of protein sequence. It can, therefore, be combined with existing gene-finding tools and provide a highly informative score for any interval considered.

4. Reannotation of yeast

A classification approach is particularly well suited for the yeast genome. The general scarcity of introns makes it hard to rely on splicing signals to discover genes. Thus, the annotation of *Saccharomyces cerevisiae* has traditionally relied solely on the length of predicted proteins to annotate genes, resulting in 6062 annotated open reading frames (ORFs), which potentially encode proteins of at least 100 amino acids (aa). Additionally, a tentative functional annotation has been inferred for as many as 3966 of these ORFs, on the basis of classical genetic experiments and systematic genome-wide studies of gene expression, deletion phenotype, and protein–protein interaction. Together, the interval-based annotation and the large set of well-known genes make it possible to apply the RFC test systematically to evaluate the functional significance of the remaining ORFs.

To apply the test, we constructed multiple sequence alignments for every ORF and every intergenic region across four closely related species. We sequenced and assembled the genomes of *S. paradoxus*, *S. mikatae*, *and S. bayanus* (Kellis *et al*., 2003), and defined genome-wide synteny blocks with *S. cerevisiae*, on the basis of discrete anchors provided by unique protein blast hits (Kellis *et al*., 2004). Within these synteny blocks, we constructed global alignments of orthologous genes and intergenic regions across the four species using CLUSTALW (Higgins and Sharp, 1988), and systematically evaluated each alignment. We compared *S. cerevisiae* to

each species in turn, and every comparison cast a vote based on its overall RFC and a species-specific cutoff. A decision was then reached for each gene by tallying the votes from all comparisons.

We evaluated the sensitivity and specificity of the approach based on the 3966 genes with functional annotations ("known genes"), and 340 randomly chosen intergenic sequences with lengths similar to the annotated ORFs. The RFC test correctly accepted 3951 known genes (99.6%) and rejected only 15 known genes; upon manual inspection, these 15 are indeed likely to be spurious (most lack experimental evidence, and deletion phenotypes of the rest are likely due to their overlap with the promoter of other known genes). The method also correctly rejected 326 intergenic regions (96%), accepting only 14 intergenic regions (of which 10 appear to define short ORFs or extend annotated ORFs, suggesting that at most 1% of true intergenic regions failed to be rejected by the RFC test). In summary, the RFC test shows a very strong discrimination between genes and intergenic regions, with sensitivity and specificity values greater than 99%.

We then applied the RFC test to all previously annotated genes, leading to a major revisiting of the yeast gene set. For ORFs with lengths greater than 100 aa, our analysis accepted 5538 ORFs and rejected 503 (of which 376 were immediately rejected, 105 were rejected with additional criteria, and 32 were merged with neighboring ORFs); the classifier abstained from making a decision in 20 cases. The rejected ORFs show an abundance of frame-shifting indels across their entire length, in-frame stop codons, and low conservation of protein sequence, in addition to the low RFC score; their length distribution and atypical codon usage additionally suggest that they are likely occurring by chance (Goffeau, 1996; Dujon, 1994; Sharp and Li, 1987); furthermore, previous systematic experimentation showed no compelling evidence that these may encode a functional gene. Thus, it appears that more than 500 previously annotated ORFs in the yeast genome are spurious predictions.

Below the 100 amino acid cutoff, no previous systematic annotation or experimentation was available. Thus, to validate our method, we compared the results of the RFC test with an independent metric: the proportion of the *S. cerevisiae* ORF that was also free of stop codons in the other three species. Between 50 and 99 aa, we found that the method is still reliable and reports 43 candidate new genes at that length. As ORFs smaller than 50 aa were tested, we found that the specificity of the test decreased, since small intervals tend to be devoid of indels by chance rather than by presence of selective pressure. Thus, additional constraints, and additional species, will be needed to discover genes reliably at such short lengths.

In addition to the discovery of genes themselves, the comparative analysis allowed us to refine gene boundaries. Once an interval was determined to be under selective pressure for RFC, the boundaries of that interval were adjusted on the basis of the conservation of start/stop/splice signals and the boundaries over which the reading frame is conserved. This led to a large-scale reannotation of gene structure, which affected hundreds of genes (146 start codon changes, 67 intron changes, 32 merges of consecutive ORFs, and 45 changes of ORF ends). It is worth noting that in 134 cases, the inferred boundary changes pinpointed sequencing errors in the primary sequence of *S. cerevisiae*, ~50 of which were tested and

corrected by resequencing. These boundary changes reveal the true location of promoter regions, new protein domains in elongated ORFs, and previously overlooked functional relationships in the case of merged ORFs

In summary, the RFC test was able to reliably distinguish between genes and intergenic regions, leading to a systematic reannotation of the yeast genome. The comparative analysis led to the rejection of more than 500 previously annotated genes, and to the discovery of many novel genes. The results agree with similar comparative analyses carried out from a number of yeast species (Cliften *et al*., 2003; Blandin, 2000; Wood *et al*., 2001; Brachat, 2003). In addition, it allowed us to refine the gene structure of hundreds of genes, adjusting start and stop boundaries, merging consecutive ORFs, and discovering many new introns. Moreover, by using multiple species, the signals leading to gene identification were powerful enough to pinpoint sequencing errors in any individual species, including *S. cerevisiae*.

5. Implications for the human genome

The challenge of gene finding in the human genome is far greater than in yeast, due to the vastly larger intergenic regions, numerous and large introns, small exons, and alternatively spliced genes. Most approaches to gene finding in higher eukaryotes have relied on Hidden Markov Models (Burge and Karlin, 1997; Majoros *et al*., 2004; Kulp *et al*., 1996; Krogh, 1997; Henderson *et al*., 1997; Stanke and Waack, 2003), which inherently emphasize the importance of exon chaining and rely on knowledge of the expected length distributions for both exons and introns. These have been recently extended to use sequences from multiple species (Yeh *et al*., 2001; Siepel and Haussler, 2004a; Siepel and Haussler, 2004b; Batzoglou *et al*., 2000; Korf *et al*., 2001; Dewey *et al*., 2004; Rinner and Morgenstern, 2002; Parra, 2003; Meyer and Durbin, 2002; Novichkov *et al*., 2001). These approaches have limitations however, in cases of alternative splicing (Brett, 2000; Mironov *et al*., 1999; Cawley and Pachter, 2003), differences in splicing between species (Nurtdinov *et al*., 2003), widely varying exon and intron lengths, and noncanonical splice sites.

An exon-based classification approach to gene finding can help overcome these limitations. By exhaustively enumerating and testing all candidate protein-coding intervals, classification approaches make exon detection independent of chaining. Splice site models can be applied to each species independently and then compared with each other, which is much stronger than the evidence in any one species alone. The intervals can be then tested on the basis of discriminating variables that distinguish genes from noncoding regions (Fickett and Tung, 1992; Goldman and Yang, 1994). New discriminating variables can be defined as alignments of each type of region are systematically compared, and these can be combined and weighted, leveraging traditional machine-learning techniques for feature selection and classification (Moore and Lake, 2003). Once relevant intervals have been identified, their boundaries can be adjusted for optimal chaining into complete genes. In particular, chaining can also leverage an inferred frame of translation for each exon, based on the higher mutation rate observed in largely degenerate third codon positions (Hurst, 2002). The exon-chaining step produces full gene models,

and is able to cope with alternative splicing and missing exons, since it is not constrained to a single optimal path through the exons.

By systematically observing alignment properties of large sequence regions, we can build new rigorous approaches for sequence analysis. These are widely applicable beyond coding exons, and similar classification-based approaches can be used to distinguish CpG islands, 5′- and 3′-untranslated regions, promoter regions, regulatory islands, and other functional elements. Through the lens of evolutionary selection, our ability to directly interpret genomes is revolutionized. Coupled with systematic experimentation and validation, these analyses can lead to a systematic catalog of functional elements in the human genome, forming the future foundations of biomedical research.

References

Bai C and Elledge SJ (1997) Gene identification using the yeast two-hybrid system. *Methods in Enzymology*, **283**, 141–156.

Batzoglou S, Pachter L, Mesirov JP, Berger B and Lander ES (2000) Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, **10**, 950–958.

Blandin G, Durrens P, Tekaia F, Aigle M, Bolotin-Fukuhara M, Bon E, Casaregola S, de Montigny J, Gaillardin C, Lepingle A, *et al*. (2000) Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Letters*, **487**, 31–36.

Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T and Philippsen P (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biology*, **4**, R45.

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J and Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**, 83–86.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Cawley SL and Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19**(Suppl 2), II36–II41.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA and Johnston M (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Dewey C, Wu JQ, Cawley S, Alexandersson M, Gibbs R and Pachter L (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Research*, **14**, 661–664.

Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, Ballesta JP, Banrevi A, Bolle PA, Bolotin-Fukuhara M and Bossier P (1994) Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.

Fairbrother WG, Yeh RF, Sharp PA and Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

Fickett JW and Tung CS (1992) Assessment of protein coding measures. *Nucleic Acids Research*, **20**, 6441–6450.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Goldman N and Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.

Henderson J, Salzberg S and Fasman KH (1997) Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, **4**, 127–141.

Higgins DG and Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Hurst LD (2002) The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends in Genetics*, **18**, 486.

Jackman JE, Montange RK, Malik HS and Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *RNA*, **9**, 574–585.

Kellis M, Patterson N, Birren B, Berger B and Lander ES (2004) Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, **11**, 319–355.

Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(Suppl 1), S140–S148.

Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proceedings/International Conference on Intelligent Systems for Molecular Biology*, **5**, 179–186.

Kulp D, Haussler D, Reese MG and Eeckman FH (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proceedings/International Conference on Intelligent Systems for Molecular Biology*, **4**, 134–142.

Majoros WH, Pertea M and Salzberg SL (2004) TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

McAlister L and Holland MJ (1982) Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. *The Journal of Biological Chemistry*, **257**, 7181–7188.

Meyer IM and Durbin R (2002) Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309–1318.

Mironov AA, Fickett JW and Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Research*, **9**, 1288–1293.

Moore JE and Lake JA (2003) Gene structure prediction in syntenic DNA segments. *Nucleic Acids Research*, **31**, 7271–7279.

Novichkov PS, Gelfand MS and Mironov AA (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, **17**, 1011–1018.

Nurtdinov RN, Artamonova II, Mironov AA and Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, **12**, 1313–1320.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW and Guigo R (2003) Comparative gene prediction in human and mouse. *Genome Research*, **13**, 108–117.

Rezaul K, Wu L, Mayya V, Hwang SI and Han DK (2005) A systematic characterization of mitochondrial proteome from human T leukemia cells. *Molecular and Cellular Proteomics*, **4**, 169–181.

Rinner O and Morgenstern B (2002) AGenDA: Gene prediction by comparative sequence analysis. *In Silico Biology*, **2**, 195–205.

Salzberg S, Delcher AL, Fasman KH and Henderson J (1998) A decision tree system for finding genes in DNA. *Journal of Computational Biology*, **5**, 667–680.

Sharp PM and Li WH (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**, 1281–1295.

Siepel A and Haussler D (2004a) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, **11**, 413–428.

Siepel A and Haussler D (2004b) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, **21**, 468–488.

Stanke M and Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl 2), II215–II225.

Thanaraj TA and Robinson AJ (2000) Prediction of exact boundaries of exons. *Briefings in Bioinformatic*, **1**, 343–356.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B and Kinzler KW (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M and Burge CB (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.

Wood V, Rutherford KM, Ivens A, Rajandream M-A and Barrell B (2001) A re-annotation of the *Saccaromyces cerevisiae* genome. *Comparative and Functional Genomics*, **2**, 143–154.

Yeh RF, Lim LP and Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Research*, **11**, 803–816.

Yeo G and Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, **11**, 377–394.

Dynamic programming for gene finders

William H. Majoros
The Institute for Genomic Research, Rockville, MD, USA

1. Introduction

Dynamic programming (DP), a technique commonly used in bioinformatics algorithms to reduce the evaluation time for recurrence relations, is especially prevalent in the implementation of gene-finding software. The task of gene finding has traditionally been formulated as that of choosing a single *parse*, or collection of zero or more nonoverlapping gene models, of a sequence such that the chosen parse maximizes some objective function. In general, the number of such parses can grow exponentially with the length of the sequence, owing to the combinatorial nature of the parses – that is, n nonoverlapping and frame-consistent exons can, in theory, be combined independently to produce 2^n different gene models. Thus, the explicit enumeration and evaluation of all possible parses of a sequence is generally too computationally expensive to be undertaken in practice. However, many of the objective functions favored by practitioners, particularly those that are probabilistic, can be expressed as a recurrence relation in which the value of a parse is defined recursively in terms of the values of subparses. It is this class of formulations that are amenable to dynamic programming, or DP.

There are two main approaches to DP: *bottom-up* and *top-down*. The latter, also called *memoization*, simply caches the value returned by a recursive call the first time it is made, so that subsequent calls with the same parameters can retrieve the value from the cache rather than recomputing it from scratch. By contrast, the bottom-up approach, which is more commonly used in gene finding and is slightly more efficient, involves the proactive computation of simpler subproblems before they are explicitly needed for the evaluation of more complex subproblems. The values of these subproblems are stored in an array or matrix, with the evaluation order of the cells in the matrix being determined by the recursive structure of the recurrence relation. Once the matrix has been evaluated, the optimal value can be extracted from the appropriate cell in the matrix, and the corresponding optimal solution can be obtained by tracing back through the recurrence links that were used to compute the value of each cell (assuming such links were explicitly stored or can be reconstructed during trace-back) (Cormen *et al*., 2001; Durbin *et al*., 1998).

Thus, the main tasks in devising a DP implementation of a new gene-finding strategy are (1) to formulate the objective function as a recurrence relation, (2) to map the recursive structure of that recurrence relation to the elements of a multidimensional matrix, and (3) to determine the evaluation order of the cells in that matrix. Optimization of the objective function is then achieved by evaluating all the cells of the DP matrix and then identifying the optimal cell.

2. Example uses

Examples of the actual use of DP for gene finding are legion, particularly in Markovian systems. It is standard to employ DP in both the training and application of Hidden Markov Models (HMMs) (*see* Article 98, **Hidden Markov models and neural networks**, Volume 8). The Forward, Backward, and Baum–Welch algorithms for training an HMM involve matrices wherein each cell corresponds to a conditional or joint probability on sequences and their emission by particular states. The Viterbi decoding algorithm for parsing a sequence with an HMM likewise involves a matrix with cells denoting conditionally optimal paths through the states of the HMM and their corresponding conditional probabilities. The global optimum is identified by the cell that pairs the HMM's designated final state with the end of the sequence (Durbin *et al*., 1998).

Other uses of DP for gene finding include the decoding algorithms used in applying Generalized Hidden Markov Models (GHMMs) and Nonstationary Markov Chains, and the procedures used to train and apply neural networks (Xu and Uberbacher, 1998; Rumelhart *et al*., 1986).

3. Practical considerations

Various practical considerations can complicate the task of devising and implementing a DP algorithm. Among the most prominent are those involving space/speed trade-offs. The space requirements of multidimensional DP matrices can grow very quickly as the lengths of the dimensions increase, but often it is possible to eliminate portions of the matrix from consideration right at the outset, thereby permitting the adoption of a "sparse matrix" representation in which memory is allocated only for the portions of the matrix actually needed. Such a representation may or may not improve speed and memory efficiency, but it can also require significantly more effort during the development and maintenance of the software.

Even if the complete matrix is fully allocated in memory, it may still be advantageous to eliminate from consideration certain cells in the matrix that are unlikely to contribute to the optimal solution, similar to "banded" alignment techniques. When the probabilities of these eliminated cells are small but nonzero, the resulting procedure becomes a heuristic that is no longer guaranteed to find the optimal solution. In many cases this might be acceptable, but doing this typically involves empirical studies to establish an acceptable range of values for the cells being eliminated. This may need to be done every time the gene finder is trained for a new organism.

Devising the original recurrence relations can itself be a difficult task (Sankoff, 2000). For any given set of proposed equations, a worthwhile exercise is the derivation of a rigorous proof that the equations exhibit the intended semantics and do not permit greedy behavior. Unintended greediness could allow the algorithm to maximize the objective function as encoded without strictly adhering to the intended semantics, and could thereby compromise prediction accuracy. For example, failure to properly separate the inductive scores of the three coding phases until the end of a CDS in a Markov chain promotes greedy behavior, because the phase is not selected in a globally optimal manner (Wu, 1996; Salzberg, 1998).

Validation of the final software implementation is important as well, but can be very difficult when the expected predictions of a given DP formulation are not precisely known and would be impractical to produce manually. In these cases, subtle programming errors can go undetected unless one undertakes the time-consuming exercise of rigorously tracing through a number of sample computations to verify correct operation (Geigerich, 2000).

Another complication that not uncommonly arises in DP implementations applied to long sequences is the occurrence of *numerical underflow*, in which a value in the computation becomes so small that it cannot be represented by the host machine, thereby causing a *floating-point exception* or simply rendering gene models indistinguishable by assigning them all a score of zero. This is especially common in probabilistic formulations involving the multiplication of large numbers of probabilities. Several methods for avoiding underflow have been devised, such as arbitrarily rescaling the numbers or applying the log transformation, but some of these methods, such as rescaling, can significantly complicate the implementation and validation of the program, and generally must be derived anew for each gene finder (Durbin *et al*., 1998).

In an attempt to counter some of these obstacles, several recent works have attempted to provide tools for the automatic construction of DP algorithms and software, though whether these techniques will see widespread adoption and tangibly impact the field remains to be seen (Geigerich, 2000; Birney and Durbin, 1997).

4. Future challenges

Gene-finding researchers now face an array of challenges for which new DP algorithms need to be developed. As more genomic information becomes available, the successful migration of gene-finding techniques into the realm of comparative genomics becomes increasingly important. The use of models such as Pair HMMs (*see* Article 17, **Pair hidden Markov models**, Volume 7) for gene finding entails a potentially significant increase in computational complexity that must be addressed. Recent attempts at using these models have drawn on techniques such as the use of sparse matrices, heuristic optimization, and approximate alignments (themselves relying on DP methods), and generally have involved one or more simplifying assumptions, such as strict conservation of intron/exon structure (Ureta-Vidal *et al*., 2003).

Much current work now focuses on gene prediction in two species simultaneously; generalizing this to three or more species will require yet additional

innovation to handle the enormous computational load. More generally, one might imagine more flexible DP frameworks in which suboptimal parses are made available in an efficient manner for perusal "on demand" by human annotators or for automatic integration with external evidence, perhaps in the context of a tightly coupled ensemble of separate gene-finding programs.

One possible avenue for future progress is in the parallelization of DP algorithms to enable their deployment on multiprocessing hardware. This in itself represents a considerable challenge, as many DP algorithms are inherently difficult to parallelize, due to pervasive dependencies in the recurrence structure.

References

Birney E and Durbin R (1997) Dynamite: a flexible code-generating language for dynamic programming methods. *Proceedings of the 5th Conference on Intelligent Systems for Molecular Biology*, **5**, 56–64.

Cormen TH, Leiserson CE, Rivest RL and Stein C (2001) *Introduction to Algorithms*, Second Edition, MIT Press: Cambridge, pp. 323–369.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis*, Cambridge University Press: Cambridge, pp. 46–99.

Geigerich R (2000) A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, **16**, 655–677.

Rumelhart DE, Hinton GE, Williams RJ and McClelland and the PDP Research Group (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *Vol. 1: Foundations*, Rumelhart DE (Ed.), MIT Press: Cambridge, pp. 318–362.

Salzberg SL (1998) Decision trees and Markov chains for gene finding. In *Computational Methods in Molecular Biology*, Salzberg SL, Searls DB and Kasif S (Eds.), Elsevier: Amsterdam, pp. 187–203.

Sankoff D (2000) The early introduction of dynamic programming into computational biology. *Bioinformatics*, **16**, 41–47.

Ureta-Vidal A, Ettwiller L and Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Reviews Genetics*, **4**, 251–262.

Wu TD (1996) A segment-based dynamic programming algorithm for predicting gene structure. *Journal of Computational Biology*, **3**, 375–394.

Xu Y and Uberbacher EC (1998) Computational gene prediction using neural networks and similarity search. In *Computational Methods in Molecular Biology*, Salzberg SL, Searls DB and Kasif S (Eds.), Elsevier: Amsterdam, pp. 109–128.

Basic Techniques and Approaches

Gene structure prediction by genomic sequence alignment

Burkhard Morgenstern
University of Göttingen, Göttingen, Germany

1. Introduction

Gene prediction is the first and most fundamental step to genome analysis and annotation (*see* Article 13, **Prokaryotic gene identification *in silico***, Volume 7, Article 14, **Eukaryotic gene finding**, Volume 7, Article 21, **Gene structure prediction in plant genomes**, Volume 7, and Article 26, **Dynamic programming for gene finders**, Volume 7). Consequently, the development of gene-finding software continues to be an important area of research in Bioinformatics. Gene-structure prediction is particularly challenging for eukaryotes, in which the density of genes is low and most genes consist of multiple exons separated by introns of varying length; see Guigó (2004) for an overview on existing software for gene prediction in eukaryotes. During the last years, a large number of computer programs have been developed to identify genes in eukaryotic genome sequences. Recent studies show, however, that the reliability of these methods is limited. While most methods work well on small sequences with single genes, their performance decreases considerably when applied to BAC-sized sequences including multiple genes (Guigó *et al*., 2000).

Traditional gene-prediction programs rely on information derived from *known* genes. *Intrinsic* methods use statistical features such as ORF length or codon usage and try to identify splice junctions and other signals. The underlying models are trained using previously known genes from the same or a closely related organism. Traditional *extrinsic* methods work by comparing genomic sequences to EST or protein sequences.

With an ever-increasing number of completely or partially sequenced genomes, new *comparative* approaches have been proposed to predict genes and other functional elements. The idea is that functionally important parts of the genome are under selective pressure, so they are usually more conserved than nonfunctional regions where random mutations can be tolerated without affecting the fitness of the organism. Consequently, local sequence conservation among genomic sequences usually indicates biological functionality. This *phylogenetic footprinting* principle has been successfully applied to various problems such as detection of regulatory elements, and it is now generally accepted that comparative sequence analysis is a powerful and universally applicable tool for functional genomics.

Gene prediction by comparative sequence analysis means that instead of comparing a given sequence to previously *known* genes, one compares genomic sequences from evolutionarily related organisms to each other. Regions of local sequence conservation are then searched for relevant signals to identify possible exons and gene structures. Note that, in contrast to traditional approaches, comparative gene finding is possible without any previous knowledge about existing genes and their statistical composition – except for simple models of splice signals. Therefore, these new approaches are able to detect genes with unusual statistical features that can be missed by traditional methods.

2. Gene prediction using CHAOS, DIALIGN, and AGenDA

Comparative gene-finding approaches have been developed, for example, by Bafna and Huson (2000), Batzoglou *et al*. (2000), and Wiehe *et al*. (2001). At MIPS/GSF research center and at the University of Bielefeld, we developed a comparative gene finder called AGenDA (*Alignment-based Gene-Detection Algorithm*, Rinner and Morgenstern, 2002). The first step in comparative sequence analysis is to calculate an *alignment* of the sequences under study as the results of any comparative method can be only as good as the underlying alignment. For AGenDA, we used the alignment program *DIALIGN* (Morgenstern *et al*., 1996). This program integrates local and global alignment features by constructing alignments based on pairwise local homologies. In a number of recent research projects, DIALIGN has been applied to identify functional elements in genomic sequences.

In a preliminary study, we found that local homologies identified by DIALIGN in genomic sequences are highly correlated to protein-coding exons (Morgenstern *et al*., 2002); in this regard, DIALIGN proved to be superior to alternative alignment programs. However, since DIALIGN has originally been designed as a general-purpose aligner, the standard version of the program is too slow to align large genomic sequences. Thus, we implemented an *anchored-alignment* approach to speed up the alignment procedure. Here, we are using the fast database search tool CHAOS to identify local sequence similarities; these similarities are then used as prealigned *anchor points* to reduce the alignment search space and program running time for DIALIGN, as described by Brudno *et al*. (2003).

On a given pair of input sequences, for example, from human and mouse, AGenDA performs the following operations (for details, see Rinner and Morgenstern, 2002):

1. High-scoring sequence similarities identified by CHAOS and DIALIGN are *clustered* by bridging small gaps between them.
2. Conserved splice junctions and start/stop codons are identified around the cluster boundaries. For splice signals, standard matrices proposed by Salzberg (1997) are used.
3. A *candidate exon* (CE) is defined as a region of local sequence similarity bounded by conserved splice sites or start/stop codons, respectively. Note that a region of local sequence conservation can be bounded by more than one conserved splice signal, so it can give rise to multiple overlapping CEs, as

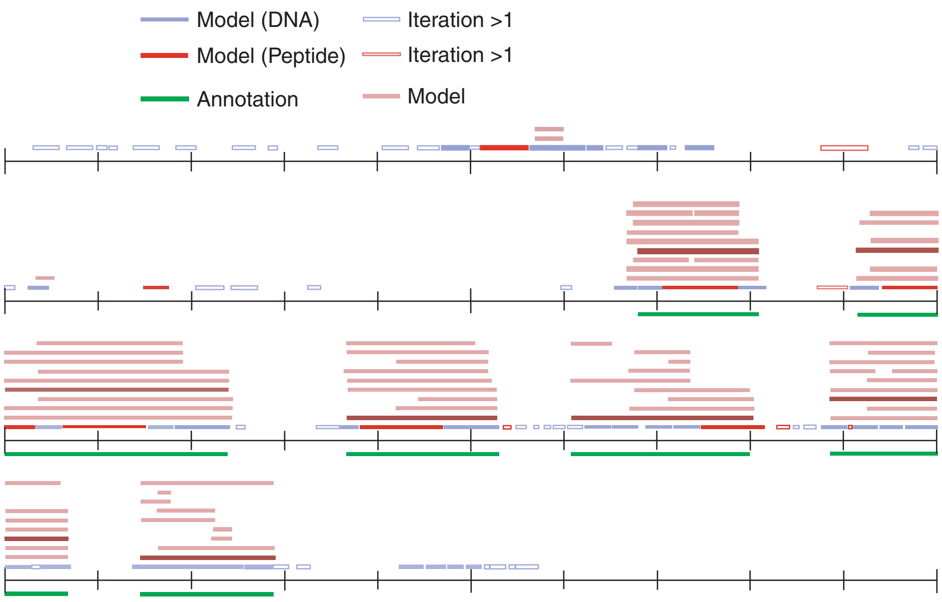


Figure 1 Gene prediction by AGenDA. A human genomic sequence of 40 kb in length has been aligned to its counterpart in the murine genome (not shown in the figure). Green lines below the sequence are known protein-coding exons. Blue and red lines above the sequence correspond to regions of local sequence similarity as identified by DIALIGN. Pink lines correspond to *candidate exons (CEs)*, that is, to local sequence similarities bounded by conserved splice sites or start/stop codons. A complete gene model is obtained as an optimal *chain* of CEs (purple lines). Local sequence conservation between human and mouse *roughly* corresponds to true exons, but is not very *specific* as many similarities outside the coding regions have been found. CEs bounded by conserved signals more accurately reflect the coding regions. Searching for optimal *chains* of CEs further reduces the noise of false-positive predictions; in our example, the optimal gene model identified by AGenDA *exactly* corresponds to the real gene

illustrated in Figure 1. Each CE is assigned a *score*, essentially depending on the level of sequence similarity and the quality of the splice signals.

4. A *candidate gene* is a chain of CEs meeting some obvious formal conditions for splice signals and start/stop codons and certain length restrictions for introns. Candidate genes are considered on the forward as well as on the reverse strand.
5. The program uses a dynamic-programming algorithm to identify a set of candidate genes with maximum total score.

3. Results and discussion

AGenDA has been evaluated on standard reference sequences from human and mouse (Batzoglou *et al*., 2000). We found that the prediction quality of AGenDA is similar to the Hidden Markov Model (HMM)-based program *GenScan* (Burge and Karlin, 1997), which is currently the most popular software tool for gene prediction in eukaryotes. It is remarkable that a method solely based on evolutionary sequence conservation yields results that are comparable to the output of sophisticated intrinsic methods. While GenScan uses species-dependent statistical models to

distinguish coding from noncoding regions, comparative methods are based on simple and universally applicable measures of local sequence similarity and on basic models for splice signals. Since these two gene-finding approaches are based on completely different types of input information, they complement each other and genes missed by one method can be detected by the other approach. A web interface for AGenDA has been set up at the Bielefeld Bioinformatics Server *BiBiServ* (Taher *et al*., 2004, **http://bibiserv.techfak.uni-bielefeld.de/agenda/**).

An obvious way of improving gene-prediction accuracy is to *combine* the predictive power of stochastical and comparative approaches. The effectiveness of such an integrated approach has been demonstrated by Korf *et al*. (2001) with their TWINSCAN program. In short, they reimplemented GenScan but also included homology information obtained from high-scoring BLAST alignments of syntenic sequences. This resulted in markedly increased sensitivity and specificity. Other combinations of intrinsic and comparative methods have been proposed by Meyer and Durbin (2002) and by Parra *et al*. (2003).

We are planning to integrate our alignment-based approach with the intrinsic gene-prediction program AUGUSTUS (Stanke and Waack, 2003). AUGUSTUS is based on a generalized HMM, with a new submodel for intron length distribution. With this new model, AUGUSTUS is superior to standard gene-finding programs if large input sequences are to be analyzed. Especially for genes with long introns, AUGUSTUS is far more accurate than other intrinsic methods. The stochastic model used by AUGUSTUS can incorporate external information in a natural way (Stanke, 2003). This way, we will use genomic alignments calculated by CHAOS and DIALIGN to further improve the performance of AUGUSTUS and to combine the advantages of these two approaches.

Acknowledgments

I would like to thank Oliver Rinner and Leila Taher for developing AGenDA, Saurabh Garg and Alexander Sczyrba for their work on the AGenDA web server, and Michael Brudno for helping us with his CHAOS software. Mario Stanke made useful comments on the manuscript.

References

Bafna V and Huson DH (2000) The conserved exon method for gene finding. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, **8**, 3–12.

Batzoglou S, Pachter L, Mesirov JP, Berger B and Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, **7**, 950–958.

Brudno M, Chapman M, Göttgens B, Batzoglou S and Morgenstern B (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Guigó R (2004) Eukaryotic gene finding. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

Guigó R, Agarwal P, Abril JF, Burset M and Fickett JW (2000) An assessment of gene prediction accuracy in Large DNA sequences. *Genome Research*, **10**, 1631–1642.

Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.

Meyer I and Durbin R (2002) Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309–1318.

Morgenstern B, Dress A and Werner T (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 12098–12103.

Morgenstern B, Rinner O, Abdeddaïm S, Haase D, Mayer K, Dress A and Mewes H-W (2002) Exon prediction by comparative sequence analysis. *Bioinformatics*, **18**, 777–787.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW and Guigó R (2003) Comparative gene prediction in human and mouse. *Genome Research*, **13**, 108–117.

Rinner O and Morgenstern B (2002) AGenDA: Gene prediction by comparative sequence analysis. *In Silico Biology*, **2**, 195–205.

Salzberg SL (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Computer Applications in the Biosciences*, **13**, 365–376.

Stanke M (2003) *Gene Prediction with a Hidden Markov Model*, PhD dissertation, University of Göttingen.

Stanke M and Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215–ii225.

Taher L, Rinner O, Gargh S, Sczyrba A and Morgenstern B (2004) AGenDA: gene prediction by cross-species sequence comparison. *Nucleic Acids Research*, **32**, W305–W308.

Wiehe T, Gebauer-Jung S, Mitchell-Olds T and Guigó R (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Research*, **11**, 1574–1583.

Computational motif discovery

Martin Tompa
University of Washington, Seattle, WA, USA

Given the vast amounts of sequence data being generated by numerous genome and proteome projects, on which regions should biologists focus their attention? The goal of computational motif discovery is to predict relatively short subsequences that are good candidates to serve some biological function. This article focuses on the computational prediction of protein-binding sites in nucleotide sequences (for instance, transcription factor binding sites in DNA promoter regions) as a specific example. However, most of the ideas carry over to other applications such as predicting protein domains, splice sites, and so on.

The binding sites to be discovered are typically quite short, between 6 and 25 nucleotides in length. If one tabulated the binding sites of a single DNA-binding protein, one would find that they are not all identical because the protein typically tolerates some variation in the sequences it binds. Instead, a typical collection of binding sites for a single protein might look something like the hypothetical example shown in Figure 1, in which some positions are well conserved (for instance, C in position 2 of each binding site) and other positions less conserved (for instance, position 1).

It is this variation among binding sites that makes computational motif discovery challenging. There are many different motif "models" that can be adopted in an attempt to characterize the variation. One of the simplest such models is the *consensus string*. For instance, the eight binding sites in Figure 1 can be described as being near the consensus ACGTCGT, allowing up to two nucleotide substitutions from this consensus. A second simple motif model is the *IUPAC* description. Each of the binding sites in Figure 1 is an instance of the IUPAC string RCGWYGW, where R stands for either of the purines A or G, Y stands for either of the pyrimidines C or T, and W stands for either of the weakly binding nucleotides A or T.

Both the consensus and IUPAC motif models have the advantage that motif discovery algorithms can systematically enumerate all possible motifs, since a motif is characterized by a single consensus or IUPAC string. Along with this advantage comes a disadvantage that these models do not have the nuanced expressiveness that sometimes is needed to describe a collection of binding sites accurately. Possibly, a more realistic motif model is the *weight matrix*, sometimes called *position-specific scoring matrix*. An example of a weight matrix for the binding sites of Figure 1 is shown in Figure 2. This matrix has one column for each of the seven nucleotide positions of the binding site. Column p assigns scores to each of the nucleotides

```
ACGACGA
GCGTTGT
ACGATGT
ACGACGT
ACGTTGT
GCGACGT
GCGTCGA
ACGTTGA
```

Figure 1 Eight hypothetical binding sites

A	**1.2**	–3.2	–3.2	**0.9**	–3.2	–3.2	**0.5**
C	–3.2	**1.9**	–3.2	–3.2	**0.9**	–3.2	–3.2
G	**0.5**	–3.2	**1.9**	–3.2	–3.2	**1.9**	–3.2
T	–3.2	–3.2	–3.2	**0.9**	**0.9**	–3.2	**1.2**

Figure 2 Log likelihood ratio weight matrix for example of Figure 1

that might appear in position p of a binding site. The higher the total score over all seven positions, the closer a potential binding site is to the collection of binding sites in Figure 1. For instance, the aforementioned consensus string ACGTCGT has a score of $1.2 + 1.9 + 1.9 + 0.9 + 0.9 + 1.9 + 1.2 = 9.9$, the highest possible score for this matrix.

Once a motif model has been selected, the next question is the type of motif discovery application. The input nucleotide sequences might be genes from a single genome that are each suspected to contain binding sites of a single protein, perhaps collected from knockout or gene expression experiments. Alternatively, the sequences might be homologous genes, typically one each from multiple related genomes. These two applications are distinguished by the names *statistical overrepresentation* and *phylogenetic footprinting*, respectively.

In the case of statistical overrepresentation, three different types of motif discovery algorithms have proved successful. The first type is *enumerative*, using either the consensus or IUPAC motif model enumeratively as outlined above. Examples of this approach are the programs RSAT (van Helden *et al*., 2000) and YMF (Sinha and Tompa, 2002). A second type of algorithm might be described as *iterative improvement*. These typically use the weight matrix model and alternate between (1) improving the current set of binding sites, given the current weight matrix, and (2) improving the current weight matrix, given the current set of binding sites. Successful programs that use this approach are the Gibbs sampler (Lawrence *et al*., 1993) and MEME (Bailey and Elkan, 1995). The third approach also uses the weight matrix model and what computer scientists call the *greedy* method, progressively adding to the current set of binding sites that short sequence that is closest to them. This approach is exemplified by the program Consensus (Hertz and Stormo, 1999). Which of these approaches is most accurate depends to a large extent on which motif model most accurately reflects the biological sites sought.

In the case of phylogenetic footprinting, there have been two approaches used. In the traditional approach, the entire homologous sequences are first globally aligned

using a multiple alignment program such as Clustal W (Thompson *et al*., 1994) or Dialign (Morgenstern, 1999). From this global multiple alignment, one then selects short sequences of contiguous columns that show unusually high conservation, by some measure closely related to one of the motif models discussed above. The disadvantage of this approach is that, if the sequences are too diverged, the short conserved binding sites are overwhelmed by the "noise" of the surrounding random sequence, and the alignment tool, in an attempt to align all the noise well, will fail to align the desired binding sites. To avoid this problem, the program FootPrinter (Blanchette and Tompa, 2002) was designed to search directly for well conserved short motifs without resorting to global multiple alignment at all. FootPrinter uses the phylogeny relating the sequences to guide this search.

The References section provides websites, where all of these motif discovery programs are available to users.

Related articles

Article 33, **Transcriptional promoters**, Volume 3; Article 18, **Information theory as a model of genomic sequences**, Volume 7; Article 19, **Promoter prediction**, Volume 7; Article 22, **Eukaryotic regulatory sequences**, Volume 7

References

Bailey TL and Elkan C (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51–80. **http://meme.sdsc.edu/meme/website/intro.html**.

Blanchette M and Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, **12**, 739–748. **http://bio.cs.washington.edu/software.html**.

Hertz GZ and Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577. **http://ural.wustl.edu/~jhc1/consensus/**.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214. **http://bayesweb.wadsworth.org/gibbs/gibbs.html**.

Morgenstern B (1999) DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218. **http://www.genomatix.de/cgi-bin/dialign/dialign.pl**.

Sinha S and Tompa M (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, **30**, 5549–5560. **http://bio.cs.washington.edu/software.html**.

Thompson J, Higgins D and Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680. **http://www.ebi.ac.uk/clustalw**.

van Helden J, Rios A and Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, **28**, 1808–1818. **http://rsat.scmbb.ulb.ac.be/rsat/**.

Introductory Review

In silico approaches to functional analysis of proteins

L. Aravind

National Institutes of Health, Bethesda, MD, USA

1. Introduction

Proteins are unrivaled as the primary functional agents of all biological systems. Hence, understanding how protein sequence and structure relate to its function is tantamount to decoding the most basic aspects of any given biological system. The emergence of the genomic era has resulted in an explosive growth of available protein sequences and other forms of protein-related information. This has not only placed considerable pressure on the conventional techniques of biochemistry and molecular biology to scale up their range of operations but has also created a strong demand for theoretical prediction of biological and biochemical functions from the available data. In particular, the immense preponderance of protein sequence data over all other forms of information pertaining to proteins places a considerable premium on the need to use sequence information as the central platform for exploring the core aspects of biology. Accordingly, reconstructing protein function requires an understanding of the dependencies between the sequence of a protein, and its structure, cellular localization, interaction with functional partners, and its contributions to the fitness of the organism. Fortunately, the ever-increasing flux of data, combined with the basic principles of protein biochemistry and evolutionary biology are only improving our understanding of these dependencies and our ability to ask new questions, which were previously impossible. In this article, we shall briefly consider the founding principles required to tackle the issue of proceeding from protein sequence/structure to biological function.

2. Information in protein sequence: domain identification and function

Pioneering explorations of the protein universe by the early structural biologists resulted in the identification of an organizational level beyond the α-helix, the β-strands, and other basic elements of protein secondary structure. These higher-level structural features were characterized by the presence of independent folding units that were spatially well delineated from each other and often mapped to a collinear

stretch of sequence in the polypeptide. These units, termed *domains*, proved to be the most relevant organizational aspect of protein in terms of understanding protein function. A striking example that illustrated the domain concept in the early days was provided by the structure of the immunoglobulins (Doolittle *et al*., 1966). These proteins had a number of structurally similar domains per polypeptide, each of which was largely composed of β-strands. Comparisons of these structurally similar immunoglobulin domains revealed a similarity even at the sequence level. This suggested their origin from an ancestral sequence unit that specified a single domain, through the process of duplication followed by sequence divergence. Identification of domains also provided the first clues regarding events beyond simple mutations in the evolution of proteins, such as recombination between unrelated genes. Thus, the genesis of large, complex proteins was made more tangible by explaining it in terms of duplications and shuffling of the smaller units, namely, the domains.

Furthermore, there also appeared to be a strong correspondence between these domains and specific biochemical activities or functions of proteins that contained them. Early examples of this principle were provided by the flavin or nicotinamide dinucleotide-binding Rossmann fold domain (Rao and Rossmann, 1973) and the DNA-binding Helix-turn-Helix domain (HTH) (Matthews *et al*., 1982; Sauer *et al*., 1982). These domains were respectively identified as the structural common denominator in otherwise unrelated proteins that bound either FAD/NAD or DNA. Experimental evidence from these proteins showed that their respective substrate binding abilities were essentially dependent on the Rossmann fold or the HTH domains (Figure 1). This fairly consistent, unitary relationship between the biochemical activities of proteins and their constituent domains suggested that the delineation of domains in a polypeptide could lead to the prediction of its biochemical and biological properties. Additionally, the presence of sequence conservation corresponding to the structural similarity of the domains implied that protein sequence information could be used as the basis for these predictions (Figure 1).

From the earliest days of protein sequence comparisons, it became clear that several proteins with shared biochemical properties, such as DNA-binding, or catalytic activities, such as ATP hydrolysis or proteolysis, often shared collinear sequence patterns or motifs that correlated with a given activity (Bork and Koonin, 1996). Experimental analysis of these protein sequence motifs (PSMs) through site-directed mutagenesis, suggested that they often corresponded to the structural core of the fold, the active site, or the substrate interaction surfaces of a given protein (Bork and Koonin, 1996). Furthermore, superposition of these conserved motifs on to the three-dimensional structures of proteins, when available, showed that the motifs could also be used in defining domains in proteins through sequence analysis. While these labors strengthened the association of sequence motifs with protein domains and their biochemistry, they also emphasized the need for objective methodologies to weed out false relationships (patterns) (Iyer *et al*., 2001).

The models, such as the extreme value distribution, originally developed for pairwise ungapped local alignments obtained in similarity searches between a query protein and the target database provided the first objective means for evaluating the statistical significance of sequence alignments (Karlin and Altschul,

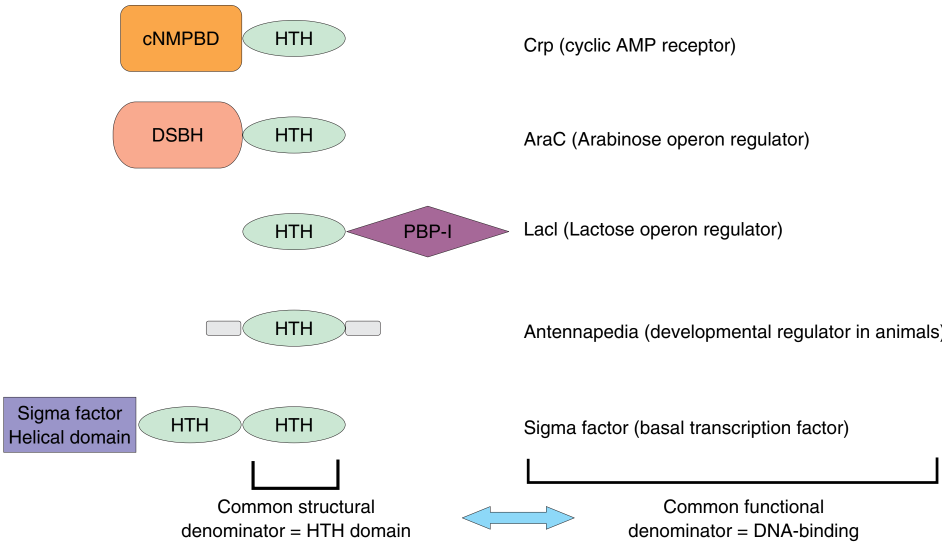


Figure 1 Correspondence between shared conserved domains and common biochemical function. Various transcription factors that share the common biochemical property of DNA binding also share a Helix-turn-Helix domain (HTH). This indicates that the correspondence between the HTH domain and DNA-binding function. The other conserved domains fused to the HTH domain are cNMPBD: the cyclic nucleotide-binding domain; DSBH: the double-stranded beta-helix domain, a sugar binding domain; PBPI: the periplasmic-binding protein I domain (also sugar binding). The grey boxes in the Antennapedia protein are the nonglobular regions

1990). Subsequently, more sensitive sequence profile methods for the detection of remote sequence similarity were developed on the basis of either position-specific score matrices (PSSM) (Altschul *et al*., 1997) or hidden Markov models (HMM) (Durbin *et al*., 1998). Simulations showed that the original extreme value model, developed for the ungapped alignments, was also valid for gapped alignments and those obtained in the sequence profile searches (Altschul *et al*., 1997). These developments enabled the robust definition of PSMs, detection of distant structural relationships and protein domains using sequence information available in the protein databases (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7 for details). Thus, with the basic procedure for the objective identification of protein domains through sequence analysis in place, the path for the in silico journey from protein sequence to function was firmly established.

The foundational principles for protein sequence analysis are deceptively straightforward, but their actual application is often bedeviled with numerous caveats, and overlying modifiers coming from other sources of information. We shall briefly consider some of these issues in the reminder of the article.

3. Sequence conservation and its functional significance

The mere detection of evolutionary (genuine) sequence similarity (termed *homology*) between two proteins does not mean that they are necessarily functionally

equivalent. For a proper evaluation of the relevance of the homology between a set of proteins, a detailed understanding of the qualitative nature of the relationship and its significance is required. A set of homologous proteins or domains are best represented as a multiple sequence alignment of the various independent occurrences of the domain or protein. The degree of relationship, defined on the basis of an objective method of sequence similarity, is typically a continuum between the extremes for any comprehensive multiple alignment of an assemblage of homologous proteins. Some copies of the domain are very similar, others more divergent, and so on, till we reach the limit of objective detection of relationships. This most inclusive set of evolutionarily related domains is defined as a *superfamily*. Natural subdivisions within the superfamily such as families and subfamilies may also be definable on the basis of sequence similarity. Typically, all members of a superfamily might share a generic biochemical property, such as catalysis of a common class of reactions or binding to a particular class of substrates. For example, members of the GNAT acetyltransferase superfamily usually catalyze the transfer of the acetyl functional group to amino groups on various substrates. The identification of the GNAT acetyltransferase domain in a protein allows prediction of such a biochemical activity in that protein, but it does not automatically specify the substrate or the actual biological role played by the protein.

In many cases two or more superfamilies may show generic similarity in terms of the topology of secondary structure elements and features of their structural arrangements, despite the absence of any detectable sequence similarity between them. These superfamilies are then described as sharing a common *fold* (Andreeva *et al*., 2004). The identification of a shared fold in different superfamilies is usually not indicative of any specific shared biochemical properties, though they might show some general commonalities in terms of the spatial position of the interaction with substrates or location of active sites. Hence, the detection of a common fold in different superfamilies by itself is usually not a good predictor of similar biochemical aspects of their function.

An understanding of the levels at which natural selection sculpts a protein helps in gleaning more specific information from a set of homologous proteins. In most proteins from a given organism, natural selection acts in the purifying sense to maintain three different levels of information in the sequence:

1. The most stringent level is the structural level – selection weeds out any changes that could destabilize the basic folding pattern of the domain. Thus, the bulk of sequence conservation, particularly patches of hydrophobic residues, metal-chelating residues, and disulfide-bond-forming cysteines, corresponds to the residues that are likely to be critical for the folding and stability of a protein. These patterns are typically preserved throughout a given superfamily of proteins.
2. At the next level of stringency, natural selection safeguards the "active sites" of proteins. In the case of enzymes, these sites are the constellation of residues required for the catalytic activity and substrate interactions, while in the case of other proteins they are other key residues required for their characteristic biochemical activities. These are less critical as compared to the structure-conferring features because two proteins descending from a common ancestor

might adopt different biochemical properties as they diverge. For example, one of the descendent proteins might be selected for its enzymatic activity and retain the catalytic residues, while the protein belonging to a sister clade descending from the same ancestor might be selected for its ligand-binding properties and would not be under any pressure to retain the catalytic residues, though both proteins are selected to retain the same structure. The identification of such residues is most critical for making meaningful prediction regarding the biochemical functions that might be shared by a group of proteins (Aravind *et al*., 2002).

3. At the lowest level in the hierarchy, natural selection preserves the features of a protein that contain information relevant to its biological context. These might include the residues necessary to interact with other proteins in the pathways or complexes in which they function. Such features are not preserved when proteins that diverge from a common ancestor adapt to different functional milieus. However, identification and characterization of such residues help in narrowing down on actual biological role of proteins to a much greater degree.

4. Low entropy and nonglobular structures in proteins

Beyond the well-defined domains that tend to form compact globular domains with regular secondary structure elements, there is an entire spectrum of low entropy structures ranging from superstructure forming repeat units to random coils. These nonglobular structures, while not following the same structure–function relationships as the globular domains, might still provide useful biological information. Amongst the most common nonglobular structures are transmembrane (TM) helices and signal peptides, which are characterized by compositionally biased regions of proteins that are highly enriched in hydrophobic residues (*see* Article 37, **Signal peptides and protein localization prediction**, Volume 7 and Article 38, **Transmembrane topology prediction**, Volume 7). Another class of nonglobular proteins includes the fibrous proteins, like keratin, which form long alpha-helical structures. This helix dimerizes with another such helix and forms a structure termed the *coiled coil*. Shorter coiled coils, also known as *leucine zippers* due to a periodic pattern of leucines, are found in a wide variety of proteins and serve as important determinants of protein dimerization (Lupas, 1996). Likewise, more complex repeats found in proteins, such as the tetratricopeptide repeat (TPR), typically organize into superhelical or propeller-like structures that serve as surfaces for interactions with other protein complexes (*see* Article 33, **Protein repeats**, Volume 7).

Yet another nonglobular segment observed in proteins is the unstructured or random-coil segment that generally does not assume a single stable configuration and occurs in solution in a highly mobile disordered state. These regions of proteins correspond to sequences with a low complexity – they are enriched in certain amino acids but lack others. In the extreme case, they might contain homopolymeric stretches of a single amino acid such as glutamine or proline, while in other cases di- or tripeptide repeats (*see* Article 32, **Sequence complexity of proteins and its significance in annotation**, Volume 7). These nonglobular regions are particularly

common in the eukaryotes, and often contain the sites for modification by a variety of enzymes such as kinases, glycosyltransferases, and NH_2-group acetyltransferases or binding sites for specific peptide-binding domains such as SH3 or WW domains (Zarrinpar *et al*., 2003). They might also contain signals for nuclear localization and other subcellular sorting signals. Given these associations, the identification of nonglobular segments often serves as a means to predict important aspects of a protein's function, such as its subcellular localization and its potential to mediate protein–protein interactions. Thus, combining the information obtained from particular low entropy regions with the biochemical insights provided by the conserved globular domains often improves the precision of functional predictions.

5. The "society" of domains and the domain-Lego principle

The structural independence of domains allows different domains to associate or dissociate from each other within a single polypeptide in the course of evolution. The majority of domains show a certain degree of "social behavior", that is, co-occurrence with other domains. As a result, a significant fraction of the globular proteins encoded by an organism are collections of domains fused together. In most cases, these domains occur successively, next to each other, or are separated from each other by short or long nonglobular segments. In a smaller number of cases, a domain might occur inserted in a loop of another domain. The order and the combination of domains found in a polypeptide is termed the *domain architecture* of the protein, and is generally schematically depicted by filled shapes standing for domains distributed along the length of the protein (Figures 1 and 2). Thus, the evolutionary process in which domains conglomerate in a protein can be likened to the construction of complex objects using simple Lego bricks (groove-bearing interlocking Danish bricks that is a recreational toy in parts of the Western hemisphere). Just as artists using a relatively limited repertoire of Lego bricks have been able to generate a remarkable diversity of forms, natural selection and recombination have fashioned an astonishing diversity of domain architectures.

Despite the variety of architectures, most of them can be grouped into a limited set of categories with specific teleological principles.

The chief architectural categories include:

1. *Solo domains*: The occurrences of a domain all by itself in a polypeptide may or may not have a definite adaptive explanation (Figure 2). It can be safely assumed that this is the basal state of any domain in evolution, and any adaptively advantageous association with another domain is likely to be fixed. Thus, domains could stochastically flip in and out of neutral combinations with other domains, but any domain that has remained solo across evolutionarily distant organisms is likely to be under certain selective pressure to remain independent.
2. *Architectures involving homodimeric or multimeric assemblies of the same domain*: Individual domains that frequently function as dimers or multimers are selected to exist in domain architectures with two or more copies of the

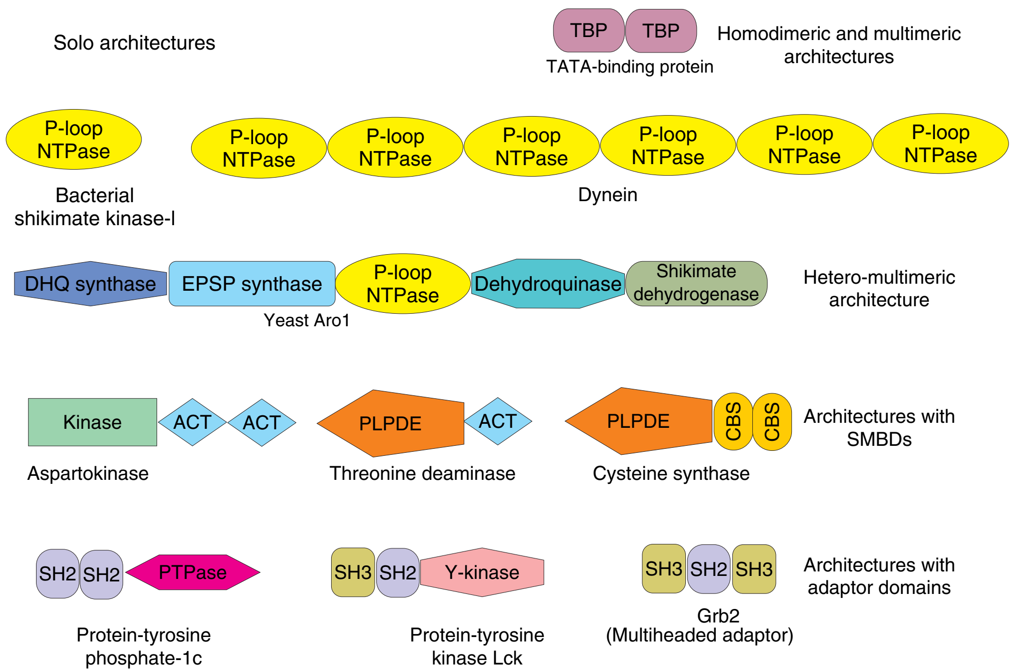


Figure 2 Categories of domain architecture. The domains depicted in the figure are P-loop NTPase: NTP utilizing catalytic domain of the P-loop fold; TBP: the DNA-binding domain of the TATA-binding proteins; DHQ synthase: 3-dehydroquinate synthase; EPSP synthase: 5-enolpyruvylshikimate-3-phosphate; ACT: a small molecule–binding domain; CBS: the so-called cystathionine beta synthase domains (also small molecule–binding domains); SH2: an adaptor domain that binds phosphotyrosine containing peptides; SH3: a polyproline peptide-binding adaptor domain; PTPase: phosphotyrosine phosphatase domain; Y-kinase: tyrosine kinase domain

same domain in a polypeptide. The TATA-binding protein (TBP) that ancestrally bound asymmetrically exists in all its extant forms with a homodimeric architecture of two tandemly repeated TBP DNA-binding domains (Figure 2).

3. *Hetero-multidomain proteins*: Proteins that are composed of domains that participate in either the same or successive steps in a pathway are frequently encountered in a variety of biological systems. The selective advantages of such architectures are clearly linked with the necessity for close functional interactions of the proteins in a pathway. Such fusions are particularly common in metabolic enzymes from eukaryotes, like the multifunctional aromatic amino acid biosynthesis protein (Aro1) that combines at least five distinct enzymatic domains into a single protein, including the shikimate kinase of the kinase family of P-loop ATPase and the dehydrogenase of the Rossmann fold (Figure 2).
4. "*Regulatory architectures" involving small molecule binding domains (SMBDs)*: A variety of low molecular weight compounds regulate responses to environmental stimuli, reaction rates in basic cellular metabolism, nutrient transport, and signal transduction cascades in biological systems. Some of the best-studied forms of small molecule–dependent regulation include allosteric control of enzyme activity, feedback regulation of biochemical pathways, and catabolite

repression. All these processes have selected for a variety of "regulatory architectures" that typically combine an SMBD with an effector domain (Figure 2). These latter domains are the actual agents of a certain biological activity, which is altered by the binding of a small molecule to the associated SMBD. Thus, we have SMBDs such as the ACT domain, which binds amino acids and other related small molecules, fused to a variety of catalytic domains in proteins such as homoserine dehydrogenase and aspartokinase. In these cases, the ACT domain binds substrates or downstream products and mediates the allosteric or feedback effect on the catalytic domains (Anantharaman *et al*., 2001).

5. "*Adaptor architectures" or fusions of adaptor domains with effector domains*: Very often, the performance of a certain biological function involves the targeting of a specific activity that resides in the effector domains to a certain cellular context. This context may be a subcellular location, or a precise protein complex or other biopolymers such as DNA or carbohydrates. The domains that recognize these contexts by binding specific target molecules pertaining to these contexts are generally termed adaptor domains. The most common of these architectures involve fusions of specific peptide-binding or protein-interacting domains with catalytic domains. These are abundantly represented by the fusions between adaptor domains such as the SH2, SH3, WW, and BRCT domains and catalytic domains such as protein kinase, GTPase activating protein, phospholipase, acetyltransferase, and ubiquitination-related domains (*see* Article 31, **Protein domains in eukaryotic signal transduction systems**, Volume 7). Yet other proteins are made up entirely of such adaptor domains, and act as "multiheaded" adaptors that bring together distinct protein complexes (Figure 2). Homophilic interactions are another common mode of action by which two proteins bearing the same adaptor domains, such as the DEATH superfamily of domains, associate with each other in a signaling pathway.

The above classes of domain architectures that are repeatedly reinvented in entirely unrelated proteins might be considered convergently favored solutions for problems in biological engineering in a domain-based world. Given that the above-discussed architectural classes are associated with specific teleological explanations, they represent a subtle source of contextual information that allows refinement of functional prediction of proteins when combined with more straightforward protocols of domain identification and nonglobular segment analysis (Figure 1).

6. Apprehending the "meta-picture": extraneous sources of contextual information in establishing protein function

Most biological systems can be conceptualized as graphs (networks) whose nodes represent individual proteins and whose edges represent the functional connectives that radiate out from a protein at a given node to its partners occupying adjacent nodes (*see* Article 30, **Contextual inference of protein function**, Volume 7). These connectives encompass a considerably diverse set of functional interactions:

1. *Physical interactions*: This includes a variety of direct interactions such as the tendency of two or more proteins to form a functional complex or enzyme–substrate interactions between two proteins.
2. *Indirect regulatory interactions*: This includes transcriptional control of the gene encoding a given protein by another protein, which is a transcription factor, or indirection regulation of a target protein through the synthesis or delivery of a small molecule messenger.
3. *Coexpression and colocalization*: Coexpression is specified by congruent expression patterns of two or more genes whose products are likely to participate in a similar biological process. Colocalization implies the similar subcellular targeting of two proteins, which in turn points to their potential functional interaction in a common organelle or cellular compartment. These functional connections are not mutually exclusive of the earlier-discussed connections.
4. *Genetic interactions*: These connections might arise because of any of the above types of interactions or the general participation of two gene products in a common pathway. Genetic interactions can be ultimately decomposed into a specific type of underlying biochemical interaction, but even in the absence of specific biochemical explanations, genetic interactions can provide useful contextual information for functional inference.

These functional interactions that anchor a protein to its place in a biological network can often help in providing contextual information that goes over and beyond the information provided by intrinsic features like the domain architectures. This extraneous contextual information is usually critical in translating the biochemical inferences gleaned from sequence analysis to the actual biological roles of proteins.

Prior to the postgenomic era, the establishment of these functional connections was the exclusive realm of focused biochemical and genetic experimentation. However, these days, a large amount of contextual information is directly available and can be used for different kinds of in silico analysis to determine biological functions of uncharacterized proteins through the principle of "guilt by association" (*see* Article 30, **Contextual inference of protein function**, Volume 7). The basic idea here is to establish a link between an uncharacterized protein and one or more functionally characterized proteins by means of one or more of the above functional connections. This can then be used to implicate a protein in a particular functional pathway or biological process (Figure 3).

The simplest forms of contextual information arise directly from genome sequence data and are usually in the form of phyletic profiles of orthologous proteins, conserved gene neighborhoods, and lineage-specific expansions of protein families. Yet other forms of contextual information have been obtained from a whole range of high-throughput experimental studies on model organisms such as *Escherichia coli, Saccharomyces cerevisiae*, *Caenorhabditis elegans,* and *Drosophila melanogaster.* These include construction of protein–protein interaction maps, development of single- and multigene knockout strains and microarray-based gene expression profiles under diverse physiological conditions and cell types. Additionally, the data from green fluorescent protein fusion studies have also

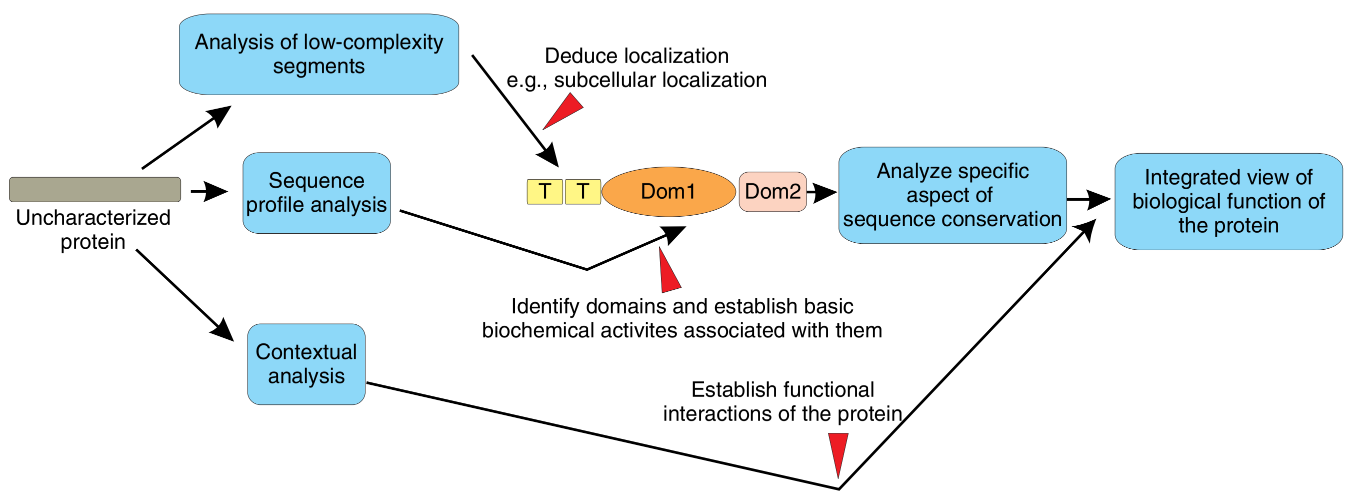


Figure 3 A general scheme for in silico functional inference for uncharacterized proteins. The grey box denotes the sequence of an uncharacterized protein prior to domain identification. The yellow boxes with T represent TM segments, while “Dom1” and “Dom2” represent two conserved globular domains that were detected in the protein

provided considerable information regarding subcellular localization of proteins in model organisms such as *S. cerevisiae* and *Schizosaccharomyces pombe*.

7. Overview of the section on protein sequence analysis and annotation

The confluence of various streams of data holds considerable promise in developing a unified basis for understanding biology in terms of the constituent modules of proteins and their functions (Figure 3). Accordingly, we have attempted to assemble a collection of articles in this section that provide primers regarding the various computational approaches that can be used to glean functional information from proteins. Some of the articles specifically explore the means of investigating nonglobular regions of proteins such as TM regions, signal peptides, and low complexity regions in making functional inferences and predicting protein localization (*see* Article 32, **Sequence complexity of proteins and its significance in annotation**, Volume 7, Article 33, **Protein repeats**, Volume 7, Article 37, **Signal peptides and protein localization prediction**, Volume 7, and Article 38, **Transmembrane topology prediction**, Volume 7). Other articles discuss the significance of conserved domains that are prevalent in specific cellular processes and systems such as signal transduction and chromatin organization (*see* Article 31, **Protein domains in eukaryotic signal transduction systems**, Volume 7 and Article 35, **Measuring evolutionary constraints as protein properties reflecting underlying mechanisms**, Volume 7). There is also a discussion of the application of modern sequence profile methods for the detection of distant relationships and domain discovery, and the role of curated protein classification databases in large-scale annotation of protein functions (*see* Article 36, **Large-scale, classification-driven, rule-based functional annotation of proteins**, Volume 7 and Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7). Finally, there is a perspective on the integration of various forms of contextual information to grasp function at the organismic level (*see* Article 30, **Contextual inference of protein function**, Volume 7).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Anantharaman V, Koonin EV and Aravind L (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *Journal of Molecular Biology*, **307**, 1271–1292.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, 226–229.

Aravind L, Mazumder R, Vasudevan S and Koonin EV (2002) Trends in protein evolution inferred from sequence and structure analysis. *Current Opinion in Structural Biology*, **12**, 392–399.

Bork P and Koonin EV (1996) Protein sequence motifs. *Current Opinion in Structural Biology*, **6**, 366–376.

Doolittle RF, Singer SJ and Metzger H (1966) Evolution of immunoglobulin polypeptide chains: carboxy-terminal of an IgM heavy chain. *Science*, **154**, 1561–1562.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press: Cambridge, pp. 356.

Iyer LM, Aravind L, Bork P, Hofmann P, Mushegian AR, Zhulin IB and Koonin EV (2001) Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biology*, **2**.

Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2264–2268.

Lupas A (1996) Coiled coils: new structures and new functions. *Trends in Biochemical Sciences*, **21**, 375–382.

Matthews BW, Ohlendorf DH, Anderson WF and Takeda Y (1982) Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 1428–1432.

Rao ST and Rossmann MG (1973) Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, **76**, 241–256.

Sauer RT, Yocum RR, Doolittle RF, Lewis M and Pabo CO (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*, **298**, 447–451.

Zarrinpar A, Bhattacharyya RP and Lim WA (2003) The structure and function of proline recognition domains. *Science's STKE*, **2003**.

Specialist Review

Contextual inference of protein function

Aswin Sai Narain Seshasayee

Anna University, Chennai, India

M. Madan Babu

MRC Laboratory of Molecular Biology, Cambridge, UK

1. Introduction

Ever since the first genome sequence of *Haemophilus influenzae* was determined in 1995, there has been an explosion in the number of organisms whose genomes have been completely sequenced and made available in public databases. As on December 2004, there were 235 organisms whose complete genome sequences are available, including that of human (Bernal *et al*., 2001). Such an exponential increase in the number of available sequences has led to the realization that to understand the biology of an organism, one has to be able to make sense of this sequence data. One way of reaching this end is by identifying the function of proteins and RNAs encoded in its genome.

Traditionally, computational methods used to assign protein function are based on the simple assumption that proteins with similar sequences (homologs) perform similar molecular function. In such homology-based methods, function of a protein is inferred by comparing its sequence against a database such as the nonredundant database or Swiss-Prot using powerful tools such as PSI-BLAST (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7) to pick up homologs and then making good use of any available information on the homologous protein (Aravind and Koonin, 1999b).

Even though homology-based methods, such as those described above, have been quite effective at predicting molecular functions of remote homologs, large gaps still exist in our knowledge. Several genes in any given organism remain uncharacterized, where even sophisticated homology-based methods fail to suggest any function (Madera *et al*., 2004). Moreover, higher-order function of a protein, such as the pathway in which it plays a role, cannot, in principle, be assigned using such methods (Huynen *et al*., 2000). It is here that homology-independent methods, which primarily utilize the context in which a gene exists, become useful in inferring function. Such methods, more generally termed contextual methods, make use of information about gene organization in multiple genomes, and data

obtained from large-scale functional genomics experiments such as gene expression, transcriptional interaction, and protein interaction studies.

In this article, we have classified homology-independent contextual methods to infer protein function into two major groups: methods that use genome sequence data and those that use information from large-scale functional genomics data.

2. Inferring function from genomic data

With a large number of genome sequences determined and many more in the pipeline, there is no paucity of data for gaining useful insights into protein function. In the following section, we discuss the different homology-independent contextual methods that use genome sequence data for inferring protein function.

2.1. Gene fusion

In a set of well-characterized proteins, it has been observed that pairs of interacting proteins or those that are functionally linked (e.g., involved in the same metabolic pathway) in one organism are fused into a single polypeptide chain in another organism. If two proteins are required to interact physically or are functionally linked, then it is likely that there is a selective pressure to produce them at the same time and at the same place within the cell, so that the chances that they interact are maximized. The best way to do it would be to have them as a part of the same gene product so that they are transcribed and translated at the same time and at the same place, thus allowing them to carry out their function efficiently. This logic has been exploited to predict pairs of interacting proteins by identifying instances of protein pairs that are fused in one organism but are found to be separate in another organism (see Figure 1a). A well-known example is the case where the genes encoding GyrA and GyrB in *Escherichia coli* are fused into a single-gene product, DNA topoisomerase, in yeast. Also illustrative is the case of gene fusions between evolutionarily mobile helicase and nuclease domains in DNA repair proteins, implying a general tendency of these domains to interact physically. An example is the fusion between PolIII subunit-like nuclease and DinG helicase in *Bacillus subtilis* (Aravind, 2000).

In one of the early computational analysis, Marcotte *et al*. (1999a) used genome sequence data and found that there are 6809 candidate interacting protein pairs in *E. coli.* The authors also used a three-step validation for their prediction: (1) ask whether the predicted pair of interacting proteins share a keyword between their annotations, which they show is very rare in randomly chosen pairs, (2) validate against experimental data, and (3) compare with results obtained from phyletic profiles (see Section 2.3 and Article 41, **Phylogenetic profiling**, Volume 7). Enright *et al*. (1999) showed that there are 215 proteins in the genomes of *Escherichia coli, Methanococcus jannaschii*, and *Haemophilus influenzae,* which are involved in 64 gene fusion events. In another similar study, it was shown that metabolic enzymes exhibited a profound 300% preference to exist as fused proteins over a control set of proteins (Tsoka and Ouzounis, 2000). This observation allowed them to conclude

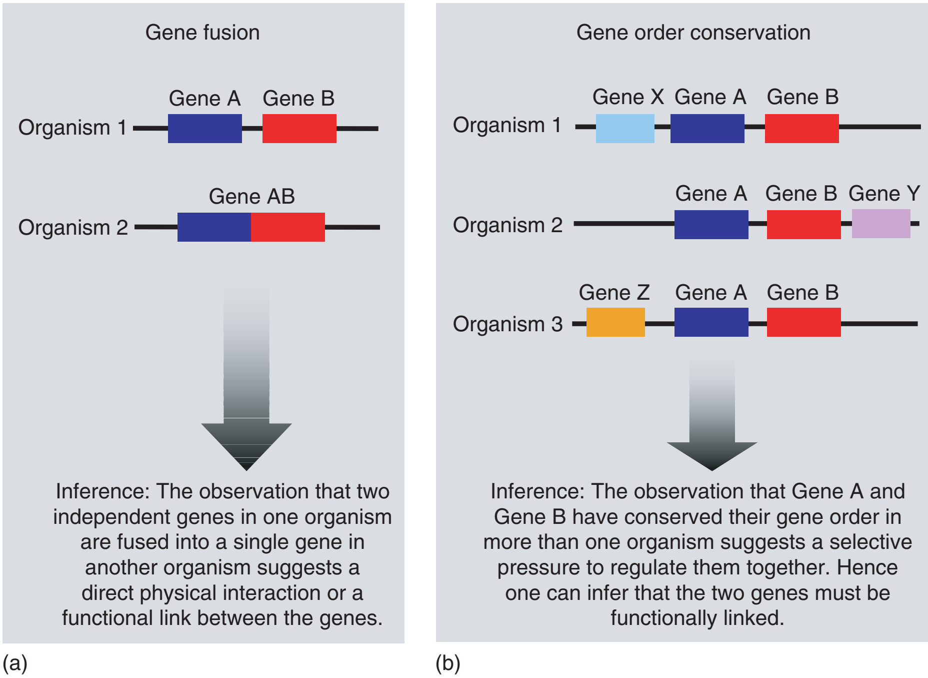


Figure 1 Figure illustrating inference of protein function from (a) gene fusion events and (b) conservation of gene order across multiple genomes

that proteins predicted to interact by detection of gene fusion events are more likely to be involved in metabolic pathways.

This method, while being useful, has a few pitfalls. In particular, this method could be confounded by evolutionarily mobile promiscuous domains that show fusions to other domains in a wide range of functional contexts. False-positives, with regard to prediction of physical interactions, may also crop up because it is perfectly possible that two fused protein domains do not interact physically but only functionally. The method cannot differentiate between interacting and noninteracting homologs. Another important problem associated with this method and the other genome sequence–based methods, which will be described later, is the difficulty in identifying true orthologs. If the stand-alone polypeptides and the individual components of the composite protein under consideration are really paralogs and not orthologs, we will end up with a false prediction (Marcotte *et al*., 1999a; Galperin and Koonin, 2000).

2.2. *Gene order conservation and co-occurrence of genes in operons*

Comparative genomic analyses among closely related species revealed an important fact that it is very rare for organisms to conserve gene order. Given the extremely high rate of recombination in bacterial genomes, the probability that the order of

genes is maintained across many organisms is negligible. However, when gene order conservation is indeed observed, it only means that it is due to a selective pressure that the genes are kept together. This gene order conservation might mean that the two genes concerned encode products that either interact functionally or physically (see Figure 1b). In fact, Demerec and Hartman (1959) observed that the "mere existence of such arrangements shows that they must be beneficial, conferring an evolutionary advantage on individuals and populations which exhibit them" (Overbeek *et al*., 1999a). When interacting proteins are coded for by a polycistronic mRNA, the collection of adjacent genes is called an *operon*. Operons are common in prokaryotes and proteins involved in a particular biochemical pathway are generally clustered together to form an operon.

While this method is conceptually simple, in silico identification of operons in genome sequence is not trivial (*see* Article 20, **Operon finding in bacteria**, Volume 7). It was observed that the number of instances of functionally linked gene pairs increased with the increase in the number of organisms considered for the analysis. Overbeek *et al*. (1999a) have shown that the number of pairs of close bidirectional best hits (PCBBHs) identified in local alignment sequence searches is related to the square of the number of genomes considered. Data from 24 organisms pointed to the presence of 34 644 PCBBHs. More recent information on function inference from operons has shown that there are 58 498 PCBBHs in 31 genomes (Overbeek *et al*., 1999b). With the number of genome sequences being made available increasing exponentially, the predictive power of this method should increase at a much higher rate.

2.3. *Correlated and anticorrelated occurrence of genes across many genomes*

The idea behind this method described by Pellegrini *et al*. (1999) is that functionally related genes evolve in a correlated fashion across genomes. This implies that these genes should coexist across a set of organisms displaying the function of interest. In this technique, each gene is represented by a "phyletic profile" (*see* Article 41, **Phylogenetic profiling**, Volume 7), which is a string of n characters, where n is the number of organisms considered. Each of these characters may be either + or − representing the presence or absence of the gene of interest in the corresponding organism. The genes are then clustered into groups on the basis of their phyletic profiles. Gene products that belong to a cluster are then predicted to be functionally linked (see Figure 2).

In their study, Pellegrini *et al*. (1999) characterized 4217 genes from *E. coli* using this method. The number of organisms used to create the phyletic profile was 16. As an example, proteins with functions associated with the ribosome are discussed (Pellegrini *et al*., 1999). The ribosomal protein RL7 has homologs in 10 of the 11 eubacterial genomes and in yeast, but not in archaea. More than 50% of *E. coli* genes with function associated with the ribosome share the phyletic profile with RL7. Several uncharacterized proteins were found to fall under this cluster, and the authors state that it is likely that these uncharacterized proteins have functions associated with the ribosome. These proteins whose function has been

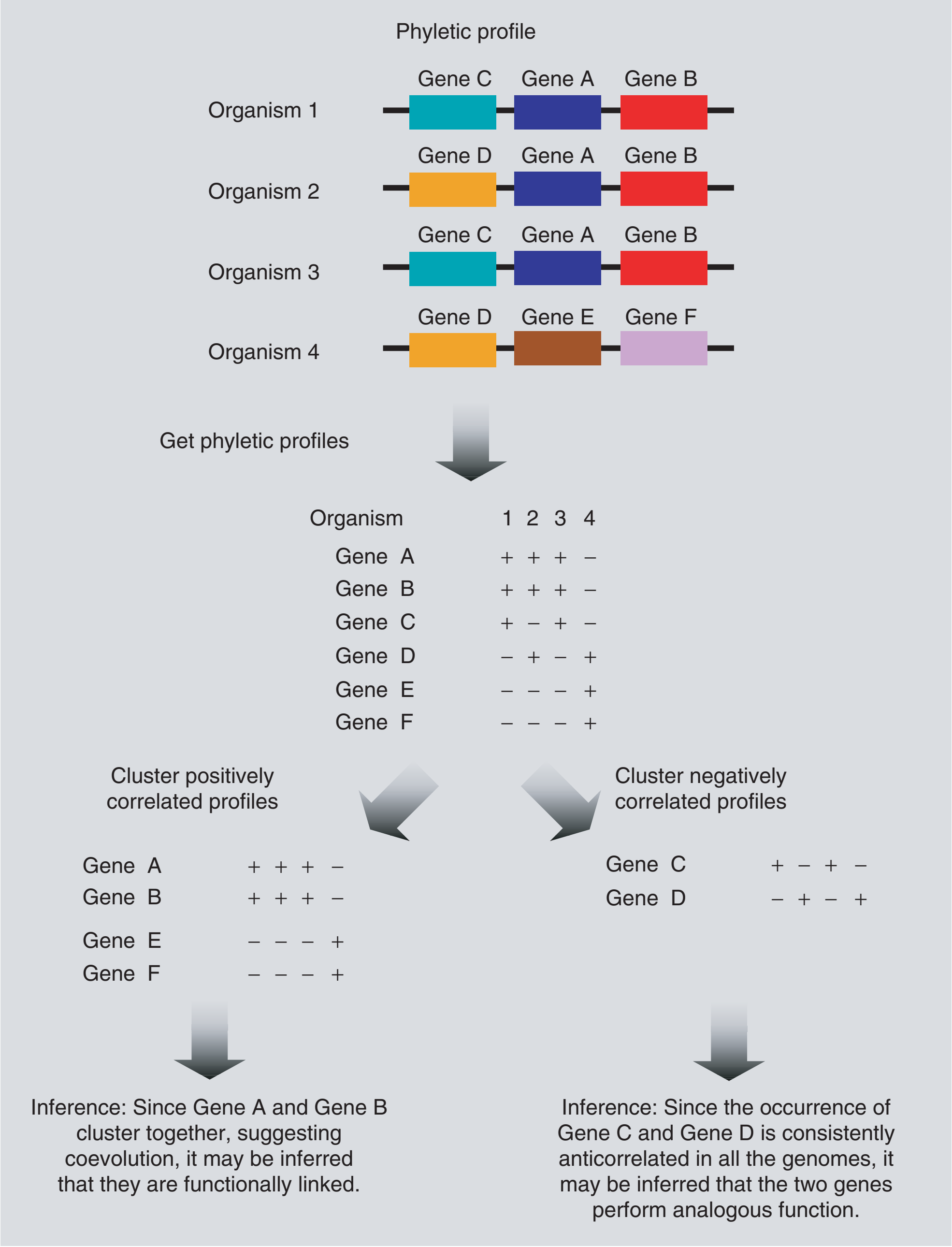


Figure 2 Inference of protein function using correlated or anticorrelated phyletic profiles of clusters of genes. "+" represents presence of the gene and "–" represents absence of the gene

inferred by this method share no sequence similarity to the characterized protein such as RL7, and hence their role in the functioning of the ribosome could not have been inferred by homology-based methods. Another example of application of this method to inferring function is the subcellular localization of proteins. It is based on the idea that proteins that localize to a particular cellular compartment have a similar phyletic profile. This method, when applied to *Saccharomyces cerevisiae* identified 361 nucleus-encoded mitochondrial proteins with 50% accuracy and 58% coverage (Marcotte *et al*., 2000).

Just like how similarity in phyletic profiles of genes can be used to infer the function of a protein, anticorrelation of phyletic profiles can also be used to identify instances of nonorthologous gene displacement, that is, instances in which nonhomologous proteins perform the same function in different organisms. The principle behind this method relies on identifying cases in which two genes have dissimilar phyletic profiles, that is, if the first gene is present in an organism, then the second gene is absent in the same organism. If this observation is consistent for a large number of organisms considered, then one can infer that the two proteins perform analogous function and evolution has selected for one of the two proteins, discarding the other (see Figure 2).

An example of how anticorrelated phylogenetic profiles have enabled us to understand a well-characterized metabolic pathway such as glycolysis involves the enzyme fructose-1,6-bisphosphate aldolase (FBA). This enzyme catalyzes the step where fructose-1,6-bisphosphate is cleaved to glycerol-3-phosphate and dihydroxy-acetone phosphate. While this enzyme is present in most bacteria and eukaryotes, it is absent in Chlamydiae and Archaea. However, a second enzyme DhnA in these organisms forms an almost-perfect complementary phyletic pattern against a set of reference genomes, indicating that this enzyme is the only FBA in Chlamydiae and Archaea (Galperin and Koonin, 2000). In another case, seven enzymes belonging to the 15-step thiamin biosynthesis pathway were found to have been displaced by analogous proteins. These predictions have been verified experimentally. Importantly, this led to the assignment of function for three proteins, till then uncharacterized. An example is the displacement of thiamin phosphate synthase, first described in *E. coli,* by genes orthologous to the hypothetical ORF MTH861 in *M. thermoautotropicum* (Morett *et al*., 2003).

2.4. *Conservation of bidirectionally transcribed gene pairs*

Yet another contextual method for inferring function from genomic data is the identification of conservation of gene orientation of adjacent, bidirectionally transcribed gene pairs across genomes (Korbel *et al*., 2004). These bidirectionally transcribed gene pairs may be regulated by a pair of overlapping promoter elements. Such gene organization is beneficial as it offers a method of transcriptional regulation distinct from operons (Korbel *et al*., 2004; Warren and ten Wolde, 2004). It is also worthwhile to note that while organization of convergently transcribed gene pairs is rapidly lost during evolution that of divergently transcribed gene pairs is not. Systematic analysis carried out by Korbel *et al*. (2004) shows that over 5000 divergently transcribed gene pairs are conserved across distantly related

organisms and that 26% of all *E. coli* genes are divergently transcribed pairs, a quarter of which are conserved in a distant evolutionary clade. There are 6.5 times more divergently transcribed gene pairs than convergently transcribed ones. Such conservation of bidirectionally transcribed gene pairs may be due to the pressure of coexpression, similar to the pressure that maintains the operon structure. About 87% of gene pairs conserved across distant genomes have correlated gene expression, with a Pearson correlation coefficient value greater than 0.6 (see Figure 3).

Bidirectionally transcribed gene pairs may more than just be coexpressed. Over 71% of such pairs identified in *E. coli* were classified as "RX" by the authors, where one gene product is a transcriptional regulator (R) and the other belongs to some other class of protein (X). Interestingly, most of such RX pairs have been shown to be autoregulatory. Thus, if a protein has been classified as a transcription factor by homology-based methods (Aravind and Koonin, 1999a; Perez-Rueda and Collado-Vides, 2000; Madan Babu and Teichmann, 2003), then its transcriptional target in vivo can be determined using this method. For example, using this method, it was shown that prokaryotic members belonging to the orthologous group KOG2969 (*see* Article 90, **COGs: an evolutionary classification of genes and proteins from sequenced genomes**, Volume 6) are involved in regulating the expression of ribosome-associated genes in alpha-proteobacteria. It has been experimentally shown that KOG2926, whose gene is divergently transcribed from the nearby mitochondrial ribosomal protein S2 (RpsU) gene, is indeed the transcriptional regulator

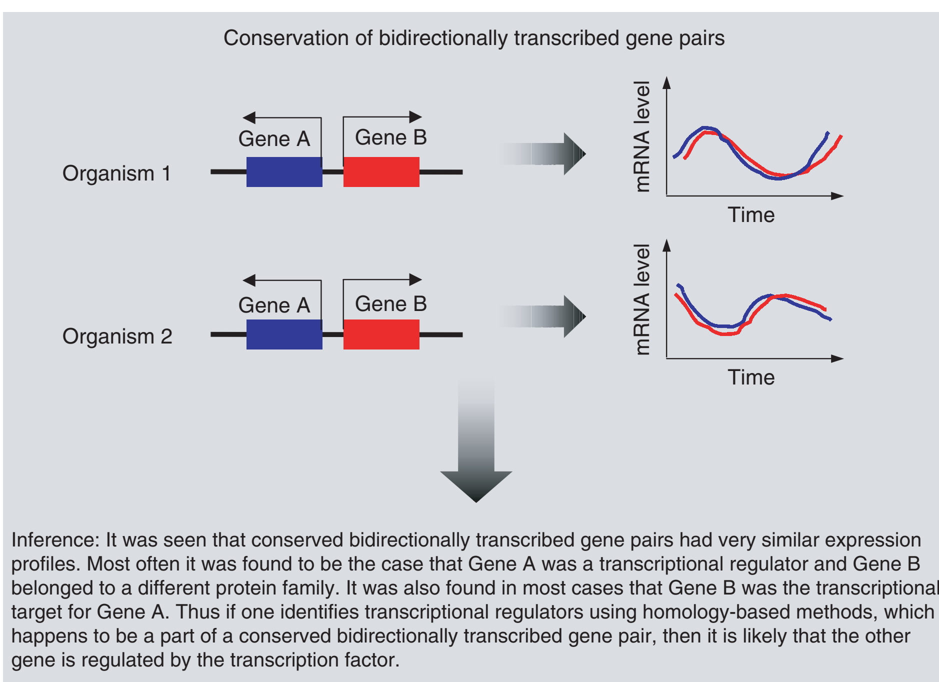


Figure 3 Determining the function of proteins from conservation of bidirectionally transcribed gene pairs

of rpsU. The efficiency of this method, like that of any other genomic context method, is dependent on the number of genomes sampled (Korbel *et al*., 2004).

3. Inferring function from large-scale experimental data

Coupled with the explosion in the availability of complete genome sequence for several organisms is public access to large-scale experimental data on interactions between the cellular components and to data on gene expression for several organisms obtained using microarrays. These interactions may be protein–protein, protein – metabolite, or protein–nucleic acid.

The following section will focus on methods that utilize data on the interactions between the cellular components and gene-expression studies to infer functions of uncharacterized genes. The set of all interactions mediated by cellular components can be conceptualized as a graph or a network, in which each cellular component is represented as node and interaction between the two components as edges or arcs. Such cellular networks can be used to infer protein function in the context of a biological process (Barabasi and Oltvai, 2004; Xia *et al*., 2004). For instance, a link between a functionally uncharacterized protein X and well-characterized protein(s) will enable one to place the function of X in context of the function of the characterized protein(s) to which it is linked (see Figure 4).

In the following section, we review four different methods of inferring protein function using data obtained from the different large-scale experiments.

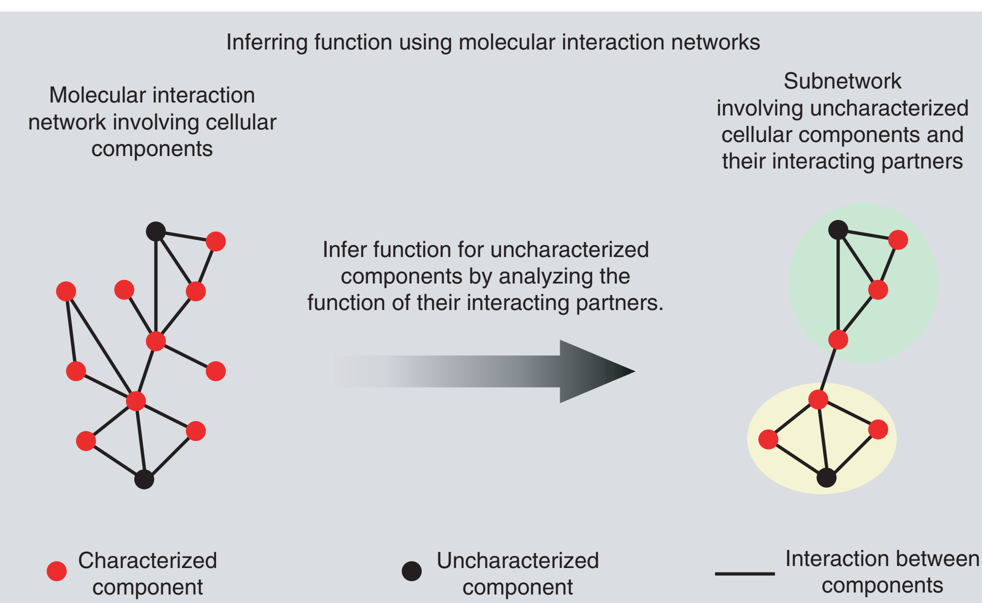


Figure 4 Figure illustrating the general idea of inferring function of an uncharacterised cellular component based on its links with other characterized components in a molecular interaction network

3.1. Protein–protein interaction network

Identification of cellular protein–protein interaction networks has become possible due to the development of high-throughput techniques such as the yeast two-hybrid experiment and the tap-tag method (*see* Article 37, **Inferring gene function and biochemical networks from protein interactions**, Volume 5). Protein–protein interaction networks derived using the two-hybrid method are available for *D. melanogaster* (Giot *et al*., 2003), *C. elegans* (Walhout *et al*., 2000), *S. cerevisiae* (Uetz *et al*., 2000), *H. pylori* (Rain *et al*., 2001), and the T7 bacteriophage (Bartel *et al*., 1996). Such experimentally derived networks can be analyzed using computational approaches and useful functional information can be obtained from them.

When a protein–protein interaction network is rendered in such a way that specific functional categories are highlighted, it will be observed that proteins belonging to the same functional category cluster together. It has been shown that 72% of interactions between experimentally characterized proteins in such networks are between pairs belonging to the same functional class. The significance of this number is illuminated when we see that this percentage is only 12% when the network is randomly divided into different categories. Linkages between proteins belonging to different functional categories are also possible. The difficulty with interpreting such linkages is that they may be false-positives or genuine cross talk between related pathways. It was also determined that 78% of interactions between proteins with known cellular localization involved proteins sharing at least one cellular compartment (Schwikowski *et al*., 2000; Tucker *et al*., 2001).

Protein–protein interaction networks have been used to assign functions to a number of previously uncharacterized proteins. A common approach for assigning function using protein–protein interaction network is what is known as the majority-rule assignment. This technique assigns function on the basis of most common functions present among the characterized interacting partners (see Figure 5). Vazquez *et al*. (2003) have used interactions between uncharacterized proteins in an iterative majority-rule technique to yeast and identified 2238 interactions among 1826 proteins (Vazquez *et al*., 2003).

Map of 957 interactions involving 1004 yeast proteins generated by Uetz *et al*. (2000) has yielded novel insights into functions of several proteins. For instance, two yeast proteins of unknown function, which were seen to interact with each other, also bind to ornithine aminotransferase. This implies that they may be involved in arginine metabolism. Novel interactions between proteins involved in the same biological function were also shown. Their data also shows that novel interactions exist between proteins involved in diverse biological processes. For example, three snRNPs were found linked to the ribosomal protein S28. This might indicate a role for S28 in RNA splicing or, alternatively, a novel role of snRNPs in translation or ribosome biosynthesis (Uetz *et al*., 2000).

This method, like any other, is not free from operational difficulties and errors. The yeast two-hybrid method can only detect binary interactions, but in reality many proteins function as protein complexes. This may partly be overcome by the tap-tag method, which is used to purify protein complexes and characterize them. Even in such cases, interactions between the individual subunits of the complexes are not known. Yeast two-hybrid also gives rise to many false-positive interactions

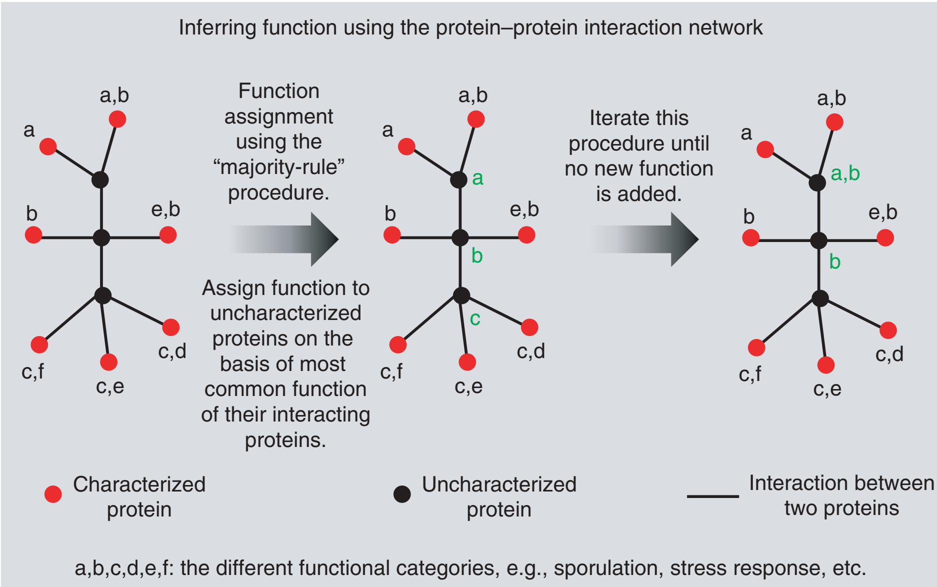


Figure 5 Figure demonstrating the use of majority-rule assignment (see text) of function from protein–protein interaction networks. In this figure, labels a–f in black represent arbitrary functional class assigned to well-characterized proteins (nodes) in the network. Labels a–c in green are the functions of uncharacterized proteins (nodes) inferred using the majority-rule method

because of the nature of the experiment – since the interaction is tested in the nucleus, the local cellular environment may be different and it is possible that the interaction is mediated by a third protein, in which case the two proteins shown to interact by the two-hybrid system do not really interact directly (Aloy and Russell, 2002; von Mering *et al*., 2002).

3.2. *Transcriptional regulatory network*

Transcriptional regulatory networks are deciphered by carrying out ChIP-chip experiments, which specifically identify DNA sequences that can interact with transcription factors. Data generated are called location data, and are again useful in assigning function to proteins (see Figure 6).

A transcriptional regulatory network has been generated by Lee *et al*. (2002) for yeast. This study used 141 transcription factors (TFs) that were listed at that time in the Yeast Proteome Database. Out of these 141 TFs, only 106 gave useful information, as experimental modification caused loss of function in 17 proteins and the others could simply not be detected. This network showed that 2343 of 6270 yeast genes were bound by one or more of the 106 TFs when the cells were grown in rich media. More than one-third of these promoters were bound by more than one TF. In the assembled network, six network motifs were identified. Network motifs are short patterns of interconnections that recur at many places in the network at frequencies much higher than seen in random networks of a similar

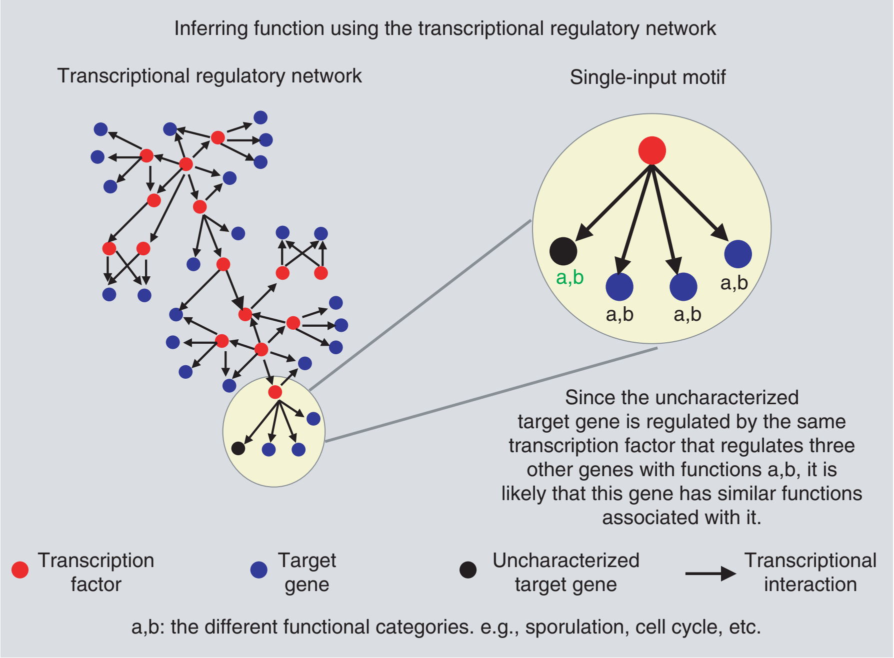


Figure 6 Determination of protein function using the context in which it occurs in the transcriptional regulatory network

size (Lee *et al*., 2002; Shen-Orr *et al*., 2002). In the yeast network, the motifs identified were the following: autoregulation, multicomponent loops, single input, multi-input, feed-forward loops, and regulator chain. Since each network motif has a specific information-processing task (Shen-Orr *et al*., 2002), such regulatory motifs could be used to make functional assignments. For example, Fhl1, a protein with function unknown was found to be involved in a single-input motif-regulating multiple genes involved in ribosome biosynthesis. This protein was also involved in a multi-input motif. Experiments have showed that a mutation in this protein caused serious defects in the ribosome synthesis machinery (Lee *et al*., 2002).

Knowledge about the function of proteins could be extended using such networks. For example, the protein Phd1 that is involved in pseudohyphal growth in nutrient stress conditions was shown to interact with proteins involved in general stress response and in regulation of metabolism (Lee *et al*., 2002).

More recent transcriptional regulatory network generated for yeast by Harbison *et al*. (2004) has revealed 11 000 interactions mediated by 203 TFs. In this study, predictions of TF-promoter pairs were validated using comparisons across four species of yeast. This work discovered and rediscovered promoter elements. Functional correlations, to confirm previous functional assignments, could be made to the network connections. For example, six cell-cycle transcriptional regulators were found to bind to the promoter for YHP1, which is involved in the regulation of the G1 phase of the cell cycle.

3.3. *Gene coexpression network*

Since proteins involved in the same pathway or those that are part of the same protein complex are coregulated at the transcriptional level, gene-expression data can be used to make functional inferences. However, coexpression alone will not necessarily mean functional relationship (Mateos *et al*., 2002). On the other hand, when a pair of gene is consistently coexpressed across large evolutionary distances, this method becomes a powerful tool to infer protein function (Teichmann and Madan Babu, 2002; Stuart *et al*., 2003; van Noort *et al*., 2003; McCarroll *et al*., 2004) (see Figure 7).

Stuart *et al*. (2003) have constructed a gene coexpression network (*see* Article 60, **Extracting networks from expression data**, Volume 7) for a set of metagenes. Metagenes are defined as sets of genes that are evolutionarily conserved across diverse organisms (effectively the term metagene is identical to a group of orthologous genes). In such a network, metagenes are represented as nodes and gene coexpression between a pair of genes as links. In their work, they first characterized 6307 metagenes using the following organisms: human, fly, worm, and yeast. For each pair of metagenes, coexpression was studied and gene pairs whose expression was significantly correlated across multiple organisms were identified. Expression was studied only for metagenes because the evolutionary conservation of these genes implies that the network covers only core biological processes. This

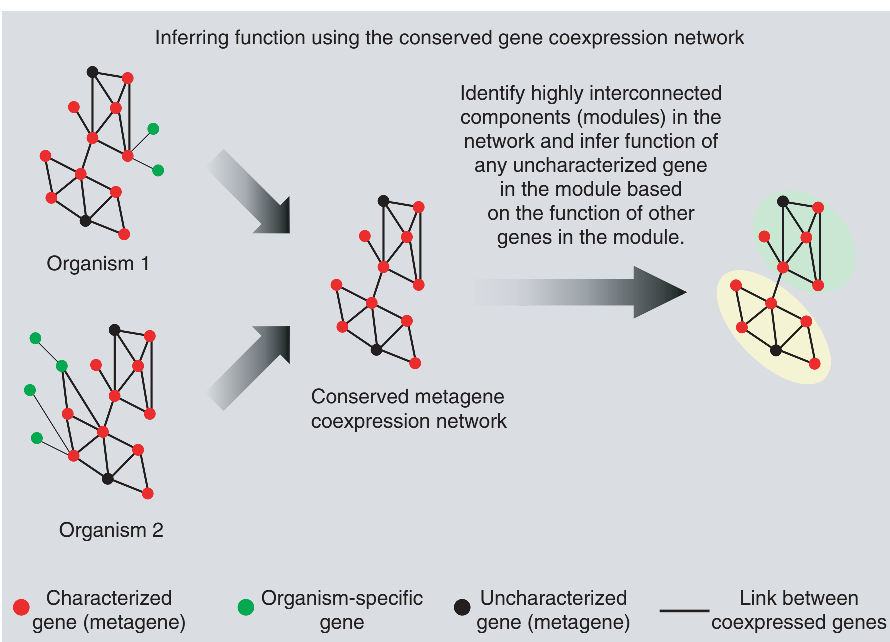


Figure 7 Use of conserved gene coexpression networks to determine protein function. The method initially identifies metagenes, which are groups of orthologous proteins across the different species. For each organism, a gene coexpression network is determined and the part of the network that is conserved in the set of organisms considered is referred to as the conserved metagene coexpression network

coexpression network comprised 3416 metagenes connected by 22 163 interactions. As a next step, the authors used a visualization technique in which metagenes were placed next to each other at a distance in proportion to their level of coexpression. Highly connected and conserved regions in the network could be visualized as peaks in the map. Such visualization led to the identification of 12 "components" – regions containing a large number of interconnected metagenes. These components were the following: signaling, ribosome biogenesis, energy generation, proteasome, cell cycle, general transcription, animal specific, translation, ribosomal subunits, secretion, neuronal, and lipid metabolism (Stuart *et al*., 2003).

The function of three uncharacterized genes was determined using this method. It was found that these proteins were coexpressed with genes involved in cell proliferation and cell cycle. Two other proteins known to function in other cellular processes also coexpressed with the above set of metagenes. Experimental verification of these results showed that all these five proteins were overexpressed in human pancreatic cancers relative to normal tissue and hence are involved in cell proliferation. Thus, by analyzing such coexpression networks, one can identify previously unknown proteins involved in well-characterized biological processes.

3.4. Integration of different data types

The methods discussed in the previous sections make use of a single type of data to infer protein function. The next logical step is to integrate different types of data to achieve this objective.

Marcotte *et al*. (1999b) have integrated data from phyletic profiles, gene fusion, and gene expression with the experimentally determined protein–protein interaction networks in yeast and identified 4130 links of very high confidence. They use this network to assign function to proteins. The yeast ORF, YGR021 W, was assigned a function related to mitochondrial protein synthesis. The function of Sup35 was extended to beyond its role as a translation release factor and in guiding nascent proteins to their cellular locations. Sup35 shows both correlated evolution and expression pattern with a yeast chaperonin system thought to aid in the folding of newly synthesized actin and microtubules.

Bar-Joseph *et al*. (2003) have combined gene-expression data and location analysis data to assign functions to proteins. They have done this because they realized that location analysis, while identifying what interacts with what, does not throw light on the type of interaction, that is, whether the interaction activates or represses transcription. To infer this, it is best to use gene-expression data. The algorithm that they have developed, namely, GRAM (Genetic Regulatory Modules), clusters sets of genes to which a common TF is bound. From this set, a subset of genes with correlated expression is derived. Positive correlation between TF binding and expression level indicates that the regulator protein is an activator. The integration of the two types of data allows a relaxation of the conditions set to identify a binding event. This will bring down the number of false-negatives while not substantially increasing the rate of false-positives. This reduction in stringency allowed six new genes to be identified as targets of Hap4, a well-characterized protein involved in regulation of oxidative phosphorylation and respiration.

In another recent work, Luscombe *et al*. (2004) developed an approach called SANDY (Statistical Analysis of Network Dynamics), which integrates gene-expression data with the transcriptional regulatory network to identify condition-specific transcriptional regulatory networks. By integrating these two sources of information, novel insights about the dynamic nature of transcriptional regulatory networks were obtained. For instance, this approach led them to identify “master regulators” that are important in particular cellular processes, such as sporulation and cell cycle. This approach additionally helped in identifying condition-specific transcription factors and their transcriptional targets for a given cellular condition such as stress and DNA damage.

4. Conclusions

The above, recently developed methods for protein function prediction have created a paradigm shift in our perspective toward understanding proteins. The classical view of protein function in terms of its molecular properties such as catalysis or ligand binding is being complemented by what is known as contextual or cellular function (*see* Article 29, ***In silico*** **approaches to functional analysis of proteins**, Volume 7). It is defined as the understanding of protein function in terms of the pathways or intracellular subnetworks in which the protein might be expected to play a role (Fraser and Marcotte, 2004; Palsson, 2004). The actual role it plays in this context would need to be elucidated by further experimentation. As in the case of homology-based methods, it is again due to the availability of a large volume of organized data that nonhomology-based methods to infer protein function have become feasible. Such large-scale methods for predicting protein function becomes more important in light of the situation where over a thousand genes in the most characterized bacterial genome, *E. coli*, remain known only as hypothetical ORFs.

Related articles

Article 37, **Inferring gene function and biochemical networks from protein interactions**, Volume 5; Article 90, **COGs: an evolutionary classification of genes and proteins from sequenced genomes**, Volume 6; Article 20, **Operon finding in bacteria**, Volume 7; Article 29, ***In silico*** **approaches to functional analysis of proteins**, Volume 7; Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7; Article 41, **Phylogenetic profiling**, Volume 7; Article 60, **Extracting networks from expression data**, Volume 7

Acknowledgments

MMB acknowledges the MRC Laboratory of Molecular Biology, Cambridge Commonwealth Trust and Trinity College, Cambridge for financial support. We would like to thank Karthik Sathiyamoorthy for critically reading the manuscript.

References

Aloy P and Russell RB (2002) The third dimension for protein interactions and complexes. *Trends in Biochemical Sciences*, **27**, 633–638.

Aravind L (2000) Guilt by association: contextual information in genome analysis. *Genome Research*, **10**, 1074–1077.

Aravind L and Koonin EV (1999a) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Research*, **27**, 4658–4670.

Aravind L and Koonin EV (1999b) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *Journal of Molecular Biology*, **287**, 1023–1040.

Barabasi AL and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**, 101–113.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **21**, 1337–1342.

Bartel PL, Roecklein JA, SenGupta D and Fields S (1996) A protein linkage map of Escherichia coli bacteriophage T7. *Nature Genetics*, **12**, 72–77.

Bernal A, Ear U and Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, **29**, 126–127.

Demerec M and Hartman PE (1959) Complex loci in microorganisms. *Annual Review of Microbiology*, **13**, 377–406.

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Fraser AG and Marcotte EM (2004) A probabilistic view of gene function. *Nature Genetics*, **36**, 559–564.

Galperin MY and Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology*, **18**, 609–613.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Huynen M, Snel B, Lathe W and Bork P (2000) Exploitation of gene context. *Current Opinion in Structural Biology*, **10**, 366–370.

Korbel JO, Jensen LJ, von Mering C and Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology*, **22**, 911–917.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Luscombe NM, Madan Babu M, Yu H, Snyder M, Teichmann SA and Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Madan Babu M and Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, **31**, 1234–1244.

Madera M, Vogel C, Kummerfeld SK, Chothia C and Goug J (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**, Database issue, D235–D239.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Marcotte EM, Xenarios I, van Der Bliek AM and Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12115–12120.

Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M and Stolovitzky G (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Research*, **12**, 1703–1715.

McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI and Li H (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics*, **36**, 197–204.

Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B and Bork P (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, **21**, 790–795.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999a) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biology*, **1**, 93–108.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999b) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–2901.

Palsson B (2004) Two-dimensional annotation of genomes. *Nature Biotechnology*, **22**, 1218–1219.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Perez-Rueda E and Collado-Vides J (2000) The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research*, **28**, 1838–1847.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al*. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.

Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1257–1261.

Shen-Orr SS, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**, 64–68.

Stuart JM, Segal E, Koller D and Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Teichmann SA and Madan Babu M (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends in Biotechnology*, **20**, 407–410, discussion 410.

Tsoka S and Ouzounis CA (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genetics*, **26**, 141–142.

Tucker CL, Gera JF and Uetz P (2001) Towards an understanding of complex protein networks. *Trends in Cell Biology*, **11**, 102–106.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

van Noort V, Snel B and Huynen MA (2003) Predicting gene function by conserved co-expression. *Trends in Genetics*, **19**, 238–242.

Vazquez A, Flammini A, Maritan A and Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, **21**, 697–700.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N and Vidal M (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.

Warren PB and ten Wolde PR (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. *Journal of Molecular Biology*, **342**, 1379–1390.

Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H and Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annual Review of Biochemistry*, **73**, 1051–1087.

Protein domains in eukaryotic signal transduction systems

Kay Hofmann
Memorec Biotec GmbH, Köln, Germany

1. Functional domains

Originally, the concept of protein domains was derived from the analysis of three-dimensional structures (*see* Article 68, **Protein domains**, Volume 7). While small proteins typically have a "monolithic" structure consisting of a single fold, larger proteins can follow two different architectural principles. Some of them just form larger monolithic structures, while the majority of large proteins consist of several smaller folding units, the so-called domains. Structural domains can fold independently from the rest of the proteins, and each one has its own hydrophobic core region (*see* Article 69, **Complexity in biological structures and systems**, Volume 7). As a consequence of their autonomous folding capabilities, domains can often be excised from their host protein and pasted into a different protein context, without major changes in fold or function. Evolutionary processes such as exon shuffling and gene duplication, -fusion or -fission have created the multidomain "mosaic" structure found in many protein classes, a prime example being proteins involved in eukaryotic signal transduction pathways (Bork *et al*., 1997; Ponting *et al*., 1999; *see also* Article 9, **Modeling protein evolution**, Volume 1).

In the course of domain evolution, the structural fold and key functional aspects have been kept largely intact. Some domain types have proven to be very successful as mediators of a particular unit function, and variants are found over and over again in extant proteins. By a combination of bioinformatical and experimental approaches, it has been possible to detect protein domains even in the absence of a three-dimensional structure (Copley *et al*., 2002). Here, functional protein domains appear as "homology domains", that is, as regions of local sequence similarity in proteins that are otherwise unrelated. The exhaustive detection of homology domains in signal transduction proteins has been – and continues to be – an important prerequisite for understanding the physiological importance of these proteins (Attwood, 2000; Hofmann, 1998; Hofmann, 2000).

As many homology domains require sophisticated sequence analysis method for their detection, a number of tools, databases, and WWW servers have been set up to help the nonspecialist with that important task. Information on homology domains can be stored in the form of so-called Hidden Markov Models (HMMs)

Table 1 Web-based resources for domain detection

Database	URL of search page
InterPro	**http://www.ebi.ac.uk/InterProScan**
PROSITE	**http://hits.isb-sib.ch/cgi-bin/PFSCAN**
Pfam	**http://www.sanger.ac.uk/Software/Pfam/search.shtml**
SMART	**http://smart.embl-heidelberg.de**

Table 2 Functional domains in phospho-protein signaling. Only the most relevant domain types are shown, together with their INTERPRO accession number

Role	InterPro	Subtype
Protein kinases	IPR000719	Classical type
	IPR000403[a]	Lipid-kinase like
	IPR000687	Rio-type
	IPR004166	alpha-kinase
Protein phosphatases	IPR001932	Ser /Thr, PP2 C type
	IPR004843[a]	Ser/Thr, calcineurin-type
	IPR000106	Tyr, LMW-type
	IPR000242	Tyr, PTP-type
	IPR001763[a]	Tyr, Cdc25-type
	IPR000340	dual specificity
Recognition	IPR000980	SH2 (pTyr-recognition)
	IPR006020[a]	PTB/PID (pTyr)
	IPR000253	FHA (pSer/pThr)

[a]Domain family also includes members with different activities.

or generalized profiles (GPs) (Hofmann, 2000; *see also* Article 91, **Classification of proteins by sequence signatures**, Volume 6). Relevant databases, which maintain information on all domain types found in signal transduction proteins, are discussed in more detail elsewhere in this book (*see* Article 83, **InterPro**, Volume 6, Article 84, **Functionally and structurally relevant residues in PROSITE motif descriptors**, Volume 6, and Article 86, **Pfam: the protein families database**, Volume 6), an overview of the most important resources is given in Table 1. InterPro is a meta-database combining information stored in several domain databases (Mulder *et al*., 2003). Thus, the signaling domains described in the remainder of the text will be referenced by their InterPro accession numbers, listed in Tables 1 to 5.

2. Eukaryotic signaling systems

Eukaryotic cells must respond to a large number of external stimuli, which are typically sensed by means of specific receptors localized on the cell surface. As a result of such stimulation, cells have a large repertoire of possible responses, including increased or decreased proliferation, differentiation, or apoptosis. On the molecular level, these responses can be mediated by increased or decreased expression of specific genes, and also by changes in the translation level of

Table 3 Functional domains in small G-protein signaling

Role	InterPro	Subtype
Small GTPase	IPR003577	Ras
	IPR003578	Rho
	IPR003579	Rab
	IPR002041	Ran
Exchange factor	IPR001895	For Ras and similar
	IPR001331	For Rho/Rac/Cdc42
	IPR007515	For Rab proteins
	IPR001194	For Rab proteins
	IPR000408[a]	For Ran
GTPase activating	IPR001936	For Ras and similar
	IPR000198	For Rho/Rac/Cdc42
	IPR000195	For Rab proteins
	IPR000331	For Ran
Recognition	IPR003116	RAFRBD for Ras
	IPR000159	RA domain for Ras
	IPR000095	Crib domain for Rho etc.
	IPR003315[b]	RABPH for Rab proteins
	IPR004012[b]	RABRAP for Rab and Rap
	IPR000156	RanBP for Ran

[a]Domain family also includes members with different activities.
[b]Two out of many Rab-binding domains that recognize only selected Rab proteins.

Table 4 Functional domains in ubiquitin signaling

Role	InterPro	Subtype
Modifier	IPR000626[a]	Ubiquitin-like
E1, E2, E3 factors	IPR000011	Ubiquitin activating (E1)
	IPR000608	Ubiquitin conjugating (E2)
	IPR000569	HECT-type (E3)
	IPR001841	RING-finger (E3)
Deubiquitinating	IPR001394	USP type (for ubiquitin)
	IPR001578	UCH type (for ubiquitin)
	IPR003653	ULP type (for sumo etc.)
	IPR003323	OTU type (for ubiquitin?)
	IPR006155	Josephin type (for ubiquitin?)
Recognition	IPR000449	UBA (for ubiquitin, NEDD8)
	IPR003903	UIM (for ubiquitin)
	IPR003892[a]	CUE (for ubiquitin)

[a]Domain family also includes members with different activities.

particular mRNAs or in the stability of specific protein products. Eukaryotic cells have developed a number of pathways for conveying signals from the cell surface to the transcription-, translation-, or degradation machinery (Pawson, 2004). In some instances, there is a simple and straightforward pathway leading from a cell surface receptor to a transcriptional regulator. More frequently, however, signal transduction pathways look – at first sight – unnecessarily complicated and involve a multitude of intermediary components. The reason for this complexity lies most

Table 5 Functional domains in signaling adaptors

Role	InterPro	Subtype
Proline-based motif recognition	IPR001452	SH3
	IPR001202	WW
Death domain like 6-helix bundle	IPR000488	Death domain (DD)
	IPR001875	Death effector domain (DED)
	IPR001315	Caspase recruitment domain (CARD)
	IPR004020	Pyrin domain (PYD)

probably in the requirement for cells to integrate the signals coming from different sources before mounting the most appropriate response. The components of signal transduction pathways have evolved to allow for a well-balanced amount of cross talk between pathways. In many instances, the whole concept of an isolated "signal transduction pathway" is not really tenable, and it appears more appropriate to talk of a "signal transduction network" instead.

2.1. Modularity in signal transduction

A hallmark of eukaryotic signal transduction proteins, starting from the cell surface receptor down to the final effector component, is their modular architecture (Bork *et al*., 1997). As a common mode of action, the receptors transform the external signal (e.g., the binding of a ligand) into a different, intracellular signal, which is then recognized by a downstream component and potentially converted into a third signal and so forth. Canonical signal transduction proteins comprise several receptor- and effector-functionalities, sometimes augmented by additional regulatory components, all of which are encoded by different domains. Besides the generation and the detection of a signal, the termination of the signal can be of crucial importance. Thus, most signal transduction systems also include dedicated components for signal termination, which are also encoded by specific domain types. It would be outside the scope of this chapter to mention, or even to discuss all signal transduction pathways employed by eukaryotic cells. Figure 1 shows three well-studied signaling paradigms that will be used as examples for the following paragraphs.

Growth factor signaling, shown in Figure 1(a), was the first signaling pathway to be understood in detail and is still among the most complex pathways known (Pawson, 2004; *see also* Article 117, **EGFR network**, Volume 6). This pathway makes use of three widely used signaling paradigms: proximity activation of enzymes, protein phosphorylation, and GTPase signaling. In brief, the cell surface receptor binds the growth factor by its extracellular domain, leading to a receptor dimerization that activates the tyrosine kinase domains found in the cytoplasmic portion of the receptor. The resulting tyrosine phosphorylation of the receptor is sensed by the SH2-domains of a downstream component, which transduces the signal via a GDP/GTP exchange protein to the small GTPase Ras. The GTP-bound Ras in turn transduces the signal to a downstream cascade of kinases, eventually resulting in MAP-kinase activation. Finally, MAP-kinase phosphorylates a large number of target proteins, including transcriptional regulators. Each of the generated downstream signals has a specialized system for signal termination.

Cytokine signaling, shown in Figure 1(b), employs a simpler pathway from receptor dimerization to transcriptional regulation (Shuai and Liu, 2003). Here, the receptor does not contain a kinase domain but rather binds noncovalently to Jak-kinases and activates them upon receptor dimerization. The resulting autophosphorylation of Jak-kinases recruits factors of the STAT family via an SH2 domain. Subsequently, the STAT proteins themselves become phosphorylated, which allows them to translocate into the nucleus where they work as transcription

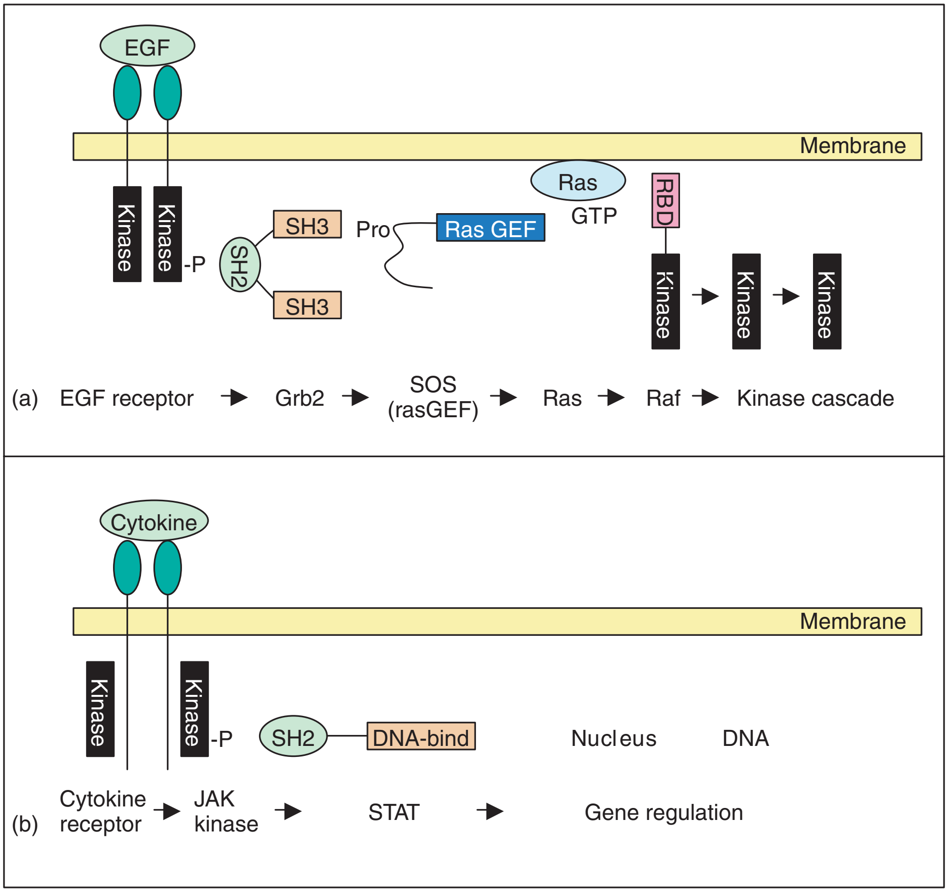


Figure 1 Simplified representation of three example pathways discussed in Section 2.1. (a) Growth factor signaling. The growth factor EGF and the membrane-associated Ras proteins are indicated by colored bubbles. The other signal transduction proteins are represented by their differently colored domains. The names of the corresponding proteins are given at the bottom of the figure. Interacting domains are shown juxtaposed to each other. Full domain names are given whenever possible. Abbreviated domain names are "RasGEF" for Ras GDP/GTP exchange domain, and "RBD" for the Ras-binding domain in the Raf kinase. (b) Cytokine signaling. Representation analogous to Figure 1(a). (c) Death receptor signaling. The trimeric death-ligand (e.g., FasL) and the mitochondrial proteins Bid and Cytochrome C are indicated by colored bubbles. The signal transduction proteins are represented by their differently colored domains, where DD, DED, and CARD mean "death domain", "death effector domain", and "caspase recruitment domain" respectively. Interacting domains are shown juxtaposed to each other. Caspase-8 cleaves Bid and Caspase-9 cleaves proCaspase-3, all other interactions are binding events

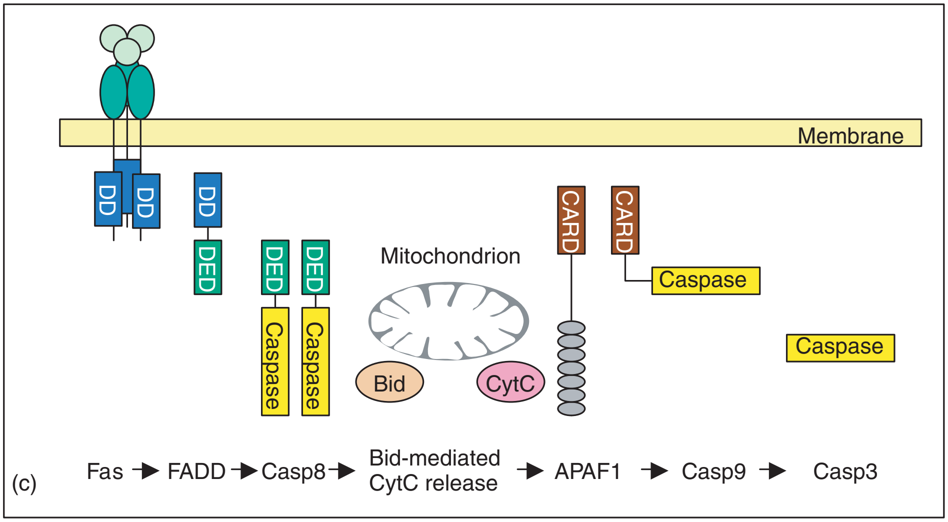


Figure 1 (*continued*)

factors by virtue of their DNA-binding and transactivation domains. Important regulators of this pathway work by selective ubiquitinating and thus downregulating proteins of the cytokine signaling pathway. Similar regulation modes by ubiquitination and selective protein degradation are known in many other signaling pathways as well.

Finally, the signaling pathway from death receptors to cell apoptosis is unusual in that it does not involve protein phosphorylation (Figure 1c) (Wajant, 2003). Central to this pathway is the activation of a protease class, the caspases, by induced proximity (Salvesen and Dixit, 1999). Caspases do not bind directly to the receptors, but rather use a cascade of unique adaptor protein containing domains of the death-domain 6-helix bundle superfamily (Aravind *et al*., 2001; Hofmann, 1999; Hofmann, 2003). Another unusual feature is the liberation of cytochrome C from the mitochondrion as an intermediate signal, which is sensed by the multidomain protein APAF-1. This binding event triggers the downstream branch of apoptosis signaling, which again involves caspases and members of the death-domain superfamily.

2.2. *Proximity activation of enzymes*

One of the recurrent motifs in eukaryotic signal transduction pathways is the activation of enzymes by induced proximity. Enzymatic activities that can be activated by this mechanism include protein kinases and proteases such as caspases. Frequently, these activities are encoded by functional domains, which are brought into close contact by the enzymatically inactive remainder of the protein. As a prerequisite for being susceptible to proximity activation, the enzyme must be inactive in its isolated form, or must be highly specific for a particular substrate to which it normally does not have access. Signaling processes, like for example receptor dimerization, bring two enzyme molecules or two enzymatic domains

into close contact. In the case of caspases, which are synthesized as virtually inactive proenzymes, this close proximity is sufficient for the mutual proteolytic removal of the inactivating prodomain by the very weak residual activity of the two proenzyme molecules (Salvesen and Dixit, 1999). As the caspase prodomains are also responsible for their dimerization, the activated caspases are liberated from the complex. In the case of receptor tyrosine kinases, the two associating kinase domains phosphorylate each other at a position that is not accessible for intramolecular auto-phosphorylation.

2.3. Protein phosphorylation

Protein phosphorylation is probably the most widely used mechanism used in eukaryotic signal transduction (Pawson, 2004; *see also* Article 63, **Protein phosphorylation analysis by mass spectrometry**, Volume 6). The signal itself consists of a phosphate group that becomes attached to the hydroxyl group of a tyrosine, serine, or threonine residue. A large variety of different signals exist, which are distinguished by the nature of the phosphorylated protein and by the site of phosphorylation. Many signaling components can become phosphorylated at multiple positions, sometimes with opposing effects. The phosphorylation signal is generated by protein kinases, whose enzymatic activities are typically encoded by functional kinase domains (see Table 2). Most known kinase domains are evolutionarily related. There are subtle differences between kinases with preference for tyrosine and those preferring either serine or threonine. One group, the so-called dual specificity kinases, are able to modify both tyrosine and serine/threonine residues, their catalytic domains closely resemble the pure tyrosine kinases. Besides the large number of "classical" protein kinases, there also exists a number of different protein domains associated with protein kinase activity (Table 2). Despite their heterogeneity, bioinformatical analyses suggest that all of those domains are distantly related to the canonical protein kinase domains.

Unlike the kinases, which are encoded by recognizable domains, there does not seem to be such a preference for phosphorylatable residues. Most kinases have a preference for certain residues flanking the modified residues, and these conditions appear sufficient to observe phosphorylation in vitro. However, these preferences are far too weak to explain the high phosphorylation specificity observed in vivo. In most cases, a specific substrate recognition by the noncatalytic part of the kinase molecule is required for efficient in vivo phosphorylation. Phosphorylation signals are not necessarily permanent, but can be removed by enzymes belonging to the class of protein phosphatases. In terms of protein relationship, phosphatases are a much more heterogeneous protein class than kinases. A number of different phosphatase domains have been described and are summarized in Table 2.

Besides the mechanisms for generating and terminating phosphorylation signals, cells must have a sophisticated system for the specific recognition of protein phosphorylation; this is typically performed by specialized phosphopeptide-recognition domains (Pawson, 2004; Tsai, 2002). The best known of these domains is the SH2 domain, which recognizes tyrosine phosphorylation. The PTB/PID domain has a similar specificity, but is more rare. The FHA domain seems to

perform an analogous task for the recognition of phosphorylated serine/threonine residues (Table 2).

2.4. Small G-protein signaling

Another important signaling system uses the nucleotide association status of small GTPases of the Ras superfamily (Geyer and Wittinghofer, 1997; Sprang, 1997). This system appears to be optimally tuned for reversibility and hence is frequently referred to as a "molecular switch". Small G-proteins, such as Ras, Rho, Rab, Ran, and others, can associate with either GTP or GDP. In many cellular systems, the GTP-associated form has an activating role, while the GDP-bound form is either inactive or even opposes the role of the GTP-form. Small G-proteins have a very low intrinsic GTPase activity, and without external influences both forms can be stable over a prolonged period of time. An activating signal can be generated by the GDP/GTP exchange factors, which load the small proteins with GTP. This activity is typically encoded by specialized domain classes, each of them responsible for one type of G-protein (see Table 3). Conversely, the activating signal can be terminated by a number of GTPase activating proteins (GAPs). GAP domains specifically bind to their cognate G-proteins and stimulate their intrinsic GTPase activity. Similar to the exchange factors, a particular class of GAP domains acts on each subclass of small G-proteins.

Owing to the large diversity of small G-proteins, there exist a large number of different systems for sensing their GTP/GDP status (Table 3). GTP-associated Ras can be recognized by the RAFRBD domain found in the Raf kinase of the growth hormone signaling pathway (Wittinghofer and Nassar, 1996). The RA domain is distantly related to the RAFRBD and occurs in a large number of Ras effectors. Similar systems exist for other G-proteins, including the Crib domain for Rho/Rac recognition and the RanBP domain for Ran recognition (Table 3).

2.5. Ubiquitination

The covalent modification of proteins with ubiquitin has initially been interpreted exclusively as an earmark for protein degradation by the proteasome. During recent years, however, it has become increasingly clear that protein ubiquitination is used for a variety of signaling purposes, including – but not limited to – different mechanisms for tightly regulated protein degradation (Pickart, 2004). Ubiquitination comes close to, or even surpasses protein phosphorylation in terms of flexibility of the generated signal and also in the number and complexity of signaling components involved (Di Fiore *et al*., 2003) (Table 4). While protein phosphorylation can modify serine, threonine, of tyrosine side chains, ubiquitination is restricted to the ε-amino groups of lysine, and possibly to the protein N-terminus. In contrast to the simple phosphorylation signal, there exists a multitude of ubiquitin-based signal, as ubiquitin itself can be ubiquitinated at various lysine residues, resulting in the formation of poly-ubiquitin chains of different lengths and linkage topologies. There is accumulating evidence that mono-ubiquitin and the different poly-ubiquitin chains all have different signaling capabilities (Schnell and

Hicke, 2003). An additional layer of complexity is provided by several ubiquitin-like modifiers that are distinct from ubiquitin but have similar capabilities for protein modification (Schwartz and Hochstrasser, 2003).

The transfer of ubiquitin and its relatives onto a protein requires a cascade of three enzyme classes termed "ubiquitin activating enzymes" (E1), "ubiquitin conjugating enzymes" (E2) and "ubiquitin ligases" (E3). There are two different classes of E3 enzymes, one of them based on HECT domains, the other one on RING fingers, which are a specific form of complex Zn-finger domains. E3 components clearly harbor most of the diversity, as these components confer specificity to the ubiquitination reaction. Analogous to the phosphatases in kinase signaling, there are a large number of proteases that specifically remove ubiquitin from target proteins (Wing, 2003). Similar to the ubiquitination enzymes, the deubiquitinating activity is typically encoded by dedicated domains, which are surrounded by protein interaction domains conferring target specificity. The ubiquitin signal can be recognized by a number of ubiquitin-sensing domains, the most prominent examples being the UBA, UIM, and CUE domains (Di Fiore *et al*., 2003).

2.6. *Signaling adaptors*

Not all intermediary steps in signal transduction involve the modification of proteins. As exemplified by the death receptor signaling pathway described above, "adaptor proteins" can play a crucial role, in particular, for mediating a controlled amount of pathway cross talk (Hofmann, 1999; Hofmann, 2003). A typical adaptor protein contains two different protein interaction domains, one of them binding to the receptor or another upstream component, and the other one recruiting a downstream signaling component into a "signaling complex".

A prime example is the DISC (for "death inducing signaling complex"), which forms at the cytoplasmic face of death receptors and is crucial for apoptosis signaling (Wajant, 2003). The cytoplasmic tail of death receptors like Fas and TNF-R55 contain a specific six-helix bundle domain termed "death domain" (DD), which has a high propensity to bind to other death domains. This property is used by FADD, a typical adaptor protein that has a death domain binding to the Fas receptor, combined with a "death effector domain" (DED) used for downstream signaling. The DED is distantly related to the death domain, but specifically recognizes other DEDs, but not DDs. The next downstream component, Caspase-8, contains two N-terminal DEDs, which recruit the enzyme to the receptor/adaptor complex, resulting in proximity activation of the caspase (Figure 1c). A third domain class belonging to the same six-helix fold is the "caspase recruitment domain" (CARD), which is found in other adaptor proteins and caspases, which work further downstream in apoptosis signaling. Recently, a fourth class of six-helix interaction domain has been described, the "pyrin domain" (PYD). All four domain classes share the propensity to interact with multiple other domains belong to the same class (Table 5).

Specific protein interactions are also crucial for other signaling events, including the growth factor pathway. For reasons that are not entirely clear, multiple signaling interactions make use of the specific recognition of short proline based motifs by dedicated domain classes, including the intensively studied SH3 and WW

domains (Macias *et al*., 2002; Zarrinpar *et al*., 2003) (Table 5). Although both domain types have similar recognition capabilities, the SH3 domain is prevalent in phosphorylation-based pathways, while the WW domain is more frequent in ubiquitination-based processes.

3. Signaling specificity and complexity

The previous paragraphs have briefly discussed some of the most important domain types found in eukaryotic signaling proteins. Obviously, this domain list is far from being complete. Important signaling systems, such as the heterotrimeric G-proteins and protein domains involved in their regulation have been omitted for space constraints. The same is true for domains that generate or sense small-molecule second messengers. Important examples are adenylate- and guanylate cyclases generating cAMP and cGMP, respectively, and phospholipase C enzymes involved in the generation of inositol- and lipid messengers. Analogous to protein signals, the small-molecule messengers are sensed by effector proteins, which typically also employ modular domains for that task. Several signaling pathways rely on the localization of its components, in particular subcellular compartments, such as the plasma membrane, the nucleus, or even some kind of nuclear body. Protein domains that mediate subcellular targeting, like for example the C2 domain, have also not been discussed.

Finally, one important question should be briefly addressed: If all members belonging to one domain class perform the same unit function, how is the necessary signaling specificity provided? There does not appear to be a general answer applying to all instances. In most cases, the unit function of a domain class is only weakly defined, like for example, “phosphorylation on a tyrosine residue” or “binding to a death domain protein”. There are marked differences between the various tyrosine kinase domains, not only in their sequence but also in their molecular properties. Thus, it is well possible to encode the specificity within the nonessential part of the homology domain. A second source of specificity comes from the use of specific targeting domains. Catalytic domains, which may be quite promiscuous in isolated form, are frequently found in a protein context next to a specificity-conferring protein interaction domain. The possibilities offered by this modular architecture are probably the major reason why the evolutionary shuffling of functional domains in signal transduction systems has turned out to be such a success story.

Related articles

Article 40, **Human signaling pathways analyzed by protein interaction mapping**, Volume 5; Article 83, **InterPro**, Volume 6; Article 91, **Classification of proteins by sequence signatures**, Volume 6; Article 108, **Functional networks in mammalian cells**, Volume 6; Article 43, **Evolution of regulatory networks**, Volume 7

References

Aravind L, Dixit VM and Koonin EV (2001) Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science*, **291**, 1279–1284.

Attwood TK (2000) The quest to deduce protein function from sequence: the role of pattern databases. *The International Journal of Biochemistry & Cell Biology*, **32**, 139–155.

Bork P, Schultz J and Ponting CP (1997) Cytoplasmic signalling domains: the next generation. *Trends in Biochemical Sciences*, **22**, 296–298.

Copley RR, Ponting CP, Schultz J and Bork P (2002) Sequence analysis of multidomain proteins: past perspectives and future directions. *Advances in Protein Chemistry*, **61**, 75–98.

Di Fiore PP, Polo S and Hofmann K (2003) When ubiquitin meets ubiquitin receptors: a signalling connection. *Nature Reviews. Molecular Cell Biology*, **4**, 491–497.

Geyer M and Wittinghofer A (1997) GEFs, GAPs, GDIs and effectors: taking a closer (3D) look at the regulation of Ras-related GTP-binding proteins. *Current Opinion in Structural Biology*, **7**, 786–792.

Hofmann K (1998) Protein classification & functional assignment. *Trends Guide to Bioinformatics*, **16**(Suppl 1) 18–21.

Hofmann K (1999) The modular nature of apoptotic signaling proteins. *Cellular and Molecular Life Sciences*, **55**, 1113–1128.

Hofmann K (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Briefings in Bioinformatics*, **1**, 167–178.

Hofmann K (2003) Chapter 4: Functional domains in apoptosis proteins. In *Genetics of Apoptosis*, Grimm S (Ed.), BIOS Scientific Publishers: Milton Park, pp. 67–91.

Macias MJ, Wiesner S and Sudol M (2002) WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Letters*, **513**, 30–37.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**, 315–318.

Pawson T (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell*, **116**, 191–203.

Pickart CM (2004) Back to the future with ubiquitin. *Cell*, **116**, 181–190.

Ponting CP, Aravind L, Schultz J, Bork P and Koonin EV (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *Journal of Molecular Biology*, **289**, 729–745.

Salvesen GS and Dixit VM (1999) Caspase activation: the induced-proximity model. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 10964–10967.

Schnell JD and Hicke L (2003) Non-traditional functions of ubiquitin and ubiquitin-binding proteins. *The Journal of Biological Chemistry*, **278**, 35857–35860.

Schwartz DC and Hochstrasser M (2003) A superfamily of protein tags: ubiquitin, SUMO and related modifiers. *Trends in Biochemical Sciences*, **28**, 321–328.

Shuai K and Liu B (2003) Regulation of JAK-STAT signalling in the immune system. *Nature Reviews. Immunology*, **3**, 900–911.

Sprang SR (1997) G proteins, effectors and GAPs: structure and mechanism. *Current Opinion in Structural Biology*, **7**, 849–856.

Tsai MD (2002) FHA: a signal transduction domain with diverse specificity and function. *Structure (Cambridge)*, **10**, 887–888.

Wajant H (2003) Death receptors. *Essays in Biochemistry*, **39**, 53–71.

Wing SS (2003) Deubiquitinating enzymes–the importance of driving in reverse along the ubiquitin-proteasome pathway. *The International Journal of Biochemistry & Cell Biology*, **35**, 590–605.

Wittinghofer A and Nassar N (1996) How Ras-related proteins talk to their effectors. *Trends in Biochemical Sciences*, **21**, 488–491.

Zarrinpar A, Bhattacharyya RP and Lim WA (2003) The structure and function of proline recognition domains. *Science's STKE*, **2003**, RE8.

Short Specialist Review

Sequence complexity of proteins and its significance in annotation

Birgit Eisenhaber and Frank Eisenhaber

Institute of Molecular Pathology, Vienna, Austria

1. Introduction

The concept of complexity of protein sequences originated from the consideration of sequences as strings of symbols that can be studied with linguistic methods. Simple descriptors such as amino acid compositional properties attracted early attention. These studies suggested that small globular proteins can be classified in accordance to amino acid composition and this classification corresponds roughly to secondary structural class, cellular localization, and enzyme function (Nishikawa and Ooi, 1982; Nishikawa *et al*., 1983a; Nishikawa *et al*., 1983b). Nevertheless, the predictive value of this correlation is limited, as later reanalyses with more abundant structural data have shown (Clementi *et al*., 1997; Eisenhaber *et al*., 1996b; Eisenhaber *et al*., 1996a).

With increased efficiency of sequencing in the eighties and nineties of the twentieth century, sequences of multidomain proteins became readily available. Some of their sequence regions differed markedly from previously known globular proteins by obvious abundance of a single or a combination of only very few amino acid types; therefore, these sequence segments were considered “unusual” or “extraordinary”. From the sequence point of view, these compositionally biased regions can be homopolymeric runs, irregular mosaics of a few residue types, or short-period (almost) regular repeats. Genome sequencing projects have shown that such proteins occur frequently, especially in eukaryote proteomes. The fly protein “brakeless” (acc. AAF76322, 2302 AA; its loss of function is associated with unlimited growth of optical axons) is an extreme example. The N- and C-terminal halfs (each of almost 1000 residues) are serine- and glutamine-rich respectively and only three tiny, scattered islands in the middle of the sequence are apparently of globular structure (Senti *et al*., 2000).

The discovery that amino acid compositional bias over domain-size segments is associated with the property of having a globular or nonglobular structure has fuelled the development of criteria for sequence complexity and methods to delineate simple, potentially nonglobular sequence segments as low complexity regions.

2. Criteria for measuring sequence complexity and methods for delineating compositionally biased regions

Whereas sequences from globular domains appear visually as "good" mixtures of residue types, the biased sequence segments are simple because of (1) their nonrandom, highly biased amino acid composition and (2) their tendency of repeating monotonously a small sequence motif. Various methods address these two points but indirect approaches are also possible. Sequence simplicity is often but not always associated with conformational flexibility (invisibility in X-ray structures or only with high B-factor), absence of regular secondary structure (coil status), and high solvent accessibility over long stretches; thus, tools that search for these properties might also be useful to segment a query protein.

The compositional aspect can be evaluated with compositional complexity measures calculated over sequence windows. Given an amino acid composition $\{n_i\}_{i=1,\ldots,20}$ and a sequence window of length $L = \Sigma n_i$, there are $N = L!/\Pi n_i$ possible sequences. In the SEG algorithm (Wootton, 1994a; Wootton and Federhen, 1996), subthreshold segments for the Shannon entropylike term

$$K_2 = -\sum \frac{n_i}{L} \log \frac{n_i}{L} \qquad (1)$$

are searched for and concatenated to a raw region first, after which the subsegment of the raw region with lowest probability $K_1 = \log N/L$ is found. Depending on the window length and thresholds, the sensitivity of SEG can be increased and it can find even regions with subtle compositional deviations such as coiled coil regions.

Other methods such as CAST (Promponas *et al*., 2000) or P-SIMPLE (Alba *et al*., 2002) search for simple segments with bias toward a single amino acid type or a small motif (up to four residues), the first one by assessing alignments with homopolymeric runs and the second by counting motif occurrences in sequence windows. The POPP approach (Wise, 2002) focuses on the frequency of specific mono-, di-, and tripeptides in whole proteins or specified segments and compares them with distributions in sequence databases.

The Globplot tool (Linding *et al*., 2003b) is a propensity-based predictor and the DisEMBL package (Linding *et al*., 2003a) is a suite of neural networks, both parametrized from coil segments in known 3D (three dimensional) structures and can serve for finding extended unordered regions.

Any of the above techniques misses a notable share of nonglobular regions (Kreil and Ouzounis, 2004). With parametrizations for high sensitivity, the false–positive prediction rate is considerable. Thus, uncurated computer-generated annotations of protein sequence databases with these methods are not advisable.

3. Occurrence of compositionally biased regions in protein sequences, their classification, and their role in human disease

It is an early observation that amino acid sequence segments with amino acid compositional bias can occur in natural proteins but, originally, they were

considered exceptional and rare. This view originated from the restricted availability of sequences, mostly of short, single-, or two-domain proteins with well-characterized 3D structure (typically, with metabolic function and of prokaryote origin). Indeed, globular domains with known 3D structure contain a well-balanced composition of various hydrophobic and different polar residues with functional groups. Otherwise, folding with tight packing in the core and a polar, solvatable surface would not have been possible. Not surprisingly, compositionally biased segments are rare in globular domains and comprise only about 0.5% of the sequence (Wootton, 1994b). They are usually short (<30 AA, ∼15 residues on average). Most of the shorter ones appear as moderately polar stretches as part of long solvent-accessible loops or at the polypeptide termini, whereas the longer segments typically represent hydrophobic or amphipatic helices and are involved in the structural packing (Saqi, 1995).

With more efficient sequencing technologies and especially with the genome sequencing projects, the available subset of protein sequences has become more representative. Usually, proteins consist of functionally different sequence regions. Some of them may represent globular domains but the others have compositional bias and, typically, have a nonglobular nature. The amount of low complexity regions in large protein databases has been evaluated with the SEG program and was found to comprise about a quarter of all residues in known proteins. Interestingly, the fraction of low complexity sequence is higher in eukaryotes (an estimated one-third of the total proteome in *Drosophila melanogaster*) compared to ∼10% in prokaryotes (Wootton, 1994b; Wootton and Federhen, 1996; Huynen *et al*., 1998).

The status of structural and functional characterization of low complexity regions is highly diverse. Some examples are listed in Wootton (1994b). Reasonably long spans (often rich in proline and small polar residues) between globular domains apparently serve just as conformationally flexible linkers. Low complexity segments with bias toward hydrophobic residues or with a repetitive hydrophobic pattern are also not so problematic. They are typically found in membrane-attachment regions, are buried in protein complexes, and/or form fibrillar structures.

In contrast, the structure forming potential and the molecular function of long low complexity regions with primarily polar residues remains insufficiently understood. Apparently, many of them carry sites for posttranslational modifications and interactions with other biomacromolecules and are important for cellular regulation and developmental processes (Wootton, 1994b). For example, mutational expansions of CAG triplets coding polyglutamine regions in human genes beyond a certain length cause neurodegenerative disorders such as Huntington's disease, spinocerebellar ataxia 6, or Kennedy disease (Perutz, 2004). Surprisingly, polyasparagine repeats are suppressed in mammalian proteins (Kreil and Kreil, 2000). Many intrinsically unstructured proteins such as titin, prion, tau, and others appear to evolve via repeat expansion (Tompa, 2003).

Polar low complexity regions are especially abundant in the known part of the proteome from *Plasmodium falciparum*, *Plasmodium berghei*, and *Dictyostelium discoideum* – even relative to the background of other eukarya (Pizzi and Frontali, 2001). Possibly, variable immuno-dominant epitope loops of transmembrane proteins and the variety of glycosylation variants in polyasparagine tracts have

clinical relevance since they are part of *Plasmodium*'s strategy to evade the immune response of the host (Duffy *et al*., 2003; Newbold, 1999).

The strict criteria for selecting low complexity segments in standard tools overlook many instances of regions with milder compositional bias that are nevertheless not globular either. The environment of sites for many posttranslational modifications (PTM) and subcellular translocation signals are examples of the more general theme of interaction of a globular protein with an extended and conformationally malleable region of another one (Eisenhaber *et al*., 2004; Iakoucheva *et al*., 2004). In the case of a PTM, only a small region of the binding site of the substrate protein directly interacts with the active site cavity of the modifying enzyme. To allow this interaction physically, the site of the PTM has to be surrounded by a sufficiently solvent-interacting sequence region without own inherent conformational preferences that links the site of the PTM with the remainder of the substrate protein (Eisenhaber *et al*., 2004).

4. The homology concept and the role of compositionally biased regions in sequence similarity searches

The most successful and widely known approach for inferring protein function with nonexperimental means, the annotation transfer from homologs, is based on the observation of similarity between protein sequences having similar 3D structure and molecular function. In evolutionary terms, even distantly related sequences are considered to originate from a common ancestor sequence in a mutational process.

The distance metrics used in sequence similarity searches, for example, the E-values in the popular BLAST/PSI-BLAST suites (Altschul *et al*., 1994; Altschul *et al*., 1997) that evaluate the probability of having a common ancestor, are correctly applicable only for globular sequence segments. The amino acid type substitution matrices such as BLOSUM62 (Henikoff and Henikoff, 2000) are derived from datasets of point mutations in secondary structural elements in globular domains of well-studied protein families and are, therefore, designed to detect subtle sequence similarities between distantly related sequence instances of the same family. Not surprisingly, false–positive hits in database searches are typically caused by compositionally biased regions, for example, cysteine- or proline-rich regions, long hydrophobic stretches of transmembrane helices/signal peptides or monotonous hydrophobic patterns of fibrillar structures.

Hence, it is necessary to remove compositionally biased regions from query sequences before sequence similarity searches. As a standard option, BLAST offers the most stringent standard parametrization of the SEG program for masking obviously low complex segments. Especially in the case of frequently interspersed homopolymeric runs, CAST is a good alternative (Kreil and Ouzounis, 2004). Often, both approaches are insufficient to remove the cause of collecting false hits and it becomes necessary to carefully determine membrane-embedded regions, fibrillar segments, peptides that are responsible for targeting to subcellular localizations (for example, signal peptides), regions responsible for posttranslational modifications (Eisenhaber *et al*., 2003; Nielsen *et al*., 1999), and other types of nonglobular segments in the query prior to the similarity search.

References

Alba MM, Laskowski RA and Hancock JM (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672–678.

Altschul S, Boguski M, Gish W and Wootton JC (1994) Issues in searching molecular sequence databases. *Nature Genetics*, **6**, 119–129.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Clementi M, Clementi S, Cruciani G, Pastor M, Davis AM and Flower DR (1997) Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Engineering*, **10**, 747–749.

Duffy MF, Reeder JC and Brown GV (2003) Regulation of antigenic variation in Plasmodium falciparum: censoring freedom of expression? *Trends in Parasitology*, **19**, 121–124.

Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G and Wildpaner M (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. *Nucleic Acids Research*, **31**, 3631–3634.

Eisenhaber F, Frömmel C and Argos P (1996a) Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins:Structure, Functions, Genetics*, **25**, 169–179.

Eisenhaber F, Imperiale F, Argos P and Frömmel C (1996b) Prediction of secondary structural content of proteins from their amino acid composition alone. I. New vector decomposition methods. *Proteins:Structure, Functions, Genetics*, **25**, 157–168.

Eisenhaber B, Eisenhaber F, Maurer-Stroh S and Neuberger G (2004) Prediction of sequence signals for lipid post-translational modifications: insights from case studies. *Proteomics*, **4**, 1614–1625.

Henikoff S and Henikoff JG (2000) Amino acid substitution matrices. *Advances in Protein Chemistry*, **54**, 73–97.

Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev SR, Yuan Y and Bork P (1998) Homology-based fold predictions for *mycoplasma genitalium* proteins. *Journal of Molecular Biology*, **280**, 323–326.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z and Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, **32**, 1037–1049.

Kreil DP and Kreil G (2000) Asparagine repeats are rare in mammalian proteins. *Trends in Biochemical Sciences*, **25**, 270–271.

Kreil DP and Ouzounis C (2004) Comparison of sequence masking algorithms and the detection of biased protein sequence regions. *Bioinformatics*, **19**, 1672–1681.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ and Russell RB (2003a) Protein disorder prediction: implications for structural proteomics. *Structure (Cambridge)*, **11**, 1453–1459.

Linding R, Russell RB, Neduva V and Gibson TJ (2003b) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, **31**, 3701–3708.

Newbold CI (1999) Antigenic variation in Plasmodium falciparum: mechanisms and consequences. *Current Opinion in Microbiology*, **2**, 420–425.

Nielsen H, Brunak S and von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, **12**, 3–9.

Nishikawa K, Kubota Y and Ooi T (1983a) Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *Journal of Biochemistry*, **94**, 981–995.

Nishikawa K, Kubota Y and Ooi T (1983b) Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *Journal of Biochemistry*, **94**, 997–1007.

Nishikawa K and Ooi T (1982) Correlation of the amino acid composition of a protein to its structural and biological characters. *Journal of Biochemistry*, **91**, 1821–1824.

Perutz MF (2004) Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends in Biochemical Sciences*, **24**, 58–63.

Pizzi E and Frontali C (2001) Low-complexity regions in plasmodium falciparum proteins. *Genome Research*, **11**, 218–229.

Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C and Ouzounis CA (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.

Saqi M (1995) An analysis of structural instances of low complexity sequence segments. *Protein Engineering*, **8**, 1069–1073.

Senti K, Keleman K, Eisenhaber F and Dickson BJ (2000) brakeless is required for lamina targeting of R1-R6 axons in the Drosophila visual system. *Development*, **127**, 2291–2301.

Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*, **25**, 847–855.

Wise MJ (2002) The POPPs: clustering and searching using peptide probability profiles. *Bioinformatics*, **18**(Suppl 1), S38–S45.

Wootton JC (1994a) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers & Chemistry*, **18**, 269–285.

Wootton JC (1994b) Sequences with 'unusual' amino acid compositions. *Current Opinion in Structural Biology*, **4**, 413–421.

Wootton JC and Federhen S (1996) Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*, **266**, 554–571.

Short Specialist Review

Protein repeats

Carolina Perez-Iratxeta and Miguel A. Andrade

Ottawa Health Research Institute, Ottawa, ON, Canada

1. Definition

Some protein domains are composed of units of similar structure (see Figure 1). Often, but not always, these units are also similar in sequence, which explains why they have a similar structure. Therefore, they can be considered protein repeats that originated by duplications from a single ancestral sequence.

These small units are large enough to form secondary structural elements but too small to be stable by themselves. They acquire stability by folding together in a repetitive structure. This is the defining feature of protein repeats. They occur repeated, never alone.

Repeats tend to be tandemly repeated (one next to the other in sequence), but the occurrence of insertions of variable length in between repeats is sometimes observed. Those insertions tend to lack structure and to be short, although there seems to be no obvious reason for that. We can speculate that because folding of a structure composed of repeats is more delicate than that of globular folds, large insertions would make the fold unstable or difficult to form.

Since repeats are consecutive in sequence, a condition on their structure is that the C-terminal must be spatially near the N-terminal in order to produce units that can fold together in a compact continuum of repeat units. Given the fact that the elements of secondary structure are quite linear, a repeat must contain a minimum of two of them to bring the C-terminal of the repeat close to the N-terminal. This imposes a minimum length of about 12 amino acids on these repeats. An upper bound of about 50 amino acids can be explained because a larger unit could fold into an independent domain that does not need another repeat to form a stable fold.

2. Classification

Repeats can be classified, like protein folds, depending on the secondary structural elements that compose them. Accordingly, most widespread repeats can be grouped in three groups, all-alpha, all-beta, and alpha/beta:

- Armadillo/HEAT repeats and tetratricopeptide repeats (TPR-like) in the all-alpha group.

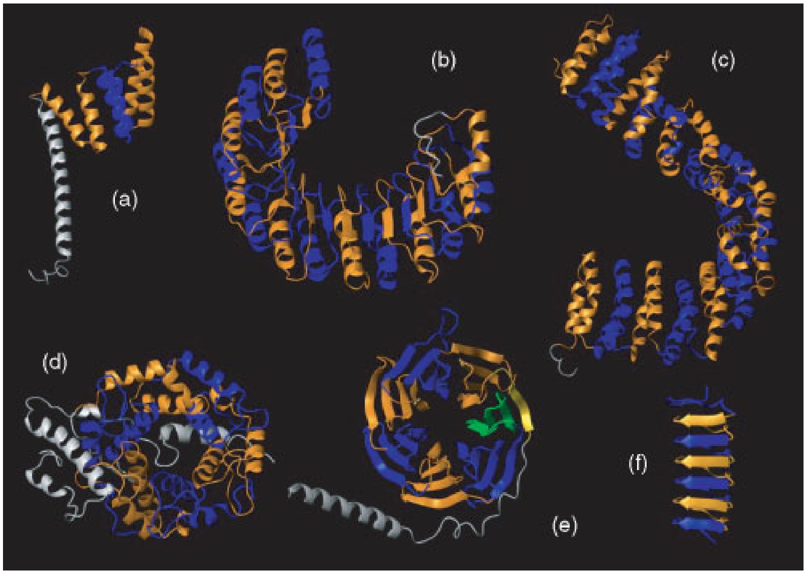


Figure 1 Examples of protein domains composed of repeats. (a) TPR repeats (PDB id: 1a17). (b) Leucine-rich repeats (PDB id: 1dfj, chain I). (c) HEAT repeats (PDB id: 1b3u). (d) PFTA repeats (PDB id: 1ft2, chain A). (e) WD40 repeats (PDB id: 1gp2, chain B). (f) Beta-loop repeats in an antifreeze protein (PDB id: 1ezg). Blue and orange represent alternating repeats. The rest of the structure is colored in grey. In the case of WD40 (e), one of the repeats is composed of a beta strand (in yellow) previous to the first repeat and three C-terminal beta strands (in green); this way of closing a barrel composed of beta-sheets is observed in other protein families of similar structure. The figure was composed using the MOLMOL software (Koradi *et al*., 1996)

- Beta propellers (as WD40 and Kelch) and beta trefoils in the all-beta group.
- Leucine-rich repeats (LRR) and ankyrin repeats in the alpha/beta group.

There are many more repeat families that occur less frequently (Andrade *et al*., 2001a), such as the beta-loop 12-residue repeats found in an insect antifreeze protein.

The shape of the superstructure can also define the function and dynamics of the repeat unit and, therefore, must also be taken into account. Repeats can form open structures (see examples in Figures 1a, 1b, 1c, and 1f) or closed structures (see examples in Figures 1d and 1e). Open structures formed by repeats are observed with variable copy number, whereas close structures are more constrained to a copy number appropriate to the closing of the superstructure (Andrade *et al*., 2000).

3. Structure and function

In general, domains composed of repeats confer to proteins an enlarged binding surface area with the possibility of multiple binding and structural roles. This is the reason protein–protein interaction is the most prevalent function of proteins with repeats (Andrade *et al*., 2001b). The periodicity of the three-dimensional structure

favors spatially periodic or symmetric interactions that are critical for structural integrity in various cellular contexts (for example, spectrin in the submembrane lamina or HEAT repeats in the vesicular coat structures). Among other functions performed by proteins with repeats, we can find enzymatic activities, ice binding (in antifreeze proteins), or energy storage proteins in plant seeds.

4. Evolution and distribution

Repeats have appeared multiple times along evolution. They originate by duplication and recombination within a single gene. The evolution of repeats from a common ancestor that necessarily must have contained a single repeat seems paradoxical, given that apparently more than one repeat is required for folding. One possible explanation is that the ancestor formed homo-oligomers where two identical chains provided the repetition (Ponting *et al*., 2000).

It is estimated that 14% of all proteins have repeats (Marcotte *et al*., 1999). Their actual number is probably higher, because the estimate is based on sequence comparison, and therefore missing those that have no detectable sequence similarity.

Repeats are present over all taxa but are more abundant in eukaryotic organisms and among them most frequent in metazoans. This distribution shows a correlation between the increase of organism complexity and the already mentioned principal function of domains composed of repeats, binding proteins.

5. Detection

Identifying tandem repeats with high sequence similarity is relatively easy. However, it happens that the constraints in sequence conservation among repeats are relatively lax. For example, the HEAT repeats of the structure displayed in Figure 1c average only 18% of sequence identity. Their short length complicates their detection too, and the number of repeats per protein may differ between members of the same family. In the absence of a reference structure, it is especially difficult to detect the boundaries of each repeated unit. Several methods have been published dealing with the detection and analysis of repeats (REP, **http://www.embl.de/~andrade/papers/rep/search.html**, Andrade *et al*., 2000; REPRO, **http://ibivu.cs.vu.nl/programs/reprowww/**, George and Heringa, 2000; Pellegrini *et al*., 1999).

References

Andrade MA, Perez-Iratxeta C and Ponting CP (2001a) Proteins repeats: structures, functions and evolution. *Journal of Structural Biology*, **134**, 117–131.

Andrade MA, Ponting CP, Gibson TJ and Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *Journal of Molecular Biology*, **298**, 521–537.

Andrade MA, Petosa C, O'Donoghue SI, Muller CW and Bork P (2001b) Comparison of ARM and HEAT protein repeats. *Journal of Molecular Biology*, **309**, 1–18.

George RA and Heringa J (2000) The REPRO server: finding protein internal sequence repeats through the web. *Trends in Biochemical Sciences*, **25**, 515–517.

Koradi R, Billeter M and Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *Journal of Molecular Graphics*, **14**, 51–55.

Marcotte EM, Pellegrini M, Yeates TO and Eisenberg D (1999) A census of protein repeats. *Journal of Molecular Biology*, **293**, 151–160.

Pellegrini M, Marcotte EM and Yeates TO (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins*, **35**, 440–446.

Ponting CP Schultz J, Copley RR, Andrade MA and Bork P (2000) Evolution of domain families. *Advances in Protein Chemistry*, **54**, 185–244.

Short Specialist Review

Large-scale protein annotation

Sarah K. Kummerfeld
MRC Laboratory of Molecular Biology, Cambridge, UK

1. Introduction

A bacterial genome encodes anywhere between 500 and 5000 proteins, whereas a large eukaryotic genome encodes as many as 25 000 proteins. Making sense of a whole-genome set of amino acid sequences may sound like a daunting task; if your approach were to consider each protein in turn through literature review and traditional experimental biochemistry, you would be right to be worried. However, computational approaches have been developed to allow automatic prediction of structure, function, and evolutionary relationships from amino acid sequence data.

This review provides an overview of large-scale computational techniques and databases by tracing the annotation of a hypothetical proteome. Our approach uses catalogs of experimentally determined protein annotation and homology methods to accurately annotate as much of the proteome as possible. Finally, we discuss how whole-proteome annotation is being used to address novel biological questions.

2. Catalogs of experimentally determined protein annotation

To begin our hypothetical proteome annotation, suppose we are given 5000 polypeptides derived from genome sequencing and gene prediction of our genome of interest. A first step is to annotate proteins for which we have experimentally derived information. Increasingly, databases are being established to store details about protein structure and function in a way that can be automatically searched.

The main repository for experimentally determined structures is the Protein Data Bank (PDB) (**http://www.rcsb.org/pdb/**, Berman *et al*., 2000), which included almost 30 000 structures at the time of writing. For functional annotation, there are many experimentally derived databases. These differ in the taxonomic source or functional class of the proteins they include and the type of experiments used to obtain the functional information.

One of the largest databases of functional annotation for proteins from across all kingdoms of life is SwissProt (**http://us.expasy.org/sprot/**, Boeckmann *et al*., 2003). Where available, annotation includes details about each protein's function, posttranslational modifications, structure, domains, similarity to other proteins, and associated diseases. There are also many system- or genome-specific catalogs

including Drosophila genome database, FlyBase (**http://flybase.bio.indiana.edu/**, The FlyBase Consortium, 2003), Biochemical pathway database, EcoCyc (**http://ecocyc.org/**, Karp *et al*., 2002), Kyoto Encyclopaedia of Genes and Genomes (**http://www.genome.jp/kegg/**, Kanehisa, 1997), which includes a comprehensive section on pathways and signal transduction, BIND biomolecular interaction network database for protein interactions (**http://bind.ca/**, Bader *et al*., 2003), and Online Mendelian Inheritance in Man for human disease genes (**http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM**, McKusick, 1998).

An example of a higher-level functional annotation is the GeneOntology database (GO) (**http://www.geneontology.org/**, The Gene Ontology Consortium, 2000). GO provides a treelike functional hierarchy, with mappings to various protein annotation databases, for example, FlyBase and PDB. The hierarchy provides a common language for comparison between species or datasets that individually have different systems of annotation.

Directly searching structural and functional databases described above would yield annotation for a small subset of the 5000 sequences in our proteome of interest; representing proteins that had been experimentally characterized before the genome was sequenced. These annotations would be expected to be extremely accurate as they are limited only by the representation of the data in the database and experiments themselves. However, this leaves the majority of proteins unannotated. The next section describes how we can use homology to predict the structure and/or function of proteins for which no experimental results are available.

3. Homology methods for annotation

Proteins with very similar amino acid sequence tend to adopt a similar fold and consequently have similar functions. This principle is used to infer structure/function by homology: proteins with no known structure/function are compared with proteins from experimentally characterized catalogs. In cases where the unknown protein is matched as being similar to one in the characterized set, we can infer that their structures and/or functions are similar. This is the principle for the protein annotation part of many genome annotation systems, for example, the Ensembl analysis pipeline (Potter *et al*., 2004). In this section, we outline a standard set of methods that may be used by an annotation pipeline.

There is a large range of methods that can be used to assess homology. One of the simplest is pairwise sequence comparison using a program such as BLAST (Altschul *et al*., 1990) or FASTA (Pearson and Lipman, 1988). Returning to our hypothetical proteome, we consider each protein in turn as a query sequence, and we use BLAST to search against the catalogs of experimentally characterized proteins as the database. BLAST returns a list of matches between our query and the database, indicating the likelihood that this match occurred by chance. For example, by comparing each protein from our set of 5000 with the SwissProt functionally characterized proteins, we are able to predict the function for all the proteins that have significant matches to a SwissProt annotated protein. These database searches are fast and allow immediate inference of structure or function for proteins with a characterized homolog.

Pairwise sequence comparison can also be used to assess the evolutionary relationship between proteins in our set and other proteomes. For example, searching each of the 5000 proteins in turn with BLAST against a database containing the proteomes of three closely related organisms will allow identification of orthologs and paralogs between the genomes. There are purpose-built tools to perform these evaluations including INPARANOID (**http://inparanoid.cgb.ki.se/**, Remm *et al*., 2001) for prediction orthologs and Clusters of Orthologous Groups (**http://www.ncbi.nlm.nih.gov/COG/**, Tatusov *et al*., 1997) for grouping functionally related paralogs.

While the details of pairwise sequence comparison are beyond the scope of this review, it is important to understand that as a method of homology detection, pairwise sequence comparison is only useful for detecting relatively close homologs. Even if two proteins adopt an identical fold and function, BLAST will not recognize them as homologs if their protein sequences are sufficiently different. A second set of homology methods uses sequence profiles to match structurally homologous proteins with lower sequence similarity than would be possible to identify with BLAST. Instead of comparing the query sequence to a sequence database, profile methods compare the query to a set of profiles. Profiles are based on alignments of closely related sequences; they indicate the probability that each amino acid should occur at a particular position in the sequence. This effectively encodes the underlying information about the conservation pattern across an entire group of related proteins, and provides a model that allows detection of more divergent homologs.

Some of the most extensive domain profile databases are Pfam (**http://www.sanger.ac.uk/Software/Pfam/**, Bateman *et al*., 2004), SUPERFAMILY (**http://supfam.org**, Gough *et al*., 2001), ProDom (**http://protein.toulouse.inra.fr/prodom/current/html/home.php**, Corpet *et al*., 2000), and Tigrfams (**http://www.tigr.org/TIGRFAMs/**, Haft *et al*., 2003). They can be used to identify functionally relevant and evolutionarily conserved sequence features as well as globular domains in proteins. The Pfam database provides a curated multiple sequence alignments for functional domain families. Using the HMMER (**http://hmmer.wustl.edu/**) sequence-profile search software, it is possible to search each of our 5000 sequences against the library of profiles. This generates a list of regions where the query sequence has a significant match to a particular domain family profile. Similarly, the SUPERFAMILY database provides profiles for structural, evolutionary-related domains. Probabilistic models are also used for the prediction of signal peptides in SignalP (**http://www.cbs.dtu.dk/services/SignalP/**, Bendtsen *et al*., 2004) and transmembrane helices in TMHMM (**http://www.cbs.dtu.dk/services/TMHMM/**, Krogh *et al*., 2001). The predicted location of Pfam (functional) domains, SUPERFAMILY (structural/evolutionary) domains, signal peptides, and transmembrane helices further contribute to the annotation of our proteome.

As with pairwise sequence comparison, profile methods can tell us a lot about the evolutionary history of proteins. In particular, a database such as SUPERFAMILY, whose domain definition is based on evolution, allows comparison of the domain repertoire of different proteomes.

The final type of database that we will discuss is based primarily on patterns. Examples include PROSITE (**http://us.expasy.org/prosite/**, Hulo *et al*., 2004) and

PRINTS (**http://bioinf.man.ac.uk/dbbrowser/PRINTS/**, Attwood *et al*., 2003). Rather than grouping homologous domains or proteins, these databases include motifs or patterns that have a particular functional role. For example, PROSITE patterns indicate biologically important functional sites. The results of searching for these patterns in our set of 5000 proteins can then be added to the functional annotation.

4. Interpreting whole-proteome annotation

Having used homology methods to infer the structure and function for as many proteins in our set of 5000 as possible, this leaves us asking: what does this tell us? Now that we have a wealth of functional, structural, and evolutionary information, we can begin to ask questions about the biology of our newly sequenced genome and how it compares with other organisms. Here we present two examples of biologically important discoveries that have been made on the basis of annotation from freshly sequenced proteomes: assessment of a newly sequenced organism's potential as a model system and exploring the evolutionary relationship of protein domains.

Identifying model organisms to study human diseases is a fundamental step in medical research. *Dictyostelium discoideum*, slime mold, has long been held up as a useful experimental system because it is relatively simple to work with, but has features that are not found in early-branching eukaryotes, including complex signal transduction, pattern formation, phagocytosis, and some aspects of development. On completion of the genome sequence, the proteome was annotated (using a similar procedure to our hypothetical example), providing a parts list of domains present in the proteome. The Pfam domain assignments, serving as a functional annotation, were used to carry out a protein-level comparison between Dictyostelium and other eukaryotes. This allowed identification of proteins common to Dictyostelium and animals, including domains related to signaling and cell adhesion. The proteins found to have these higher-eukaryote-like domains are now candidates for use as experimental systems to study intercell signaling and adhesion.

A second example of interpreting large-scale protein annotation data aims to understand how modular proteins have evolved. A recent study (Vogel *et al*., 2004) compared the patterns of domains found in 131 completely sequenced genomes. By considering the N-C terminal arrangement of domains, Vogel *et al*. found that particular two and three domains patterns are overrepresented. Using confirmation from structural information, they were able to show that these groups of domains are themselves evolutionary units, termed "supradomains", which have been duplicated as a set. Not only did this study extend our understanding of how proteins have evolved, it also provided a list of high-priority targets for structural genomic projects. Since supradomains are overrepresented, solving structures for representatives will significantly increase structural coverage of protein space.

Large-scale computational protein annotation provides feasible methods for predicting the structure, function, and evolutionary history of proteins from their amino acid sequence. Using remote homology detection methods, we can make accurate (less than 1% error) predictions about the domain structure of between 30

and 80% of proteins in individual genomes and 60% of proteins in the UNIPROT database. Experimentally determined functional annotation is available for a smaller fraction of proteins; through homology, this information can be applied to whole genomes. Current genome annotations are sufficient to answer questions about the biology of whole organisms, their systems, and evolutionary relationships.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, *et al*. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, **31**, 400–402.

Bader GD, Betel D and Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, **31**(1), 248–250.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783–795.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al*. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

Corpet F, Servant F, Gouzy J and Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, **28**, 267–269.

Gough J, Karplus K, Hughey R and Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, **313**, 903–919.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**, D134–D137.

Kanehisa M (1997) A database for post-genome analysis. *Trends in Genetics*, **13**, 375–376.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C and Gama-Castro S (2002) The EcoCyc Database. *Nucleic Acids Research*, **30**, 56–58.

Krogh A, Larsson B, von Heijne G and Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

McKusick VA (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, *Twelfth Edition*, Johns Hopkins University Press: Baltimore.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R and Clamp M (2004) The Ensembl analysis pipeline. *Genome Research*, **14**(5), 934–941.

Remm M, Storm CEV and Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**, 1041–1052.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research*, **31**, 172–175.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Vogel C, Berzuini C, Bashton M, Gough J and Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. *Journal of Molecular Biology*, **336**, 809–823.

Measuring evolutionary constraints as protein properties reflecting underlying mechanisms

Andrew F. Neuwald
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Jun S. Liu
Harvard University, Cambridge, MA, USA

1. Introduction

Certain protein families are very highly conserved across the major divisions of life. This is illustrated, for example, by the alignment in Figure 1(a) of Ran GTPases, which demonstrate a remarkable degree of sequence conservation across metazoans, fungi, plants, and protozoans. Ran is a member of the Ras superfamily of small GTPases (Hall, 2000), which function as molecular binary switches involved in cellular signaling pathways. Ran's specific role involves signaling subcellular location with reference to chromosomes and the nucleus (for a recent review, see Quimby and Dasso, 2003).

Inasmuch as proteins such as Ran are highly conserved and appear to be required by all eukaryotes, these constitute the core of the eukaryotic cellular machinery, and thus have more or less ceased to evolve one to two billion years ago. The functions of most of the highly conserved residues in such proteins are unknown, yet clearly important, considering that often even minor modifications, such as side-chain attachment of a methyl or hydroxyl group, are consistently eliminated by natural selection. Indeed, at some residue positions (e.g., 75 and 81 in Figure 1a), even isomeric swapping between leucine and isoleucine is strongly selected against. Such residues are likely to form functionally critical interactions with specific geometric and/or chemical constraints corresponding to subtle and currently poorly understood principles.

Much of our understanding of the principles of protein function comes through biophysical, biochemical, and genetic analyses, which reveal basic properties, such as sequence, structure, catalytic activity and kinetics, and subcellular location. Likewise, conserved patterns in multiple sequence and structural alignments constitute another important source of biological information inasmuch as these reflect selective constraints imposed by the underlying mechanisms during protein evolution.

Indeed, these selective constraints – as quantified through statistical analysis of co-conserved patterns (see below) – may be viewed as basic properties of proteins.

Interpreting the biological significance of conserved patterns is confusing, however, when nearly all residues in a protein family are well conserved, as for Ran. In this case, it is helpful to first decompose conserved residues into functional

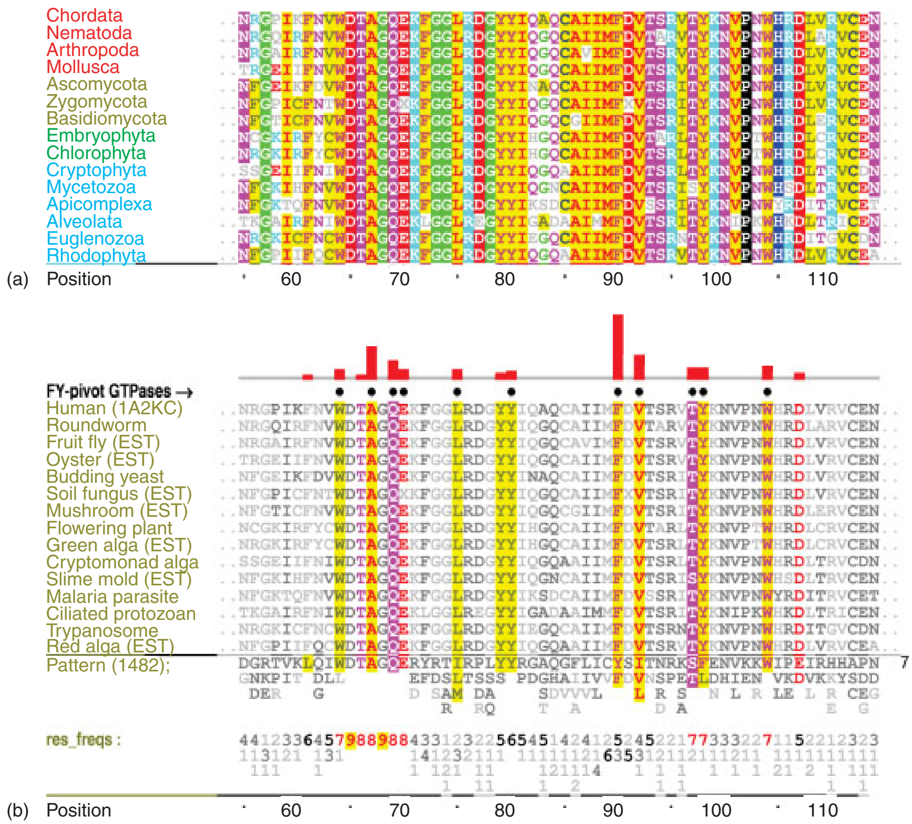


Figure 1 (a) Ran family multiple alignment. The leftmost column specifies each sequence's phylum, which are colored by major eukaryotic taxa as follows: metazoans, red; fungi, dark yellow; plants, green; protozoans, cyan. Only a subregion of the Ran alignment is shown. The top sequence is the query. (b) CHAIN analysis Ran family contrast hierarchical alignment. Ran family aligned residues are displayed directly, while the "foreground" sequences (as defined in the text) consists of FY-pivot GTPases, which are too numerous to display directly and thus are represented below the alignment as position-specific conserved patterns. (The total number of aligned sequences is shown in parentheses.) The corresponding residue frequencies ("res_freq") are given in integer tenths below the conserved patterns. For example, a "5" in integer tenths indicates that the corresponding residue directly above it occurs in 50–60% of the (weighted) sequences. Histogram bar heights are approximately logarithmically proportional to our measure of selective constraints, as defined by a ball-in-urn statistical model at each position in the alignment. This model defines the selective constraint as the difficulty of drawing by chance at least as many of the same or similar balls (residues) from a given position in an alignment of all P-loop GTPases as are observed at that position for FY-pivot GTPases. Dots below the histogram (and directly above the alignment) indicate residue positions specifically assigned to the FY-pivot category. (c) Contrast hierarchical alignment showing co-conserved residues in various FY-Pivot GTPase families of known structure. The leftmost column gives the pdb code and protein name. (Reproduced from Neuwald *et al*. (2003) by permission of Cold Spring Harbor Laboratory Press)

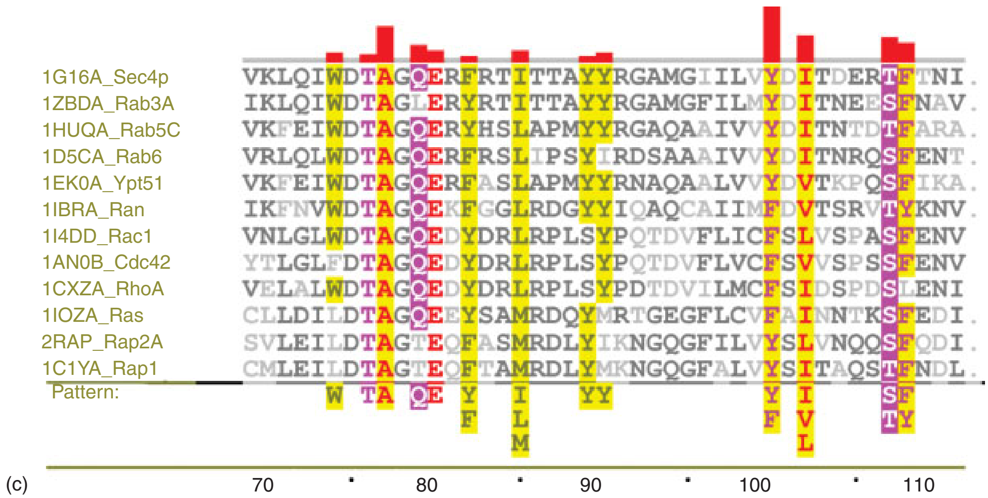


Figure 1 (*continued*)

categories on the basis of their patterns of conservation within a large class of homologous proteins. Identifying residues in Ran involved in nucleotide binding and hydrolysis is straightforward, as these are highly conserved across all P-loop GTPases. The functions of residues with more complex patterns of conservation, however, are less easily ascertained.

2. CHAIN analysis

Contrast hierarchical alignment and interaction network (CHAIN) analysis (Neuwald *et al*., 2003) identifies groups of optimally co-conserved residues and corresponding molecular interactions. Such co-conserved patterns presumably correspond to category-specific functional constraints that are "layered on top of" the weak constraints maintaining that protein's structural fold – which may be conserved in the absence of detectable pairwise sequence similarity (*see* Article 72, **Threading for protein-fold recognition**, Volume 7). CHAIN analysis works best on proteins belonging to a large class consisting of many well-conserved families and subfamilies and for which vast amounts of sequence, structural, and other experimental data are available. In this case, powerful statistical procedures can be used to accurately align the sequences and to detect subtle, yet biologically significant characteristics.

CHAIN analysis was initially applied to Ran and related small GTPases (Neuwald *et al*., 2003). This revealed a group of co-conserved residues (Figure 1b,(1)c) and structural interactions (Figure 2) within Rab, Rho, Ras, and Ran GTPases, which were termed FY-pivot GTPases after a strikingly conserved phenylalanine or tyrosine that appears to serve as a pivot. Because these residues are strikingly *unconserved* in other GTPases (including the Ras-like GTPases Arf and Sar), they presumably perform a function specific to FY-pivot GTPases. Moreover, these

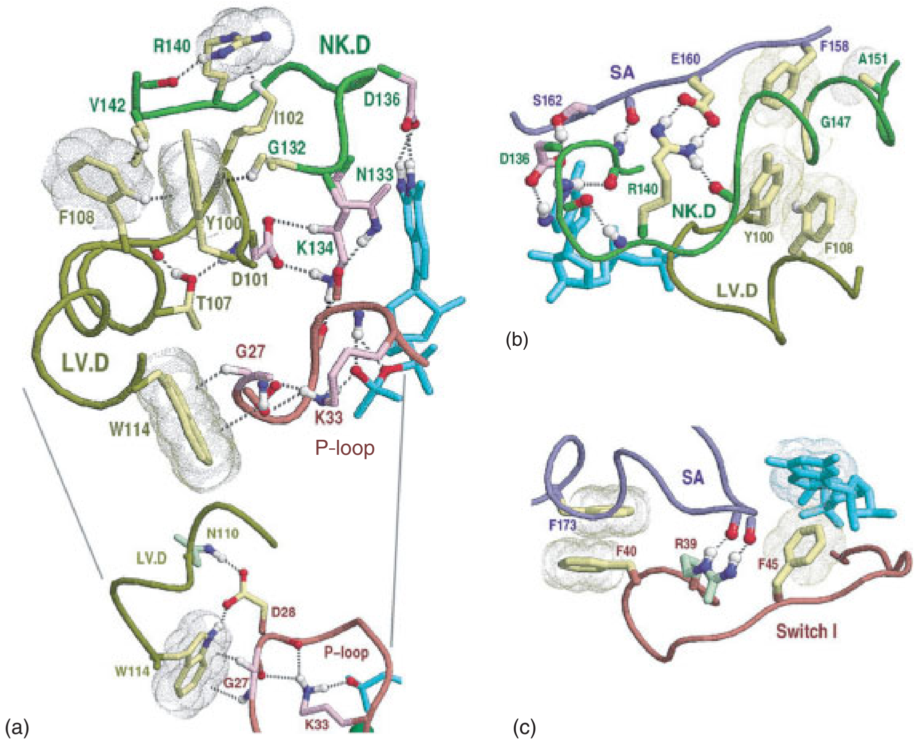


Figure 2 Sec4p as a structural prototype of FY-pivot GTPases. The structure of Sec4p is shown in complex with GDP (pdb code: 1G16) (Stroupe and Brunger, 2000). Hydrogen bonds are depicted as dotted lines, and aromatic–aromatic and van der Waals interactions as dot clouds. Dotted lines into clouds depict CH-π or NH-π hydrogen bonds (Weiss *et al*., 2001). Color scheme: GDP (cyan); main-chain traces and residue designations (colored by regions); residue side chains and canonical glycine main chains (yellow, co-conserved in FY-pivot GTPases; magenta, conserved in all GTPases; green, intermediate); oxygen, nitrogen, and hydrogen atoms establishing hydrogen bonds (red, blue, and white, respectively); hydrogen bonding carbons (colored as their corresponding side chains) (Reproduced from Neuwald *et al*. (2003) by permission of Cold Spring Harbor Laboratory Press). (a) Residues surrounding the P-loop. (b) Residues near the FY-pivot (Y100). (c) Residues in the Switch I region

residues, along with other residues specifically conserved in Ran (not shown), appear to play important roles in Ran nucleotide exchange (Neuwald *et al*., 2003). Rather than focusing on the biological implications of this and similar analyses, however, which are described in detail elsewhere (Kannan and Neuwald, 2004; Neuwald, 2003; Neuwald *et al*., 2003), this article focuses on the methodological aspects of CHAIN analysis.

3. A mathematical description of selective constraints

Although qualitative descriptions can be useful, mathematically rigorous descriptions reveal subtle properties that otherwise would be missed. For this reason, CHAIN analysis defines protein-selective constraints on the basis of classical and

Bayesian statistical models. Conceptually, these formulations are roughly analogous to the notion of free energy in statistical thermodynamics: Just as the energy of a chemical reaction may be defined by the degree to which products have shifted away from reactants, CHAIN analysis defines selective constraints by the degree to which residue positions in a subset of aligned homologous sequences (termed the foreground set) have shifted away from (or contrast with) the residue compositions observed at those positions in the remaining sequences (termed the background set). This is illustrated for Ran in Figure 1(b).

Put another way, just as ΔG defines the difference in free energy between distinct thermodynamic states, our statistical models define differences in selective constraints between two distinct protein "functional states": those corresponding to the foreground and background sets. CHAIN analysis optimally defines the foreground and background sets using a Gibbs sampling (Liu, 2001) procedure called Bayesian partitioning with pattern selection (BPPS) (Neuwald *et al*., 2003). This procedure defines a foreground set such that these sequences share a strikingly conserved pattern that is strikingly unconserved in the remaining, background sequence set (Figure 3).

The input alignment may be defined in various ways depending on the nature of the selective constraints one seeks to characterize. For example, when analyzing an active ATPase, one may choose to remove inactive homologs from the input alignment while retaining active ATPases. (This may be done using certain CHAIN analysis routines, including the BPPS procedure itself.) Indeed, the most useful information is obtained by measuring and comparing selective constraints associated with biologically meaningful foreground and background "states" in this way. Thus, just as a chemist describes a protein's physical properties by measuring changes in free energy between various thermodynamic states, the BPPS procedure describes a protein's evolutionary properties by measuring changes in selective constraints between various functional categories.

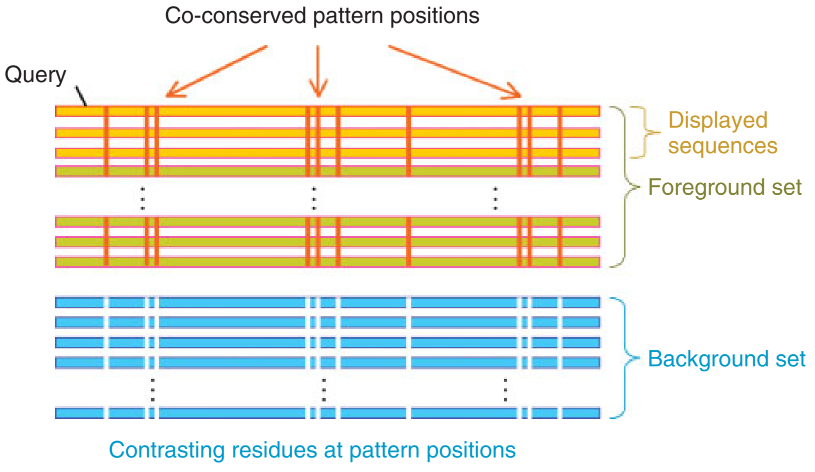


Figure 3 Schematic representation of a "contrast hierarchical alignment" generated by the BPPS procedure. See discussion in text

4. Analogy between CHAIN analysis and protein crystallography

Algorithmically, CHAIN analysis infers information about underlying protein mechanisms from conserved sequence and structural patterns in a manner analogous to inferring protein structural information from crystalline X-ray diffraction patterns (Figure 4). Moreover, as is true of crystallography, CHAIN analysis is empirical inasmuch as it models selective constraints based solely on what may be justified statistically by the input data. It lets the data speak.

The first step (analogous to protein purification) is to apply iterative search procedures – which, in large part, are based on the PSI-BLAST algorithm (Altschul *et al*., 1997; *see also* Article 93, **Detecting protein homology using NCBI tools**, Volume 8) – to gather many, often distantly related protein sequences and corresponding available structures. For enhanced search sensitivity, CHAIN analysis also may rely on profile hidden Markov models (HMMs) (Eddy, 1998; *see also* Article 98, **Hidden Markov models and neural networks**, Volume 8) applied in conjunction with Markov chain Monte Carlo (MCMC) optimization methods (Neuwald and Liu, 2004; Neuwald and Poleksic, 2000).

The second step (analogous to crystal formation) is to multiply align the sequences and structures very precisely in order to identify biologically meaningful co-conserved patterns. CHAIN analysis currently multiply aligns (often thousands

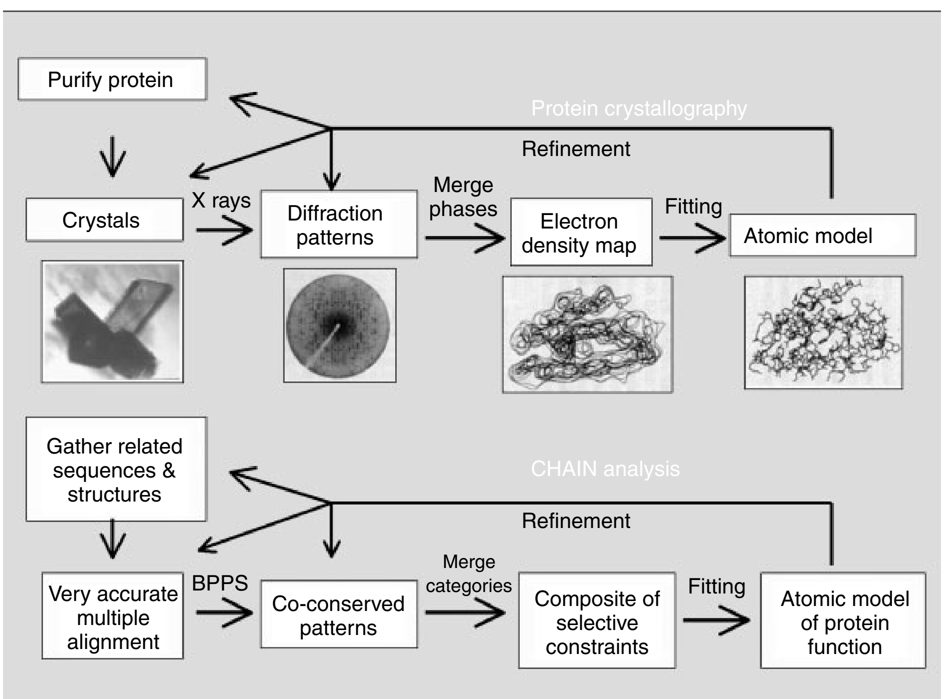


Figure 4 Analogy between CHAIN analysis and protein crystallography. See discussion in text

or tens of thousands of) sequences using a modified version of the PSI-BLAST algorithm that incorporates statistically based HMM and MCMC sampling strategies (Neuwald *et al*., 2003; Neuwald and Liu, 2004; Neuwald and Poleksic, 2000) that are extensions of our earlier Gibbs sampling methods (Lawrence *et al*., 1993; Liu *et al*., 1995; Liu *et al*., 1999; Neuwald *et al*., 1995; Neuwald *et al*., 1997; *see also* Article 106, **Gibbs sampling and bioinformatics**, Volume 8). These statistical procedures help correct misaligned regions, such as those associated with protein domains harboring sizable insertions. Unlike most multiple alignment methods, all of these methods align sequence regions only when there is statistical support for evolutionary conservation – an essential characteristic for accurately estimating selective constraints. CHAIN analysis also relies on programs, such as CE (Guda *et al*., 2004; Shindyalov and Bourne, 1998), for structural alignment (*see* Article 75, **Protein structure comparison**, Volume 7), which – though often viewed as a gold standard – can also introduce errors owing, for example, to rotation of homologous substructures relative to each other. For this reason, "intervention strategies" (Neuwald and Liu, 2004) that can tap into expert knowledge provided by the user, yet still rely on objective statistical criteria to assess alignment quality, are also employed.

The third step (analogous to gathering X-ray diffraction patterns) involves repeated application of the BPPS procedure (Neuwald *et al*., 2003) (see above) to optimally define multiple categories of co-conserved patterns. For each of these categories and at each pattern position, a classical ball-in-urn statistical model is used to quantify selective constraints relative to the background set (Figure 1b, 1c).

The fourth step (analogous to constructing an electron density map) is to merge various categories of selective constraints into a composite picture of the entire protein class. Within a specific protein family, this often reveals a hierarchy of increasingly specialized co-conserved functional "components" that presumably arose through a process of evolutionary divergence and specialization. Such components appear to be layered, so to speak, one on top of another and to function together in a coordinated manner. Comparing and contrasting the hierarchical characteristics of various members of a class provides a global perspective of the selective constraints acting on each protein and on the class as a whole. It is also helpful to analyze in the same way those proteins that functionally interact with the proteins of interest.

The fifth step is to "fit" the various categories of selective constraints to other available experimental data in order to identify likely molecular mechanisms associated with the protein class. Conceptually, the fitting is done probabilistically rather than deterministically, as is illustrated by the Venn diagram in Figure 5. Each point within the Venn diagram represents a possible mechanistic model, while each circle represents a set of high-probability models given the indicated data source. For example, the currently available biochemical data weighs in favor of some models, disfavors others, and totally excludes others still. In principle, by combining these sources of data (using bioinformatics tools) with inferred evolutionary constraints, one may identify those models that are most compatible with the available data (i.e., the area labeled "high-probability models" in Figure 5). Ideally, we would like to represent all of these data sources using rigorous Bayesian statistics, just as we have done for estimating selective constraints within sequences. Though this

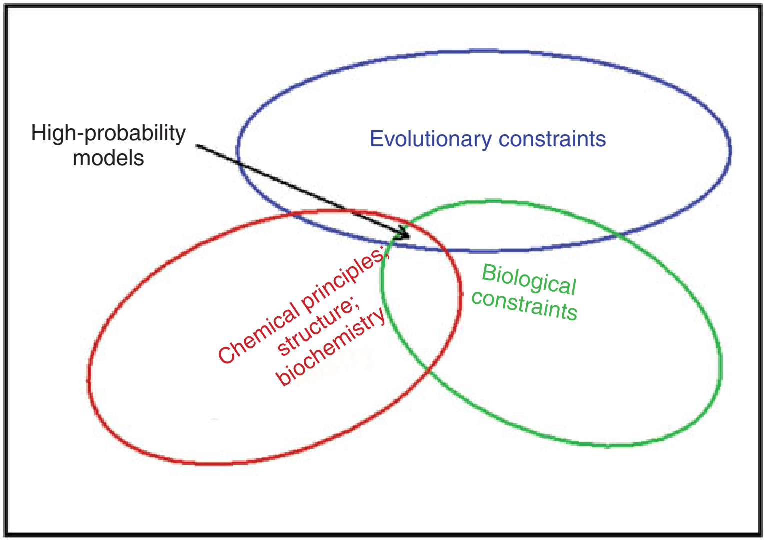


Figure 5 Venn diagram illustrating how Bayesian analysis of evolutionary constraints, in conjunction with other data sources, helps identify likely models of protein function. See discussion in text

currently is unfeasible for most forms of data, we are developing such an approach for structural data.

Finally, iteratively recycling through this entire process helps refine likely models of protein function. After an initial analysis, earlier steps may be improved, for example, by adding previous false-negative sequences to the alignment, by correcting alignment errors on the basis of detailed structural comparisons, or by defining more appropriate input sequence sets for the BPPS procedure.

5. Related methods

Other computational approaches – such as hierarchical analysis of residue conservation (Livingstone and Barton, 1993), principal component analysis (Casari *et al*., 1995), evolutionary trace (Lichtarge *et al*., 1996), positional entropy (Hannenhalli and Russell, 2000), site-specific rate shifts (Gaucher *et al*., 2002), and specificity-determining residues (Li *et al*., 2003; Mirny and Gelfand, 2002) – likewise obtain evolutionary insights into protein function based on multiply aligned sequences. CHAIN analysis complements most of these methods in that it focuses on co-conserved residues that stopped evolving long ago rather than on evolutionary changes within the context of a phylogenetic tree. It also differs in that it combines Bayesian and classical statistical inference, MCMC procedures, and molecular interaction analyses to detect – on the basis of thousands of distantly related sequences – subtle structural characteristics in atomic detail.

6. Accessibility over the Web

Because automation of these procedures is nontrivial, making CHAIN analysis programs directly available is not feasible at this time. Instead, a Web database of curated results from various analyses is being developed in collaboration with Lincoln Stein at Cold Spring Harbor Laboratory. This will facilitate experimental follow-up on CHAIN analysis predictions, primarily through mutational studies. While some mutational analyses are straightforward, we predict that other compelling hypotheses likely will require very clever experimental designs, perhaps involving unnatural amino acid substitutions (Hendrickson *et al*., 2004) or the use of drug-like molecular probes, to unambiguously confirm predicted functional roles for individual atoms.

Acknowledgments

We thank Jack Chen, Andy Reiner, and Ji-Joon Song for critical reading of the manuscript and helpful suggestions. This work was supported by NIH grant LM06747 to AFN and by NSF grants DMS0204674 and DMS0244638 to JSL.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Casari G, Sander C and Valencia A (1995) A method to predict functional residues in proteins. *Nature structural biology*, **2**, 171–178.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Gaucher EA, Gu X, Miyamoto MM and Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends in Biochemical Sciences*, **27**, 315–321.

Guda C, Lu S, Scheeff ED, Bourne PE and Shindyalov IN (2004) CE-MC: a multiple protein structure alignment server. *Nucleic Acids Research*, **32**, W100–W103.

Hall A (Eds.) (2000) *GTPases*, Oxford University Press.

Hannenhalli SS and Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology*, **303**, 61–76.

Hendrickson TL, Crecy-Lagard VdV and Schimmel P (2004) Incorporation of nonnatural amino acids into proteins. *Annual Review of Biochemistry*, **73**, 147–176.

Kannan N and Neuwald AF (2004) Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2α. *Protein Science*, **13**, 2059–2077.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Li L, Shakhnovich EI and Mirny LA (2003) Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4463–4468. Epub 2003 April, 4464.

Lichtarge O, Bourne HR and Cohen FE (1996) Evolutionarily conserved $G_{\alpha\beta\gamma}$ binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 7507–7511.

Liu JS (2001) *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag: New York.

Liu JS, Neuwald AF and Lawrence CE (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal. American Statistical Association*, **90**, 1156–1170.

Liu JS, Neuwald AF and Lawrence CE (1999) Markovian structures in biological sequence alignments. *Journal. American Statistical Association*, **94**, 1–15.

Livingstone CD and Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer applications in the biosciences: CABIOS*, **9**, 745–756.

Mirny LA and Gelfand MS (2002) Using orthologous and paralogous proteins to identify specificity determining residues. *Genome biology*, **3**, PREPRINT0002.

Neuwald AF (2003) Evolutionary clues to DNA polymerase III β clamp structural mechanisms. *Nucleic Acids Research*, **31**, 4503–4516.

Neuwald AF, Kannan N, Poleksic A, Hata N and Liu JS (2003) Ran's C-terminal, basic patch and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras and Ran GTPases. *Genome Research*, **13**, 673–692.

Neuwald AF and Liu JS (2004) Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics*, **5**(1), 157.

Neuwald AF, Liu JS and Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, **4**, 1618–1632.

Neuwald AF, Liu JS, Lipman DJ and Lawrence CE (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Research*, **25**, 1665–1677.

Neuwald AF and Poleksic A (2000) PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of β-propellers in UV-damaged DNA-binding protein. *Nucleic Acids Research*, **28**, 3570–3580.

Quimby BB and Dasso M (2003) The small GTPase Ran: interpreting the signs. *Current Opinion in Cell Biology*, **15**, 338–344.

Shindyalov IN and Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, **11**, 739–747.

Stroupe C and Brunger AT (2000) Crystal structures of a Rab protein in its inactive and active conformations. *Journal of Molecular Biology*, **304**, 585–598.

Weiss MS, Brandl M, Suhnel J, Pal D and Hilgenfeld R (2001) More hydrogen bonds for the (structural) biologist. *Trends in Biochemical Sciences*, **26**, 521–523.

Short Specialist Review

Large-scale, classification-driven, rule-based functional annotation of proteins

Darren A. Natale, C. R. Vinayaka and Cathy H. Wu
Georgetown University Medical Center, Washington DC, USA

1. Introduction

The high-throughput genome projects have resulted in a rapid accumulation of predicted protein sequences for a large number of organisms. Meanwhile, scientists have begun to tackle protein functions and other complex regulatory processes using global-scale data generated at various levels of biological organization, ranging from genomes and proteomes to metabolomes (metabolites synthesized by a biological system) and physiomes (the physiological dynamics of whole organisms). To fully realize the value of the data, scientists need to understand how proteins function in making up a living cell. With experimentally verified information on protein function lagging behind, bioinformatics methods are needed for reliable and large-scale functional annotation of proteins.

A general approach for functional characterization of unknown proteins is to infer function based on sequence similarity to annotated proteins in sequence databases. While this is a powerful method, numerous genome annotation errors have been detected, many of which have been propagated throughout molecular databases. There are several sources of errors (Galperin and Koonin, 1998; Koonin and Galperin, 2003). Errors often occur when identification is made on the basis of local domain similarity or similarity involving only parts of the query and target molecules. Furthermore, the similarity may be to a known domain that is tangential to the main function of the protein or to a region with compositional similarity, such as transmembrane domains. The best hit may be statistically insignificant. Errors also occur when the best hit entry is an uncharacterized or poorly annotated protein, or is itself incorrectly predicted, or simply has a different function. Aside from erroneous annotation, database entries may be underidentified, such as a "hypothetical protein" with a convincing similarity to a protein or domain of known function, or may be overidentified, such as ascribing a specific enzyme activity when a less specific one would be more appropriate.

These problems can be addressed by using a curated, hierarchical, whole-protein classification database, especially in conjunction with a rule-based system

designed specifically for large-scale annotation. Annotation rules define not only the conditions that must be met but also the fields (including function, ontology, and site-specific features) that could be confidently propagated.

2. Hierarchical whole-protein classification

2.1. The advantage of protein classification

There is an immediate advantage to using a protein classification database as a basis for annotation. This advantage is conferred by “strength in numbers”. Instead of relying on the (hopefully) accurate annotation of a single (hopefully related) protein (usually, the BLAST best hit), using curated classification databases allows reliance on the collected wisdom of multiple proteins, or at least the assurance that the members are truly related. (For the sake of this discussion, it is assumed that the query protein has been accurately assigned to a protein family of interest. This, typically, is done via specialized BLAST search, (e.g., the COG database, Tatusov *et al*., 2003), HMMer search (e.g., Pfam, Bateman *et al*., 2004), or combination of both (e.g., PIRSF, Wu *et al*., 2004). *Streptococcus agalactiae* protein gbs1797 (UniProt accession Q8E3G2) provides a good example of the usefulness of protein families for annotation. It is likely that the simple “Hypothetical protein” annotation attached to this protein is derived from the fact that its best hit has the same annotation. However, long before this protein was submitted to the sequence databases, the COG database had its predicted family annotated as “Galactose-1-phosphate uridylyltransferase” (COG1085). Similarly, *Mannheimia succiniciproducens* “Hypothetical protein MS0372” (UniProt accession Q65VN1) is predicted to be a member of PIRSF family PIRSF000378 (“Glycyl radical cofactor protein YfiD”), PFAM family PF01228 (“Glycine radical”), and InterPro family IPR001150 (“Formate C-acetyltransferase glycine radical”), all of which were available at the time of submission. Thus, several of the problems associated with propagating meaningful annotation to new proteins, including misannotated or statistically insignificant best hits, are averted by the use of protein classification databases.

2.2. Whole proteins

Are whole proteins equal to the sum of their parts? Mostly, yes. However, this is not always the case, at least for annotation purposes. For example, PIRSF000142 contains “glycerol-3-phosphate dehydrogenase (anaerobic), subunit C” proteins. The annotation of such proteins, if one were to rely on predominantly domain-centric databases alone, would likely be “Cysteine-rich iron-sulfur binding protein” (based on hits to PF02754, “Cysteine-rich domain” and PF00037, “4Fe-4S binding domain”). Worse yet are the possibilities that a protein is composed of multiple domains, but only one is described (thereby causing an underannotation by means

of omission), or the converse, where a protein is composed of only one domain but hits proteins of much longer length and likely different function (*see* Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6). The use of a whole-protein classification database, such as PIRSF, combined with an insistence that predicted members of a given family exhibit (near) end-to-end similarity, obviates such problems. Other predominantly whole-protein classification databases (which may or may not also contain domain families) include COGs, TIGRFAMs (Haft *et al*., 2003), PANTHER (Mi *et al*., 2005), and HAMAP (Gattiker *et al*., 2003).

2.3. Hierarchies

The annotation power of protein classification databases is rendered even more powerful if a single database contains families with progressively greater levels of similarity (i.e., hierarchies), or if different databases (with different levels) are searched. Such databases, most notably PIRSF and the database-integrating InterPro (Mulder *et al*., 2005), allow annotation at an appropriate level of specificity. Theoretically, one query protein could be confidently predicted to be a member of a parent family, but not a child family, while a different query might be confidently assigned to both levels. In such cases, the most-specific possible annotation could be propagated. For example, PIRSF001370, a large family that contains acetolactate synthase-like thiamine diphosphate-dependent enzymes, can be further divided into seven subfamilies. Two *Pseudomonas putida* members of this family (UniProt accessions Q88DY8 and Q88 N22) are annotated as the large subunit of acetolactate synthase (the latter as putative). However, only Q88DY8 belongs to subfamily PIRSF500108, which contains verified acetolactate synthase large subunits. Q88 N22 can only be confidently assigned to the parent family. Pending confirmation of activity, "thiamine diphosphate-dependent enzyme" would be a safer annotation. As these examples demonstrate, the use of a hierarchical database helps prevent over- or underannotation (*see* Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6).

2.4. The PIRSF system

The PIRSF protein classification system combines all of the approaches described above (Wu *et al*., 2004). The system provides protein classification from superfamily to subfamily levels in a network structure based on evolutionary relationships of whole proteins. Such classification allows identification of probable function for uncharacterized sequences. PIRSF classification, which considers both full-length similarity and domain architecture, discriminates between single- and multidomain proteins where functional differences are associated with the presence or absence of one or more domains. Furthermore, classification based on whole proteins, rather than on the component domains, allows annotation of both generic biochemical and specific biological functions.

In this system, unassigned proteins are tested against a set of hidden Markov models (HMMs) representing curated full-length protein families at different hierarchical levels. The following criteria, used by PIRSF Scan (**http://pir.georgetown.edu/pirsf**), must be met:

1. The query protein must be recognized by an HMM with a score that falls above the minimum threshold score appropriate for that family. The threshold is based on the minimum score exhibited by current curated members of the family.
2. The length of the query protein must fall within the range exhibited by true members of the family. Both the score and length thresholds are adjusted on the basis of characteristics of the protein family – that is, by taking into account standard deviations – to enable the scan program to capture divergent members for further testing.
3. The set of BLAST hits for the query must be assigned predominantly to the same PIRSF as predicted by HMM. Here, the "majority rule" principle is used, meaning that the query must hit at least 10 or one-third of the members of the PIRSF.

Fine-tuning of the match parameters is an ongoing process. Exact parameters used are described in the help document linked at the URL given above.

3. Rule-based automated annotation

Curated, hierarchical, whole-protein classification databases are, by design, well suited to large-scale protein annotation. First, the classification of whole proteins – not parts – enables specific biological function to be accurately propagated rather than only generic biochemical function (or worse, inaccurate biological function). Second, a hierarchical system of classification allows the propagation of generic biochemical function when a protein cannot confidently be assigned to a subfamily with known specific biological function. However, it is still possible to further refine large-scale automatic annotation systems. Such refinement is afforded by using annotation rules.

3.1. Annotation rules defined

An annotation rule is a set of condition/action statements. The conditions can range from the sequence-based, such as "member of family X", "contains domain Y", or "motif Z present", to the organism-based, such as "member of taxonomic lineage A" or "encodes metabolic function B". The action, typically, is the propagation of appropriate information to the query protein, but it can also include the flagging of preexisting annotation as inappropriate. Rules are most useful (and accurate) when associated with a particular family or set of families. For example, a rule written to propagate the name "Pyruvate (flavodoxin) dehydrogenase" to proteins that match its active site motif could end up wrongly propagating that term to Midasin or other

inappropriate proteins. However, if this same rule were applied only to members of the appropriate family (e.g., PIRSF000159), then only suitable proteins would be so annotated, assuming they match the motif.

Annotation rules have thus far been developed by the three members of the UniProt Consortium – the Protein Information Resource (PIR), the Swiss Institute of Bioinformatics (SIB), and the European Bioinformatics Institute (EBI). PIR has developed manually curated Site Rules (PIRSRs) and Name Rules (PIRNRs), each of which (described in detail below) is based on curated protein families of the PIRSF system. SIB has developed curated rules based on the curated HAMAP family system (Gattiker *et al*., 2003). EBI has multiple systems – the manually curated RuleBase system (Biswas *et al*., 2002) and the automatically generated Spearmint (Kretschmann *et al*., 2001) and Xanthippe (Wieser *et al*., 2004) rules. Each of these rule sets are integrated into the annotation pipeline for proteins in the UniProt Knowledgebase (Bairoch *et al*., 2005).

3.2. Advantages of rule-based propagation of functional annotation

Annotation rules add significant advantages when used in conjunction with protein classification systems for the automated propagation of information from a family to an individual protein:

- *Increased specificity*: For example, both archaea and eukarya contain homologs to the DNA polymerase sliding clamp (PIRSF002090), commonly called Proliferating Cell Nuclear Antigen (PCNA). However, though the designation PCNA is perfectly reasonable for eukarya, it is not for archaea. Furthermore, the archaeal and the eukaryal versions are not easily separable based on sequence. Therefore, the only recourse for proper naming would be a taxonomy-based rule. Another example is provided by PIRSF000532, a protein family composed of members with the same general function but different specificities. The proteins are phosphofructokinases that are either ATP dependent, pyrophosphate-dependent, or of unknown dependency. The specificity is encoded in a very small number of residues (Bapteste *et al*., 2003). Multiple instances of convergent evolution render the division of this family into subfamilies based on whole-protein similarity to be difficult if not impossible. However, rules can test for the known amino acid combinations conferring ATP or pyrophosphate dependence, thereby allowing the propagation of "ATP-dependent phosphofructokinase" or "Pyrophosphate-dependent phosphofructokinase" as appropriate. In addition, a fallback rule can be formulated such that entries failing both of the specific rules could be named the less-specific, but still accurate, "Phosphofructokinase".
- *Maintenance*: This attribute works on two levels. First, maintaining a single rule for multiple proteins is easier than maintaining those proteins individually. Second, at least as applied in UniProt, the annotation of proteins that fit a particular rule can be periodically updated to reflect changes in the rule actions. The reapplication of rule actions means that entries previously annotated on the basis of a particular rule would immediately be fitted with the new information from the updated rule.

- *More annotation fields*: The ease of maintenance fosters an increase in the number of functional annotation fields that can be "touched" by automated means. Most current methods for applying automatic annotation to individual proteins focuses specifically on the protein name (even though, in principle, other fields can be annotated as well; in practice, this is not done). However, rules afford an easy mechanism for more accurate propagation of other important annotation fields, such as position-specific sequence features, Enzyme Commission (EC) name and number, keywords, references, and Gene Ontology (GO) terms. Function-based classifications, such as EC and GO, provide useful complements to sequence-based classifications enabling, for example, studies on analogous enzymes (Galperin *et al*., 1998).
- *Standardization*: The uniform application of a rule to proteins in a given family, by definition, will create a uniformity in annotation (when warranted). When applied to protein names, such annotation becomes another controlled vocabulary, thereby significantly aiding text-based searches.
- *Evidence attribution*: Rules can themselves be annotated with information that describes the source for the rule and whether the propagatible information is based on experimental evidence or computational prediction. Storing rules with unique identifiers allows information propagated by a rule to be tagged with the rule identifier. This, in effect, forms the basis for evidence attribution describing data source, types of evidence, and methods for annotation. Such evidence attribution provides an effective means to avoid misinterpretation of annotation information and propagation of annotation errors.
- *Validation*: A large amount of inaccurate information can be (and has been) propagated to proteins using nonstringent criteria. These erroneous data are difficult to find, so nearly impossible to eradicate completely. However, annotation rules – used to propagate reliable information – can also be used to flag unreliable information through "caution" statements. Each of the annotation rule systems indicated above has this component; however, the Xanthippe system performs this function exclusively.

3.3. Annotation rules at PIR

The PIRSF classification serves as the basis for a rule-based approach that provides standardized and rich automatic functional annotation. In particular, annotation can be reliably propagated from sequences containing experimentally determined properties to closely related homologous sequences based on curated PIRSF families. PIR rules are manually defined and curated for several annotation fields, as described below. Two types of annotation rules have been developed. PIR Site Rules focus on sequence-specific features (UniProt "FT" lines). Originally designed to capture and propagate protein name (UniProt "DE" line) information, PIR Name Rules have expanded in scope to include synonyms and acronyms, Enzyme Commission (EC) name and number, Gene Ontology terms, and comment fields such as Function, Pathway, and Caution.

3.4. PIRSR Site Rules

To assure correct functional assignments, protein identifications must be based on both global (whole protein) and local (domain and motif) sequence similarities. Position-specific Site Rules enable the annotation of active site residues, binding site residues, modified residues, or other functionally important amino acid residues. To exploit known structure information, Site Rules are defined starting with PIRSF families that contain at least one known 3D structure with experimentally verified site information. The active site information on proteins is taken from the PDB (Bourne *et al*., 2004) SITE records, the LIGPLOT of interactions available in PDBSum database (Laskowski, 2001), and published scientific literature. The dataset of catalytic residues (Bartlett *et al*., 2002) is used as an authoritative source for catalytic residues in enzyme active sites.

Site rule curation involves manually editing multiple sequence alignments of representative protein family members (including the template PDB entry), building hidden Markov models (HMMs) from the conserved regions containing the functional site residues, and visualizing sequences and structures. The rules are defined using appropriate syntax and controlled vocabulary for site description and evidence attribution. The profile HMM thus built allows one to map functionally important residues from the template structure to other members of the PIRSF family that do not have a solved structure. To avoid false positives, site features are only propagated automatically if all site residues match perfectly in the conserved region by aligning both the template and target sequences to the profile HMM using HmmAlign (Eddy *et al*., 1995). Potential functional sites missing one or more residues or containing conservative substitions are only annotated after expert review with evidence attribution. Table 1 shows six example Site Rules, three for PIRSF000077 and three for PIRSF000097.

Automatic annotation generated by the PIR Site Rules will be attributed with the rule ID, which links to the rule report containing additional information. For Site Rules, the rule report will include multiple sequence alignments with highlighted site residues (Figure 1).

3.5. PIRNR Name Rules

To capture the full annotation capability of PIRSFs, PIR Name Rules (PIRNRs) were developed. As with Site Rules, each Name Rule is defined with conditions for propagation (Table 1). While most Name Rules assign protein names to all family members, many families require more specialized rules with additional conditions to propagate appropriate names and avoid overidentification or underidentification. In addition to the PCNA example of taxonomically restricted names (or activities), PIR Name Rules provide the means to account for functional variations within one PIRSF, including instances in which a protein lacks the active site residue(s) necessary for enzymatic activity, and cases in which evolutionarily related proteins

Table 1 PIR Site Rules and Name Rules for automated annotation of functional sites and standardized protein names

Rule type/ID	Condition	Propagatible annotation	Evidence
Site rule/ PIRSR000077-1	PIRSF000077 member and HMM site match	Feature: Active site	Template: UniProt:P00274; PDB:2TRX
		Residues: Cys 33, Cys 36	Status: Validated [PMID:2181145]
Site rule/ PIRSR000077-2	PIRSF000077 member and HMM site match	Feature: Site	Template: UniProt:P00274; PDB:2TRX
		Residues: Gly 34, Pro 35	Status: Validated [PMID:9099998]
Site rule/ PIRSR000077-3	PIRSF000077 member and HMM site match	Feature: Site	Template: UniProt:P00274; PDB:2TRX
		Residue: Asp 27	Status: Validated [PMID:9374473]
Site rule/ PIRSR000097-1	PIRSF000097 member and HMM site match	Feature: Active site	Template: UniProt:P15121; PDB:1US0
		Residue: Tyr 49	Status: Validated [PMID:8245005]
Site rule/ PIRSR000097-2	PIRSF000097 member and HMM site match	Feature: Binding site	Template: UniProt:P15121; PDB:1US0
		Residue: His 111	Status: Validated [PMID:8245005]
Site rule/ PIRSR000097-3	PIRSF000097 member and HMM site match	Feature: Site	Template: UniProt:P15121; PDB:1US0
		Residue: Lys 78	Status: Validated [PMID:15146478]
Name Rule/ PIRNR001555-1	PIRSF001555 member	Name: aspartate – ammonia ligase	Template: UniProt:P00963
		Synonym: asparagine synthetase A	Status: Validated [PMID:1369484]
		EC: aspartate – ammonia ligase (EC 6.3.1.1)	
Name Rule/ PIRNR000881-1	PIRSF000881 member and vertebrates	Name: *S*-acyl fatty acid synthase thioesterase	Template: UniProt:P00633
		EC: oleoyl-[acyl-carrier-protein] hydrolase (EC 3.1.2.14)	Status: Validated [PMID:2415525]
Name Rule/ PIRNR000881-2	PIRSF000881 member and not vertebrates	Name: Type II thioesterase	Template: UniProt:Q08788
		EC: thiolester hydrolases (EC 3.1.2)	Status: Validated [PMID:9560421]
Name Rule/ PIRNR025624-1	PIRSF025624 member	Name: ACT domain protein	Status: Predicted
		Misnomer: chorismate mutase	

have known differences in biochemical activities (such as the phosphofructokinases of PIRSF000532) or domain organization.

PIRNRs are part of a two-tier system. The first tier is the "zero-level" rule. Such rules are designed without the ability for discrimination. That is, the information propagated by such a rule will apply to all members of the PIRSF to which the rule belongs (each PIRNR will apply only to members of its cognate PIRSF, and

```
             1        10        20        30  *   *  40        50        60
P00274   MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQG-KLTVAKLN: 60
P57653   -MNKIIELTDQNFEEQVLNSKSFFLVDFWAQWCNPCKILAPILEEISKEYSN-KVIVGKLN: 59
P51225   --MSVSQVTDASFKQEVINNDLPVLVDFWAPWCGPCRMVSPVVDAIAEEYES-SIKVVKIN: 58
P09857   -SDSIVHVTDDSFEEEVXKSPDPVLVDYWADWCGPCKMXAPVXDEIADEYAG-RVKXAKXN: 59
Q9ZEH4   --MAIVKVTDADFDSKVESG--VQLVDFWATWCGPCKMIAPVLEELAADYEG-KADILKLD: 56
Q9CF37   ---MEYNITDATFDEETKEG--LVLIDFWATWCGPCRMQAPILEQLSEELDESELKICKMD: 56
P14949   --MAIVKATDQSFSAETSEG--VVLADFWAPWCGPCKMIAPVLEELDQEMGD-KLKIVKID: 56
Q97EM7   ---MVKEINESIFDEEIKTSGEPVIVDFWAPWCGPCKMLGPIIDELSEDLDG-KAKFTKVN: 57
Q97IU3   ---MIQEINDKSFVNVISNSKKVVVVDFWATWCEPCKMIAPILEEIAGELKE-SLEIIKIN: 57
P56430   -MSHYIELTEENFESTIKKG--VALVDFWAPWCGPCKMLSPVIDELASEYEG-KAKICKVN: 57
Q9X2T1   MSEHIVNVTDASFEQDVLKADGPVLVDYWAEWCGPCKMIAPVLDEVARDYQG-KLKVCKLN: 60
Q9K8A8   --MAIVNVTDQTFAQETSEG--LVLADFWAPWCGPCKMIAPVLEELDGEMGD-KVKIAKLD: 56
Q9KV51   MSDKILQLTDDGFENDVIKAAGPVLVDFWAEWCGPCKMIAPILDEVADEYAG-KLTIGKLN: 60
P43785   -MSEVLHINDADFESVVVNSDIPILLDFWAPWCGPCKMIAPVLDELAPEFAG-KVKIVKMN: 59
Q8ZAD9   MSDKIIHLSDDSFDTDVLKASGLVLVDFWAEWCGPCKMIAPILDEIAEEYEG-RLTIAKLN: 60
Q9S386   ---MVKEITDATFEQETSEG--LVLTDFWATWCGPCRMVAPVLEEIQEERGE-ALKIVKMD: 55
Q7M0Y9   ---MVKDINDSNFQEEVKAG--TVVVDFWAAWCGPCKMLGPVIDEDLSEEQG-KLDIAKVN: 55
```

Figure 1 Multiple sequence alignment of PIRSF family thioredoxin (PIRSF000077), the two active site cysteines (site rule PIRSR000077-1) are boxed in red. Thioredoxins are small redox proteins that catalyze dithiol-disulfide exchange reactions. The source organisms, with the UniProt accession numbers in parentheses, are *Escherichia coli* (P00274), *Buchnera aphidicola* (P57653), *Porphyra purpurea* (P51225), *Chromatium vinosum* (P09857), *Staphylococcus aureus* (Q9ZEH4), *Lactococcus lactis* (Q9CF37), *Bacillus subtilis* (P14949), *Clostridium acetobutylicum* (Q97EM7), *Clostridium acetobutylicum* (Q97IU3), *Helicobacter pylori* (P56430), *Pseudomonas aeruginosa* (Q9X2T1), *Bacillus halodurans* (Q9K8A8), *Vibrio cholerae* (Q9KV51), *Haemophilus influenzae* (P43785), *Yersinia pestis* (Q8ZAD9), *Listeria monocytogenes* (Q9S386), and *Clostridium pasteurianum* (Q7M0Y9)

to no other PIRSF members). This ensures that every classified protein will get some annotation (assuming any information is known for the PIRSF members, and such information is deemed applicable to all members). The "higher-order" rules will fit only those members that pass additional tests. The additional tests may include taxonomic placement, presence or absence of critical domains, fit to a PIR Site Rule, or presence on manually curated inclusion (false negative) or exclusion (false-positive) lists. By definition, higher-order rules are more specific than zero-level rules. Therefore, the propagation system is designed to prevent propagation based on the zero rule when a higher-order rule applies.

Creating a PIRNR, which seeks to capture known or predicted information about a PIRSF, follows the completed curation of each PIRSF:

1. Compile known (literature) or predicted (sequence analysis) information about members of the PIRSF of interest.
2. Indicate match conditions (higher-order rules only), including
   (a) taxonomic range to which the rule should apply;
   (b) essential domains (in the form of matches to PFAM);
   (c) essential residues (in the form of matches to PIR Site Rules);
   (d) false-negatives (list known entries that should get the indicated annotation, but would otherwise fail the rule tests).
3. Indicate exclusion conditions (higher-order rules only), including
   (a) taxonomic range that should NOT get propagation;

   (b) domains that should NOT be present;
   (c) residues that should NOT be present;
   (d) false-positives (list known entries that pass all the rule tests, but should NOT get propagated information).
4. Indicate propagatible information, including
   (a) protein name (DE line) and acronym that adheres to UniProt standards;
   (b) alternative names or synonyms used in the literature, with acronyms;
   (c) comments on function, pathway, or cautions about nomenclature (as in the PIRNR025624-1 example; Table 1);
   (d) EC number;
   (e) GO terms. Only the leaf-most term is propagated.
5. Outline the scope of the rule, or provide reasons why a particular protein name was chosen as standard for the group.

4. Conclusions

Critical to our understanding of biology is accurate and up-to-date information. The process of evolution affords us the ability to make inferences about the nature of the proteins that govern biological processes, since like proteins often perform like (if not exact) functions. Unfortunately, this same process has been far from a smooth transition from state to state. The result is that inferences made about one protein based on similarity to another protein using automated methods are often suspect. This is more than a mere annoyance. The lack of rigorous methods for propagating appropriate information hampers knowledge discovery by either reducing the associations that can be made, or by producing associations that should not be made. However, the recent development of methods for better annotation hold much promise for preventing – and even reversing – the previous trend toward rampant misinformation. The combination of hierarchical, whole-protein classifications and rule-based large-scale annotation pipelines is a significant step in the right direction.

Related article

Article 87, **The PIR SuperFamily (PIRSF) classification system**, Volume 6

Acknowledgments

The project is supported by grant U01-HG02712 from National Institutes of Health and grant DBI-0138188 from National Science Foundation.

Further reading

Wu CH, Yeh L-S, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, *et al*. (2003a) The protein information resource. *Nucleic Acids Research*, **31**, 345–347.

Wu CH, Huang H, Yeh LS and Barker WC (2003b) Protein family classification and functional annotation. *Computational Biology and Chemistry*, **27**, 37–47.

References

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2005) The universal protein resource (UniProt). *Nucleic Acids Research*, **33**, D154–D159.

Bapteste E, Moreira D and Philippe H (2003) Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. *Gene*, **318**, 185–191.

Bartlett GJ, Porter CT, Borkakoti N and Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, **324**, 105–121.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Biswas M, O'Rourke JF, Camon E, Fraser G, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, *et al*. (2002) Applications of Interpro in protein annotation and genome analysis. *Briefings in Bioinformatics*, **3**, 285–295.

Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, *et al*. (2004) The distribution and query systems of the RCSB protein data bank. *Nucleic Acids Research*, **32**, D223–D225.

Eddy SR, Mitchison G and Durbin R (1995) Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology*, **2**, 9–23.

Galperin MY and Koonin EV (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology*, **1**, 55–67.

Galperin MY, Walker DR and Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, **8**, 779–790.

Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, *et al*. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Computational Biology and Chemistry*, **27**, 49–58.

Haft DH, Selengut JD and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373.

Koonin EV and Galperin MY (2003) *Sequence – Evolution – Function: Computational Approaches in Comparative Genomics*, Kluwer Academic Publishers: Boston.

Kretschmann E, Fleischmann W and Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.

Laskowski RA (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, **29**, 221–222.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284–D288.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Research*, **33**, D201–D205.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Wieser D, Kretschmann E and Apweiler R (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20**(Suppl 1), I342–I347.

Wu CH, Nikolskaya A, Huang H, Yeh L-S, Natale D, Vinayaka CR, Hu Z, Mazumder R, Kumar S, Kourtesis P, *et al*. (2004) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Research*, **32**, D112–D114.

Signal peptides and protein localization prediction

Henrik Nielsen
Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

1. Introduction

In 1999, the Nobel prize in Physiology or Medicine was awarded to Günther Blobel "for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell". Since the subcellular localization of a protein is an important clue to its function, the characterization and prediction of these intrinsic signals – the "zip codes" of proteins – has become a major task in bioinformatics.

Here, I will review the most important methods for the prediction of subcellular localization, also known as protein sorting. Owing to the limited space, this review is far from complete; especially, applications that are not publicly available on-line are ignored. Generally, there are two approaches to protein localization prediction: signal detection, that is, prediction of the sorting signals themselves, and prediction based on global properties (amino acid composition and/or physicochemical variables) that are characteristic of different subcellular compartments.

2. Secretory signal peptides

The best known "zip code" is the secretory signal peptide, which targets a protein for translocation across the plasma membrane in prokaryotes and across the endoplasmic reticulum (ER) membrane in eukaryotes (von Heijne, 1990). It is an N-terminal peptide, typically 15–30 amino acids long, which is cleaved off during translocation of the protein across the membrane. There is no simple consensus sequence for signal peptides, but they typically show three distinct compositional zones: an N-terminal region that often contains positively charged residues, a hydrophobic region of at least six residues, and a C-terminal region of polar uncharged residues with some conservation at the −3 and −1 positions relative to the cleavage site.

In a very early bioinformatics application, von Heijne (1986) developed a weight matrix for recognition of the signal peptide cleavage site. This weight matrix has found an extremely wide usage. It does not exist as a WWW-server, but it is included in PSORT (see below).

A newer method is SignalP (**http://www.cbs.dtu.dk/services/SignalP**), which is based on neural networks (NNs) and hidden Markov models (HMMs) (Bendtsen *et al*., 2004). See Article 98, **Hidden Markov models and neural networks**, Volume 8 for an introduction to these machine-learning methods. A comparison of signal peptide prediction methods showed that both NNs and HMMs outperform the weight matrix (Menne *et al*., 2000). This still seems to be the case, even though a newer application using weight matrices has become available (Hiller *et al*., 2004) (**http://www.predisi.de/**).

The HMM included in SignalP is a complex architecture that does not adhere to the well-known profile HMMs. However, Zhang and Wood (2003) showed that the task can be done with an only slightly lower performance using a profile HMM implemented in the standard HMMER package (the model can be downloaded from **http://share.gene.com/**).

It should be noted that far from all proteins with secretory signal peptides are actually secreted to the outside of the cell. In gram-negative bacteria, they by default end up in the periplasmic compartment, and a separate mechanism is needed to secrete them to the growth medium (Pugsley *et al*., 1997). In eukaryotes, proteins translocated across the ER membrane are by default transported through the golgi apparatus and exported by secretory vesicles, but some proteins have specific retention signals that hold them back in the ER, the golgi or the lysosomes. In general, these retention signals are poorly characterized, one exception being the ER retention signal that has the consensus sequence KDEL or HDEL (van Vliet *et al*., 2003).

Some transmembrane proteins also have a cleavable secretory signal peptide that initiates translocation, whereafter the translocation is halted by a transmembrane α-helix that acts as a stop-transfer signal, leaving the protein integrated in the membrane. For a comparison of various publicly available methods for predicting transmembrane helices, see Chen *et al*. (2002) (*see* Article 38, **Transmembrane topology prediction**, Volume 7 and Article 65, **Analysis and prediction of membrane protein structure**, Volume 7).

Transmembrane helices often lead to false-positives in signal peptide prediction and vice versa. Recently, a combined HMM that deals with this problem by modeling both these signals, Phobius, (**http://phobius.cgb.ki.se**) has become available (Käll *et al*., 2004).

Other membrane proteins do not have transmembrane domains, but are linked to the membrane by a covalently attached lipid group. Prokaryotic lipoproteins have signal peptides that are cleaved by a special signal peptidase, and their cleavage site has a characteristic consensus signal with a 100% conserved cysteine in position +1. Two publicly available signal peptide prediction methods are designed to recognize prokaryotic lipoprotein signal peptides: LipoP, (**http://www.cbs.dtu.dk/services/LipoP/**), which is based on a combination of NNs and HMMs (Juncker *et al*., 2003); and SPEPlip, (**http://gpcr.biocomp.unibo.it/predictors/**), which is based on NNs combined with a simple pattern matching (Fariselli *et al*., 2003). In eukaryotes, some proteins are linked to the membrane by a glycosylphosphatidylinositol (GPI) anchor at the C-terminus or a myristoyl anchor at the N-terminus. These can be predicted with the big-Π and NMT tools (**http://mendel.imp.univie.ac.at/mendeljsp/sat/index.jsp**) (Eisenhaber *et al*.,

2003), and the myristoylation also with Myristoylator (**http://www.expasy.org/tools/myristoylator/**) (Bologna *et al*., 2004).

3. Other localization signals

The target peptides of chloroplasts and mitochondria are also N-terminal cleavable peptides (Schatz and Dobberstein, 1996). They are less well characterized than the secretory signal peptide, but they are both rare in negatively charged residues and able to form amphiphilic α-helices (Bannai *et al*., 2002; Bruce, 2000).

A widely used method to predict mitochondrial transit peptides (mTPs) is Mitoprot (**http://websvr.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter**) (Claros and Vincens, 1996). It is a feature-based method, using a linear combination of a number of sequence characteristics such as amino acid abundance, maximum hydrophobicity, and maximum hydrophobic moment (α-helix amphiphilicity) that are combined into an overall score.

A newer method, MITOPRED (**http://mitopred.sdsc.edu**), does not rely on mitochondrial targeting signals, but is based on Pfam domain occurrence patterns and the amino acid compositional differences between mitochondrial and non-mitochondrial proteins (Guda *et al*., 2004).

For chloroplast transit peptides, there is a NN-based method available, ChloroP (**http://www.cbs.dtu.dk/services/ChloroP/**) (Emanuelsson *et al*., 1999). A successor of ChloroP is TargetP (**http://www.cbs.dtu.dk/services/TargetP/**), which provides prediction of both chloroplast transit peptides, mitochondrial transit peptides, and secretory signal peptides (Emanuelsson *et al*., 2000). Both ChloroP and TargetP use a combination of NNs to calculate a transit or signal peptide score, and a weight matrix to locate the transit peptide cleavage sites.

Another NN-based method is Predotar (**http://genoplante-info.infobiogen.fr/predotar/**) (Small *et al*., 2004). In contrast to TargetP that uses moving windows to calculate the transit peptide score, Predotar uses a fixed window comprising the first 40–60 amino acids of the sorting signal. Like TargetP, it predicts mitochondrial, chloroplast, and secretory signals.

Bannai *et al*. (2002) tested a large number of physicochemical features of N-terminal parts of proteins with signal or transit peptides and obtained a combination of simple rules that yielded a discriminative performance fairly close to that of TargetP. Interestingly, a simple hydrophobicity scale even outperformed the NN-based TargetP on plant signal peptides. The resulting method is called iPSORT (**http://hypothesiscreator.net/iPSORT/**).

Not all localization signals are N-terminal and cleavable. Nuclear localization signals can occur internally in the sequence and are not cleaved. The method PredictNLS (**http://cubic.bioc.columbia.edu/predictNLS/**) (Cokol *et al*., 2000) predicts nuclear localization by comparing the query sequence to a database of experimentally verified NLS sequences and derived signals.

Peroxisomes also have their own protein import machinery. Two uncleaved signals are known: the C-terminal PTS1 and the N-terminal PTS2. PTS1 with the consensus sequence -SKL is best known, and a predictor is available (**http://mendel.imp.univie.ac.at/PTS1/**) (Neuberger *et al*., 2003).

4. Global property methods

In addition to the recognition of the sorting signals, prediction of protein sorting can exploit the fact that proteins of different subcellular compartments differ in global properties, reflected in the amino acid composition. Andrade *et al*. (1998) found that the signal in the total amino acid composition, which makes it possible to identify the subcellular location, is due almost entirely to surface residues. While the signal-prediction methods are probably closer to mimicking the information processing in the cell, methods based on global properties can work also for genomic or EST sequences where the N-terminus of the protein has not been correctly predicted. One drawback is that such methods will not be able to distinguish between very closely related proteins or isoforms that differ in the presence or absence of a sorting signal.

The NNPSL method (**http://predict.sanger.ac.uk/nnpsl/**) (Reinhardt and Hubbard, 1998) uses NNs trained on overall amino acid composition to predict localization. The method distinguishes between three bacterial compartments (cytoplasmic, periplasmic, and extracellular) and four eukaryotic compartments (cytoplasmic, extracellular, mitochondrial, and nuclear). Interestingly, plant proteins were found to be very poorly predicted, and are not included in the present method.

Nair and Rost (2003), also working with NNs, found that prediction could be improved by using information from protein structure. Specifically, they calculated amino acid composition separately for three categories of secondary structure (helix, sheet, and coil) and for surface-accessible residues. Naturally, the improvement was most pronounced when applied to proteins of known structure, but even a predicted secondary structure (according to an NN) was able to enhance prediction. The resulting method is implemented in a database, LOC3D (**http://cubic.bioc.columbia.edu/db/LOC3d/**), and a web server, LOCtarget (**http://cubic.bioc.columbia.edu/services/LOCtarget/**) (Nair and Rost, 2004).

There is a rapidly growing number of subcellular localization prediction methods based on amino acid composition and related features. The SubLoc method (**http://www.bioinfo.tsinghua.edu.cn/SubLoc/**) (Hua and Sun, 2001) is based on support vector machines (SVMs, *see* Article 110, **Support vector machine software**, Volume 8). The data set used to train SubLoc is that of Reinhardt and Hubbard (1998), but the predictive performance is significantly better than the NN. Three newer SVM applications are Esub8, PLOC, and ESLpred. Esub8 (**http://bioinfo.tsinghua.edu.cn/CoupleLoc/eu8.html**) (Cui *et al*., 2004) uses the amino acid composition of the first and last half of each sequence and distinguishes between eight subcellular locations. PLOC (**http://www.genome.ad.jp/SIT/ploc.html**) (Park and Kanehisa, 2003) uses, in addition to amino acid composition, amino acid pairs (adjacent or separated by one to three positions) to enhance prediction. It distinguishes between as many as twelve subcellular locations. ESLpred (**http://www.imtech.res.in/raghava/eslpred/**) (Bhasin and Raghava, 2004) is a four-location predictor based on the NNPSL data, which uses both amino acid composition, adjacent amino acid pairs, PSI-BLAST output, and physicochemical properties.

It should be stressed that no prediction method is better than the data set used to train it. One problem that is rarely properly addressed in global property methods is homology in the data. If the data used to test a method has sequences

that are significantly homologous to sequences in the training data, the apparent performance of the method is an overestimate. To compute a true generalizable performance, the data set should be reduced so that no homologous pairs remain (Hobohm *et al*., 1992). Reinhardt and Hubbard (1998) reduced their data set, but only removed sequences with more than 90% identity, which is clearly much higher than the homology threshold. Newer methods, with the exception of PLOC, are trained with the Reinhardt and Hubbard (1998) data or without homology reduction at all. Therefore, care should be taken when comparing performance measures.

5. Integrated methods

PSORT (**http://psort.nibb.ac.jp** or **http://www.psort.org/**) (Nakai and Kanehisa, 1991, 1992; Nakai and Horton, 1999) is an integrated system of several prediction methods, using both sorting signals and global properties. Some of the components are developed within the PSORT group, others are implementations of methods published elsewhere, including selected PROSITE patterns. PSORT is the only publicly available system that shows this degree of integration. In addition to localization (up to 16 different possible locations in plant cells), it also predicts motifs for posttranslational modifications such as lipid attachment.

All the constituent predictors provide feature values, which are then integrated to produce a final prediction. In the original version, PSORT I, the integration was done in the style of a conventional knowledge base using a collection of "if-then" rules, while the newer PSORT II version uses quantitative machine-learning techniques, such as probabilistic decision trees and the k nearest neighbors classifier to integrate scores from all the features. PSORT II is available for animal and yeast proteins (11 locations), while plant proteins still have to rely on PSORT I.

For gram-negative bacteria, there is a recently improved version named PSORT-B (**http://www.psort.org/psortb/index.html**) (Gardy *et al*., 2003) discriminating between five possible locations (cytoplasm, inner membrane, periplasm, outer membrane, and extracellular). It uses a combination of homology searches to proteins of known localization, PROSITE motifs, signal peptide, and transmembrane helix predictors based on HMMs, and a SVM-based predictor using amino acid composition.

Drawid and Gerstein (2000) developed a different integrated system for localizing all the proteins in the yeast genome to one of the five possible compartments (cytoplasm, nucleus, mitochondria, membrane, or secretory pathway). It is a Bayesian system integrating 30 features comprising both specific motifs (e.g., signal sequences or the HDEL motif), overall properties of a sequence (e.g., surface composition or isoelectric point), and whole-genome data (e.g., absolute mRNA expression levels or their fluctuations). The method is not available for submission of new sequences, but predictions for all known yeast genes can be retrieved (**http://bioinfo.mbb.yale.edu/genome/localize/**).

Acknowledgments

I thank Gunnar von Heijne and Jacob Engelbrecht for comments on the manuscript.

References

Andrade MA, O'Donoghue SI and Rost B (1998) Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology*, **276**, 517–528.

Bannai H, Tamada Y, Maruyama O, Nakai K and Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.

Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783–795.

Bhasin M and Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, **32**, W414–W419.

Bologna G, Yvon C, Duvaud S and Veuthey A-L (2004) N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics*, **4**, 1626–1632.

Bruce BD (2000) Chloroplast transit peptides: structure, function and evolution. *Trends in Cell Biology*, **10**, 440–447.

Chen CP, Kernytsky A and Rost B (2002) Transmembrane helix predictions revisited. *Protein Science*, **11**, 2774–2791.

Claros MG and Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, **241**, 779–786.

Cokol M, Nair R and Rost B (2000) Finding nuclear localization signals. *EMBO Reports*, **1**, 411–415.

Cui Q, Jiang T, Liu B and Ma S (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, **5**, 66.

Drawid A and Gerstein M (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, **301**, 1059–1075.

Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G and Wildpaner M (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Π, NMT and PTS1. *Nucleic Acids Research*, **31**, 3631–3634.

Emanuelsson O, Nielsen H, Brunak S and von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**, 1005–1016.

Emanuelsson O, Nielsen H and von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, **8**, 978–984.

Fariselli P, Finocchiaro G and Casadio R (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.

Gardy JL, Spencer C, Wang K, Ester M, Tusnády GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, *et al*. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, **31**, 3613–3617.

Guda C, Fahy E and Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.

Hiller K, Grote A, Scheer M, Munch R and Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, **32**, W375–W379.

Hobohm U, Scharf M, Schneider R and Sander C (1992) Selection of representative protein data sets. *Protein Science*, **1**, 409–417.

Hua S and Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H and Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, **12**, 1652–1662.

Käll L, Krogh A and Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, **338**, 1027–1036.

Menne KML, Hermjakob H and Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.

Nair R and Rost B (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.

Nair R and Rost B (2004) LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *Nucleic Acids Research*, **32**, W517–W521.

Nakai K and Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, **24**, 34–35.

Nakai K and Kanehisa M (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.

Nakai K and Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A and Eisenhaber F (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of Molecular Biology*, **328**, 581–592.

Park K-J and Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Pugsley AP, Francetic O, Possot OM, Sauvonnet N and Hardie KR (1997) Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria – a review. *Gene*, **192**, 13–19.

Reinhardt A and Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, **26**, 2230–2236.

Schatz G and Dobberstein B (1996) Common principles of protein translocation across membranes. *Science*, **271**, 1519–1526.

Small I, Peeters N, Legeai F and Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.

van Vliet C, Thomas EC, Merino-Trigo A, Teasdale RD and Gleeson PA (2003) Intracellular sorting and transport of proteins. *Progress in Biophysics and Molecular Biology*, **83**, 1–45.

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*, **14**, 4683–4690.

von Heijne G (1990) The signal peptide. *The Journal of Membrane Biology*, **115**, 195–201.

Zhang Z and Wood WI (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, **19**, 307–308.

Transmembrane topology prediction

Lukas Käll and Erik Sonnhammer
Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden

1. Introduction

Transmembrane (TM) proteins make up about 20% of all protein sequences known, yet less than 1% of all the known structures. This discrepancy is due to the fact that TM proteins are hard to overexpress and crystallize, and therefore difficult to examine with X-ray diffraction or NMR. It is, however, much easier to determine the *TM topology*. By this, one normally means localizing all TM segments as well as determining which subcellular compartment the loops in between the TM segments are exposed to. The problem can be addressed either experimentally using antibodies or fusing the examined gene with reporter genes or, as will be discussed here, by using a TM topology predictor.

TM topology predictors normally assume that all cellular membranes can be treated equally, regardless of which organelle or cell type they surround. The location of a loop can be classed as being on the originating side, meaning the side of a membrane from which the TM protein was inserted (normally the cytoplasm), or the translocated side, the opposite side of the membrane. The commonly used "inside" and "outside" notation is confusing and should be avoided. For instance, the inside of an organelle is the translocated side, which is normally considered the "outside."

2. TM prediction principles

Early TM helix prediction methods were based on theoretically or experimentally determined hydropathy indices of hydrophobic properties for each amino acid. For the examined protein, a hydropathy plot was calculated by summing the hydropathy indices over a window of a fixed length. A heuristically determined cutoff value was then used to indicate possible TM segments, see for instance (Argos *et al*., 1982; Kyte and Doolittle, 1982). An important improvement to this strategy came from the observation that positively charged amino acids (arginine and lysine) are overrepresented on the originating side loops of TM proteins (von Heijne and Gavel,

1988). This gives a hint about the location of the loops and led to the development of the first automated full TM topology prediction methods, for example, TOPPred (von Heijne, 1992). This method first scans the sequence for certain putative TM segments and then selects the putative segments that maximize the difference in charged amino acids in loops, summed over each side separately. Instead of only using a hydropathy index, some methods use a combination of this and the indices for amino acids known to be frequent near the end of membrane helix ends, for example, SOSUI (Hirokawa *et al*., 1998). Other methods are letting a sequence profile, for example DAS (Cserzo *et al*., 1997), or an Artificial Neural Network, for example PHDhtm (Rost *et al*., 1996), detect potential TM segments.

Instead of first scanning the sequence for TM segments and then sorting out the topology as a second step, the search for TM segments can be integrated with the evaluation of possible topologies in one step. The amino acid distribution of the investigated sequence is compared to precalculated expected amino acid distributions in each type of topologically distinct region (TM helices, originating side loops, and translocated side loops) of a TM protein (see Figure 1). Given correlation measurements between the amino acid distributions of the examined protein and the expected amino acid distributions in different topological regions, the most likely topology can be predicted. A nice feature of this approach is the ability to model all parts of the protein so that all topogenic signals are properly weighted, which is preferable to giving priority to the hydrophobic signal. This was first done by a dynamic programming algorithm in the method Memsat (Jones *et al*., 1994). Pure probabilistic approaches to the problem have been taken as well. A commonly used probabilistic framework for such tasks is the hidden Markov model (HMM). Some popular HMM-based predictors are TMHMM (Sonnhammer *et al*., 1998; Krogh *et al*., 2001) and HMMTOP (Tusnady and Simon, 1998).

A common problem for the methods mentioned above is that they have a tendency to falsely classify signal peptides as TM segments. This is a natural consequence of the fact that signal peptides, as well as TM segments, contain a long hydrophobic region (*see* Article 37, **Signal peptides and protein localization**

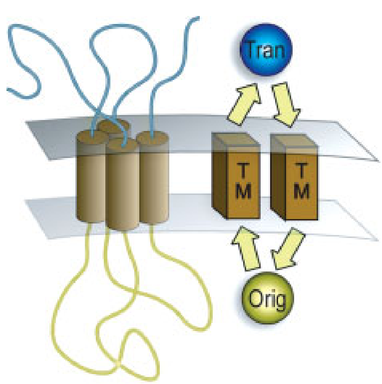


Figure 1 Modern TM predictors compare the amino acid distributions of a query protein (left) with the different regions (TM segments, originating side loops and translocated side loops) of a cyclic model (right) trained on known TM proteins

prediction, Volume 7). To discriminate between the two types of sequence features, a recently published method, Phobius (Käll *et al*., 2004), employs a combined HMM model of both signal peptide and TM structure.

β-barrel TM proteins seem to be hard to predict with the classical TM prediction methods since their TM segments are generally shorter and have a different amino acid composition than the α-helical TM segments. Lately, some methods to predict such structures have been published (Martelli *et al*., 2002; Zhai and Saier, 2002).

A common way to improve the performance of a predictor is not only to look at the examined sequence but instead to first find homologs using alignment tools such as Blast (Altschul *et al*., 1990) and then predicting the topology of the whole alignment. The idea is that topology should be conserved in a family, and by looking at the entire family there is less chance of mispredicting single, atypical members. Examples of such methods are TMAP (Persson and Argos, 1997) and PHDhtm (Rost *et al*., 1996).

A good practice when predicting TM topology is to compare the results from several different predictors. This is most easily done by running a Meta server, that is, a server that runs a number of different prediction programs. The results may be delivered in the form of e-mails from the different underlying predictors, like for Meta-PP (Eyrich and Rost, 2003), or they may be displayed side by side graphically, as by Sfinx (Sonnhammer and Wootton, 2001) (see Figure 2). The results from multiple methods may also be combined by a consensus predictor. Such a predictor only contains a weight for each method and heuristics for combining the results. An example is ConPred II (Xia *et al*., 2004).

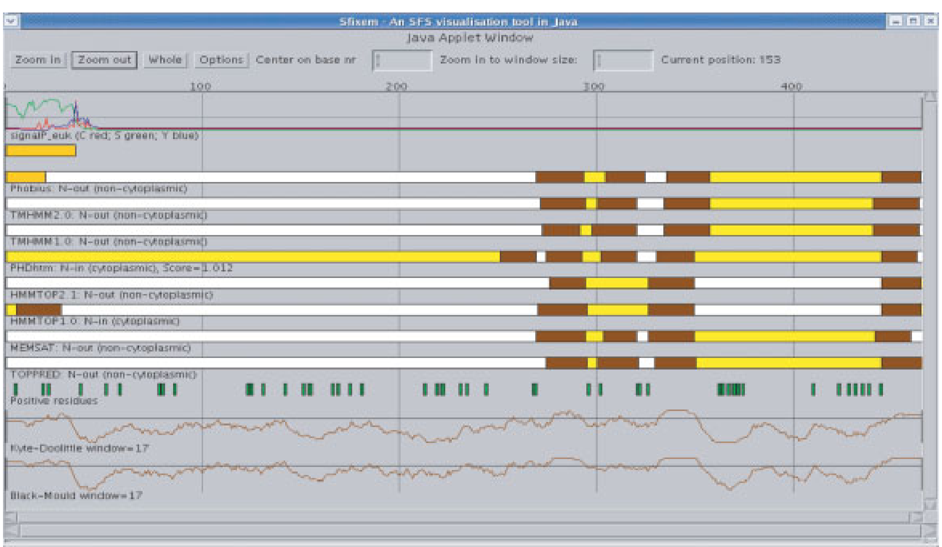


Figure 2 A screen shot from the Sfinx workbench analyzing the sequence GAC1_RAT (Swissprot: P23574) with multiple prediction methods. TM topology predictions are shown as colored boxes, and the used method is indicated underneath. Brown color indicates a TM segment, white a loop on the translocated side, yellow a loop on the originating side, and orange a signal peptide. Using a Meta server like this shows that different predictors often produce different results. The consensus topology is easily found

Table 1 URLs to the most popular TM prediction methods

Server	Single sequence	Multiple sequence	Meta server	URL
TOPPred	+			**http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html**
Sosui	+			**http://sosui.proteome.bio.tuat.ac.jp/sosuimenu0.html**
DAS	+			**http://www.sbc.su.se/∼miklos/DAS/**
Memsat	+			**http://saier-144-37.ucsd.edu/memsat.html**
TMHMM	+			**http://www.cbs.dtu.dk/services/TMHMM/**
Phobius	+			**http://phobius.cgb.ki.se/**
TMAP		+		**http://www.mbb.ki.se/tmap/**
HMMTOP	+	+		**http://www.enzim.hu/hmmtop/**
PHDhtm	+	+		**http://cubic.bioc.columbia.edu/predictprotein/**
ConPred	+		+	**http://bioinfo.si.hirosaki-u.ac.jp/∼ConPred2/**
Meta-PP	+		+	**http://cubic.bioc.columbia.edu/meta/**
Sfinx	+		+	**http://sfinx.cgb.ki.se/**

Some of the methods mentioned above have publicly available web interfaces, see Table 1.

3. Common problems

There are a number of sequence features that are problematic for most TM topology predictors. As mentioned before, signal peptides are known to easily get mixed up with TM segments. Some TM prediction methods, therefore, require the user to remove potential signal peptides by the usage of a signal peptide predictor such as SignalP (Bendtsen *et al*., 2004). This is, however, not trivial because signal peptide predictors are known to often falsely predict TM segments as signal peptides.

In TM proteins containing many TM helices, a few of the helices may be shielded from the membrane by the other TM helices. Such shielded helices are usually less hydrophobic than normal TM helices and are therefore easily missed by some TM topology predictors.

Some TM proteins contain α-helical regions that are buried in the membrane but do not cross the membrane. Such regions are often present in voltage-gated ion channels. Most predictors predict those regions as TM segments, and therefore get their topology wrong. Likewise, long hydrophobic α-helical regions buried within globular domains of soluble proteins can easily be mistaken as TM segments.

4. Benchmarking

To be able to compare different TM prediction methods, one normally assembles a test set with known topologies and examines how well each prediction agrees with it. For such benchmark experiments, it is important to collect adequate test sets since a biased test set could easily favor some of the predictors. Over the years, a couple of such test sets have been assembled, for example, the MPtopo (Jayasinghe *et al*., 2001), Möller (Möller *et al*., 2000), and TMPDB (Ikeda *et al*., 2003) sets.

Table 2 The number of correctly predicted topologies according to different benchmarking studies. Values marked with asterisk (*) were measured on older versions of the method

Method	Möller *et al*.	Ikeda *et al*.	Chen *et al*. High resolution	Chen *et al*. Low resolution
TMHMM 2.0	47%	48.4%	45%*	85%*
HMMTOP 2.1	45%*	54.1%	61%	79%
PHDhtm - single	18%	–	54%	68%
PHDhtm - multiple	–	–	66%	67%
Memsat 1.5	41%	45.1%	–	–
Number of proteins	188	122	36	165

Most modern TM topology predictors are based on machine-learning algorithms, that is, they use training sequences to assemble general knowledge about the sequence features of TM proteins. When comparing the performance of different methods, it is therefore essential to not include the training data of any of the compared methods in the test. This is, however, problematic, because so few TM topologies are known that when removing all the training set proteins, only a few proteins with dubious topologies are left. In addition, benchmarking sets generally seem to be easier to predict than genomic data (Käll and Sonnhammer, 2002).

It, therefore, may not be so surprising that different benchmark studies come to different conclusions about which TM topology prediction method is better. Möller and colleagues (Möller *et al*., 2001; Möller *et al*., 2002) rate TMHMM as the best method, while Ikeda and colleagues (Ikeda *et al*., 2002) rate HMMTOP as the best. Chen and colleagues (Chen *et al*., 2002) use two different test sets, one containing topologies with known 3D structures (high resolution) and one containing topologies without known 3D structure (low resolution). PHDhtm scored best on the high-resolution set, while TMHMM scored best against the low-resolution set. Table 2 lists a few different assessments of accuracy by a number of prediction methods. As the reader can see, the reported performance differs significantly. Hence, the reported performance figures for TM prediction methods should be interpreted with caution, both in terms of absolute and relative accuracy.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Argos P, Rao JK and Hargrave PA (1982) Structural prediction of membrane-bound proteins. *European Journal of Biochemistry*, **128**, 565–575.

Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783–795.

Chen CP, Kernytsky A and Rost B (2002) Transmembrane helix predictions revisited. *Protein Science*, **11**, 2774–2791.

Cserzo M, Wallin E, Simon I, von Heijne G and Elofsson A (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Engineering*, **10**, 673–676.

Eyrich VA and Rost B (2003) META-PP: single interface to crucial prediction servers. *Nucleic Acids Research*, **31**, 3308–3310.

Hirokawa T, Boon-Chieng S and Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.

Ikeda M, Arai M, Lao DM and Shimizu T (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biology*, **2**, 19–33.

Ikeda M, Arai M, Okuno T and Shimizu T (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Research*, **31**, 406–409.

Jayasinghe S, Hristova K and White SH (2001) MPtopo: A database of membrane protein topology. *Protein Science*, **10**, 455–458.

Jones DT, Taylor WR and Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Krogh A, Larsson B, von Heijne G and Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

Kyte J and Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**, 105–132.

Käll, L, Krogh, A and Sonnhammer, EL (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, **338**, 1027–1036.

Käll L and Sonnhammer EL (2002) Reliability of transmembrane predictions in whole-genome data. *FEBS Letters*, **532**, 415–418.

Martelli PL, Fariselli P, Krogh A and Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**(Suppl 1), S46–S53.

Möller S, Croning MD and Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Möller S, Croning MD and Apweiler R (2002) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **18**, 218.

Möller S, Kriventseva EV and Apweiler R (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Persson B and Argos P (1997) Prediction of membrane protein topology utilizing multiple sequence alignments. *Journal of Protein Chemistry*, **16**, 453–457.

Rost B, Fariselli P and Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science*, **5**, 1704–1718.

Sonnhammer EL, von Heijne G and Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, **6**, 175–182.

Sonnhammer EL and Wootton JC (2001) Integrated graphical analysis of protein sequence features predicted from sequence composition. *Proteins*, **45**, 262–273.

Tusnady GE and Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of Molecular Biology*, **283**, 489–506.

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology*, **225**, 487–494.

von Heijne G and Gavel Y (1988) Topogenic signals in integral membrane proteins. *European Journal of Biochemistry*, **174**, 671–678.

Xia JX, Ikeda M and Shimizu T (2004) ConPred_ elite: a highly reliable approach to transmembrane topology predication. *Computational Biology and Chemistry*, **28**, 51–60.

Zhai Y and Saier MH Jr. (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Science*, **11**, 2196–2207.

IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis

Yuri I. Wolf
National Institutes of Health, Bethesda, MD, USA

1. Introduction: philosophy of profile-based analysis

Sequence evolution is a largely stochastic process. Random undirected mutations occur during the DNA replication. Depending on the effect of these mutations on the structure and function of the protein, they may become fixed in the population. According to a neutral theory of molecular evolution (Kimura, 1983), the majority of the fixed mutations are effectively neutral. As two proteins diverge, following either speciation or gene duplication within the same genome, they accumulate substitutions and their similarity gradually declines over time. Random, mostly neutral, character of protein evolution ensures practical impossibility of sequence convergence of unrelated proteins.

Given a reasonable mathematical model of amino acid composition and molecular evolution of proteins, one can obtain precise quantitative estimates of several important measures. In particular, the distribution of sequence similarity measures (fraction of identical amino acids or alignment score) for random (i.e., unrelated) sequences can be computed (Karlin and Altschul, 1990). If a pair of sequence produces an alignment, whose similarity exceeds random expectations, it is a strong indication of the homology of the proteins. However, if the proteins diverged long ago, their similarity drops to the level where it is undistinguishable from the statistical "noise". For such proteins, one cannot justify the claim of homology on the basis of their alignment score. The level of similarity where confident inference of homology becomes impossible is often referred to as "twilight zone" (Doolittle, 1981).

Evolution of proteins is characterized by remarkable nonuniformity of mutability across sequence sites. Some parts of the polypeptide chain are extremely tolerant to amino acid substitutions, while others require specific residues to preserve the function and/or structure of the protein (Doolittle, 1981). The overall nonuniform shape of the distribution of evolution rates between sites can be taken into account in the sequence evolution models, but there is no general way to account for properties of particular sites in particular protein families.

```
tufB    MSKEKFERTKPHVNVGTIGHVDHGKTTLTAAITTVLAKTY----GGAARAFDQIDNAPEEKARG
                          * *  ** *        *                        * **
coaA    EQFLGTNGQRIPYIISIAGSVAVGKSTTARVLQALLSRWPEHRRVELITTDGFLHPNQVLKERG
(a)

tufB    MSKEKFERTKPHVN-VGTIGHVDHGKTTLTAAITTVLAKTYGGAARAFDQIDNAPEEKARGITI
        *            *       *  *      **           * *      *     *
fimI    MLLMRMRPSRFSINNLPRFRDVITGRDAHPCAIKITMKR-----KRLFLLASLLPMFALAGNKW
(b)
```

Figure 1 (a) A fragment of an alignment of two distantly related homologous sequences (tufB and coaA). (b) A fragment of an alignment of two unrelated sequences (tufB and fimI). Asterisks indicate residues identical between the two sequences

Two pairwise alignments with the same (low) level of sequence similarity are equivalent from a statistical point of view; however, they might have very different biological meaning, readily distinguishable by an expert in the particular group of proteins. Both alignment fragments in Figure 1 have 14–16% of identical residues. The relevance of alignment A is self-evident to anyone familiar with P-loop ATPases (Saraste *et al*., 1990), as it correctly juxtaposes important residues of the so called Walker A motif; alignment B is a spurious match of unrelated proteins.

The distinction comes easy if the proteins are well studied and the relative importance of different fragments and residues of the polypeptide chain is known. For less-characterized proteins, this approach is, by definition, unproductive.

The difference in mutability of different protein sites is easy to observe comparing multiple proteins separated by various evolutionary distances. Sites, evolving at slower rates, are more often seen as amino acid matches in pairwise comparisons; fast-evolving sites do not display conservation beyond the random expectation. Compiling a multiple alignment of distant but unambiguously related proteins allows one to assess relative prominence of different sites. Taking into account multiple alignments built for two proteins that display low sequence identity could help answer the question about the relevance of the similarity: if the alignment of these sequences, however poor, largely follows sites, conserved in each of the multiple alignments, then the alignment is likely to be biologically significant (Figure 2a); if the two multiple alignments are discordant, the juxtaposition is probably spurious (Figure 2b). Going from closely related proteins to more and more distant ones, one can learn features specific for a given level of evolutionary relationships and extend the range of confident biological conclusions far into the “twilight zone” (Doolittle, 1981).

It is possible to formalize this observation introducing the notion of sequence profile (Gribskov *et al*., 1987). Each position in a profile corresponds to a site in a protein; but, unlike a straightforward protein sequence, each site is represented not by a single amino acid but rather by a spectrum of possible variants. Highly variable sites are represented by a widely dispersed distribution of amino acids; for conserved sites, the distribution is sharply biased toward a particular amino acid (or class of amino acids, e.g., “aromatic” or “positively charged”). The simplest form of a profile representing an alignment of length L is a matrix of $L \times 20$ dimension where each column corresponds to a site in alignment, each row corresponds to an amino acid, and each matrix element shows a frequency of a given amino acid in a given alignment position. In practice, instead of straightforward frequencies,

```
MAKGEFIRTKPHVNVGTIGHVDHGKTTLTAALTYVAAAENPNV--------------EVKDYGDIDKAPEERARG
AVEASASKRKPHLNVGGMGHVDHGKTTLAAAITKVLAETGGAR---------------YTAYEEIDKAPEERARG
KVHELMMNQKNIRNISVIAHVDHGKSTLTDCLVIKAKIVSKDSG----------------GGRYMDSREDEQQRG
------MITKDIRNIAVVAHVDHGKTTLISAMLQQTSDSTT------------------AEDRVLDSHELEREKG
KMMGLQQKAAGIRNICVLAHVDHGKTTLADCLISSNGIISSRLA---------------GKLRYLDSREDEQVRG
-----MGKEKTHVNVVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAELGKGSFKYAWVLDKLKAERERG
QCERLMDNPEQIRNIAIAAHVDHGKTTLTDNLLAGAGMIS---------------EDTAGQQLAMDTEEDEQERG
MSKEKFERTKPHVNVGTIGHVDHGKTTLTAAITTVLAKTY---------------GGAARAFDQIDNAPEEKARG
                  * *  ** *        *                                   * **
EQFLGTNGQRIPYIISIAGSVAVGKSTTARVLQALLSRWP-----------EHRRVELITTDGFLHPNQVLKERG
EQFLGTNGQRIPYIISIAGSVAVGKSTTARVLQALLSRWP-----------EHRRVELITTDGFLHPNQVLKERG
GIFLQRESKSQPFIIGVSGSVAVGKSTTSRLLQILLSRTF-----------TDATVELVTTDGFLYPNQTLIEQG
KRVRERGKVGKSFIIGITGSVAAGKSTLASDLAKMLD---------------GISVEVISTDGYLYPNKILEQKG
QTFYNEKWQKIPFIIGVTGSVAVGKSTFAKKITRLFERLE-----------PEKTIAQVSADGFLMSNAELKAKN
LKRRSSILGDVPYLLGITGSVAAGKTVMSKTFSRIFAKIF-----------K-KNVDIISTDSFLKKNNELINEG
EILARLAKTEGRLIVAIAGPPGAGKSTLSDYLLHAINKGG-----------NAPSI-IVPMDGFHIDDVILEQRG
ARKCFAGSNDRRVMIAIAGAPGSGKSTLAERVVDALA--G-----------EGVSAALFPMDGFHYDDAVLEAMN
LAALQTVNPQRRTVVFLCAPPGTGKSTLTTFWEYLAQQDP-----------ELPAIQALPMDGFHHYNSWLDAHQ
```

(a)

```
MAKGEFIRTKPHVN-VGTIGHVDHGKTTLTAALTYVAAAENPNV--------------EVKDYGDIDKAPEERARG
AVEASASKRKPHLN-VGGMGHVDHGKTTLAAAITKVLAETGGAR---------------YTAYEEIDKAPEERARG
KVHELMMNQKNIRN-ISVIAHVDHGKSTLTDCLVIKAKIVSKDSG----------------GGRYMDSREDEQQRG
------MITKDIRN-IAVVAHVDHGKTTLISAMLQQTSDSTT------------------AEDRVLDSHELEREKG
KMMGLQQKAAGIRN-ICVLAHVDHGKTTLADCLISSNGIISSRLA---------------GKLRYLDSREDEQVRG
-----MGKEKTHVN-VVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAELGKGSFKYAWVLDKLKAERERG
QCERLMDNPEQIRN-IAIAAHVDHGKTTLTDNLLAGAGMIS---------------EDTAGQQLAMDTEEDEQERG
MSKEKFERTKPHVN-VGTIGHVDHGKTTLTAAITTVLAKTY---------------GGAARAFDQIDNAPEEKARG
*            *       *  *      **                           * *      *     *
MLLMRMRPSRFSINNLPRFRDVITGRDAHPCAIKITMKR--------------------KRLFLLASLLPMFALAG
---------------------------------------------------------MAIMLFLLS----PAALAG
-----------------------------------------------------------MIRKGAALVGLVLMSPV
-----------------------------------------------------MKHKLMTSTIASLMFVAGAAVAA
-----------------------------------------------------------MTRTSNPCAVVLAFAAI
-------------------------------------------------------MKKTIMSLAVVSALVSGVAFA
------------------------------------------------------------MKRLHKRFLLATFCAL
----------------------------------------------------------MKRSIIAAAVFSSFFMSA
```

(b)

Figure 2 (a) Alignment of sequences in Figure 1(a) juxtaposed with multiple alignments for tufB and coaA. (b) Alignment of sequences in Figure 1(b) juxtaposed with multiple alignments for tufB and fimI. Asterisks indicate the same pairs of residues as in Figure 1. Alignments are colored according to physicochemical properties of amino acid residues

profiles often contain precomputed weights, associated with occurrence of a given amino acid in a given position of the alignment; such profile is often referred to as position-specific score matrix (PSSM). PSI-BLAST (Altschul *et al*., 1997) and HMMer (Durbin *et al*., 1998) program families are among the most widely used profile-based sequence analysis software.

2. PSI-BLAST: Position-specific Iterated BLAST

Position-specific iterated BLAST was developed at the National Center for Biotechnology Information (NCBI) (Altschul *et al*., 1997). Its profile-based features dramatically improved sensitivity and specificity of protein sequence similarity search.

PSI-BLAST generalizes the regular BLAST scoring scheme for a comparison between a query PSSM and a single target sequence. Query PSSM is computed from the query-anchored multiple alignment (sometimes referred to as "master-slave alignment" with the query in the role of the master sequence). Scores in each column of the PSSM reflect the frequencies of different amino acids in the corresponding alignment position "diluted" by background amino acid frequencies to dampen the effect of sampling error. Query-target alignment is scored using the PSSM instead of a presupplied scoring matrix, as it is done in regular BLAST. To alleviate the effect of many highly similar sequences "outvoting" less abundant divergent forms, the amino acid frequencies used to compute the PSSM are adjusted according to a sequence weighting scheme by Henikoff and Henikoff (1994).

The primary mode of PSI-BLAST application is the iterative search starting with a single sequence as the query. The first iteration consists of a regular BLAST search against a given database. Sequences, found to be confidently similar with the query (i.e., with e-values below a specified threshold), contribute to a query-anchored multiple alignment. Weighted amino acid frequencies and position-specific scores are computed for each alignment position. Well-conserved sites, naturally, tend to favor a particular amino acid (or a narrow range of amino acids), producing position-specific scores that severely discriminate between "right" and "wrong" amino acids. Amino acid frequencies in highly variable sites approach background frequency distribution, resulting in scores uniformly close to zero. In the subsequent iterations, target sequences are compared against the query PSSM. Matches or mismatches against sites conserved in the previous iteration are scored in a sharply contrasted manner; variable sites are scored neutrally regardless of the target amino acid. Thus, target sequences that preserve the pattern of conserved sites are scored higher than those that have the same number of matching sites, but spaced in different manner. In each iteration, the signal contained in the conserved sites is reinforced by new matches; the noise in variable sites is further diluted. Highly diverged homologous sequences that were scored low in previous iterations usually get progressively higher scores, rising in the ranked list of hits.

Obviously, the success of PSI-BLAST search strongly depends on the database content and on the query choice. If the database contains a great number of sequences, related to the query, with gradually declining similarity, there is a good chance that the iterated procedure will eventually retrieve them all. It is especially helpful if the selected query is equidistant from the other homologs and does not contain too many insertions and deletions compared to the majority of its relatives. If there is a wide gap between the immediate relatives of the query and its more distant homologs, the search is likely to prematurely converge after a few iterations (i.e., no new sequences can be recognized as significantly similar because the narrow group of the closest relatives does not allow for sufficient discrimination between important and irrelevant sites). Another problem with PSI-BLAST search is the potential of "PSSM explosion". If an unrelated sequence somehow makes it into profile (usually because of its compositional bias), it is possible that it influences the scores strong enough that at the next iteration its own relatives would also be scored high. Subsequent iteration would bring more hits, unrelated to the original query. It often leads to a widespread "flattening" of

the profile, with loss of distinction between the conserved and variable sites, further reducing the search specificity.

The PSSM, constructed in the course of the iterated search, can be saved and later reused, provided that the query and the search parameters, affecting statistical calculations, remain the same. If this search is performed against the same database, it has simply the effect of bypassing the previous (often time-consuming) iterations. The PSSM, however, can be used to search a different database. Typically, one constructs family-specific profiles by iterative searches against a sequence database, containing maximum available diversity of sequences. These profiles are later used for searches in particular genomes or other specialized datasets. Protein PSSM queries can be also used in a search against a nucleotide sequence database, the latter being dynamically translated in all six frames.

PSI-BLAST can initiate a search with a PSSM constructed from a multiple sequence alignment. One of the alignment sequences (with gap characters removed) must be designated as the master query; thus many different, although similar, PSSMs can be obtained from the same multiple alignment. This approach has an advantage of the possibility to use a high-quality expert-curated alignment to make a profile.

Recent modifications of PSI-BLAST introduce further modification of the scoring matrices depending on the amino acid frequency bias in the query (composition-based statistics). Practically, this feature greatly increases the search specificity for the price of a certain reduction of sensitivity; this option is especially attractive for fully automated projects.

PSI-BLAST is distributed as a stand-alone program that can be used for a search against a local database. Web-based PSI-BLAST service, provided by the National Center for Biotechnology Information, contains an important additional feature: a possibility of direct user control over the sequences, contributing to the PSSM construction. Checking and unchecking control boxes, the user can include or exclude sequences regardless of their formally computed statistical significance. Allowing an expert opinion to influence the PSSM composition gives the approach much greater flexibility.

3. IMPALA: Integrating Matrix Profiles and Local Alignments

Accumulation of a collection of high-quality profiles poses a technical challenge of reversing a profile search scenario. Instead of using PSSM queries in a search against a sequence database, it is often practical to run a (series of) single-sequence query against a library of profiles. IMPALA, released in 1999, provided this capability (Schaffer *et al*., 1999). A collection of PSI-BLAST profiles and corresponding master query sequences needs to be processed into a searchable library for use with IMPALA. Unlike BLAST family of programs, which use empirical “X-dropoff” algorithm to produce and score local alignment between the query and the target (Altschul *et al*., 1997), IMPALA implements rigorous Smith–Waterman algorithm (Smith and Waterman, 1981), producing provably optimal results. This, of course, trades off the execution speed for the accuracy. Additionally, IMPALA included several advanced techniques of PSSM handling

(notably, finer scaling of the scores), later incorporated into mainstream BLAST software. IMPALA is distributed as a stand-alone suite of programs.

4. RPS-BLAST: Reverse PSI-BLAST

RPS-BLAST (Marchler-Bauer *et al*., 2002) is a direct implementation of BLAST search algorithms in a reversed manner – with a protein sequence query and a PSSM library as a target. Like with IMPALA, PSI-BLAST profiles are postprocessed into a searchable database for use with RPS-BLAST. RPS-BLAST has a significant advantage in running speed over IMPALA, with only a minor degradation of sensitivity. RPS-BLAST is the search engine of the CD-search service (Marchler-Bauer *et al*., 2002) of the National Center for Biotechnology Information, available through a Web interface.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry. *Science*, **214**, 149–159.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press: Cambridge.

Gribskov M, McLachlan AD and Eisenberg D (1987) Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 4355–4358.

Henikoff S and Henikoff JG (1994) Position-based sequence weights. *Journal of Molecular Biology*, **243**, 574–578.

Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2264–2268.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press: Cambridge.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY and Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, **30**, 281–283.

Saraste M, Sibbald PR and Wittinghofer A (1990) The P-loop – a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Sciences*, **15**, 430–434.

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L and Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Introductory Review

The domains of life and their evolutionary implications

Carl R. Woese

University of Illinois at Urbana-Champaign, Urbana, IL, USA

1. Introduction

Over a century ago, Darwin wrote: "The time will come I believe, though I shall not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of nature" (Burkhardt and Smith, 1990). From his phrasing, it is clear that Darwin had more in mind than just the Plant and Animal Kingdoms. Indeed, by Darwin's time bacteria ("infusoria" as they were then called) were not only recognized but also recognized as evolutionarily distinct from both plants and animals. Yet, twentieth century biology did not seem particularly interested in Darwin's vision of an all-embracing phylogeny. Biology was now headed in an entirely new (reductionist) direction, one in which evolution was far from a central concern. The classical evolutionists of the twentieth century, Darwin's descendants, were seriously interested only in animals and plants. And the microbiologists, who should have been vitally interested in bacterial phylogenies, became discouraged over the difficulty of determining a phylogenetic classification for bacteria (on the basis of classical criteria), and threw in the towel – just as the problem was about to yield (Woese, 1994; Stanier and van Niel, 1962). Before doing so, however, they invented a guesswork solution to the question of bacterial relationships, the "prokaryote" (Stanier and van Niel, 1962). All bacteria were arbitrarily declared to be of a kind; all were "prokaryotes", which meant that (1) all shared a common general type of "prokaryotic" organization and (2) all had come from common "prokaryotic" ancestral stock (Stanier and van Niel, 1962; Stanier *et al*., 1963). Although no facts supported it, the guesswork "prokaryote" rapidly became doctrine, thereby allowing microbiologists to wash their hands of the vexing problem of bacterial genealogies. To this day, Biology remains saddled with the "prokaryote", although the concept has long ago been exposed for what it is – pure rhetoric (Woese, 2004).

However, Darwin's vision of a comprehensive genealogy, one that embraced all life, did not permanently fade. Unknowingly, molecular technology came to the rescue (Sanger and Tuppy, 1951; Sanger and Thompson, 1953). While their gross properties – shapes, physiologies, motilities – reveal little about bacterial genealogies, their genealogies are just as apparent as any others on the molecular

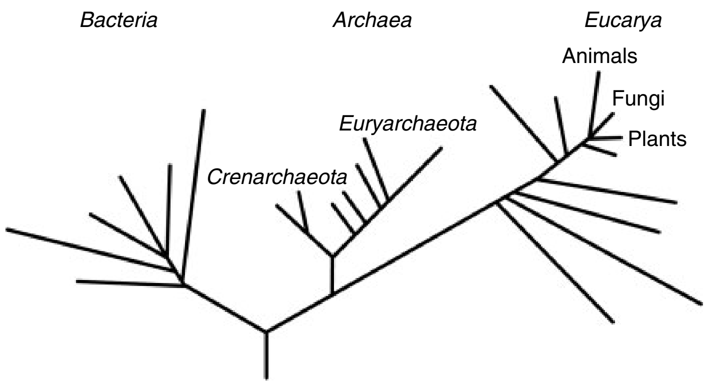


Figure 1 The universal phylogenetic as determined by ribosomal RNA sequences analyses (Reproduced from Woese CR (2002) by permission of National Academy of Sciences, U.S.A.)

level. Once this was recognized (which took a while), techniques for sequencing macromolecules (the first of which were developed in the 1950s) provided more than enough information to establish all genealogical relationships. By the early 1970s, a program to determine bacterial genealogies through a molecular sequencing was underway (Fox *et al*., 1980). The result was nothing short of spectacular; see Figure 1. For the first time, all the "great kingdoms of nature" were brought together in a common evolutionary framework. Evolutionary study had burst the bounds of the Animal and Plant Kingdoms.

2. The primary lineages

As Figure 1 demonstrates, life on this planet divides in the first instance into three primary lineages, now known as the Bacteria (or eubacteria), the Archaea (which, like Bacteria, comprises single-celled organisms), and the Eukarya (or eukaryotes) (Woese *et al*., 1990). As the tree structure also indicates, the distinctions (phylogenetic distances) among them is a striking one, far greater than are any seen within the individual primary groupings themselves. However, this strictly quantitative linear measure cannot capture the distinction's full quality. Each of the three primary lineages is associated with a unique type of cellular organization. It has long been obvious that this is true of the eukaryotic cell (the classical textbook picture of a cell), but it is only recently that it has also been recognized for the organization of the archaeal cell (Graham *et al*., 2000). The tripartite division of life is easily seen in various of the cell's subsystems. The information processing systems – translation, transcription, and genome replication – exhibit this most strongly.

Consider translation. Most of the mechanism's componentry is universal, but the bacterial version easily distinguishes itself from the other two in two ways: (1) in terms of novel (unique) componentry and (2) in terms of the degree of sequence dissimilarity among components that are universal in distribution (Olsen and Woese, 1996). With regard to novel componentry, the bacterial ribosome contains a relatively small number ribosomal proteins (∼5 in the small subunit ribosome; ∼15

in the large) not found in the ribosomes of either the archaea or eukaryotes. On the other hand, the archaeal ribosome shows a (somewhat larger) number of ribosomal proteins (~12 and ~20, small and large subunits respectively) not found in the bacteria, *all* of which (save one) *are* found in the eukaryotic ribosome! (The eukaryotic ribosome in its turn has a small number of additional ribosomal proteins that the others lack.) With regard to universal componentry, the bacterial ribosomal protein sequences are, in almost all cases, readily distinguishable from their archaeal and eukaryotic counterparts (Olsen and Woese, 1996; Vishwanath *et al*., 2004). The difference is great enough to have a qualitative aspect to it, a difference in "genre". The eukaryotic versions of the universal ribosomal proteins are almost exclusively of the archaeal genre.

From the evolutionist's perspective, the most intriguing aspect of the translation componentry is the tRNA charging enzymes, the so-called aminoacyl-tRNA synthetases (aaRSs) (Woese *et al*., 2000). Unlike others of the ribosomal componentry, these enzymes have obviously been subject to extensive horizontal gene transfer (HGT). But the pattern characteristic of the ribosomal proteins is, nevertheless, evident here too: for the majority of the 20-odd aaRS types, there exists a bacterial and an archaeal genre (the latter found in most of their eukaryotic counterparts as well). What is seen among the aaRSs but not with the ribosomal proteins, however, are cases where the archaeal genre of a given aaRSs shows up in one or a few bacterial taxa – an obvious result of HGT. The "generic" distinction between the bacterial and the archaeal/eukaryotic versions of the aaRSs (and by extension all of the ribosomal componentry) has not been caused by, but has persisted in spite of, extensive horizontal mixing of genes among the three primary lineages (Woese *et al*., 2000). This generic distinction stands witness to some unknown ancient evolutionary transition.

The evolutionarily pattern seen in the case of translation reappears in transcription, but in a somewhat more blatant form. The generic difference between bacteria and the other two domains is most extreme in the genome replicating apparatus. In the case of transcription, the three principal subunits of the bacterial DNA-dependent RNA polymerase have counterparts in the other two domains, but a pronounced difference in "genre" again distinguishes the bacterial from the archaeal (and related eukaryotic) sequences thereof (Langer *et al*., 1995; Olsen and Woese, 1996). Moreover, the archaeal version of the transcription polymerase contains additional subunits (~6) not found in the bacteria, and these (as now might be expected) have homologs in the eukaryotic RNA polymerases. Continuing the similarity with translation, the eukaryotic versions of the transcription apparatus contain still further subunits unique to themselves (Langer *et al*., 1995). The transcription initiation complex used by Bacteria is simply different from those found in the Archaea and Eukarya, with the archaeal version again looking like a simpler form of the eukaryotic (Bell and Jackson, 2001).

What all this means is uncertain at present. I do not feel one should jump to the simplistic conclusion that the archaeal/eukaryotic similarity implies specific relationship between the two at the organismal level. It is best to hold the matter open, leaving us with a stimulating question.

A different impression emerges when relationships among the three organismal domains are viewed from the perspective of cellular metabolism (Rogozin *et al*.,

2003). Here, much of eukaryotic intermediary metabolism seems to have been derived from the bacteria. Some of the eukaryotic intermediary metabolism, of course, was acquired from the endosymbiont that evolved into the mitochondrion: and the corresponding genes readily trace back to an ancestry within the alpha proteobacterial phylum (the taxon from which the original mitochondrial endosymbiont came). Yet, the bulk of the eukaryote's acquired metabolic capacity traces phylogenetically to a region deep on the stem of the bacterial lineage (Canback *et al*., 2002), significantly below the earliest bifurcations known for the recognized bacterial lineages. In other words, the type of "bacteria" that might have given rise to the bacterial-type metabolism found in eukaryotes today predated what we take to be the most recent common ancestor of all extant bacterial lineages; which asks the question of what kinds of organisms these ancient "proto-ancestral eubacteria" were.

All this raises the interesting question as to why eukaryotic metabolism does not reflect archaeal metabolism, as one sees with the archaeal information processing systems. The answer may be that the archaea have little to offer in the way of metabolic versatility, at least of a kind that has been indigenously generated. The euryarchaeota, for example, exhibit five general classes of phenotype: the methane producers; the anaerobic thermophilic sulfate reducers (Archaeglobales); the haloarchaea (heterotrophs that grow aerobically in high salt environments); the Thermoplasmales (a collection of heterotrophs growing at elevated temperatures and low pH); and the Thermococcales (again, a rather nondescript high-temperature heterotrophic phenotype) (Olsen *et al*., 1984). The native euryarchaeal phenotype is clearly methanogenesis, for it is phylogenetically spread throughout the group, in four separate phyla. The other four euryarchaeal phenotypes, however, are each confined to a particular (major) lineage within the overall group, all but one of them, the Thermococcales, being sister groups to one of the methanogenic lineages (the Methanomicrobiales). The interesting thing about these non-methanogenic phenotypes is that the evolution of their diverse metabolisms is the result of acquiring key genes from the eubacteria (Baliga *et al*., 2004; Deppenmeier *et al*., 2002). It would appear that only the eubacterial lineage has been metabolically inventive, and that the other two domains, archaea and eukarya, have fashioned the bulk of their metabolic diversity by importing bacterial genes.

3. Probing the depths of evolutionary time

Molecular sequence-based analyses take us into virgin territory, early evolutionary stages that could not even begin to be conceptualized by the classical biologist. Sequence comparisons provide evidence of evolutionary events, possibly predating cells as we know them – evidence we have yet to fully understand (Woese, 2002; Woese, 2004). For example, many protein families (formed through duplication and functional diversification of some ancestral gene) had to have arisen before the stage represented by the "root" of the universal tree. The protein couple EF-Tu and EF-G are an example. The two play complementary roles in translation and both are essential to the process – being found in all organisms. Yet, the two have clearly evolved from the same common ancestral protein (gene) (Olsen and Woese,

1996). (Evidence also suggests that neither of them has been subject to significant HGT.) In detecting their relatedness, one is sensing evolution at stages prior to the "most recent common ancestor" of extant life. A similar conclusion can be adduced for others of the cellular systems, for example, certain multisubunit enzymes of universal distribution in which two or more of the subunits come from the same family; or certain universal metabolic pathways in which enzymes at different steps in the pathway contain subunits whose sequences have familial relationship; and so on. All of this is raising the important question of what the "root" of the universal organismal tree actually represents (Woese, 2002; Woese, 2004). Traces of some mysterious "Wonder Land" of primitive evolution lie hidden in the sequences (and structures) of extant molecules.

4. Gene shuffling

The strange phenomena being unearthed by genomic probings into the depths of evolution are raising serious questions about our simplistic picture of evolution, which served so well when we had only to explain the evolutionary ebb and flow for "modern cells" and organisms (particularly the plants and animals). Unfortunately, when reported, new and revolutionary findings often come wrapped in hyperbole, so that a disinterested appreciation of the facts usually requires deconstructing the exaggerated interpretations that accompany them. Fortunately, there exists a touchstone of evolutionary common sense, and that, as always, remains Darwin. It is nothing short of incredible that words written well over a century ago (in "Origin of Species") are directly relevant and essential to our conceptualization of evolution today (Darwin).

The phenomenon that most confounds molecular evolutionists is HGT. Although HGT has been known for several decades, it is only with the accumulation of large amounts of genomic sequence information that its full evolutionary impact has been appreciated. It seems that, given sufficient time, HGT can affect the majority of (if not all) genes in the genome (Doolittle, 1999a; Gogarten *et al*., 2002; Bapteste *et al*., 2004). Not only does HGT provide a major source of evolutionary novelty, it can also result in the subtle replacement of a cellular function by some alien equivalent. This latter consequence of HGT becomes increasingly problematic as phylogenetic inference is extended further into the past (to the deep branches in the universal tree). The question, of course, is *how* problematic this type of HGT becomes in the limit; can it actually erase the entire organismal genealogical record, which would mean that we are deceiving ourselves in trying to infer universal phylogenetic trees? There are many who think this the case, and have proposed alternative evolutionary scenario to account for the genomic sequence data (Doolittle, 1999a; Gogarten *et al*., 2002; Bapteste *et al*., 2004; Doolittle, 2000; Doolittle, 1999b). It is common to read about "rerooting", "uprooting", or outright "felling" the universal organismal tree (Doolittle, 2000). Chief among the alternative interpretations is the "Web of Life" (see Figure 1; Bapteste *et al*., 2004; Sapp, 1994), which not only casts doubt on the validity of the current Tree of Life but also questions the scientific value of representing evolutionary descent as a tree.

This confusion is only apparent. It disappears with a dose of Darwinian insight. "Descent with variation" is a central Darwinian tenet. It is important to note that Darwin did not specify the nature of variation, other than viewing it as occurring in small steps, that is, any given variation is only a minor perturbation on the existing organismal design. It was the twentieth century neo-Darwinian geneticists who defined variation, establishing unequivocally that it can be brought about by point mutations and the like. They made one mistake, however, which has carried through to this day (and underlies the present confusion); and that was to assume that localized mutations are the *only* source of variation. Now we know they are not: variation is also the product of horizontal acquisition of genes (an idea long ago entertained, historians tell us (Bateson, 1913; Sapp, 1994), by the great English geneticist William Bateson and others). In other words, HGT is perfectly consistent with Darwin's original view of variation – and that is how we should view it.

Although adaptation to new environments often involves HGT (among eubacteria or archaea), there have been no documented cases of horizontal gene acquisition changing the basic nature of the organism. HGT drastic enough to do that is something that has only been *postulated*. HGT and outright cellular fusion events are central to almost all speculation concerning the origin of the eukaryotic cell (Doolittle, 1998; Hartman, 1984; Martin and Muller, 1998). The endosymbiotic origin of the mitochondrion is often cited as an example of this; but it is not. Although the endosymbiotic bacterium that ultimately evolved into the mitochondrion was indeed degenerated almost to the point of unrecognizability in the process, nothing of the sort happened to the eukaryotic cell itself; acquisition of a mitochondrion left the basic eukaryotic cellular organization intact (Kurland and Andersson, 2000).

Let us consider separately the two main types of HGT, that which introduces new functionality and that which surreptitiously replaces an endogenous function by an alien equivalent (homolog). Variation of the first type is clearly adaptive and that of the second, adaptively neutral. Darwin eschewed the use of adaptive characters in genealogical reconstruction, saying " . . . adaptive characters, though of paramount importance to the being, are of hardly any importance in classification . . . " (Darwin, 1859). Although Darwin understood that the accumulation of variations over long periods of time drastically change the character of an organism, he held that, despite this, genealogically telling (essential) characteristics persisted (at least among animals or plants) although obviously decreasing in number the further back in time the genealogy is traced. Looking at Figure 2 of Doolittle (1999b), it is apparent that the huge genetic flows suggested by the (thick) arrows for the most part represent (horizontally acquired) variations that are adaptive in nature are involved with what Darwin called "the general place of each being in the economy of nature" (Darwin, 1859). To suggest, as Figure 2 of Doolittle (1999b) does, that these horizontal flows of adaptive characters are genealogically telling is misleading.

The *history* of (organismal) descent (whose essence lies in variation) should not be confused with the genealogical *record* thereof (whose essence lies *in* variant characteristics). Strictly speaking (and practically speaking in the case of distant relationships), a genealogy is merely a record of descent, no matter how

uninformative it otherwise may be. Darwin again: " . . . we have to discover the lines of descent by the most permanent characters, however slight their vital importance may be" (Darwin, 1859). Thus, the key question here is whether there exist organismal genes (characters) that have permanently resided within the genomes of particular organismal lineages long enough to serve as reliable genealogical markers over the interval of genealogical interest – whether the interval be, for example, as short (recent) as the descent of the mammals or vertebrates in general, or as long (ancient) as that of all archaea. Darwin was under no illusion that the number of genealogically informative characters (genes, in today's terms) had to be large or that the genealogical record diminished the further back in time genealogy is traced. (That is only common sense.) Thus, the modern notion that genealogical inference must somehow be based upon the entire genome (and reflect some kind of genomic consensus) is wrong headed, especially when it is applied to the very ancient genealogical record. (Common sense alone dictates that ancient records should be scant, minimal – and they will involve only the most permanent features of the organisms, as Darwin said.) A representation of descent that lumps the overwhelming number of adaptive characters with genealogically informative ones, as indeed the reticulated "thicket" of Figure 2 in Doolittle (2000) does, completely obscures any organismal genealogical trace that may be present (especially at the "base" of the phylogenetic tree). This then gives the false impression that an ancient genealogically trace of organismal descent does not actually exist. Could this be true?

As we have seen above, there exists strong evidence that genes are capable of tracing organismal lineages at least as far back into the past as the "root" of the universal tree (Figure 1). Recall the above discussion of the (20-odd) aminoacyl-tRNA synthetases, all of which have been subject to extensive HGT, but despite this, exhibit (albeit in eroded form) the same "generic" pattern of phylogenetic relationship among the primary lines of descent that characterizes the other components of the translation apparatus (which latter have not undergone extensive, wide ranging HGT) (Olsen and Woese, 1996). This pattern (generic distinction) cannot have been caused by any HGT that occurred subsequent to the evolutionary stage represented by the "root" of the universal tree. The pattern is indeed a genetic trace of a major organismal evolutionary transition that predated the stage represented by the tree's "root". Thus, Darwin's original statement (quoted above) that ultimately we would "have very fairly true genealogical trees of each great kingdom of nature" seems remarkably prescient today.

In summary, given that rRNA is a central, integral part of the cell (and so relatively unlikely to have undergone significant HGT; Fox *et al*., 1977), there can be little doubt that the overall validity of the branching orders in the universal phylogenetic tree based upon rRNA sequence analyses does indeed approximate the true organismal genealogical tree. The sole issue that remains in doubt – and a challenging one it is – is what the "root" of this (r-RNA-based) tree represents. Is it a "root" (ancestor) in the classical sense of that word (as, for example, when biologists speak of the common ancestor of all vertebrates), or does the "root" not signify an organismal ancestor at all, but some evolutionary transition in the process of evolving cellular life – a kind of "origin of species" that Darwin in his day could have known nothing about?

5. Darwin, common descent, and the root of the universal tree

Darwin's position on the common ancestry of life is relevant here, for we credit him with declaring that all life arose from one "primordial form". Actually, Darwin did not require this to be the case; he merely entertained it as a possibility (Darwin, 1859). It was his twentieth century successors who confidently asserted it as a doctrine, "The Doctrine of Common Descent". (Darwin explicitly knew the number of ancestral "primordial forms" to be irrelevant to the Theory of Evolution, and was comfortable with both "a few forms or ... one" (Darwin, 1859).) In the twentieth century, a single ancestral form seemed the unavoidable consequence of the fact that all organisms shared a common complex core of biochemistry. No one at the time entertained the alternative that this commonality could also have been achieved through HGT. Today, "Common Descent" must be treated as an important question, but no longer a Doctrine.

The key to understanding what the "root" of the universal tree signifies is understanding what defines the (phylogenetic) span and frequency of HGT. Clearly, the function encoded by a gene is the strongest determinant. But that cannot be all, for among the most important functions in the cell, there exists a great difference in their susceptibility to horizontal gene displacement. Analyses shows that the primary defining factor here is the ways in and degree to which a given component is integrated into the cellular fabric. Look, for example, at the differences in how HGT has affected the componentry of the cellular translation mechanism. Almost all of that mechanism's central componentry is highly resistant to horizontal gene displacement *except* for the aminoacyl-tRNA synthetases (as discussed above). What distinguishes the latter componentry from the other parts of the mechanism is modularity: aminoacyl-tRNA synthetases interact with cellular componentry in general only minimally (Woese *et al*., 2000), which means that there exist relatively few constraints upon the shapes, sizes, and compositions of these enzymes. This, plus the fact that the molecules upon which these enzymes operate, the tRNAs, are universal in distribution and very uniform in shape, makes the aminoacyl-tRNA synthetases relatively susceptible to (neutral) horizontal gene displacement. In other words, it is the integration of componentry into the cellular fabric that largely determines the degree to which they are subject to horizontal gene displacement (and the character of the that displacement). From this it follows that the simpler the design of a cell, the more susceptible its componentry is to HGT. In the early stages (proteinaceous) of cellular evolution, cell design (their organization) must have been far simpler than it is in modern cells. Hence, while HGT plays a relatively minor (but still important) role in (microbial) evolution today, in earlier stages of cellular evolution, it must have been an absolutely dominant factor in their evolution.

Carrying this line of reasoning to the extreme, one can envision an early stage in the evolution of cells wherein genetic exchanges among all evolving entities were high enough that evolution of cellular organization behaved as though it were a communal rather than an individual (individual lineages) affair, which when viewed in retrospect would give the *appearance* of a unitary common ancestor, a typical "root" to a universal organismal tree (Woese, 2002). All evolving cell designs could usefully partake of the inventions created by any other evolving cell design. At this

stage, there would be no such thing as a permanent genealogical trace residing in the organismal genome; any such trace would be ephemeral only.

The root of the universal tree may then represent a transitional stage, one in which a primitive evolutionary dynamic (dominated by HGT) gives way to the modern type of evolution dynamic (having much lower levels and restricted range of HGT) – when a "pre-Darwinian" evolutionary world transforms into the familiar "Darwinian" one. It may well be that the three primary cell types, the (eu)bacteria, archaea, and eukarya, each existed in some primitive kind of form in this pre-Darwinian world. Evidence suggesting this comes from certain gene families, ones that reasonably had to have existed prior to the stage represented by the "root" of the universal tree, but that are confined to only one of the three primary cell lineages – a fact hard to explain otherwise (Woese, 2004). Since the transition from a pre-Darwinian condition to a Darwinian one would be a function of the cell design undergoing that transition, this would almost certainly mean that each of the three primary cell types underwent the transition from the pre-Darwinian evolutionary world to the Darwinian one at a different time. In other words, there were, in effect, at least three different "primordial forms" from which extant life arose.

It is clear that molecular approaches to evolution are going to show us a very different world of evolution than that early twentieth century biologists envisioned. But wonderfully, a world more in keeping with, and with the spirit of, Darwin's simple original conception, than with the twentieth century neo-Darwinian evolution that replaced it.

Acknowledgments

The authors work is supported by his grants from the Department of Energy and the National Aeronautics and Space Administration.

References

Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, Shannon P, Chiu Y, Weng RS, Gan RR, *et al*. (2004) Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. *Genome Research*, **14**, 2221–2234.

Bapteste E, Boucher Y, Leigh J and Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. *Trends in Microbiology*, **12**, 406–411.

Bateson W (1913) *Problems of Genetics*, Yale University Press: New Haven.

Bell SD and Jackson SP (2001) Mechanism and regulation of transcription in archaea. *Current Opinion in Microbiology*, **4**, 208–213.

Burkhardt F and Smith S (1990) *The Correspondence of Charles Darwin*, Vol. 6, Cambridge University Press: Cambridge, pp. 786–820.

Canback B, Andersson SG and Kurland CG (2002) The global phylogeny of glycolytic enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6097–6102.

Darwin C (1859) *On the Origin of Species 1859*, A facsimile of the first edition, Cambridge University Press: Cambridge.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, *et al*. (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *Journal of Molecular Microbiology and Biotechnology*, **4**, 453–461.

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, **14**, 307–311.

Doolittle WF (1999a) Lateral genomics. *Trends in Cell Biology*, **9**, M5–M8.

Doolittle WF (1999b) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129. [Web shown here].

Doolittle WF (2000) Uprooting the tree of life. *Scientific American*, **282**, 90–95.

Fox GE, Pechman KR and Woese CR (1977) Comparative cataloging of 16 S ribosomal RNA: molecular approach to procaryotic systematics. *International Journal of Systematic Bacteriology*, **27**.

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, *et al*. (1980) The phylogeny of prokaryotes. *Science*, **4455**, 457–463.

Gogarten JP, Doolittle WF and Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, **19**, 2226–2238.

Graham DE, Overbeek R, Olsen GJ and Woese CR (2000) An archaeal genomic signature. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3304–3308.

Hartman H (1984) The origin of the eukaryotic cell. *Speculations in Science and Technology*, **7**, 77–81.

Kurland CG and Andersson SG (2000) Origin and evolution of the mitochondrial proteome. *Microbiology and Molecular Biology Reviews*, **64**, 786–820.

Langer D, Hain J, Thuriaux P and Zillig W (1995) Transcription in archaea: similarity to that in eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5768–5772.

Martin W and Muller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature*, **6671**, 37–41.

Olsen GJ and Woese CR (1996) Lessons from an Archaeal genome: what are we learning from *Methanococcus jannaschii*? *Trends in Genetics*, **12**, 377–379.

Olsen GJ, Woese CR and Overbeek R (1984) The winds of (evolutionary) change: breathing new life into microbiology. *Journal of Bacteriology*, **176**, 1–6.

Rogozin IB, Babenko VN, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Mirkin BG, *et al*. (2003) Evolution of eukaryotic gene repertoire and gene structure: discovering the unexpected dynamics of genome evolution. *Cold Spring Harbor Symposia on Quantitative Biology*, **68**, 293–301.

Sanger F and Thompson EOP (1953) The amino-acid sequence in the glycyl chain of insulin. *The Biochemical Journal*, **53**, 353–374.

Sanger F and Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. *The Biochemical Journal*, **49**, 481–490.

Sapp J (1994) *Evolution by Association*, Oxford University Press: Oxford.

Stanier RY, Doudoroff M and Adelberg EA (1963) *The Microbial World*, *Second Edition*, Prentice-Hall: Engelwood Cliffs.

Stanier RY and van Niel CB (1962) The concept of a bacterium. *Archiv Fur Mikrobiologie*, **42**, 17–35.

Vishwanath P, Favaretto P, Hartman H, Mohr SC and Smith TF (2004) Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Molecular Phylogenetics and Evolution*, **33**, 615–625.

Woese CR (1994) There must be a prokaryote somewhere: microbiology's search for itself. *Microbiological Reviews*, **58**, 1–9.

Woese CR (2002) On the evolution of cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 8742–8747.

Woese CR (2004) A new biology for a new century. *Microbiology and Molecular Biology Reviews*, **68**, 173–186.

Woese CR, Kandler O and Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4576–4579.

Woese CR, Olsen GJ, Ibba M and Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, **64**, 202–236.

Introductory Review

Phylogenetic profiling

Matteo Pellegrini, Todd O. Yeates and David Eisenberg

Howard Hughes Medical Institute, University of California at Los Angeles, Los Angeles, CA, USA

Sorel T. Fitz-Gibbon

IGPP Center for Astrobiology, University of California at Los Angeles, Los Angeles, CA, USA

1. Introduction

Biology has been profoundly changed by the development of techniques to sequence DNA. The advent of rapid sequencing in conjunction with the capability to assemble sequence fragments into complete genome sequences enables researchers to read and analyze entire genomes of organisms. Parallel progress has been made in algorithms to study the evolutionary history of proteins. The techniques rely on the ability to measure the similarity of protein sequences in order to determine the likelihood that different proteins are descended from a common ancestor. It is therefore possible to reconstruct families of proteins that share a common ancestor.

Combining these two capabilities, we can now not only determine which proteins are coded within an organism's genome but we can also discover the evolutionary relationships between the proteins of multiple organisms. Phylogenetic profiling is the study of which protein types are found in which organisms.

In order to perform phylogenetic profiling, one must first establish a classification of proteins into families. An example of such a classification scheme across a broad range of fully sequenced organisms is the Clusters of Orthologous Groups (Tatusov, 1997), where an attempt is made to group together proteins that perform a similar function. Next, each organism is described in terms of which protein families are coded or not coded in its genome.

As we will see in this review, this simplified representation is useful for exploring the evolutionary history of an organism as well as for studying the function of protein families and how they may be related to observable phenotypes.

2. Genome phylogeny

Species phylogenies have traditionally been constructed by measuring the evolutionary divergence in a particular family of proteins or RNAs (Fitch, 1967). The most commonly used sequence for such phylogenetic reconstructions is that of the small subunit ribosomal RNA. The advantages of using this RNA gene are that it is found in all organisms, and it has evolved relatively slowly, thus permitting the construction of phylogenies between distant organisms.

Access to the complete genomes of organisms offers a new approach to phylogenetic reconstruction. Rather than looking at the evolution of a single protein or RNA family, it is now possible to compare the gene content of two organisms. This general approach to phylogenetic reconstruction has been applied in a variety of ways (Fitz-Gibbon, 1999; Snel, 1999; Tekaia, 1999; Lin, 2000; Montague, 2000; Wolf, 2001; Bansal, 2002; Clarke, 2002; House, 2002; Li, 2002).

Several metrics have been used to measure the similarity of two organisms on the basis of their gene contents, including the percentage of genes shared by the two species. Furthermore, phylogenetic trees may be reconstructed using several techniques including distance-based phylogenies and parsimony. In general, the trees constructed using whole genome comparisons are similar to those using small subunit rRNA sequences, with occasional discrepancies of interest (Figure 1).

3. Coevolution of protein families

Before fully sequenced genomes became available, the computational study of protein function relied entirely on the detection of sequence similarity. The general notion upon which these studies are based is that proteins with detectable sequence similarity are likely to have evolved from a common ancestor and thus by definition are homologs. Furthermore, such proteins are likely to have preserved common structure and function. Therefore, similarity detection may be used to assign a putative structure and function to proteins that have a sufficient degree of sequence similarity to an experimentally characterized protein. The definition of "sufficient degree" of similarity has been at the center of much research. Depending on the methodology used to determine sequence similarity, various statistical tests have been devised to determine whether two proteins have truly evolved from a common ancestor.

Although techniques based on sequence similarity are powerful, they are unable to inform us about a possible structure or function of a protein family that does not contain experimentally characterized members. This is a significant limitation because a large fraction of all protein families currently fall within this category. Phylogenetic profiling may be used to address this problem, and give us at least partial functional information on these protein families by determining the pathway or complex to which a protein belongs.

Unlike the application of phylogenetic profiling to genome phylogeny where we were interested in measuring the similarity of organisms based on their profile of gene families, here we wish to measure the similarity between the profiles of the families themselves. To accomplish this, we measure the co-occurrence or coabsence of pairs of protein families across genomes (see Figure 2). The underlying

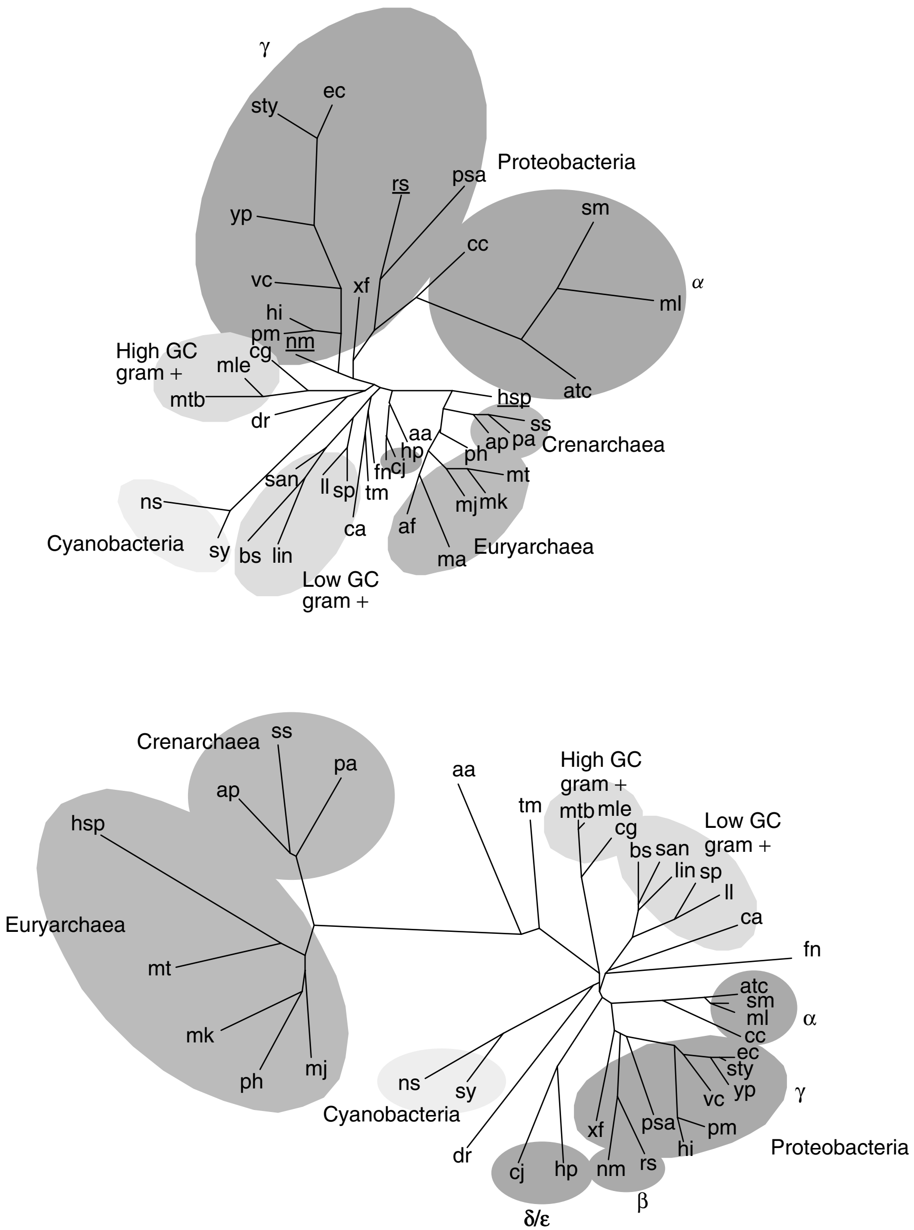


Figure 1 Phylogenetic trees of prokaryotes, based on gene content (upper tree; House, 2002), and small subunit ribosomal RNA sequence (lower tree), constructed using on-line analysis tools at the Ribosomal Database Project (**http://rdp.cme.msu.edu**) (Cole, 2003). A few notable discrepancies are shown in the gene content tree as underlined taxa



Figure 2 Clustered phylogenetic profiles of human HMBS (hydroxymethylbilane synthase), ALDH3 (aldehyde dehydrogenase), and FTHFD (formyltetrahydrofolate dehydrogenase) genes. The profiles are computed over 83 organisms shown on the top. Red indicates that a homolog of the human gene was found in the corresponding organism and black that it was not. The profiles have been clustered using hierarchical clustering (Eisen, 1998)

assumption of this method is that pairs of nonhomologous proteins that are present together in genomes, or absent together, are likely to have coevolved; that is, the organism is under evolutionary pressure to encode both or neither of the proteins within its genome and encoding just one of the proteins lowers its fitness.

It has been observed that coevolved protein families are likely to be members of the same pathway or complex (Huynen, 1998; Pellegrini, 1999). This is not surprising since it is more efficient for an organism to retain all or none of the subunits of a complex, or members of a pathway, since preserving only a fraction of these would not retain the function of the complex or pathway yet would entail their wasteful synthesis. Phylogenetic profiling has therefore emerged as a powerful method to group proteins together into cellular complexes and pathways.

Notice that protein families clustered on the basis of their phylogenetic profiles need not possess any sequence similarity. Therefore, phylogenetic profiling is able to determine functions for proteins families with no experimentally characterized members, thus going beyond the capabilities of conventional sequence similarity–based techniques.

4. Computing phylogenetic profiles

To compute phylogenetic profiles for each protein coded within a genome, one can use several approaches. One of these is to first define orthologous proteins across genomes. Orthologs are proteins that have descended from a common ancestor by way of speciation. Although the actual calculation of orthologs is not trivial, an estimate of groups of orthologous proteins has been compiled in the Clusters of Orthologous Groups (COG) database (Tatusov, 1997). Armed with these clusters, a profile may be trivially calculated by enumerating the organisms that are represented in each COG.

Another approach to establishing a phylogenetic profile is to identify homologs of a protein using a sequence alignment technique. Along these lines, a popular method is to define a homolog of a query protein to be present in a secondary genome if the alignment, using BLAST (Altschul, 1997), of the query protein

with any of the proteins encoded by the secondary genome generates a significant alignment. The result of this calculation across N genomes yields an N-dimensional phylogenetic profile of ones and zeroes for the query protein. At each position in the phylogenetic profile, the presence of a homolog in the corresponding genome is indicated with a 1 and its absence with a 0.

There is no need to restrict phylogenetic profiles to contain only entries of 1's and 0's. Various methods have been used in which the entries of the phylogenetic profile measure the similarity of two proteins. As an example, one method uses the inverse of the log of the E value from a BLAST search as the similarity metric (Date, 2003).

5. Estimating the probability of coevolution

Once the phylogenetic profiles have been computed, one needs to determine the likelihood that two proteins have coevolved on the basis of the similarity of their profiles. A variety of techniques have been reported to compute these probabilities. Here, we briefly review a few of them.

The first approach is the computation of the similarity between two phylogenetic profiles using the Hamming distance (Pellegrini, 1999). The Hamming distance is the number of bits that differ between the two profiles. Although this is a simple measure to compute, it is limited by not providing a probability estimate of observing this distance.

It is possible to obtain such an estimate of the probability that two proteins coevolve by using the hypergeometric distribution. If we assume that the two proteins A and B do not coevolve, we can compute the probability of observing a specific overlap between their two profiles by chance by using the hypergeometric distribution:

$$P(k'|n, m, N) = \frac{\binom{n}{k'}\binom{N-n}{m-k'}}{\binom{N}{m}} \tag{1}$$

where N represents the total number of genomes analyzed, n the number of homologs for protein A, m the number of homologs for protein B, and k' the number of genomes that contain homologs of both A and B (Wu, 2003). Because P represents the probability that the proteins do not coevolve, $1 - P(k > k')$ is then the probability that they do coevolve.

A similar approach attempts to compute the likelihood of coevolution using the mutual information between two phylogenetic profiles (Date, 2003; Wu, 2003):

$$MI(A, B) = H(A) + H(B) - H(A, B) \tag{2}$$

where

$$H(A) = -\sum p(a) \ln p(a) \tag{3}$$

and

$$H(A, B) = -\sum p(a, b) \ln p(a, b) \tag{4}$$

Here, the sums are over the possible states that the profiles can assume. If two profiles are identical, their mutual information is zero. Dissimilar profiles have positive mutual information scores. One advantage of the mutual information approach is that it can be applied to nonbinary phylogenetic profiles, whereas the hypergeometric function cannot.

6. Recovery of pathways and complexes

Protein pairs that coevolve are likely under some evolutionary pressure because their functions are coupled: preserving one without the other disables their combined function. This scenario may occur if the proteins are subunits of cellular complexes or components of pathways.

It is possible to test this hypothesis starting from pathway annotation. Several databases have been developed that through extensive manual curation have categorized proteins into pathways (Tatusov, 2003; Kanehisa, 2004; Camon, 2004).

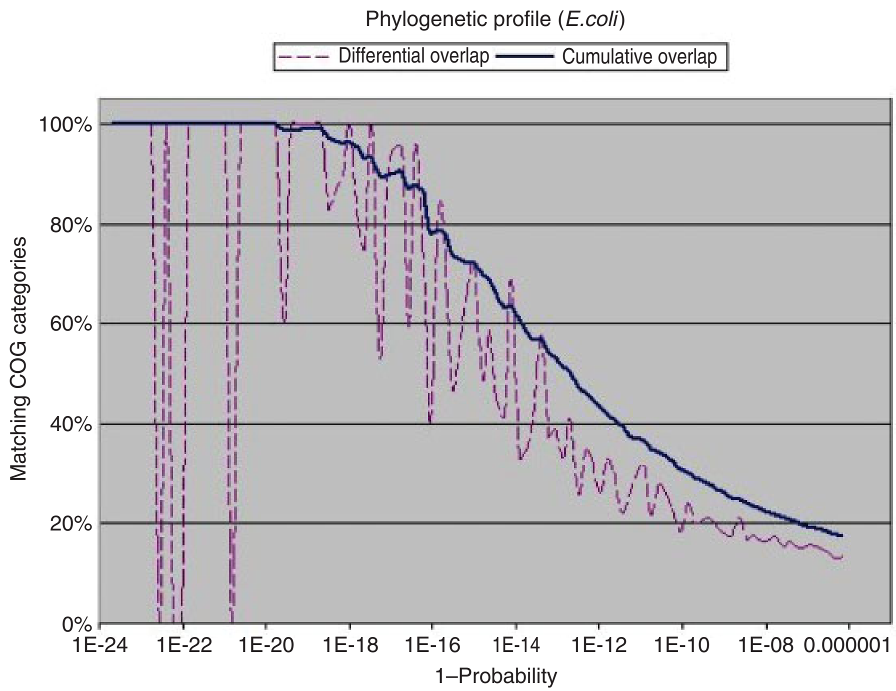


Figure 3 The probability that two genes have coevolved as a function of their likelihood to belong to the same pathway. The probability is computed using the hypergeometric function (see text). The pathways are obtained from the COG databases (Tatusov, 2003). Pairs of genes with significant *P*-values (on left) are nearly always found to belong to the same pathway

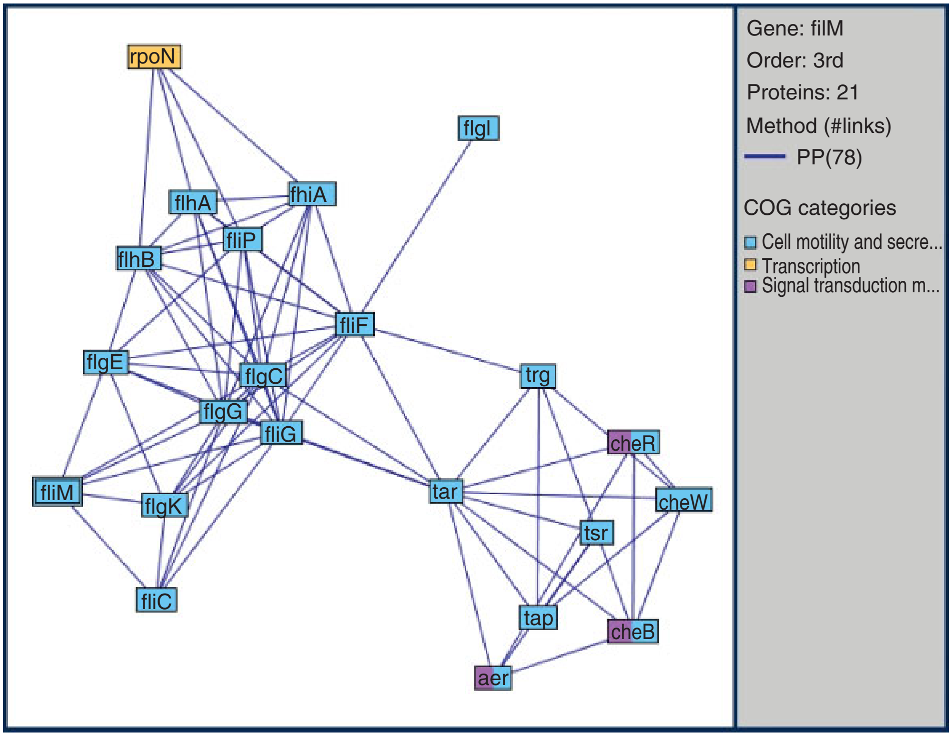


Figure 4 Clusters of *Escherichia coli* proteins that are predicted to coevolve by the phylogenetic profile analysis and that form a large network. The network shows a cluster of proteins (flg and flh genes) that are components of the bacterial flagella. A second cluster includes components of the chemotaxis pathway (che genes). These two clusters are linked to each other, indicating that flagellar and chemotaxis clusters have coevolved in bacteria

In Figure 3, we show that proteins that are likely to have coevolved (have significant *P*-values) are likely to belong to the same pathway (using the COG pathway definitions, Tatusov, 2003). In fact, we find that protein pairs with significant *P*-values nearly always belong to the same pathway. A similar curve could also be constructed using protein complexes instead of pathways, yielding similar results (Bowers, 2004).

By combining all pairs of coevolving proteins with significant *P*-values, we can generate a vast network. This is because if protein A is found to coevolve with B, and is thus said to be functionally linked to B, B may then be linked to C, C to D, and so forth. By examining clustered groups of proteins within this network, one can identify the protein components of pathways and complexes (Strong, 2003; Von Mering, 2003). An example of such a network is shown in Figure 4. Here, we see that many of the components of the flagella form a cluster, as do the components of the chemotaxis pathway. Furthermore, the network also illuminates the fact that these two clusters are coevolving. This is not surprising given the intimately coupled function of flagella and chemotaxis within the cell.

7. Phenotype profiling

We have discussed the use of phylogenetic profiling to study the evolution of genomes and to study the coevolution of encoded proteins, yielding functional clusters and networks of clusters. A third application we review is the linking of genes to phenotypes (Jim, 2004; Levesque, *et al*., 2003).

Each of the fully sequenced organisms that is used to construct phylogenetic profiles of a gene has specific phenotypes. A phenotype is any observable characteristic of the organism. Examples of phenotypes include flagella, pili, and thermosensitivity. It is possible to construct a phenotypic profile by cataloging the presence or absence of the phenotype across genomes, just as we have done for the presence or absence of genes.

By identifying the genes whose phylogenetic profiles are correlated with the phenotypic profiles, it is possible to associate a gene with the phenotype. For instance, about half of the fully sequenced organisms contain flagella. The genes whose phylogenetic profiles are correlated with a flagella profile are nearly all known components of the bacterial flagella (Levesque, 2003; Jim, 2004). The same approach may also be used to identify the components of pili, and the proteins that endow organisms with thermostability (Jim, 2004). In general, if a reliable phenotypic profile can be constructed for a trait that is found in a significant fraction of the sequenced genomes, this technique can identify the proteins that are most likely responsible for the trait.

8. Conclusions

The availability of fully sequenced genomes has enabled us to perform phylogenetic profiling by identifying the distribution of protein families across organisms. As we have discussed in this review, phylogenetic profiling may be used to study the evolution of genomes, the coevolution of proteins or the association between proteins and phenotypes.

Today, we have access to about 100 fully sequenced genomes. However, it is reasonable to assume that within the next decade this number will grow by orders of magnitude. As the data become available, phylogenetic profiling will become far more powerful than it is today. As a result, phylogenetic profiling will undoubtedly continue to expand our understanding of genome evolution and protein function.

Acknowledgments

We thank DOE, NIH, and Howard Hughes Medical Institute for support.

Further reading

Gaasterland T and Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microbial & Comparative Genomics*, **3**(4), 199–217.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Bansal AK and Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. *Journal of Bacteriology*, **184**(8), 2260–2272.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO and Eisenberg D (2004) ProLinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, **R35**, Epub 2004 Apr. 16.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R and Apweiler R (2004) The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Research*, **32**(1), D262–D266.

Clarke GD, Beiko RG, Ragan MA and Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *Journal of Bacteriology*, **184**(8), 2072–2080.

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, *et al*. (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, **31**(1), 442–443.

Date SV and Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**(9), 1055–1062, Epub 2003 August 17.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster Analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Fitch WM and Margoliash E (1967) Construction of phylogenetic trees. *Science*, **155**(760), 279–284.

Fitz-Gibbon ST and House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, **27**(21), 4218–4222.

House CH and Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Journal of Molecular Evolution*, **54**(4), 539–547.

Huynen MA and Bork P (1998) Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(11), 5849–5856.

Jim K, Parmar K, Singh M and Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Research*, **14**(1), 109–115.

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**(1), D277–D280.

Levesque M, Shasha D, Kim W, Surette MG and Benfey PN (2003) Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Current Biology*, **13**(2), 129–133.

Li W, Fang W, Ling L, Wang J, Xuan Z and Chen R (2002) Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biology Physics*, **28**, 439–447.

Lin J and Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, **10**, 808–818.

Montague MG and Hutchison CA (2000) Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(10), 5334–5339.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), 4285–4288.

Snel B, Bork P and Huynen MA (1999) Genome phylogeny based on gene content. *Nature Genetics*, **21**(1), 108–110.

Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D (2003) Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome wide functional linkage maps. *Nucleic Acids Research*, **31**(24), 7099–7109.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**(5338), 631–637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(1), 41.

Tekaia F, Lazcano A and Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Research*, **9**, 550–557.

Von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA and Bork P (2003) Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26), 15428–15433, Epub 2003 December 12.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL and Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, **1**, 8.

Wu J, Kasif S and DeLisi C (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**(12), 1524–1530.

Specialist Review

Reconstructing vertebrate phylogenetic trees

Sudhir Kumar and Alan Filipski
Arizona State University, Tempe, AZ, USA

1. Introduction

The first major scientific figure to challenge the idea of fixity of species was Jean-Baptiste Lamarck (1744–1829), the French naturalist who in 1802 coined the term *biology*. Although his name is now most commonly associated with erroneous ideas about inheritance of acquired characters, his major contribution to evolutionary thought was his promotion of the very concept of adaptive change in lineages over long periods of time. Figure 1, from Lamarck (1809) in his *Philosophie Zoologique*, depicts a portion of the first published evolutionary tree based on a hypothesized order of acquisition of a few key characters. For example, the tree showed monotremes deriving from birds because of their egg-laying characteristic, while the two classical groups of terrestrial mammals – hoofed (e.g., cow and horse) and nailed (e.g., human and mouse) – both derived from amphibious mammals as an intermediate form. Although the depiction has some problems, as showing extant groups as ancestral to each other, and incorporating Aristotle's idea of a linear progression from simple to complex, yet it is an actual evolutionary tree as stated in its original caption "to show the origin of the different animals."

Ernst Haeckel (1834–1919), inspired by Darwin, Goethe, and Lamarck, whom he called the "three fathers of the theory of descent", coined the word *phylogeny* in his *Generelle Morphologie der Organismen* of 1866 (Haeckel, 1866) and drew hundreds of trees depicting the evolutionary relationship of organisms. Haeckel constructed them largely on the basis of his theory of progress, as revealed by his biogenetic law ("ontogeny recapitulates phylogeny") with Germanic man at the pinnacle and other lineages represented as radiating side branches. (Haeckel, 1900; Dayrat, 2003).

With the introduction of phenetics and cladistics in the twentieth century, phylogenetic analyses grew more sophisticated, with investigators abandoning the idea of man as the summit of progress in favor of the Darwinian idea of adaptive divergence along all lineages, and using a larger number of more informative characters. In particular, characters relating to the skeleton and teeth were of great value in vertebrate phylogenetics, because they are easily fossilized and straightforwardly compared in living specimens. Using shared derived characters, many monophyletic groups

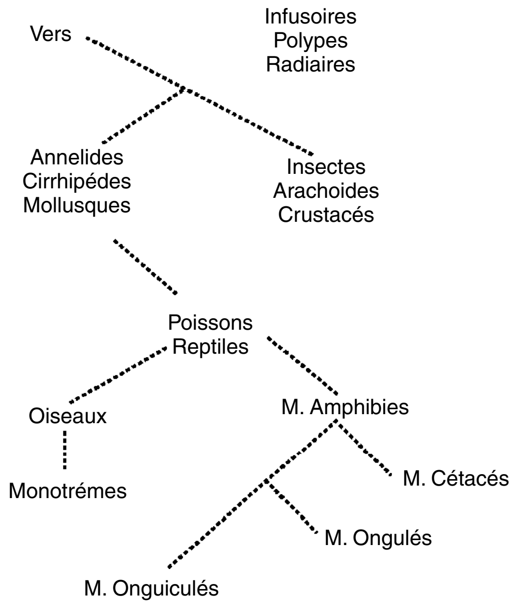


Figure 1 The first tree showing evolutionary relationships of animals, published by Lamarck (1809), including, among the vertebrates: Fish (Poissons), monotremes (Monotrèmes) birds (Oiseaux), amphibious mammals (M. Amphibies), whales (M. Cétacés), hoofed mammals (M. Ongulés), and nailed/clawed mammals (M. Onguiculés)

(groups of taxa that include all descendants of their most recent common ancestor) that are still in use today were delineated. For example, the group Tetrapoda (four-legged animals) was defined to include amphibians (e.g., frogs) and amniotes (e.g., birds and mammals) based on their many shared skeletal characteristics, such as dactyly, sacrum, stapes, and *fenestra ovalis* (Panchen and Smithson, 1988). By making successive nested groupings in this way, full phylogenetic trees have been inferred. Indeed, as data matrices have become larger, investigators have taken more algorithmic and computational approaches and increased their use of the principle of maximum parsimony.

Phylogenies based on morphological characters have clearly resolved many basic vertebrate relationships. However, some groupings have been less convincing. For example, Cetaceans (whales and dolphins), even-toed ungulates (artiodactyls), and odd-toed ungulates (perissodactyls) form monophyletic groups based on morphology, but morphological analyses did not clearly resolve their interrelationship (Novacek, Wyss *et al*., 1988). This is because in certain characters cetaceans appear to have been derived from even-toed ungulates and in others from odd-toed ungulates.

There also has been great interest in establishing a timescale of major events in vertebrate history. Although the vertebrate fossil record is the best among all known animal groups, the assignment of individual fossils into the tree based on living species is complicated and the divergence times based on the fossil record are almost always smaller than the actual time. This is because an unknown amount

of time must pass before discernibly different forms evolve, and there is a low probability that investigators will discover the earliest fossils after a divergence (e.g., Novacek, 1992; Hedges, Parker *et al*., 1996; Tavare, Marshall *et al*., 2002). For instance, fossils of placental mammals are found throughout the Cretaceous era (65–144 Mya), but identifiable members of modern orders (e.g., rodents, primates, and artiodactyls) older than the Palaeocene epoch (56–65 Mya) were found only recently (e.g., Archibald, 1996). So, although we can put a lower bound on the divergence time of these orders sometime during the Palaeocene, we cannot be sure if they existed at an earlier time (see, however, Tavare, Marshall *et al*., 2002).

While in the last decade investigations aimed at finding the relationships among species on the basis of morphological, physiological, behavioral, and other nonmolecular characters have continued to refine the vertebrate tree of life, many specific questions have remained controversial, including the interrelationship of mammalian orders. To resolve these points, molecular phylogenetics is proving invaluable (e.g., Nei and Kumar, 2000; Felsenstein, 2003).

2. The age of molecular phylogenetics

In the last decades of the twentieth century, a vast quantity of DNA and protein sequence data became publicly available, offering almost limitless opportunities for inferring phylogenetic relationships and temporal patterns of vertebrate evolution. In fact, the technological advances made in sequencing genomes, the development of sophisticated statistical and computational methods of phylogeny reconstruction, and the success of individual phylogenetic investigations in resolving long-standing issues led the National Science Foundation (USA) to initiate the Tree of Life (ToL) project to establish a phylogenetic framework (Darwin's "Great Tree of Life") for the approximately 1.75 million described species of organisms. Many groups of vertebrates are now the targets of large-scale investigations in this most promising project of modern evolutionary biology.

However, molecular data do not provide a magic lens for clarifying all phylogenetic questions, even when data are available for a large number of genes or even complete genomes. Many biological and methodological complexities affect the accuracy of the inferred phylogenies and the times of species divergence (reviewed in Nei and Kumar, 2000; Benton and Ayala, 2003; Felsenstein, 2003; Hedges and Kumar, 2003). In this chapter, we briefly discuss the techniques and data used to resolve disputed vertebrate relationships and to confirm or challenge the evolutionary trees inferred from nonmolecular data.

Figure 2 shows one view of the modern vertebrate phylogeny, with an emphasis on the evolutionary relationships of major vertebrate groups and mammalian taxa. In it, the hagfishes (Myxionoidea) and lampreys (Petromyzontoidea) form a monophyletic group, (Cyclostomata), that is an outgroup to all other vertebrates (jawed vertebrates or Gnathostomes). However, the exact topology of this deepest part of the vertebrate tree was controversial until recently. Traditionally, nonmolecular data supported the idea that Cyclostomata was a monophyletic group, but morphological analyses in the late 1980s suggested that lampreys were the closest relatives of jawed vertebrates. Then, in the early 1990s, molecular

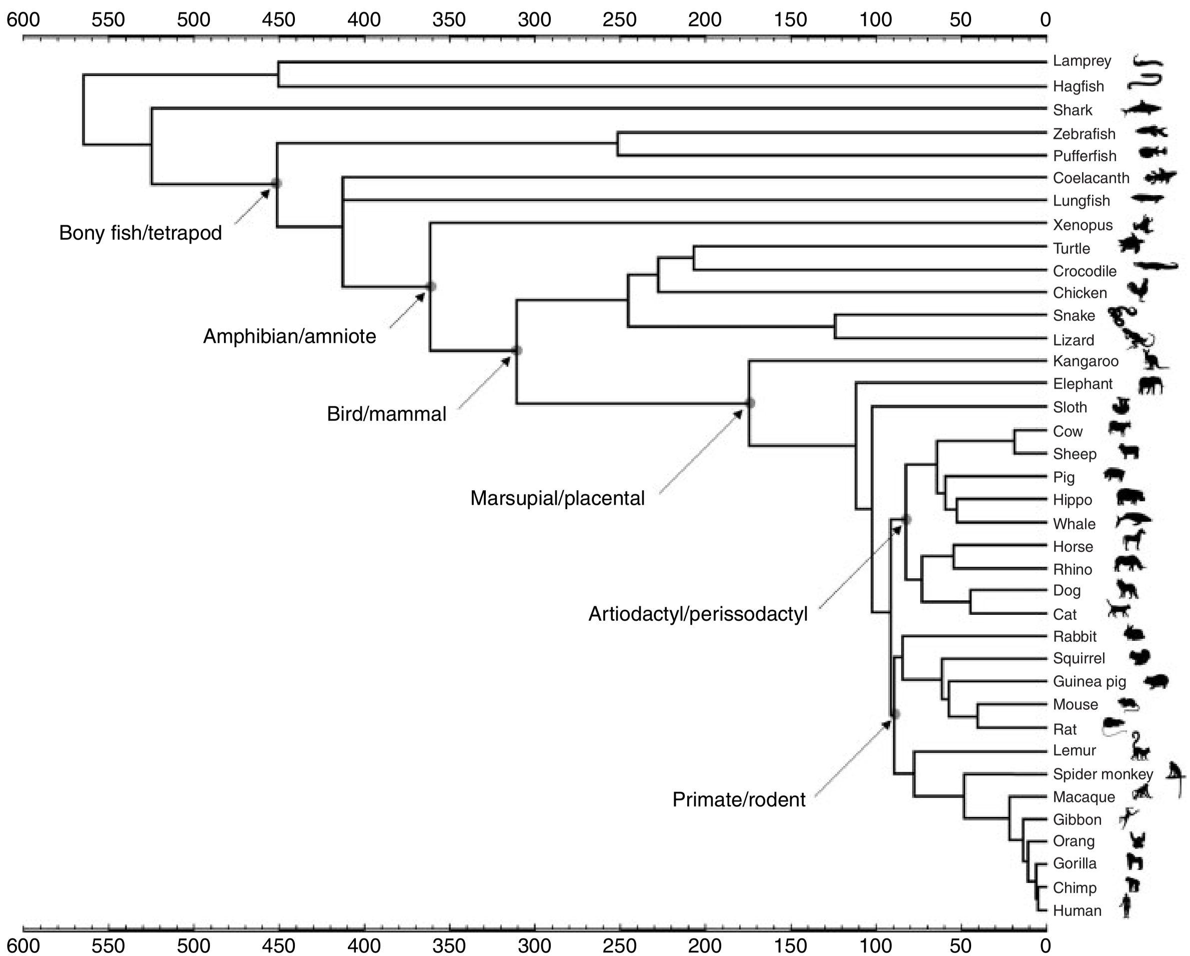


Figure 2 The current understanding of evolutionary relationship and timescale of major classes of vertebrate and groups of mammals based on fossil record and molecular data

phylogenies inferred using a single gene (18 S ribosomal RNA) again supported the traditional view of a monophyletic Cyclostomata (Stock and Whitt, 1992), although this view was not immediately accepted due, in part, to its use of only a single gene. Recent analyses of many nuclear proteins and complete mitochondrial DNA (mtDNA) have reaffirmed the monophyly of Cyclostomata (Hedges, 2001; Delarbre, Gallut *et al*., 2002; Takezaki, Figueroa *et al*., 2003).

While the cyclostome phylogeny based on the mtDNA genome agreed with that established from nuclear proteins (Delarbre, Gallut *et al*., 2002), the use of these genomes has led to suggestions that the classical position of the cartilaginous fishes (sharks and their kin) be revised (Rasmussen and Arnason, 1999). For compelling morphological reasons, the cartilaginous fishes traditionally have been placed in a basal position with respect to the rest of the jawed vertebrates (including bony fish and tetrapods). But early mtDNA phylogenies placed sharks, with significant statistical support, as a sister group to bony fish. However, a subsequent analysis of many nuclear proteins reversed the mtDNA phylogeny (Cotton and Page, 2002; Takezaki, Figueroa *et al*., 2003). This raises questions about the appropriateness of mtDNA for inferring the deepest branching patterns in the vertebrate tree. These results underscore the need for genomic data for more organisms, a more extensive analysis of statistical support, and a consideration of sources of error and biases that prevent robust molecular phylogenetic inference.

The next major divergence of vertebrates concerns the relationship of tetrapods with lungfish and coelacanth (Figure 2). The traditional view is that coelacanth and tetrapods cluster together. This is based on the inclusion of coelacanths and certain extinct fish into the group Crossopterygii, whose members share several important characters with certain early amphibians, such as limb bone patterns and labyrinthodont teeth. However, there is also some morphological support for a close association of lungfish and tetrapods. Molecular phylogenies have found support for all three configurations, coelacanth with lungfish, tetrapod with lungfish, and tetrapod with coelacanth. (Hedges, Hass *et al*., 1993; Zardoya and Meyer, 1996; Zardoya, Cao *et al*., 1998; Venkatesh, Erdmann *et al*., 2001). This trichotomy remains unresolved and will probably require much more molecular data for a satisfactory resolution. This shows that it is not easy, using either traditional or molecular methods, to resolve events that took place more than 360 million years ago within a narrow window of perhaps 30 million years. Morphological homology can become confused with homoplasy and sequence similarities can become effaced with time, especially if subsequent speciation events occur over a short time period.

The classification of mammals has occupied naturalists since the time of Aristotle. Although early taxonomists had a fairly good idea of how to partition the placental mammals into orders (rodents, primates, etc.) and superorders (e.g., Gregory, 1910), an understanding of the interordinal relationships among placental mammals and even the relation of the three main groups of mammals (placentals, marsupials, and monotremes) has remained elusive. The relation of the scuirognaths (e.g., mouse, squirrel), hystricognaths (e.g., guinea pig, porcupine), and lagomorphs (e.g., rabbit, hare) has been extensively debated for decades, recently with the publication of articles such as "Is the Guinea pig a rodent" and "What, if anything, is a rabbit?" (Graur, Hide *et al*., 1991; Graur, Duret *et al*., 1996). In both of these instances, molecular data overturned the classical relationships (see Liu and

Miyamoto, 1999), yet the use of different methods of phylogenetic inference and larger, more taxonomically dense datasets now supports the classical phylogenies (Murphy, Eizirik *et al*., 2001). Modern analyses also have placed rodents and primates as sister groups, a departure from the more basal position of rodents in the placental mammal phylogenies of the 1990s. Still, all superordinal mammalian divergences are not resolved with high statistical confidence (Scally, Madsen *et al*., 2001; Springer and de Jong, 2001).

The possibility that higher-level placental divergences occurred within a very short time period (<5 Myr) has suggested the use of alternative molecular markers to bring more data to bear on the problem (e.g., Shedlock and Okada, 2000). For instance, initial gene sequence analyses placed hippos as a sister group to cetaceans with the combined group appearing as a sister lineage to the ruminants within Artiodactyla (Gatesy, 1997; Gatesy, Milinkovitch *et al*., 1999). However, the widespread acceptance of this scheme came about only after a cladistic analysis of SINE elements, the short sequences of 80–400 nucleotides that appear as repeats in the genome (Shimamura, Yasue *et al*., 1997; Nikaido, Rooney *et al*., 1999). Recent fossil discoveries support this reevaluation of the relationships of hippos, cetaceans, and artiodactyls (Thewissen and Madar, 1999; Thewissen, Williams *et al*., 2001).

A useful feature of molecular phylogenetics is its ability to assign dates to speciation events. Although the rates of molecular evolution vary greatly for different kinds of sequence data and for genes under different selective constraints, it is possible to identify and test data sets for a constant rate of evolution (molecular clock property), and sequences that deviate from this standard may be excluded. If we obtain one or more calibration points – divergence times in a tree whose date is confidently known – then the molecular evolution rate can be established and all divergence events in that tree can be dated. In addition, sophisticated methods of incorporating complex models of DNA and protein evolution and distributions of evolutionary rates along lineages are now available (see review and references in Hedges and Kumar, 2003). These molecular timescales complement those based on the fossil record and many molecular-based estimates show excellent correspondence with fossil-based estimates (Kumar and Hedges, 1998); however, others show significant differences, with molecular time estimates being considerably larger in some cases. For instance, initial molecular clock studies suggest that major orders of mammals diverged more than 90 Mya (Hedges, Parker *et al*., 1996; Kumar and Hedges, 1998) and that continental breakup, rather than the dinosaur extinction's vacating of ecological niches for mammalian radiation, may be the primary mechanism. Such estimates have been confirmed in subsequent analyses, which have used different datasets and statistical methodologies (Springer, Cleven *et al*., 1997; Hasegawa, Thorne *et al*., 2003). On the other hand, the discrepancy between the molecular evolutionary timescales within rodents (e.g., mouse–rat divergence), which indicate a much older divergence than suggested by the fossil record (>30 Mya versus ~12 Mya) has recently been revised, bringing both molecular divergence estimates much closer to the fossil record (Easteal, 1990; Kumar and Hedges, 1998; Springer, Murphy *et al*., 2003; Hedges and Kumar, 2004). Thus, more extensive molecular clock analyses are in some cases confirming early divergence estimates based on small data sets, and, in other cases, reconfirming fossil-based estimates.

While the value of using molecular data to discern evolutionary relationships and timescales is clear, the design of the analysis is important. Some guiding principles that can be useful in generating more robust results, based on our personal experience, are as follows. First, it is necessary to use as much data as possible. It is particularly important to use all possible genes for which data are available for all taxa (see results and references in Rosenberg and Kumar, 2001a,b; Hillis, Pollock *et al*., 2003; Rosenberg and Kumar, 2003). It is also important to use a large number of species per group, especially when there is a possibility of long-branch attraction (Felsenstein, 1978) or of evolutionary rate speed-up in some lineages (Springer, Murphy *et al*., 2003). To accomplish this, multiple genes or other homologous regions can be combined by simple concatenation (e.g., Rokas, Williams *et al*., 2003), although it is then necessary to account for rate variation among sites and to use gene-specific nucleotide and protein substitution matrices (e.g., Nei, Xu *et al*., 2001; Springer, Murphy *et al*., 2003). Second, only orthologous gene sequences can be used to represent species, that is, they must derive from speciation events and not gene duplications. This is crucial for both molecular phylogeny as well as time estimation. Third, it is important to realize that there are no invincible methods for phylogenetic inference (Nei and Kumar, 2000). Generally, it is important to try several applicable methods and look for congruence (although congruence should not be taken as a necessary predictor of accuracy of the inferred phylogenies, but used to pursue method-specific biases that might be responsible for some observed branching patterns). Such efforts are, however, often hampered by the number of species in the datasets, because very large scale analyses are cost prohibitive for methods other than the neighbor-joining method (Saitou and Nei, 1987; Tamura, Nei *et al*., 2004) or fast heuristic searches under maximum likelihood and parsimony. Fourth, when examining the robustness of the inferred phylogeny, it is important to use approaches such as the bootstrap method to generate conservative estimates of statistical support (reviewed in Nei and Kumar, 2000; Felsenstein, 2003). Fifth, the timing of phylogenetic events should be determined by using one or more highly constrained and reliable reference times for calibrating molecular clocks. Also, a variety of statistical methods should be used to account for rate variation among lineages, either eliminating them by using relative rate tests (Wu and Li, 1985; Takezaki, Rzhetsky *et al*., 1995; Kumar and Hedges, 1998) or by explicitly modeling lineage-specific rate variations (e.g., Thorne and Kishino, 2002; Hasegawa, Thorne *et al*., 2003).

Molecular sequences provide a vast new data stream for vertebrate phylogenetics. These data complement and provide reciprocal validation for analyses based on morphological data. Although morphological data are more easily available for long-extinct taxa, molecular sequence analyses can provide a new view of divergence times and can assist with the resolution of some important questions. Indeed, phylogenetic analysis based on molecular sequence data seems poised to address many of the long-standing and difficult problems of vertebrate phylogenetics. At the same time, molecular phylogenetics is becoming increasingly important in a broader sense as it informs biological application areas such as medicine, animal husbandry, and conservation biology. We now have complete sequenced genomes for human, mouse, dog, rat, chimpanzee, chicken, and several fish, with many more vertebrates in the pipeline (see review in Filipski and Kumar, 2004). Although we

have identified and located many genes in these species, our understanding of the location and nature of the critical control regions in these genomes that direct gene interaction and regulate expression and development is still quite limited. For such purposes, comparative genomics from an evolutionary perspective is essential (Hedges and Kumar, 2002; Thomas and Touchman, 2002). An important factor in deciding which genomes to sequence next is the question of which ones will best contribute to our understanding of the function and evolution of the human genome (Thomas, Touchman *et al*., 2003). Reciprocally, as more genomes are mapped, knowledge of the function of homologous areas in related genomes can be applied, for example, to animal husbandry and the identification of genes of agricultural importance. Recently, for example, sequencing project for the cow genome has begun (Lewin, 2003). Besides the obvious agricultural applications relating to milk and beef production, this genome will provide a useful comparison to the human genome, as a relatively closely related mammalian relative. Therefore, as more molecular sequences and the tools to analyze them become available, the associated improved understanding of evolutionary relationships will continue to help guide both pure and applied biological research.

References

Archibald JD (1996) Fossil evidence for a late cretaceous origin of “Hoofed” mammals. *Science*, **272**(5265), 1150–1153.

Benton MJ and Ayala FJ (2003) Dating the tree of life. *Science*, **300**(5626), 1698–1700.

Cotton JA and Page RDM (2002) Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proceedings: Biological Sciences*, **269**(1500), 1555–1561.

Dayrat B (2003) The roots of phylogeny: how did Haeckel build his trees? *Systematic Biology*, **52**(4), 515–527.

Delarbre C, Gallut C, Barriel V, Janvier P and Gachelin G (2002) Complete mitochondrial DNA of the hagfish, Eptatretus burgeri: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. *Molecular Phylogenetics and Evolution*, **22**(2), 184–192.

Easteal S (1990) The pattern of mammalian evolution and the relative rate of molecular evolution. *Genetics*, **124**(1), 165–173.

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**(4), 401–410.

Felsenstein J (2003) *Inferring Phylogeny*, Sinauer Associates: Sunderland.

Filipski A and Kumar S (2004) Comparative genomics of eukaryotes. *The Evolution of the Genome*, Ryan TG (Ed.), Academic Press: New York.

Gatesy J (1997) More DNA support for a cetacea/hippopotamidae clade: the blood-clotting protein gene gamma-fibrinogen. *Molecular Biology and Evolution*, **14**(5), 537–543.

Gatesy J, Milinkovitch M, Waddell V and Stanhope M (1999) Stability of cladistic relationships between cetacea and higher-level Artiodactyl taxa. *Systematic Biology*, **48**(1), 6–20.

Graur D, Duret L and Guoy M (1996) Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). *Nature*, **379**(6563), 333–335.

Graur D, Hide WA and Li WH (1991) Is the Guinea-Pig a rodent. *Nature*, **351**(6328), 649–652.

Gregory WK (1910) The orders of mammals. *Bulletin of the American Museum of Natural History*, **27**, 1–524.

Haeckel EHPA (1866) *Generelle morphologie der organismen. Allgemeine grundzèuge der organischen formen-wissenschaft, mechanisch begrèundet durch die von Charles Darwin reformirte descendenztheorie*, Reimer G: Berlin.

Haeckel EHPA (1900) *The Evolution of Man: a Popular Exposition of the Principal Points of Human Ontogeny and Phylogeny*, The Werner Company: Akron.

Hasegawa M, Thorne JL and Kishino H (2003) Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems*, **78**(4), 267–283.

Hedges SB (2001) Molecular evidence for the early history of living vertebrates: paleontology, phylogney, genetics, and development. *Major Events in Early Vertebrate Evolution*, Ahlberg PE (Ed.), Taylor & Francis: London.

Hedges SB Hass CA and Maxson LR (1993) Relations of fish and tetrapods. *Nature*, **363**(6429), 501–502.

Hedges SB and Kumar S (2002) Genomics. vertebrate genomes compared. *Science*, **297**(5585), 1283–1285.

Hedges SB and Kumar S (2003) Genomic clocks and evolutionary timescales. *Trends in Genetics*, **19**(4), 200–206.

Hedges SB and Kumar S (2004) Precision of molecular time estimates. *Trends in Genetics*, **20**(5), 242–247.

Hedges SB, Parker PH, Sibley CG and Kumar S (1996) Continental breakup and the ordinal diversification of birds and mammals. *Nature*, **381**(6579), 226–229.

Hillis DM, Pollock DD, McGuire JA and Zwickl DJ (2003) Is sparse taxon sampling a problem for phylogenetic inference? *Systematic Biology*, **52**(1), 124–126.

Kumar S and Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature*, **392**(6679), 917–920.

Lamarck JB and de Monet de PA (1809) *Philosophie Zoologique, ou, Exposition des Considerations Relative ·L'histoire Naturelle des Animaux*, Chez Dentu [et] L'Auteur: Paris.

Lewin HA (2003) The future of cattle genome research: the beef is here. *Cytogenetic and Genome Research*, **102**(1–4), 10–15.

Liu FG and Miyamoto MM (1999) Phylogenetic assessment of molecular and morphological data for eutherian mammals. *Systematic Biology*, **48**(1), 54–64.

Murphy WJ, Eizirik E, O′Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ and de Jong WW (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**(5550), 2348–2351.

Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press: New York.

Nei M, Xu P and Glazko G (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(5), 2497–2502.

Nikaido M, Rooney AP and Okada N (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(18), 10261–10266.

Novacek MJ (1992) Mammalian phylogeny: shaking the tree. *Nature*, **356**(6365), 121–125.

Novacek MJ, Wyss AR and McKenna MC (1988) The major groups of eutherian mammals. In *The Phylogeny and Classification of the Tetrapods*, *Vol. 2, Mammals*, Benton MJ (Ed.), Clarendon Press: Oxford, pp. 31–71.

Panchen AL and Smithson TR (1988) The relationships of the earliest tetrapods. In *The Phylogeny and Classification of the Tetrapods*, *Vol. 1, Amphibians, Reptiles, Birds*, Benton MJ (Ed.), Clarendon Press: Oxford, pp. 1–32.

Rasmussen AS and Arnason U (1999) Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes. *Journal of Molecular Evolution*, **48**(1), 118–123.

Rokas A, Williams BL, King N and Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**(6960), 798–804.

Rosenberg MS and Kumar S (2001a) Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(19), 10751–10756.

Rosenberg MS and Kumar S (2001b) Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Molecular Biology and Evolution*, **18**(9), 1823–1827.

Rosenberg MS and Kumar S (2003) Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology*, **52**(1), 119–124.

Saitou N and Nei M (1987) The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Scally M, Madsen O, Douady CJ, de Jong WW, Stanhope MJ and Spring MS (2001) Molecular evidence for the major clades of placental mammals. *Journal of Mammalian Evolution*, **8**, 239–277.

Shedlock AM and Okada N (2000) SINE insertions: powerful tools for molecular systematics. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **22**(2), 148–160.

Shimamura M, Yasue H, Oshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I and Okada N (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature*, **388**(6643), 666–670.

Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM and Stanhope MJ (1997) Endemic African mammals shake the phylogenetic tree. *Nature*, **388**(6637), 61–64.

Springer MS and de Jong WW (2001) Phylogenetics: which mammalian supertree to bark up? *Science*, **291**, 1709–1711.

Springer MS, Murphy WJ, Eizirik E and O′Brien SJ (2003) Placental mammal diversification and the Cretaceous-tertiary boundary. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(3), 1056–1061.

Stock DW and Whitt GS (1992) Evidence from 18 S ribosomal RNA sequences that lampreys and hagfishes form a natural group. *Science*, **257**(5071), 787–789.

Takezaki N, Figueroa F, Zaleska-Rutczynska Z and Klein J (2003) Molecular phylogeny of early vertebrates: monophyly of the agnathans as revealed by sequences of 35 genes. *Molecular Biology and Evolution*, **20**(2), 287–292.

Takezaki N, Rzhetsky A and Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution*, **12**(5), 823–833.

Tamura K, Nei M and Kumar S (2004) Prospects for inferring very large phylogenies using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11030–11035.

Tavare S, Marshall CR, Will O, Soligo C and Martin RD (2002) Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, **416**(6882), 726–729.

Thewissen JG and Madar SI (1999) Ankle morphology of the earliest cetaceans and its implications for the phylogenetic relations among ungulates. *Systematic Biology*, **48**(1), 21–30.

Thewissen JG, Williams EM, Roe LJ and Hussain ST (2001) Skeletons of terrestrial cetaceans and the relationship of whales to artiodactyls. *Nature*, **413**(6853), 277–281.

Thomas JW and Touchman JW (2002) Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends in Genetics*, **18**(2), 104–108.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**(6950), 788–793.

Thorne JL and Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*, **51**(5), 689–702.

Venkatesh B, Erdmann MV and Brenner S (2001) Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(20), 11382–11387.

Wu CI and Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the United States of America*, **82**(6), 1741–1745.

Zardoya R, Cao Y, Hasegawa M and Meyer A (1998) Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data. *Molecular Biology and Evolution*, **15**(5), 506–517.

Zardoya R and Meyer A (1996) Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Molecular Biology and Evolution*, **13**(7), 933–942.

Specialist Review

Evolution of regulatory networks

Erich Bornberg-Bauer and Amelie Veron
The Westphalian Wilhelm's University of Münster, Münster, Germany

1. Introduction

Recurrent development of similar morphological features in response to evolutionary adaptation are often seen as a prime challenge in explaining the evolution of biological complexity. One well-known example is the camera eye. In essence, it has a similar construction scheme in the octopus and in higher vertebrates but does not exist in the last common ancestor. Other examples are the repeated adaptation of fish, saurian, birds, and mammals to life in water resulting in similar streamline shape and fins, or the repeated development of wings from the frontal limbs in birds and bats. These events are often termed phenotypically convergent evolution and can be seen as an incarnation of Gould's metaphoric question of "What would be conserved if the tape were played twice" (Gould, 1989). Computer experiments at the level of an abstract chemistry (Fontana and Buss, 1994) and the simulation of primordial evolution in an RNA world (Fontana and Schuster, 1998) showed that similar complex structures may repeatedly arise under evolutionary pressure, when starting from completely different initial conditions.

However, organismal development is more difficult to understand since it depends on many conflicting fitness parameters. It is intimately related to genetic regulation since, for example, organisms or metabolic response require the concerted action of not just one but many regulatory proteins. *So, how then does genetic regulation evolve?*

In Eukaryotes, gene regulation happens predominantly at the level of transcription and the number of transcription factors increases from a few hundred in Bacteria such as *Escherichia coli* to well over 3000 in humans (Levine and Tjian, 2003; van Nimwegen, 2003). Consequently, the number of genetic circuits or regulatory networks that arise as a result of the combinatorics, also increases. Therefore, minor changes in single genes may well propagate along such networks and may have, in the end, quite drastic effects on gene expression in response to external stimuli and during development. But how do these changes occur, considering that the emergence of, for example, new organs requires the concerted interplay between many transcription factors and with their binding sites during development? Do networks evolve in big leaps, for example, by duplicating a whole module during a genome duplication? Do new transcription factors always emerge together with their up- and downstream binding sites?

Here we summarize briefly some of the most recent insights from genomics, proteomics, and transcriptomics studies on regulatory network evolution. We break the subject down into sections covering the evolution of transcription factors and their domains, binding sites, their coevolution and the graph-based "global" analysis. Owing to space limitations and the fuzziness of available information, the section on promoter evolution will be kept very brief. We close with some speculations that attempt to provide a coherent picture on how complex networks may have evolved. It can only be preliminary since the field is moving rapidly. Therefore, much caution is required in the interpretation of findings and current knowledge and technologies are just about to reveal a comprehensive picture.

2. Definitions

At a most basic, conceptual level, a regulatory network consists of three components (see also Figure 1):

1. the transcription factor (TF), or its gene itself;
2. the upstream-regulatory region on the DNA by which (through another transcription factor) transcription of the gene is mediated;

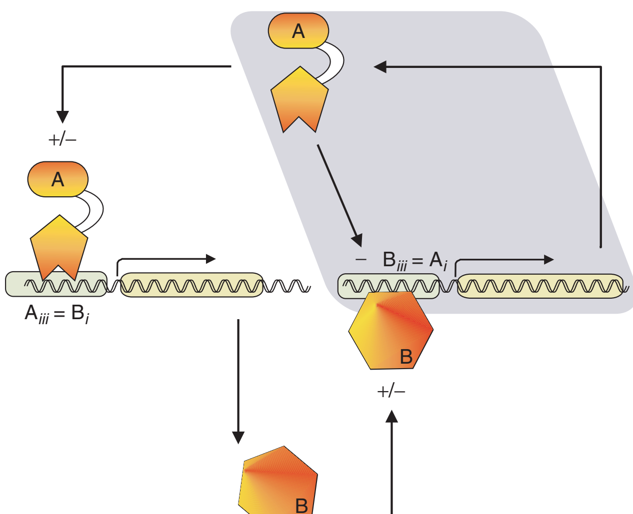


Figure 1 Schematic drawing of the basic construction principle of an elementary unit of a hypothetical genetic regulatory network: Transcription factor *A* (a two-domain protein) binds to its downstream binding site A_{iii}, which is also the upstream binding site B_i for transcription factor *B*, transcribed from gene *B*. *B* (orange hexagon) in turn acts on the binding site ($A_i = B_{iii}$). This gives a simple 2-component regulatory networks. It has a feedback structure with many possible variations since either binding can activate or repress transcription or both, depending on physiological conditions. When *A* also binds to its own upstream-regulatory region, for example, as a repressor blocking its own transcription and competing with *B* of binding to A_i, this results in a negative autoregulatory feedback loop. A single, autoregulatory negative feedback loop is shaded grey. The activity of *A* may be further regulated by stimuli, for example, via signaling pathways acting on the secondary regulatory domain

3. the regulatory region to which the transcription finally binds to, the downstream-binding region.

As we will detail later on, these links are mostly one-to-many, that is, each transcription factor regulates more than one gene and most genes are controlled by some, albeit generally few, transcription factors. Every transcription factor, with the exception of maternal transcription factors, is itself regulated by another one. This combinatorics gives rise to quite complex regulatory networks.

Each of these elements can be decomposed into subelements. For example, regulatory regions are often modular. The transcription factor itself typically contains a DNA-binding domain, and other regulatory domains. These are dimerization domains to form homo- or heterodimers and other protein interaction domains that mediate interaction with other proteins such as signaling proteins in order to react to physiological or environmental changes.

3. Evolution of transcription factors

3.1. Evolution of DNA-binding domains

Generally, the DNA-binding domains are among the most ancient domains, and they are derived from a relatively small set of folds (Perez-Rueda and Collado-Vides, 2000; Aravind and Koonin, 1999; Riechmann *et al*., 2000; Babu *et al*., 2004). These domains can thus be used to define protein families. They are differentially combined with other domains (Amoutzias *et al*., 2004a; Ledent and Vervoort, 2001; Morgenstern and Atchley, 1999). Examples are the families of bZIP proteins and NR (nuclear receptors), which comprise an ancient DNA-binding domain and some additional domains (see Table 1). Another, slightly more involved example is the family of bHLH transcription factors. There, the DNA-binding basic region is adjacent to the HLH domain. It is an ancient homo- and heterodimerizing domain and defines the family. The combination of additional domains (basic DNA-binding domain, leucine zipper, PAS or orange, see Figure 2) defines the monophyletic subgroups with distinct functions, many of which resemble organismal development and have emerged subsequently during evolution.

At least for most TFs in *E. coli* (Babu and Teichmann, 2003) and for bHLH, NR, and bZIP proteins (Amoutzias *et al*., 2004a,b), it was shown that their major means of proliferation was single-gene duplication of one ancestral protein, that is, they are monophyletic. Many TF families have spread over all major kingdoms but there are remarkable characteristics (see Table 1): some TFs appear mainly in Bacteria, like numerous Helix-Turn-Helix (HTH) families (Huffman and Brennan, 2002); some, although found in all metazoan, have particularly spread in some phyla such as the nuclear receptors, which are very frequent in *C. elegans* (Sluder *et al*., 1999) and Zn_2/Cys_6-type zinc cluster, which is restricted to fungal genomes (Akache *et al*., 2001). The MADS-box genes, only present in the Eukarya, were duplicated before the plant-animal divergence (Alvarez-Buylla *et al*., 2000). Both subsequent types of genes are found in all Eukarya, but in plants, the type II genes

Table 1 Occurrences of DNA-binding domains of transcription factors across taxa as defined in Pfam version 16.0 (Reproduced from Bateman *et al*. (2004) by permission of Oxford University Press)

Domain name	*E. coli*	*S. cerevisiae*	*C. elegans*	*D. melanogaster*	*H. sapiens*	*A. thaliana*
Homeodomain[a]	0	22	147	300	533	224
Nuclear receptors[b]	0	0	363	46	132	0
Zinc finger[c]	3	4	233	597	1574	512
Zinc cluster (Zn2-Cys6)	0	62	0	(1)	0	0
Basic leucine zipper[d]	0	16	37	46	102	121
MADS-box	0	4	3	6	6	171
Basic HLH	0	10	55	86	196	255
HTH[e]	131	2	2	22	9	4
wHTH[f]	9	9	44	66	159	36
P53	0	0	1	3	29	0
Runt	0	0	1	13	3	0
Paired box (PAX)	0	0	16	18	22	0
Myb-like DNA-BD	0	16	22	44	102	604

[a]Homeodomain: homeobox, POX, POU, TF_Oxt, HD_ZipN, OAR, PBX, PHD, ELK, CUT.
[b]Nuclear receptors: FTZ, Zf_C4, Androgen Oestrogen, Glucocorticoid, Progesterone.
[c]Zinc finger: GATA, CBFD_NFYB_HMF, zf-CCHC, zf-CXXC, C2 H2, zf-Dof, zf-MIZ, zf-NF-X1, zf-DBF.
[d]Basic leucine zipper: bZip_1, bZip_2 and bZip_Maf.
[e]Helix-Turn-Helix: iclR, DeoR, HTH_1,2,3,5,6,8, lacI, Crl, LexA DNA-BD, AbrB-like, gntR, MerR.
[f]Winged Helix-Turn-Helix: subfamily of the HTH (Huffman and Brennan, 2002), fork-head, ETS, MarR, E2F, LytTr, and La.

have particularly proliferated, giving rise to the flourishing MIKC family (Becker and Theissen, 2003).

Generally, the distribution over Bacteria is more even than over Eukarya. This has been interpreted as a result of frequent horizontal transfer (Lespinet *et al*., 2002; Coulson *et al*., 2001; Babu *et al*., 2004). In Eukarya, there exist around 12 to 15 families according to structural classification of the DNA-binding domain. The most frequent family in Eukarya are the zinc-fingers. They are strongly overrepresented in animals and plants.

3.2. *Evolution of dimerization domains*

Many transcription factors bind to DNA-binding sites as dimers. Notwithstanding the debate whether dimerization precedes DNA binding or vice versa, the resulting network is particularly interesting since it can be studied by phylogenetic trees and abundant protein-interaction data (Amoutzias *et al*., 2004a). In the family of bHLH, proteins members of the “D-group” act as modulators. They have a dimerization (HLH) domain but no DNA-binding domain. Therefore, they can act as repressors by deactivating the dimerization partner. Both from reasoning and from phylogenetic profiling, it was concluded that the ancestral proteins are probably the homodimerizing ones (Amoutzias *et al*., 2005). They have proliferated through series of single-gene duplications. This gave rise to a complex interaction network with the descendant of the ancestral proteins becoming the hubs.

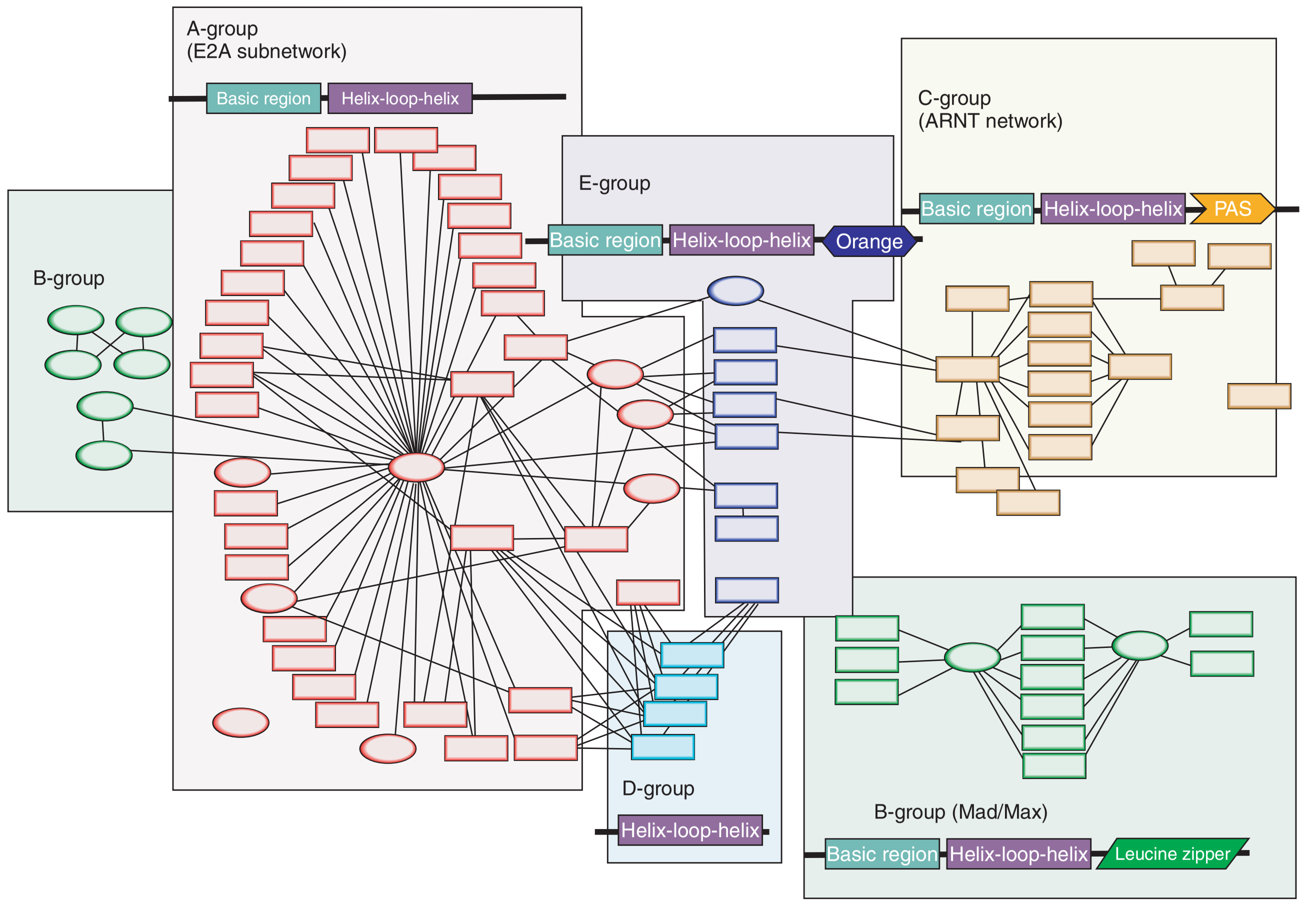


Figure 2 **Schema of the protein-interaction network of bHLH proteins**. Homodimerizing proteins are drawn as ellipses, heterodimerizing ones as rectangles, and heterodimerizing interactions as connecting lines. All data are based on multiple sources of information as described in Amoutzias *et al*., (2004a,b), Amoutzias *et al*. (2005) and available on **www.uni-muenster.de/Biologie.Botanik/ebb/html/research/AROB-EMBO-031210-supplementary-material.pdf**. Some proteins appear as hubs and, typically, they are descendants of ancient homodimerizing proteins. The corresponding domain architectures are given in corresponding boxes. Each group has evolved separately from an ancestral gene, primarily through series of gene duplication (Reproduced from Amoutzias *et al*. (2004a), Bornberg-Bauer *et al*. (2005) by permission of Birkhäuser Verlag AG. Copyright Birkhäuser Publishing Limited, Basel, Switzerland)

Evolution of dimerization domains allowed the MADS-box genes to interact with many different partners and achieve a specificity for the target gene and physiological conditions. A large proportion of animal and fungal MADS-box genes interact with phylogenetically unrelated partners such as homeodomain or zinc-finger proteins, while plant MADS-box genes form higher-order complexes, mainly with closely related transcription factors, that is, other MADS-box genes (Messenguy and Dubois, 2003). Multimerization of some plant MADS-domain proteins (termed MIKC-type) may have been facilitated by the acquisition of a second protein-interaction domain during evolution (Kaufmann *et al*., 2005).

3.3. Evolution of effector domains

Generally, effector domains are domains that relay a signal, for example, from a signaling pathway, to the DNA-binding property of the TF.

To the best of our knowledge, no studies so far have linked genetic networks and their regulation with other interactions such as signaling pathways, although many effector domains are protein-interaction domains. However, the aforementioned studies (Amoutzias *et al*., 2004a) have suggested that in the bHLH system, the loss (leucine zipper) and accretion (PAS, Orange) of domains has supported the structural differentiation of duplicated genes. As a consequence, these proteins lost their binding ability to their parental network and became free to differentiate into their own networks. In other words, the domain architecture mediates structural changes that help to form networks.

Nuclear receptors contain, apart from the DNA-binding domain (DBD), an effector domain that binds ligands (LBD) such as hormones as external signals. This domain binds, for example, steroids tightly and specifically in response to a hormone signal, which has migrated to the nucleus. A structural change in this LBD induces a structural change in the DBD, activating or deactivating its regulatory function (DNA binding). Lacking the signal, NRs are inert or repressors. According to phylogenetic analysis and structural considerations, some LBDs have probably lost their ligand binding ability during evolution (Schwabe and Teichmann, 2004).

3.4. Evolution of cofactors

Cofactors are, for example, ligands or other proteins that mediate transcriptional activity via interaction, either directly or indirectly, with the TFs. They confer time and gene specificity, and can be seen as the bearers of news from the outside world, bringing them to the transcription factors. They can modify the effector or the DNA-binding domains of their target transcription factors (Messenguy and Dubois, 2003) or trigger their migration from the cytoplasm to the nucleus (Kawana *et al*., 2003). Generally, only very few systematic results on this aspect have been disseminated, probably because of the tight involvement with metabolic pathways, signaling pathways and signaling molecules.

The cofactors could consequently play a role in linking different networks, such as the LIM proteins, which interact with bHLH and zinc-finger transcription factors,

protein kinases, and numerous other types of proteins (Matthews and Visvader, 2003).

Transcription factors, such as the above-mentioned D-group of the bHLH proteins, can also be viewed as cofactors because they block other bHLH proteins through dimerization (see Figure 2).

The function (activator or repressor) and strength of a transcription factor can change according the interacting partners, be they cofactors or other TFs. The information contained in these interactions is integrated by the modules composing the promoter regions of the target genes (Wray *et al*., 2003; Emberly *et al*., 2003).

4. Evolution of promoters

Unlike coding regions, the evolution of upstream-regulatory regions cannot be easily evaluated by means of sequence alignments and comparisons. Indeed, the structure, organization, and function of promoters differ widely from that of the coding sequences, resulting in totally different evolution rates and consequences of sequence variation (Rodriguez-Trelles *et al*., 2003). Generally, promoter organization underlies more variation than the organization of coding sequences and is more context dependent. It is influenced by DNA structure (Wang *et al*., 2004) and other epigenetic effects such as chromatin structure (Bode *et al*., 2003), competing binding events, neighborhood relations on the DNA, availability of the appropriate cofactors, and so forth (for a comprehensive review, see (Wray *et al*., 2003). Eukaryotic promoter sequences, or upstream-regulatory regions, contain numerous transcription factor binding sites (TFBS), and are typically a few hundred bp to more than 100 kb long. Within these regions, only 10 to 20% of the nucleotides actually are TFBS. Most binding sites range only 5–8 bp and are grouped into functional modules (Yuh *et al*., 2001). These modules are interspersed with regulatory regions that contain no binding site.

Several transcription factors and cofactors often interact in alternate combinations, depending on signals and physiological or environmental requirements. This gives rise to activating or repressing complexes and requires a modular organization of the binding sites (Messenguy and Dubois, 2003). However, even regions without TFBS may influence the local DNA conformation and modulate interactions within or in between such modules (Wray *et al*., 2003).

While the genetic operations acting on DNA are the same for coding and non-coding regions, the complexity of promoter organization induces different selection criteria since consequences of mutations on organismal development are different (Rodriguez-Trelles *et al*., 2003). Systematic studies are rare due to the aforementioned involvement of subtle signals and context dependency. While, for example, only ≈2–3% of expression variation in yeast duplicates can be explained by motif divergence in upstream-regulatory regions (Zhang *et al*., 2004), point mutations in essential binding sites can have consequences on the phenotype, all the more if the target gene is a transcription factor itself (Carroll, 2000). Different phenomena can explain this apparent contradiction. Compensatory mutations in binding sites or modules have been shown to exist by Tautz and coworkers (Tautz, 2000). Moreover, not all TFBS computationally identified are functional, and subsequently have less

importance for gene regulation (Rodriguez-Trelles *et al*., 2003). Mutations within the modules can modify the affinity of a binding site for its ligand(s), or might even create or delete binding sites resulting in modified transcription profiles that are specific for the affected module and conditions while mutations in coding regions will always be expressed the same (Wray *et al*., 2003).

The regulation of transcription factors plays a key role in morphological diversity. This is illustrated by *Hox* genes, which govern early organismal development. Their regulation is spatially diverged between species (Carroll, 2000). Simple modifications within the upstream-regulation region of TF can explain both minor and major changes between species, without involving any disruption of gene structure, contrary to what would occur in the effector genes (Carroll, 2000). Therefore, evolution of regulatory regions is thought to be a major source of diversity.

Some more general principles were inferred from studies on large-scale "omics" data. When a target gene is duplicated it will only be of use if it also has a functioning regulatory region. Unless it has been copied together with its previous regulatory region (or is copied accidentally into a position of another transcribed region), it will be inert and thus rapidly become a pseudogene. Consequently, target genes with a close homology (as inferred by sequence similarity) can be expected to have a tendency to be more strongly coregulated than those that have drifted more away since, on average, promoter changes will be correlated with genetic drift. Indeed, a positive correlation is observed between the degree of sequence similarity and the coregulation of two duplicated genes (van Noort *et al*., 2004). This has been explicitly shown for yeast target genes that are controlled by the same TFs (Papp *et al*., 2003; Maslov *et al*., 2004). The yeast coexpression network properties have been successfully explained on the basis of this model of coduplication of genes with their TFBSs, deletion and duplication of individual TFBSs and gene loss (van Noort *et al*., 2004).

5. Coevolution and decoupled evolution

From the standpoint of the neutral theory, it is expected that a universally valid and exact molecular clock would exist if, for a given molecule, the mutation rate for neutral alleles per year were exactly equal among all organisms at all times (Gojobori *et al*., 1990). On the basis of this assumption, Wagner tested for a correlation between divergence in sequence and divergence in expression patterns, and found that there was no significant association for yeast duplicates (Wagner, 2000). More recent studies, using more comprehensive data sets (Hughes *et al*., 2000; Lee *et al*., 2002), suggested that a positive correlation existed between sequence similarity of duplicated genes and their coexpression (van Noort *et al*., 2004).

Another study by the same group (Snel *et al*., 2004) combined coexpression and TFBS data, revealing a high conservation of gene coregulation between *C. elegans* and *S. cerevisiae*. It was observed that, in the case of a gene duplication in one of the species, only one of the duplicated genes had a conserved coexpression. Snel *et al*. (2004) then proposed a model for gene duplication in which one copy would maintain the relations of the ancestral gene, while the other would be "free" of selection constraints and could differentiate and/or undergo subfunctionalization.

Recently, the question of how network genes diverge in their transcriptional regulation after duplication was asked and tested (Evangelisti and Wagner, 2004) on the basis of the data set by Lee *et al*. (2002). The authors found that divergence after duplication is often rapid and that there is a frequent net loss of TFBS. However, this study has to be balanced with the observation that although the number of shared modules decreased, the number of modules in regulatory regions of duplicated genes in yeast is stable whatever be the age of the duplication event (Papp *et al*., 2003).

Evolution of regulatory units are extensively studied in Bacteria, but transferring this knowledge to Eukarya is problematic since eukaryotic regulation is more complex (in particular, as much as the binding sites are concerned), and the high frequency of horizontal gene transfer in Bacteria has a strong influence on network organization (McAdams *et al*., 2004).

6. A more holistic perspective: motifs, modules, and graphs

As a consequence of the abundance of data which became recently available, the idea has emerged to conceptualize genetic networks as directed graphs with nodes corresponding to the transcription factors linked by edges to their target genes.

The above-mentioned most basic elements (TF, upstream-binding site, downstream-binding site) can be combined into a next higher level, so-called network motifs (Guelzim *et al*., 2002; Babu *et al*., 2004; Shen-Orr *et al*., 2002). Probably, the most basic motif is the autoregulatory loop (see Figure 1, shaded box): a transcription factor regulates its own expression. In particular, for the inhibition autofeedback loop, Becskei and Serrano (2000) have elegantly shown in theory and experiment that it is an important basic building block enabling stability against perturbation. Generally, these motifs should not be seen as separate units of transcription since they often overlap (Dobrin *et al*., 2004) but rather as basic architectures with a certain response behavior. For example, the feed-forward loop (FFL, see Figure 3, also known as oriented triangles (Guelzim *et al*., 2002)) was shown experimentally to be designed such that noise from the upstream signals is eliminated while there is still rapid response to the target genes (Mangan *et al*., 2003).

Alon and coworkers have investigated the structure of both genetic and protein–protein interaction networks using graph analysis algorithms (Yeger-Lotem *et al*., 2004). As a main conclusion, they found that certain topologies of small subnets are statistically very much overrepresented (Shen-Orr *et al*., 2002; Lee *et al*., 2002).

Conant and Wagner have followed up this idea and introduced the notion of common ancestry for gene circuits or motifs: two motifs share a common ancestor if every pair of genes in the two circuits is derived from a common ancestor: all pairs in the circuits must be duplicated genes (Conant and Wagner, 2003). Finding that essentially no pair of motifs with identical topology had common ancestry, they concluded that their emergence is the result of convergent evolution and not duplication of one or a few ancestral circuits, and noted that convergent evolution was more likely to be important in module topology than in protein sequence

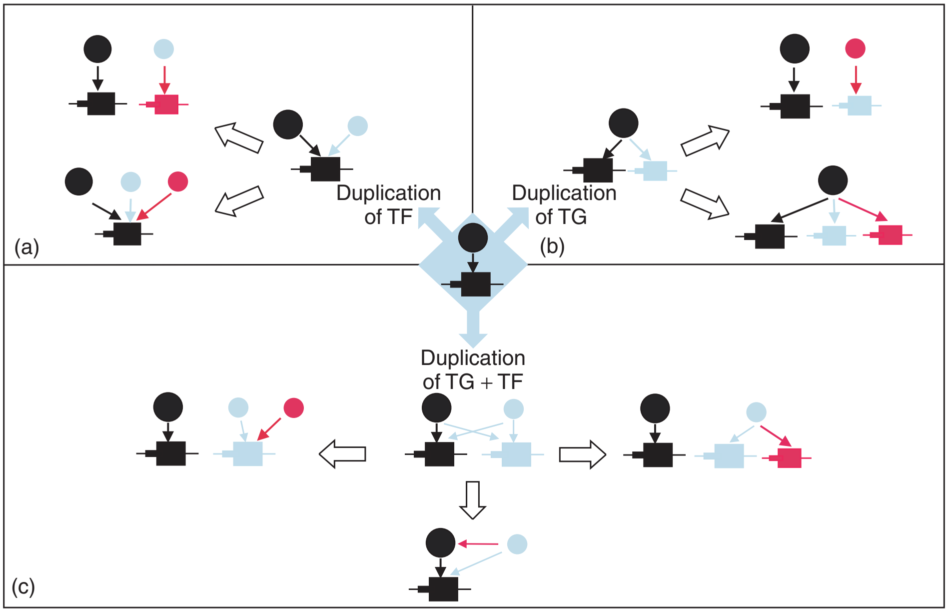


Figure 3 Growth and evolution of regulatory networks by duplication. The circles represent transcription factors (TF), the rectangles stand for the target genes (TG), the thin arrows symbolize regulation, the thick ones represent evolution. Light blue elements are duplication events and products (which are paralogous to their black counterparts), and red symbolizes gain of interaction (generally, this gain can happen by another duplication, except when mentioned). Initially, the basic unit (1st level) of the network is the pair TF/TG (in the center, light blue background). Three kinds of duplication can occur. (a) Duplication of the TF: the TG is regulated by both TF immediately after duplication, in a competitive way. The duplicated TF can then lose its affinity for the TG and gain a new one (red, top). Another solution is the gain of new TF binding sites in the promoter region of the TG, which could lead to the recruitment of yet a third TF, and form a Multiple Input Motif (MIM) (red, bottom). (b) Duplication of the TG: both TG are then regulated by the same TF. Eventually, the duplicated TG can lose its TFBS for the initial TF and gain a new one (red, top); or a third TG could join the motif, thus creating a SIM, a Single Input Motif (red, bottom). Motifs are the 2nd level of regulatory networks. (c) Duplication of both the TF and the TG: immediately after the duplication, both TF regulate both TG. This particular motif is also a MIM, but more closely connected that the one in (a). Through loss and gain of interactions, the new pair can acquire a new TF (red, left), of a new gene (red, right). Finally, after subsequent loss of interaction, if one of the TF acquires a TFBS corresponding to its duplicate (this cannot come from a duplication process), the bottom motif can appear, forming a FFL, or Feed-forward Loop (bottom) (Reproduced from Teichmann and Babu (1999) by permission of Nature Publishing Group)

(Conant and Wagner, 2003). Intuitively, such a conclusion suggests that there is a positive selection on these motifs.

The possible scenarios for the evolution of elementary networks, such as motifs, require the investigation of loss and gain of interactions in terms of duplication or coduplication of transcription factors and target genes as Teichmann and Babu (2004) have elaborated in a seminal paper. Duplication of a target gene alone will result in two genes (co-)controlled by the same TF, duplication of the TF alone will mean the target genes are controlled by the two TFs while coduplication will lead to genes coregulated by both TFs (see Figure 3). Changes in sequences of regulatory regions and/or binding domains or secondary domains will then remove or add further links. Clearly, these basic mechanisms will lead to SIMs and MIMs,

while FFLs require different mechanisms such as the gain of a new module in the promoter of a preexisting gene (see Figure 3, also for definitions). Intuitively, one might assume that the aforementioned observation of convergent motif evolution indicates that they have arisen predominantly by duplication of the whole unit. However, although many (≈45%) of the regulatory interactions in *E. coli* and yeast have arisen by gene duplication and conservation of the corresponding interactions (Teichmann and Babu, 2004), the main mechanisms of generating motifs are frequent losses and gains of interactions following the more elementary events mentioned afore (van Noort *et al*., 2004).

The third level of network organization according to Babu *et al*. (2004) comprises transcriptional modules. These are collections of motifs that are fairly independently regulated. They correspond to functional units as has been concluded from several studies using expression data (Lee *et al*., 2002). Modules represent collections of transcription factors that are expressed under distinct experimental conditions (Ihmels *et al*., 2002) and largely controlled by one (or very few) regulators as was shown by hierarchical clustering of expression profiles (Segal *et al*., 2003). Tavazoie *et al*. (1999) showed experimentally that the clusters of coexpressed genes largely correspond to sets of genes with similar function according to database classifications.

It is worthwhile noting here that modules of networks have also been identified for metabolic pathways, however, in terms of smallest stoichiometric units that cannot be further decomposed (Schuster *et al*., 2000; Schilling and Palsson, 1998) and can be defined for signaling pathways as well. Future research will show if these units refer to evolutionary related units and if their genes, too, are highly coregulated. For metabolic pathways, genomic clustering indicating strong evolutionary relationship has been shown (Lee and Sonnhammer, 2003).

For genetic networks, first indirect evidence of common evolutionary origin was provided by Schwikowski *et al*. (2000) who showed that functionally related proteins tend to have overrepresented protein–protein interactions and that most interactions are confined to proteins acting in the same cell compartment, including the nucleus and the transcription factors contained therein.

We can now return to one of the initial questions on convergent evolution of the camera eye. A recent study based on EST data suggested that in spite of differences in the actual topology, eye development in both organisms is controlled by sets of pairwise homologous genes although they are not involved in eye development in the common ancestor (Ogura *et al*., 2004). Assuming now that these regulatory genes correspond to modules in the sense mentioned above (which remains to be proven) and are controlled by few transcription factors, it is conceivable that the modules arose repeatedly from this (or these) ancestral gene(s) through the mechanism outlined above.

Further analysis of global features of genetic networks and protein-interaction networks were done on the basis of concepts borrowed from graph analysis and has revealed a scale-free topology (Barabasi and Albert, 1999; Barabasi and Oltvai, 2004). This means there are few genes, so-called hubs, controlling many others and many genes with only few links. However, regulatory networks are dynamic: it was shown that large-scale topological changes happened in the yeast genome and that although a few TFs serve as permanent hubs, most act transiently only during

certain conditions (Luscombe *et al*., 2004). The formula $C_I = \frac{2n_I}{k(k-1)}$, defines a clustering coefficient, where n_I is the number of links connecting the k_I neighbors of node I to each other (Barabasi and Oltvai, 2004). A value close to 0 with many neighbors indicates a hub, a value close to 1 a cluster. Protein-interaction networks are very useful since they can help investigate networks emerging from dimerization and secondary regulatory domains of TFs. In contrast to gene-regulatory networks, they induce nondirected graphs.

Observing the directed graph of gene-regulatory networks in yeast (Guelzim *et al*., 2002) from two distinct perspectives revealed a fundamentally different behavior of the two subgraphs : (1) the graph of incoming interactions, that is, the population of target genes being regulated by transcription factors, where the degree of connectivity (in-degree) is the number of regulatory proteins binding to a promoter region and (2) the graph of outgoing interactions, that is, the population of transcription factors regulating target genes, where the degree of connectivity (out-degree) is the number of target genes to which a transcription factor binds. It appears (Guelzim *et al*., 2002) that the in-degree has a very narrow distribution: around the mean number of transcription factors, the probability decreases with an exponential rate (nearly all target genes have the same number of transcription factors), while the out-degree has a broader distribution and follows a power-law.

Knockout studies in yeast suggested that target genes regulated by more than nine TFs are, in general, less essential (they contain proportionally less lethal genes), while the most important transcription factors are the ones binding the most target genes (Yu *et al*., 2004).

It is also worthwhile noting that more complex organisms have a higher number of regulatory genes per target gene (van Nimwegen, 2003), suggesting that it is mostly the evolution by duplication and diversification of transcription factors and of their interactions that increases organismic complexity as a whole.

7. Evolution of expression

The observation, in numerous rescue studies, that orthologous TFs could maintain their function despite a significant sequence divergence might have been overestimated (Hsia and McGinnis, 2003). Both changes in the TF expression (upstream regulation) and changes in its coding sequence can explain the evolution of transcription factor function and thus play an important role in the evolution of regulatory pathways (Hsia and McGinnis, 2003).

Recently, it was suggested that evolution with respect to expression levels is neutral. Khaitovich *et al*. (2004) observed similar divergence of gene expression levels for real and pseudogenes. However, Jordan *et al*. (2005) recently reported on a study including tissue specific expression levels in rat, mouse, and human. They suggested that while there may be a degree of neutrality in genes that are less important for an organ, selective constraints prevail on gene expression divergence. While clocklike rates of neutral evolution were found for sequence divergence of genes, the divergence of profiles with respect to tissue specificity across species suggested that selective constraints dominate.

8. Conclusions

Over the last years, the availability of genomics, proteomics, and transcriptomics data have significantly enhanced our understanding of evolution of complex networks and, at the same time, enabled us to transfer knowledge from better studied model organisms (such as *E. coli* and yeast) to those for which fewer data are available but, for example, a genome has been analyzed by low shot-gun coverage. In essence, the following picture, which will undoubtedly undergo revisions in the near future, is beginning to emerge:

- Most ancestral to genetic networks as we observe them today were probably a small group of DNA-binding domains that, while conserving their structure (fold), diverged into a large variety of TFs. Ancestral proteins are widely believed to have been single domain proteins.
- More recently, most proteins, among them probably TFs, underwent many cycles of domain rearrangements (Bornberg-Bauer *et al*., 2005). Additional dimerization and sensoring domains were added and often lost again.
- More ancestral regulatory genes are, in the light of today's situation, often retained and have maintained a key role, for example, in regulating function-specific modules. They evolve further, preferentially by series of single-gene duplications, thus generating networks of regulatory genes that arrange into these modules. They can be colocalized and systematically build up units of higher complexity. These events can be quite recent and lineage specific as we have learned from the uneven distribution of some TF families.
- A growing number of findings suggests that structurally similar or even identical motifs can arise repeatedly and thus represent a simple level of convergent evolution.
- More complex modules, which may also have preferentially arisen by series of single-gene duplications, gave rise to similar topologies. They are, typically, collections of regulatory genes with a tightly linked function and correlated expression that are well controlled by the descendants of the ancestral gene.
- Large-scale duplications may have been important in generating specific network structures as we see them today and increase complexity but probably they were not instrumental for establishing the networks.
- The evolution of promoter regions is less understood, although of great importance. Genotypic changes at this level are probably among the main reasons why, in spite of minor interorganismic differences at the level of proteins, major changes in the topologies of genetic networks during development induce a wide morphological difference and diversity in todays creatures.
- While there is no evidence for selection on topological motifs and modules, expression studies suggest that selection of expression levels per se does take place.

In summary and, although intuitively difficult to grasp, the evolutionary mechanisms in generating diverse networks with dramatic phenotypic changes are taken from the well-known standard repertoire of evolutionary operators: gene duplication, redundancy, functionalization, neutral and selective evolution, gene loss,

and modular rearrangements. Taken together, these insights are "Taking the mystery out of biological networks" (Aloy and Russell, 2004). So, although we are nowhere close to rationally constructing or predicting the complex behavior of gene-regulatory networks in the cell, we are beginning to understand their evolutionary dynamics.

Acknowledgments

We gratefully acknowledge useful comments from Georg Fuellen, Francois Kepes, Sarah Teichmann, and Guenter Theissen.

Further reading

Yanai I, Graur D and Ophir R (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, **8**, 15–24.

References

Akache B, Wu K and Turcotte B (2001) Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic Acids Research*, **29**, 2181–2190.

Aloy P and Russell R (2004) Taking the mystery out of biological networks. *EMBO Reports*, **5**, 349–350.

Alvarez-Buylla E, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, de Pouplana LR, Martínez-Castilla L and Yanofsky MF (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5328–5333.

Amoutzias G, Robertson D and Bornberg-Bauer E (2004a) The evolution of protein interaction networks in regulatory proteins. *Comparative and Functional Genomics*, **5**, 79–84.

Amoutzias G, Robertson D, Oliver S and Bornberg-Bauer E (2004b) Convergent networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, **5**, 274–279.

Amoutzias G, Weiner J III and Bornberg-Bauerg E (2005) Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. *Gene*, **347**(2), 247–253.

Aravind L and Koonin E (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Research*, **27**, 4658–4670.

Babu M and Teichmann S (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, **31**, 1234–1244.

Babu M, Luscombe NM, Aravind L, Gerstein M and Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, **14**, 283–291.

Barabasi A and Albert R (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barabasi A and Oltvai Z (2004) Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, **5**, 101–113.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, *et al*. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**(Database issue), D138–D141.

Becker A and Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution*, **29**, 464–489.

Becskei A and Serrano L (2000) Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593.

Bode J, Goetze S, Heng H, Krawetz SA and Benham C (2003) From DNA structure to gene expression: mediators of nuclear compartmentalization and dynamics. *Chromosome Research*, **11**, 435–445.

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA and Weiner J III (2005) The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences*, **62**(4), 435–445.

Carroll S (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, **101**, 577–580.

Conant G and Wagner A (2003) Convergent evolution of gene circuits. *Nature Genetics*, **34**, 264–266.

Coulson R, Enright A and Ouzounis C (2001) Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, **17**, 95–97.

Dobrin R, Beg Q, Barabasi A and Oltvai Z (2004) Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, **5**, 10.

Emberly E, Rajewsky N and Siggia E (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics*, **4**, 57.

Evangelisti A and Wagner A (2004) Molecular evolution in the yeast transcriptional regulation network. *Journal of Experimental Zoology. Part B. Molecular and Developmental Evolution*, **302**, 392–411.

Fontana W and Buss L (1994) What would be conserved if "the tape were played twice"? *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 757–761.

Fontana W and Schuster P (1998) Continuity in evolution: on the nature of transitions. *Science*, **280**, 1451–1455.

Gojobori T, Moriyama E and Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 10015–10018.

Gould S (1989) *Wonderful Life: The Burgess Shale and the Nature of History*, W.W. Norton: New York.

Guelzim N, Bottani S, Bourgine P and Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, **31**, 60–63.

Hsia C and McGinnis W (2003) Evolution of transcription factor function. *Current Opinion in Genetics & Development*, **13**, 199–206.

Huffman J and Brennan R (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Current Opinion in Structural Biology*, **12**, 98–106.

Hughes T, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H and He YD (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y and Barkai N (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**, 370–377.

Jordan I, Marino-Ramirez L and Koonin E (2005) Evolutionary significance of gene expression divergence. *Gene*, **345**(1) 119–126.

Kaufmann K, Melzer R and Theissen G (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene*, **347**(2), 183–198.

Kawana K, Ikuta T, Kobayashi Y, Gotoh O, Takeda K and Kawajiri K (2003) Molecular mechanism of nuclear translocation of an orphan nuclear receptor, SXR. *Molecular Pharmacology*, **63**, 524–531.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W and Pääbo S (2004) A neutral model of transcriptome evolution. *PLoS Biology*, **2**, E132.

Ledent V and Vervoort M (2001) The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. *Genome Research*, **11**, 754–770.

Lee J and Sonnhammer E (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research*, **13**, 875–882.

Lee T, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Lespinet O, Wolf Y, Koonin E and Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research*, **12**, 1048–1059.

Levine M and Tjian R (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Luscombe N, Babu MM, Yu H, Snyder M, Teichmann SA and Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Mangan S, Zaslaver A and Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, **334**, 197–204.

Maslov S, Sneppen K, Eriksen K and Yan K (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evolutionary Biology*, **4**, 9.

Matthews J and Visvader J (2003) LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. *EMBO Reports*, **4**, 1132–1137.

McAdams H, Srinivasan B and Arkin A (2004) The evolution of genetic regulatory systems in bacteria. *Nature Reviews. Genetics*, **5**, 169–178.

Messenguy F and Dubois E (2003) Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene*, **316**, 1–21.

Morgenstern B and Atchley W (1999) Evolution of bHLH transcription factors: modular evolution by domain shuffling? *Molecular Biology and Evolution*, **16**, 1654–1663.

Ogura A, Ikeo K and Gojobori T (2004) Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human. *Genome Research*, **14**, 1555–1561.

Papp B, Pal C and Hurst L (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends in Genetics*, **19**, 417–422.

Perez-Rueda E and Collado-Vides J (2000) The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research*, **28**, 1838–1847.

Riechmann J, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, *et al*. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Rodriguez-Trelles F, Tarrio R and Ayala F (2003) Evolution of cis-regulatory regions versus codifying regions. *The International Journal of Developmental Biology*, **47**, 665–673.

Schilling C and Palsson B (1998) The underlying pathway structure of biochemical reaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 4193–4198.

Schuster S, Fell D and Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, **18**, 326–332.

Schwabe J and Teichmann S (2004) Nuclear receptors: the evolution of diversity. *Science's STKE*, **2004**, pe4.

Schwikowski B, Uetz P and Fields S (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**, 1257–1261.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.

Shen-Orr S, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**, 64–68.

Sluder A, Mathews SW, Hough D, Yin VP and Maina CV (1999) The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Research*, **9**, 103–120.

Snel B, van Noort V and Huynen M (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Research*, **32**, 4725–4731.

Tautz D (2000) Evolution of transcriptional regulation. *Current Opinion in Genetics & Development*, **10**, 575–579.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281–285.

Teichmann S and Babu M (2004) Gene regulatory network growth by duplication. *Nature Genetics*, **36**, 492–496.

van Nimwegen E (2003) Scaling laws in the functional content of genomes. *Trends in Genetics*, **19**, 479–484.

van Noort V, Snel B and Huynen M (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, **5**, 280–284.

Wagner A (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6579–6584.

Wang H, Noordewier M and Benham C (2004) Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. *Genome Research*, **14**, 1575–1584.

Wray G, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV and Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377–1419.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U and Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 5934–5939.

Yu H, Greenbaum D, Xin Lu H, Zhu X and Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends in Genetics*, **20**, 227–231.

Yuh C, Bolouri H and Davidson E (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, **128**, 617–629.

Zhang Z, Gu J and Gu X (2004) How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends in Genetics*, **20**, 403–407.

Short Specialist Review

Phylogenomic approaches to bacterial phylogeny

Vincent Daubin

University of Arizona, Tucson, AZ, USA

Université Claude-Bernard Lyon 1, Villeurbanne, France

1. Introduction

Because phylogeneticists are in a constant quest for more phylogenetic characters, the approaches aimed at building trees using multiple genes and complete genome sequences have become quite popular. Genome sequences can be used in different ways to generate pertinent phylogenetic information. The most widely applied approach consists in concatenating several sequence alignments in order to increase the statistical power of phylogenetic methods (*see* Article 42, **Reconstructing vertebrate phylogenetic trees**, Volume 7). Other characters such as the presence of genes, gene families, or protein domains in genomes have also been used. While the benefits of such methods are evident, their application supposes that certain hypotheses are respected. In bacteria, for instance, it is quite certain that the first of these hypotheses – that genes follow a pattern of strict vertical inheritance and thus provide reliable markers of species phylogeny – has been repeatedly violated. Intertwined with limitations due to such biological phenomenon are a number of methodological and theoretical pitfalls that need to be identified. In this review, I will discuss the importance of these limitations with regard to recent literature and the applicability of a phylogenomic approach to bacterial history. Because an unrooted tree of bacteria has very little to say about their history, I will mostly discuss attempts to reconstruct the tree of life.

2. The organismal phylogeny: are we chasing after Chimeras?

The first tree of life based on molecular data has been proposed by Woese *et al*. (1985) on the basis of a single gene, the RNA of the small subunit of the ribosome (SSU rRNA) (*see* Article 40, **The domains of life and their evolutionary implications**, Volume 7). This gene has the advantage of being ubiquitous and extremely well conserved in cellular organisms. Although the whole classification of bacteria is based on this single gene, its use as a unique reference has been questioned. Moreover, the SSU rRNA phylogeny defines a number of strongly

cohesive bacterial groups, but the relationships among these groups are poorly supported. The only robust information about the ancient history of bacteria provided by the SSU rRNA phylogeny is the basal position of the hyperthermophilic bacteria *Aquifex* and *Thermotoga*, but this result has been recently challenged (Brochier and Philippe, 2002). Therefore, there is a need to investigate the history of bacteria (and the reliability of the current classification) using independent information.

Bacteria, and more generally prokaryotes, are certainly very special organisms when it comes to phylogenetics. Lateral gene transfer (LGT) is frequent in bacteria, and its impact on phylogenetic inference can be dramatic. The abundance of LGT is supported by two main observations: first, incongruence is frequently observed among gene phylogenies when considering genes from distantly related bacteria (Brown *et al*., 2001; Brochier *et al*., 2002; Gogarten *et al*., 2002); second, the differences in gene repertoires among bacteria are striking at every phylogenetic level, suggesting constant gene acquisitions and nonhomologous replacements (Ochman *et al*., 2000; Daubin *et al*., 2003a, b; Koonin, 2003). Traditionally, genes encoding proteins involved in intimate relationships with the cell apparatus have been considered resistant to LGT, because their sequences result from a long coevolution during which they have established precise interactions with other proteins (Jain *et al*., 1999). However, cases of LGT have been described in the heart of the ribosome (Brochier *et al*., 2000), and have affected ribosomal RNA, at least in some groups (Ueda *et al*., 1999; Schouls *et al*., 2003). Thus, it is likely that virtually every gene tree will show evidence for LGT if the species sampling considered is wide enough.

These results have led some phylogeneticists to suggest that, because of LGT, "treelike phylogenies are inadequate to represent the pattern of prokaryotic evolution at any level" (Gogarten *et al*., 2002; see also Doolittle, 1999). However, this view is reducing the organisms to their genes. In the absence of cell fusion and despite LGT, the history of cell lineages is a sequence of successive divisions, which can be appropriately described as a tree. In contrast to eukaryotes, where endosymbiosis seems to have occurred repeatedly, there is no evidence suggesting that free-living bacterial species are hybrids or chimeras. Therefore, contrary to previous claims, the abundance of LGT in bacteria simply questions genes as adequate phylogenetic markers for inferring the organismal phylogeny and not the fact that such a phylogeny has existed. Although the pervasiveness of LGT can be estimated without reference to the history of cell lineages, it is only by considering this history that we can understand the biological significance of LGT.

Although the debate has been presented as a challenge to a Darwinian treelike vision of evolution, it appears that the relevant question is not whether the tree metaphor is appropriate to represent the history of bacteria but whether the gene sequences carry pertinent information about the history of cell lineages?

A high degree of LGT is not in contradiction with the concept of organismal phylogeny, but it may forbid its reconstruction. To address this question, it is possible to systematically compare the phylogenetic information of different genes to evaluate the extent of conflict. Different approaches have been proposed to address this issue. A first method consists in comparing the topologies of orthologous gene trees to a reference tree, believed to be the species phylogeny. Such an approach

has been used by Jain *et al*. (1999) using the data from six complete prokaryotic genomes (four bacteria and two archaea). These authors found that most trees derived from orthologous genes were incongruent with the reference and interpreted this result as an evidence for continual horizontal transfers. However, the same approach applied to six complete eukaryotic genomes (three animals, two fungi, and one plant) has revealed a very similar degree of conflict between trees (Wolf *et al*., 2004). Because such a frequency of LGT can hardly be hypothesized, especially among different eukaryotic kingdoms, it appears that the test is probably inappropriate to measure reliably the extent of LGT. Other methodologies that evaluate the adequacy of alternative topologies with gene alignments using Maximum Likelihood approaches have suggested high degree of conflict (Zhaxybayeva and Gogarten, 2002), but again, these approaches may be too sensitive to phylogenetic artifacts because they only consider four genomes at a time (Daubin and Ochman, 2004). These studies have addressed the question using few anciently diverged species, and probably suffered from insufficient taxon sampling, with the result that the incongruences observed among genes could be interpreted as due to phylogenetic artifacts rather than LGT. In contrast, analysis of more closely related species (Daubin *et al*., 2003b) or using better species sampling (Lerat *et al*., 2003) have revealed remarkable congruence between orthologous gene histories. Although these studies addressed the history of a group of species rather than the entire bacterial domain, they demonstrated that sequences contain exploitable information to infer the organismal phylogeny and that given the rates at which LGT has been reported to occur (Ochman *et al*., 2000; Daubin *et al*., 2003a), orthologous genes constitute surprisingly good phylogenetic markers. However, these studies do not prejudge of the applicability of phylogenomic methods to the inference of relationships between bacterial phyla.

3. Phylogenetics in the genomic era

Several attempts to reconstruct the tree of bacteria using the data from complete genomes have been published. The different methods proposed can be classified into two groups: those using the phylogenetic signal of numerous genes to increase the statistical power of the phylogenetic methods and those defining global indexes of similarity between genomes, for instance by comparing gene content (*see* Article 49, **Chromosome phylogeny**, Volume 7).

There are two main approaches using the phylogenetic signal contained in gene sequences. It is possible to apply these methods without having complete genome sequences. The most intuitive of them is the concatenation of genes (Teichmann and Mitchison, 1999; Brown *et al*., 2001; Wolf *et al*., 2001; Brochier *et al*., 2002; Matte-Tailliez *et al*., 2002). It anticipates that large sequence alignments will increase the robustness of the phylogenetic signal and consists in concatenating all the orthologous genes available into a "super-alignment". Every concatenation approach admits a priori that individual genes may not, for a reason or another, represent the species phylogeny but may contain information that can be revealed when confronted to other genes. In conducting such an analysis, it is necessary to use a method that allows different sites of the alignment to evolve at different rates

because genes may be informative at different evolutionary scale. The principal limitation of this approach is that it requires the genes to be present in all or much of the species considered. This restriction results in the limited number of genes that can be used, especially in the case of the tree of life, with the consequence that one or a few long genes may rule over the phylogenetic signal of other genes in the "super-alignment". Such problems can have dramatic effect when genes having undergone LGT are included in the "super-alignment" (Teichmann and Mitchison, 1999; Brown *et al*., 2001). The most current practice in this case is to exclude the alignment of the genes presenting evidence for LGT, even if only one or a few species are concerned. In a study of the tree of life by Brown *et al*. (2001), as much as nine gene alignments (over 23) were excluded (because their individual trees presented the bacteria as not monophyletic), resulting in removing more than 40% of the phylogenetic information.

A possible alternative is the "supertree" method (Daubin *et al*., 2001, 2002). Here, the genes trees are combined instead of the alignments. The great advantage of this method over the concatenation approach is that it allows combining trees that share only a few species. For the inference of the tree of life, it raises the number of informative genes from two dozen to several hundreds. It also resolves the problem of the dominance of a gene over the others because each gene tree is given approximately the same weight in the analysis. The potential impact of LGT is thus reduced. The main limitation of the method is that it reduces the phylogenetic information of a gene to a single tree that is not guaranteed to be the true tree. Therefore, in contrast to the concatenation approach, there is no possibility for a weak phylogenetic signal present in several genes to be valorized by the analysis. The "supertree" approach will rather deal with conflicting information in the manner of a consensus method. Both the concatenation and the "supertree" approach have confirmed most of the evolutionary groups proposed by Woese *et al*. (1985) and have even brought some new insights into the ancient history of bacteria.

The other class of methods defines a global index of resemblance among genomes and requires having complete genome sequences. One of the first trees based on global comparison of genomes has been proposed by Snel *et al*. (1999), where the index of similarity between two genomes was defined as the ratio of the number of shared orthologs and the number of genes present in the smallest of them. It is important to note that such a method uses phylogenetic characters that are independent from a concatenation approach, which considers genes that are shared among all species. In a study of genome content, only genes that have limited phylogenetic distribution are considered as informative characters. This approach has been surprisingly successful in recovering most of the groups that had been defined on the basis of the SSU rRNA phylogeny. Numerous other approaches that instead of shared orthologs use the presence of gene families or protein domains families have been proposed (Fitz-Gibbon and House, 1999; Tekaia *et al*., 1999; Lin and Gerstein, 2000; House and Fitz-Gibbon, 2002). These characters have proved to be informative at some phylogenetic level, allowing to recover the monophyly of Bacteria, Archaea, and Eukaryotes, but generally fail to infer more precise relationships. These differences with the results from Snel *et al*. (1999), can be explained by the reduction in the number and the quality of informative characters (protein domain families may contain several gene families that themselves contain

several families of orthologs). Other authors have defined the resemblance of two genomes as a function of the BLAST scores obtained when comparing all genes from a pair of genomes (Clarke *et al*., 2002). These approaches, again, confirmed most of the groups defined by Woese.

4. Reconstructing the history of bacteria: despite or thanks to LGT?

The similarities in the trees obtained from independent phylogenomic methods have been suggested to be the result of LGT. In this view, the grouping observed in these trees would only reflect the propensity of bacterial species to exchange genes (Gogarten *et al*., 2002). Indeed, an expected consequence of LGT is the illegitimate grouping of species, both in gene trees and gene content–based phylogenies. However, such a hypothesis predicts that gene phylogenies will show conflict between each other at the narrowest phylogenetic level because the species that "seem" closely related exchange genes frequently. This prediction is clearly rejected by the analysis: genes sampled from closely related species show remarkable phylogenetic congruence, demonstrating that vertical transmission of orthologous genes is prevalent in the short term (Daubin *et al*., 2003b; Lerat *et al*., 2003). Therefore, the congruence of the results obtained from different phylogenomic approaches implies that the phylogenetic signal has also been sufficiently conserved for a phylogeny to be inferred, at least within bacterial phyla. Interestingly, genome content, if used appropriately, appears to be suitable for phylogenetic inference, indicating that the acquisition of new genes by LGT, though frequent, is not an overwhelming source of homoplasy.

Most of the difficulties in reconstructing the bacterial phylogeny reside in the deepest nodes of the tree, and particularly in the position of its root. The antiquity of bacteria is a major barrier in reconstructing their history and it is possible that most of the phylogenetic information has been erased. However, despite the growing number of complete bacterial sequenced genomes, the current sampling of their diversity in genome databases is still very limited. The availability of more genomes should increase the reliability of the approaches described here. Ultimately, the combination of phylogenomic approaches will allow documentation of the events, including vertical inheritance, LGT, and gene loss, that have shaped contemporary and ancestral genomes.

Acknowledgments

I would like to thank Emmanuelle Lerat for comments on this paper.

References

Brochier C, Bapteste E, Moreira D and Philippe H (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends in Genetics*, **18**, 1–5.

Brochier C and Philippe H (2002) Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature*, **417**, 244.

Brochier C, Philippe H and Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends in Genetics*, **16**, 529–533.

Brown JR, Douady CJ, Italia MJ, Marshall WE and Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics*, **28**, 281–285.

Clarke GD, Beiko RG, Ragan MA and Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *Journal of Bacteriology*, **184**, 2072–2080.

Daubin V, Gouy M and Perriere G (2001) Bacterial molecular phylogeny using supertree approach. *Genome Informatics Series: Workshop on Genome Informatics*, **12**, 155–164.

Daubin V, Gouy M and Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research*, **12**, 1080–1090.

Daubin V, Lerat E and Perriere G (2003a) The source of laterally transferred genes in bacterial genomes. *Genome Biology*, **4**, R57.

Daubin V, Moran NA and Ochman H (2003b) Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**, 829–832.

Daubin V and Ochman H (2004) Quartet mapping and the extent of lateral transfer in bacterial genomes. *Molecular Biology and Evolution*, **21**, 86–89.

Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129.

Fitz-Gibbon ST and House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, **27**, 4218–4222.

Gogarten JP, Doolittle WF and Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, **19**, 2226–2238.

House CH and Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Journal of Molecular Evolution*, **54**, 539–547.

Jain R, Rivera MC and Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 3801–3806.

Koonin E (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews. Microbiology*, **1**, 127–136.

Lerat E, Daubin V and Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biology*, **1**, 19.

Lin J and Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, **10**, 808–818.

Matte-Tailliez O, Brochier C, Forterre P and Philippe H (2002) Archaeal phylogeny based on ribosomal proteins. *Molecular Biology and Evolution*, **19**, 631–639.

Ochman H, Lawrence JG and Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Schouls LM, Schot CS and Jacobs JA (2003) Horizontal transfer of segments of the 16 S rRNA genes between species of the Streptococcus anginosus group. *Journal of Bacteriology*, **185**, 7241–7246.

Snel B, Bork P and Huynen MA (1999) Genome phylogeny based on gene content. *Nature Genetics*, **21**, 108–110.

Teichmann SA and Mitchison G (1999) Is there a phylogenetic signal in prokaryote proteins? *Journal of Molecular Evolution*, **49**, 98–107.

Tekaia F, Lazcano A and Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Research*, **9**, 550–557.

Ueda K, Seki T, Kudo T, Yoshida T and Kataoka M (1999) Two distinct mechanisms cause heterogeneity of 16 S rRNA. *Journal of Bacteriology*, **181**, 78–82.

Woese CR, Stackebrandt E, Macke TJ and Fox GE (1985) A phylogenetic definition of the major eubacterial taxa. *Systematic and Applied Microbiology*, **6**, 143–151.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL and Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, **1**, 8.

Wolf YI, Rogozin IB and Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Research*, **14**, 29–36.

Zhaxybayeva O and Gogarten JP (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics*, **3**, 4.

Short Specialist Review

Phylogenomics for studies of microbial evolution

Siv G. E. Andersson and Hans-Henrik Fuxelius
Uppsala University, Uppsala, Sweden

1. Introduction

The classification of microorganisms represents a major challenge, the basis of which has been an issue of debate for the past 100 years (Woese, 1994). Phylogenetic inference based on rRNA sequences is currently the gold standard for studies related to microbial evolution and classification. As more and more microbial genomes are added to the rapidly expanding pool of sequence data, automatic systems are being developed to meet the growing demand for large-scale phylogenetic analyses. However, the development of such tools and their application to biological problems has raised more questions than has been solved, and the debate about the extent to which microorganisms are related by vertical descent versus horizontal transmission is more heated than ever before. Here, we describe the recent development of methods and tools for phylogenomic studies and their application in microbiological research.

2. Ribosomal RNA trees

More than 100 years ago, attempts were made to construct a bacterial classification scheme, following the Linnean tradition. However, morphological and biochemical characters were found to be too simple, and the phylogenetic approach to microbial classification was explicitly abandoned in the mid-1960s. Ironically, only 10 years later, Carl Woese revolutionized biology by constructing a phylogenetic tree based on rRNA sequences, the so-called Tree of Life, which suggested that the prokaryotes (cells without a nucleus) belong to two genetically different groups, Bacteria and Archaea, each of which is equidistant from the eukaryotes (cells with a nucleus) (Woese *et al*., 1990). Since then, rRNA sequence data has been collected from thousands of organisms, and rRNA trees have modernized microbiology by providing a reference system for diagnosis, classification, and evolutionary investigations.

3. Genome annotation by sequence similarity searches

The first task in the analysis of a newly sequenced genome is to predict the location of genes and assign putative functions to the identified genes. Since experimental data is in most cases missing, the functional assignment (*annotation*) is based on sequence similarity to genes in other species with an already assigned function that was experimentally determined. The most widely used program for such sequence similarity searches is BLAST (Altschul *et al*., 1997), which is a rapid and simple method to detect closest matches in database searches.

This method is ideal for the analysis of genome sequence data from the point of view that it can easily be automated for the annotation of thousands of genes. The results may be graphically mapped onto circles that display the order of genes, with colors representing the functional categories to which the annotated genes belong. Additional circles can be added to detail the species to which a hit for any given gene was identified. Regions without similarity to genes in the databases are visualized as "holes" in the circles, as seen, for example, in the genome circle of *Francisella tularensis*, the causative agent of tularemia (Larsson *et al*., 2005). Such species-unique segments, also called genomic islands, often represent horizontally acquired DNA that putatively code for proteins involved in pathogenesis and/or environmental adaptation.

Despite the many virtues of the BLAST program, a limitation is that it does not distinguish orthologs related by vertical descent from members of large gene families that share a relationship due to gene duplication events. Within a family, all gene members share some sequence similarity. Yet, the encoded proteins may perform related but not necessarily identical tasks due to functional divergence following the gene duplication events. Hence, it is very difficult to infer functional equivalence solely on the basis of sequence similarity inferred with the aid of the BLAST program.

4. Phylogenetic methods

Phylogenetic methods differ from BLAST methods in that they were developed to provide exact measures of the internal relationships of a limited number of sequences. To this end, the sequences to be analyzed need first to be extracted from the public databases with the aid of BLAST, after which they are aligned using programs such as CLUSTALW (Thompson *et al*., 1994). The evolutionary relationships of the taxa represented by the aligned sequences can then be explored using a variety of tree reconstruction methods, such as, for example, distance-, parsimony-, maximum likelihood, and Bayesian methods. Finally, the statistical support for each cluster of taxa is estimated.

Thus, phylogenetic inference includes a series of steps; (1) manual extraction of sequence data from the databases, (2) alignment of the selected sequences, (3) editing of the alignment, (4) selection of substitution models, (5) inference of the tree topology, (6) statistical testing of the identified clusters, (7) visualization, and (8) inspection of the resulting trees. Until recently, each of these steps required manual intervention and it is only with the recent development of automatic

tools that phylogenetic methods have become a realistic alternative to the simpler sequence similarity methods used for large-scale analyses of genome sequence data.

5. Phylogenomics

The term "phylogenomics" refers to the application of phylogenetic methods to whole-genome sequence data. Automatic tools for phylogenetic inferences are of practical use in genome annotation projects and are also of great interest to explore the occurrence of horizontal gene transfer events. With the aid of such methods, sequences are extracted and aligned automatically, and trees are built and visualized in a fully automated manner. One of the first such tools was applied to the analysis of seven microbial genomes and resulted in the construction of more than 8000 individual phylogenetic trees (Sicheritz-Ponten and Andersson, 2001). An interesting observation from this early study was that although a majority of trees supported the rRNA division of bacteria and archaea, a substantial fraction of trees deviated from the expected pattern (Sicheritz-Ponten and Andersson, 2001).

6. How to describe life?

The incongruence between the many individual gene trees constructed in phylogenomic studies has been used as an argument in favor of the hypothesis that horizontal gene transfer events occur frequently in the microbial world. Indeed, mechanisms for DNA transfer are well known for decades: transformation, transduction, and conjugation all contribute to the exchange of genes within and among bacterial populations. Among genes typically encountered on transmissible plasmids are those for antibiotic resistance and environmental adaptation. Some people argue that horizontal gene transfer events occur so frequently that the tree of life is not a proper description of the relationships among microbial species. For example, the proponents of the web-of-life model argue that life should instead be represented as a web, in which microbial cells freely exchange DNA within and among populations (Doolittle, 1999).

Although phylogenomic studies are indicative of extensive gene exchange, they support the notion that the separation of Bacteria and Archaea into two separate domains is robust, that is, not distorted by horizontal gene transfer. This suggests that there is a backbone of genes with an evolutionary history that is in harmony with the "universal tree of life" (Sicheritz-Ponten and Andersson, 2001). This is also consistent with the results from comparative genome analyses, which have shown that microbial genomes consist of vertically inherited regions interspersed with horizontally acquired blocks of DNA. The sequences inside these islands are turning over much more rapidly than the remainder of the genome. This may be because most of such genes are not essential or only required under certain environmental conditions.

7. Phylogenomics – problems and pitfalls

A pertinent question that needs to be asked is whether the many gene trees obtained in phylogenomic studies accurately record the evolutionary history of the individual genes? Phylogenetic inference is based on a number of assumptions, such as, for example, that the genes are conserved to an extent that matches the studied divergences and that they are functionally equivalent so that they evolve under identical selective constraints. Ideally, the genes under investigation should also evolve at similar rates and have similar base composition patterns.

In real case studies, these assumptions are often violated, which may result in tree artifacts, that is, the obtained tree topology may not necessarily reflect the evolutionary history of the gene. Because evolution has only happened once and since we do not know the order of divergences a priori, there are few test grounds under which the assumptions and models used in phylogenetics can be tested. Such experiments would be particularly relevant for phylogenomic studies in which there are few opportunities to manually inspect the reconstructed trees.

Fortunately, there are a few biological systems that may be used to examine the accuracy of whole-genome phylogenetic methods. These include endosymbionts that have coevolved with their hosts and for which horizontal gene transfer events are expected to be rare or nonexistent. A recent examination of the incongruence of a few hundred gene trees from *Buchnera aphidicola* suggests that biased mutation pressures and enhanced substitution rates may severely distort tree topologies (Canbäck *et al*., 2004). Incorrect nodes are often supported by high bootstrap support values, which further complicates the interpretations of the phylogenomic analyses. Great care should therefore be taken before drawing conclusions from phylogenomic studies.

8. Conclusions

Phylogenomic methods offer a wonderful opportunity to perform sophisticated analyses of large sequence data sets in an automatic manner, studies that have until recently been considered unrealistic. Now that the bioinformatic methods are in place, biological expertise is required to pose interesting questions about the structure of life – and to interpret the results obtained in a meaningful manner.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Canbäck B, Tamas I and Andersson SGE (2004) A phylogenomic study of endosymbiotic bacteria. *Molecular Biology and Evolution*, **21**, 1110–1112.

Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2128.

Larsson P, Oyston PC, Chain P, Chu MC, Duffield M, Fuxelius HH, Garcia E, Halltorp G, Johansson D, Isherwood KE, *et al*. (2005) The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nature Genetics*, **37**, 153–159.

Sicheritz-Ponten T and Andersson SGE (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, **29**, 545–552.

Thompson JD, Higgins DG and Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Woese CR (1994) There must be a prokaryote somewhere: microbiology's search for itself. *Microbiological Reviews*, **58**, 1–9.

Woese CR, Kandler O and Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4576–4579.

Short Specialist Review

Mapping mutations on phylogenies

Rasmus Nielsen
Cornell University, Ithaca, NY, USA
Bioinformatics Centre, University of Copenhagen, Copenhagen, Denmark

1. Introduction

The role of comparative genomic data analysis is often to make inferences regarding the pattern and distribution of mutations. For example, conserved nucleotide or amino acid sites can easily be identified if the number of mutations occurring on a phylogeny can be inferred (*see* Article 41, **Phylogenetic profiling**, Volume 7). Likewise, studies aimed at detecting correlated amino acid substitutions (e.g., Shindyalov *et al*., 1994) or positively selected sites (e.g., Bush *et al*., 1999) are easily conducted if mutations can be inferred on a phylogeny. In the following, we will refer to a set of mutations inferred on a phylogeny (a character history) as a *mapping of mutations*. In an alignment of n sequences, each sequence is associated with a leaf node (terminal node) in a binary unrooted tree with $2n - 3$ branches (edges). An example of such an unrooted tree for $n = 5$ is depicted in Figure 1. For a single site (a column in the alignment), each branch in the phylogeny is associated with zero or more mutations, such that the pattern of mutations in the tree is compatible with the sequence data. In the example of Figure 1, the observed data for a single site is *GCCGT*. For the particular tree shown in Figure 1, any mapping of mutations requires a minimum of three mutations. One of the (infinitely) many possible mappings are shown in Figure 1, requiring two $C \Leftrightarrow G$ and a $G \Leftrightarrow T$ mutation. In this notation, mutations are bidirectional because the tree is unrooted. In rooted trees, the direction of the mutation can be inferred and mutations of type $i \rightarrow j$ will be distinguished from mutations of type $j \rightarrow i$.

2. Parsimony mappings

It is clear that if the true mapping of mutations on a phylogeny is known it is easy to make inferences regarding rates and patterns of substitution. Standard statistical tests can be used to determine if certain branches in the phylogeny or certain sites in the alignment have an overrepresentation of mutations. Likewise, we can test if certain types of mutations occur more frequently than other mutations or

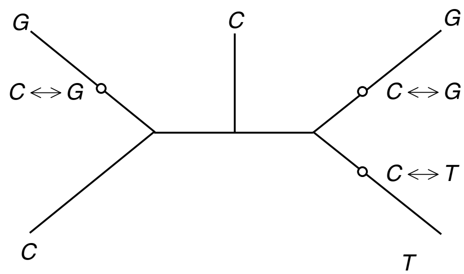


Figure 1 An example of a mutational mapping on a tree with five leaf nodes

if mutations in certain sites tend to co-occur with each other. Such methods have been used extensively in the literature, for example, for detecting selection (e.g., Fitch *et al*., 1997; Bush *et al*., 1999; Wyckoff *et al*., 2000), detecting correlated evolution (e.g., Shindyalov *et al*., 1994), or for making inferences regarding the distribution and pattern of mutations among sites (e.g., Wakeley, 1993, 1994). A real and important challenge in this regard is that the true mapping of mutations will not be known. It must instead be inferred using a statistical methodology.

The most commonly used approach for mapping mutations on a phylogeny based on a given sequence alignment, is to use the maximum parsimony principle. According to this idea, a mapping of mutation should be chosen that minimizes the total number of mutations in the tree. The mapping shown in Figure 1 is a maximum parsimony mapping because no other valid mapping exists requiring fewer mutations than three. However, another parsimony mapping would assign the two $C \Leftrightarrow G$ mutations to the branches leading to the leaf nodes associated with C. A maximum parsimony mapping is not necessarily unique.

Mutational mappings have been used in a number of exciting publications. For example, by mapping mutations on phylogenies based on sequences from the influenza hemagglutinin gene, Fitch *et al*. (1997) and Bush *et al*. (1999) were able to demonstrate that the ancestral lineages of the influenza virus, which lead to new epidemics, have particular properties with respect to the mutations that occur on them. This opened up the possibility that it might be possible to predict which of the current influenza strains may lead to the next epidemic.

Another example of the use of parsimony mappings is in the study of correlated evolution. If certain amino acid (aa) residues in a protein interact, they may be expected to evolve in a correlated fashion. There has, therefore, been considerable interest in identifying such correlated aa residues (e.g., Altschuh *et al*., 1987; Shindyalov *et al*., 1994; Pollock *et al*., 1999). In evolutionary biology, there has long been considerable interest in identifying methods for detecting correlated character evolution, of morphological as well as molecular characters (e.g., Ridley, 1983; Felsenstein, 1985; Maddison, 1990; Harvey and Pagel, 1991). Mapping of mutations are highly useful for this purpose because they may help identify pairs of sites (characters) in which mutations (character changes) tend to co-occur in the same lineages of a phylogeny (e.g., Ridley, 1983; Maddison, 1990; Pagel, 1994; Shindyalov *et al*., 1994).

3. Probabilistic approaches

Although parsimony mapping has been a very useful tool for many years in the study of molecular evolution, it is, from a statistical perspective, not ideal. The most likely mapping may in many cases not be a parsimony mapping, and the use of a methodology that always tends to choose the fewest number of mutations, may in some cases bias the results of a study (e.g., Nielsen, 2002). Even more importantly, the use of parsimony mappings usually does not take into account the uncertainty associated with the mappings, that is, the inferred mutations are treated as real data even though they are associated with some statistical uncertainty. Many of these problems can be circumvented by instead focusing on methods that implicitly sum over all possible mappings. Among such methods are the, now classical, likelihood methods for making inferences about molecular evolution (e.g., Felsenstein, 1981; Yang, 1994). In these methods, the evolution of a nucleotide site or an amino acid site is described as a continuous time Markov chain model. The Markov model is defined by the state space, $\Phi = \{A, C, T, G\}$, and an infinitesimal generator (rate matrix) $\mathbf{Q} = \{q_{ij}\}$, $i,\ j \in \Phi$, where q_{ij} is defined as $q_{ij} = -\sum_{k:k \neq i} q_{ik}$ for $i = j$, and the stationary distribution of the process is given by $\pi = (\pi_A, \pi_C, \pi_T, \pi_G)$. Notice that $-q_{ij}$ is the rate of substitution out of state i. Inferences are made by superimposing this Markov process along the branches of the tree. By multiplying transition probabilities, from the root, along the branches of the tree, summing over all possible unknown ancestral states in the internal nodes, the likelihood function under any Markov model of evolution can be calculated. When transition probabilities of the Markov chain are calculated, this calculation implicitly sums over all possible mapping of mutations. However, in many cases it may still be desirable to focus on mapping of mutations, for example, for exploratory data analysis or for analysis of models or hypotheses that are not computationally tractable using methods that analytically sums over all possible mappings. In such cases, it may, from a statistical standpoint, be more attractive to focus on the probability distribution of mappings of mutations instead of a single parsimony mapping.

Given a fixed phylogeny (topology and branch lengths) and a specified Markov model of evolution, the probability of observing a particular mapping of mutations M given the sequence data D, can easily be calculated (Nielsen, 2001, 2002). For a single site, D consists of a sequence of nucleotides of lengths n and M is a set of mutations in which each mutation is identified with respect to type (e.g., $C \Leftrightarrow G$), branch and time (position on the branch). The conditional probability density of M can be calculated as

$$p(M|D) = \frac{p(M, D)}{p(D)} \tag{1}$$

where $p(D)$ is the usual likelihood function, although the parameters have been suppressed in the notation here, and it can be calculated using standard methods (e.g., Felsenstein, 1981). $p(M, D)$ can be calculated by noting that if the current state of the chain is i, the waiting time to the next mutation is exponentially distributed with parameter $-q_{\text{ii}}$, and the probability that the mutation is of type

$i \to j$, given that a mutation happens, is $-q_{\mathrm{ij}}/q_{\mathrm{ii}}$. For time-reversible models, $p(M, D)$ can then be calculated by placing the root of the tree arbitrarily in any node of the tree and multiplying the appropriate density functions and probabilities along the branches of the tree starting in the root node. For example, consider the tree and mapping of mutations in Figure 2. By multiplying densities and probabilities along the edges of the tree from the root, we find for this particular tree and mapping of mutations

$$\begin{aligned}\Pr(M, D) &= \pi_G \left(\mathrm{e}^{qGG^{t_1}}\right)\left(\mathrm{e}^{qGG^{t_2}}\right)\left(\frac{qGC}{-qGG}\right)\left(-qGG\mathrm{e}^{qGG^{t_3}}\right)\left(\mathrm{e}^{qCC^{t_{42}}}\right)\\ &= \pi_{GqGC}\mathrm{e}^{qGG}(t_1 + t_2 + t_3) + qCC^{t_{42}} \qquad (2)\end{aligned}$$

In general, the probability density of a particular mapping is given by

$$p(M, D) = \pi_a \mathrm{e}^{\sum_{i \in \Phi} q_{ii} T_i} \prod_{(j,k):j \neq k} q_{jk}^{x_{jk}} \qquad (3)$$

where a is the ancestral nucleotide in the root imposed by M, x_{jk} is the number of mutations in the rooted tree of type $j \to k$, and T_{i} is total time in tree in which the Markov process is in state i. Under models of evolution where the embedded point process is a Poisson process, that is, $q_{ii} = q_{jj}$, for all i and j, Pr(M,D) reduces to

$$p(M, D) = \pi_a \mathrm{e}^{q_{aa} T} \prod_{(j,k):j \neq k} q_{jk}^{x_{jk}} \qquad (4)$$

where T is the total tree length, that is, $T = \sum_{i \in \Phi} T_i$. Notice that in these models, the position of a mutation along a branch does not influence the probability of the mapping.

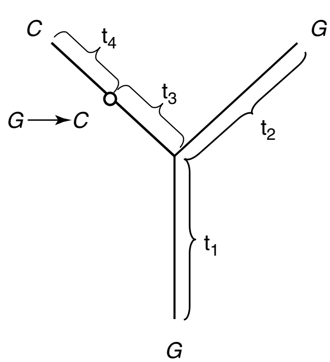


Figure 2 A mutational mapping on a tree with three leaf nodes and times of mutations indicated on the tree

4. A numerical example

As an example, consider the phylogeny in Figure 2 and assume $t_1 = t_2 = t_3 + t_4 = 1$. For simplicity of the example, further assume a Jukes and Cantor (1969) model of evolution, in which $q_{ij} = 1/3$ for all $i \neq j$ and $\pi_i = 1/4$ for all i. Then, using standard methods (e.g., Felsenstein, 1981) we find that the probability of the observed data (C, G, G) is $p(D) = 0.0161$. Likewise, we can calculate the joint probability density of the depicted mapping and the data as $p(M, D) = (1/4)e^{-1\times 3}(1/3) = 0.0041$. So the probability density associated with the mapping depicted in Figure 2 is $0.0041/0.0161 = 0.2571$, for any particular value of t_3. Integrating over all possible values of t_3, we find that the probability that a mapping with a single $G \Leftrightarrow C$ mutation is the correct mapping is 0.2571. Had we instead assumed $t_1 = t_2 = t_3 + t_4 = 0.1$, we would have found this probability to be 0.9290. The probability of different mappings depends on the branch lengths, and with sufficiently long branches the probability that any of the parsimony mappings is the right mapping can become very small. Nielsen (2002) gave an example in which a nonparsimonious mapping was more than 1000 as likely to be the right mapping as the parsimony mapping.

This basic approach for assigning probabilities to particular mappings also allows simulation of mappings of mutations and statistical inferences based on mapping of mutations. An efficient simulation algorithm for simulating mutational mappings from $p(M, D)$ was described in Nielsen (2002). In brief, using Felsenstein's (1981) pruning algorithm, the conditional probabilities of the nucleotide states in all nodes are calculated. Starting from the root, the ancestral states at each node are then simulated on the basis of these probabilities. Finally, given the ancestral state at each node, the mutational path along an edge can easily be simulated from the correct distribution using standard Markov chain theory.

In the probabilistic approach, inferences regarding the pattern of substitutions are done by focusing on a distribution of mappings instead of a single estimated mapping. In this way, statistical uncertainty in the mapping can be taken into account. By simulating mappings of mutations from the posterior, $p(M|D)$, estimates of parameters are obtained, with associated measures of statistical uncertainty. Hypothesis testing can be done by comparing the inferred distribution of mappings to the distributions expected under a given null hypothesis. Uncertainty in the parameters of the model of molecular evolution, and uncertainty regarding tree topology and branch lengths, can also be taken into account by combining the mapping of mutations approach with the now classical Bayesian approaches in phylogenetics. These concepts will be illustrated in the following sections using an application developed by Huelsenbeck *et al*. (2003) for detecting correlated character evolution.

5. Application to the detection of correlated evolution

Huelsenbeck *et al*. (2003) analyzed molecular data and the evolution of flower morphology and self-incompatibility in the family Pontederiaceae (Kohn *et al*., 1996; Graham *et al*., 1998). Although this example involves nonmolecular data,

the application of the method would work similarly for molecular data. The data consisted of *rbcL* and *ndhF* genes and data on flower morphology and self-incompatibility versus self-compatibility for each of 25 species. To distinguish the molecular from the nonmolecular data, we will use the notation D_{M} for the molecular data and D_{N} for the nonmolecular data. The major biological question investigated was to which degree the flower morphology and self-compatibility/incompatibility tend to coevolve along the branches of a phylogeny (Kohn *et al*., 1996). In particular, it was of interest to examine if self-incompatible species tend to have tristylous flowers.

To take into account uncertainty regarding tree branch length and topology, Huelsenbeck *et al*. first estimated (sampled from) the posterior distribution of trees (T) based on chloroplast *rbcL* and *ndhF* genes using the Markov Chain Monte Carlo (MCMC) algorithm implemented in MrBayes (Huelsenbeck and Ronquist, 2001). For details regarding models of molecular evolution and MCMC parameters, please see Huelsenbeck *et al*., 2003. This analysis generated a sample of k (= 8000) trees sampled from the posterior distribution, $p(T|D_M)$, that is, the distribution of trees given the molecular data. For each tree, mappings of character changes of the nonmolecular characters were simulated from the posterior distribution, $p(M_i|T_i, D_{NM})$, $i = 1, 2, \ldots, k$. This was done under a Markov model of evolution applicable to nonmolecular characters (Lewis, 2001). These mappings represented samples from $p(M|D_{NM}, D_M)$, and more than one mapping of character histories could have been sampled from each simulated tree. For each of these character mappings, a test statistic ($S^{(i)}$) was calculated. In this case, the degree of correlated evolution between the two characters was of interest. Given character mappings of two characters, the amount of time two character states are co-occurring on the tree can be directly observed. Likewise, the time the two characters states are expected to co-occur can be calculated under the hypothesis of independence in the evolution of the two characters. The test statistic was then defined as the difference between the expected and the observed proportion of times the two character states co-occur on the phylogeny. Finally, the average value of the statistic was calculated, $S = \sum_{i=1}^{k} S^{(i)}$. This procedure is illustrated in Figure 3(a).

To determine the significance of this test statistic, new data sets of the nonmolecular characters are simulated under the posterior distribution of trees (and other parameters). This was done by first simulating a set of trees from $p(T|D_M)$ and then simulating a set of new nonmolecular data sets on these trees. For each of the new simulated data sets, the previous procedure for calculating S was applied. This was again done by sampling from the posterior distribution of trees given the molecular data. This procedure is illustrated in Figure 3(b).

The results of this procedure on the Pontederiaceae data are shown in Figure 4. Because the observed value of S lies to the right of the simulated values of S, we reject the null hypothesis of independence in the evolution of the two characters and interpret this as evidence that the two characters evolve correlated.

This procedure has several advantages over previous procedures for detecting correlated evolution. Most importantly, it takes into account uncertainty both in the phylogeny and in the assignment of character mappings to the phylogeny. The exact

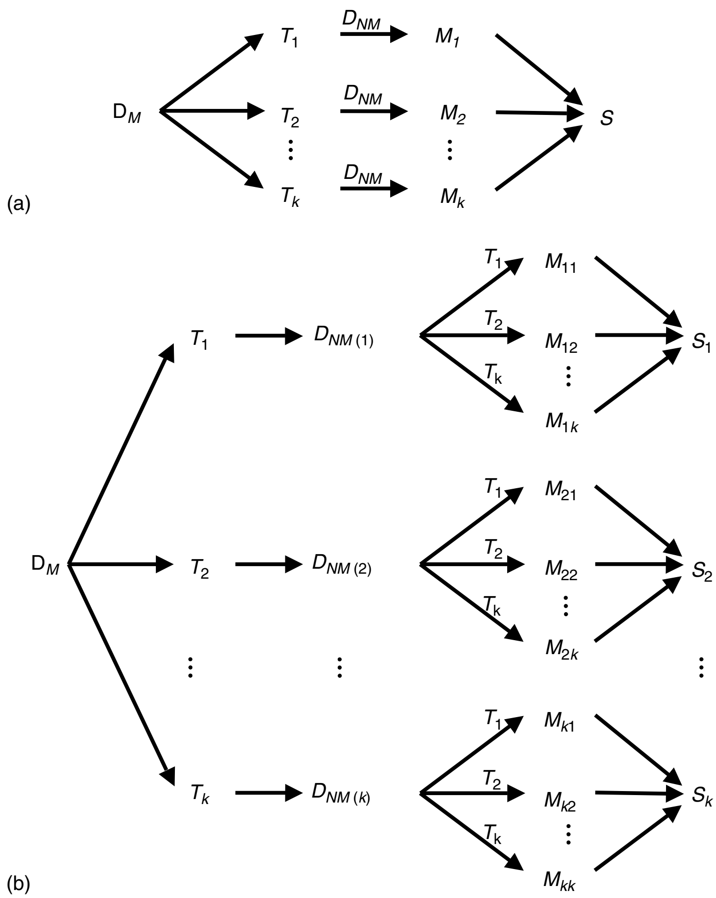


Figure 3 The procedure used to (a) calculate the test statistic and to (b) approximate its posterior predictive distribution. In (a), a sample of trees is generated from the posterior distribution given the molecular data. For each tree, a mapping of character histories are simulated from the distribution of mappings given the tree and the nonmolecular data and the test is then calculated for each mapping and averaged over mappings. (b) A similar procedure is used to generate the predictive distribution, but here the test statistic is calculated from each of k data sets of nonmolecular simulated under the posterior distribution of trees (and other parameters), under the assumption of independence between the two characters

same procedure could be applied to molecular data, replacing the nonmolecular character data with nucleotide or amino acid characters.

6. Conclusions

Mutational (character) mappings have been of tremendous importance in evolutionary biology. New probabilistic approaches can incorporate uncertainty regarding the mutational mappings into the existing approaches for analyzing data on the basis of character mappings. The new approaches are attractive because they can incorporate statistical uncertainty, but they have the disadvantage of relying on simulation,

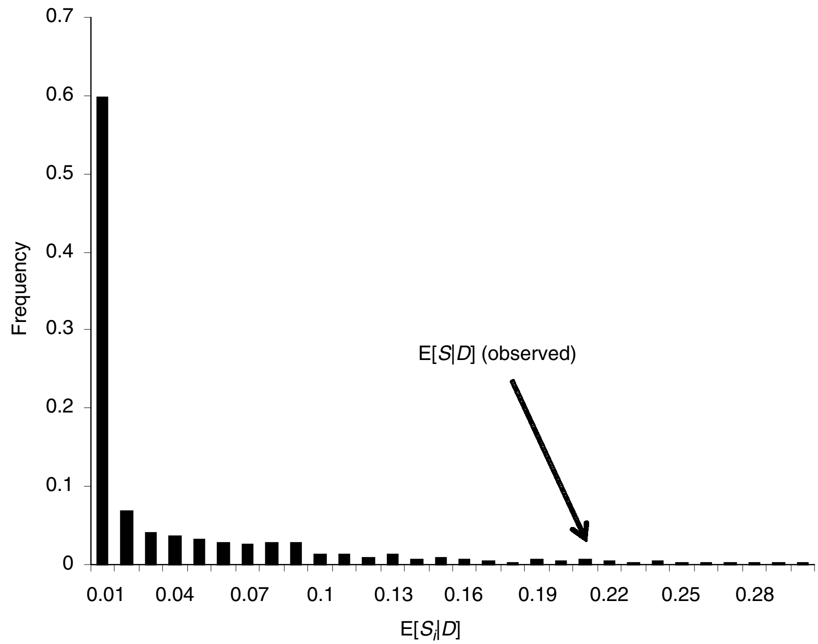


Figure 4 The simulated distribution and observed value of $E[S|D]$. $E[S|D]$ is the expected value of the correlation statistic S, where the expectation is taken with respect to the posterior distribution of mutational mappings. The simulated distribution is obtained by simulating data sets under the posterior predictive distribution

and therefore are less easy to implement and more computationally demanding than simpler methods. The parsimony-based approaches are implemented in many available computer programs including the Maclade package (Maddison and Maddison, 2000) and the Mesquite package (Maddison and Maddison, 2004), both of which offer convenient graphical user interfaces. Two computer programs are distributed that allow inferences regarding mutational mappings based on the described probabilistic approaches: Mutonphyl (Nielsen and Todd, 2004) with an interface by Nuin and Golding 2005), and SIMMAP (Bollback, 2005). The latter program has a graphic user interface that allows the user to explore distributions of mutational mappings on trees.

With the availability of new methods for mapping mutations that can take statistical uncertainty into account, mutational mappings are likely to keep their importance as one of the main inferential tools in the evolutionary analysis of molecular data.

Acknowledgments

The author gratefully acknowledges the HFSP, NSF, NIH, and the Danish National Science Foundation for grant support.

References

Altschuh D, Lesk D, Bloomer AC and Klug A (1987) Correlation of coordinated amino-acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, **193**, 643–707.

Bollback JP (2005) *SIMMAPv1.0*, Available at **http://www.simmap.com**.

Bush RM, Fitch WM, Bender CA and Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution*, **16**, 1457–1465.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist*, **125**, 1–15.

Fitch WM, Bush RM, Bender CA and Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences*, **94**, 7712–7718.

Graham SW, Kohn JR, Morton BR, Eckenwalder JE and Barrett SCH (1998) Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. *Systematic Biology*, **47**, 545–567.

Harvey PH and Pagel M (1991) *The Comparative Method in Evolutionary Biology*, Oxford University Press: Oxford.

Huelsenbeck JP, Nielsen R and Bollback JP (2003) Stochastic mapping of morphological characters. *Systematic Biology*, **52**, 131–158.

Huelsenbeck JP and Ronquist F (2001) MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, **17**, 754–755.

Jukes TH and Cantor CR (1969) Evolution of protein molecules. In H. N. Munro, ed., *Mammalian Protein Metabolism*, pp. 21–132, Academic Press, New York.

Kohn JR, Graham SW, Morton B, Doyle JJ and Barrett SCH (1996) Reconstruction of the evolution of reproductive characters in Pontederiaceae using phylogenetic evidence from chloroplast DNA restriction-site variation. *Evolution*, **50**, 1454–1469.

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

Maddison WP (1990) A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree. *Evolution*, **44**, 539–557.

Maddison WP and Maddison DR (2004) *Mesquite: A Modular System for Evolutionary Analysis*. Version 1.01 Available at **http://mesquiteproject.org**.

Maddison DR and Maddison WP (2000) *MacClade 4: Analysis of Phylogeny and Character Evolution*, Sinauer Associates: Sunderland.

Nielsen R (2001) Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics*, **159**, 413–422.

Nielsen R (2002) Mapping mutations on phylogenies. *Systematic Biology*, **51**, 729–739.

Nielsen R and Todd MJ (2004) *Mutonphyl*. Available at **http://evol.biology.mcmaster.ca/paulo/mutonphyl.php**.

Nuin PAS and Golding BG (2005) *GUI for Mutonphyl*. Available at **http://evol.biology.mcmaster.ca/paulo/mutonphyl.php**.

Pagel M (1994) Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London, Series B*, **255**, 37–45.

Pollock DD, Taylor WR and Goldman N (1999) Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal Molecular Biology*, **287**, 187–198.

Ridley M (1983) *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating*, Oxford University Press: Oxford.

Shindyalov IN, Kolchanov NA and Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, **7**, 349–358.

Wakeley J (1993) Substitution rate variation among sites in hypervariable region-1 of human mitochondrial DNA. *Journal Molecular Evolution*, **37**, 613–623.

Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. *Molecular Biology and Evolution*, **11**, 436–442.

Wyckoff GJ, Wang W and Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309.

Yang Z (1994) Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, **39**, 105–111.

Short Specialist Review

Phylogenetic analysis of BLAST results

Fiona S. L. Brinkman
Simon Fraser University, Burnaby, BC, Canada

BLAST (Basic Local Alignment Search Theorem; Altschul *et al*., 1990; Altschul *et al*., 1997) is the most widely used bioinformatics application developed to date (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7). It has been used extensively by both large genome projects and individual researchers alike to perform rapid heuristic identification of regions of sequences in a database that are similar to a user-provided query sequence. Its speed, ability to identify short regions of similarity (rather than global alignment of full length sequences), and free distribution have played a large role in its popularity. However, research literature is fraught with examples of ways in which BLAST analyses may be misused or misinterpreted (Peri *et al*., 2001; Pertsemlidis and Fondon, 2001; Ponting, 2001).

As with most bioinformatic analyses, BLAST uses evolutionary theory as the basis for its main assumptions. Therefore, an understanding of evolutionary theory is critical to the appropriate interpretation, and further analysis, of its results. Results should be interpreted in an evolutionary context, to fit with the assumptions being made about sequence divergence over time. Therefore, subsequent phylogenetic analysis of a BLAST result is frequently a good complement to a BLAST-based analysis. Phylogenetic analysis essentially allows one to more accurately view the relationship between sequences provided by a "one-dimensional" BLAST result. A BLAST result is essentially a unidirectional catalog of how similar a sequence is to a list of other sequences, while phylogenetic analysis, which presents results in a tree-based format, allows one to compare the degree of relationship between all sequences in the set. The "two-dimensional" view of a phylogenetic analysis can be helpful in, for example, identifying whether there are subsets of sequences in a BLAST output that could be grouped as a family. BLAST alone, for example, cannot identify whether three sequences found to be equally similar to a sequence are (Option 1:) actually all highly related to each other, or are (Option 2:) quite divergent from each other (but still equally related to the query sequence). Such information provided by phylogenetic analysis is critical when inferring function based on sequence similarity.

It has also been well documented that "the closest BLAST hit is not often the nearest neighbour" (Koski and Golding, 2001). That is, the sequence that is listed

first in a BLAST output as being similar to a query sequence is not necessarily the most similar sequence according to a phylogenetic analysis. Knowing how to handle such issues and appreciating the need to perform robust evolutionary analyses after a BLAST analysis is an important component of high-caliber bioinformatics research.

In addition to the accurate identification of the true "top hit" (most similar sequence to your query) and delineation of gene families, phylogenetic analysis is also useful for the functional assignment of orthologs and paralogs. Orthologs, representing the genes that diverged because of speciation, tend to share a more similar function (at some level) versus paralogs that have diverged because of a gene duplication event (for an overview, see Brinkman, 2004). While this issue is still being actively studied, it only makes sense that paralogs would diverge in function to some degree, since after a gene duplication event, selection for new traits in a duplicate gene may occur with less risk of the changes being deleterious, since there is another duplicate gene still performing the same function. Taken another way, there may be more selection for a paralog that has diverged in function and supplies a new, useful function versus one that stays the same, and so, again, paralogs may tend to diverge more in function versus orthologs. Of course, the specific divergence in function that may occur, and degree of divergence, can vary greatly for paralogs, and so this is a complex issue. Regardless, it has been found useful to differentiate such paralogous genes from orthologs because of their differing potential to diverge in function and because ortholog comparisons also facilitate more accurate cross-species comparisons. However, a simple BLAST analysis does not provide a clear indication of potential orthologs and paralogs. Phylogenetic analysis of BLAST results, coupled with a comparison of associated organismal phylogenetic relationships, can help reveal which genes are the likely orthologs and paralogs and aid both gene function assignment and analysis of the evolutionary history of the genes involved.

Performing phylogenetic analyses of BLAST results has become easier, now that semiautomated, Web-based methods have been developed that facilitate the creation of rough phylogenetic views of the data with a few clicks. PhyloBLAST (Brinkman *et al*., 2001) and WebPHYLIP (Lim and Zhang, 1999) provide user-friendly implementations of PHYLIP (Felsenstein, 1995) for downstream phylogenetic analysis. PhyloBLAST permits one to click on entries in a BLAST analysis (BLASTp only) and then perform an automated ClustalW multiple sequence alignment followed by phylogenetic analysis using a selection of programs from the PHYLIP package. WebPHYLIP is another Web-based implementation of the PHYLIP package, and there are many more similar resources available through an extensive collection of phylogenetic software listed with the PHYLIP website (**http://evolution.genetics.washington.edu/phylip/software.html**). Implementations focusing more on ortholog identification have also been initiated (**http://www.bork.embl-heidelberg.de/Blast2e/**; Yuan *et al*., 1998); however, often, ortholog identification is performed using programs unlinked to BLAST analysis. In general, there is still a lack of software available to facilitate rapid phylogenetic analysis of a BLAST result, or similar database search results (FASTA etc.).

Though programs such as PhyloBLAST can be useful for initial analysis, it must be emphasized that phylogenetic analysis must often be customized for a given experiment or research question (Brinkman, 2004; Brinkman and Leipe, 2001). While the above tools can be useful for giving a general representation of similarities between sequences, often, more customized analyses must be performed to attempt to answer a particular research question. For example, a query protein may be similar to a set of proteins according to a BLAST analysis, and subsequent PhyloBLAST analysis indicates that a subset of the sequences reflect a possible gene family that contains a mixture of putative orthologs and paralogs, relative to the query sequence. However, you may need to perform further phylogenetic analyses of subsets of these sequences, perhaps using a relevant domain of the protein only, and using other phylogenetic methods (i.e., maximum likelihood) and model tests to more accurately identify the set of sequences that may belong to the given gene family. To identify orthologs, this analysis would also need to be compared to a phylogenetic tree of the relevant organisms to compare divergence of the proteins with the species divergence. Finally, one may wish to examine particular indels (insertions and deletions) or other specific changes in the sequences in more detail and compare them to the phylogenetic analysis to identify functionally important indels/sequences and deduce how they may have evolved. Models of evolution may need to be changed to suit the data (reviewed in Brinkman and Leipe, 2001).

The benefits of phylogenetic analysis are relevant not only for further analysis of a BLAST output but also for similar database searches performed using other algorithms. Such downstream analysis of database search results is becoming increasingly important as sequence databases increase in both size and complexity. Now it is becoming common for a BLAST output to contain many copies of similar/redundant sequences. Databases searched also contain strong biases in terms of what sequences are available from what organisms (for example, there is a dearth of protozoan genomic sequence versus animal sequences). Such biases can be more easily viewed in some cases through phylogenetic analysis. However, while sequence databases are growing in size at significant rates, we are only beginning to understand the true diversity of life that exists on earth. Environmental genomic sampling, and other studies of diversity, are confirming that we have only scratched the surface of the true complexity and variety of sequences that exist in nature (Tyson *et al*., 2004; Venter *et al*., 2004). Therefore, rapid database searches, coupled with linked phylogenetic analysis and additional evolutionary analysis, will become increasingly necessary to help us grapple with the continually increasing onslaught of sequence data.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Brinkman FSL (2004) Phylogenetic analysis. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, *Third Edition*, Baxevanis AD and Ouellette BFF (Eds.), John Wiley & Sons, pp. 365–392.

Brinkman FSL and DD Leipe (2001) Phylogenetic analysis. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, *Second Edition*, Baxevanis AD and Ouellette BFF (Eds.), John Wiley & Sons, pp. 323–358.

Brinkman FSL, Wan I, Hancock REW, Rose AM and Jones SJ (2001) PhyloBLAST: a tool facilitating phylogenetic analysis of BLAST results. *Bioinformatics*, **17**, 385–387.

Felsenstein J (1995) *PHYLIP (Phylogeny Inference Package)*, University of Washington: Seattle.

Koski LB and Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, **52**, 540–542.

Lim A and Zhang L (1999) WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*, **15**, 1068–1069.

Peri S, Ibarrola N, Blagoev B, Mann M and Pandey A (2001) Common pitfalls in bioinformatics-based analyses: look before you leap. *Trends in Genetics*, **17**, 541–545.

Pertsemlidis A and Fondon JW III (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biology*, **2**, reviews2002.1–reviews2002.10.

Ponting CP (2001) Issues in predicting protein function from sequence. *Briefings in Bioinformatics*, **2**, 19–29.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Yuan YP, Eulenstein O, Vingron M and Bork P (1998) Towards detection of ortholgues in sequence databases. *Bioinformatics*, **14**, 285–289.

Short Specialist Review

Connecting genes by comparative genomics

Itai Yanai
Weizmann Institute of Science, Rehovot, Israel

A result of deep importance to modern molecular genetics is the observation that most biological functions are carried out by the products of multiple genes and that most genes are associated with multiple functions (Hartwell *et al*., 1999). Unraveling the functional relationships among genes thus provides a promising route to the molecular understanding of biological functions.

Examples of tight functional relationships among genes are ubiquitous. Metabolic pathways are composed of multiple gene products each corresponding to a step along the pathway. Genes of the same metabolic pathway or signaling cascade can be said to be functionally linked. Importantly, there is frequent cross-talk among the pathways, such that a gene may subserve multiple pathways: human gene GPI, glucose phosphate isomerase, is found in glycolysis, pentose phosphate, and starch and sucrose pathways. To complicate matters further, complexes of several gene products, each of which can also have membership in multiple other types of such complexes (Gavin *et al*., 2002), may actually accomplish a single metabolic step. Besides pathways, cohorts of genes regularly form functional systems such as the oxidative phosphorylation system and the flagellum. These systems are intricately structured complexes and here also we find that members of one system are actually modular and can be used in the context of multiple functional systems as in the type III secretion system and the flagellum (Gophna *et al*., 2003).

Ideally, all such interactions would be identified by experiments testing for all possible interactions among the genes. Indeed, high-throughput methods – such as yeast two-hybrid (Ito *et al*., 2001; Uetz *et al*., 2000) and tandem-affinity purification (TAP) coupled with mass spectrometry (Gavin *et al*., 2002) – are extremely useful in identifying functional links on a large scale. However, as high-throughput as these methods may be, they cannot be truly complete. Since there are on the order of N^2 possible interactions, where N is the number of genes in a genome, it is not feasible to actually test all interactions. Moreover, these methods exhibit high false-positive rates, are highly work-intensive, and to date they mostly have been used to identify protein–protein physical interactions, a subset of the more general type of functional relationships described above.

An alternative complementary approach to the brute force experiments is not unlike the one applied to comparative anatomy in the nineteenth century. In 1843,

Sir Richard Owen, the Superintendent of the British Museum, compared the skeletons of the various classes of vertebrates – fish, reptiles, birds, and mammals. From his comparisons, Owen boldly concluded that vertebrates have a common structural plan he called an “archetype”. Owen introduced the term *homology* to mean “the same organ in different animals under every variety of form and function” (1843). Thus, human fingers are homologous to bird fingers. While originally interpreting similarities as evidence of a creator’s “archetypes”, with Darwin’s introduction of evolutionary theory, homology took on a new significance. In light of evolution, homology is interpreted as similarity due to common ancestry. By comparing increasingly distant organisms, it is thus possible to trace the order of appearances of anatomical features. We find that an understanding of the historical narrative of a biological phenomenon can be of tremendous insight toward its functioning. Essentially, the relationships among a skeleton’s bones are revealed by exploiting the common ancestry of extant organisms through a comparison of the disposition of their homologies.

Relationships among a genome’s genes may similarly be revealed by analyses of their disposition in other genomes. As expected, the operational unit of comparison in such an approach is a specific kind of homology called *orthology*, most easily defined as “the same gene in different genomes”. More exactly, an inference of orthology between two genes signifies a relationship of speciation. We infer that the two genes stem from one precursor gene that was present in the most recent common ancestor before the speciation event that dispersed the two lineages.

The ability to identify orthologous relationships proved to be of revolutionary importance to genetics. The reason is that an orthologous relationship provides a strong hint toward functional conservation. Although the length of time since divergence is a correlating factor, the notion that orthologs generally maintain the same function has found wide-range experimental support. Thus, orthology is the great simplifying notion of genetics; it offers the hope that each gene in every organism will not need to be studied separately to be understood. The knowledge that one has acquired by hard work in one organism can thus be transferred to its ortholog in another organism. For example, frog histones need not be thoroughly studied because their structure and function in yeast has already been worked out in detail and the general mechanisms can be transferred. An important caveat, however, is that orthologous genes often evolve to take on vastly different functions. Nonetheless, the delineation of orthologous gene families is of such fundamental importance that it should be considered the first law of functional genomics.

Working with orthologous families of genes allows for the reformulation of the goal of identifying gene relationships. Instead of seeking to link individual genes, it is most efficient to link orthologous families. Three methods have been proposed to detect relationships among families based upon complete genomes.

The phylogenetic distribution of the members of an ortholog family is informative. A given gene family may be, for example, specific to bacteria, scattered across the microbes, or present in all organisms examined. Consider that with 100 genomic sequences, there are 2^{100} (on the order of 10^{30}) possible distributions. Specific enough to be informative, a family’s phylogenetic distribution may be taken as its fingerprint. An ortholog’s phylogenetic distribution – or phylogenetic profile – however is often not unique to it alone. In cases where another family’s profile is

identical, or comparable, it may be a lucky incidence of no biological importance or it may actually serve as a hint of a relationship between the two families.

Correlated presences and absences in the same genomes of members of a pair of ortholog families may be explained by a functional link. In other words, if the pair of genes is required for a certain function, a genome would not be expected to have only one of them. Thus, tightly coupled genes would indeed be expected to have identical profiles of occurrences (Date and Marcotte, 2003; Huynen and Bork, 1998; Pellegrini *et al*., 1999; Tatusov *et al*., 1997; Wu *et al*., 2003). The bacterial flagellum is a good example of a functional system with conserved phylogenetic profiles (Pellegrini *et al*., 1999). The principle of phylogenetic profiling has been used to identify novel pathways (Date and Marcotte, 2003).

The physical proximity of a pair of genes along the chromosome may also be of functional importance. Chromosomal proximity of genes may be maintained by selection due to a shared mechanism of expression such as operons, or perhaps, as is common in Bacteria and Archaea, because their proximity offers a better survival rate for horizontal transfer of the paired genes (Lawrence, 1999). In both instances, a functional link is appropriate. On the other hand, proximity between two genes may be of no biological importance. Adopting a comparative genomic approach, these two instances may be distinguished. It has been shown that the number of genomes in which proximity was found is correlated to the accuracy of the functional relationship prediction (Yanai *et al*., 2002). Thus, for a given pair of ortholog families, we might examine the relative proximity of the pair across genomes (Overbeek *et al*., 1999). If in a certain number of genomes the genes are proximate, we consider the two gene families linked. Interestingly, the families may be linked even if the proximities are not held in all of the genomes and, consequently, a pair of genes in one genome may be said to be linked by chromosomal proximity even if they are not actually proximate (Yanai *et al*., 2002).

Unfortunately, functional links by chromosomal proximity seem to hold only for Bacteria and Archaea examined thus far. This may be due to the fact that operon organization is most significantly a phenomenon that is limited to these kingdoms. Eukaryotic genomes tend to be less streamlined, and thus chromosomal organization may be dominated by different evolutionary forces. Regardless of this limitation, the general applicability to Bacteria and Archaea is very useful.

A third type of functional link takes advantage of an extreme case of chromosomal proximity. It has been observed that two genes that are distinct in one organism may be found as one contiguous fusion gene in another organism (Enright *et al*., 1999; Marcotte *et al*., 1999). A fusion gene generally unites genes of the same metabolic pathway or protein complex. Thus, even if nothing is known about the two genes, a functional link based upon a fusion elsewhere could be inferred on the basis of the fusion event in one (or more) of the genomes. Indeed, it has been statistically shown that fusion links tend to pair up genes of the same broad functional category (Yanai *et al*., 2001).

The major difficulty with the fusion method is that it often appears to be a special case of a general mechanism of modular genes. Most human genes, as well as those of many other genomes, are composed of multiple domains – independent structures serving as evolutionary conserved building blocks (Ponting and Russell,

2002). Since domains, such as SH2 domains for example, appear in many different kinds of genes, there is no reason to suppose a direct functional link between the domain partners. Thus, the specificity of the fusion method can be increased if fusion links are constrained to only those domains that do not appear to be promiscuous (Marcotte *et al*., 1999).

A major confounding phenomenon of these three methods involves the issue of gene duplication. Very frequently, a gene has multiple orthologs in another genome, signifying that the original gene underwent duplication since speciation (Jordan *et al*., 2001). The question of how gene function evolves following a duplication event is, in fact, one of the most interesting open questions in the genomics field today. Since functional conservation within an orthologous family is the first assumption in all the methods, the applicability of the three genomic context methods may be called into question when gene duplications have occurred across the lineages.

How applicable are phylogenetic profiling, chromosomal proximity, and gene fusion, to a genome's genes? On average, functional links in microbial genomes contain an average of 57% of an organism's complete genetic complement (Yanai and DeLisi, 2002). Interestingly, there is very little overlap in terms of links generated by the methods but considerable overlap in the genes involved in the links (Yanai and DeLisi, 2002). Thus, the links are largely additive to form networks of interactions. The links uncover substantial portions of known pathways, and suggest the function of previously unannotated genes.

The three comparative genomics methods described here apply evolutionary theory to unravel the functional organization of genomes. Comparative genomics can essentially be seen as a decent substitute for a time machine. While ancestral states of an organism would only be available to us with the use of a time machine, comparing extant genomes leads to inferences of these very states. As in other realms, knowledge of the past makes way for understanding the present.

References

Date SV and Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**, 1055–1062.

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gophna U, Ron EZ and Graur D (2003) Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, **312**, 151–163.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Huynen MA and Bork P (1998) Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5849–5856.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574.

Jordan IK, Makarova KS, Spouge JL, Wolf YI and Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Research*, **11**, 555–565.

Lawrence J (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current Opinion in Genetics & Development*, **9**, 642–648.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–2901.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Ponting CP and Russell RR (2002) The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, **31**, 45–71.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wu J, Kasif S and DeLisi C (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**, 1524–1530.

Yanai I and DeLisi C (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biology*, **3**, research0064.

Yanai I, Derti A and DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7940–7945.

Yanai I, Mellor JC and DeLisi C (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends in genetics*, **18**, 176–179.

Chromosome phylogeny

Guillaume Bourque
Genome Institute of Singapore, Singapore

1. Introduction

Impressive sequencing and comparative mapping endeavors have made available detailed whole-genome sequences and maps (Lander *et al*., 2001; Venter *et al*., 2001; Waterson *et al*., 2002). One of the stated goals of these projects is to better our understanding of organismal biology through comparative analyses (O'Brien *et al*., 1999). The rationale is that insights into the dynamics of evolution can help ascertain the underlying adaptive biochemical and physiological responses. Comparisons can be at the raw sequence level, where the focus is typically on the analysis of individual genes, or at the coarse genomic level where the focus is on large-scale rearrangement events (*see* Article 42, **Reconstructing vertebrate phylogenetic trees**, Volume 7 and Article 44, **Phylogenomic approaches to bacterial phylogeny**, Volume 7). Dobzhansky and Sturtevant (1938) pioneered the latter type of analysis almost 70 years ago in a study of inversions in Drosophila pseudoobscura.

Since point mutations and rearrangement events act as parallel modes of evolution, they can be exploited concurrently to provide complementary perspectives on the relationships between the genomes under study. There are two important advantages of studying rearrangements, or chromosomal mutations: first, rearrangement analyses capture the evolution of a chromosome or a genome as a whole and not just locally. Second, because chromosomal mutations are rare events (Rokas and Holland, 2000), and because they are selected from a large set of possible candidates (e.g., there is a quadratic number of potential inversions that can affect a genome), they are less likely to suffer from homoplasy. This allows the reconstruction of scenarios even for deep phylogenies.

2. Algorithms and recent applications

There are different types of rearrangement events: some affect gene order (inversions, translocations, fusions, fissions, transpositions and inverted transpositions); while others affect gene content (insertion, deletion and duplication). Various methods have been developed to efficiently compute the rearrangement distance and sort a pair of genomes with equal gene content under different sets of operations; for

instance, with inversions only (Hannenhalli and Pevzner, 1999; Bader *et al*., 2001; Bergeron, 2001), with inversions, translocations, fusions, and fissions (Hannenhalli and Pevzner, 1995; Tesler, 2002), and with transpositions (Meidanis *et al*., 1997; Bafna and Pevzner, 1998). Some of these methods were then further adapted for multiple genome comparisons and phylogeny reconstruction under a maximum parsimony criterion: with inversions only (Siepel and Moret, 2001) and with inversions, translocations, fusions, and fissions (Bourque and Pevzner, 2002). See Figure 1 for an example.

Even without specifying a model of rearrangement events, the relative gene order can be used to define the breakpoint distance that can then be exploited, again under a parsimony criterion, for phylogenetic tree reconstruction (Blanchette *et al*., 1997; Blanchette *et al*., 1999). Although models that are not restricted to equal gene content are more comprehensive (see El-Mabrouk, 2005 for a review), they are typically more challenging algorithmically and have been restricted to a limited number of phylogenetic applications (Sankoff *et al*., 2000; Earnest-DeYoung *et al*., 2004).

Prior to the availability of detailed large genome sequenced data, rearrangement studies were restricted to the analysis of gene orders in genomes like mitochondria or chloroplasts (Palmer and Herbon, 1988; Sankoff *et al*., 1992; Bafna and Pevzner, 1995; Blanchette *et al*., 1999; Cosner *et al*., 2000). With the significant progress in large-scale sequencing and comparative mapping, rearrangement studies had to be reconfigured into a two-step process:

1. Identification of homologous syntenic blocks (HSBs) shared by the set of genomes under study.
2. Comparison of the respective arrangements of these common blocks in the different genomes.

The HSBs that need to be identified in step 1 can be obtained either from information on orthologous genes (Zdobnov *et al*., 2002) or directly from homologous sequences (Kent *et al*., 2003; Pevzner and Tesler, 2003a). In both cases, a threshold needs to be set to allow the HSBs to extend over small local discrepancies that could originate from different sources: erroneous assignment of orthologous genes (especially in large gene families), assembly errors, smaller rearrangement events falling outside the model (e.g., transposons), and so on. There are advantages to using orthologous genes to identify HSBs: it focuses the analysis on important regions of the genome, the thresholds are length independent and the blocks are less sensitive to repeat sequences. Similarly, there are advantages to using raw sequence data: it avoids annotation problems, it works in noncoding regions, it is less sensitive to gene families and finally it preserves more information on microrearrangements (rearrangements within HSBs) that can then be used as additional phylogenetic characters (Bourque *et al*., 2005).

Once the HSBs are identified, they can be processed in step 2 using algorithms for genome rearrangements in the same way genes were used previously. This two-step methodology was used to compare the human with the mouse genome (Pevzner and Tesler, 2003a) and suggested a larger number of rearrangements than previously expected (mostly intrachromosomal rearrangements). It also helped refine a model

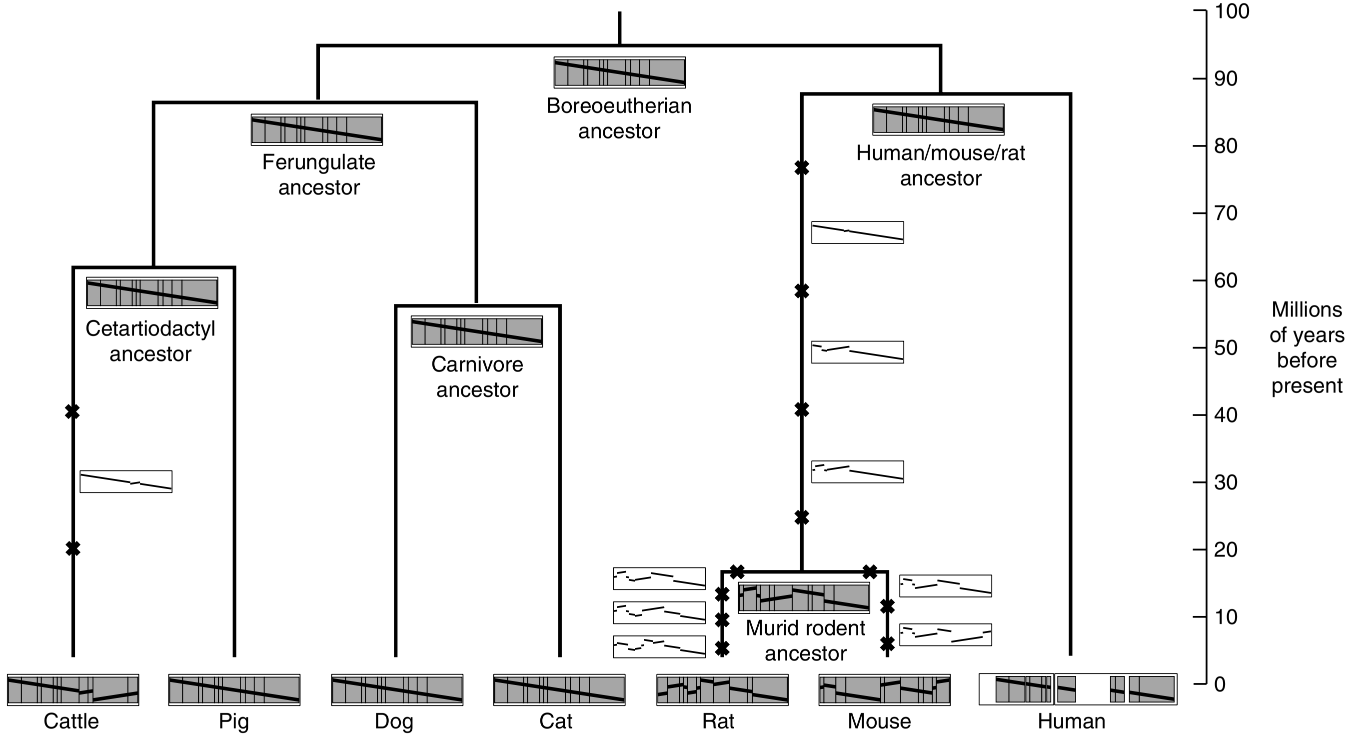


Figure 1 Mammalian X chromosome phylogeny. The arrangements of 11 homologous syntenic blocks, identified on the X chromosome of seven contemporary mammalian genomes (human, mouse, rat, cat, dog, pig, and cattle), are shown at the bottom of the tree. Blocks are drawn proportionally to their size in human and gaps are only shown in human to display coverage. A diagonal line traverses the blocks to show their order and orientation relative to human. The top of the tree exhibits the putative X chromosome ancestors in a most parsimonious inversion scenario as recovered by MGR (Bourque and Pevzner, 2002). The occurrence of an inversion is shown with a small cross on a branch of the tree but the exact timing of events on that branch is unknown. Hypothetical intermediate X chromosomes on a path are displayed using a white background. Data adapted from Murphy *et al*. (in press)

for chromosome evolution in which some breakpoints are reused nonrandomly (Pevzner and Tesler, 2003b; Larkin *et al*., 2003).

When more than two genomes are compared, rearrangement scenarios not only lead to the inference of phylogenetic relationships, they also provide rate estimates for the different branches of the tree and allow for the reconstruction of the putative architecture of ancestral genomes. A comparison of the human, mouse, and rat genomes (Bourque *et al*., 2004) confirmed the accelerated rate of interchromosomal rearrangements previously observed in lower-resolution studies of rodent genomes and predicted the genomic architecture of the murid rodent ancestor. The addition of the chicken genome acting as an outgroup (Bourque *et al*., 2005), allowed the reconstruction of the putative mammalian ancestral genome and further localized the lineage-specific chromosomal mutations. The analysis further revealed highly variable rates of genomic rearrangements across different branches of the tree with a particularly slow rate of interchromosomal rearrangements in the chicken lineage, in the early mammalian lineage, or in both.

Applications focusing on specific areas of the genomes allow for the identification of very detailed scenarios and possibly provide clues into the rearrangement mechanisms themselves. In the comparison of large vertebrate genomes, the X chromosome is a very interesting subproblem for studying chromosome evolution as it rarely exchanges genetic material with other chromosomes. Figure 1 shows the evolution of the X chromosome in seven mammalian genomes (human, mouse, rat, cat, dog, pig, and cattle); an example adapted from Murphy *et al*. (in press). This example highlights once again the unstable rate of genome rearrangements: about 85% (11/13) of the large-scale inversions affecting the X chromosome are found on the rodent branches whose total length only represents about 20% of the total length of the branches of the tree (corresponding to 500 Mya of evolution). What triggered and what are the consequences of this accelerated rate of inversions on the rodent X chromosome? These questions will require further work.

3. Limitations and future prospects

Many of the algorithms and applications described above still rely on a very simplistic model of evolution in which a limited set of equally likely operations are considered. Given the wealth of genomic data now available, it would be interesting, and challenging, to revisit these models, parameterize the different operations (in a way similar to what was done for mitochondria genomes (Blanchette *et al*., 1996)), and estimate their preponderance from the data itself. Such an experiment could potentially lead to a more realistic biological model of rearrangements but more importantly, it might provide new insights into the dynamics of genome evolution. In terms of making the rearrangement models more realistic, integrating other sources of information such as centromere and telomere positions could also prove valuable.

A related topic involves the development of a Bayesian framework for genome rearrangements. The methods described above have relied on a most parsimonious criterion but, in some cases, finding a most parsimonious scenario is insufficient.

For difficult problems (i.e., highly nonadditive trees), there are typically many optimal solutions and even the assumption that the actual history of rearrangement corresponds to a most parsimonious scenario becomes weak. A Bayesian approach has already been suggested for unichromosomal genomes (Larget *et al*., 2005) but has yet to be adapted to multichromosomal genomes. Such an approach would provide a richer description of the solution space but it would also allow a more natural parameterization of the rearrangement operations.

An emerging approach for chromosome phylogeny, similar in spirit to the analysis of breakpoints in that it does not require the specification of a rearrangement model, involves the analysis of conserved intervals in unichromosomal genomes (Bergeron *et al*., 2004). Although this approach does not analyze rearrangements directly, it can also identify phylogenetic relationships and reconstruct ancestral architectures by maximizing the conservation of the relative order of subsets of genes. It has the potential to be applicable to a wide variety of questions.

Another desirable improvement would involve the efficient expansion of genome rearrangement studies to genomes with unequal gene, or sequence, content. Although some preliminary studies have been moderately successful (Sankoff *et al*., 2000; Earnest-DeYoung *et al*., 2004), a tool to compare systematically whole eukaryotic genomes, not only on a set of common HSBs, is still needed. An intermediate solution could be to adapt multigenome studies to only require pairwise identification of common blocks.

As alternative methodologies for the analysis of genome rearrangements emerge, new systematic ways for comparing not only recovered phylogenies but also prediction at ancestral nodes will be required. Putative ancestors are difficult to compare because, in different studies, they are reconstructed using different sets of HSBs. So far, quality assessments have been limited to the analysis of predicted chromosomal associations (Bourque *et al*., 2004; Bourque *et al*., 2005; Murphy *et al*., in press) but it would be interesting to analyze more thoroughly the robustness of the rearrangement scenarios recovered when different choices of parameters, or completely different approaches, are used to generate HSBs.

4. Conclusion

Finally, as more and more detailed sequences and maps are released, lineage-specific breakpoint regions will be more systematically defined simply by the identification of HSBs. A recent study by Murphy *et al*. (in press) confirms that around 20% of breakpoint regions are reused and that those breakpoint regions are enriched for centromeres. The same study also finds interesting correlations between frequently rearranged regions, gene density, segmental duplications and recurring cancer breakpoints. Obviously, these evolutionary breakpoints represent an extremely rich source of information on the mechanisms regulating genome rearrangements and what can lead to their fixation in a population. As in-depth analyses of breakpoints regions are pushed further, the hope is that crucial information on these mechanisms will be uncovered. The challenge will then be to refine models and algorithms for genome rearrangements to employ that information proficiently.

Acknowledgments

The author is funded by the Agency for Science, Technology and Research (A*STAR) of Singapore.

Further reading

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Milan D, Beever JE, Chowdhary BP, Galibert F, *et al Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps*, In press.

References

Bader DA, Moret BME and Yan M (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology*, **8**(5), 483–491.

Bafna V and Pevzner PA (1995) Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, **12**, 239–246.

Bafna V and Pevzner PA (1998) Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, **11**(2), 224–240.

Bergeron A (2001) A very elementary presentation of the Hannenhalli-Pevzner theory. *The 12th Annual Symposium on Combinatorial Pattern Matching*, LNCS 2089, Springer-Verlag: Berlin, pp. 106–117.

Bergeron A, Blanchette M, Chateau A and Chauve C (2004) Reconstructing ancestral gene orders using conserved intervals. *WABI 2004*, Bergen, Norway, LNBI 3240, pp. 14–25.

Blanchette M, Bourque G and Sankoff D (1997) Breakpoint phylogenies. In *Genome Informatics Workshop (GIW 1997)*, Miyano S and Takagi T (Eds.), Universal Academy Press: Tokyo, pp. 25–34.

Blanchette M, Kunisawa T and Sankoff D (1996) Parametric genome rearrangement. *Gene*, **172**, GC11–GC17.

Blanchette M, Kunisawa T and Sankoff D (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, **49**, 193–203.

Bourque G and Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, **12**, 26–36.

Bourque G, Pevzner PA and Tesler G (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Research*, **14**(4), 507–516.

Bourque G, Zdobnov EM, Bork P, Pevzner PA and Tesler G (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Research*, **15**(1), 98–110.

Cosner ME, Jansen RK, Moret BME, Raubeson LA, Wang L, Warnow T and Wyman SK (2000) A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. *Proceedings of ISMB*, San Diego, pp. 104–115.

Dobzhansky T and Sturtevant A (1938) Inversions in the chromosomes of Drosophila pseudoobscura. *Genetics*, **23**, 28–64.

Earnest-DeYoung JV Lerat E and Moret BME (2004) Reversing gene erosion – reconstructing ancestral bacterial genomes from gene-content and order data. *Workshop on Algorithms in Bioinformatics*, Bergen, Norway, pp. 1–13.

El-Mabrouk N (2005) Genome rearrangements with gene families. In *Mathematics of Evolution and Phylogeny*, Gascuel O (Ed.), Oxford University Press.

Hannenhalli S and Pevzner PA (1995) Transforming men into mice: polynomial algorithm for genomic distance problem. *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, Milwaukee, Wisconsin, pp. 581–592.

Hannenhalli S and Pevzner PA (1999) Transforming cabbage into turnip: polynomial algorithm for sorting by reversals. *Journal of ACM*, **46**, 1–27.

Kent WJ, Baertsch R, Hinrichs A, Miller W and Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11484–11489.

Lander E, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Larget B, Simon DL, Kadane JB and Sweet D (2005) A Bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, **22**(3), 486–495.

Larkin DM, Everts-van der Wind A, Rebeiz M, Schweitzer PA, Bachman S, Green C, Wright CL, Campos EJ, Benson LD, Edwards J, *et al*. (2003) A cattle–human comparative map built with cattle BAC-Ends and human genome sequence. *Genome Research*, **13**, 1966–1972.

Meidanis J, Walter ME and Dias Z (1997) Transposition distance between a permutation and its reverse. *WPM '97*, Valparaiso, Chile, pp. 70–79.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE and Graves JA (1999) The promise of comparative genomics in mammals. *Science*, **286**, 458–481.

Palmer JD and Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **27**, 87–97.

Pevzner PA and Tesler G (2003a) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, **12**, 37–45.

Pevzner PA and Tesler G (2003b) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 7672–7677.

Rokas A and Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution*, **15**, 454–459.

Sankoff D, Leduc G, Antoine N, Paquin B, Lang B and Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 6575–6579.

Sankoff D, Deneault M, Bryant D, Lemieux C and Turmel M (2000) Chloroplast gene order and the divergence of plants and algae, from the normalized number of induced breakpoints. In *Comparative Genomics*, Sankoff D and Nadeau JH (Eds.), Kluwer, pp. 89–98.

Siepel A and Moret BME (2001) Finding an optimal inversion median: experimental results. *Proceedings of the First International Workshop on Algorithms in Bioinformatics*, LNCS 2149, Aarhus, Denmark, pp. 189–203.

Tesler G (2002) Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences*, **65**(3), 587–609.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1352.

Waterson RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, *et al*. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, **298**, 149–159.

Integrating statistical approaches in experimental design and data analysis

Ernst Wit and Raya Khanin
University of Glasgow, Glasgow, UK

1. Introduction

The main objectives of experimental design are reducing variation, eliminating bias, and making the analysis of the data and the interpretation of the results as simple and as powerful as possible. This should occur with the purpose of the experiment in mind and within the practical constraints. Principles of statistical design of experiments were first introduced by Ronald A. Fisher in the 1920s for designing agricultural field trials. The relevance of his ideas has only increased since then. This article elucidates the main principles of designing efficient experiments and the statistical analysis thereof. It uses microarray experiments as an example.

2. Design principles: between variation and bias

The common aim of many scientific experiments is to establish the activity of a *response* of interest across several *conditions* or *subpopulations*. For example, how do DNA sequence repeats vary across different positions on the human genome? Sometimes, the conditions or populations of interest are defined through an active intervention of the part of the scientist. In this case, conditions are typically referred to as *treatments*. For example, plant scientists may be interested in the change in expression profiles for Arabidopsis plants as a result of fertilizer, seed dormancy, or vernalization treatments. Sometimes, a proxy for the response is necessary if the response itself is not directly measurable. Average spot intensity on a microarray is typically taken as the operative definition of gene expression. Different proxies might have different characteristics, and sometimes it is useful to consider several proxies simultaneously.

Despite having a measurable quantity, it is typically infeasible to measure that quantity for the whole population of interest, nor is it likely that this quantity for all members of the population will be the same. This necessitates taking a *sample* and, at least as important, one that is *representative* of the population as a whole.

Representative means that it should embody the population variation, without any bias toward any particular group within it.

The ideas behind a representative sample are crucial for experimental design and can be described by means of several *design principles* (*see* Article 52, **Experimental design**, Volume 7):

1. *Replication:* When access to experimental material is limited or requires substantial resources, there might be the temptation to make several measurements on the same or a nearby experimental unit. Unfortunately, such observations are almost certainly very similar and therefore fail to reflect the full variation present in the sample. True replication, sometimes called *biological* replication in the context of biological experiments, is a necessary condition for obtaining a representative sample.
2. *Randomization:* If experimenters are let to decide for themselves which experimental units receive which treatment, it is possible that intentionally or unintentionally one condition is favored over another. By making the assignment of the units to the conditions subject to, for example, a coin flip, any such bias of the effects and their variances can be avoided.
3. *Blocking:* Often, it is impossible to get homogeneous experimental materials, in which case a subdivision in more or less homogeneous *blocks* should be considered. Microarrays come in print batches, DNA pools come from PCR amplification runs, and so on. Even though blocking factors are not of primary interest, it is important not to ignore the variation they introduce. Typically, it is prudent to model blocking factors such as print batches explicitly.
4. *Crossing:* If one experimental condition is always observed jointly with another, then it is impossible to decide whether the observed effect in the response is due to one or the other. This is called *confounding*. In order to avoid this type of ambiguity, it is essential to *cross* conditions with each other, that is, making sure that the levels of experimental conditions do not always appear in tandem.
5. *Balance:* It makes intuitive sense that in order to get the best results, each condition should be measured more or less an equal amount of times. Although not essential for inference, this type of *balance* has many computational advantages and is generally recommended wherever possible.

Following these general principles, one converts unplanned systematic variability into planned, chance-like variability. It protects against possible bias, and makes subsequent statistical analysis possible.

3. Several standard designs

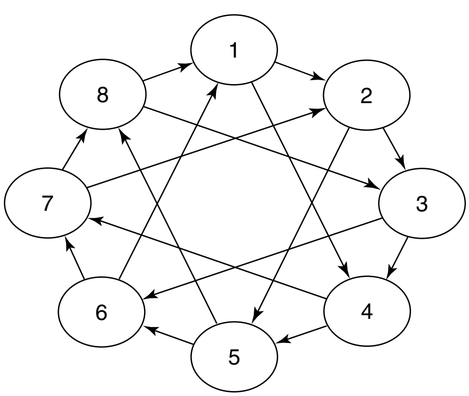
To implement the design principles, one should be led first and foremost by the requirements of the experimental situation. Sometimes, it is possible to make use of several standard designs that satisfy automatically all the necessary design conditions.

If the experimental sample is from a homogeneous population, then no blocking is needed and the conditions can be randomly assigned to the experimental units.

This is called a *completely randomized design* and is most useful in only the simplest of scientific experiments. A commonly used design is the *randomized complete block design*, developed in 1925 by R. A. Fisher. In this design, the experimental units are subdivided into more or less homogeneous "blocks" and within each block the treatments are randomly assigned to the units.

Sometimes, the blocks are too small in order to apply all the conditions of interest within each block. This is typically the case in microarray studies, where each array is a block and at most two conditions can be measured within each array. *Incomplete block designs* (IBD) are aimed to deal with this situation. A Latin Square design is an elegant example of an IBD. It crosses the condition of interest with two blocking factors in such a way that the design is balanced, but not each combination of factors is observed.

Two-channel microarray experiments can be regarded as experimental designs with block-size two. If more than two conditions are of interest, then the assignment of conditions to arrays can be particularly tricky. Many standard incomplete block designs do not apply in this case. A simple class of interwoven loop designs has been suggested (Kerr and Churchill, 2001b) and shown to have excellent efficiency properties (Wit *et al*., 2005). Figure 1 shows an example of an interwoven loop design for eight conditions, which consists of lining up the conditions in the an imaginary circle (in any order) and by hybridizing each condition with the condition that is 1 step and with the condition that is 3 steps down the circle respectively. An interwoven loop design should always contain a 1-step circle in order to guarantee that all the conditions are connected. The other jump sizes can either be chosen intuitively or with some optimization function, such as od in smida library in R (Wit and McClure, 2004). An additional advantage is that by its very definition interwoven loop designs are dye-balanced and avoid the need of using inefficient dye-swap designs.



A-optimality score for contrasts: Tr[Inv(X′X)] = 13.
No of arrays = 16 No of conditions = 8.

Figure 1 An interwoven loop design is a special case of an *incomplete block design* with block-size two. This design has excellent efficiency properties and is balanced in the "nuisance" dye factor

4. Normalization and estimation of effects

If an experiment has been designed according to the above design principles, then *Analysis of Variance* (ANOVA) is a coherent technique to untangle the effects of interest from the nuisance effects (*see* Article 53, **Statistical methods for gene expression analysis**, Volume 7 and Article 59, **A comparison of existing tools for ontological analysis of gene expression data**, Volume 7). For example, in a microarray experiment, one would like to estimate the contribution of a particular condition to the gene expression of a particular gene, while correcting for the dye used for hybridization, the pin used for printing and the position of the spot on the array, and perhaps the batch number from which the microarray came. Classical ANOVA is by no means the only technique that is able to estimate the effects of interest. More recent (hierarchical) Bayesian modeling approaches are capable of doing very much the same thing. In this discussion, we focus on the classical ANOVA.

First of all, it is essential to determine the response of interest. Gene expression is typically measured as an optical intensity across several pixels (dual-channel arrays) or several probes (Affymetrix Gene Chip). Typically, these values are summarized in a single expression value, either by taking the mean, median, or by performing a somewhat more complicated operation (e.g., Li *et al*., 2003). Intensity is an essentially positive and often a multiplicative scale. We recommend applying a logarithmic transformation to the intensities before analysis (see also Kerr and Churchill, 2001a), although generalized log-transformations have also proved useful (Huber *et al*., 2003). For two-channel arrays, each array gives two expression values for each gene. Some studies interpret each channel as an independent observation (Kerr and Churchill, 2001a), whereas others prefer to take the log-ratio of each of these two values (Wit and McClure, 2004). The latter is safer, although more conservative.

Anything that contributes to variation in the response is known in the statistical literature as an *effect*. The idea behind ANOVA is to write the observed experimental quantity as a sum of contributing elements, which are the result of the varying experimental conditions. An *ideal* microarray experiment across n_c conditions and n_g genes would be modeled as

$$y_{ij} = \mu + G_i + C_j + (GC)_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, n_g,\ j = 1, \ldots, n_c \tag{1}$$

where one typically assumes that the noise term ε is normally distributed with a fixed variance. The quantity μ stands for the overall expression level across the arrays G_i, measures the average deviation from that expression level μ for gene i, C_j, measures the average deviation from μ for condition j. Most importantly, $(GC)_{ij}$ stands for the differential expression of gene i with respect to level j. In particular, if for a particular gene i the quantities $(GC)_{ij}$ are zero across all j, then this gene is not differentially expressed under the experimental conditions. The term $(GC)_{ij}$ is an example of an *interaction effect*. This interaction effect allows the response to vary *differently* across the conditions for each of the genes.

Clearly, the model in equation (1) is too simplistic. Besides a random error term, practitioners know that in many experiments there are also systematic nuisance

effects. These nuisance terms are related to experimental conditions only partially under the control of the experimenter, such as, for example, a dye effect, a spatial effect, a pin effect, an array effect, or a batch effect. Under the assumption that each of these factors influences the signal additively, a more complete model for describing the observed gene expression involves additional terms for each of them, which is given by

$$y_{ij} = \mu + G_i + C_j + (GC)_{ij} + S_{s(i)} + B_{b(j)} + \cdots + \varepsilon_{ij}, \tag{2}$$

where $s(i)$ and $b(j)$ are functions that indicate the position of gene i on the array and the batch number of array j respectively. If the experiment uses printed dual-channel arrays, then also a dye term and a print pin term can be added. One can also add interaction terms of these nuisance effects. For example, a pin-batch interaction effect would model the fact that in a different microarray batch, the pin effects are probably different as the pins have most likely changed between different batch runs (Kerr and Churchill, 2001a).

The key element in an ANOVA is the F-test, evaluating the effect of each of the terms. These F-tests are typically summarized in an ANOVA table. See Table 1 for an example. Given that the number of terms in an ANOVA is limited, multiple testing is not particularly an issue and we recommend using the usual 0.05 cutoff for significance. Terms that have an associated p-value (last column) less than the cutoff are deemed important. Biologically most important are the individual estimates of the differential expression terms, that is, the condition-gene interactions. The individual differential expression terms can be put to further scrutiny.

There are two types of effects: so-called fixed effects and random effects. Fixed effects are most common and assume that the condition or treatment underlying the effect is some individual entity. Examples are gene effects, dye, and spatial effects. Random effects are used to model the effects of elements that can be considered to come from a larger population. The microarray batch used in an experiment can be considered as a particular sample from all possible batches made by the particular producer. Random effects are used when the effect of the individual levels of the condition or treatment are not of any particular interest, but the overall influence of them collectively is something to be taken into consideration. When both fixed

Table 1 ANOVA table for Arabidopsis example, using log gene expression ratios for 100 genes across the four distinct plant types and two different regions involving 16 microarrays

	Df	Sum sq	Mean sq	F-value	Pr(>F)
Dye	1	1314.42	1314.42	1302.02	0.000
Pin	3	1.09	0.36	0.36	0.782
Condition	7	72.81	10.40	10.30	0.000
Gene	96	101.01	1.05	1.04	0.378
Array	8	90.70	11.34	11.23	0.000
Condition* gene	693	3930.93	5.67	5.62	0.000
Type*gene	297	3474.29	11.70	11.58	0.000
Region*gene	99	134.99	1.36	1.35	0.018
Type*region*gene	297	321.65	1.08	1.07	0.235
Residuals	792	799.54	1.01		

and random effects are used in modeling a particular, one speaks of a mixed effects model.

Wolfinger *et al*. (2001) introduced a mixed effect model in a microarray context. The difference between fixed and random effects is not always clear-cut – for example, should pins be considered fixed or random effects? Using random effects results in somewhat more conservative estimates (larger standard errors), although typically it does not affect the estimates of the fixed effects, for example, the differential expression parameters GC_{ij}.

5. Design and analysis of a microarray experiment

We illustrate all the ideas in this article by means of a practical design and analysis of a stylized dual-channel microarray experiment. An experimenter wants to compare four genetically distinct types of Arabidopsis plants across two different regions of a country. It is of interest to compare the gene-expression values both across the genetic types as well as across region. We can therefore distinguish $4 \times 2 = 8$ separate conditions. Given the availability of 16 dual-channel microarrays, the way to obtain the best estimates of differential expression for each condition-pair is as shown in Figure 1 (Wit *et al*., 2005). Typical reference designs would merely measure each condition twice, whereas this interwoven loop design measures each condition four times. The microarrays contain 100 probes, associated with 100 different genes.

The resulting dataset contains 1200 log-expression ratios. The ANOVA table, shown in Table 1, can be obtained by most standard statistical software. We first consider the nuisance effects. By evaluating the p-values, it is clear that there is an important dye and array effect, although there is no evidence that the four pins used to print the arrays have any distorting effect on the gene expressions. After correcting for the array and dye effect, it is evident from the highly significant condition-gene interaction that there are at least some genes that are differentially expressed across the eight conditions. In particular, there is a lot of evidence for differentially expressed genes across the four Arabidopsis types. There is also evidence that the region from where the plants were harvested affects the gene expression, although no synergetic effect on gene expression of genetic type and region is observed.

6. Conclusions

In order to avoid obtaining useless data, it is essential to integrate the statistical analysis with the design of any complex scientific experiment. For example, sources of variation cannot be estimated if they are confounded in the experimental design. Using ANOVA, or Bayesian equivalents, has the advantage of combining the normalization process with the data analysis. Considerations of optimal design can be employed to obtain the most accurate estimates of the parameters of interest.

References

Huber W, von Heydebreck A, Seultmann H, Poustka A and Vingron M (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, **2**(1), A3.

Kerr MK and Churchill GA (2001a) Statistical design and the analysis of gene expression microarray data. *Genetical Research*, **77**, 123–128.

Kerr MK and Churchill GA (2001b) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Li C, Tseng GC and Wong WH (2003) Model-based analysis of oligonucleotide arrays and issues in cDNA microarray analysis. In *Statistical Analysis of Gene Expression Microarray Data*, Inter-disciplinary Statistics, Spped T (Ed.), Chapman & Hall/CRC, pp. 1–34.

Wit EC and McClure JD (2004) *Statistics for Microarrays: Design, Analysis and Inference*, John Wiley & Sons: Chichester.

Wit EC, Nobile A and Khanin R (2005) Near-optimal designs for dual-channel microarray studies. accepted for publication. *Applied Statistics*, **54**(5).

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C and Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **6**(6), 625–637.

Introductory Review

Mass spectrometry and computational proteomics

Vineet Bafna
University of California, San Diego, CA, USA

Knut Reinert
Free University Berlin, Berlin, Germany

1. Introduction

Computation occupies a central role in the interpretation of high-throughput biological data, and proteomics is no exception. It can also be argued that *mass spectrometry* is now the tool of choice for proteomics, with applications to peptide sequencing, protein structure prediction, protein–protein interactions, (relative) protein expression, and many others. Continued improvements in instrumentation and computational technologies will only help accelerate this trend. Indeed, the chemistry Nobel prize in 2002 was awarded for the development of protein ionization methods in mass spectrometry. While rewarding individual accomplishments, the awards were also a recognition of the immense potential of this field.

In this overview, we summarize algorithms for interpreting mass spectrometry (MS) data. This short overview is not intended as an introduction to the technology itself or to proteomics (see for example (Siuzdak, 1996; Wilkins *et al*., 1997)). Instead, we will provide an abstract overview of MS data in order to describe key algorithmic ideas required for its interpretation. Owing to space limitations, we will concentrate on three applications: *protein identification, protein interactions*, and *differential analysis of protein expression*.

2. MS technology

The mass spectrometer is a device that measures the mass (actually, the mass-to-charge ratio) of an ionized molecule. Its key components are a source for sample introduction and ionization (typically MALDI (Matrix Assisted Laser Desorption Ionization) or ESI (ElectroSpray Ionization)), and a *mass analyzer* for measuring the mass (e.g., Ion Trap, TOF (time of flight), Quadrupole, and Fourier transform (FT_MS)). The conceptual question one might ask is the following: how can a

device that essentially measures mass be used in diverse applications like protein sequencing, expression, and structure?

To answer this, one must note that while these devices are conceptually simple, they now offer extremely high mass accuracy and resolving power. Various protocols can be used to obtain a set of characteristic masses that would be diagnostic for a protein. In *Peptide Mass Fingerprinting* (PMF), the protein is enzymatically digested and the peptide masses are recorded using a single mass spectrometric measurement (MS). Every protein will have a characteristic set of digested peptides, and correspondingly, peptide masses. For more complex mixtures, this procedure is not sufficient and instead multiple stages of mass spectrometry are applied (tandem MS (MS/MS) or MS^n).

In tandem mass spectrometry (*see* Article 4, **Interpreting tandem mass spectra of peptides**, Volume 5), the peptides (from an enzymatically digested protein mixture) are ionized with one or more units of charge, as in single-stage MS, and a specific peptide is chosen for fragmentation by *collision-induced dissociation* (CID). Fragments retaining the ionizing charge after CID have their mass–charge ratio measured in a second stage of mass spectrometry. Since peptides typically break a peptide bond when they fragment by CID, the resulting spectrum contains information about the constituent amino acids of the peptide.

The fragmentation of the peptide in CID is a stochastic process governed by the physicochemical properties of the peptide and the energy of collision. The charged fragment can be inferred by the position of the broken bond and the side retaining the charge. In Figure 1(b), the N-terminal a_1, b_1 and c_1 fragment ions and the C-terminal x_{n-1}, y_{n-1}, and z_{n-1} fragment ions are shown. While *a,b,y* represent the commonly occurring fragments, a high-energy collision often

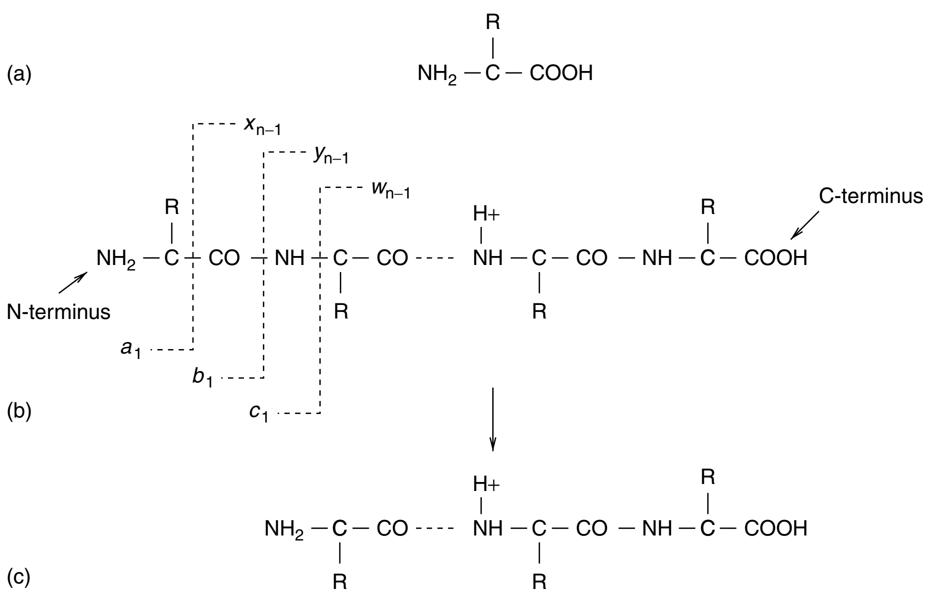


Figure 1 (a) The structure of an amino acid. (b) An ionized peptide. (c)y_{n-1}^{+} ion

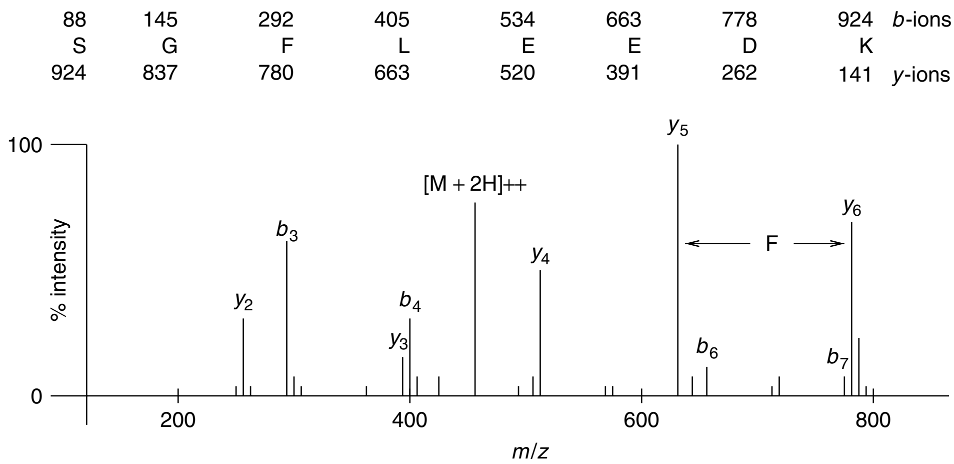


Figure 2 MS/MS spectrum for peptide SGFLEEDK

results in other fragments, including *internal* fragments formed by breakage at two points, fragments formed by breaks in side chains, and neutral molecule losses from fragments, including H_2O, and NH_3. One or more of these fragments retain the charge unit(s) and their mass–charge ratio is registered. Figure 1(c) shows the single charge being retained by y_{n-1}. In a single experiment, many charged fragments are formed by CID of multiple copies of the same peptide. The aggregate of the mass–charge ratios detected is called the *MS/MS spectrum*. A cartoon MS/MS spectrum for the peptide SGFLEEDK is shown in Figure 2. It helps illustrate how the MS/MS spectrum can be used to determine the sequence of amino acids of a peptide. Note that the difference in mass–charge ratio of the adjacent singly charged y-ions, y_5 and y_6, is exactly the mass of the residue F. The algorithmic challenge is to use the MS/MS data in the presence of noise, incomplete fragmentation, and mixed prefix, suffix, and internal ions, to provide unambiguous identification.

3. Protein identification

MS/MS might not be necessary if a relatively simple mixture of proteins is being analyzed. Instead, PMF can be used with every protein containing a characteristic set of peptide masses. Two problems complicate this simple picture. First, only a small fraction of the peptides are typically ionized and detected by MS. Second, the observed peptides are sometimes from multiple (2–10) dominant proteins in the mixture. These issues are partially resolved using a Bayesian probabilistic model that assigns a likelihood of a protein sequence, given a spectrum of peptide masses (Perkins *et al*., 1999). The protein with the highest likelihood is then reported. In order to deal with multiple proteins, the process is repeated after “removing” the peaks that matched the first protein. For many proteomic applications with complex mixtures of proteins, single-stage MS is gradually being supplanted by tandem and higher-order mass spectrometry as the more reliable method for protein sequencing.

The core of most of the software programs for analyzing tandem MS data contain an implementation of the following three modules: *interpretation, filtering*, and *scoring*. In the interpretation module, the input is an MS/MS spectrum and the output is an *interpreted spectrum* containing meta-information that can be reliably inferred from the spectrum. It may include parent peptide mass, partial or complete sequence tags, and combinations of sequence tags and molecular masses. The filtering module takes the interpreted spectrum and a peptide sequence database as input and filters out most of the peptides, leaving a small list of *candidate peptides* that might have generated the MS/MS spectrum. Finally, the scoring module takes the list of candidate peptides MS/MS spectrum as input and returns a ranking of the candidate peptides along with a score and possibly a p-value (probability that the score was achieved by random chance). If significant, the highest scoring peptide is the correct interpretation. As the peptide sequences have relatively low redundancy, the identification of the sequence of one or two peptides is usually sufficient to identify the protein. Certainly, *scoring* is the key module and is under active research (Bafna and Edwards, 2001; Elias *et al*., 2004; Eng *et al*., 1994; Tabb *et al*., 2001). However, most existing toolkits use some aspects of all three modules. With an increase in the size of databases, and growing interest in posttranslational modifications, algorithms for interpretation and filtering are coming to the fore.

3.1. De novo interpretation

We take a general view by defining interpretation to be any sequence information gleaned from an analysis of the spectrum. The class of *de novo sequencing* algorithms attempt to reconstruct the entire peptide sequence via interpretation without the use of a peptide database (see for example (Bafna and Edwards, 2003; Bartels, 1990; Chen *et al*., 2001; Dancik *et al*., 1999; Fernández de Cossio *et al*., 1995; Johnson and Biemann, 1989; Taylor and Johnson, 1997)). To understand how this is done, consider the simple case in which all observed fragments are b-type prefix ions. If all the fragment ions were present, then the sequence could be read simply by reading the ladder of increasing masses and assigning residues to mass differences between adjacent peaks. The complexity comes from the fact that (1) there are multiple fragment ion types, including internal ions and different charge states; (2) the prefix ions cannot a priori be separated from the suffix ions; and (3) not all positions are along the peptide chain fragment.

A key algorithmic idea that deals with mixed prefix and suffix ions is the prefix residue graph, first described in Dancik *et al*. (1999), but implicit in (Bartels, 1990; Taylor and Johnson, 1997). If we knew the ion type for a spectral peak, it would uniquely define the residue mass of a prefix (PRM) of the peptide. Thus, for all possible interpretations of a peak, a node labeled with the PRM value is added to the graph and a directed edge is drawn from node u to node v if PRM[v]-PRM[u] corresponds to a residue mass (adjacent fragments) or if it is close to 0 (identical fragments). Each edge is labeled with the ϵ or the appropriate amino acid. Thus, any path in this graph corresponds to a de novo sequence interpretation simply by concatenating the non-ϵ edge labels. A multitude of paths are possible, and scoring and ranking these paths is critical to correct interpretation.

One argument against this approach is that a peak may be used multiple times in a path with different interpretations. Resolving this problem in a general scenario is equivalent to constructing paths in which certain pairs of nodes are forbidden, which is known to be computationally hard (Garey and Johnson, 1979). However, Chen *et al*. (2001) made the observation that the forbidden suffix–prefix pairs for tandem MS are nonintersecting. This observation allowed them to give a dynamic programming solution for finding paths with forbidden pairs. Bafna and Edwards (2003) extended this approach to include multiple ion types (*a,b,y*, and all the neutral losses) in the interpretation. While pure de novo sequencing of spectra has seen many improvements, tandem mass spectra usually do not have enough information to make an unambiguous identification. Searching a database of candidate peptides (as described in the next section) constrains the possibilities enough to make such identification possible. However, de novo interpretation is still useful in some situations. It can be used to generate sequence tags that can then be used as additional filters to improve database search (Mann and Wilm, 1994; Tabb *et al*., 2003; Taylor and Johnson 1997), especially for modified peptides. Also, for pure proteins whose sequence is not yet available in databases, a *de novo shotgun* sequencing of overlapping peptides obtained via multiple enzymatic digestion holds promise (MacCoss *et al*., 2002).

3.2. Scoring spectra against peptides

When the protein sequence of interest is present in a database, one can interpret an MS/MS spectrum by computing the correlation between the spectrum and a hypothetical spectrum of each peptide. The so-called database searching algorithms (Bafna and Edwards, 2001; Eng *et al*., 1994; Fenyo *et al*., 1998; Mann and Wilm, 1994; Pevzner *et al*., 2000; Tabb *et al*., 2003) rely on this technique for interpretation and have been extremely successful. Sequest (Eng *et al*., 1994) is the prototypic database search method. It presents a model for the generation of hypothetical spectra, and a correlation function for scoring. If the spectrum is of poor quality, there is no guarantee that the top scoring peptide is the correct interpretation. Subsequent algorithms (Bafna and Edwards, 2001; MacCoss *et al*., 2002; Perkins *et al*., 1999) therefore included a p-value along with the raw score to give a probability of that score arising from a randomly chosen peptide. Further improvements in scoring have come from an analysis of the physicochemical rules of fragmentation, such as "neutral losses are more likely in the presence of acidic or basic residues", and "proline-directed fragmentation". The algorithm in Scope (Bafna and Edwards, 2001) presents a model for quantifying these rules as probabilities and efficient scoring with the probability functions. Recent work is directed toward data mining and learning techniques to optimize a score function, as well as the use of intensity values in scoring (Elias *et al*., 2004).

3.3. Filtering

The goal of filtering is to scan a database of peptide candidates and quickly filter out the vast majority of them while retaining the true peptides for detailed scoring.

While most algorithms include simple filters typified by parent mass, immonium ions, and matching peaks, this topic was not actively researched until recently. With exponential growth in the number of spectra and sequence databases, this is only now beginning to see active research. Pre-indexing sequence databases is useful in removing redundant peptide information and efficient search for candidates. One approach to indexing is the use of suffix trees (Edwards and Lippert, 2002; Lu and Chen, 2003). Other approaches include the use of sequence tags as filters in a database search (Mann and Wilm, 1994; Tabb *et al*., 2003; Taylor and Johnson, 1997).

4. Differential analysis of protein mixtures

Differential analysis of proteome expression levels has been developing rapidly over the last years. The current standard separation technique in proteomics is undoubtedly gel electrophoresis, which is typically used together with mass spectrometry for identification of the proteins separated on the gel. While this technique is well established and is in use in most major proteomics facilities, it has disadvantages in the context of high-throughput settings. In particular, the difficult handling of the gels prompted several proteomics facilities (mostly in commercial settings (Domon *et al*., 2002), but also in academia (MacCoss *et al*., 2003; Sickmann *et al*., 2003)) to use High Performance Liquid Chromatography (HPLC)-/MS-based techniques instead (*see* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5). In these approaches, the liquid-chromatographic separation in an HPLC column replaces the separation on the gel. In contrast to gels, HPLC systems allow for direct coupling to a mass spectrometer, thus greatly simplifying the automation of the whole analytical process. In the shotgun approach, the proteins are usually fully digested in order to simplify and unify sample preparation. The combined use of the complementary expression techniques (mRNA expression, gel-based MS, and HPLC-based MS) will yield further progress in diagnostics and systems biology.

The analysis of HPLC/MS data for differential analysis of protein expression poses many challenging computational problems. Note that measuring protein expression using MS techniques is generally difficult. Different peptides have different capacity to retain charge, and so peptides with identical concentration might show up with very different intensities in the spectrum. However, the relative intensity of peaks for the *same* peptide is predictive of relative expression of that peptide in different samples. To use this fact in differential analysis, consider the output of an HPLC/MS run. As the peptides elute off the column, MS spectra are acquired in real time. The data can be represented as a two-dimensional (2D) spectrogram, or map, with the two dimensions being LC retention time (RT) and m/z. As a peptide typically elutes over a fixed time span and has a fixed m/z, a "spot" on this map corresponds to the elution of a peptide. The intensity of the spectra provides a third dimension. Two maps from different samples (normal and diseased) will have similar spots corresponding to the commonly expressed peptides and spots of differential intensity corresponding to proteins whose expression levels

have changed. Geometric matching algorithms are used to match the similar spots, and their intensities are used for normalization, sometimes with internal standards. Once this is done, the relative intensities of differentially expressed peptides can be computed. This basic idea can be made to be statistically robust by choosing multiple samples from both categories. This is analogous to the use of 2D gels for separation with important differences. It is the peptides and not proteins that are being separated. As the retention time and mass measurements are typically accurate, the reproducibility of maps is much higher than 2D gels. Algorithms for creating and comparing various 2D maps and for computing relative intensities of a few thousand differentially expressed peptides have been used to successfully identify differentially expressed peptides in tumor cells (Domon *et al*., 2002). The key components of these algorithms include the deconvolution of peaks from different but overlapping peptides and creation, normalization, and comparison of multiple maps to identify differential expression.

An alternative approach uses some type of differential mass labeling of the two samples to provide uniform experimental conditions (*see* Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5). The two samples are labeled differently, for example, by introducing stable isotopes such as ^{15}N, ^{2}H, ^{13}C, or ^{15}O into the proteins of one of the samples (Berger *et al*., 2002; Oda *et al*., 1999). Other approaches are mass-coded abundance tags (MCAT (Cagney and Emili, 2002)) and isotope-coded affinity tags (ICAT, (Gygi *et al*., 1999)), where the mass difference is introduced by derivatization of specific amino acid residues. The samples are then mixed and subjected to HPLC/MS. In the resulting maps, each peptide should occur in paired "spots" that are very similar in RT and have a mass offset that corresponds exactly to the differential mass of the labeling tags. This simplifies the computation somewhat, as comparison of two LC/MS maps is not required. For statistical robustness, one might still need to perform multiple such experiments and compare maps in order to identify differentially expressed peptides unambiguously (Domon *et al*., 2002).

5. Protein structure: cross-linking

MS technologies are clearly creating an impact on protein identification and quantification. Can we also use them for protein structure determination? X-ray crystallography and NMR remain the dominant techniques in this field, but much work remains to be done. These techniques require large amounts of pure analyte, and even if this is available, it can take many months until the structure is determined. On the other hand, it is theoretically possible to determine the tertiary structure of a protein computationally, given enough interatomic distance information (Cohen and Sternberg, 1980). Even partial information often proves to be helpful. To acquire this distance information, the mass of *cross-linked* peptides can be measured (Young *et al*., 1997), and using this information, the sequence of the cross-linked complex can be determined. Cross-linking is a method in which certain molecules are used to specifically link two peptides. The identification of cross-linked peptides in a folded molecule then provides constraints on the spatial distance of the peptides in a folded state. One approach to identifying cross-linked

peptides is simply to find all pairs of peptides whose mass sum (plus the mass of the linker) equals the mass of the parent cross-linked molecules. Among these, a peptide pair is chosen whose theoretical spectrum best correlates with the measured spectrum (Chen *et al*., 2001). These additional constraints can be used to improve tertiary structure prediction.

6. Conclusion

We conclude by reiterating that mass spectrometric techniques are key to proteomic explorations, and computational algorithms for analyzing MS data are crucial to further development of this technology. Mass spectrometry is a dynamic and evolving field and it is likely that many of the data sets and applications described here will be outdated in the coming years. Nevertheless, it is our hope that the basic understanding of MS principles and key algorithmic components will continue to be useful for future proteomic applications of mass spectrometry.

References

Bafna V and Edwards N (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17**(Suppl 1), S13–S21. June Appeared in Intl. Conference on Intelligent Systems for Molecular Biology.

Bafna V and Edwards N (2003) *On de novo* interpretation of peptide spectra. In *International Conference on Computational Molecular Biology (RECOMB)*, The Association for Computing Machinery: New York, pp. 9–18.

Bartels C (1990) Fast algorithm for peptide sequencing by mass spectrometry. *Biomedical and Environmental Mass Spectrometry*, **19**, 363–368.

Berger SJ, Lee S-W, Anderson GA, Pasa-Tolic L, Tolic N, Shen Y, Zhao R and Smith RD (2002) High-throughput global peptide proteomic analysis by combining stable isotope amino acid labeling and data-dependent multiplexed-MS/MS. *Analytical Chemistry*, **74**, 4994–5000.

Cagney G and Emili A (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures mass-coded abundance tagging. *Nature Biotechnology*, **20**, 163–170.

Chen T, Jaffe JD and Church GM (2001) Algorithms for identifying protein cross-links via tandem mass spectrometry. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)*, The Association for Computing Machinery: New York, pp. 95–102.

Chen T, Kao MY, Tepel M, Rush J and Church GM (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **8**(6), 571–583.

Cohen FE and Sternberg MJ (1980) The use of chemically derived distance contraints in the prediction of protein structure with myoglobin as an example. *Journal of Molecular Biology*, **137**, 9–22.

Dancik V, Addona T, Clauser K, Vath J and Pevzner PA (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **6**, 327–342.

Domon B, Alving K, He T, Ryan TE and Patterson SD (2002) Enabling parallel protein analysis through mass spectrometry. *Current Opinion in Molecular Therapeutics*, **4**, 577–586.

Edwards, N and Lippert, R (2002) Generating peptide candidates from amino-acid sequence databases for protein identification via mass spectrometry. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, Springer-Verlag: Berlin; Heidelberg; New York, pp. 68–81.

Elias JE, Gibbons FD, King OD, Roth FP and Gygi SP (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, **22**, 214–219.

Eng J, McCormack A and Yates J (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of American Society of Mass Spectrometry*, **5**, 976–989.

Fenyo D, Qin J and Chait BT (1998) Protein identification using mass spectrometric information. *Electrophoresis*, **19**(6), 998–1005.

Fernández de Cossio J, Gonzales J and Besada V (1995) Protein identification using mass spectrometric information. *Computer Applications in the Biosciences*, **11**, 427–434.

Garey MR and Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company: New York.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Johnson RJ and Biemann K (1989) Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical and Environmental Mass Spectrometry*, **18**, 945–957.

Lu, B and Chen, T (2003) A suffix tree approach to the interpretation of tandem mass spectra:applications to peptides of non-specific digestion and post-translational modification, *Bioinformatics*, **19**(Suppl. 2), ii113–ii121.

MacCoss MJ, McDonals WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, *et al*. (2002) Shotgun identification of protein modifications from protein complexes and lens tissues. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(12), 7900–7905.

MacCoss MJ, Wu CC, Liu H, Sadygov R and Yates III JR (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Analytical Chemistry*, **75**(24), 6912–6921.

Mann M and Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399.

Oda Y, Huang K, Cross FR, Cowburn D and Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6591–6596.

Perkins DN, Pappin DJ, Creasy DM and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18), 3551–3567.

Pevzner PA, Dancik V and Tang CL (2000) Mutation-tolerant protein identification by mass-spectrometry. In *International Conference on Computational Molecular Biology (RECOMB)*, Shamir R, Miyano S, Istrail S, Pevzner PA and Waterman MS (Eds.), ACM Press: New York pp. 231–236.

Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, *et al*. (2003) The proteome of saccharomyces cerevisiae mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13207–13212.

Siuzdak G (1996) *Mass Spectrometry for Biotechnology*, Academic Press: San Diego; New York; Boston; London; Sydney; Tokyo; Toronto.

Tabb DL, Eng JK and Yates JR 3rd (2001) *Protein Identification by SEQUEST*, Vol. 1, Springer: New York, pp. 125–142.

Tabb DL, Saraf A and Yates JR (2003) GutenTag: High-throughput sequence tagging via am empirically derived fragmentation model. *Analytical Chemistry*, **75**, 6415–6421.

Taylor JA and Johnson RS (1997) Sequence database searches via de novo peptide sequencing by mass spectrometry. *Rapid Communications in Mass Spectrometry*, **11**, 1067–1075.

Wilkins MR, Williams KL, Appel RD and Hochstrasser DF (1997) *Proteome Research: New Frontiers in Functional Genomics*, Springer Verlag: Berlin; Heildelberg; New York.

Young M, Tang N, Hempel JC, Oshiro CM, Taylor EW and Kuntz ID (1997) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5802–5806.

Specialist Review

Experimental design

Kevin K. Dobbin and Richard M. Simon
National Cancer Institute, Bethesda, MD, USA

1. Introduction

Experimental design issues for dual-label and single-label microarray experiments are discussed, including identification of research objectives, avoidance of confounding, allotment of samples to arrays and dyes in dual-label experiments, dye bias, pooling RNA before labeling, and determination of the number of arrays required to achieve the study objectives.

2. Objectives

The first step in designing a microarray experiment is to identify the goals of the experiment. There is no one design that will be appropriate for every experiment, but the optimal design for a particular experiment will depend on the research questions being addressed. As the size and scope of microarray studies grow, so does the range of questions that researchers are asking. It is not possible to be comprehensive here in discussing study objectives, but it is useful to identify a few general types of objectives often seen in microarray research.

Class comparison objectives apply to studies of collections of specimens that come from two or more predefined classes, types, or conditions. The defining aspect is that the classes are known ahead of time, independent of the gene expression data, and that the research question is: which genes are expressed differently in the different classes? For example, Hedenfalk *et al*. (2001) compared primary tumors from two classes of women, those who carried the BRCA1 mutation genotype and those who carried the BRCA2 mutation genotype. One goal of the study was a class comparison goal, namely, to discover a set of genes that were expressed differently in the two genotypes. Other examples of studies with class comparison objectives include: (1) Golub *et al*. (1999) identified genes expressed differentially between specimens of acute myelogenous leukemia (AML) and specimens of acute lymphocytic leukemia (ALL); (2) Ross *et al*. (2000) compared expression profiles of cancer cell lines from different tissues of origin. In these studies, class comparison was often not the only goal, but there were other goals as well, such as class prediction.

Class prediction objectives can apply either to studies of collections of specimens from predefined classes or to studies in which some clinical outcome is measured

for each individual from which a specimen was obtained. The defining aspect is that the research question is: can we construct a prediction rule for the specimens that will predict, from the gene expression data alone, the likely class or outcome for an individual? The ultimate goal is usually to develop a rule that can be used on future individuals for whom the classification or outcome data is not available. The Golub *et al.* (1999) paper also provides an example of a class prediction objective; the differentially expressed genes discovered in the class comparison phase were used to develop a multigene class predictor that distinguished between AML and ALL, and the predictor was applied to an independent set of tumors to assess the predictive performance. Rosenwald *et al.* (2002) developed a molecular predictor of survival after chemotherapy on the basis of biopsy samples from diffused large-B-cell lymphoma patients.

Class discovery objectives can apply either to studies in which there are no predefined classes for the specimens, or in which the current classification system is deemed inadequate. The defining aspect is that the research question is: can we discover a new classification system for these specimens based solely on gene expression data? Bittner *et al.* (2000) applied cluster analysis techniques (*see* Article 54, **Algorithms for gene expression analysis**, Volume 7) to the gene expression profiles of a collection of melanoma samples to identify a novel classification system for this otherwise homogeneous group of specimens, suggesting a gene expression–based taxonomy.

Another type of class discovery study addresses the research question: can we use what we know about the function of some genes to figure out the function of other genes (*see* Article 60, **Extracting networks from expression data**, Volume 7)? Or, what genes are coregulated in these samples? These questions are typically approached by applying cluster analysis to the genes instead of the samples.

3. Confounding

In an ideal microarray experiment, tissue handling and cutting, RNA extraction, reverse transcription and labeling, and array hybridization would all be performed at about the same time, under identical experimental conditions. For such an experiment in a class comparison setting, observed significant differences between the classes in gene expression could safely be attributed to real biological differences between the classes. But it is common that this type of ideal experiment is not possible, that not all microarrays used will be analyzed at the same time, that the microarray chips used may not all be uniform, that reagents may vary over the course of the experiment, and so on.

If the microarray assay conditions vary, then it is still possible to construct a well-designed experiment by ensuring that the assay conditions are not confounded with the goals of the experiment. For instance, suppose that one has two classes of specimens one wishes to compare, that there are 10 specimens from each class, and that there are also 10 microarrays from each of the two different chip versions that are to be used. An experiment that assigns the samples from one class to one chip version and the samples from the other class to the other chip version would be poorly designed because one would not know whether to attribute observed differences

to chip versions or to the classes. Chip version and classes would be completely confounded, that is, their effects would be inextricably mixed together. A well-designed experiment, in this situation, would be one in which each class is assigned five chips from each version. Then observed differences between the classes could not be attributed to the chip version, and one could separate out the effects of the different chip versions from the biological differences between the classes.

4. Dual-label microarray experiment designs

Dual-label systems present design issues not encountered by single-label systems. Each array in a dual-label system has two channels, one corresponding to each dye label. The motivation for using two labels instead of one is that it allows one to eliminate spot-to-spot variation – due to quality, size, and location of spots – from comparisons of interest. The result has much greater power in class comparison experiments and greater ability to identify true clusters in class discovery than if a single dye was used. The resulting data structure, with a large variation between measurements from different spots and a small variation between the two measurements on the same spot, is called a block structure by statisticians, with the spots on the microarray serving as the blocking factor whose variability is blocked out of the comparisons of interest.

We will focus here on two types of designs for dual-label experiments, reference designs and balanced block designs, because for most experiments where the goal is class comparison, class prediction, and/or class discovery, one of these two designs will be optimal. Variations on these designs, such as dye swaps and technical replicates, will be discussed below. Other designs have been proposed, such as all-pairs designs (Yang and Speed, 2002), and loop designs (Kerr and Churchill, 2001).

4.1. The reference design

Reference design dual-label microarray experiments utilize multiple aliquots from a single RNA source, called the reference, which are applied to each microarray and are usually all labeled with the same dye. The role of the reference is sometimes not clearly understood. The size and location of a spot on a microarray is a major source of variability in these types of experiments. The role of the reference is to provide an estimate of this “spot” effect for every spot on every array. Consider the situation for a single gene. In all of the reference aliquots, this gene has the same expression level; therefore, differences in the expression level for this gene on different arrays can be attributed to the combination of spot size, quality, and location effects that together make up the spot-to-spot variation. Hence, for this gene, the spot-to-spot variation is well represented by the variation in the expression of the reference sample over the spots (assuming that the gene is expressed at a sufficiently high level in the reference sample). This allows one to correct for this source of variability when comparing the samples in the nonreference channels. Without the reference, it would not be possible to correct for this source of variation,

and the noise would effectively drown out much of the effects of interest for which one is looking.

The reference sample does not have to have a biologically meaningful interpretation, but it should be selected so that, in general, the genes expressed in the nonreference samples are also expressed in the reference sample. In order to get a good correction for the spot-to-spot variation, one needs some gene expression in the reference sample for that gene, otherwise there will be a universally low gene expression in the reference channel regardless of differences in the size, quality, and location of the spots, and the reference channel will not reflect the "spot" effects well. Therefore, genes with low or no expression in the reference channel will provide relatively noisy comparisons between the nonreference samples.

4.2. The balanced block design

The reference design may appear ideal because of the simple and intuitive way in which it allows one to estimate the spot-to-spot variation and eliminate it from comparisons of interest. But, in fact, spot-to-spot variation can also be estimated and eliminated from other types of designs. Although the estimates in these other types of designs may not be as intuitive as in the reference design, they are equally valid and can in fact result in considerable improvement in efficiency for class comparison experiments. The most efficient designs for class comparisons are balanced block designs. Other types of designs that have been proposed in the literature are not as efficient (Dobbin and Simon, 2002).

A balanced block design does not use a reference sample. Instead, a single aliquot from each biological sample (e.g., from each person or each mouse) is taken, and the samples are arranged so that samples from any two classes are paired together the same number of times over the microarrays. In the case of just two classes, this means that a sample from each class is paired together on each array. As discussed in the dye bias section below, half the samples from each class should be tagged with Cy3 dye, and the other half with Cy5 dye.

Examples of a reference design and a balanced block design are given in Tables 1 and 2.

4.3. When to use a reference design and when to use a balanced block design

The reference design is the most commonly used design in dual-label microarray experiments. There are several advantages to the reference design that make it particularly appealing to scientists who may not have access to a trained biostatistician or bioinformatician to consult with on the design or analysis of their experiment: (1) some microarray data analysis software packages assume that a reference design was used, and analyzing data from a different type of design with these packages may not be straightforward; (2) reference designs do not require that the investigator stipulate ahead of time what particular comparisons are going to be of primary interest, and hence allow for greater flexibility than designs (such as the balanced block design), which do require the classes to be identified ahead of time

Table 1 Reference design example. R is the reference sample. A, B, C, D, E are the five different classes being compared. There are two different samples from each class

Array	1	2	3	4	5	6	7	8	9	10
Cy3	R	R	R	R	R	R	R	R	R	R
Cy5	A	B	C	D	E	A	B	C	D	E

Table 2 Balanced block design example. A, B, C, D, E are the five different classes being compared. There are four different samples from each class

Array	1	2	3	4	5	6	7	8	9	10
Cy3	A	C	A	E	B	D	B	C	E	D
Cy5	B	A	D	A	C	B	E	D	C	E

and which may also lock the investigator into one particular type of comparison to the exclusion of other possible comparisons (for instance, one may be locked into a comparison of different tumor grades when the real interesting comparison turns out to be different tumor stages – but it may not even be possible to make this comparison after a balanced block experiment is run with grade as the comparison of interest); (3) if samples are analyzed at different times, then there may be no easy way to adjust for these differences unless one has the reference sample to serve as a baseline for the comparisons.

If class discovery is a goal of the experiment, and the samples are to be analyzed using cluster analysis, then the reference design has been shown to be superior to other designs that have been proposed (Dobbin and Simon, 2002). The reason for this is that effective cluster analysis depends on having good estimates of the distance between every pair of samples. In a reference design, the distance between any two samples is measured with the same efficiency because that distance only involves measurement error related to two arrays, corresponding to the arrays for each of the samples; this is because the repeated reference sample on each array can be used to "connect" any two arrays. Other designs have less direct "connections" between the samples that involve more arrays and hence more measurement error. Because some of the distances in these alternative designs will be measured quite inefficiently and so will be very poorly estimated, a cluster analysis algorithm will have a hard time picking up the structure in the data, as we have shown in simulations (Dobbin and Simon, 2002). Importantly, the balanced block design will have very poor class discovery performance because the spots, that is, the sources of greatest noise in these experiments, are confounded with the individual sample effects, completely obscuring the true distances between samples on different arrays.

Things are less straightforward if class comparison is the major goal of the experiment. In this case, the best design depends on what the limiting factor is in the experiment. Arrays are the limiting factor of the experiment if one has plenty of samples (or could produce plenty of samples), but owing to expense or other logistics, one can only run a fixed number of arrays. On the other hand, samples are the limiting factor if one only has access to a fixed number of samples, and

Table 3 Relative efficiencies of balanced block and reference designs. Efficiency of block design divided by efficiency of reference design, so that a relative efficiency over 1 indicates a block design that is more efficient. Ratio of biological variation to experimental error variation assumed to be 4 (see Dobbin and Simon, 2002)

Relative efficiencies		Number of varieties	
		2	3
Limiting factor	Same number of arrays used	2.4	1.8
	Same nonreference samples used	1.2	0.9

array expense or logistics is a relatively minor concern – one just wants to measure these samples as well as possible.

For a class comparison experiment in which arrays are the limiting factor, a balanced block design can be significantly more efficient than a reference design. This means that the resulting list of differentially expressed genes will potentially have far fewer "false positives" – genes that are not truly differentially expressed – and will be missing far fewer "false-negatives" – genes that truly are differentially expressed but which do not show up on the gene list because the difference is drowned out by experimental error variation. Table 3 shows the relative efficiencies for a balanced block compared to a reference design. For two classes, the relative efficiency is 2.4, indicating that it would require more than twice as many arrays with a reference design to achieve the same efficiency as a balanced block design.

For a class comparison experiment in which the samples are the limiting factor, the efficiency differences between a balanced block and reference design are much less dramatic (see Table 3, second row). This is partly because the reference design uses twice as many arrays, and hence it will be more expensive and labor intensive than the balanced block design. With three or more classes, the reference design is more efficient, and seems preferable overall. With two classes, the reference design is slightly less efficient, but this may not offset some of the advantages described in the first paragraph of this subsection.

5. Dye bias and the use of dye swap designs

Several definitions of dye bias have been used in the literature: (1) the tendency of one dye to appear brighter overall across genes (although this should be removed by proper normalization); (2) the tendency of spots with different overall intensities to display different relative efficiencies of the two dyes (which can be adjusted for using intensity dependent normalization such as lowess); (3) bias for a particular subset of genes caused by an interaction between the gene and the dye being incorporated. This creates dye bias for certain genes, but the dye bias for a gene is the same in all the samples; (4) bias caused by an interaction between genes and the dyes that is different for different samples. We will restrict attention to definition (3), and we refer to this type of dye bias as a gene-specific dye bias to distinguish it from (1) and (2). Bias of type (4), which could be called gene-and-sample-specific dye bias, could not be removed statistically from the comparisons

of the different samples, so we will not discuss this type of dye bias here. Several authors have investigated gene-specific dye bias and found that it does seem to exist, but generally tends to be small in quantity (Dobbin *et al*., 2003a,b Tseng *et al*., 2001).

For class comparison experiments, gene-specific dye bias will not affect comparisons between classes of samples labeled with the same dye. Hence, gene-specific dye bias will not affect the identification of differentially expressed genes in single-label experiments or in dual-label reference design experiments (for comparison of classes of nonreference samples). Gene-specific dye bias may affect cluster analyses in class discovery experiments in either single-label or dual-label platforms with a correlation metric because intensity differences between genes may be influenced by gene-specific dye incorporation efficiency.

For class comparison experiments using a balanced block design, gene-specific dye bias can be eliminated from the class comparisons in dual-label systems by labeling half (or nearly half, if odd number) of the samples from every class with each dye, and adding a dye bias term to the model. There is no need to dye swap individual arrays (i.e., run the same two samples with the labeling reversed) to eliminate the dye bias. In fact, dye swapping individual arrays will result in a loss of efficiency compared to running new arrays with different samples (Dobbin *et al*., 2003a).

Finally, if, in a dual-label system, comparisons between the reference sample and the nonreference samples are of interest, then dye bias adjustment can best be made by running dye swaps on some, but not all, arrays (Dobbin *et al*., 2003a).

6. Pooling samples

In some situations, there is not enough RNA available from individual specimens to run the microarray assay. This problem can be overcome either by RNA amplification, or by pooling different RNA samples together until there is enough RNA for an array. Pooling samples, in this way, can be a viable alternative to RNA amplification. In order to perform statistical inference in a class comparison setting, it is necessary to construct several independent pools from each class or condition. Two pools are independent if there is no overlap in the sources from which the pools are constructed, for example, if RNA from three mice is used to form each pool, then no two pools have a mouse in common.

Sometimes the motivation for pooling is not that there is inadequate RNA from individual samples to run the microarrays, but that by pooling it is hoped that the cost of the experiment will be reduced because fewer microarrays are required. For instance, an experiment with 12 samples from each of two classes would require 24 single-label microarrays; by pooling pairs of RNA samples together, one can reduce the number of arrays required to 12, 6 for each class. The pooled samples will also show less variation because, by pooling samples, the biological variation is reduced. But the improvement in power usually associated with such a reduction in variance is offset somewhat because power is also related to degrees of freedom for error, and the degrees of freedom are reduced from 22 in the 24-array experiment to 10 in the 12-array experiment. In fact, in order to get the same power as in the

Table 4 Number of arrays and samples required for various pooling levels. An independent pool is constructed for each array, so that no sample is represented on more than one array. Significance level used was 0.001, and power chosen was 95% to detect a twofold change in expression. Variance of log-ratios assumed to be 0.25 and the ratio of biological variation to experimental error variation assumed to be 4 (see Dobbin and Simon, 2005)

Number of samples pooled on each array	Number of arrays required	Number of samples required
1	25	25
2	17	34
3	14	42
4	13	52

24-array experiment, one will need to use more samples in the pooled experiment. This results in a trade-off between the cost of the microarrays and the cost of sample acquisition, which is displayed in Table 4. In general, unless the samples are very available and inexpensive relative to the microarrays, pooling does not appear to be an effective way to reduce cost (Dobbin and Simon, 2005; McShane *et al*., 2003; Shih *et al*., 2004).

7. Sample size

Here we will focus on sample size determination for class comparison experiments. Sample size for prognostic studies was treated in Simon *et al*. (2002), and sample size calculations for class discovery or class prediction remain open research questions.

In class comparison problems, it is common to cycle through, gene by gene, to determine which genes are differentially expressed (although in small studies, gene variance estimates may borrow information across genes (Wright and Simon, 2003)). Although multivariate permutation tests can be more effective (Simon *et al*., 2004), it is reasonable to power the study based on multiple univariate analyses. Assuming decisions about differential expression for individual genes are based on t-test statistics, a sample size formula for a two-class class comparison experiment is $Min\left\{n : n \geq 4\sigma^2 \frac{(t_{n-2,\alpha/2}+t_{n-2,\beta})^2}{\delta^2}\right\}$, where *Min* indicates the smallest positive integer n satisfying the equation, which is found iteratively. This formula can be used either for a reference design dual-label experiment or a single-label experiment. n is the total number of arrays required, with $n/2$ for each class. σ^2 is the variance of the base 2 log-ratios for the dual-label reference design, or the variance of the base 2 log-intensities for the single-label experiment. $t_{n-2,\alpha/2}$ $t_{n-2,\beta}$ and are the $\alpha/2$ th percentile and the βth percentile of the t-distribution with $n-2$ degrees of freedom, respectively. α is the significance level of the test, and $1-\beta$ is the power to detect a difference of size δ in the class means on the base 2 log scale. Selection of an appropriate variance σ^2 is somewhat problematic because different genes are usually assumed to have different error variances, but reasonable estimates can be derived using prior data from a similar experiment (Yang and Speed, 2002).

For $n \geq 60$, the sample size formula can be based on the simpler standard normal approximation $n = 4\sigma^2 \frac{(z_{\alpha/2}+z_\beta)^2}{\delta^2}$, where $z_{\alpha/2}$ and z_β are the $\alpha/2$ th and βth percentiles of the standard normal distribution, respectively.

Sample size formulas for balanced block designs, and for experiments with technical replicates, have been presented elsewhere (Dobbin and Simon, 2005).

References

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Dobbin K, Shih JH and Simon R (2003a) Statistical design of reverse dye microarrays. *Bioinformatics*, **19**, 803–810.

Dobbin K, Shih JH and Simon R (2003b) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute United States of America*, **95**, 1362–1369.

Dobbin K and Simon R (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438–1445.

Dobbin K and Simon R (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Cligiuri MA, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression profiling. *Science*, **286**, 531–537.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Raffeld M, *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, **344**, 539–548.

Kerr MK and Churchill GA (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

McShane LM, Shih JH and Michalowska AM (2003) Statistical issues in the design and analysis of gene expression microarray studies of animal models. *Journal of Mammary Gland Biology and Neoplasia*, **8**, 359–374.

Rosenwald A, Wright G, Chan WC, Connors JM, Camp E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB and Staudt LM (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, **346**, 1937–1947.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**, 227–235.

Shih JH, Michalowska AM, Dobbin K, Ye Y, Qiu TH and Green JE (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*, **20**, 3318–3325.

Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW and Zhao Y (2004) *Design and Analysis of DNA Microarray Investigations*, Springer-Verlag: New York.

Simon R, Radmacher MD and Dobbin K (2002) Design of studies using DNA microarrays. *Genetic Epidemiology*, **23**, 21–36.

Tseng GC, Oh M, Rohlin L, Liao J and Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549–2557.

Wright GW and Simon RM (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.

Yang YH and Speed T (2002) Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**, 579–588.

Specialist Review

Statistical methods for gene expression analysis

Shibing Deng
SAS Institute Inc., Cary, NC, USA
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Tzu-Ming Chu and **Russ D. Wolfinger**
SAS Institute Inc., Cary, NC, USA

Young K. Truong
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

1. Introduction

Microarray gene expression data provide considerable challenges to statistics owing to their high-dimensional gene space and small sample size, tremendous variation in the data, and correlations among genes (*see* Article 90, **Microarrays: an overview**, Volume 4). Recognizing the analytical challenges of microarray data, researchers are seeking and developing appropriate statistical tools for data analysis. Much research effort has been drawn to this area, and publications are growing exponentially in the past several years. Supported by these analysis methods, DNA microarrays are a revolutionary high-throughput technology for biological researches, addressing a wide variety of problems such as finding or predicting subclasses of cancer (Golub *et al*., 1999; Perou *et al*., 2000; Bittner *et al*., 2000).

The measurements of interest in a microarray experiment are usually log-transformed intensities or ratios (for two-color arrays). Before statistical analysis, some data preprocessing is necessary to ensure data quality; however, for this chapter, we assume the data have been processed and are ready for analysis. One of the basic statistical questions in microarray data analysis is the identification of differentially expressed genes. Statistically, this can be viewed as a hypothesis-testing problem, and traditional parametric methods such as t-tests can be useful for simple comparisons. When distributional properties of the data are in question, nonparametric permutation methods can also be adopted to determine significant changes (Tusher *et al*., 2001; Dudoit *et al*., 2002). For more complex experimental designs including covariates, ANOVA-type models are appropriate (Kerr *et al*.,

2000; Wolfinger *et al.*, 2001; Chu *et al.*, 2002). Classification, or supervised learning, of microarray data can potentially lead to clinical diagnosis and prognosis using DNA microarray technology. The goal is to predict clinical outcomes of new samples on the basis of their gene expression profiles using classification techniques built from existing data. The ideal methods need to be robust and sensitive in prediction. To make microarray data clinically applicable, it is desirable to select a small subset of genes that predict similarly well or even better than when all genes are used. This feature-selection problem is typically a critical part of the classification problem. Finding new patterns in microarray data is another interesting problem, as it can help researchers to identify the molecular signature of diseases. Clustering, or unsupervised learning, is the primary statistical tool for this purpose (Eisen *et al.*, 1998). Clustering can be performed on both samples and genes, and can help identify subtypes of tumors and groups of genes associated with clinical outcomes (Perou *et al.*, 2000).

2. Identifying differentially expressed genes

Several basic techniques for identifying differentially expressed genes are addressed here.

2.1. T-tests

Dudoit *et al.* (2002) use a two-sample t-statistic for each gene when comparing two conditions in a microarray experiment: $t = \frac{\overline{x_2} - \overline{x_1}}{s}$, where $\overline{x_1}$ and $\overline{x_2}$ are the mean expression measure for control and test group, s is their pooled standard deviation. $s^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$; s_1^2 and s_2^2 are sample variances for the control and test groups, respectively. Tusher *et al.* (2001) propose the SAM (Significance Analysis of Microarrays) method (*see* Article 59, **A comparison of existing tools for ontological analysis of gene expression data**, Volume 7) that is a shrunken t-statistic $t_s = \frac{\overline{x_2} - \overline{x_1}}{s + s_0}$, where s_0 is a small inflation factor, added to stabilize the t-statistic and control false-positives for low-expression measurements. The t-statistic provides us with a relative significance of differential expression. Ordering t-statistics by their magnitude provides one way to identify the most differentially expressed genes in a list. When overall statistical significance is important, computing p-values from the t-statistics is the standard procedure. The p-values can be computed as tail areas from appropriate t-distributions or by permutation methods.

2.2. Rank-based methods

Rank-based methods only consider the order statistics of the data and make minimal distributional assumptions. They are also less sensitive to location shift and scale changes, but usually have less power than parametric methods to detect differentially expressed genes, especially as many microarray experiments have limited sample size. Park *et al.* (2001) propose a nonparametric score statistic based on Kendall's tau and assess significance through permutation of the sample class labels.

2.3. AVOVA models

When more than two treatments conditions are present in an experiment, as well as for complex experimental designs, ANOVA models are reasonable choice and provide a rigorous method for extending two-sample t-tests. Kerr *et al*. (2000) fit the ANOVA model

$$Y_{ijkg} = \mu + A_i + D_j + T_k + G_g + (AG)_{ig} + (TG)_{kg} + \varepsilon_{ijkg} \qquad (1)$$

where Y_{ijkg} is the log intensity of array i, dye j, treatment k, and gene g; μ represents the overall mean; A_i, D_j, T_k, and G_g represent array, dye, treatment, and gene main effect respectively; AG and TG give gene specific array effect and treatment effect; and ε_{ijkg} is error and assuming i.i.d. with mean zero. In the ANOVA model, the main effects A, D, T are used to account for the overall systematic effects attributable to these factors, and thus serve as normalization factors. The interaction term TG models the genes that have differential expression across treatment levels. The significance of the differences can be evaluated parametrically using t- and F-tests or nonparametrically by bootstrapping the errors ε.

2.4. Linear mixed models

Wolfinger *et al*. (2001) extend the ANOVA model by modeling array and its interaction with other effects as random effects. Random effects are typically assumed to arise from a normal distribution and allow inferences to be made on the general population from which arrays are sampled. They can also be viewed as an empirical Bayes technique. Two stages of modeling make the computations more feasible. The first stage fits the main effects not involving genes and then subtracts their estimates from the data in order to normalize them. The second modeling stage fits a separate mixed model to data for each gene and the significance of treatment effects is used as assessment of differential expression for gene g. Under this framework, other covariates can also be included in the models. One advantage of mixed model is its ability to directly estimate the variability of different random factors, such as subject and array. Chu *et al*. (2002) extend this framework to Affymetrix array data.

2.5. Multiple comparisons

Microarray experiments typically involve simultaneously testing thousands of hypotheses, and so adjusting for multiple comparisons becomes a very important issue. The common approach is to control family-wise error rate (FWER), defined as the probability of having at least one false-positive error. There are many ways to control FWER and the simplest one is Bonferroni adjustment, which divides the desired FWER by the number of tests to form a conservative significance cutoff. Holm (1979) proposes a step-down procedure to adjust p-values. For m p-values, rank p_j such that $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$. The step-down Holm adjusted p-values are then $p^*_{(j)} = \max_{i=1,\ldots j} \{\min[(m - i + 1)p_{(i)}, 1]\}$.

Holm's procedure is less conservative than Bonferroni procedure, and the adjusted p-values still keep the order $p^*_{(1)} \leq p^*_{(2)} \leq \cdots \leq p^*_{(m)}$. Dudoit *et al*. (2002) use Westfall and Young's (1993) step-down procedure to estimate p-values by permutation. The advantage of this procedure is that it keeps the gene–gene correlations intact and it can often outperform simpler methods.

Instead of controlling FWER, another approach is to control false discovery rate (FDR). It defines the significance level on the basis of the proportion of false rejection of null among all rejection of null. In general, controlling FDR is less conservative than controlling FWER for microarray data (Storey and Tibshirani, 2003).

3. Classification of microarray data

In general, classification starts with fitting a model $\hat{y} = f(x)$ to minimize a predefined loss function, and then using this fitted model to predict new observations. In recent years, various classification methods have been applied to microarray data, particularly in the area of cancer classification. We now describe several commonly used methods.

3.1. Nearest centroid

Define the centroid for class k as $\overline{x}_k = \sum_i^{n_k} x_{k,i}/n_k$, which is the mean vector of class k, then classify sample x to the class whose centroid is nearest. $y(x) = \arg\min_k ||x - \overline{x}_k||$. van't Veer *et al*. (2002) apply a similar method to breast cancer outcomes with correlation distance.

3.2. Shrunken centroid (PAM)

Tibshirani *et al*. (2002) propose a modified version of nearest centroid. They define a t-statistic type of score for each gene as $d = \frac{\overline{x}_k - \overline{x}}{m_k(s+s_0)}$, where $\overline{x}_k$ is the class k centroid, $\overline{x}$ is the overall centroid, s is the pooled within-class standard deviation, and s_0 is a fudge factor set to control the possible large d score when intensity is low, $m_k = \sqrt{1/n_k + 1/n}$. The idea is to shrink the d score to zero if the class centroid is very close to the overall centroid. By doing so, the class centroid will not be affected by genes too close to the overall mean (more likely to be noise). Samples are classified to the nearest shrunken centroid:

$$y(x) = \arg\min_k ||\delta_k|| \tag{2}$$

with distance defined as weighted Euclidean distance, $\delta_k = \sum_{i=1}^{p} \frac{(x_i - \overline{x}'_{ik})^2}{(s_i + s_0)^2} - 2\log \pi_k$, where $\overline{x}'_{ik}$ is the shrunken centroid for ith gene at kth class and π_k is the prior probability of class k.

3.3. *K nearest neighbor (KNN)*

KNN classification method is based on the idea that data points close to each other in the feature space neighborhood should more likely come from the same class. Therefore, by counting class labels from the k nearest neighbors, we can predict the new sample's class. Letting x_i, $i = 1, \ldots k$ be the k nearest points to the new sample x, the predicting class of x is simply the mode of $y(x_i)$. Alternatively, one can use weighted voting with weights based on the distance for classification. Dudoit *et al*. (2001) compare KNN with several other classification methods in application to several cancer microarray datasets.

3.4. *Support vector machines (SVM)*

SVM, developed in machine-learning theory (Vapnik, 1998), has been widely used in microarray classification. In contrast to other common classification methods, the dimension of SVM calculations is a function of the number of observations, not the number of predictors. Therefore, it suffers much less from the curse of dimensionality than most other methods. For a given gene expression $p \times n$ matrix X, SVM finds a hyperplane $w{\bullet}X{+}b$ that separates two classes in the input space defined by X. If the two classes cannot be completely separated with linear boundaries in the input space, we can map the input space to a higher dimension feature space $(\phi_1(X), (\phi_2(X), \ldots \; (\phi_m(X))$, where the two classes can be separated by a hyperplane $w{\bullet}\phi(X){+}b$. The actual computation involves the dot product of $\phi^T(X){\bullet}\phi(X)$, which defines a kernel function. As illustrated in Figure 1, the smallest distance between the data points in each class and the hyperplane defines the margin, and the hyperplane is found by maximizing the margin. The final classification decision only depends on the data points on the boundaries, that is,

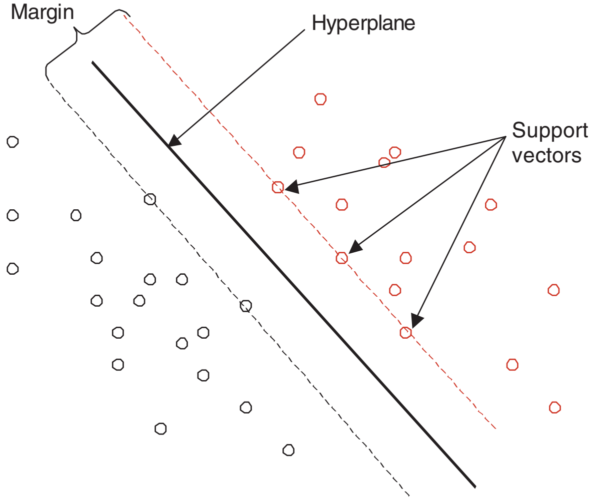


Figure 1 Illustration of SVM

the support vectors. Although SVM is developed for two-class classifications, it can be applied to multiclass problems through one-vs-other or one-vs-one algorithms (Yeang *et al*., 2001).

3.5. Neural networks (NN)

NN were developed as statistical tool for pattern recognition (Ripley, 1996). It is a two-stage classification model with a hidden layer between output Y and input data X. The hidden layer consists of M units that is a nonlinear transformation of the input X: $Z_m = s(\beta_{0m} + X\beta_m)$ with $s(x)$ serving as an activation function (usually a sigmoid function). Then Y is modeled by $g(\gamma_{0k} + Z\gamma_k)$ with $k = 1, \ldots K$ for the level of Y. The model has many weights parameters $(\beta_{0m}, \beta_m, \gamma_{0k}, \gamma_k)$, and they are learned from the training data by minimizing a loss function. Without proper regularization, a neural network can lead to overfitting problem owing to its large number of parameters. Khan *et al*. (2001) use an artificial NN to classify the small round blue cell tumor (SRBCT).

3.6. Partial least squares (PLS)

Traditional linear regression seeks a linear function of predictors X to explain as much the variation in the outcome Y as possible. The objective is achieved by minimizing the error function and obtaining least square estimate of parameters. In principal component analysis (PCA), the goal is to find a linear combination of X that accounts for the most variance in X. This is achieved through eigen-analysis. PLS balances these two objectives and finds a linear function of X that explains as much variation in both X and Y as it can. It achieves this goal by finding a linear function $t = Xw$ that maximizes its covariance with Y.

$$w = \arg\max\{Cov[t(X, w), Y]\} \tag{3}$$

The resulting PLS model can be used for predictions.

3.7. Logistic discrimination (LD)

Logistic regression models the probability of a class using a logit link function and assumes that outcomes are generated from a binomial distribution. A problem with traditional logistic regression is its limit in handling a large number of predictors. The model saturates when the number of predictors approaches the number of observations (samples). So it is difficult to directly use LD in the classification of microarray data where parameters outnumber the samples. However, LD can be used in conjunction with other dimension-reduction techniques, such as stepwise variable selection, SVD (Ghosh, 2002) and PLS (Nguyen and Rocke, 2002) for sample classification using microarray data.

3.8. *Other classification methods*

Other classification methods are also used in the literature for classification of microarray experiments. Dudoit *et al.* (2001) compare several discriminant methods for classifying several publicly available tumor datasets. Besides the KNN, they also compare Fisher Linear discriminant analysis (FLDA), quadratic linear discriminant (DLDA), classification tree (CART), bagging, and boosting. Their results showed that simple methods such as DLDA and KNN perform as good as more sophisticated methods, CART, bagging, and boosting. Golub *et al.* (1999) use a weighted voting scheme as a classifier to predict leukemia subtype AML and ALL. It is a weighted version of nearest centroid. They weight the correlation-based distance by a weight that is the Euclidean distance between the sample and the mean vector of class centroid. West *et al.* (2001) propose a Bayesian statistical regression model for the prediction of breast cancer outcome (estrogen receptor status). The model first performs an SVD to reduce the dimensionality of the data and then regresses $p(y)$ – the probability of ER positive – on the eigen genes through a probit link function, using Bayesian analysis for the model fitting. Romualdi *et al.* (2003) include recursive partition in the methods they compare.

3.9. *Criteria to evaluate classification*

The goodness of fit for a classification model is classification error rate. Without testing data, error rate is based on the prediction of training data itself. The so-called apparent error rate has significant optimistic bias. A better alternative is *cross-validation* (CV) *error* rate, where the training sample is randomly split into k roughly equal-sized groups, and for the ith group, we train the data on the other $k-1$ groups and predict the ith group. Repeat this for all k groups and estimate the error rate. In the literature, $k = n$ (leave-one-out) and $k = 10$ (10-fold) are the common choices for k. Another method for error rate estimation is bootstrap. We can bootstrap the training set and use it to predict the samples that are not in the bootstrap sample. One more advanced estimate is the ".632 estimator" (Efron, 1986).

Which method should we use? There is no universally best method in classification, and the choice of methods depends on the data, such as the number of classes, sample size, and the number of genes. SVM has shown popularity in the literature due to its overall good performance in microarray classification. Dudoit *et al.* (2001) and Romualdi *et al.* (2003) both compare several different classification methods in gene expression analysis with respect to classification performance. Dudoit *et al.* find that simple traditional methods such as LDA can perform better than the newer and more sophisticated methods. Romualdi *et al.* find through simulations that when the number of classes is less than four and sample size per group is >50, all methods perform comparably; SVM and KNN are more robust to noises in the data; LDA performs better than others when the number of classes is more than 5; and when sample size is small, NN has better performance than others. Again, these are based on simulation with a particular dimension-reduction technique and the real data are usually more complicated. It is always helpful to try several methods and compare their performance in terms of

CV error rate. Furthermore, the choice of classifiers also depends on the availability of software packages.

4. Clustering analysis

There is a large literature on clustering analysis of microarray gene expression data; Shannon *et al*. (2003) provide a good review. We here briefly discuss several commonly used methods. All methods employ some kind of distance or similarity measure; common ones include Euclidean distance and Pearson correlation. Nonparametric measures such as Spearman correlation and Kendall's tau can also be used. SAS/Proc Distance (SAS Institute Inc, 2004) describes many other alternatives. Once a distance measure has been selected, commonly used clustering methods include hierarchical clustering, K-means clustering, and self-organized maps. Model-based clustering assumes the data are from a mixture of some probability distributions.

4.1. *Hierarchical clustering (HC)*

HC is a bottom-up clustering method that starts from a distance matrix. Depending on the analysis objective, the distance can be defined between genes or samples. Let us assume we intend to cluster on genes. The HC algorithm proceeds as follows: (1) Initially, each observation forms a cluster, (2) identify the pair of clusters with the smallest distance and combine them into a new cluster, (3) update the new distance matrix, (4) repeat this process until one cluster is left. The distance matrix can be calculated by using the average distance (average linkage algorithm), the smallest distance (single linkage algorithm), or the largest distance (complete linkage algorithm) between two clusters. In practice, average linkage works quite well with standardized microarray data (Shannon *et al*., 2003). The hierarchical tree (dendrogram) resulting from HC analysis provides a very useful tool for data visualization. In combination with heat map of gene expression measures, it gives a graphical presentation of gene expression data, and is widely used by researchers. Despite its success, HC has its limitations. It is not obvious where the cutoff should be on the dendrogram to determine the number of clusters. HC can also suffer from lack of robustness, nonuniqueness, and a tendency to break large clusters. The deterministic nature of HC algorithms can result in groupings based on local decision with no opportunity to reevaluate the clustering that already is done (Tamayo *et al*., 1999).

4.2. *K-means clustering*

K-means clustering is a top-down method starting from the final required number (K) of clusters. The algorithm begins by randomly selecting K genes and defines them as the initial centroids. Then, the distances between each gene and these centroids are computed and each gene is grouped to its nearest centroid. The

centroid's position is recalculated and each gene again is regrouped to the updated nearest centroid. The iterative process continues until a prespecified convergence criterion is met. K-means clustering generates the required number of clusters and is useful when some prior knowledge of the number of clusters exists. Tavazoie *et al.* (1999) show that it is effective in analyzing microarray gene expression data on *Saccharomyces cerevisiae* with 30 clusters. Unfortunately, K is unknown in most cases. The results from K-means clustering also depend on the choice of initial centroids that are usually picked randomly. In practice, it is helpful to repeat the K-means clustering for a number of initial centroid selections and choose the best solution based on the tightness of clusters. K-means clustering is found to produce tighter clusters than hierarchical clustering (Tibshirani *et al.*, 1999).

4.3. *Self-organizing maps (SOM)*

SOM map data points from a high dimensional space to a geometrical configuration in a low dimensional space (Kohonen, 1989). The SOM algorithm starts with a simple topology, such as a grid of nodes in a two-dimensional plane and a distance function $d(N_i, N_j)$ between nodes. Each node is mapped into the original data space with a reference vector. Tamayo *et al.* (1999) and Golub *et al.* (1999) successfully apply SOMs to microarray data.

There are similarities between SOM, K-means clustering, and multidimensional scaling (Tibshirani *et al.*, 1999). Like K-means clustering, SOMs begin with the required number of clusters (nodes). Unlike K-means clustering, the positions of all nodes (not just the nearest node) are updated by each data point.

4.4. *Model-based clustering*

Unlike the above heuristic algorithms, model-based clustering algorithms start with the assumption that the data is a mixture of a finite number of distributions with a density function

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i | \mu_k, \Sigma_k) \tag{4}$$

where $x_i (i = 1, \ldots, n)$ is the data to be clustered on and π_k is a probability that an observation comes from component $k (k = 1, \ldots, K)$. Each component in the mixture determines a cluster, and the objective is to search for the parameters that define the components. The standard approach assumes multivariate normal distributions for f_k and utilizes EM algorithm to estimate parameters μ_k, Σ_k, and π_k. The covariance parameters (Σ_k) determine the shape of the clusters. One of the advantages of model-based clustering is that the determination of the number of clusters K becomes a model-selection problem that can be easily estimated by the well-known Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Another advantage of the model-based clustering is its ability to assign posterior probabilities of observations belonging to each cluster. In practice,

data transformations are often necessary to meet distributional assumptions, and requisite computations can be very intensive. Yeung *et al*. (2001b) apply model-based clustering algorithms to gene expression data.

4.5. Two-way clustering

Two-way methods cluster genes and samples simultaneously, and can identify a cluster of genes grouped with a subset of samples. Tibshirani *et al*. (1999) use a block clustering algorithm to split the data into two pieces, either in rows or in columns, whatever gives the largest reduction of total within-block variation. A similar approach based on mean square residue was introduced by Cheng and Church (2000).

4.6. Evaluation of clustering methods

Evaluating the goodness of clustering results has been an active research topic for decades. For microarray data, the best evaluation may come from the understanding of the biological processes associated with the data; however, in most cases, this is not at all possible due to the lack of full biological knowledge. Statistical evaluation can be helpful in absence of that knowledge. The first aspect of the evaluation is to determine the right number of clusters. For model-based methods, this can be solved from model selection using AIC or BIC, as discussed before. For heuristic methods, Tibshirani *et al*. (2001a) propose a gap statistic to compare a within-cluster variability measure to its expectation and choose the number of clusters on the basis of the relative size of the gap statistic and its standard deviation. Ben-Hur *et al*. (2002) determine the cluster number by a stability-based resampling approach. Good clusters are unlikely to change dramatically on the leave-one-out subsamples.

Good clusters should have relatively small variance within clusters and relatively large separations between clusters. Some scores based on these measures are developed to evaluate the quality of the clusters. The silhouette index of a cluster measures the compactness of the cluster and how far it is from the next cluster. For gene i, it is defined as

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{5}$$

where $a(i)$ is the average distance of gene i to other genes in the same cluster, $b(i)$ is the average distance of gene i to genes in its nearest neighbor cluster. Another similar index is Dunn's index discussed by Azuaje (2002). Famili *et al*. (2004) propose a stability score to evaluate the quality of the clusters. Chen *et al*. (2002) compare several scores including silhouette in evaluating clustering performance.

Most of the methods of evaluating the quality of clustering results are based on resampling or perturbation methods. Tibshirani *et al*. (2001b) propose a prediction strength method that splits the data into a training set and a test set, and each set is

clustered separately and then assessed on how well the training set cluster centers predict the comemberships in the test set. Yeung *et al*. (2001a) introduce a figure of merit (FOM) method by predicting one sample using clusters generated from all other samples, on the basis of the idea that a good clustering scheme should have a good predictive power. Kerr and Churchill (2001) assess the stability of clustering analysis by bootstrapping the residuals from ANOVA results. Bittner *et al*. (2000) evaluate the robustness of clustering methods by adding noise to the data and compare the sensitivity of the clustering results to the level of noise. Each clustering method has its strengths and weaknesses, and the choice of a clustering method depends on the data and the analysis objective. It is helpful to have some alternative approaches to validate clustering results.

References

Azuaje F (2002) A cluster validation framework for genome expression data. *Bioinformatics*, **18**, 319–320.

Ben-Hur A, Elisseeff A and Guyon I (2002) A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, **7**, 6–17.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Chen G, Jaradat SA, Banerjee N, Tanaka TS, Ko MSH and Zhang MQ (2002) Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica*, **12**, 241–262.

Cheng Y and Church GM (2000) Biclustering of expression data. *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology*, San Diego, pp. 93–103.

Chu T, Weir B and Wolfinger RD (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, **176**, 35–51.

Dudoit S, Fridlyand J and Speed TP (2001) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77.

Dudoit S, Yang YH, Callow MJ and Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron B (1986) How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, **81**, 461–470.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Famili AF, Liu G and Liu Z (2004) Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics*, **20**, 1535–1545.

Ghosh D (2002) Singular value decomposition regression models for classification of tumor from microarray experiments. *Pacific Symposium on Biocomputing*, **7**, 18–29.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Kerr MK and Churchill GA (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8961–8965.

Kerr MK, Martin M and Churchill GA (2000) Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819–837.

Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C, *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression pro_ling and arti_cial neural networks. *Nature Medicine*, **7**, 673–679.

Kohonen T (1989) *Self-Organization and Associative Memory*, *Third Edition*, Spring-Verlag: Berlin.

Nguyen DV and Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39.

Park PJ, Pagano M and Bonetti MA (2001) Nonparametric scoring algorithm for identifying informative genes from microarray data. *Pacific Symposium on Biocomputing*, **6**, 52–63.

Perou CM, Surlie T, Eisen MB, van de Rijn M, Jeffreyk SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al*. (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Ripley BD (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press: Cambridge.

Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G and Lanfranchi G (2003) Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Human Molecular Genetics*, **12**, 823–836.

SAS Institute Inc (2004) *SAS/Stat Manual*, SAS Institute Inc: Cary.

Shannon W, Culverhouse R and Duncan J (2003) Analyzing microarray data using cluster analysis. *Pharmacogenomics*, **4**, 41–51.

Storey JD and Tibshirani R (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907–2912.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281–285.

Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D and Brown P (1999) *Clustering Methods for the Analysis of DNA Microarray Data*, Technical Report, Standford University, October 1999, **http://www-stat.stanford.edu/~tibs/research.html**.

Tibshirani R, Hastie T, Narasimhan B and Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567–6572.

Tibshirani R, Walther G, Botstein D and Brown P (2001b) *Cluster Validation by Prediction Strength*, Technical Report No.2001-21, Stanford University.

Tibshirani R, Walther G and Hastie T (2001a) Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society Series B*, **63**, 411–423.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Vapnik V (1998) *Statistical Learning Theory*, Wiley: New York.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan R, Olson JA Jr, Marks JR and Nevins JR (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 11462–11467.

Westfall PH and Young SS (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C and Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed model. *Journal of Computational Biology*, **8**, 625–637.

Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, Reich M, Lander E, Mesirov J and Golub T (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17**(Suppl 1), S316–S322.

Yeung KY, Fraley C, Murua A, Raftery AE and Ruzzo WL (2001b) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yeung KY, Haynor DR and Ruzzo WL (2001a) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Specialist Review

Algorithms for gene expression analysis

Alvis Brazma
European Bioinformatics Institute, Cambridge, UK

Aedín C. Culhane
The Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland

1. Gene expression data

Gene expression data characterize the absolute or relative abundance of mRNA molecules in a biological sample. Ideally, we would like to measure gene expression in some objective units, such as the number of mRNA molecules of each species that are present. In reality, this is rarely possible. Usually, the measurements are more qualitative, and even if quantitative measurements are given, the noise levels are typically rather high and often unknown. This has to be taken into account when analyzing expression data.

There are many technologies used to measure gene expression such as microarrays, serial analysis of gene expression (SAGE), RT-PCR, and northern blots. Tissue-specific gene expression may be assessed by quantification of expressed sequence tag (EST) libraries. Spatial patterns of gene expression are obtained by in situ hybridization. Of all of these technologies, microarrays are both the most high throughput and most widely used. Therefore, in this chapter, we will concentrate exclusively on the analysis of microarray gene expression data (although properly formatted data from other sources could be analyzed in a similar fashion).

To understand the rationale behind various microarray gene expression data algorithms, we need to understand some of the basic principles on which microarray technology is based. Microarrays exploit preferential binding of mRNA to their complementary nucleotide sequence. A microarray chip consists of thousands of DNA molecules attached in fixed locations onto a solid surface. The abundances of mRNA molecules in a biological sample are assessed by fluorescently labeling its mRNA extract, applying this to the microarray chip, and measuring the fluorescence intensities from each location on the array. This processed array is often called a *hybridized array*, because on this array each mRNA molecule is bound or

hybridized to its complementary DNA strand. Microarrays are often used for the comparison of two samples representing, for example, a treatment and a control, in which case the treatment and the control molecules are labeled with different dyes. Gene expression levels measured in this way are relative and expressed as ratios between the sample and control (or more commonly, their logarithms base 2).

Each individual microarray produces digital images obtained by scanning the hybridized arrays, each of which has to be transformed into a numeric representation that we call the spot quantitation matrix. In this matrix, each row represents a spot (feature) on the array, while each column represents a particular quantitation type, such as the mean of the pixel intensities in a spot. Unfortunately, the relationship between the fluorescence intensity and the abundance of a given RNA molecule is not straightforward. Moreover, some microarray technological platforms, for example Affymetrix, have specific additional data preprocessing requirements (*see* Article 57, **Low-level analysis of oligonucleotide expression arrays**, Volume 7). The data for a particular experiment consists of the collection of spot quantitation matrices for all hybridizations.

A collection of spot quantitation matrices from related hybridization experiments is the starting point for the next data transformation step. The set of spot quantitation matrices are summarized into a single gene expression data matrix, where each row represents a gene (as opposed to a feature on the array), each column represents a particular experimental condition, and each position contains one or more quantities characterizing the absolute or relative abundance of each particular gene product.

To transform fluorescence intensity measurements into relative nucleotide abundances, we have to remove systematic noise originating from a variety of sources. The transformations that correct for the systematic noise and derive the abundance measurements are known as *data normalization* (Quackenbush, 2002; Huber *et al*., 2003; Irizarry *et al*., 2003). Data normalization methods can be based either on external controls (specific control features spotted on the array in combination with control molecules added to the sample) or on various assumptions about the measurement data itself (which may or may not be reasonable for a particular experiment). One of the most common assumptions is that the total amount of nucleic acid hybridized in all the samples is equal, and this serves as the basis for widely used normalization methods such as quantile normalization, locally weighted linear regression (Lowess), and variance stabilizing normalization. External control-based normalization is necessary in some cases, particularly if we do not know how the total amount of nucleic acid changes in the sample (van de Peppel *et al*., 2003). Various software packages and tools are available for microarray data normalization, of which Bioconductor (Gentleman *et al*., 2004) is among the most popular one (*see* Article 55, **Differential expression with the Bioconductor Project**, Volume 7).

2. Extracting meaning from gene expression data

One of the first problems one faces when analyzing gene expression data matrix is the presence of missing values, that is, some of the values in the gene expression

matrix are undefined. Various methods can be used for imputing a value from the information in the existing data. These include the row average method, which uses the average expression of the gene of interest in the other samples. Other approaches that may produce more reliable estimates of missing values include weighted k-nearest neighbors (Troyanskaya *et al*., 2001), least squares, or expectation maximization (Bo *et al*., 2004).

Gene expression matrices are usually too large to enable a manual gene-by-gene study, for instance, a human microarray gene expression matrix may contain over 25 000 rows – a row per gene. Therefore, automated data analysis methods can be used to reduce the dimensionality. These methods roughly can be classified into two categories – supervised and unsupervised analysis methods (Brazma and Vilo, 2000). The two main classes of unsupervised data analyses applied to microarrays are various *clustering algorithms* and ordination-based methods such as *principal component analysis* (PCA).

3. Clustering

The rows in the gene expression data matrix are often called gene expression profiles, while the columns are called sample expression profiles. The goal of expression data clustering is to group together genes or samples that have similar expression profiles. Clustering is a well-established field, and various clustering algorithms have been invented. These include *hierarchical agglomerative clustering*, which is based on iteratively grouping together the objects that are most similar to each other, and *k-means clustering* (Hartigan, 1975), in which the number of clusters is defined a priori, and the clustering is iteratively improved by adjusting the cluster centers in Euclidean space. There are also newer clustering algorithms, such as Kohonen's self-organizing maps (Toronen *et al*., 1999; Wang *et al*., 2003), and graph-theory-based algorithms (Sharan *et al*., 2003). A simple clustering algorithm based on binning, that is, discretizing the expression profile space and clustering together the profiles that map to the same bin, has been shown to be useful for grouping data in situations in which the number of experimental conditions is relatively small (Brazma *et al*., 1998).

Before choosing a clustering algorithm, one has to define how to measure the distance between the expression profiles (Causton *et al*., 2003). Various distance measures are used, of which those based on Pearson's correlation are the most popular. Pearson's correlation quantifies the linear correlation between two expression profiles. It ranges from $+1$ to -1, where a value of $+1$ means that there is a perfect positive linear relationship between the two expression profiles (all points are on the diagonal of the scatterplot). Among popular alternatives are Euclidean distance, which takes into account the absolute expression levels, or Spearman's rank correlation, which is based on ranking the expression levels, and then measuring the correlation between the rank vectors. Some clustering algorithms are only applicable for particular distance measures, for example, k-means assumes that the distance measure has Euclidean properties.

An important consideration when selecting a clustering algorithm is its appropriateness and performance. Clustering may impose a hierarchical or flat structure on

the data, which may be built using divisive (top-down) or agglomerative (bottom-up) approaches. Divisive methods start out with the complete set of data as one large group or cluster, and proceed to add partitions that separate the objects starting with those that are most dissimilar. In the case of gene expression analysis, an object is either a gene or sample expression profile of a gene expression data matrix. *Hierarchical agglomerative clustering* is a process in which each object starts out as a separate "cluster" and these objects are successively fused, until typically all objects are included. Hierarchical agglomerative clustering starts by considering each object as a separate or singleton cluster; starting with n objects or singleton clusters, where n is equal to the total number of objects. The first iteration of clustering will group together the two objects that are most similar to form a single cluster, and leave $(n-1)$ clusters. The distances between the newly formed cluster containing two objects and all of the remaining clusters or objects are then updated and then the next most similar objects are grouped together as a single cluster. This process is carried out iteratively until there is a single large cluster, and produces a hierarchical tree (Dendrogram), as shown in Figure 1.

Hierarchical clustering was one of the first clustering algorithms applied to gene expression data (Eisen *et al*., 1998), and remains a very popular microarray analysis method. The results of hierarchical clustering are frequently represented in a hierarchical tree, also known as *dendrogram*. The branch lengths of the tree may represent the degree of similarity between the data. Note that the threshold values, that is, where to split the tree to obtain distinct clusters, have to be established independently. Usually, it is done quite arbitrarily by picking the clusters that seem either to be the "tightest", or that include genes with similar functional annotation.

Nonhierarchical, or flat, clustering groups the data into clusters, which may either be a partitioning of the data in nonoverlapping clusters or a set of overlapping clusters. The number of clusters can either be given in advance or estimated from the data by applying various criteria. The most common nonhierarchical methods are k-means clustering, self-organizing maps, and various so-called Bayesian approaches. Self-organizing maps impose a partial structure on the data, such that adjacent clusters are related. Bayesian clustering permits the incorporation of priors, that is, additional information about our knowledge of the data.

K-means is the most common method of flat partition–based clustering. It starts with the given number of cluster centers, chosen either randomly or by applying some heuristics. Next, the distance from the centroids to every object is calculated, and each object is assigned to the cluster defined by the closest centroid. Then, for each cluster a new centroid is found. The distance from each object to each of the new centroids is calculated, and in this way, the boundaries of the partitioning are revised. This is repeated until either the centroids stabilize (which is not guaranteed), or until an a priori defined maximum number of iterations have been reached. It is common that as many as 20 000 to 100 000 cycles are necessary before the position of the nodes stabilizes.

There is no compelling evidence that more sophisticated clustering algorithms perform better than the simplest ones with respect to the biological insights that have been obtained. At the same time, this does not mean that all clustering algorithms are appropriate for all datasets. It is possible that in the future, when we

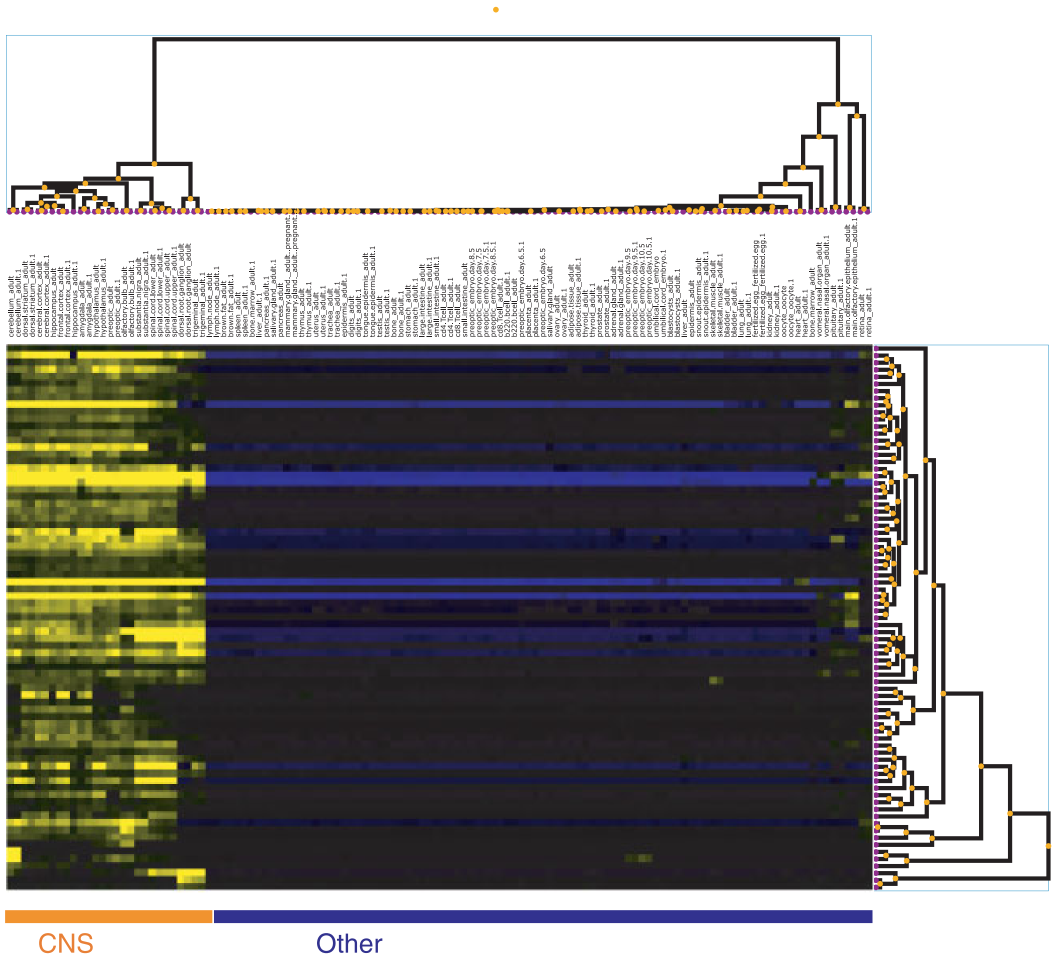


Figure 1 Hierarchical clustering of selected 77 mouse genes and 122 samples representing 61 tissue types each in two replicates (Su *et al*., 2002). The data were retrieved from ArrayExpress database (**http://www.ebi.ac.uk/arrayexpress**), accession number E-AFMX-4, and clustered and visualized using hierarchical clustering in expression profiler (**http://www.ebi.ac.uk/expressionprofiler**). Each row represents a gene and each column a sample expression profile. Genes that are overexpressed when compared to the mean expression level are in yellow, and those that are underexpressed are in blue. The genes shown were preselected using the between-group analysis component of expression profiler. These genes were the top 77 genes that discriminated most between central nervous system (CNS) and other tissue types

have a better understanding of the nature of gene expression data, more specialized algorithms will be developed that significantly outperform those now in use.

4. Ordination-based methods

Ordination is a complementary method to clustering and should not be confused with it. This is because ordination considers the variability of the whole data matrix, bringing out general gradients or trends in the data, whereas clustering investigates

pairwise distances among objects looking for fine relationships (Legendre and Legendre, 1998). Since ordination methods do not partition data into discrete clusters, they permit the detection of both gradients and discrete groupings in data. PCA and correspondence analysis (COA) are two of the many ordination methods that have been applied to microarray data analysis (Raychaudhuri *et al*., 2000; Crescenzi and Giuliani, 2001; Fellenberg *et al*., 2001).

The idea behind *PCA* and ordination is quite intuitive – if there are correlations between measured quantities, such as gene expression vectors, we can combine correlated objects and reduce the "dimensionality" of the problem. Mathematically, this means that if a linear relationship exists between gene (or sample) expression profiles in a data matrix, these can be expressed as a linear combination of the original such that colinear variables are regressed onto a new set of coordinate axes. These new axes are often called *eigenvectors*, *principal components*, *principal factors*, or *principal axes*, as they explain the principal trends in the data. These are generally ordered on the basis of their ability to explain the primary trends in the data, and the first few of these new axes (usually two or three) are used to present a dimensionally reduced representation of the data. The maximum number of new axes is equal to the number of rows or columns (whichever is smaller) of the original data matrix minus 1.

One expects coregulated (colinear) gene expression changes in response to a particular biological perturbation. For example, a microarray expression pattern "diseased" might equal x.gene1 + y.gene2 + z.gene3, where x, y, z are the loadings or contribution of each gene. Typically, there are only a few important gene trends and these are described by the first few principal components. If the study was a simple control versus treated experiment, one or two principal components might explain the major gene trends (expressed in particular tissue types). However, if the experiment was a time course, or complicated diagnosis of multiple cancer types, more principal components would be required to explain the main trends in the data. Although selection of the number of principal components is quite subjective, it can be simplified by examining the *eigenvalue* associated with each principal component. An eigenvalue scores the amount of information or variance explained by a principal component. In PCA and COA, eigenvalues are ranked from highest to lowest. Thus, the most important trends in the data (highest variance) will be described by the first few principal components.

The above described methods, clustering and ordination, do not assume that the order in which samples have been presented is important. This is not true for time-series data. Although these methods can still be applied, they will not help one to analyze order-specific properties of time-series data, such as periodicity. There are specialized methods, such as autocorrelation or Fourier analysis to perform this analysis.

After reducing thousands of profiles into a few clusters of genes exhibiting "interesting" expression behavior, a human investigator can study these genes in the context of prior knowledge to draw conclusions about the underlying biological processes and propose new experimentally testable hypotheses. Automated software tools can help in finding clusters of genes sharing the same gene annotation (Cheng *et al*., 2004; Harris *et al*., 2004), or in analyzing gene regulatory regions to identify

shared sequence patterns. Such approaches have been successful, for instance, in identifying transcription factor binding sites in yeast.

5. Supervised methods

Supervised data analysis methods exploit our prior knowledge about the genes or experimental conditions right from the beginning. For instance, if we are given a set of gene expression measurements for diseased and healthy samples, we can look for expression values in the matrix that allow the prediction of "diseased" or "normal" state of a new sample given its expression profile. This can then be used in diagnostics, but additionally, if the predictors can be described by simple rules, these can also help us in understanding the biological mechanisms of the disease. For instance, if a small subset of genes are highly expressed in the disease state and lowly expressed in normal or vice versa, these can be hypothetically implicated in the mechanisms of the particular disease. This approach has been successfully used, for instance, in cancer classification and class prediction (Hampton and Frierson, 2003).

There are a number of different techniques used in supervised gene expression data analysis, including *linear regression* or *linear discriminant analysis, nearest neighborhood analysis, support vector machines, neural networks,* and *decision trees* (Simon, 2003; Kuo *et al*., 2004). In principle, any supervised analysis or machine-learning techniques can be applied. However, it has to be noted that in gene expression analysis, the number of objects that we want to classify (samples) typically is much smaller than the number of parameters that can be used in classification (genes). This is known as the "curse of dimensionality", and is driving the development of new data analysis methods.

This problem has been most commonly circumvented by selecting subsets of genes in advance or iteratively during training. Typically, these subsets use about 50–100 genes, although it has been suggested that just a few genes might be statistically preferable (Li and Yang, 2001). Such gene selection may be cumbersome to produce, possibly involving arbitrary selection criterion, or may miss highly informative combinations of genes. To this end, Culhane *et al*. (2002) described a powerful yet simple supervised method called *between-group analysis* (BGA), a multiple discriminant approach that could be safely used when the number of genes exceeds the number of samples (Dolédec and Chessel, 1987). The basis of BGA is to ordinate the group means rather than the individual samples. For N groups (or classes of samples), we find $N-1$ eigenvectors or axes that arrange the groups so as to maximize the between-group variances (Figure 2). The individual samples are then plotted along them. Each eigenvector can be used as a discriminator to separate one of the groups from the rest. New samples are then placed on the same axes and can be classified on an axis-by-axis basis or by proximity to the group centroids. BGA was shown to be a fast and simple approach to use, yet it produced accurate discrimination as judged by the performance on gene expression test data or by a jack-knife leave-one-out cross-validation analysis (Culhane *et al*., 2002).

Frequently, in microarray studies, researchers have insufficient resources to generate adequate numbers of samples to provide a training and test dataset

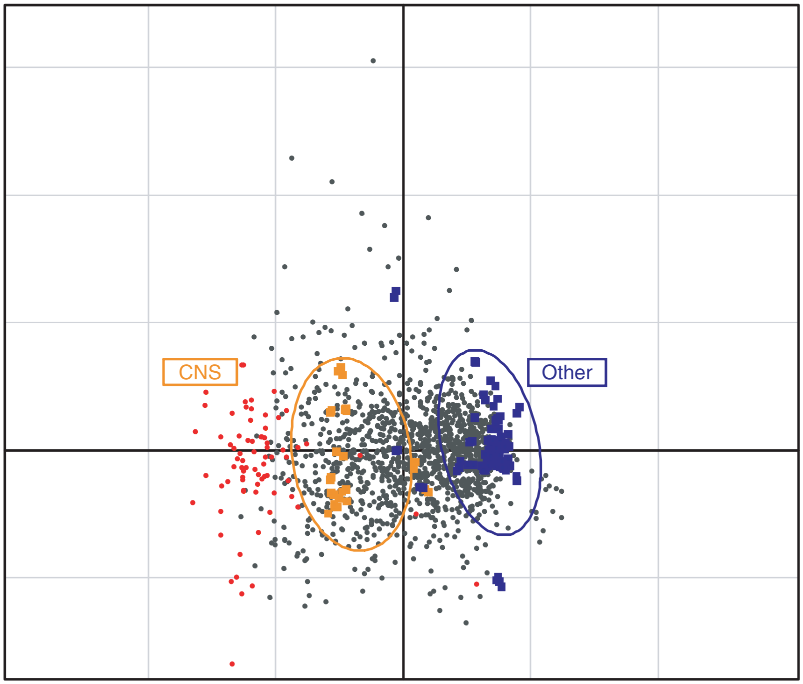


Figure 2 Visualization of two eigenvectors of a correspondence analysis of gene expression profiles of mouse tissues (Su *et al*., 2002). Correspondence analysis was performed on 1389 mouse genes and 122 samples representing 61 tissues types (as in Figure 1). These 1389 genes were the top 1389 genes that discriminated most between central nervous system (CNS) and other tissue types, as defined using the between-group analysis component of expression profiler. Genes (closed circles) and array samples (closed squares) are plotted on the one graph. A gene and array with a high association are project in the same direction from the origin. The top 77 genes (Figure 1) that discriminate between these groups are highlighted (red closed circles). The figure was obtained using the made4 package in Bioconductor. Correspondence analysis is also available in Expression Profiler

that are necessary in supervised methods such as those described above. In this case, researchers may simply rank the genes that are most differentially expressed between classes, and subsequently will use experimental approaches to validate these genes. Unfortunately, simple approaches to ranking genes are often complicated by the level of noise in the data, the low number of experimental replicates, and the high number of genes. Simple methods that rank genes using fold difference or the classical t-statistic are very susceptible to high numbers of false-positives. This has led to the development of new gene-ranking methods, of which *significance analysis of microarrays* (SAM), (Tusher *et al*., 2001) and *limma* (Smyth, 2004) are among the most popular (*see* Article 59, **A comparison of existing tools for ontological analysis of gene expression data**, Volume 7 and Article 62, **Large error models for microarray intensities**, Volume 7). SAM ranks genes using a moderated t-statistic, and the significance of these values are adjusted for multiple testing by permutation. Limma fits a linear model to the data,

but moderates the gene standard error using an empirical Bayes model to obtain a moderated t-statistic.

Automated methods can go beyond simple clustering or classification approaches. For instance, one can try to reverse-engineer the gene regulation network from gene expression data. Though it is unlikely that reverse engineering of gene regulation networks is possible purely from microarray data, nevertheless, these attempts are providing valuable insights into different properties of gene regulation networks (Schlitt and Brazma, 2002).

There exist a large number of commercial and free software tools implementing various analysis methods. Some of the most popular free software tools are Cluster and TreeView (Eisen *et al*., 1998), TIGR Multi-Experiment Viewer (Saeed *et al*., 2003), J-Express (Dysvik and Jonassen, 2001), and Bioconductor (Dudoit *et al*., 2003). Expression Profiler (Kapushesky *et al*., 2004), which is provided by the European Bioinformatics Institute, is available from **www.ebi.ac.uk/expressionprofiler** and is an on-line-based microarray data analysis tool that implements many of the methods described above. To use this tool, the user only needs a web browser.

6. Meta-analysis

Microarray experiments are relatively expensive, but they generate data that can continue producing new knowledge long after the experiments have been completed. Large, systematic, high-quality datasets are proving to be a valuable resource for the research community, similar in value to genome sequence data. Meta-analysis of large expression datasets, particularly when combined with other relevant datasets, and analyzed using new methods, may bring new insights that go beyond the scope of the original studies. Simple methods, such as *coinertia analysis*, can be used to compare the global correlation between gene expression profiles of the same tissues or cell lines obtained in different studies, even if these studies have used arrays with different catalogs and numbers of genes (Culhane *et al*., 2003). Much information can be obtained by comparing gene profiles across studies or even between species. For instance, the comparison of the expression profiles of homologous genes across a range of organisms can help in predicting the orthologous genes (Grigoryev *et al*., 2004), or comparing similar processes, such as the cell cycle or other core biological processes in different organisms (Stuart *et al*., 2003). To make such gene expression data meta-analysis possible, it is important that all the experimental details and sample properties are recorded. *Minimum Information About a Microarray Experiment* or MIAME is a microarray data annotation standard developed by the Microarray Gene Expression Data Society for this purpose (Brazma *et al*., 2001).

Several public repositories have been established to collect the publicly available microarray data – *Gene Expression Omnibus* at the NCBI, *CIBEX* in DDBJ, and *ArrayExpress* at the European Bioinformatics Institute (Brazma *et al*., 2003). A researcher can query these databases for all experiments of a given type for a given organism and retrieve the respective data, which then can be combined with the researcher's own data, or used for designing new experiments. With growing

amounts of microarray data, improving quality, and new analysis methods being developed, such databases are becoming an essential resource for functional and comparative genomics research.

References

Bo TH, Dysvik B and Jonassen I (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, **32**, e34.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

Brazma A, Jonassen I, Eidhammer I and Gilbert D (1998) Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, **5**, 279–305.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, *et al*. (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**, 68–71.

Brazma A and Vilo J (2000) Gene expression data analysis. *FEBS Letters*, **480**, 17–24.

Causton HC, Quackenbush J and Brazma A (2003) *Microarray Gene Expression Data Analysis a Beginner's Guide*, Blackwell Publishing: Oxford, p. 192.

Cheng J, Sun S, Tracy A, Hubbell E, Morris J, Valmeekam V, Kimbrough A, Cline MS, Liu G, Shigeta R, *et al*. (2004) NetAffx gene ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.

Crescenzi M and Giuliani A (2001) The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS Letters*, **507**, 114–118.

Culhane AC, Perriere G, Considine EC, Cotter TG and Higgins DG (2002) Between-group analysis of microarray data. *Bioinformatics*, **18**, 1600–1608.

Culhane AC, Perriere G and Higgins DG (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.

Dolédec S and Chessel D (1987) Rhythmes saisonniers et composantes stationelles en milieu aquatique I- description d'un plan d'observations complet par projection de variables. *Acta Oecologica Oecologica Generalis*, **8**, 403–426.

Dudoit S, Gentleman RC and Quackenbush J (2003) Open source software for the analysis of microarray data. *Biotechniques*, **34**(Suppl), S45–S51.

Dysvik B and Jonassen I (2001) J-Express exploring gene expression data using Java. *Bioinformatics*, **17**(4), 369–370.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD and Vingron M (2001) Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10781–10786.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Grigoryev DN, Ma SF, Irizarry RA, Ye SQ, Quackenbush J and Garcia JG (2004) Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biology*, **5**, R34.

Hampton GM and Frierson HF (2003) Classifying human cancer by analysis of gene expression. *Trends in Molecular Medicine*, **9**, 5–10.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**(Database issue), D258–D261.

Hartigan J (1975) *Clustering Algorithms*, Wiley: New York.

Huber W, von Heydebreck A and Vingron M (2003) Analysis of microarray gene expression data. In *Handbook of Statistical Genetics*, *Second Edition*, Balding DJ, Bishop M and Cannings C (Eds.), John Wiley & Sons: Chichester.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kapushesky M, Kemmeren P, Culhane AC, Durinck S, Ihmels J, Korner C, Kull M, Torrente A, Sarkans U, Vilo J, *et al*. (2004) Expression profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Research*, **32**, W465–W470.

Kuo WP, Kim EY, Trimarchi J, Jenssen TK, Vinterbo SA and Ohno-Machado L (2004) A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics*, **37**, 293–303.

Legendre P and Legendre L (1998) *Numerical Ecology*, *Second English Edition*, Elsevier: Amsterdam.

Li W and Yang Y (2001) How many genes are needed for a discriminant microarray data analysis? In *Methods of Microarray Data Analysis*, Lin SM and Johnson KF (Eds.), Kluwer Academic Publishers: Boston, pp. 137–150.

Quackenbush J (2002) Microarray data normalization and transformation. *Nature Genetics*, **32**(Suppl), 496–501.

Raychaudhuri S, Stuart JM and Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing*, **5**, 452–463.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Schlitt T and Brazma A (2002) Learning about gene regulatory networks from gene deletion experiments. *Comparative and Functional Genomics*, **3**, 499–503.

Sharan R, Maron-Katz A and Shamir R (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.

Simon R (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*, **89**, 1599–1604.

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), Article 3. **http://www.bepress.com/sagmb/vol3/iss1/art3**

Stuart JM, Segal E, Koller D and Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4465–4470.

Toronen P, Kolehmainen M, Wong G and Castren E (1999) Analysis of gene expression data using self-organizing maps. *FEBS Letters*, **451**, 142–146.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D and Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D and Holstege FC (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports*, **4**, 387–393.

Wang J, Bo TH, Jonassen I, Myklebost O and Hovig E (2003) Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics*, **4**, 60.

Specialist Review

Differential expression with the Bioconductor Project

Anja von Heydebreck

Max-Planck-Institute for Molecular Genetics, Berlin, Germany

Wolfgang Huber

German Cancer Research Center, Heidelberg, Germany

Robert Gentleman

Dana-Farber Cancer Institute, Boston, MA, USA

1. Introduction

The measurement of transcriptional activity in living cells is of fundamental importance in many fields of research, from basic biology to the study of complex diseases such as cancer. DNA microarrays provide an instrument for measuring the mRNA abundance of tens of thousands of genes. Currently, the measurements are based on mRNA from samples of hundreds to millions of cells, thus expression estimates provide an ensemble average of a possibly heterogeneous population. Comparisons within gene across sample are supported by the current state of the technology, whereas comparisons within samples and between genes are not. In this article, we focus on detecting probes that reflect differences in gene expression associated with specific phenotypes of samples. The Bioconductor Project (**http://www.bioconductor.org**) provides a variety of tools for this purpose. We will illustrate general principles of the analysis, rather than discussing detailed features of individual software packages; for this, we refer the reader to the Bioconductor website and the *vignettes* provided with each of the packages.

Gene expression is a well-coordinated system and hence measurements on different genes are, in general, not independent. Given a more complete knowledge of the specific interactions and transcriptional controls, it is plausible, and likely, that more precise comparisons between samples can be made by considering the joint distribution of specific sets of genes. However, the high dimension of gene expression space prohibits a comprehensive exploration, while the fact that our understanding of biological systems is only in its infancy means that we do not know which relationships are important and should be studied. Practitioners have adopted two different strategies to deal with this situation. One ignores the dependencies and treats each gene as a separate and independent experiment, while

the other strategy attempts to use relevant biological knowledge to reduce the set of genes to a manageable number.

The first strategy ignores the dependencies between genes and analyses the data gene-by-gene, for example, through statistically testing for association between each gene's expression levels and the phenotypic data. This approach has been popular, largely because it is relatively straightforward and a standard repertoire of methods can be applied. However, the approach has a number of drawbacks: most important is the fact that a large number of hypothesis tests will be carried out. This involves a problematic trade-off between specificity and sensitivity. p-value correction methods provide one mechanism for redress, by focusing on specificity, but they are not a panacea, as the price will usually be a large loss in sensitivity.

The second approach is to reduce the number of hypotheses to a more manageable number that directly address the questions of biological interest. Ideas such as nonspecific filtering also fall under this rubric. But the main tool here is the use of prior biological knowledge to focus on a small number of genes and of their associations with each other. Specific examples include examining particular pathways or specific transcription factors (looking for downstream correlates) that have been associated with a disease or condition of interest.

We note that it does not have to be an either-or world. It is entirely possible, and acceptable, to first test specific hypotheses of interest and to then adopt a more exploratory approach as a basis for the next round of experimentation and hypothesis testing.

Carrying out these explorations requires the use of software analysis tools. There are very many different projects and commercial enterprises that provide software solutions. Many of the gene-expression analysis tools present themselves as a graphical user interface to a predefined set of operations on the data. Such systems are easy to use and provided that they incorporate the methodology that the user wants, they perform satisfactorily. In contrast, the Bioconductor project provides an extensible and flexible set of tools that are associated with a full-fledged programming language that has a large collection of numerical and statistical libraries. While more difficult to use, the software is also much more powerful and can accommodate virtually any chosen analysis.

Bioconductor is an international open source and open development software project for the analysis and comprehension of genomic data. Its main engine is *R* (Gentleman and Ihaka, 1996), a language and environment for statistical computing and graphics (**http://www.r-project.org**). The software runs on all major computing platforms, Unix/Linux, Windows, and Mac OS X. Through its flexible data-handling capabilities and well-documented application programming interface (API), it is easy to link it to other applications, such as databases, web servers, numerical or visualization software. *R* and Bioconductor support the concept of *compendiums* (Gentleman and Temple Lang, 2004), which are interactive documents that bundle primary data, processing methods (computational code), derived data, and statistical output with the textual documentation and conclusions.

2. Gene selection

Fundamental to the task of analyzing gene expression data is the need to select those genes whose patterns of expression are related to a specific phenotype of interest. This problem is hampered by the issues mentioned above, in that we would prefer to model the joint distribution of gene expression but the biological knowledge is lacking and instead genes will be treated as independent experiments. In this section, we describe some of the analysis strategies that are in widespread use. Since our own bias is to reduce the set of genes under consideration to those that we consider to be expressed in a substantial subset of the population we will first prefilter the genes using a nonspecific approach (i.e., we do not use phenotypic information to aid in our decision). In addition to a short description of nonspecific filtering, we consider several different testing paradigms.

One popular approach is to simply conduct a standard statistical test, such as the t-test, for each gene separately. As Kendziorski *et al*. (2003) indicate, analyses that treat genes as separate, independent experiments tend to be less efficient than those that adopt a Bayesian approach in order to take advantage of the shared information between genes. Baldi and Long (2001), Tusher *et al*. (2001), Lönnstedt and Speed (2002), Smyth (2004) proposed moderated versions of the t-statistic where the gene-specific variance estimator in the denominator is augmented by a constant that is obtained from the data of all genes. This is especially useful when few replicated samples are available and typically gene-specific variance estimates are highly variable. The moderated t-statistics may be seen as interpolations between a fold-change criterion and the usual t-statistic. Except for the variant of Tusher *et al*. (2001), they coincide with the latter in the case of many replicates. Irrespective of the number of replicates, the use of a fold-change criterion may be beneficial, as this helps to screen out genes whose effect sizes are small in absolute terms, even though they may be statistically significant.

To demonstrate some of the differences between these general approaches, we consider expression data from 79 samples from patients with acute lymphoblastic leukemia (ALL) that were investigated using HGU95AV2 Affymetrix GeneChip arrays (Chiaretti *et al*., 2004). The data were normalized using quantile normalization, and expression estimates were computed using RMA (Irizarry *et al*., 2003). There are a lot of different choices for the preprocessing. This can have much impact on the results of subsequent analyses (Huber *et al*., 2002). The appropriate choice of method depends on the technology and experimental design, and this is still a subject of active research. Here, we sidestep these issues and instead focus on the expression values themselves.

Of particular interest is the comparison of samples with the BCR/ABL fusion gene resulting from a chromosomal translocation (9;22) with samples that are cytogenetically normal. There are 37 BCR/ABL samples and 42 normal samples (labeled NEG).

We will demonstrate some aspects of differential gene expression analysis with this example data set. Code chunks using functionality from *R* and Bioconductor illustrate the calculations.

2.1. Nonspecific filtering

This technique has as its premise the removal of genes that are deemed to be not expressed according to some specific criterion that is under the control of the user. One may also want to eliminate from consideration the genes that do not show sufficient variation in expression across all samples, as they tend to provide little discriminatory power. Many of the genes represented by the 12 625 probe sets on the array are not expressed in B-cell lymphocytes (either in their normal condition or in any of the disease states being considered), which are the cells that were measured in this experiment, and hence the probes for those genes can, and should, be removed from the analysis. In the example below, we require estimated intensities to be above 100 fluorescence units in at least 25% of the samples, and the interquartile range (IQR) across the samples on the log base 2 scale to be at least 0.5. We start with an object, `eset`, that contains the expression data and the phenotypic information for the samples.

```
> library(genefilter)
> f1 <- pOverA(0.25, log2(100))
> f2 <- function(x) (IQR(x) > 0.5)
> ff <- filterfun(f1, f2)
> selected <- genefilter(eset, ff)
> sum(selected)

[1] 2391

> esetSub <- eset[selected, ]
```

The following analysis will be based on the expression data of the 2391 selected probe sets.

2.2. Differential expression – one gene at a time

We now turn our attention to the process of selecting interesting genes by using a statistical test and treating each gene as an independent experiment. Our setting is the simple case where we compare two phenotypes, the BCR/ABL-positive and the cytogenetically normal samples. Many of the principles of the analysis presented below are also valid in the case of a more general response variable, or in a multifactorial experiment. We concentrate on the differences between the *mean* expression levels in the two groups. In other analyses, different properties of the distributions of the expression levels could also be interesting. For instance, Pepe *et al*. (2003) look at the ability of any given gene to serve as a marker for one of the groups, meaning that high (or low) expression of the gene is mainly seen in this group.

Using the *multtest* package, we perform a permutation test for equality of the mean expression levels of each of the 2391 selected genes.

```
> cl <- as.numeric(esetSub$mol == "BCR/ABL")
> resT <- mt.maxT(exprs(esetSub), classlabel = cl, B = 1e+05)
```

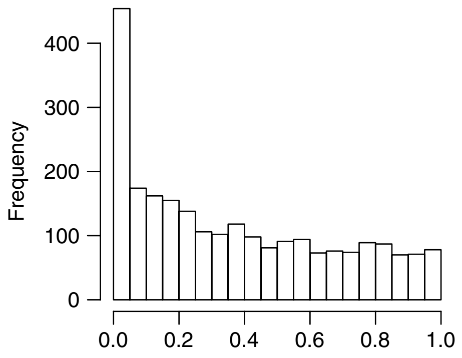


Figure 1 Histogram of *p*-values for the gene-by-gene comparison between BCR/ABL positive and negative leukemias

```
> ord <- order(resT$index)
> rawp <- resT$rawp[ord]
> names(rawp) <- geneNames(esetSub)
```

The histogram of *p*-values (Figure 1) suggests that a substantial fraction of the genes are differentially expressed between the two groups.

In order to control the family-wise error rate (FWER), that is, the probability of at least one false-positive in the set of significant genes, we use the permutation-based maxT-procedure (Westfall and Young, 1993). We obtain 18 genes with an adjusted *p*-value below 0.05. Compare this number to the size of the leftmost bar in the histogram – clearly we are missing out on a large number of differentially expressed genes.

Looking at the description of the probe sets with smallest adjusted *p*-values, we see that the three most significant ones represent the ABL1 gene, which, in the form of the BCR/ABL fusion gene, defines one of the phenotypes we are studying.

```
1636_g_at 39730_at  1635_at  40202_at  37027_at
    "ABL1"     "ABL1"    "ABL1"     "BTEB1"     "AHNAK"
```

As illustrated above, the FWER is a very stringent criterion, and in some microarray studies, only few genes may be significant in this sense, even if many more are truly differentially expressed. A more flexible criterion is provided by the false discovery rate (FDR), that is, the expected proportion of false-positives among the genes that are called significant. We use the procedure of (Benjamini and Hochberg, 1995) as implemented in *multtest* to control the FDR at a level of 0.05, which leaves us with 109 significant genes (note however that this procedure makes certain assumptions on the dependence structure between genes):

```
> res <- mt.rawp2adjp(rawp, proc = "BH")
> sum(res$adjp[, "BH"] < 0.05)

[1] 109
```

2.3. Multiple probe sets per gene

The annotation package *hgu95av2* provides information about the genes represented on the array, including LocusLink identifiers (**http://www.ncbi.nlm.nih.gov/LocusLink**), Unigene cluster identifiers, gene names, chromosomal location, Gene Ontology classification, and pathway associations. While the term *gene* has many aspects and can mean different things to different people, we operationalize it by identifying it with entries in the LocusLink database. One problem that does arise is that some genes are represented by multiple probe sets on the chip. The multiplicities for the HGU95AV2 chip are shown in the following table.

```
Multiplicity            1    2    3   4    5    6    7    8    9
No. LocusLink IDs    6719 1579  505 121   29   18   10    9    1
```

This leads to a number of complications, as we discuss in the following. Of the 2272 LocusLink IDs that have more than one probe set identified with them, we found that in 510 cases our nonspecific filtering step of Section (2.1) selected some, but not all corresponding probe sets.

In Section (2.2), we found that the three top-scoring probe sets all represented the ABL1 gene. However, there are five more probe sets on the chip that also represent the ABL1 gene, none of which passed our filtering step from Section (2.1). The permutation p-values of all eight probe sets are:

```
  1635_at   1636_g_at  39730_at  1656_s_at   32974_at
  0.00001     0.00001   0.00001    0.05800    0.23000
32975_g_at  2041_i_at 2040_s_at
    0.53000     0.59000   0.76000
```

Some caution must be used in interpreting this particular example. The BCR/ABL-positive samples have an mRNA that is the fusion of the BCR gene and the ABL1 gene, and this fused gene is generally highly expressed. The heterogeneity in the p-values listed might be due to the specific location of the probes in the ABL1 gene and whether or not they are able to detect the fused gene. This question could be further investigated using the *hgu95av2probe* package from Bioconductor, which contains information about the probe sequences and their location.

Figure 2 shows a comparison of the t-statistics calculated in Section (2.2) between 206 *pairs* of probe sets that represent the same gene. While in many cases they are approximately concordant, there are a number of cases where a *significant* p-value would be obtained for one of the probe sets but not for the other.

The consistency of the behavior of different probe sets that are intended to represent the same gene is seldom systematically reported. Here, we found some striking discrepancies. How can these be interpreted? In some cases, different probe sets may represent alternative transcripts of the same gene, which could be expressed differently. However, the magnitude of the problem suggests that

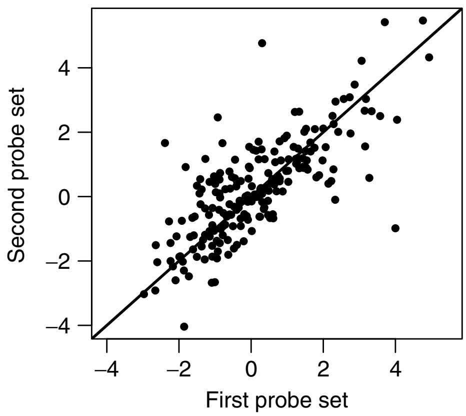


Figure 2 Comparison between the t-statistics for pairs of probe sets that represent the same gene

there may also be errors in the mapping of the probes to LocusLink records. A better understanding of the set of all possible transcripts and their mapping to the genome, but also further and better quality control procedures are needed.

2.4. *The relation between prefiltering and multiple testing*

As explained above, the aim of nonspecific filtering is to remove genes that, for example, because of their low overall intensity or variability, are unlikely to carry information about the phenotypes under investigation. The researcher will be interested in keeping the number of tests/genes as low as possible while keeping the interesting genes in the selected subset.

If the truly differentially expressed genes are overrepresented among those selected in the filtering step, the FDR associated with a certain threshold of the test statistic will be lowered due to the filtering. This appears plausible for two commonly used global filtering criteria: *Intensity-based filtering* aims to remove genes that are not expressed at all in the samples studied, and these cannot be differentially expressed. Concerning the *variability across samples*, a higher overall variance of the differentially expressed genes may be expected, because their between-class variance adds to their within-class variance.

To investigate these presumed effects, we compare the scores for intensity and variability that we used in the beginning for gene selection with the absolute values of the t-statistic, which we now compute for all 12 625 genes.

```
> IQRs <- esApply(eset, 1, IQR)
> intensityscore <- esApply(eset, 1, function(x)
   quantile(x, 0.75))
> abs.t <- abs(mt.teststat(exprs(eset), classlabel=cl))
```

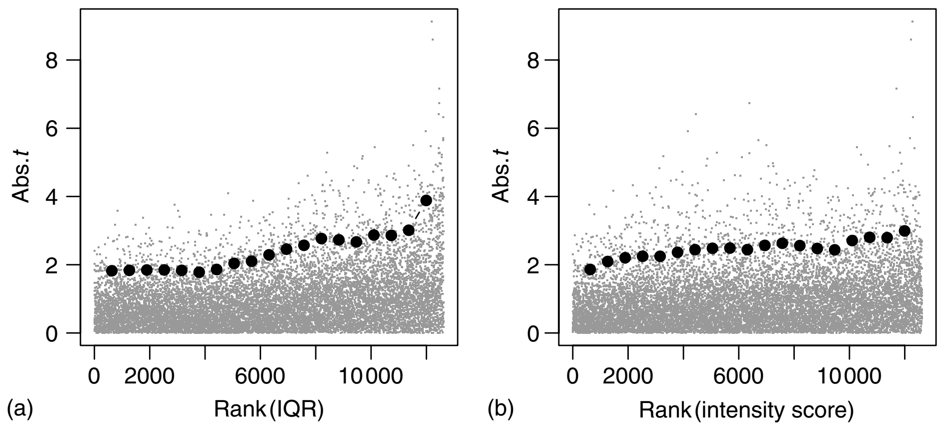


Figure 3 Plots of the absolute values of the t-statistic (y-axis) against the ranks of the values of the two filtering criteria: (a) interquartile range (IQR), (b) an overall intensity score. The larger dark dots indicate the 95%-quantiles of the absolute value of the t-statistic computed for moving windows along the x-axis

The result is shown in Figure 3. Gene selection by the interquartile range (IQR) seems to lead to a higher concentration of differentially expressed genes, whereas for the intensity-based criterion, the effect is less pronounced.

2.5. *Moderated t-statistics*

Many microarray experiments involve only few replicated samples, which makes it difficult to estimate the gene-specific variances that are used, for example, in the t-test. In the *limma* package, an Empirical Bayes approach is implemented that employs a global variance estimator s_0^2 computed on the basis of all genes' variances. The resulting test statistic is a moderated t-statistic, where instead of the single-gene estimated variances s_g^2, a weighted average of s_g^2 and s_0^2 is used. Under certain parametric assumptions, this test statistic can be shown to follow a t-distribution under the null hypothesis with the degrees of freedom determined by the data (Smyth, 2004).

With 79 samples at hand, there is no big difference between the ordinary and the moderated t-statistic. Here we use the ALL data to illustrate the behavior of the different approaches for small sample sizes: We repeatedly draw random small sets of arrays from each of the two groups and apply different statistics for differential expression. The results are compared to those of the analysis of the whole data set. As an approximation, we declare the 109 genes with an FDR below 0.05, based on the whole set of samples, as truly differentially expressed genes.

```
> is.diff <- res$adjp[order(res$index), "BH"] < 0.05
```

Now we consider a random subset of four arrays per group and compute ordinary as well as moderated t-statistics for them.

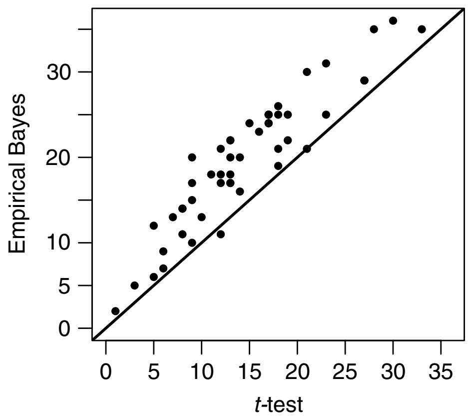


Figure 4 Number of true-positives among the top 100 genes selected by the t-test (x-axis) and a test based on a moderated t-statistic, as implemented in the *limma* package (y-axis)

```
> groupsize <- 4
> design <- cbind(c(1, 1, 1, 1, 1, 1, 1, 1), c(0, 0, 0, 0, 1, 1,
+ 1, 1))
> g1 <- sample(which(esetSub$mol== "NEG"), groupsize)
> g2 <- sample(which(esetSub$mol== "BCR/ABL"), groupsize)
> subset <- c(g1, g2)
> fit <- lm.series(exprs(esetSub)[, subset], design)
> eb <- ebayes(fit)
> tsub <- mt.teststat(exprs(esetSub)[, subset], classlabel=cl[subset],
    +test = "t.equalvar")
> rawpsub <- 2 * (1 - pt(abs(tsub), df=2 * groupsize - 2))
```

For a comparison, we define true-positives as those genes that are among the 100 most significant ones in the tests based on the small data set, and which were also selected in the analysis of the full data set.

```
> TPeb <- sum(is.diff[order(eb$p.value[, 2])[1:100]])
> TPtt <- sum(is.diff[order(rawpsub)[1:100]])
```

Repeating this procedure for 50 random subsets of arrays, we then compare the numbers of true-positives between the two methods. The results are shown in Figure 4.

It appears that for small sample sizes, the moderated t-statistic, which may be seen as an interpolation between the t-statistic and a fold-change criterion, has a higher power than the simple t-test.

3. Asking specific questions – using metadata

We now turn our attention to analyses that are based on asking more specific and detailed questions that relate directly to the known biology of the system or disease

under study. We examine three different biological aspects. First, we ask whether differentially expressed genes are enriched on specific chromosomes.

3.1. Chromosomal location

In the following, we look at all genes with an unadjusted $p < 0.01$ (taking the median p-value in the case of several probe sets per gene), and conduct a Fisher test for each chromosome to see whether there are disproportionally many differentially expressed genes on the given chromosome. The information on the chromosomal location of genes is taken from the annotation package *hgu95av2*.

```
> ll <- getLL(geneNames(esetSub), "hgu95av2")
> chr <- getCHR(geneNames(esetSub), "hgu95av2")
> names(chr) <- names(rawp) <- NULL
> chromosomes <- unique(chr[!is.na(chr)])
> ll.pval <- exp(tapply(log(rawp), ll, median))
> ll.chr <- tapply(chr, ll, unique)
> ll.diff <- ll.pval < 0.1

> p.chr <- sapply(chromosomes, function(x) {
+      fisher.test(factor(ll.chr == x), as.factor(ll.diff))$p.value
+ })

> signif(sort(p.chr), 2)

      7      17       X      15       8      21       Y       3
 0.0062 0.1100 0.1800 0.1800 0.2000 0.2900 0.3300 0.3600
      6      22       4      12       5       9       1      10
 0.4400 0.4800 0.5600 0.5900 0.6100 0.6700 0.6900 0.6900
      2      16      18      20      11      19      14      13
 0.7000 0.8200 0.8300 0.8700 0.9100 0.9200 1.0000 1.0000
```

We identify chromosome 7 as being of potential interest. Out of the 87 genes that are mapped there, 35 were differentially expressed according to our criterion. We further explore this by visualizing gene expression on Chromosome 7. In Figure 5, we present the output of the `plotChr` function from the *geneplotter* package. In this plot, gene expression values for each sample are separated according to the strand of the chromosome that the gene is located on. Then a lowess smoother is applied (across genes, within sample) and the resultant smooth estimate is plotted. The top part of the plot represents the positive strand, while the bottom part of the plot represents genes on the negative strand. High expression values are near the outside of the plotting region, while low expression values correspond to the center (i.e., the expression values are mirrored). One can readily detect regions of heterogeneity between samples (such as the left end of both strands and again around `38299_at` on the negative strand).

3.2. Using Gene Ontology data

A second source of valuable biological data that is easily accessible through Bioconductor software is data from the Gene Ontology (GO). It is known that

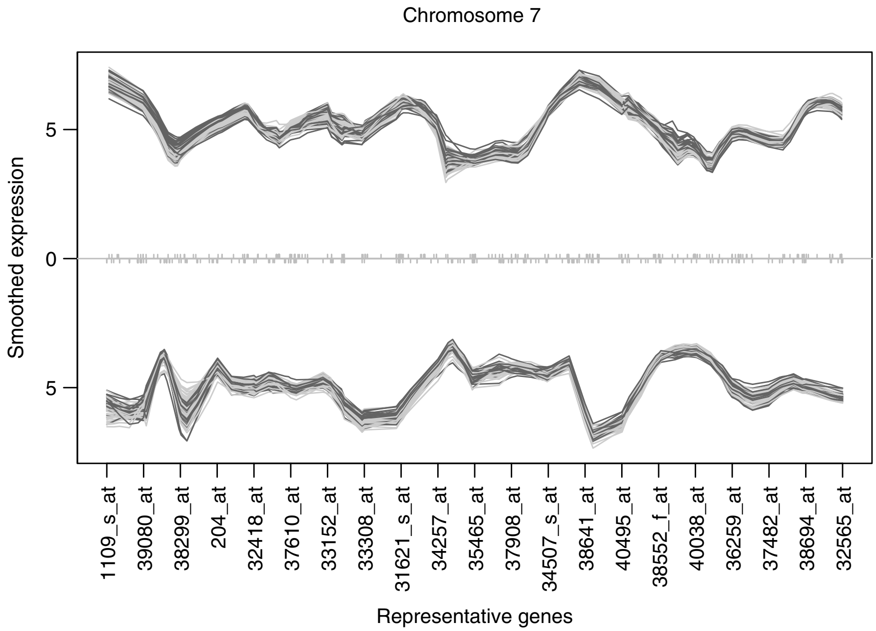


Figure 5 Smoothed expression for chromosome 7, both strands are plotted separately

many of the effects due the BCR/ABL translocation are mediated by tyrosine kinase activity. It will therefore be of interest to examine genes that are known to have tyrosine kinase activity. We examine the set of GO terms and identify the term, `GO:0004713` from the *molecular function* portion of the GO hierarchy as referring to `protein-tyrosine kinase activity`. We can then obtain all Affymetrix probes that are annotated at that node, either directly or by inheritance, using the following command.

```
> tykin <- unique(lookUp("GO:0004713", "hgu95av2", "GO2ALLPROBES"))
```

We see that 234 probe sets are annotated at this particular term. Of these, only 41 were selected by the nonspecific filtering step of Section (2.1). We focus our attention on these probes and repeat the permutation t-test analysis of Section (2.2).

In the analysis of the GO-filtered data, seven probe sets have FWER-adjusted p-values less than 0.1. They are printed below, together with the adjusted p-values from the first analysis that involved 2391 genes.

```
[1] "GO analysis"

   1635_at  1636_g_at  39730_at  40480_s_at  2039_s_at
   0.00001     0.00001   0.00001     0.00002    0.00030
   36643_at  2057_g_at
    0.02383    0.08282
```

```
[1] "All Genes"

   1635_at  1636_g_at  39730_at  40480_s_at  2039_s_at
   0.00001     0.00001   0.00001     0.00095    0.01407
   36643_at    2057_g_at
    0.46938      0.82884
```

Owing to the reduced number of tests in the analysis focused on tyrosin kinases, we are left with more significant genes after correcting for multiple testing. For instance, the probe set `36643_at`, which corresponds to the gene DDR1, was not significant in the unfocused analysis, but would be if instead the investigation was oriented toward studying tyrosine kinases.

3.3. *Using pathways*

In a closely related disease, chronic myeloid leukemia, there is evidence of a BCR/ABL-induced loss of adhesion to fibronectin and the marrow stroma. This observation suggests that there may be substantial differences between the BCR/ABL samples and the NEG samples with respect to the expression of genes involved in the integrin-mediated cell adhesion pathway. A version of this pathway was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) as Pathway 04510. The relevant genes were identified and mapped to the corresponding Affymetrix identifiers using the *hgu95av2* and *KEGG* packages from the Bioconductor Project.

There were 110 probes that correspond to 67 LocusLink identifiers. Now we can ask whether any of these are differentially expressed between the two groups. Users can either apply our nonspecific filter first, or not, depending on their particular point of view.

We did not apply the nonspecific filtering of Section (2.1) and simply applied the `mt.maxT` function to obtain permutation adjusted p-values for differences between the BCR/ABL and the NEG group. This analysis identified four probe sets that had significant p-values. Three corresponded to FYN and one to CAV1. To compare with the other analyses reported above, we determined that there were four probe sets included on the HGU95AV2 chip for FYN and of these three were selected by the nonspecific filtering in Section (2.1) and the same three were selected here. In Section (2.2), we found two of the three FYN probes had significant p-values, after adjustment. On the other hand, CAV1 is represented by a single probe set and it was not selected in Section (2.1) because it was expressed at levels above 100 in only nine samples.

A scatterplot matrix of the gene expression data for the four selected probes is presented in Figure 6. This plot provides corroboration for some of some of the points made previously about probe set fidelity. For two of the FYN probe sets, the agreement is remarkably good. But their correlation with the third set of FYN measurements is poorer. We can also see that CAV1 expression is related to the BCR/ABL phenotype, but does not seem to be related to FYN expression.

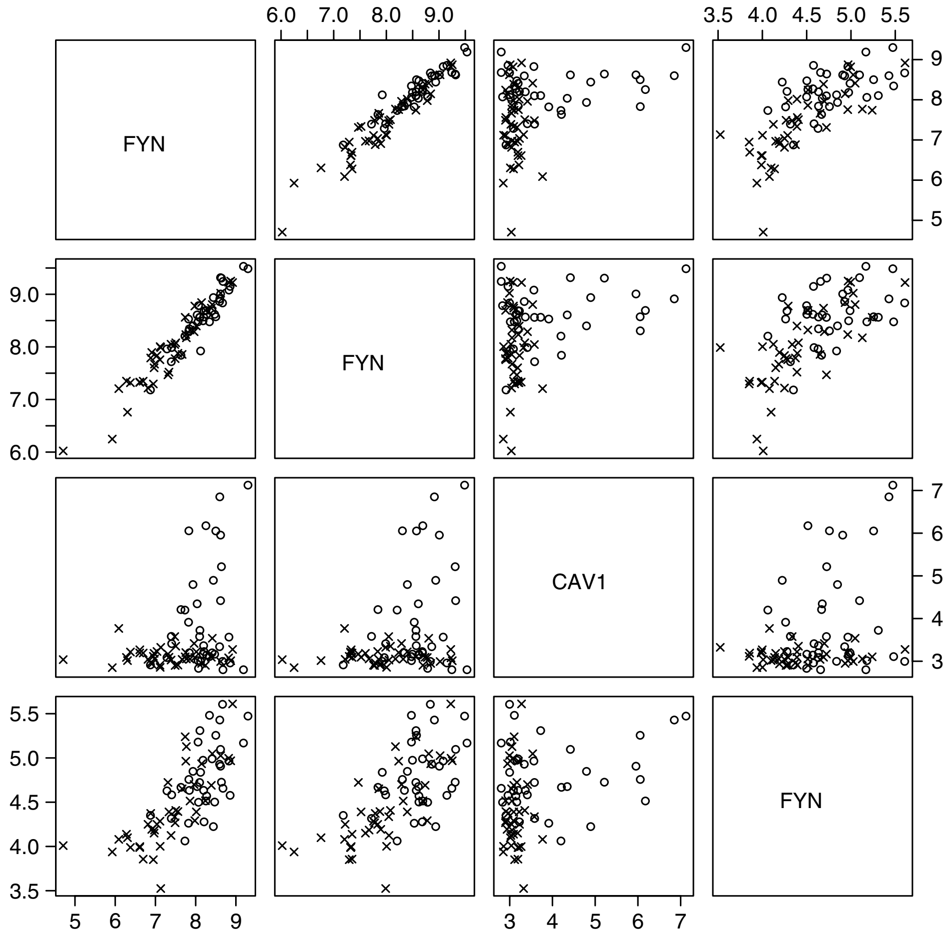


Figure 6 Pairwise scatterplots of the expression values for the four probe sets selected. Crosses indicate BCR/ABL negative samples, circles denote BCR/ABL positive samples

4. Discussion

Understanding the complex dependencies between the transcriptional activities of genes remains an important, difficult, and essentially unsolved problem. Current analysis of differential gene expression mostly works on a gene-by-gene basis, where the necessary concepts are within reach and established statistical tests can be used. Statisticians have directed a lot of attention to methods for p-value correction. This has led to widespread adoption of these methods and particularly of the false discovery rate (FDR). They rely on the assumption that those tests with the most extreme p-values are most likely to have arisen from hypotheses that are truly false. The decision boundary is shifted sufficiently to ensure that some required proportion of those hypotheses deemed false are in fact truly false. But such an approach cannot be viewed as a solution to the real problem of identifying all

truly false hypotheses. We note that in adjusting the rejection boundary, one has also taken an unknown number of truly false hypotheses and deemed them to be true. Furthermore, there is no guarantee that the biologically most relevant genes are the ones with the most extreme p-values. Genome-wide expression profiles may be dominated by secondary changes in gene expression, such that the primary effectors may be buried within long lists of statistically significant genes. There is no simple fix to this problem apart from testing fewer and more directed hypotheses and where possible this should be done.

In the examples of Section (3), we demonstrated a variety of analyses incorporating biological information that are currently available and easily carried out. We showed that such analyses have the capability of extracting more information from the data than an omnibus approach can. Increases in our understanding of the relevant biology coupled with gains in statistical knowledge will likely lead to even more promising analyses.

Acknowledgments

We are grateful to Drs. J. Ritz and S. Chiaretti of the DFCI for graciously providing their data. And also, to the developers of *R* and those of the Bioconductor software packages that have made it possible to so easily consider a wide variety of approaches to analyzing these data.

References

Baldi P and Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J and Foa R (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Gentleman R and Ihaka R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314, **http://www.r-project.org**.

Gentleman R and Temple Lang D (2004) Statistical analyses and reproducible research. *Bioconductor Project Working Papers. Working Paper 2*, **http://www.bepress.com/bioconductor/paper2**.

Huber W, von Heydebreck A, Sültmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl 1), S96–S104.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kendziorski C, Newton M, Lan H and Gould M (2003) On parametric Empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**(24), 3899–3914.

Lönnstedt I and Speed TP (2002) Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Pepe M, Longton G, Anderson G and Schummer M (2003) Selecting differentially expressed genes from microarray experiments. *Biometrics*, **59**, 133–142.

Smyth G (2004) Linear models and Empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Westfall P and Young S (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons.

Specialist Review

Mass spectrometric data mining for protein sequences

Christian Cole, Patrick J. Lester and Simon J. Hubbard
University of Manchester, Manchester, UK

1. Introduction

The ability to study genes and their protein products on a genome-wide basis is driving a paradigm shift in modern biology from hypothesis-driven to more predictive, hypothesis-generating bioscience, where the state of every gene and/or gene product in different biological conditions can potentially be studied. Driven by complete genome sequences, proteomics is one example of these techniques, which offers a key advantage over many other postgenomic functional studies in that the proteome is predominantly the functional entity in a cell or tissue. Proteins are responsible for carrying out most biological functions and, therefore, also for dysfunction and disease. Hence, proteins are frequently targeted for treatment in the form of drugs and therapeutics. Proteomic data sets can provide both qualitative and quantitative information on proteins, potentially thousands of different proteins expressed in a cell or tissue. At the heart of proteomics lies advances in mass spectrometry (see Volume 3, Section 1) and protein/peptide separation (*see* Article 2, **Sample preparation for proteomics**, Volume 5) along with, crucially, attendant bioinformatics tools. Whether the proteomics experiment is qualitative or quantitative, protein species must at some stage be identified by bioinformatics tools from mass spectra derived from the proteins themselves, or more commonly from their peptides or peptide fragment ions. This technique exploits the knowledge of the exact masses of the amino acid constituents of protein sequences along with the growing data banks of protein sequences against which the mass spectral information may be compared to match peptides (and thereby their "parent" proteins).

In this review, we discuss the data mining approaches being taken to identify proteins from characteristic mass spectra, and recent advances in both experimental and bioinformatic approaches, focusing on the two principal application areas of peptide mass fingerprinting (PMF) and tandem mass spectrometry (MS/MS). Both are widely used to identify proteome proteins from mass spectra, although MS/MS approaches can deliver protein sequence information directly. Tandem MS

permits the sequence of a peptide (and by extension, protein) to be determined *de novo* in some cases so that the protein sequence does not need to be known *a priori*. This is of particular importance when the sequence of the genome in question is not currently available or there is a difference between the real and "virtual" proteome; that is, when proteins are posttranslationally modified leading to different isoforms that are not predicted directly from the genome sequence. Indeed, when a complete genome sequence is unavailable, this leads to the problem of cross-species protein identification where sequence information is of particular importance. We also examine this issue and review approaches to cross-species proteomics, as well as current developments in tandem MS peptide database searching.

2. Mass spectrometry in proteomics

In order to appreciate how protein sequences can be studied via mass spectrometry, the basic techniques are presented here and discussed in more detail elsewhere (see Volume 3, Section 1). In using mass spectrometry for proteomics, "soft ionization" techniques are exploited whereby the analyte, usually proteolytic peptide fragments of a protein or mixture of proteins, is introduced into the instrument as a solution for electrospray ionization (ESI) (*see* Article 14, **Sample preparation for MALDI and electrospray**, Volume 5), or as a solid dissolved in a UV-absorbing crystalline matrix for matrix-assisted laser desorption/ionization (MALDI) (*see* Article 14, **Sample preparation for MALDI and electrospray**, Volume 5). The gaseous ions generated are separated by their mass-to-charge ratios (m/z) in a mass analyzer and then detected by either an electron multiplier or multichannel plate detector. This occurs on a microsecond timescale, allowing the whole process to be automated and, therefore, facilitates high-throughput analysis of many samples, yielding m/z measurements for each of the analyte ions. The masses of the proteolytic peptides (usually generated via tryptic digestion) can then be used to search protein sequence databases in order to determine the most likely protein(s) that could have generated the equivalent peptide masses; a technique usually referred to as peptide mass fingerprinting (PMF).

Mass spectrometry can also be used for peptide/protein sequencing via tandem MS. In this instance, collision-induced dissociation (CID) is used to initiate fragmentation of the protein/peptide. Here, the so-called precursor ions are selected for in an initial mass-to-charge (m/z) ratio measurement. This precursor ion is then subjected to fragmentation where some of the kinetic energy is converted into internal energy through collisions with gaseous atoms (e.g., Argon), causing the molecule to fragment. The m/z of the subsequent fragment ions, referred to as the fragment ion series, are then measured and with this information, the sequence of the precursor ion can be inferred. Other fragmentation techniques can be employed that will generate an alternative fragment ion series, from that generated in CID, of the peptide. The information obtained from the fragment ion spectra can either be annotated or used directly to search for the peptide sequence that produced it in a sequence database.

3. Data mining using PMF

3.1. Same-species protein identification

PMF remains the method of choice for single-protein identification in many laboratories as it can easily be automated into a successful high-throughput protein identification strategy compared to the more elaborate tandem MS approaches (Wilke *et al*., 2003). The basic problem is to determine which protein(s) has produced the set of proteolytic peptide ions obtained by mass spectrometry. Typically, proteins are first separated by two-dimensional polyacrylamide gel electrophoresis (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5). The resultant protein "spots" are digested with a proteolytic enzyme, usually trypsin, and the resulting peptides are extracted from the gel and analyzed on the mass spectrometer. A bioinformatic search tool then uses the experimentally derived primary (peptide ion *m/z* values) and secondary protein information to search sequence databases for an entry that best correlates with the experimental *m/z* data within a given experimental error. Ideally, the list of results will enable the user to identify the experimental protein with high confidence.

Confident protein identification by PMF broadly depends on the number of experimental proteolytic peptides that match theoretical database peptides calculated using the same cleavage specificity within the experimental error tolerance. Early algorithms ranked proteins on the basis of the total number of matches for each database entry (Mann *et al*., 1993). This relied on high-quality experimental data, and often failed to accurately discriminate between random matches from large database proteins and true-positive hits. Unambiguous identification was generally assumed for database proteins with five matching peptides, a sequence coverage of approximately 15% and a gap between the potential "true-positive" and next best "false-positive" match. However, more recent algorithms such as MOWSE/Mascot (Pappin *et al*., 1993; Perkins *et al*., 1999), ProFound (Zhang and Chait, 2000), PeptIdent2 (Gras *et al*., 1999), and MS-Fit (Clauser *et al*., 1999) typically use sophisticated scoring systems to distinguish true-positive from false-positive matches, improving upon simple ranking on the number of peptide matches (Eriksson *et al*., 2000).

Mascot incorporates probability-based scoring, extending the MOWSE algorithm (Pappin *et al*., 1993), which also forms the basis for MS-Fit (Clauser *et al*., 1999). This algorithm accounts for the dependency on protein mass of the peptide mass distributions in the database, effectively down-weighting random matches to larger database proteins. The algorithm calculates a probability that the protein match is nonrandom, which is then converted into a Mascot score by a log transform. Another popular PMF search engine, ProFound (Zhang and Chait, 2000), uses a scoring algorithm based on Bayesian probability theory to identify a specific protein in a database from its attendant peptide masses. ProFound differs from Mascot in that it calculates a probability that each protein in the database is the correct match, assuming that the true protein *is* in the database, ranking candidate matches in decreasing likelihood. ProFound also attempts to reduce the bias toward random matches of larger proteins by introducing a specific term in the scoring function and, importantly, also models the experimental error in *m/z*

determinations. Unlike Mascot, ProFound assumes that matching peptides from the true protein will be normally distributed about the theoretical database peptide mass with a standard deviation proportional to the experimentally determined mass error. Conversely, Mascot gives equal weighting to any match within the error tolerance supplied by the user. It is unclear whether these provide universal advantages or disadvantages to protein identification for PMF. For all algorithms, other background information (species of origin, molecular mass of the intact protein, enzyme cleavage chemistry, etc.) can be included and additional empirical peptide information can be submitted when available, such as limited sequence information with MS-Tag (Clauser *et al*., 1999). This background information is either explicitly or implicitly modeled by the scoring system through appropriate choice of search parameters.

A distinct advantage of the Mascot and ProFound algorithms is the significance testing that allows users to assign confidence to their hits. Extending MOWSE into a probabilistic frame allows Mascot to determine a threshold at the 5% significance level, and the software provides a simple color-coded histogram showing nonrandom protein matches above this limit. ProFound supplies significance testing via expectation values calculated from raw probability scores, based on the distribution of random scores, providing e-values in much the same way as a BLAST search (Altschul *et al*., 1997). In this case, the number of expected protein matches with such a score is returned, depending on the database searched.

Independent of algorithm choice, confident matches are often subject to user opinion, especially when several sequences return a score indicating a confident match. This can be due to database redundancy or error, or several species in the database yielding random noise from protein homologs. The objective assessment of database search results to determine the probability that the identified protein is a false-positive has been investigated (Eriksson *et al*., 2000) and remains of significant importance with the emphasis on high-throughput studies. A recent comparison of PMF algorithms (Chamrad *et al*., 2004) examines some key points comparing the algorithms, finding that Mascot and ProFound outperform MS-Fit, although they both have advantages for some proteins. Although all the algorithms attempt to account for scoring bias from high-mass proteins, an apparent bias toward large proteins was observed in identifications from ProFound and MS-Fit, but not Mascot. In addition, the same group has developed a meta-score approach based on the three algorithms, which offers some advantages (Chamrad *et al*., 2003) and appears to be a successful "consensus" strategy that has worked well in other areas of bioinformatics such as secondary structure prediction. Again, an independent significance score in the form of an expectation value is provided and appears to outperform the independent PMF algorithms.

3.2. Cross-species protein identification

Although genome-sequencing projects continue apace, there will always remain unsequenced species owing to the incredible diversity of life on this planet. The cross-species protein identification problem, where homologous (and ideally, *orthologous*) proteins should be identified from experimental data, is particularly

acute for PMF-based strategies. In this case, even a single amino acid substitution can radically alter the mass of a peptide preventing a match to orthologous peptides (and adding noise by creating a match to random, unrelated peptides). Despite this, several approaches have proven successful in cross-species PMF recently, as has been recently reviewed (Lester and Hubbard, 2002; Liska and Shevchenko, 2003a). Generally, approaches have required the use of more than just proteolytic peptide mass data as this is generally poorly conserved across species barriers (Wilkins and Williams, 1997; Lester and Hubbard, 2002). Instead, features such as amino acid composition, protein mass, and isoelectric focusing point (pI) were much better conserved, and attempts to combine these and other protein and peptide features have shown some limited success (Cordwell *et al*., 1997; Wilkins *et al*., 1998). However, this requires a comprehensive amino acid analysis of the protein in question, which is not as amenable to high-throughput studies as PMF. Some groups have described successful cross-species protein identification strategies using PMF, including marsupial proteins using a mammalian database and MultIdent (Verrills *et al*., 2000) and the characterization of *Ochrobactrum anthropi* proteins using amino acid composition and/or peptide mass fingerprinting (Wasinger *et al*., 1999). In addition to this, peptides corresponding to functionally important (conserved) motifs have been used in order to identify similar patterns in phylogenetically distant organisms (Cordwell *et al*., 1997). The authors suggest that sequences sharing above 60% sequence identity may be identified using PMF via homologous proteins and a ± 6-Da peptide mass tolerance, although this yields many false-positives. Other recent approaches have focused on high mass–accuracy determinations of peptide masses, in the range ± 0.5–5 ppm, where it is possible to reduce the number of candidate proteins to a tractable number so that single amino acid substitutions may be incorporated into the search (Clauser *et al*., 1999). Hence, coupled with high-accuracy peptide mass measurements, homologous proteins with 70% sequence identity or better can be matched with confidence. Indeed, an approximate cutoff of 70% sequence identity has been suggested by several groups as a limit beyond which cross-species PMF becomes impossible without detailed sequence information (Cordwell *et al*., 1997; Clauser *et al*., 1999) despite a considerable enrichment above random of tryptic peptide conservation even at lower identity levels (Lester and Hubbard, 2002).

When partial genomic sequence information is available, then several approaches have proven successful, relying on preliminary, unannotated genome sequences and related resources such as expressed sequence tags (ESTs) (Choudhary *et al*., 2001; Lisacek *et al*., 2001; Kwon *et al*., 2003). A recent study devised an approach to search raw genome sequences directly with PMF data, termed genome fingerprint scanning (GFS), which does not require annotation of coding regions or open reading frames (Giddings *et al*., 2003). Instead, an *in silico* digest of the whole genome is performed in all six reading frames and peptide ions matched via their *m/z* values to the putative "genomic" tryptic peptides. Genomic regions where a high density of such matches are found are then examined and an algorithm scores each sequence window according to the number of in-frame hits, consistency of frame assignment, missed cleavages, and abutment of fragments. The algorithm performs comparably to standard PMF tools but offers advantages over protein

database searches since GSF can spot unannotated proteins and cope with some limited sequencing/frameshift errors. Given the success of tandem MS, however, this approach remains the one that would normally be applied only when others fail.

EST databases, typically available prior to completion of genome sequences, are targeted toward protein coding sequences, represent the majority of all sequence entries in Genbank, and can be obtained at a fraction of the cost of a whole-genome sequence. However, they are problematic for PMF approaches since they are usually only 500–600 nucleotides in length and do not contain the whole protein sequence. One study has circumnavigated this problem by checking for consistency of EST matches in six reading frames across "contig" sequences derived from assembly of overlapping EST sequences, ensuring that matches are in the same reading frame, and using BLASTX (Altschul *et al*., 1997) homology searches to correctly identify likely peptides (Lisacek *et al*., 2001). Many EST contigs will contain alternatively spliced forms, and careful analysis of the data allows these biologically important protein variants to be identified, which would not otherwise be evident from a translated genome sequence.

4. Data mining using tandem MS

Although PMF remains the most amenable method for automated, high-throughput proteomics for most laboratories, tandem MS approaches are catching up rapidly and offer a number of advantages particularly for cross-species and *de novo* studies. The rise of tandem MS methodologies has aided protein identification by giving a partial or complete peptide sequence for searching against a database (Eng *et al*., 1994; Mann and Wilm, 1994). This additional information on a peptide, in conjunction with its *m/z*, narrows down the list of possible correct matches, and can be obtained from a variety of instruments including ESI triple quadrupole, MALDI-QTOF (quadrupole time-of-flight), ion-trapping, and Fourier Transform Ion Coupled Resonance (FT-ICR) instruments, often coupled with a liquid chromatography stage to aid the separation of complex samples (Aebersold and Mann, 2003; *see also* Article 5, **FT-ICR**, Volume 5, Article 7, **Time-of-flight mass spectrometry**, Volume 5, Article 8, **Quadrupole mass analyzers: theoretical and practical considerations**, Volume 5, and Article 9, **Quadrupole ion traps and a new era of evolution**, Volume 5). The latter is used widely with ion-trapping instruments in the Multidimensional Protein Identification Technology (MudPIT) pioneered by Yates (Washburn *et al*., 2001; *see also* Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5).

The added information available from tandem experiments generates its own set of bioinformatic challenges. Unless performing *de novo* peptide sequencing, the added sequence information generated by tandem MS requires that the database must contain peptide sequences under study in addition to its mass, as is the sole requirement for PMF. In principle, a BLAST (Altschul *et al*., 1997) or FASTA-type (Pearson and Lipman, 1988) sequence search filtered for the peptides of the right mass would rapidly give a list of possible matches. However, tandem results are rarely that precise in their description of a peptide's sequence, often containing

ambiguities such as uncertainty due to isobaric residues (Ile/Leu and Gln/Lys) or modifications (e.g., acetylation is equivalent to a change from Ser to Glu) (Creasy and Cottrell, 2002; Chamrad *et al*., 2004). Often, the peptide sequence information is limited to only a few C- or N-terminal residues, with the majority of the sequence still unknown (Mann and Wilm, 1994). Additionally, the protein/peptide sequence in a database may not be an exact match to that determined in the MS/MS experiment, due to misannotation, database errors, existence of multiple isoforms and/or paralogs in the database.

Although direct annotation/interpretation of MS/MS peptide spectra rarely determines the complete amino acid sequence, there is frequently sufficient data, coupled with the precursor ion *m/z*, to determine the peptide from a protein sequence database. Indeed, modern software tools facilitate automatic database searches using uninterpreted spectra along with constraints on the precursor ion *m/z* range at which the first mass spectrometer was "gated" (Eng *et al*., 1994; Perkins *et al*., 1999; Field *et al*., 2002). Some systems permit real-time searching in a complex LC-MS/MS run, or the database searching may be performed off-line.

There are many algorithms for mining MS/MS data for peptide/protein sequences. Of these, the two most widely used are Mascot (Perkins *et al*., 1999; Creasy and Cottrell, 2002) and SEQUEST (Eng *et al*., 1994; Yates *et al*., 1996). As mentioned for PMF searching, Mascot uses a probability-based scoring system to match database sequences to the search data. The algorithm is able to integrate probability scores from both PMF (i.e., peptide *m/z* matches) and MS/MS (i.e., peptide fragment ion *m/z* matches) searches to derive an overall probability of matching a given protein. The score calculated is the chance of the protein match being obtained at random. The significance of this probability score is then readily apparent, and Mascot reports hits above a 5% significance level as confident matches. SEQUEST uses a cross-correlation algorithm that compares the 200 most abundant experimental spectral peaks for the search peptide with proteolytic peptide ion fragment spectra generated *in silico*. The algorithm searches the top 500 database peptides that match the parent ion's mass (within a set tolerance) and then calculates a cross-correlation score (X_{corr}) between the processed experimental spectrum and the reconstructed database spectra.

Although SEQUEST does not have a significance score, as is the case with Mascot, it is claimed that if the difference in score (ΔC_n) between the first and second ranked hits is <0.1, then it is likely that the top hit is a false-positive (Yates *et al*., 1996). Despite this, in a recent paper comparing the relative performance of the two algorithms, on the basis of standard search parameters, SEQUEST was found to separate true-positives from false-positives "more obviously" than Mascot (Chamrad *et al*., 2004). Mascot's scoring system was found to be more sensitive to changes in the search parameters than SEQUEST, but overall, the initial results were not greatly improved upon. The single most effective change for improving protein matching was to reduce the database from all species to just mouse, the species of interest. Apart from improving the number of correct matches, reducing the size of the database also improves on the search speed. However, the downside is that with a reduced database the chances of a random hit are increased owing to deterioration in score separation between random and nonrandom hits, suggesting

that there is a compromise between specificity and sensitivity when altering search parameters.

Although Mascot's significance testing provides a measure of confidence in the database match obtained with the search data, this does not mean, however, that a top match that is not above the 5% significance score is necessarily an incorrect match. However, significance testing is widely valued by users, and the popularity of the SEQUEST algorithm has led to the development of a statistical framework for the assessment of peptide and attendant protein matches derived from SEQUEST hits (Keller *et al*., 2002; Nesvizhskii *et al*., 2003). In the first of these studies, expectation maximization is used to obtain parameters for a discriminant function F that is used to classify SEQUEST peptide identifications as either correct or incorrect. The algorithm is trained on a set of known peptide identifications from 18 known mammalian proteins, and also incorporates knowledge of the number of tryptic termini and considers doubly and triply charged precursor ions. This approach can increase the number of correct SEQUEST identifications by up to 46%. Building upon this, a Bayesian framework is constructed using the same data and bacterial decoy proteins to provide probability estimates for individual proteins being correctly identified. The algorithm considers both the number of sibling peptides confidently identified in an LC-MS run for each protein, as well as degenerate peptides present in multiple proteins in the database, and can offer high sensitivity (94%+) at low error rates (1.2%), frequently correctly identifying low abundance, low molecular weight proteins from only a single peptide match. Approaches such as these are critical for the high-throughput LC-MS/MS experiments where manual confirmation of thousands of protein and peptide matches is impractical.

Other algorithms tend to be similar in their approaches. MS-Tag/MOWSE (Pappin *et al*., 1993; Clauser *et al*., 1999), PepFrag (Fenyö *et al*., 1998), and PepSea (Mann and Wilm, 1994) use fragment ion data in combination with the parent ion *m/z* to generate a list of potential peptides to search against a database using homology-based search strategy. An advantage of this approach is that it is error-tolerant and thus will match similar sequences. However, it is also worth noting that error-tolerant tandem MS searches are possible directly with the latest version of Mascot (Creasy and Cottrell, 2002). Newer methods such as Sonar (Field *et al*., 2002) and Tandem (Craig and Beavis, 2003) concentrate on the speed of searching as well as accuracy of matching. Sonar has the additional advantage of being able to search genomic as well as proteomic databases. Sonar employs a scoring system based on "peptidic vectors" derived from the experimental and database peptides, and correct matches are predicted on the dot-product of the vectors. The resultant scores are cast into an expectation value framework derived from a set of simulations that leads to e-values that are assigned to each peptide, and ultimately protein, match (Field *et al*., 2002).

Recent studies have seen several new algorithms being developed. The Tandem algorithm exploits some of the features of SONAR as well as other algorithms and scoring systems and word-searching methods of BLAST to speed up tandem MS database searching (Craig and Beavis, 2003). PEP_PROBE (Sadygov and Yates, 2003) uses a hypergeometric probabilistic distribution model to improve the scoring of peptide matches to a database, assuming that fragment ion frequencies follow a

hypergeometric distribution, and thus the probability of a random match must also follow this distribution. Another recent algorithm, ProbID (Zhang *et al*., 2002), also uses a novel probabilistic model for the ranking of tandem MS data, which performs similarly to SEQUEST, but its significance testing seems to differentiate well between similar matches. Although these algorithms promise advances in methodology, they still need to be fully validated and compared to current methods to see their full utility.

4.1. JEDI approaches

Although tandem MS approaches can generate extensive sequence information, it is not always necessary or possible to generate a complete (albeit with ambiguities) sequence for a tryptic peptide. PMF is already a powerful technique in identifying proteins from their peptide masses alone, but the additional knowledge of even the presence of a single amino acid in a peptide can improve database searching significantly; theoretically, 83% of all *Saccharomyces cerevisiae* and 79% of all *Caenorhabditis elegans* proteins could be identified from a single aspartate-containing peptide (Sullivan *et al*., 2001), and partial sequence information via a "peptide sequence tag" frequently permits unambiguous protein identification from a single peptide match (Mann and Wilm, 1994). These limited sequence approaches have been labeled the JEDI approach (Just Enough Diagnostic Information) and demonstrate the power of protein identification from directed, promoted fragmentation pathways using a variety of derivatization chemistries (Brancia *et al*., 2001; Sidhu *et al*., 2001; Sullivan *et al*., 2001; Karty *et al*., 2002). This is especially useful when high-quality MS/MS data is not available due to low abundance or weak signal of the precursor ion, and fragmentation of the precursor in a controlled fashion yields a small number of detectable daughter ions rather than comprehensive but weak (or undetectable) daughter ions.

5. *De novo* peptide sequencing

Tandem MS is capable of capturing sufficient information to determine a proteolytic or other peptide sequence, sometimes in its entirety, without recourse to a database search. This is usually referred to as *de novo* peptide (or even, protein) sequencing. This is of particular use where the species of interest does not have a completed genome sequence, or the peptides/proteins in question are not represented in the known databases. Although fewer computational approaches have been developed compared to those that attempt to "look up" the peptide from a known database, several tools exist (Taylor and Johnson, 1997; Dancik *et al*., 1999; Chen *et al*., 2001), as reviewed recently by Standing (2003). The LUTEFISK algorithm uses a sequence graph derived from a filtered set of peaks from a fragment ion spectrum to construct lists of likely matching peptide sequences, which are in turn searched against a protein database with a modified version of the FASTA algorithm (Taylor and Johnson, 2001). The SHERANGA algorithm also uses a spectral graph to denote a peptide sequence, using peaks as nodes and *m/z* and edges or vertices

corresponding to mass differences of a single amino acid (Dancik *et al*., 1999). The algorithm then uses methods from graph theory to find the highest scoring paths through the graph, representing a list of candidate peptide sequences. In this case, no database sequences are required at all, and the algorithm can be considered to be truly *de novo*. A related approach (Chen *et al*., 2001) uses dynamic programming to traverse a spectral graph to find the optimal path and, hence, the most likely peptide sequence.

As mentioned, the *de novo* sequencing problem is particularly acute for cross-species peptide identification when no database sequence is available. Shevchenko and colleagues have extended the *de novo* approach with tandem data, including that from post source decay MS, to identify peptides across species (Shevchenko *et al*., 2001; Liska and Shevchenko, 2003b). Their MS-BLAST algorithm adapts standard BLAST searching to interrogate a protein database using predicted peptide sequences from tandem MS data, adapting the scoring system to cope with isobaric residues and tryptic sites to successfully identify canine peptides without their presence in the database.

6. Data archival

As with all high-throughput and data-intensive technologies, the storage, dissemination, and analysis of the resultant information becomes a real challenge. Fortunately, in proteomics, this has not been forgotten and several proposals have been published or put forward on how best to deal with this aspect of proteomics. The Proteomics Standards Initiative (PSI) (Orchard *et al*., 2004) has been set up and is in the process of defining an International data standard based on different approaches such as PEDRo (Taylor *et al*., 2003), ProDB (Wilke *et al*., 2003), PARIS (Wang *et al*., 2004), RADARS (Field *et al*., 2002), and others (Jones *et al*., 2003). The goal is to define a complete, robust and distributable standard for the publication of proteomics data (*see* Article 61, **Data standardization and the HUPO proteomics standards initiative**, Volume 7). At present, data standards are emerging and latest developments are best viewed on the PSI website (**http://psidev.sourceforge.net/**).

7. Future directions

Although there have been many interesting developments in the field of data mining for proteome proteins and peptides, there is still some way to go in this area. It is still widely acknowledged that current software approaches are still not really robust enough for high-throughput studies, and many peptide MS/MS spectra still do not produce confident identifications. There are many reasons why a tandem MS spectrum does not lead to a confident peptide/protein identification, but improvements can surely be made from a better fundamental understanding of gas-phase peptide ion chemistry. A recent study has applied statistical tests to improve this situation using a large database of peptide spectra, and reexamining the mobile proton theory for peptide fragmentation (Kapp *et al*., 2003). This builds on previous studies from Wysocki, Yates, and colleagues (Breci *et al*., 2003) who

have found similar trends in fragmentation pathways in tryptic peptides promoted by the presence of particular amino acid combinations in the amino acid sequence. Approaches such as these promise to improve peptide MS/MS interpretation and identification algorithms when this information can be incorporated.

Alternatively, developments in mass spectrometry itself could outstrip the bioinformatics, using "top-down" approaches available from Electron Capture Dissociation (ECD) and FT-ICR instruments (*see* Article 5, **FT-ICR**, Volume 5). Here, the protein ion itself is introduced into the mass spectrometer and fragmentation is induced via ECD, yielding peptide ions series from which sequence information can be inferred (Horn *et al*., 2000). Although this technique promises much, it is still far away from the routine analyses available to most labs or the high-throughput required for a comprehensive proteome map required by some labs.

Related articles

Article 13, **Multidimensional liquid chromatography tandem mass spectrometry for biological discovery**, Volume 5; Article 16, **Improvement of sequence coverage in peptide mass fingerprinting**, Volume 5; Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5

Acknowledgments

We thank Jennifer Lynch for useful comments on the manuscript and all authors acknowledge support from the BBSRC.

References

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Brancia FL, Butt A, Beynon RJ, Hubbard SJ, Gaskell SJ and Oliver SG (2001) A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis*, **22**, 552–559.

Breci LA, Tabb DL, Yates JR and Wysocki VH (2003) Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. *Analytical Chemistry*, **75**, 1963–1971.

Chamrad DC, Koerting G, Gobom J, Thiele H, Klose J, Meyer HE and Blüggel M (2003) Interpretation of mass spectrometry data for high-throughput proteomics. *Analytical and Bioanalytical Chemistry*, **376**, 1014–1022.

Chamrad DC, Körting G, Stühler K, Meyer HE, Klose J and Blüggel M (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, **4**, 619–628.

Chen T, Kao MY, Tepel M, Rush J and Church GM (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **8**, 325–337.

Choudhary JS, Blackstock WP, Creasy DM and Cottrell JS (2001) Matching peptide mass spectra to EST and genomic DNA databases. *Trends in Biotechnology*, **19**, S17–S22.

Clauser KR, Baker P and Burlingame AL (1999) Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **71**, 2871–2882.

Cordwell SJ, Wasinger VC, CerpaPoljak A, Duncan MW and HumpherySmith I (1997) Conserved motifs as the basis for recognition of homologous proteins across species boundaries using peptide-mass fingerprinting. *Journal of Mass Spectrometry*, **32**, 370–378.

Craig R and Beavis RC (2003) A method for reducing the time required to match protein sequences in tandem mass spectra. *Rapid Communications in Mass Spectrometry*, **17**, 2310–2316.

Creasy DM and Cottrell JS (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, **2**, 1426–1434.

Dancik V, Addona TA, Clauser KR, Vath JE and Pevzner PA (1999) *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, **6**, 327–342.

Eng JK, McCormack AL and Yates JR III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.

Eriksson J, Chait BT and Fenyo D (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Analytical Chemistry*, **72**, 999–1005.

Fenyö D, Qin J and Chait BT (1998) Protein identification using mass spectrometric information. *Electrophoresis*, **19**, 998–1005.

Field HL, Fenyo D and Beavis RC (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, **2**, 36–47.

Giddings MC, Shah AA, Gesteland R and Moore B (2003) Genome-based peptide fingerprint scanning. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 20–25.

Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, *et al*. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, **20**, 3535–3550.

Horn DM, Zubarev RA and McLafferty FW (2000) Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 10313–10317.

Jones A, Wastling J and Hunt E (2003) Proposal for a standard representation of two-dimensional gel electrophoresis data. *Comparative and Functional Genomics*, **4**, 492–501.

Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RAJ, Speed TP and Simpson RJ (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry*, **75**, 6251–6264.

Karty JA, Ireland MME, Brun YV and Reilly JP (2002) Defining absolute confidence limits in the identification of Caulobacter proteins by peptide mass mapping. *Journal of Proteome Research*, **1**, 325–335.

Keller A, Nesvizhskii AI, Kolker E and Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, **74**, 5383–5392.

Kwon K-H, Kim M, Kim JY, Kim KW, Kim SI, Park YM and Yoo JS (2003) Efficiency improvement of peptide identification for an organism without a complete genome sequence, using expressed sequence tag database and tandem mass spectral data. *Proteomics*, **3**, 2305–2309.

Lester PJ and Hubbard SJ (2002) Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics. *Proteomics*, **2**, 1392–1405.

Lisacek FC, Traini MD, Sexton D, Harry JL and Wilkins MR (2001) Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. *Proteomics*, **1**, 186–193.

Liska AJ and Shevchenko A (2003a) Combining mass spectrometry with database interrogation strategies in proteomics. *TrAC Trends in Analytical Chemistry*, **22**, 291–298.

Liska AJ and Shevchenko A (2003b) Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, **3**, 19–28.

Mann M, Hojrup P and Roepstorff P (1993) Use of mass-spectrometric molecular-weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, **22**, 338–345.

Mann M and Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399.

Nesvizhskii AI, Keller A, Kolker E and Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, **75**, 4646–4658.

Orchard S, Zhu W, Julian RK Jr, Hermjakob H and Apweiler R (2004) Further advances in the development of a data interchange standard for proteomics data. *Proteomics*, **3**, 2065–2066.

Pappin DJC, Højrup P and Bleasby AJ (1993) Rapid identification of proteins by peptide-mass finger printing. *Current Biology*, **3**, 327–332.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Perkins DN, Pappin DJC, Creasy DM and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Sadygov RG and Yates JR III (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical Chemistry*, **75**, 3792–3798.

Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W and Standing KG (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical Chemistry*, **73**, 1917–1926.

Sidhu KS, Sangvanich P, Brancia FL, Sullivan AG, Gaskell SJ, Wolkenhauer O, Oliver SG and Hubbard SJ (2001) Bioinformatic assessment of mass spectrometric chemical derivatisation techniques for proteome database searching. *Proteomics*, **1**, 1368–1377.

Standing KG (2003) Peptide and protein *de novo* sequencing by mass spectrometry. *Current Opinion in Structural Biology*, **13**, 595–601.

Sullivan AG, Brancia FL, Tyldesley R, Bateman R, Sidhu KS, Hubbard SJ, Oliver SG and Gaskell SJ (2001) The exploitation of selective cleavage of singly protonated peptide ions adjacent to aspartic residues using a quadrupole orthogonal time-of-flight mass spectrometer equipped with a matrix-assisted laser desorption/ionization source. *International Journal of Mass Spectrometry*, **210/211**, 665–676.

Taylor JA and Johnson RS (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, **11**, 1067–1075.

Taylor JA and Johnson RS (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, **73**, 2594–2604.

Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin ZK, Deutsch EW, Selway L, Walker J, Riba-Garcia I, *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**, 247–254.

Verrills NM, Harry JH, Walsh BJ, Hains PG and Robinson ES (2000) Cross-matching marsupial proteins with eutherian mammal databases: Proteome analysis of cells from UV-induced skin tumours of an opossum (*Monodelphis domestica*). *Electrophoresis*, **21**, 3810–3822.

Wang JH, Caron C, Mistou MY, Gitton C and Trubuil A (2004) PARIS: a proteomic analysis and resources indexation system. *Bioinformatics*, **20**, 133–135.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Wasinger VC, Urquhart BL and Humphrey-Smith I (1999) Cross-species characterisation of abundantly expressed Ochrobactrum anthropi gene products. *Electrophoresis*, **20**, 2196–2203.

Wilke A, Ruckert C, Bartels D, Dondrup M, Goesmann A, Huser AT, Kespohl S, Linke B, Mahne M, McHardy A, *et al*. (2003) Bioinformatics support for high-throughput proteomics. *Journal of Biotechnology*, **106**, 147–156.

Wilkins MR, Gasteiger E, Wheeler CH, Lindskog I, Sanchez JC, Bairoch A, Appel RD, Dunn MJ and Hochstrasser DF (1998) Multiple parameter cross-species protein identification using MultiIdent - a World-Wide Web accessible tool. *Electrophoresis*, **19**, 3199–3206.

Wilkins MR and Williams KL (1997) Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *Journal of Theoretical Biology*, **186**, 7–15.

Yates JR III, Eng JK, Clauser KR and Burlingame AL (1996) Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *Journal of the American Society for Mass Spectrometry*, **7**, 1089–1098.

Zhang N, Aebersold R and Schwilkowski B (2002) ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.

Zhang WZ and Chait BT (2000) Profound: An expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry*, **72**, 2482–2489.

Short Specialist Review

Low-level analysis of oligonucleotide expression arrays

Xuemin Fang and **Wing Hung Wong**
Harvard School of Public Health, Boston, MA, USA

Cheng Li
Dana-Farber Cancer Institute, Boston, MA, USA

1. Oligonucleotide expression microarrays

Affymetrix GeneChip® oligonucleotide expression microarrays are manufactured through a combination of photolithography and combinatorial chemistry (**http://www.affymetrix.com/technology/manufacturing/index.affx**). GeneChip® technology produces arrays with hundreds of thousands of different probes at an extremely high density, so that researchers could generate high-quality genome-wide data using small sample volumes. The features of an oligonucleotide array are represented by the probe sets. There are usually 8000–20 000 probe sets on a single array, depending on the array type. A probe set has 11–20 probe pairs (usually 16 for U95 arrays and 11 for U133 arrays), and a probe pair consists of one perfect match probe (PM) and a mismatch probe (MM), and both PM and MM probes are oligonucleotides of 25 bp. Each probe set is designed for one target gene; the PM probes have exactly the same complementary sequences with the segments of the cRNA of the target gene, and MM probes are the same as the paired PM probes, except for the middle base (13th), which is replaced by a single-base substitution.

2. Image analysis and feature extraction

In microarray experiments, arrays are hybridized to the prepared solution of sample cRNA, in which each probe hybridizes to the cRNA of its target gene, which is labeled with a fluorescent dye. After hybridization, intensity values are read from a laser scanner. To extract the features from the image intensities, the Affymetrix Microarray Suite software (MAS) first determines the preset feature boundaries and then uses a dynamic gridding algorithm to segment the image and then compute

the feature intensities using a percentile algorithm (Lockhart *et al*., 1996). Schadt *et al*., (2000, 2001) also developed an adaptive image processing algorithm that analyzes the distribution of pixel intensities for a given feature, computes the feature intensities, and estimates the blurring effects. In the presence of the image artifact, the MAS software allows users to manually mask out the problematic features; however, the process could be time-consuming. Schadt *et al*. (2000) proposed a semiautomatic way to mask out the problem regions.

3. Presence calling

An advantage of oligonucleotide array over cDNA array is that with the PM and MM intensities, it is possible to determine whether a gene is expressed in the sample or not. Affymetrix determines the presence of a gene by calculating a set of statistics (Lockhart *et al*., 1996) to estimate the parameters of a presence or absence decision tree. Schadt *et al*. (2000) determine the presence of a gene by testing for each pair of PM and MM intensities, if PM and MM are equal in distribution; and if the ratio of PM/MM is equal to the MM/PM in distribution. Liu *et al*. (2002) present a detection call algorithm based on Wilcoxon's signed-rank test by calculating the discrimination scores.

4. Normalization

When multiple arrays are used in an experiment, array intensities may not be directly comparable. One batch of arrays may be brighter or dimmer than another batch; arrays processed by one technician may be brighter or dimmer than that processed by another technician. This raises the issue of how to normalize the arrays so that the differences due to the technical factors are eliminated while the biologically differential expression patterns are still retained. Normalization can have a profound influence in determining differentially expressed genes and other subsequent high-level analysis (Hoffmann, 2002).

The MAS suite uses a simple scaling of the expression indexes (see Section 6) to normalize between the arrays, by making the trimmed average expression of each array to be a prespecified value, for example, 250. Other methods normalize at the probe level: Li and Wong (2001b) normalize arrays to a baseline array according to a running-median, smoothing curves fitted through a set of iteratively selected rank-invariant probes; Bolstad *et al*. (2003) normalize arrays by making the distribution (or quantiles) of probe intensities for each array in a set of arrays the same. Åstrand (2003) first imposes an orthogonal change of basis of the log intensity matrix, then fits a curve between the two arrays and transforms the array to be normalized, changes back to the original basis and scale according to the curve.

These normalization methods assume that the arrays to be normalized together do not have a large number of differentially expressed genes. When normalizing arrays of different sample or tissue groups, one solution is to normalize within groups separately so that the biologically meaningful differential expression patterns are preserved after normalization; Another solution is to normalize according to a set

of housekeeping genes (genes known to have a constant expression regardless of tissue types) (Hill *et al*., 2001; Baugh *et al*., 2001); also, in the newer Affymetrix array types, there are 100 control housekeeping probe sets for the purpose of normalization or quality assessment.

5. Background adjustment

Array intensity background can come from physical and biological sources (Naef *et al*., 2001): physically, the light reflection of the substrate, or the photo detector dark current can affect the array intensities. Biologically, the nonspecific cross-hybridization of the probes to the nontarget genes also adds up to the observed probe intensities and thus obscure the real intensity signal generated by the hybridization of the probes to their target genes.

How to estimate the background, especially the biological background? Is the background an additive effect? Is it probe-specific or the same for the probe set or even the whole array? How do we take advantage of the MM probe intensities? MM probes are intended to serve as the nonspecific hybridization background control of the probe pair, by assuming that the specific binding of MM probe to the target gene is greatly reduced because of the middle-base substitution, and the observed MM intensity is mostly due to the nonspecific cross-hybridization, whose magnitude does not differ significantly from the nonspecific cross-hybridization of its paired PM probe. Then if we subtract the MM intensity from the PM intensity, the remaining part should be due to the real hybridization intensity signal of the PM probe to its target gene. By utilizing this quantity, we hope to derive an expression index that reflects the target gene's expression level in the sample solution. This is the basic idea that motivates the Affymetrix MAS 4.0 algorithm and the dChip PM/MM difference model (Li and Wong, 2001a).

In reality both PM and MM, probes hybridize to their target gene as well as to various nontarget genes, which greatly complicates the situation. Researchers have made efforts to estimate the background: some estimate the distribution of the background by estimating the mean and stand deviation of the MM intensities (Naef *et al*., 2001; Irizarry *et al*., 2003b); some estimate the biological background by modeling the cross-hybridization parameter (Affymetrix Workshop, 2003). Affymetrix MAS 5.0 software uses MM to construct a robust estimate of the probe-specific background (Hubbell *et al*., 2002).

6. Expression index

An expression index is a numeric value to reflect the gene's cRNA concentration based on the observed intensities. The ultimate goal of probe-level analysis of the oligonucleotide array is to extract the expression index from the entangled information we have: PM intensities, MM intensities, background, and cross-hybridization. Most expression index calculation methods estimate and adjust cross-hybridization intensities during the background adjustment step (see Section 5). After the adjustment, the remaining probe intensities are assumed to be purely due

to the hybridization of the probes to their target gene. MAS 4.0 or 5.0 estimate the expression index by taking the average or robust average of the PM and MM intensity differences. The Model-based Expression Index (MBEI) (Li and Wong, 2001a) uses a multiplicative model of gene expression index and the probe sensitivity index to fit the background-adjusted intensities with several varieties: LWF (Li Wong Full Model) fits both PM and MM intensities, LWR (Li Wong Reduced Model) fits the PM and MM differences, and PM-only model fits the PM intensities. All of them are implemented in the dChip (Li and Wong, 2003). This model-based approach attaches standard errors to the estimated expression indexes and probe sensitivity indexes for assessing array and probe quality. The Robust Multichip Average (RMA) (Irizarry *et al*., 2003b) fits the log-transformed, background-subtracted PM probe intensities with a linear additive model, including an array effect and a probe effect.

With the recently available probe sequence information, researchers are able to go deeper in index estimation and probe selection. It has been found that the probe intensity depends on the physical properties of the probe and its target gene. (Hekstra *et al*., 2003) use a Langmuir adsorption model to reflect the sequence dependence of the probe hybridization and the chemical saturation, and estimate the gene's absolute mRNA concentration in picomolar. Zhang *et al*. (2003) use the probe sequences to model the binding interaction and measure expression after estimating the nonspecific binding. Held *et al*. (2003) construct a model using hybridization rate equations, calculated free energies and microarray data for known target concentration to estimate the gene concentration. Wu *et al*., 2004 improve array background adjustment with probe sequence information, and use an empirically motivated stochastic model to compliment the physical model for estimating expression index (Wu and Irizarry, 2004).

Mei *et al*. (2003) makes an effort in probe selection by designing a probe response metric with a thermodynamic hybridization model.

7. Comparison of different methods

How to choose one set of normalization, background adjustment, and expression index method to analyze one's data? Cope *et al*. (2004) have developed an assessment tool (AffyComp website, 2003) for Affymetrix probe-level expression measures. Various aspects related to accuracy and precision, and from a biological perspective (James *et al*., 2004), were assessed using two standard datasets: the Affymetrix spike-in dataset and the GeneLogic dilution dataset (can be downloaded from the AffyComp website). This assessment tool is available on the Web and has generated comparison results for a number of submitted methods. Knowing the pros and cons of each method helps to decide which one you may want to use. In a brief comparison of Affymetrix MAS software, dChip, and RMA, Irizarry *et al*. (2003a) pointed out that compared to MAS, both RMA and dChip add a slight bias to the expression index while reducing much more variance. Features of each software or method are also of interest to the users in making the decision. For example, dChip has a friendly user interface for the biologists, while Bioconductor has access to many statistical functions so prototype analysis can be carried out easily. Other

comparisons of the expression indexes are done by some other researchers (Lemon *et al*., 2002; Irizarry *et al*., 2003a; Chu *et al*., 2004; James *et al*., 2004).

8. Software

Several software packages are available for the low-level data analysis of oligonucleotide expression arrays. MAS software (renamed to GCOS in 2003) is designed for analysis ranging from processing raw array images to calculating the gene expression indexes; dChip (Li and Wong, 2003) implements the invariant set and MBEI methods, as well as several high-level analysis functions such as hierarchical clustering and searching for functional and chromosomal enrichment in a list of genes; Bioconductor project (**www.bioconductor.org**) has developed many microarray data analysis packages (e.g., Affy package; Gautier *et al*., 2004) based on the R-statistical environment (Ihaka and Gentleman, 1996), and it allows users to normalize, calculate gene expression indexes using various methods, and do high-level analysis such as using graphs to represent pathways and linking genes to on-line databases by HTML page.

9. Prospects

There are researchers actively working on low-level microarray data analysis on the above-mentioned issues, such as improving background estimation and better utilization of probe sequence information. These were discussed in the two major workshops on low-level microarray analysis (GeneLogic workshop, 2001; Affymetrix Workshop, 2003). In the near future, generating large-scale benchmark datasets such as arrays hybridizing to multiple tissue types or hybridizing to samples with more genes spike-in at known concentrations will be of tremendous help to many research topics, including estimating the cross-hybridization background and developing model for index calculation and searching for more housekeeping genes. We also recommend not going too quickly in reducing the probe set size, since having more probes per probe set helps researchers to better understand the probe behaviors and develop more mature low-level analysis methods.

Further reading

Affymetrix Incorporated (1999) GeneChip ® Expression Analysis Algorithm Tutorial.

Affymetrix Manual (2001) *Affymetrix Microarray Suite User Guide Version 5.0*, Santa Clara.

Affymetrix manufacturing website **http://www.affymetrix.com/technology/manufacturing/index.affx**.

Affymetrix software reference (2003) **http://www.affymetrix.com/products/software/index.affx**.

Bioconductor website: **www.bioconductor.org**.

Hill A, Hunter C, Tsung B, Tucker-Kellogg G and Brown E (2000) Genomic analysis of gene expression. in *C. elegans. Science*, **290**, 809–812.

Holder D, Raubertas RF, Pikounis BV, Svetnik V and Soper K (2001) Statistical analysis of high density oligonucleotide arrays: a SAFER approach. *Proceedings of the ASA Annual Meeting*, Atlanta, 2001.

References

AffyComp website (2003) **http://affycomp.biostat.jhsph.edu/**.

Affymetrix Workshop (2003) *The 2003 Affymetrix GeneChip Microarray Low-Level Workshop*, Berkeley, 7-8 August 2003, **http://www.affymetrix.com/corporate/events/seminar/microarray_workshop.affx**.

Åstrand M (2003) Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology*, **10**(1), 95–102.

Baugh L, Hill A, Brown E and Hunter C (2001) Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Research*, **29**, 1–9.

Bolstad BM, Irizarry RA, Astrand M and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**(2), 185–193.

Chu T-M, Weir BS and Wolfinger RD (2004) Comparison of Li–Wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays. *Bioinformatics*, **20**, 500–506.

Cope LM, Irizarry RA, Jaffee H, Wu Z and Speed TP (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**(3), 323–331.

Gautier L, Cope LM, Bolstad BM and Irizarry RA (2004) Affy – a package for the analysis of Afftmetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3), 307–315.

GeneLogic workshop (2001) *Genelogic Workshop on Low Level Analysis of Affymetrix GeneChip® Data*, Bethesda, 19 November 2001, **http://stat-www.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogi2001.html**.

Hekstra D, Taussig AR, Magnasco M and Naef F (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*, **31**(7), 1962–1968.

Held GA, Grinstein G and Tu Y (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 7575–7580.

Hill A, Brown E, Whitleyn M, Tucker-Kellogg G, Hunter C and Slonim D (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked crna controls. *Genome Biology*, **2**.

Hoffmann R, Seidl T and Dugas M (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology*, **3**, research0033.1–research0033.11.

Hubbell E, Liu WM and Mei R (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**(12), 1585–1592.

Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**.

Irizarry R, Hobbs FCB, Beaxer-Barclay Y, Antonellis K, Scherf U and Speed T (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

James AC, Veitch JG, Zareh AR and Triche T (2004) Sensitivity and specificity of five abundance estimators for high-density oligonucleotide microarrays. *Bioinformatics*, **20**(7), 1060–1065.

Lemon W, Palatini J, Krahe R and Wright F (2002) Theoretical and empirical comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**(11), 1470–1476.

Li C and Wong W (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31–36.

Li C and Wong W (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**(8), research0032.1-0032.11.

Li C and Wong WH (2003) DNA-Chip analyzer (dChip). In *The Analysis of Gene Expression Data: Methods and Software*, Parmigiani G, Garrett ES, Irizarry R and Zeger SL (Eds.), Springer.

Lipshutz RJ, Fodor SP, Gingeras TR and Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**(Suppl 1), 20–24.

Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, *et al*. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**(12), 1593–1599.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675–1680.

Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T, *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(20), 11237–11242.

Naef F, Lim DA, Patil N and Magnasco MO (2001) From features to expression: high density oligonucleotide array analysis revisited. *Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data*, 24-26 October, DIMACS Center, Rutgers University, Piscataway, NJ.

Schadt EE, Li C, Ellis B and Wong WH (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, Suppl 37, 120–125.

Schadt EE, Li C, Su C and Wong WH (2000) Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, **80**, 192–202.

Wu Z and Irizarry RA (2004) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Proceedings of RECOMB 2004*.

Wu Z, Irizarry RA, Gentleman R, Martinez Murillo F and Spencer F (2004) *A model based background adjustment for oligonucleotide expression arrays*. Technical Report, to appear in JASA.

Zhang L, Miles MF and Aldape KD (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, **21**(7), 818–821.

Short Specialist Review

CGH data analysis

Adam A. Margolin
Columbia University, New York, NY, USA

Adam A. Margolin, Joel Greshock and Barbara L. Weber
University of Pennsylvania, Philadelphia, PA, USA

1. Introduction

Genomic abnormalities, such as amplification or deletion of segments of DNA, are an important feature of cancer and other genetic disorders. Comparative genomic hybridization (CGH) has been widely employed since its introduction in 1992 to identify such chromosomal alterations (Kallioniemi *et al*., 1992). Using this technology, differentially labeled test DNA and reference DNA are hybridized simultaneously to normal chromosome spreads, the hybridization is detected with two different fluorochromes, and regions of gain or loss of DNA sequences are seen as changes in the intensity ratio of the two fluorochromes (Kallioniemi *et al*., 1992). With advances in array-based technologies, CGH has been implemented in a microarray format (Pinkel *et al*., 1998), enabling much higher resolution mapping of chromosomal alterations and providing more quantifiable results amenable to statistical and high-throughput analysis. Gene expression array platforms that have been adapted for array comparative genomic hybridization (aCGH) include two-channel arrays using genomic DNA from bacterial artificial chromosomes (BACs) (Pinkel *et al*., 1998) or cDNA clones (Greshock *et al*., 2004; Pollack *et al*., 1999), as well as single-channel oligonucleotide-based arrays (Bignell *et al*., 2004; Lucito *et al*., 2003). Although these aCGH assays share mechanical similarities with assays designed to measure gene expression, analysis and interpretation of aCGH data differs for several important reasons. First, expression arrays measure a nearly continuous spectrum of gene expression levels, whereas copy number arrays must quantify discrete changes in genomic copy number. Second, expression array probes can be treated as independent objects, each measuring the expression level of a single gene or mRNA species, whereas copy number array probes must be interpreted within the context of their genomic location and the surrounding probes, and can be used to define the overall pattern of chromosomal abnormalities throughout the genome. For copy number arrays that do not have full tiling

resolution genomic coverage, this data can be used to interpolate copy number changes for genomic regions not directly covered by a probe. Finally, normal diploid DNA provides a constant baseline value for comparison in CGH arrays. This review will discuss how to account for these differences to adapt well-studied techniques for analysis of gene expression array data to the analysis of aCGH data.

2. Probe value analysis

The test statistic output by most aCGH scanning and preprocessing software is typically the normalized spot fluorescence intensity ratio of a test-to-reference sample for two-channel arrays, where a ratio of 1.0 represents a diploid state in the test sample. For single-channel arrays, this statistic is typically the normalized spot fluorescence intensity. Ratio values are generally log-transformed, allowing amplification and deletion events to be treated symmetrically (although it should be noted that deletion events generally do not give zero signal due to low levels of "normal" contamination in patient tumor samples). To compensate for any potential dye-bias in two-channel arrays, replicate "dye swap" experiments can be performed in which the dyes used to label the test and reference sample are applied in both combinations. In this context, the geometric mean of the log values of each hybridization can be used to assign a single measure of representation to each probe in a dye swap experiment.

After obtaining measures for each probe in an experiment, this data must be transformed into interpretable statistics to most accurately determine the probability that a genomic region is amplified or deleted in a test sample. Transforming data into standard statistics also allows results from different assays and different experiments to be compared. The most widely employed and simplest approach to this problem is to use hard discretization of the data based on defined thresholds of the logarithm of the probe values to classify genomic regions as amplified, deleted, or normal, with possible additional thresholds representing homozygous deletions or multiple copy amplifications. For example, a genomic region covered by a probe may be considered deleted if the test-to-reference $\log_2$ ratio value is less then -0.25, and amplified if this value is greater than 0.25 (Veltman *et al*., 2003) (Figure 1). Optimally, a proper threshold can be determined by examining the distribution of ratios for regions of known aberrations in control assays.

A more sophisticated approach would account for systematic variation in probe values that can primarily occur from two sources (1) between-assay variation and (2) between-probe variation. First, differences in laboratory experimental conditions, variation in hybridization efficiency for different slides, the presence of contaminating normal tissue, and other sources of variation can lead to a distribution of measured representation across patients, even for regions that have identical copy numbers. One approach to address this variability is to adjust probe thresholds based on the overall distribution of probe values for a particular experiment (Snijders *et al*., 2003), although this can be problematic when analyzing highly aneuploid samples that may not produce accurate baseline distributions (Davidson *et al*., 2000). A mixture-model approach can also be used in which a mixture of three Gaussian distributions, corresponding to the normal, gained, and deleted

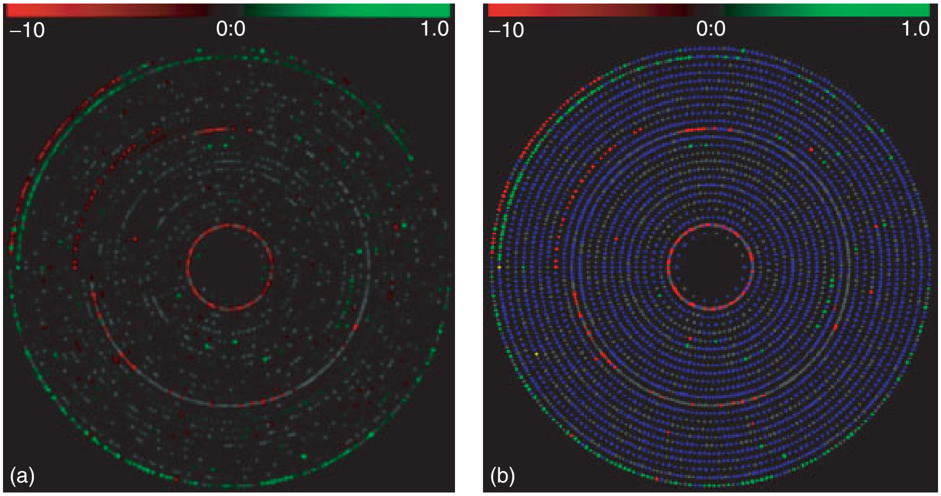


Figure 1 Circular display of the entire genome for a single aCGH experiment performed on a melanoma cell line. Chromosomes are displayed as concentric circles, with chromosome 1 presented as the outermost circle, followed by chromosome 2, and the Y chromosome presented as the innermost. Each dot represents an arrayed BAC clone probe, with the p-arm telomere of each chromosome beginning along the horizontal on the left of the display, and subsequent probes arranged clockwise by their linear order on the chromosome. (a) Probes are colored by a gradient based on the base 2 logarithm of the ratio of measured fluorescence intensities of the test sample relative to the unaltered reference. Log_2 ratios of 1 (representing likely amplifications) are displayed in bright green, and $\log_2$ ratios of -1 (representing likely deletions) are displayed in bright red, with gradients of color representing intermediate levels. Notice that a large portion of chromosome 1 appears to be amplified. The bright red circle toward the center represents the X chromosome, which appears to be "deleted" in this male-to-female hybridization. (b) Discrete determinations are made for the DNA copy number content represented by each probe using $\log_2$ threshold ratios of -0.25 and 0.25 for deletions and amplifications. Probes representing amplified and deleted DNA content are displayed as green and red dots, with probes representing normal (diploid) DNA content displayed as blue dots. Figures were produced using the CGHAnalyzer software package (Margolin *et al*., 2005)

portions of the genome, is fitted to a histogram of the probe value ratios for each experiment. This approach was employed by Hackett *et al*. (2003) using a hybrid least-squares/maximum likelihood method. Another method, employed by Paris *et al*. (2004), determined sample-specific thresholds by using a discrete-time hidden Markov model (HMM) to segment probes on individual chromosomes into the states corresponding to underlying copy numbers.

Between-probe variation arises owing to variation in the probe sequence composition, including the presence of repetitive sequences, base composition, and differences in secondary structure, all of which can affect hybridization kinetics and which have been shown to alter the distribution of hybridization intensity values for different probes throughout experiments (Snijders *et al*., 2001). To address this problem, a reference data set of normal–normal hybridizations can be used to estimate the Gaussian distribution of signal intensity for each probe representing the distribution of values for cases where the probe covers a region of diploid DNA. The mean value of all replicate spots for a probe in a single experiment can be compared to the distribution for that probe calculated from the normal–normal dataset to assign a z-score representing the probability that the segment of DNA

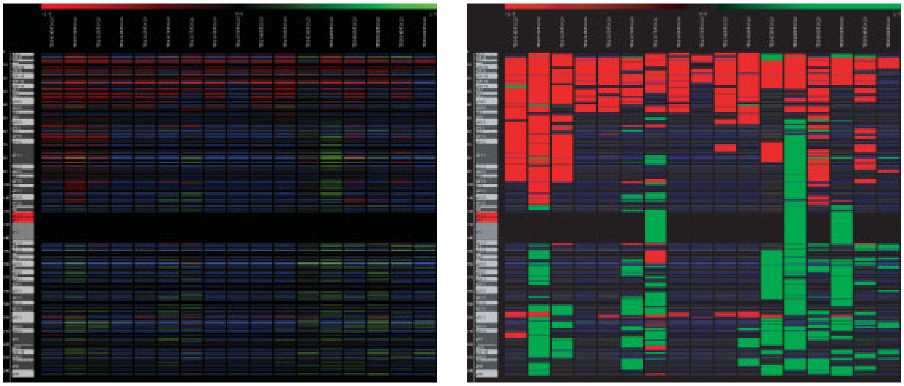
covered by the probe is amplified or deleted in the test sample. Moore *et al*. (1997) employed a similar approach for analysis of traditional CGH experiments. However, this group compute a modified t-value, as described by Kendel and Stuart (1967), by comparing the distribution of values from replicate experiments to the distribution calculated from the normal–normal dataset.

Each of the methods described above attempts to address a different source of variation in probe value measurement. The most comprehensive method of transforming probe values to probabilities of chromosomal alterations should consider aspects of all of these models. More research is needed to formally define the limitations and capabilities of each method. Fortunately, analyzing X chromosome probe values for experiments that use DNA from a male as the test sample and DNA from a female as the reference sample provide a controlled setting for investigating each method's ability to detect heterozygous deletions.

3. Profile generation

In expression arrays, which are designed to measure the expression level of distinct mRNA species, the probes for each gene can be treated as independent objects. However, in using CGH arrays, which are designed to measure the DNA copy number profile across the entire genome, the spatial distribution of the probes across the genome and the correlations between signals from nearby probes provide important information for interpreting the results. Recently, tiling arrays that use a set of overlapping probes covering the genome have been used for comprehensive estimation of copy number variation (Ishkanian *et al*., 2004). However, most CGH arrays currently in use only directly measure copy numbers for a small percentage of the genome. These arrays are generally designed such that probes are evenly distributed across the genome, and copy number is estimated across the genome by interpolating the measurements for flanking regions. Defining a copy number profile across the entire genome, and not just at the location of the probes, allows comparison between arrays performed on different platforms. While arrays with nonoverlapping probes are the most commonly used, it should be noted that they can miss microamplifications and microdeletions whose boundaries lie completely between arrayed probes.

The most conservative method of interpolation for the region of DNA between arrayed probes is to assume that any copy number change measured by a probe extends on either side of the probe until the next probe of a different measured copy number (Figure 2). The authors tend to favor this conservative approach, although it overestimates the altered regions, because there is no way to accurately determine the break-point between two probes of differing copy number, so maximizing potential regions of alteration will not exclude potentially important genes. Another potential problem of this approach is identifying single, nonaberrant probes that break up otherwise large, contiguous regions of alteration (e.g., an entire chromosome arm). Although such noncontiguous regions of alteration certainly occur in real samples, a probe value that lies just within the alteration threshold is more likely to represent a true alteration given supporting evidence from nearby probes.



(a) (b)

Figure 2 Display of chromosome 1 for 17 CGH microarrays performed on neuroblastoma cells lines (Mosse *et al*., 2005). (a) Genomic coordinates in megabases and a chromosome ideogram are displayed on the left side of the graph, with arrayed BAC clone probes displayed as horizontal bars at positions corresponding to their genomic location. A discrete copy number determination is made for each probe using $\log_2$ threshold ratios of -0.25 and 0.25 for deletions and amplifications. Probes determined to represent regions of amplified or deleted DNA content are displayed in green and red respectively, with probes determined to represent normal diploid DNA displayed in blue. (b) DNA content for genomic regions not directly represented by arrayed probes is interpolated using the "flanking regions" approach to maximize potential alterations, producing an estimated whole-genome copy number profile. Genomic regions that are determined to be amplified or deleted in each sample are represented by green and red boxes in the graph

Therefore, thresholds should be relaxed for probes that lie within contiguous alteration regions. These thresholds could be adjusted by using a simple weighted average of neighboring probe values, or by employing a more sophisticated kernel-based function of neighboring probe values. For example, Snijders *et al*., (2003) have employed an unsupervised HMM to identify groups of probes with the same underlying copy number. A more recent approach (Myers *et al*., 2004) employs an EM-based edge placement algorithm to statistically optimize predicted alteration start and end locations, combined with a window significance test to determine the statistical significance of these predictions. We note that data smoothing methods that adjust every probe value on the basis of neighboring probe values reduce the ability to detect smaller, more localized regions of alteration, which have been shown to be of consequence in tumor samples (Reiter and Brodeur, 1996). Therefore, a smoothing method that only adjusts values for probes that lie between those that represent alterations may be more appropriate for aCGH analysis. More research is needed to evaluate the efficacy of various statistical methods that can be used to infer DNA copy number for genomic regions that are not directly measured.

4. Data visualization

We describe methods of visualizing aCGH data at three levels of granularity which, when used together, can provide a comprehensive representation of the chromosomal alteration profile across multiple experiments.

The most widely employed visualization method for viewing single experiments is a scatter plot of genomic location versus probe ratio values (Snijders *et al*., 2003). Probes representing normal diploid DNA should form a baseline of ratio values, and regions of chromosomal alterations can be identified visually as deviations from this baseline. For fine-resolution analysis of an experiment, this view must be displayed for a single chromosome, as displaying all chromosomes on a single graph often becomes cluttered for high-density arrays and this view does not accurately display boundaries between chromosomes. Plots of single chromosomes are often displayed next to a chromosomal ideogram, providing an easily interpretable view of chromosomal location (Chi *et al*., 2004).

A circular display can be used to represent data from the entire genome of a single sample in a compact form (Figure 1). This display consists of a number of concentric circles, each representing a chromosome, with chromosome 1 (the largest) presented as the outermost circle. Each circle is composed of a number of ordered spots, each representing an arrayed probe, and each colored on the basis of the probe intensity ratio in the given experiment. This view facilitates identification of large-scale abnormalities and the overall aneuploidy of a sample.

The typical view for visualization of multiple gene expression array experiments is a matrix or heap graph displaying each probe as a row and each sample as a column, with each cell colored by the probe value for the corresponding experiment. A modified version of this is typical for copy number visualization and can be applied to aCGH data to explicitly display probes by their chromosomal location. This display presents a chromosome ideogram on the left side of the screen with experiment arranged as columns, and a box for each probe colored by the probe's ratio value, beginning, and ending at the chromosomal location of the probe with reference to the chromosome ideogram (Figure 2). Alternatively, a display typically used for metaphase CGH, which displays vertical bars representing chromosomal gains on one side of a chromosome ideogram and vertical bars representing losses on the other, can be adapted for the display of aCGH data after interpolating a whole genome alteration profile.

5. Higher-order analyses

Many of the higher-order analysis techniques that are widely employed for expression array analysis have been used to analyze aCGH data. Techniques employing traditional t-tests have been used to distinguish tumor classes using expression array data (Olshen and Jain, 2002), and have recently been applied to aCGH data (Veltman *et al*., 2003). Several groups have employed statistical correlation measures and hierarchical clustering (Hackett *et al*., 2003; Jain *et al*., 2001; Wilhelm *et al*., 2002) on arrayed probes to attempt to identify cooperating loci, or to classify tumors on the basis of their DNA copy number profile. Although these studies demonstrate the potential of using aCGH data for tumor classification, overweighting large aberrations or genomic regions with higher densities of arrayed probes may produce misleading results. One method of addressing this problem, as employed by Veltman *et al*. (2003), is to define a priori a group of known oncogenes or tumor suppressor genes for a given tumor type and cluster based on

the DNA copy number status of only these genes. Of course, this method is limited in that it reduces the size of the dataset and ignores potential novel aberrations that may be important for understanding the disease. Fritz *et al*. (2002) demonstrated that machine-learning approaches could also be applied effectively to aCGH data by showing that support vector machines (SVM) were able to reliably distinguish between pleomorphic liposarcomas and dedifferentiated liposarcomas, and that an SVM applied to aCGH data outperformed an SVM applied to expression array data in class differentiation.

6. Conclusion

Determination of chromosomal alterations associated with cancer and other genetic diseases is an important tool for understanding their pathogenesis. aCGH is an emerging technology aimed at answering this question at an unprecedented level of resolution. As aCGH shares many similarities to gene expression arrays, many aCGH analysis methods can be adapted from well-studied and established gene expression array analysis methods. However, analysis of aCGH data presents novel challenges, and as this technology becomes a widely employed tool in many laboratories, additional methods to analyze this data must be developed and improved. Most important will be to develop sound statistical methods for determining the likelihood that aCGH measurements represent chromosomal alterations, for using aCGH data to infer whole-genome copy number profiles, and for adapting higher-order analysis techniques to aCGH data for disease classification and identification of cooperating/associated loci.

References

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, **14**, 287–295.

Chi B, DeLeeuw RJ, Coe BP, MacAulay C and Lam WL (2004) SeeGH–a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5**, 13.

Davidson JM, Gorringe KL, Chin SF, Orsetti B, Besret C, Courtay-Cahen C, Roberts I, Theillet C, Caldas C and Edwards PA (2000) Molecular cytogenetic analysis of breast cancer cell lines. *British Journal of Cancer*, **83**, 1309–1317.

Fritz B, Schubert F, Wrobel G, Schwaenen C, Wessendorf S, Nessling M, Korz C, Rieker RJ, Montgomery K, Kucherlapati R, *et al*. (2002) Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Research*, **62**, 2993–2998.

Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M and Weber BL (2004) 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Research*, **14**, 179–187.

Hackett CS, Hodgson JG, Law ME, Fridlyand J, Osoegawa K, de Jong PJ, Nowak NJ, Pinkel D, Albertson DG, Jain A, *et al*. (2003) Genome-wide array CGH analysis of murine

neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Research*, **63**, 5266–5273.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics*, **36**, 299–303.

Jain AN, Chin K, Borresen-Dale AL, Erikstein BK, Eynstein Lonning P, Kaaresen R and Gray JW (2001) Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7952–7957.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kendel M and Stuart A (1967) *The Advanced Theory of Statistics*, Haffner Publishing Company: New York.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research*, **13**, 2291–2305.

Margolin AA, Greshock J, Naylor TL, Mosse Y, Maris JM, Bignell G, Saeed AI, Quackenbush J and Weber BL (2005). CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*.

Moore DH II, Pallavicini M, Cher ML and Gray JW (1997) A t-statistic for objective interpretation of comparative genomic hybridization (CGH) profiles. *Cytometry*, **28**, 183–190.

Mosse YP, Greshock J, Margolin A, Naylor T, Cole K, Khazi D, Hii G, Winter C, Shahzad S, Asziz MU, *et al*. (2005) High-resolution detection and mapping of genomic DNA alterations in neuroblastoma. *Genes Chromosomes Cancer*, **43**(4), 390–403.

Myers CL, Dunham MJ, Kung SY and Troyanskaya OG (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Olshen AB and Jain AN (2002) Deriving quantitative conclusions from microarray expression data. *Bioinformatics*, **18**, 961–970.

Paris PL, Andaya A, Fridlyand J, Jain AN, Weinberg V, Kowbel D, Brebner JH, Simko J, Watson JE, Volik S, *et al*. (2004) Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Human Molecular Genetics*, **13**(13), 1303–1313.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, **23**, 41–46.

Reiter JL and Brodeur GM (1996) High-resolution mapping of a 130-kb core region of the MYCN amplicon in neuroblastomas. *Genomics*, **32**, 97–103.

Snijders AM, Fridlyand J, Mans DA, Segraves R, Jain AN, Pinkel D and Albertson DG (2003) Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene*, **22**, 4370–4379.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263–264.

Veltman JA, Fridlyand J, Pejavar S, Olshen AB, Korkola JE, DeVries S, Carroll P, Kuo WL, Pinkel D, Albertson D, *et al*. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Research*, **63**, 2872–2880.

Wilhelm M, Veltman JA, Olshen AB, Jain AN, Moore DH, Presti JC Jr, Kovacs G and Waldman FM (2002) Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Research*, **62**, 957–960.

Short Specialist Review

A comparison of existing tools for ontological analysis of gene expression data

Purvesh Khatri and Sorin Draghici
Wayne State University, Detroit, MI, USA

1. Introduction

Microarrays (*see* Article 90, **Microarrays: an overview**, Volume 4, Article 92, **Using oligonucleotide arrays**, Volume 4, and Article 100, **Protein microarrays as an emerging tool for proteomics**, Volume 4) are at the center of a revolution in biotechnology, allowing researchers to simultaneously monitor the expression of tens of thousands of genes. Independent of the platform and the analysis methods used, the result of a microarray experiment is, in most cases, a list of genes found to be differentially expressed between two or more conditions under study. The more recently proposed protein microarrays are also producing a list of differentially regulated proteins that can be mapped to their respective genes. The common challenge faced by the researchers using any of these techniques is to translate such lists of differentially regulated genes into a better understanding of the underlying biological phenomena. A first step in this direction can be the translation of the list of differentially expressed genes into a functional profile able to offer insight into the cellular mechanisms relevant in the given condition. As recently as 2002, an automatic ontological analysis approach has been proposed to help with this task (Khatri *et al*., 2002). Currently, this approach is the de facto standard for the secondary analysis of high-throughput experiments, and a large number of tools have been developed for this purpose.

This paper presents a review of this approach and a comparison of 14 current tools in this area. A detailed analysis of the capabilities of these tools, of the statistical models deployed as well as of their back-end annotation databases (if applicable), is included here in order to help researchers choose the most appropriate tool for a given type of analysis.

2. Ontological analysis – problem description

2.1. Gene ontology

Manually querying available Web-based annotation databases in order to find the functions associated with various genes is inherently slow and error prone.

An alternative approach to manual database queries and literature searches is an automatic analysis performed by a computer. However, an automatic high-level analysis cannot be performed if the functional annotations are stored as free text because of issues such as the imprecision of the terms, synonymy (different terms used for the same concept), and polysemy (same term used for different concepts). A solution to this vocabulary problem was provided by the Gene Ontology (GO) (Ashburner, 1998; Ashburner *et al*., 2000; Ashburner *et al*., 2001). The goal of GO is to provide an organism-independent, controlled vocabulary as knowledge of gene and protein roles in cells is being accumulated (Ashburner *et al*., 2000). GO is organized in three principal categories: molecular functions, biological processes, and cellular components. A gene or a gene product can have one or more molecular functions, can participate in one or more biological processes, and might be associated with one or more cellular components. Currently, all existing tools for ontological analysis use GO.

2.2. *Building functional profiles*

The translation from a list of differentially expressed genes to a functional profile able to offer insight into the cellular mechanisms is a very tedious task if performed manually. Typically, one would take each regulated gene, search various public databases and compile a list with, for instance, the biological processes that the gene is involved in. This task must be performed repeatedly for each gene in order to construct a master list of all biological processes in which at least one gene is involved. Further processing of this list provides a set of those biological processes that are common between several of the regulated genes. It is expected that those biological processes that occur more frequently in this set would be more relevant to the studied condition. The same type of analysis is necessary for molecular functions and cellular components.

The following example will illustrate this ontological analysis approach (Drăghici *et al*., 2003b). Let us consider an array containing 1000 genes used to investigate the effect of a substance X. Using classical statistical and data analysis methods, we decide that 100 of these genes are differentially regulated by substance X. Let us assume that the 100 differentially regulated genes are found to be involved in the following biological processes: 80 of the 100 genes are involved in *positive control of cell proliferation*, 40 in *oncogenesis*, 30 in *mitosis*, and 20 in *glucose transport*. This type of automatic processing is tremendously useful since they save the researcher the inordinate amount of effort required to process each of the 100 genes, compile lists with all biological processes, and then cross-compare those biological processes to determine how many genes are in each process (Khatri *et al*., 2002). In comparison, a manual extraction of this information would literally take several weeks and would be less reliable and less rigorous.

2.3. *Significance analysis*

The large number of genes involved in cell proliferation, oncogenesis, and mitosis in the functional profile above, might suggest substance X is involved in cancer

Table 1 Statistical significance. The number of genes involved in a given biological process can be misleading on its own. In this example, positive control of cell proliferation may appear to be the most important biological process affected since 80 of the 100 differentially regulated genes are involved in it. However, this process loses its significance when placed in the context that 800 of the 1000 genes on the chip are involved in it

Biological process	Genes found	Genes expected	Significance
Positive control of cell proliferation	80	80	Not significant
Oncogenesis	40	40	Not significant
Mitosis	30	10	Significant
Glucose transport	20	5	Highly significant

mechanisms. However, a reasonable question is: what would happen if all genes on the array were involved in cell proliferation? Would the presence of cell proliferation at the top of the list be significant? Clearly, the answer is no. If most or all genes on the array are involved in a certain process, the fact that particular process appears at the top is not significant. In order to correct this, any software implementing this approach should be able to consider the microarray used in the experiment, calculate the expected number of occurrences of each category, and compare them with the observed numbers.

Table 1 shows this comparison for the example above. Now, the interpretation of the functional profile appears to be completely different. There are indeed 80 *cell proliferation* genes but in spite of this being the largest number, we actually expected 80 such genes, and therefore this event is not significant. The same holds true for *oncogenesis*. *Mitosis* starts to be interesting because we expected 10 genes and we observed 30, which is 3 times more than expected. However, the most interesting process is the *glucose transport*. We expected only 5 genes and we observed 20, that is, 4 times more than expected. The emerging picture changes radically: instead of generating the hypothesis that substance X is a potential carcinogen, we may consider the hypothesis that X is correlated with diabetes.

The problem is that an event such as observing 30 genes when we expect 10 can still occur just by chance. A simple enrichment analysis comparing, for instance, the proportion of genes in a certain category between the entire set of genes and the selected set of genes will not be sufficient since almost any proportion can occur with a nonzero probability. It is the magnitude of this probability that will determine whether a specific category is significant or not. Hence, it is important that any ontological analysis software calculate the probability of a certain proportion occurring just by chance.

3. Existing tools for ontological analysis

The ontological analysis approach described above was first proposed in 2001 by Onto-Express (Khatri *et al*., 2002). The statistical approach used to calculate the significance of various categories followed one year later (Drăghici *et al*., 2003b). From 2003 to 2005, 13 other tools have been proposed for this type of analysis and more tools continue to appear everyday. By and large, all these tools use the same approach. The tools discussed here are FatiGO (Al-Shahrour

et al., 2004), GOstat (Beissbarth and Speed, 2004), eGOn Explore Gene Ontology (eGOn, 2005), DAVID (Dennis *et al*., 2003), GeneMerge (Castillo-Davis and Hartl, 2002), FuncAssociate (Berriz *et al*., 2003), CLENCH (Shah and Fedoroff, 2004), GOToolBox (Martin *et al*., 2004b), GOSurfer (Zhong *et al*., 2004), EASE (Hosack *et al*., 2003), OntologyTraverser (Young *et al*., 2004), GoMiner (Zeeberg *et al*., 2003), and GOTree Machine (Zhang *et al*., 2004). Figure 1 provides a time line of the development of these tools. The following represents our best attempt to compare these tools on the basis of the review of the published manuscripts, user manuals, and our experience with each tool. In addition to functional profiling, GO has also been used for many other purposes such as predicting annotations of unknown sequences (Khan *et al*., 2003; Martin *et al*., 2004a) or measuring the semantic functional similarity between genes (**http://194.117.20.201/tools/ssm/**, 2003). Such applications are outside the scope of this paper. In the following, we first describe each tool very briefly and then compare their capabilities and features.

Onto-Express (OE) was the first tool to propose the automated functional profiling through the integration of dbEST (Boguski *et al*., 1993a,b), UniGene (Schuler, 1997), and LocusLink (Pruitt *et al*., 2000) databases in the Onto-Tools database. The current version of OE supports KEGG, Entrez Gene, PubMed, RefSeq, GenBank, and Gene Ontology databases. OE's back-end database is currently being expanded to integrate GenPept, Protein Information Resource (PIR), Protein Data Bank (PDB), Swiss-Prot, TrEMBL, Online Mendelian Inheritance in Mammals (OMIM), HomoloGene, and Eukaryotic Promoter Database (EPD) databases.

At present, OE can process lists of genes specified as GenBank accession IDs, UniGene cluster IDs, LocusLink IDs, Affymetrix probe IDs, gene symbols, or WormBase accession IDs (Figure 2). OE's output is organized in various views that include: (1) a flat view, which is a bar-graph view that only considers the annotations at the most specific level, (2) a tree view, which represents the results in a custom cut through the GO hierarchy, and (3) a synchronized view, which is

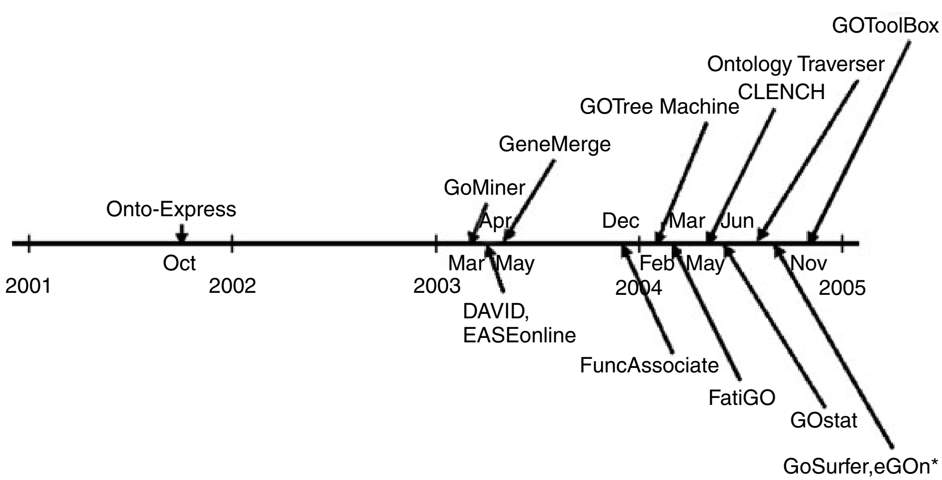


Figure 1 Evolution history of GO-based functional analysis software. The tool marked with a star (eGOn) has not been published in a peer-reviewed journal yet

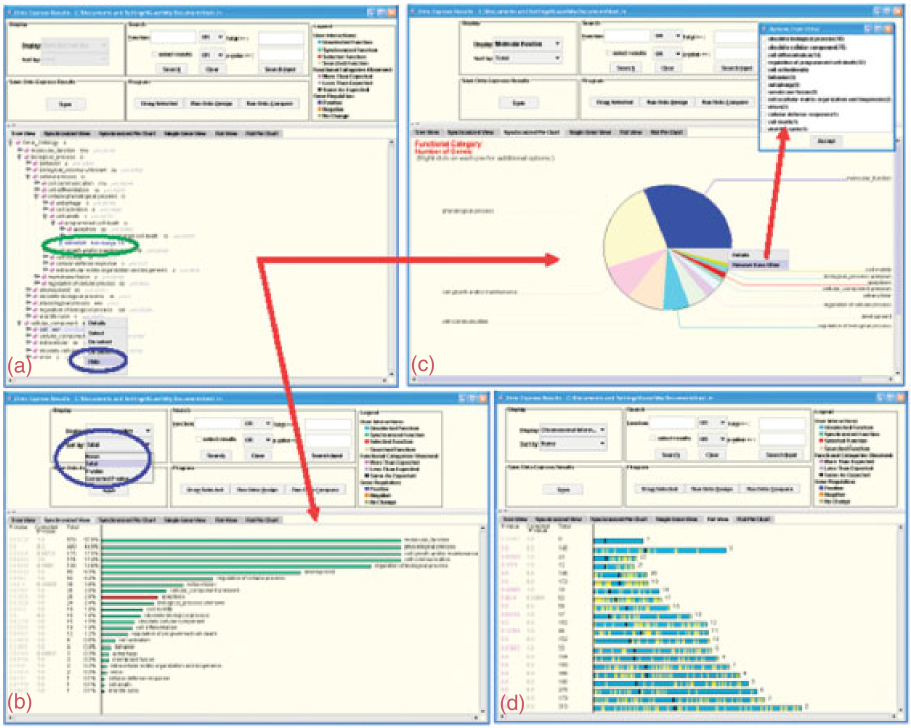


Figure 2 Onto-Express output interface: tree view (a), synchronized bar-chart view (b), synchronized pie-chart view (c), and chromosome view (d). The user can select different levels of abstraction for different branches of the GO hierarchy in the tree view by expanding or collapsing the desired nodes (a). Switching to the synchronized bar-chart view performs the analysis on the GO terms selected in the tree view (b). The tree view also displays the fold change value for an input gene as provided by the user (highlighted in a). In the synchronized view, results can be sorted by name, by total number of genes for each term, by p-value or by corrected p-value (highlighted in b). Categories in the pie chart can be added or removed by the user

a custom bar-graph synchronized with the tree view. The combination of the tree and synchronized views allows the user to select a custom level of abstraction for each branch of the Gene Ontology. This allows the user to ask specific questions about the phenomenon studied. For instance, one could ask whether the apoptotic pathway as a whole is affected without making the distinction between various subprocesses such as positive or negative regulation of apoptosis, induction of apoptosis by various causes (e.g., extracellular signals, intracellular signals, hormones), and so on. In such cases, the user can use OE's tree view to collapse the categories that are unnecessarily specific into more abstract categories. For instance, the approximately 40 subcategories of apoptosis can be collapsed into a single *apoptosis* node. Subsequently, the analysis will be done at this level of abstraction, directly providing the answer to the desired question. A more detailed discussion of this issue and its consequences are included in Section (4).

The Onto-Tools database currently includes 172 arrays from 8 manufacturers that can be selected as the reference. Any other array can be uploaded by the user as a custom reference array. A distinguishing feature of OE is the ability

to sort the entire GO tree by the total number of genes, *p*-values or terms. Another unique feature is a chromosomal location profile that shows the location of the various differentially regulated genes on various chromosomes (however, this feature is currently available only for human genes). Onto-Express also features an application programming interface (API) that allows any other bioinformatic tool to place queries and display the results of the analysis. This capability is currently used by two commercial companies (Insightful and SAS), which integrated OE with their respective software packaged for microarray data analysis: S + ArrayAnalyzer and SAS' Microarray Solution.

GoMiner relies on the Gene Ontology database and can only process gene symbols as input. The same group provides a separate tool, MatchMiner, which converts other types of IDs into gene IDs that are appropriate for use with the GoMiner. This does allow the user to perform the analysis using other types of IDs but adds a separate step to the analysis pipeline since the user has to manually take the results from MatchMiner and submit them to GoMiner.

By default, GoMiner's analysis is not organism specific (Figure 3). This means that if a gene is annotated with function A in one organism and function B in another organism, GoMiner will treat it as if it were annotated with A *and* B, independent of the organism that is actually studied. The user has to choose the

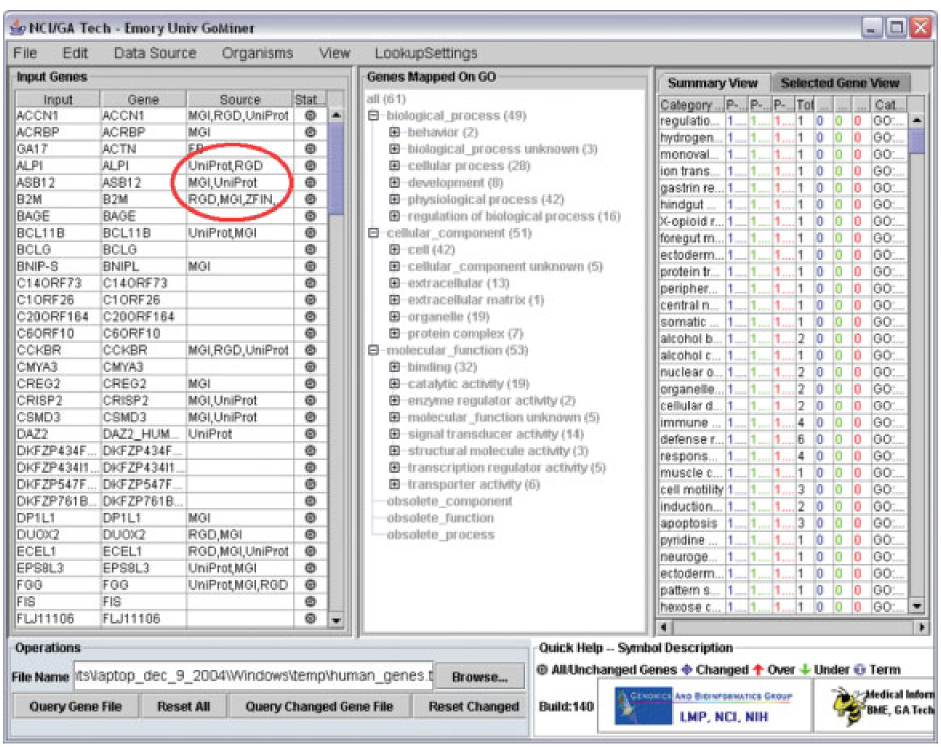


Figure 3 GoMiner output. By default, GoMiner results lack organism specificity. In response to a query with human genes, GoMiner can return annotations from either one of MGD (mouse), RGD (rat), ZFIN (*D. rerio*), and FlyBase (*D. melanogaster*)

desired organism and/or the appropriate data source from the menu to restrict the results to a specific organism. GoMiner is able to represent the results graphically as a directed acyclic graph (DAG). This is achieved using an Adobe scalable vector graphics (SVG) web browser plug-in, which is only compatible with MS Windows. In consequence, GoMiner is platform dependent (Windows only) although it is written in Java and could, in principle, be used on other platforms, as well.

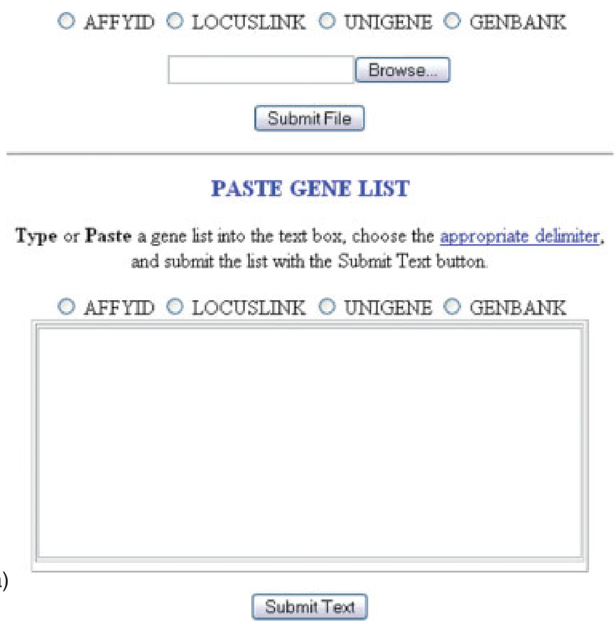
DAVID and EASEonline. DAVID (Database for Annotation, Visualization and Integrated Discovery) provides an HTML-based interface that allows the user to query using a set of genes. The results are presented as a bar chart indicating the percentage of the input genes for each GO term. The results are very simple, do not consider the GO hierarchical structure, and do not calculate any p-value to characterize the statistical significance of the observed results. On the plus side, DAVID does allow the user to specify a depth of analysis, which can be used to perform a horizontal cut through the GO hierarchy.

EASE (Expression Analysis Systematic Explorer) is a companion to DAVID. EASE allows the user to submit Affymetrix IDs, UniGene cluster ID, LocusLink ID, or GenBank accession IDs and calculates the p-values of various categories using Fisher's exact test. As shown in Figure 4, EASE does not consider or show the ontological categories in the context of the hierarchical structure of GO. EASE currently supports 27 Affymetrix arrays as possible reference arrays.

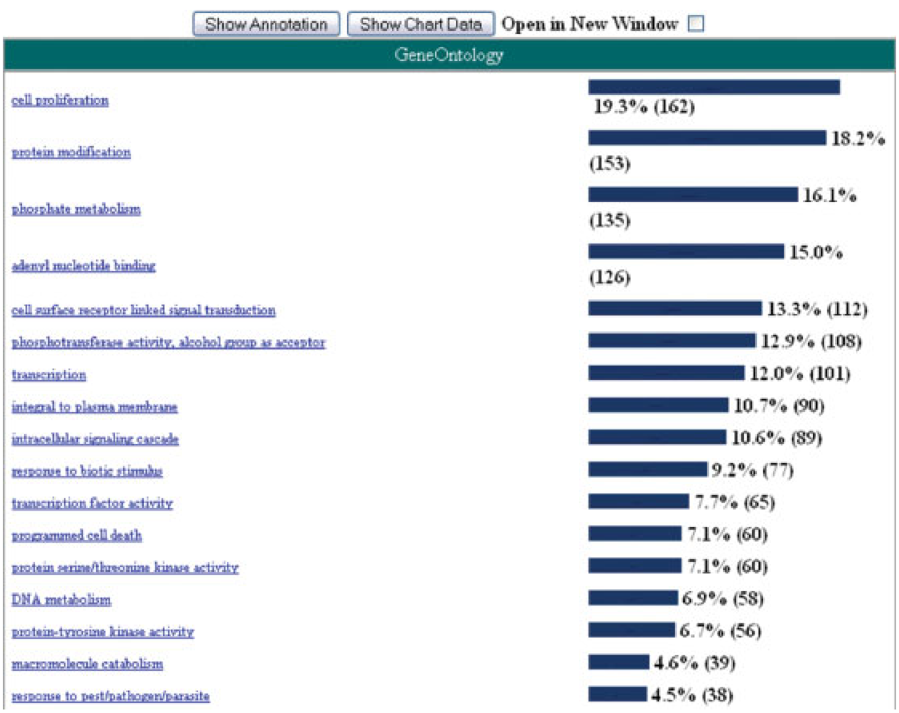
GeneMerge only accepts organism-specific gene IDs used by the various organism-specific annotation groups to deposit functional annotations in the GO Consortium database. Although in principle, it supports more types of IDs as input, its usefulness is somewhat limited since it does not support commonly used IDs such as GenBank accession numbers, RefSeq mRNA, protein IDs, and so on. In order to address this, GeneMerge's authors have recently provided a gene name converter that translates a larger set of IDs into the required format accepted by GeneMerge. As a limitation, GeneMerge only allows to query one of the three principal GO categories at a time. Outstanding features for GeneMerge include the ability to query a few organism-specific features such as KEGG metabolic and signaling pathways for yeast and fruit fly, and deletion viability data for yeast, and so on (Figure 5).

FuncAssociate uses Fisher's exact test to calculate the p-values and a Monte-Carlo simulation to correct for multiple hypothesis testing (Figure 6). FuncAssociate allows the user to submit a rank-ordered list of genes as input, such as a list of genes ranked by fold changes. Given this, FuncAssociate ranks the GO terms in the result according to which initial segment of the input list gives the most significant degree of over- or underrepresentation, where the initial segment for each gene g in the original ordered set Q, is the subset of Q consisting of g plus all the genes that precede g in the given order (Berriz *et al*., 2003).

GOTree Machine (GOTM) has the ability to visualize the results both as a hierarchical tree as well as a bar chart (Figure 7). However, GOTM requires the user to select a specific depth in the GO hierarchy, before displaying a bar chart. GOTM supports 37 Affymetrix arrays as reference arrays. A key aspect of GOTM is the performance. The output interface of GOTM is based on HTML, which means that every user interaction, such as a request to expand a GO term or a request to retrieve a list of input gene IDs, is an HTTP request over the Internet. This is reflected in a



(a)



(b)

Figure 4 EASE input (a) and output (b) interface. EASE supports GenBank accession IDs, Affymetrix probe IDs, UniGene cluster IDs, and LocusLink IDs as input. The user can either submit a file containing the input IDs or paste the input IDs in the text area (a). EASE does not consider the hierarchical structure of the GO in the analysis (b)

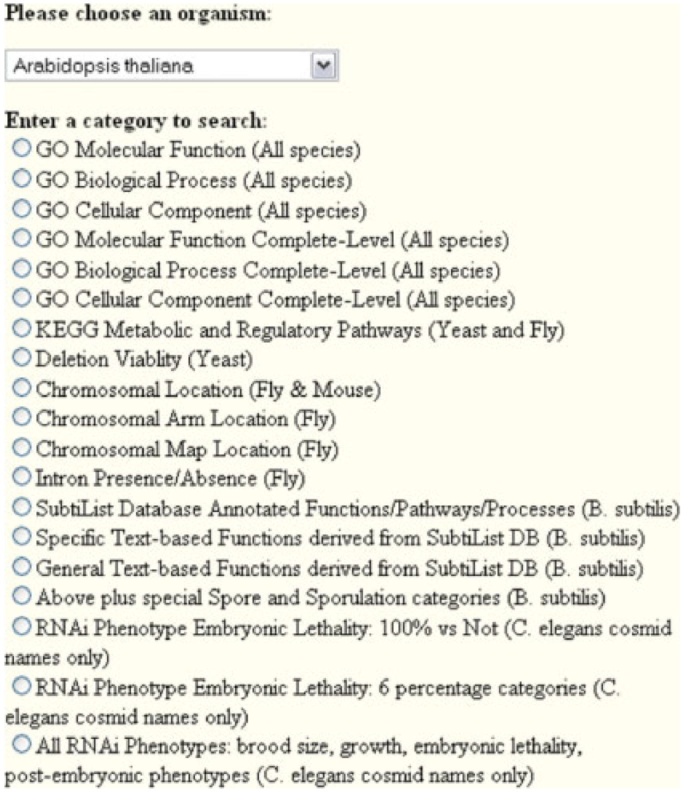


Figure 5 GeneMerge input interface. In addition to functional profiling, GeneMerge also allows to query for KEGG pathways and chromosomal location for yeast and fly, RNAi phenotypes for *C. elegans*, and functions, processes and pathways for *B. subtilis*

slower performance compared to GoMiner or to Onto-Express, which only query their respective back-end database once, after which all processing is done locally.

FatiGO is also a Web-based tool. FatiGO performs the analysis on only one of the three principal GO categories per request and requires the user to select the depth in the GO at the beginning of the analysis (Figure 8). However, the analysis can only be performed up to the 6th level in the GO hierarchy. FatiGO displays the results both as a bar chart as well as a completely expanded GO tree in a single HTML page. If the input only contains a relatively small number of genes (e.g., fewer than 50) and the number of GO terms is limited, this display is informative and convenient since little or no page scrolling is required. However, if the input list contains a few hundreds of genes, which usually yields a functional profile with a few thousand GO terms, the single page display becomes overwhelming.

CLENCH is a command-line, stand-alone tool specifically designed for *Arabidopsis thaliana*. At the moment, CLENCH uses TAIR as a unique source of annotations. In consequence, *A. thaliana* annotations originated at TIGR may not be considered in the analysis. CLENCH retrieves the annotations for *A. thaliana* at runtime from the TAIR website. Similar to FatiGO, CLENCH also outputs the

OVERREPRESENTED ATTRIBUTES

Rank	N	X	LOD	P	P-adj	GO Attribute
1	111	4678	0.703	6.6e-29	<0.001	0007154: cell communication
2	95	3847	0.670	6.3e-25	<0.001	0007165: signal transduction
3	55	1359	0.809	5.1e-23	<0.001	0006793: phosphorus metabolism
4	55	1359	0.809	5.1e-23	<0.001	0006796: phosphate metabolism
5	46	995	0.851	1.8e-21	<0.001	0006468: protein amino acid phosp
6	44	954	0.846	1.7e-20	<0.001	0004672: protein kinase activity
7	141	9175	0.563	1.7e-19	<0.001	0009987: cellular process
8	46	1132	0.790	3.1e-19	<0.001	0016310: phosphorylation
9	45	1093	0.794	4.9e-19	<0.001	0016773: phosphotransferase activ:
10	63	2246	0.650	3.1e-18	<0.001	0006464: protein modification
11	46	1299	0.725	6.3e-17	<0.001	0016301: kinase activity/phosphok:
12	47	1367	0.713	8.5e-17	<0.001	0016772: transferase activity, tr
13	79	3728	0.546	4.8e-16	<0.001	0004871: signal transducer activi
14	11	31	1.841	8.6e-16	<0.001	0004907: interleukin receptor act:
15	11	32	1.821	1.3e-15	<0.001	0019965: interleukin binding/IL b:
16	13	72	1.452	3.4e-14	<0.001	0019955: cytokine binding
17	24	413	0.914	1.3e-13	<0.001	0004713: protein-tyrosine kinase
18	13	81	1.391	1.7e-13	<0.001	0004896: hematopoietin/interferon-
19	13	88	1.348	5.2e-13	<0.001	0004714: transmembrane receptor p
20	82	4623	0.459	2.5e-12	<0.001	0019538: protein metabolism/prote:
21	32	887	0.706	4e-12	<0.001	0050794: regulation of cellular p
22	59	2733	0.512	5.2e-12	<0.001	0017076: purine nucleotide bindin
23	52	2219	0.539	7e-12	<0.001	0030554: adenyl nucleotide bindin
24	59	2767	0.506	8.7e-12	<0.001	0000166: nucleotide binding
25	13	110	1.237	9.7e-12	<0.001	0019199: transmembrane receptor p
26	51	2186	0.535	1.4e-11	<0.001	0005524: ATP binding
27	35	1207	0.609	1.5e-10	<0.001	0008283: cell proliferation
28	151	12633	0.412	1.5e-10	<0.001	0005488: binding/ligand
29	14	169	1.066	1.9e-10	<0.001	0007167: enzyme linked receptor p
30	35	1242	0.596	3.2e-10	<0.001	0005887: integral to plasma membr
31	42	1716	0.541	3.2e-10	<0.001	0005886: plasma membrane/bacteria.
32	12	120	1.155	4.5e-10	<0.001	0007169: transmembrane receptor p

Figure 6 FuncAssociate output interface. FuncAssociate allows the user to submit a rank-ordered list of genes as input. The GO terms in the results are then ordered in the increasing order of the p-values. The ranks of the GO terms are displayed in the first column and their p-values are displayed in the fifth column

results as a single HTML file. Unlike FatiGO, it does not show or consider the hierarchical structure of GO. The local installation of CLENCH requires a previous installation of Perl as well as specific Perl modules. Hence, the initial deployment of CLENCH as well as its command-line interface may be a bit uncomfortable for a typical life scientist.

GOstat is another Web-based tool (Figure 9). It supports *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Danio rerio*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Similar to EASE and FatiGO, GOstat requires the user to select the level of abstraction in the GO hierarchy. GOstat supports GenBank accession number, UniGene cluster IDs, gene symbols and various organism-specific IDs. GOstat represents the results in an HTML or tab-delimited text file, which does not consider the GO hierarchy. GOstat offers false discovery rate (FDR) or Holm's correction for multiple hypothesis testing.

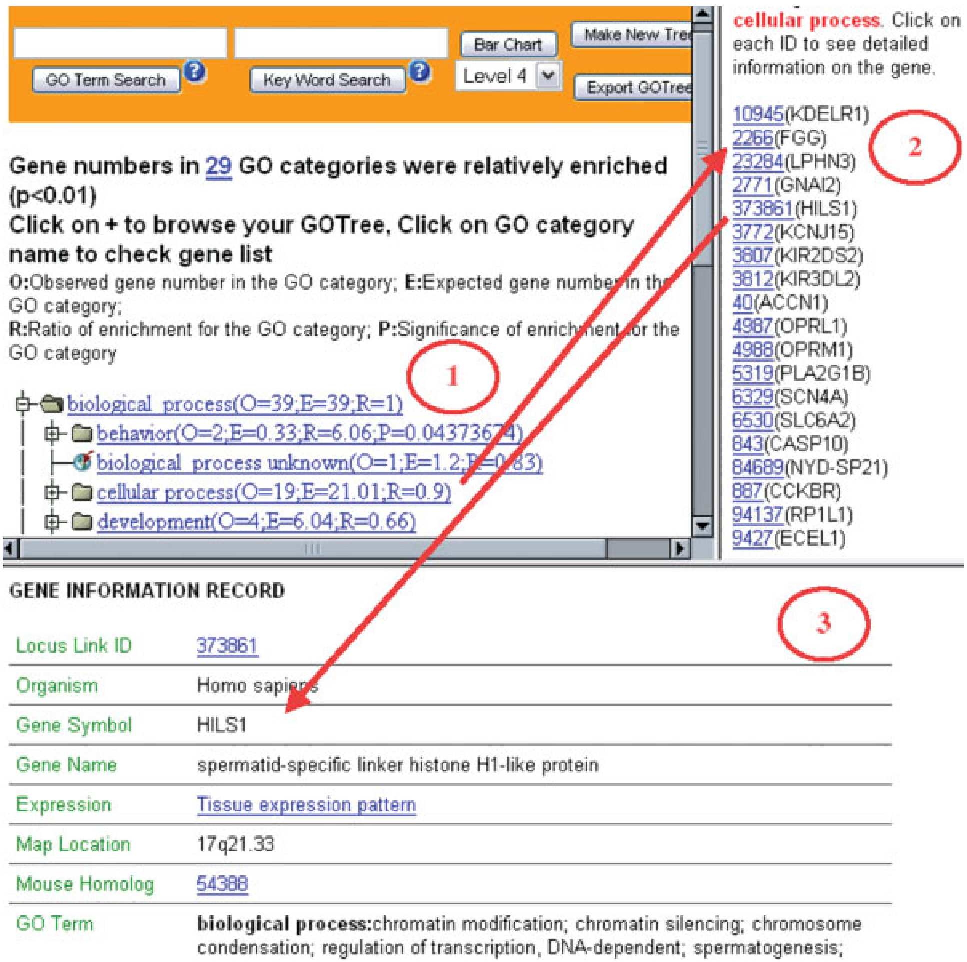


Figure 7 GOTree Machine (GOTM) output. GOTM results are organized in three frames as numbered in the figure. The first frame displays the results in the GO hierarchy. Clicking a GO category in the first frame displays the input genes annotated with the GO category in the second frame. Clicking on a gene in the second frame presents more details about the gene such as the gene name and symbol, LocusLink ID, location on a chromosome, and so on, in the third frame

GOToolBox is also a Web-based tool and supports *Arabidopsis thaliana*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (Figure 10). GoToolBox only supports the organism-specific IDs used in the GO database. GOToolBox either requires the user to select a specific depth in the GO hierarchy or considers the lowest level in the GO hierarchy.

GoSurfer is available as a stand-alone application and supports Affymetrix probe IDs, LocusLink IDs, and UniGene cluster IDs as input. It requires the user to download a specifically formatted GO structure file and at least one gene information file that is also specifically formatted. It displays the results as a DAG on which zoom operations are possible. However, none of the nodes in the DAG



(b)

Figure 8 FatiGO input (a) and output (b) interface. FatiGO only allows to query one of the three principal GO categories at a time. As seen in the highlighted area, FatiGO either uses the entire genome or only the genes annotated with GO terms as the reference. FatiGO presents the results as bar charts as well as a completely expanded GO tree in a single HTML page (b)

are labeled, which makes it very difficult for the user to interpret the results or look for a specific GO term.

OntologyTraverser is described (This is the only tool that we could not use ourselves. In spite of numerous attempts over several weeks, we always encountered the following error: “Error unmarshaling return header; nested exception is: java.io.EOFException”.) as a Web-based tool as well as an R package. Its Web-based version allows the user to query one GO category at a time and calculates the *p*-values using a hypergeometric distribution. However, the model seems to be applied rather differently: instead of considering the number of genes for a given GO term, this tool seems to use the number of genes annotated with the terms at the same level in GO (Young *et al*., 2004). If this description is accurate, this usage of the hypergeometric model may be difficult to justify from a biological point of view.

eGOn. Similar to EASE, GeneMerge, FatiGO, and Ontology Traverser, eGOn only allows the user to query one of the three main GO categories at a time (Figure 11). The analysis performed by eGOn is centered around UniGene clusters.

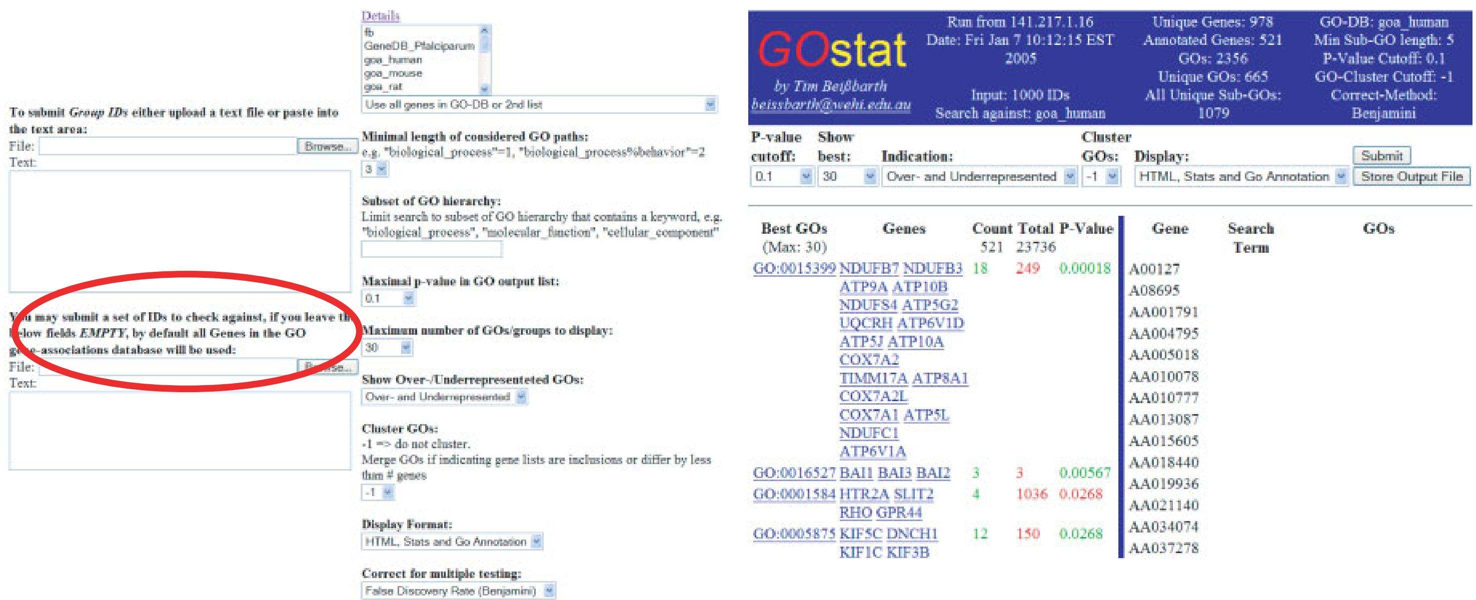


Figure 9 GOstat input and output interface. GOstat requires that the user must upload a reference gene list. Otherwise, it uses all genes in the GO database as the reference. The gene names in the GOstat output are hyperlinked to GeneCards database

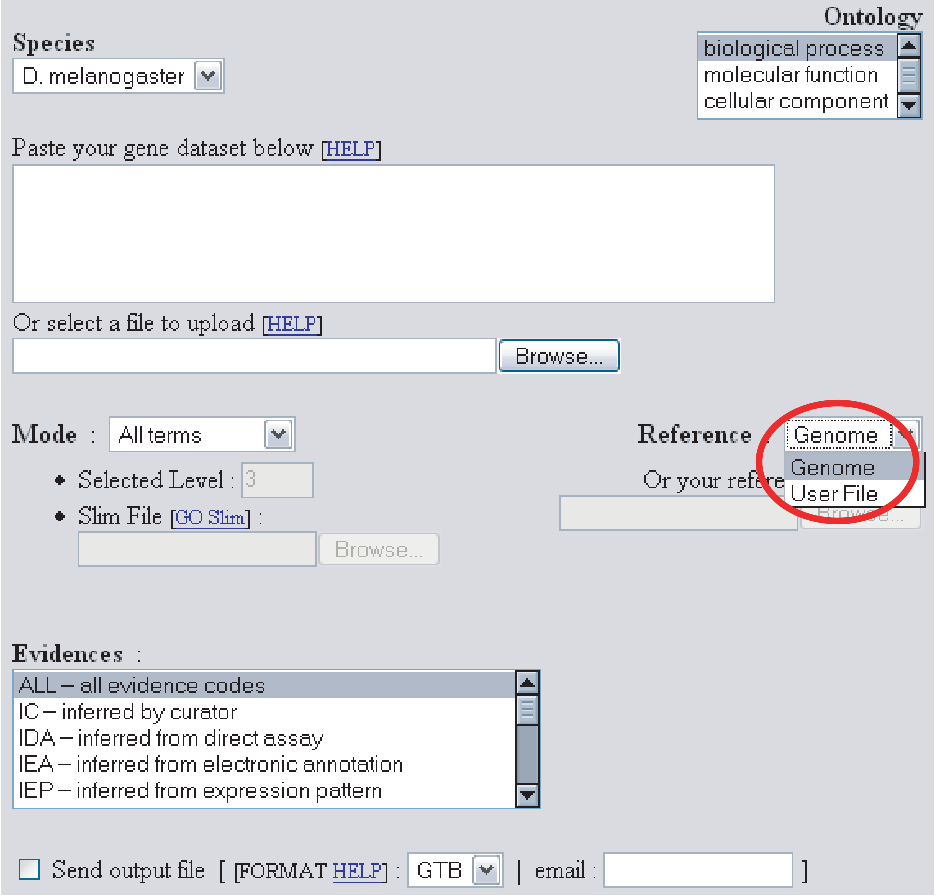


Figure 10 GOToolBox input interface. GOToolBox allows to filter the annotations by evidence codes. If the user does not specify the reference list of genes, GOToolBox uses the genome as the reference

The fundamental implicit assumption here is that each UniGene cluster corresponds to distinct genes. However, this assumption may not always be accurate. UniGene has been created by comparing expressed sequence tags (ESTs) in the dbEST database (Schuler, 1997). However, owing to the alternative splicing of the mRNA, it is entirely possible that ESTs from the same gene cluster in different groups, which will result in several UniGene clusters being associated with the same gene. As an example, the SET8 gene (*SET8: PR/SET domain containing protein 8*, LocusLink ID: 387893) is associated to UniGene clusters Hs.443735 and Hs.536369. If a study includes any such gene, treating each UniGene cluster as a distinct gene, as eGOn does, may not always be appropriate. In such circumstances, the results can be skewed toward those GO terms that are associated to genes from which more than one UniGene clusters are derived. This may become particularly important if further research will confirm the current estimates that up to 90% of the human genes may have alternative splice variants.

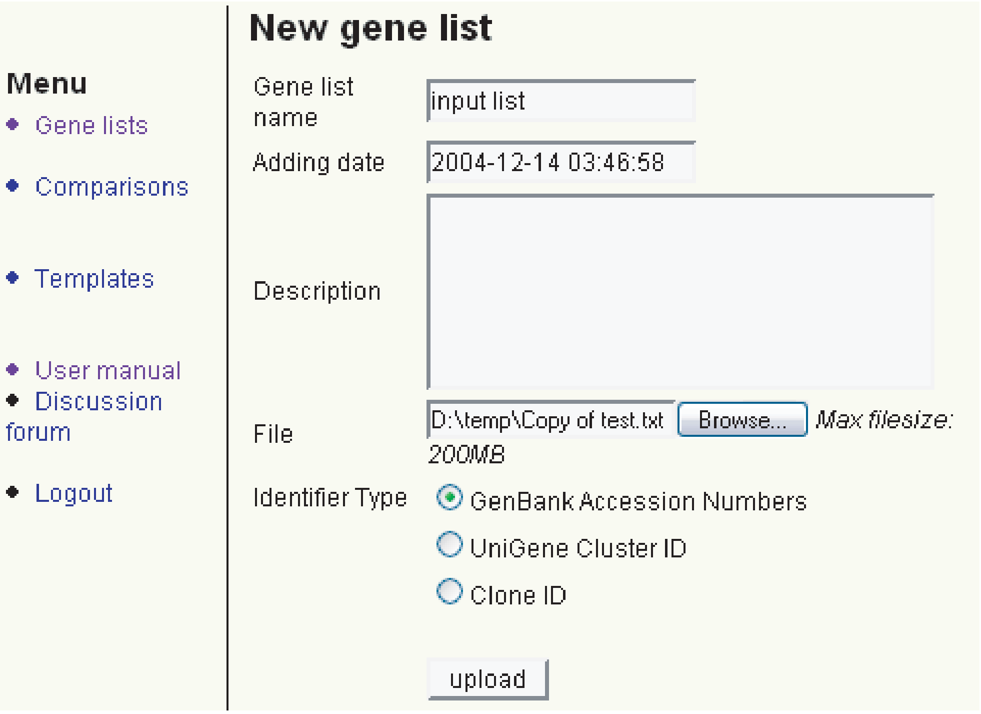


Figure 11 eGOn input interface. eGOn requires that each gene list and each data analysis is given a unique name. The gene lists and analysis are stored on the server for 30 days

4. Comparison of existing functional profiling tools

The criteria used to compare the ontological analysis tools are described in details in the following.

4.1. *The statistical model*

The ontological analysis can be performed with a number of statistical models including hypergeometric (Cho *et al*., 2001), binomial, χ^2 (chi-square) (Fisher and van Belle, 1993), and Fisher's exact test (Man *et al*., 2000). The use and applicability of these tests have been discussed in the literature (Drăghici, 2003; Drăghici *et al*., 2003a,b). More recent tools continue to use one or several of these models (Beissbarth and Speed, 2004; Berriz *et al*., 2003; Martin *et al*., 2004b; Young *et al*., 2004; Zeeberg *et al*., 2003; Zhang *et al*., 2004). GoMiner, EASEonline, GeneMerge, FuncAssociate, GOTM, GOToolBox, GOSurfer, OntologyTraverser, and eGOn only support one statistical test. GOstat allows the user to choose between two tests (χ^2, and Fisher's exact test), CLENCH allows a choice between three tests (χ^2, hypergeometric, and binomial), while Onto-Express implements all four tests (χ^2, hypergeometric, binomial, and Fisher's exact test).

4.2. The set of reference genes

An important consideration when identifying statistically significant GO terms is the choice of the reference list of genes against which the p-values for each GO term in the results are calculated. Several tools such as GOToolBox, GOstat, GoMiner, FatiGO, and GOTM (GOTM also allows the users to upload their own list of genes or use one of 37 Affymetrix arrays as the set of reference genes.) use the total set of genes in a genome as the reference (Beissbarth and Speed, 2004; Martin *et al*., 2004b; Zeeberg *et al*., 2003; Zhang *et al*., 2004) or the set of genes with GO annotations (Al-Shahrour *et al*., 2004). Either of these may be an inappropriate choice when the input list of genes to these tools is a list of differentially expressed genes obtained from a microarray experiment, since the genes that are not present on a microarray do not ever have a chance of being selected as differentially regulated. The fundamental idea here is to assign significance to various functional categories by comparing the observed number of genes in a specific category with the number of genes that might appear in the same category if a selection performed *from the same pool* were completely random. If the whole genome is considered as the reference, the pool considered when calculating the random choice includes all genes in the genome. At the same time, the pool available when actually selecting differentially regulated genes includes only the genes represented on the array used, since a gene that is not on the array can never be found to be differentially regulated. This represents a flagrant contradiction of the assumptions of the statistical models used.

4.3. Correction for multiple experiments

Another crucial factor in the assessment of a functional category is the correction for multiple experiments (see, for instance, Chapter 9 in Drăghici, 2003). This type of correction must be performed in all situations in which the functional category is *not* selected a priori and many such categories are considered at the same time. The importance of this step cannot be overstated and has been well recognized in the literature (Al-Shahrour *et al*., 2004; Beissbarth and Speed, 2004; Berriz *et al*., 2003; Castillo-Davis and Hartl, 2002; Drăghici, 2003; Shah and Fedoroff, 2004; Zeeberg *et al*., 2003). In spite of this, several of the tools reviewed here do not perform such a correction: GoMiner, DAVID, GOTM, CLENCH, and eGOn. GoMiner provides a "relative enrichment" statistic calculated as $R_e = \frac{n_f/n}{N_f/N}$, where n and N are the numbers of genes in the selected and reference sets, respectively, and n_f and N_f are the number of genes in the functional category of interest in the selected and reference sets, respectively (Zeeberg *et al*., 2003). However, this relative enrichment cannot be used in any way as a correction for multiple experiments (Note that this statistic does not take into consideration the number of experiments performed in parallel.) but rather as another indication of the significance of the given category, somewhat redundant to, but less informative than the p-value. This statistic can be misleading because the user will be tempted to assign biological meaning to all those categories that are enriched. In reality, any particular relative enrichment value can actually appear with a nonzero probability just by chance. It

is the magnitude of the probability that should be used to decide whether a category is significant or not, rather than the relative enrichment.

All remaining tools deal with the problem of multiple comparisons in some way. EASEonline, GeneMerge, and GOToolBox support the Bonferroni correction. Bonferroni and Šidák are perfectly suitable in many situations, in particular, when not very many functional categories are involved (e.g., fewer than 50). However, these corrections are known to be overly conservative if more categories are involved (Drăghici, 2003). A family of methods that allow less conservative adjustments of the p-values is the Holm step-down group of methods (Hochberg and Tamhane, 1987; Holland and Copenhaver, 1987; Holm, 1979; Shaffer, 1986).

Bonferroni, Šidák, and Holm's step-down adjustment are statistical procedures that assume that the variables are independent, which is known to be false for this type of analysis (The very hierarchy of the GO on which this type of analysis relies shows that many biological categories are very closely related, sometimes as children of the same node on the next level up.). For this type of analysis, when it is known that dependencies exist, methods such as false discovery rate (FDR) are more appropriate (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Drăghici, 2003). If the dependencies are very strong (e.g., several subprocesses of the same larger process), a bootstrap or Monte-Carlo simulation approach may be better able to capture these dependencies, but only if enough categories are present to make the simulation meaningful.

The tools offering more than one correction method effectively allow the researcher to adapt the analysis to the number of categories and degree of known dependencies between them. The one tool standing out regarding this criterion is FuncAssociate, which uses a more original Monte-Carlo simulation. FatiGO and GOStat implement Holm's and FDR corrections. Onto-Express offers Bonferroni, Šidák, Holm's, and FDR.

4.4. The scope of the analysis

An important factor in assessing the usefulness of a tool is its ability to provide a complete picture of the phenomenon studied. In terms of functional profiling using GO, a complete analysis should include all three primary GO categories: molecular function, biological process and cellular component as well as other information if available. Among the tools reviewed, eGOn, FatiGO, GeneMerge, and OntologyTraverser only analyze one category at a time. The other tools allow the user to analyze all three categories simultaneously. Extra features are present in GeneMerge and Onto-Express. GeneMerge shows KEGG metabolic and signaling pathways for yeast and fruit fly, and deletion viability data for yeast. Onto-Express also shows KEGG metabolic and signaling pathway data, as well as a chromosome location of differentially regulated genes (linking to NCBI's Mapviewer for further analysis).

4.5. Visualization capabilities

The GO is organized as a directed acyclic graph (DAG), which is a hierarchical structure similar to a tree where, unlike a tree, a node can have several parents.

However, the DAG structure may not be the best choice for navigational purposes since it tends to clutter the display (Zeeberg *et al*., 2003). An alternative way to visualize the DAG structure of the GO is to represent and visualize it as a tree structure in which a node with several parents is represented in the tree multiple times, once under each parent. Any tool using GO for functional profiling of a list of genes should be able to graphically represent the hierarchical relationships between various functional categories, which allows the user to better understand the phenomenon studied. Furthermore, the functional analysis can be continued and refined by exploring certain interesting subgraphs of the GO hierarchy.

Among the tools reviewed, CLENCH, DAVID, EASEonline, FuncAssociate, GOstat, GOToolBox, and Ontology Traverser do not show the results in the context of the hierarchical structure of GO. Onto-Express, eGOn, FatiGO, GoMiner, and GOTM represent the results in their GO context. Onto-Express also allows sorting and searching operations in the hierarchy, appropriately expanding and/or collapsing nodes if necessary.

4.6. Custom level of abstraction

In the hierarchical structure of the GO, the genes are annotated at various level of abstraction (see Figure 12). For instance, *induction of apoptosis by hormones* is a type of *induction of apoptosis*, which in turn is a part of *apoptosis*. *Apoptosis* represents a higher level of abstraction, more general, whereas *induction of apoptosis by hormones* represents a lower level of abstraction, more specific. When annotating the genes with the GO terms, efforts are made to annotate the genes with the highest level of details possible, which corresponds to the lowest level of

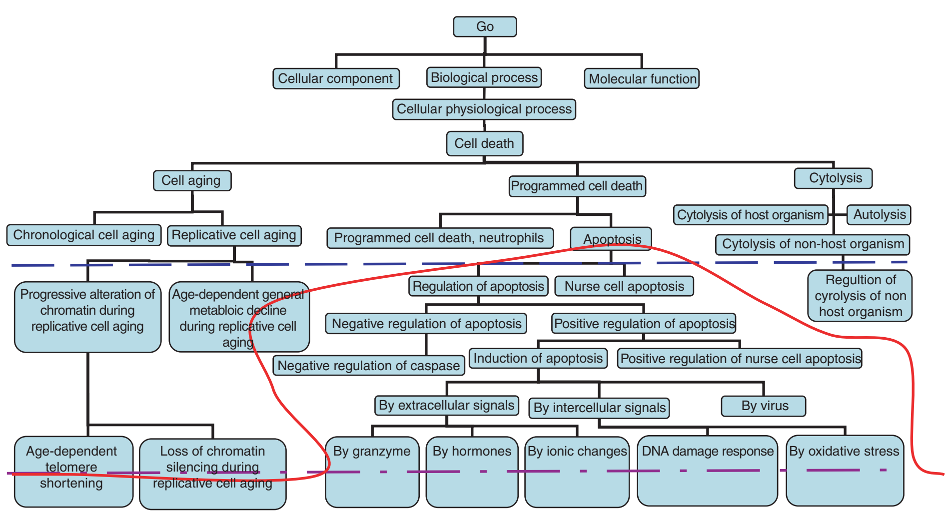


Figure 12 Levels of abstraction. The analysis can be performed at a lowest level of abstraction (purple dash-and-dot line), at a fixed level of abstraction chosen by the user (blue dashed line) or at a custom level of abstraction that can go to different depths in various subtrees of the GO (red continuous line)

abstraction. For example, if a gene is known to induce apoptosis in response to hormones, it will be annotated with the term *induction of apoptosis by hormones* and not merely with one of the higher level terms such as *induction of apoptosis* or *apoptosis*. A very valuable capability of a functional profiling tool is to let the user select a custom level of abstraction. From this point of view, the tools reviewed fall into one of the following three categories. The first category includes the tools able to perform the analysis only with the specific terms associated with each gene. This corresponds to an analysis undertaken at the lowest possible level of abstraction or the highest level of specificity (see the dash-and-dot line in Figure 12). This type of analysis is essentially a one-shot look-up into the annotation database used. Each data set can only be analyzed once since any further analysis can only provide the exact same results. The analysis cannot be directed to answer specific biological questions and cannot be refined in any way.

The second category of the tools include those tools that allow the user to select a predetermined depth, or level of abstraction in GO. Once this level is selected, these tools will consider any gene below the chosen level associated with the corresponding category at the chosen level. This is illustrated by the dashed line in Figure 12. The capability of choosing a predetermined depth allows the user to refine this analysis by performing it repeatedly, at various levels of abstraction, thus forcing various very specific terms to be grouped into more general, and perhaps more informative categories. It is often the case that each specific category does not appear to be significant because there are only few genes associated with it, while the more general category becomes highly significant once all the genes associated with specific subcategories are analyzed together as representing the more general category. Tools having this capability allow a more complex and detailed analysis that can be directed to ask specific biological questions.

Finally, the third category includes tools that allow a completely custom cut through the GO, at different levels of abstraction in different directions. If the analysis is performed at a fixed depth of 9 for instance, the analysis can distinguish between the various subtypes of apoptosis induction: by hormones, by extracellular signals, by intracellular signals, and so on (see Figure 12). However, for a fixed depth, the same analysis will also be performed on several other thousands of functional categories situated at the same level. If the results are presented in a bar graph, the interesting categories will be cluttered by all the extra categories that just happened to be at the same depth in GO, even though they may not be interesting to the researcher. At the same time, other phenomena may be missed because the chosen level may be too specific for those GO categories. A tool that allows a full customization is the most powerful since it will allow the user to perform the analysis at different depths in various parts of the GO hierarchy, as required by the specific biological hypothesis investigated. This is illustrated by the red continuous line in Figure 12.

Most of the existing tools only perform the analysis at the lowest level in GO, with the specific categories that genes have been annotated with, and do not allow any further refinement. Among the tools reviewed, FatiGO, EASEonline, and GOstat allow the user to select a specific level of abstraction before submitting the input list of genes. FatiGo and GOMiner also calculate a p-value for all nodes throughout the GO. This corresponds to a global static analysis in which all genes

under a certain node are considered to be associated with that node. Onto-Express is, at this time, the only tool that allows a fully customized analysis by allowing any node to be collapsed or expanded in the GO. Collapsing a node is equivalent to reassigning to this node all genes associated with any of its descendants. The p-value calculated for a collapsed node in Onto-Express corresponds to the p-value calculated in the global static analysis performed by GoMiner and FatiGo. Expanding a node will distinguish between genes associated with the node itself and the genes associated to any of it descendants. The p-value of an expanded node will be based only on the genes directly associated with it. This p-value is not provided by any other tool from those reviewed here. A current drawback in Onto-Express is that if a user wishes to perform the analysis at a fixed depth throughout GO, the user is required to manually expand the nodes up to this level.

4.7. Prerequisites and installation issues

Another important factor is the amount of effort necessary in order to install and use a tool. The Web-based services provide the experimental biologists a convenient solution by avoiding the problems usually associated with a local installation of a program (Zhang *et al*., 2004). On the other hand, tools available over the Web may be initially obstructed by security issues. For instance, if the tool uses a specific TCP/IP port and the researcher is behind a firewall, the required port must be open on the firewall before the tool can be used.

Stand-alone tools such as CLENCH, GoMiner, and GoSurfer force the user to understand the complexities of a software installation. For example, prerequisites for CLENCH include the prior installation of Perl modules for: (1) HTTP request handling, (2) file and console access, (3) common gateway interface (CGI), (4) database access, (5) statistical computation, and (6) graphical display (Shah, 2004). As another example, GoMiner requires the user to install the Adobe scalable vector graphics (SVG) and NCBI Cn3D browser plug-ins GoMiner. In comparison, the Web-based tools such as Onto-Express, EASEonline, DAVID, Ontology Traverser, GOTree Machine (GOTM), FuncAssociate, FatiGO only require that the user has a web browser with an Internet connection.

The most important problem in this category is the version control. From this point of view, the Web-based tools are far superior because the researcher can always be assured that they are using the very latest version of the software. The software or database updates are always done on the server, by the team who initially wrote the tool. For stand-alone tools, the burden of version control usually rests upon the user who is required to check periodically for new releases and updates. Once such an update becomes available, the burden of software or data update rests again with the user who has to go over the installation process again. In many cases, the updated version works worse than the older version due to issues related to the local environment. In principle, stand-alone tools can try to address this issue by providing automatic software updates. However, this approach means that the updating software must correctly identify, and appropriately deal with, various local software environment issues that tend to be slightly different over the potential hundreds or thousands of different installations.

4.8. Data sources

Most of the available tools use annotation data from a single public database. This has the advantage that the data is always as up-to-date as the database used. The disadvantage is that no single database offers a complete picture. The secondary analysis is more powerful if more types of data are integrated in a coherent way. A dedicated annotation database that integrates various types of data from various sources is potentially more useful than any single database. The drawback is that such a database is (1) difficult to design and (2) will need to be updated every time any one of its source databases is. This places a heavy burden on the shoulders of the team maintaining it. Given this, it is understandable that most tools use only one of the available annotation databases. EASEonline, GOSurfer, eGOn, and GOTM use LocusLink, whereas GeneMerge, GoMiner, and GOToolBox use the GO database. Onto-Express uses its own Onto-Tools database, which is currently the only attempt to integrate resources from several annotation databases. Currently, Onto-Tools uses data from, and it is linked to GenBank, dbEST, UniGene, LocusLink, RefSeq, Gene Ontology, and KEGG. Onto-Tools also uses data from NetAffx and Wormbase without being linked to them.

5. Conclusions

This paper discusses comparison of several ontological analysis tools implementing a very similar functional profiling approach. The comparison used the following criteria: statistical model(s) used, type of correction for multiple comparisons, reference microarrays available, scope of the analysis, visualization capabilities, capabilities for analysis at a custom level of abstraction, prerequisites and installation issues, and the sources of annotation data.

Acknowledgments

This work has been supported by the following grants: NSF DBI-0234806, DOD DAMD 17-03-02-0035, NIH(NCRR) 1 S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1 R21 CA10074001, 1 R21 EB00990-01, and 1 R01 NS045207-01.

Further reading

GoMiner. *GoMiner System Requirements* (2003) **http://discover.nci.nih.gov/gominer/requirements.jsp**.

References

Al-Shahrour F, Diaz-Uriarte R and Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**(4), 578–580.

Ashburner M (1998) On the representation of gene function in genetic databases. *Proceedings of ISMB, Montreal*.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Ashburner M, Ball CA, Blake JA, Butler H, Cherry JM, Corradi J, Dolinski K, Eppig JT, Harris M, Hill DP, *et al*. (2001) Creating the Gene Ontology resource: design and implementation. *Genome Research*, **11**(8), 1425–1433.

Beissbarth T and Speed T (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**(1), 289–300.

Benjamini Y and Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188.

Berriz GF, King OD, Bryant B, Sander C and Roth FP (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**(18), 2502–2504.

Boguski M, Lowe T and Tolstoshev C (1993a) dbEST–database for "expressed sequence tags". *Nature Genetics*, **4**(4), 332–333.

Boguski MS, Lowe TmJ and Tolstoshev CM (1993b) dbest - database for expressed sequence tags. *Nature Genetics*, **4**, 332–333.

Castillo-Davis CI and Hartl DL (2002) GeneMerge-post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**(7), 891–892.

Cho R, Huang M, Campbell M, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge S, Davis R and Lockhart D (2001) Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, **27**, 48–54.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA Jr (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, **4**, P3.

Drăghici S (2003) *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC Press.

Drăghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA and Tainsky MA (2003a) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, **31**(13), 3775–3781.

Drăghici S, Khatri P, Martins RP, Ostermeier GC and Krawetz SA (2003b) Global functional profiling of gene expression. *Genomics*, **81**(2), 98–104.

Explore Gene Ontology (eGOn) (2005) **http://nova2.idi.ntnu.no/egon/**.

Fisher LD and van Belle G (1993) *Biostatistics: A Methodology for Health Sciences*, John Wiley and Sons: New York.

Hochberg Y and Tamhane AC (1987) *Multiple Comparison Procedures*, John Wiley & Sons: New York.

Holland B and Copenhaver MD (1987) An improved sequentially rejective Bonferroni test procedure. *Biometrical*, **43**, 417–423.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hosack DA, Dennis G, Sherman BT, Lane HC and Lempicki RA Jr (2003) Identifying biological themes within lists of genes with EASE. *Genome Biology*, **4**(6), P4.

http://194.117.20.201/tools/ssm/. (2003) *FuSSiMeG: Functional Semantic Similarity Measure between Gene-Products*.

Khan S, Situ G, Decker K and Schmidt CJ (2003) GoFigure: automated Gene Ontology. *Bioinformatics*, **19**(18), 2484–2485.

Khatri P, Drăghici S, Ostermeier GC and Krawetz SA (2002) Profiling gene expression using Onto-Express. *Genomics*, **79**(2), 266–270.

Man MZ, Wang Z and Wang Y (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, **16**(11), 953–959.

Martin DM, Berriman M and Barton GJ (2004a) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, **5**(1), 178.

Martin D, Brun C, Remy E, Mouren P, Thieffry D and Jacq B (2004b) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, **5**, R101.

Pruitt KD, Katz KS, Sicotte H and Maglott DR (2000) Introducing refseq and locuslink: curated human genome resources at the ncbi. *Trends in Genetics*, **16**(1), 44–47.

Schuler GD (1997) Pieces of puzzle: Expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, **75**(10), 694–698.

Shaffer JP (1986) Modified sequentially rejective multiple test procedures. *Journal of American Statistical Association*, **81**, 826–831.

Shah N (2004) **http://www.personal.psu.edu/faculty/n/h/nhs109/Clench/Clench_2.0/Prerequisites.txt**.

Shah N and Fedoroff N (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, **20**(7), 1196–1197.

Young A, Whitehouse N, Cho J and Shaw C (2004) Ontologytraverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, **4**(4), R28.

Zhang, B Schmoyer, D Kirov S and Snoddy S (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**(16).

Zhong S, Tian L, Li C, Storch K-F and Wong WH (2004) Comparative analysis of gene sets in the Gene Ontology space under the multiple hypothesis testing framework. *Proceeding of the IEEE Computational Systems Bioinformatics*, 425–435.

Short Specialist Review

Extracting networks from expression data

Vassily Hatzimanikatis and Eleftherios T. Papoutsakis
Northwestern University, Evanston, IL, USA

Development of sophisticated analytical technologies has enabled the cell-wide monitoring of gene and protein expression, and protein–protein and protein–DNA interactions (*see* Article 91, **Creating and hybridizing spotted DNA arrays**, Volume 4, Article 92, **Using oligonucleotide arrays**, Volume 4, Article 100, **Protein microarrays as an emerging tool for proteomics**, Volume 4, Article 103, **SAGE**, Volume 4, Article 1, **Core methodologies**, Volume 5, Article 20, **Separation-dependent approaches for protein expression profiling**, Volume 5, Article 21, **Separation-independent approaches for protein expression profiling**, Volume 5, Article 22, **Two-dimensional gel electrophoresis**, Volume 5, Article 23, **ICAT and other labeling strategies for semiquantitative LC-based expression profiling**, Volume 5, Article 24, **Protein arrays**, Volume 5, Article 28, **Real-time measurements of protein dynamics**, Volume 5, Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5, Article 30, **2-D Difference Gel Electrophoresis – an accurate quantitative method for protein analysis**, Volume 5, Article 31, **MS-based methods for identification of 2-DE-resolved proteins**, Volume 5, Article 32, **Arraying proteins and antibodies**, Volume 5, Article 37, **Inferring gene function and biochemical networks from protein interactions**, Volume 5, Article 38, **The *C. elegans* interactome project**, Volume 5, Article 39, **The yeast interactome**, Volume 5, and Article 44, **Protein interaction databases**, Volume 5). These technologies allowed the study of the responses of the genetic and protein networks to environmental and genetic perturbations, and the identification of genes associated with diseased phenotypes. In the analysis of mRNA and protein expression data, genes or proteins are clustered with respect to their similarity in expression changes (compared to a control cellular condition or a "global" or "reference" set of transcripts) on the basis of the hypothesis that genes that serve similar cellular functions, or subject to the same regulatory actions, should display similar expression patterns. Various mathematical and computational techniques have been introduced for the clustering and analysis of expression profiles of genes (Quackenbush, 2001; *see also* Article 50, **Integrating statistical approaches in experimental design and data analysis**, Volume 7, Article 54, **Algorithms for gene expression analysis**, Volume 7, Article 57, **Low-level analysis of oligonucleotide expression arrays**, Volume 7, Article 63, **Relevance networks**, Volume 7).

However, these methods do not allow the construction of genetic or protein interaction networks either between individual genes or groups of genes (clusters). In order to infer possible regulatory architectures in a genetic or protein network, more sophisticated, model-based identification methods are required.

Model-based identification methods are based on the formulation of a mechanistic model of mRNA mass balances, that is, synthesis rate minus degradation rate. The synthesis of each mRNA is also considered to be under the regulation of multiple gene products. These models involve a set of structural parameters and a set of quantitative parameters. The structural parameters represent the number and identity of the genes that regulate the expression of each gene, and the type of regulation, for example, induction or repression. The quantitative parameters represent the strength of the regulatory interactions, the transcriptional efficiency of each gene, and the half-lives of the gene products. Once such a model has been formulated, mathematical and computational methods are employed for the identification of the structural and quantitative parameters. Estimation of the structural parameters will lead to the identification of the genetic regulatory network. Estimation of the associated quantitative parameters will allow the formulation of a mathematical model of the genetic network that could be used to predict network responses to environmental and genetic changes.

Several genetic network models have been proposed, which integrate biochemical-pathway information and expression data to trace genetic regulatory interactions (Akutsu *et al*., 2000; Dasika *et al*., 2004; Ideker *et al*., 2001; Lin *et al*., 2003; Maki *et al*., 2001; Moriyama *et al*., 1999; Noda *et al*., 1998; Thomas *et al*., 2004; Van Someren *et al*., 2000; Yeung *et al*., 2002; Gardner *et al*., 2003; *see also* Article 109, **Analyzing and reconstructing gene regulatory networks**, Volume 6, Article 110, **Reverse engineering gene regulatory networks**, Volume 6, and Article 118, **Data collection and analysis in systems biology**, Volume 6). They range from abstract Boolean descriptions to detailed mechanistic models and from Bayesian networks to neural network models (Bower and Bolouri, 2001; D'Haeseleer *et al*., 2000; Friedman *et al*., 2000). Every representation has its advantages and limitations with respect to its ability to infer structural and quantitative parameters in genetic networks.

One of the first approaches to infer the structural properties of a genetic network based on information of mRNA profiles employed a Boolean modeling framework (Somogyi *et al*., 1997). Boolean networks model the state of the gene as either ON or OFF and the input–output relationships are postulated as logical functions. Further improvements of the Boolean-based frameworks introduced Mutual Information (Liang *et al*., 1998) and an entropy-based approach (Ideker *et al*., 2000) in order to identify the experiments required for discrimination between alternate genetic networks predicted by these frameworks. Rigorous analysis has led to the identification of the number of experiments required for the identification of the Boolean model that is consistent with the data (Akutsu *et al*., 2000; Somogyi *et al*., 1997). However, in real systems, transcript levels vary in a continuous manner, implying that the key assumptions underlying Boolean network models may not be appropriate, and thus, more general models may be necessary.

Several continuous modeling frameworks have also been proposed, and they employ linear models (Chen *et al*., 1999; D'Haeseleer *et al*., 1999; Van Someren

et al., 2000; Weaver *et al*., 1999; Yeung *et al*., 2002; Gardner *et al*., 2003), hybrid Boolean and power law models (Akutsu *et al*., 2000), Bayesian models (Friedman *et al*., 2000), and linear models for inferring time delay in genetic networks (Dasika *et al*., 2004). Linear models cannot capture the inherent nonlinearities of biological systems. A possible approach then is to use the S-system, power-law framework whereby the nonlinearity of genetic regulation can be reasonably captured (Akutsu *et al*., 2000; Lin *et al*., 2003; Thomas *et al*., 2004).

The challenges in extracting networks from expression data arise from the type of data, the dimensionality of the data, the combinatorial nature of gene expression regulation, and the mathematical representation of the genetic network. Since every mathematical model is an approximation of the real biological system, methods for the identification of regulatory interactions should introduce the minimum number of simplifications and allow the identification of multiple possible regulatory interactions that are consistent with the experimental data. This would provide the expert with alternative hypothetical networks that could be evaluated on the basis of expert knowledge and literature information, while guiding experiments for hypothesis validation.

The data can be either temporal, that is, relative gene expression over time, or nontemporal, that is, relative expression over different time-invariant conditions. In most of the proposed methodologies, temporal data are discretized over time, and they are analyzed using frameworks similar to those developed for the analysis of nontemporal data. However, such discretization depends on the time interval of the sampling, and it might introduce biases that could lead to false estimation of the structural and quantitative parameters. Frequent sampling and improved identification methods tailored to the temporal nature of the data might help in improving network extraction from temporal expression data.

One of the biggest limitations in the inference of genetic networks is the use of mRNA expression data alone. While the real regulatory input is the protein product of a gene, most of the mathematical models and the associated inference methods ignore the presence of proteins and they treat mRNAs as the regulatory inputs of the mRNA expression. Ideally, protein expression data should be combined with mRNA expression data in the identification studies wherever this is feasible. However, the cost and sophistication of technologies that allow genome-wide monitoring of protein expression is limiting the availability of such information. Alternatively, it has been proposed that mathematical models could approximately describe the relation between mRNA and protein expression and thus account for lack of information about protein expression. Such approximations involve mechanistic modeling of protein expression for nontemporal data (Thomas *et al*., 2004) or using time-lag action of an mRNA (due to delay in protein synthesis) in temporal data (Chen *et al*., 1999; Dasika *et al*., 2004).

In addition, most of the DNA microarray experiments provide information about the relative expression patterns, whereas most of the models treat this information as absolute concentration information. This limitation can be resolved either by obtaining an estimate of the absolute mRNA concentration levels, or with mathematical representations that formulate the constraints imposed by the mRNA and protein mass balances as constraints on the relative expressions (Akutsu *et al*., 2000; Thomas *et al*., 2004).

A number of studies have also led to some very important conclusions regarding the dimensionality of the data. The number of experiments required for the unique identification of the correct regulatory structure depends both on the number of genes and the number of simultaneously perturbed biophysical parameters in each experiment. The number of the experiments times the number of the biophysical parameters perturbed in each experiment should be a portion of the number of the genes in the system in order to identify a small number of alternative regulatory structures that could explain the data. This result suggests that the required number of experiments could be decreased if in each experiment the number of the perturbed biophysical parameters is increased. Similar considerations suggest that identification of the regulatory structure based on single-gene knockout experiments might require the knockout of every gene (Thomas *et al*., 2004).

In addition, if there are two or more genes whose expression profiles across the different experiments are linearly dependent, any model-based identification method will fail to identify a unique regulatory structure consistent with the data. This could be the case for coregulated genes, or for genes that appear to be coregulated within the accuracy of the experimental methods. This finding led to the suggestion of clustering the coregulated genes into a single "gene" and studying the interactions of this new "gene" in the network (D'Haeseleer *et al*., 1999; Van Someren *et al*., 2000). The identified regulatory connections could then be due to interactions of single or multiple members of the cluster, and closer investigation of the elements in the cluster might help in refining the regulatory network.

Another element contributing to the complexity in reconstructing genetic regulatory networks is the combinatorial nature of regulation. The protein products of multiple genes can regulate the expression of a single gene and in multiple alternative output combinations. For a single gene with K regulatory input, $\left(2^2\right)^k$ regulatory functions are possible. Experimental methods such as chip-on-chip (chromatin immunoprecipitation on a chip (DNA array)) and sequence analysis computational methods would allow one to postulate possible regulatory interactions or to exclude genes from acting as regulators. Integration of this information into the identification frameworks will require the employment of computational frameworks that can enforce the presence or absence of regulatory interactions, as well as the maximum number of regulatory inputs per gene. Mixed-integer programming, the class of optimization methods that involves both discrete and continuous variables, offers the possibility of efficiently dealing with such constraints where the regulatory interaction of each gene with another gene is modeled using a binary variable with values 0 or 1 in the absence or presence, respectively, of this interaction (Dasika *et al*., 2004; Lin *et al*., 2003; Thomas *et al*., 2004). The Mixed-integer formulation can also allow the identification of multiple alternative regulatory networks that are consistent with the data. When such methods identify multiple structures, it is also observed that certain identified regulatory interactions are common in every alternative structure. This suggests that the regulatory interactions that are common in the identified alternative structures are highly probable to be present in the original system.

Finally, the accuracy of the mathematical models used in model-based inference methods plays an important role in the outcome of the analyses. While a detailed, nonlinear mechanistic model of transcription and translation will be able to capture

important aspects of the regulatory complexity, the accuracy of the data, the relatively small amount of experimental information, and the large size of the system do not allow meaningful computational analysis and network identification. Therefore, simpler models with a small number of quantitative parameters are necessary. A better understanding on the model reduction, from the full, nonlinear mechanistic models to simplified nonlinear or linear models is very important in developing model-based inference methods.

Future developments of frameworks for extracting networks from expression data need to address these challenges. The most powerful frameworks should be able to provide experimentalists with alternative hypotheses and molecular models that could be evaluated based on biological expertise and additional experimental studies.

References

Akutsu T, Miyano S and Kuhara S (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.

Bower JM and Bolouri H (Eds.) (2001) *Computational Modeling of Genetic and Biochemical Networks, Computational Molecular Biology Series*, MIT Press: Cambridge.

Chen T, He HL and Church GM (1999) Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, **4**, 29–40.

Dasika M, Gupta A, Maranas CD and Varner JD (2004) A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. *Pacific Symposium on Biocomputing*, **9**, 474–485.

D'Haeseleer P, Liang S and Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

D'Haeseleer P, Wen X, Fuhrman S and Somogyi R (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*, **3**, 41–52.

Friedman N, Linial M, Nachman I and Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Gardner TS, di Bernardo D, Lorenz D and Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ideker TE, Thorsson V and Karp RM (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, **5**, 305–316.

Liang S, Fuhrman S and Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, **3**, 18–29.

Lin X, Floudas CA, Wang Y and Broach JR (2003) Theoretical and computational studies of the glucose signaling pathways in yeast using global gene expression data. *Biotechnology and Bioengineering*, **84**, 864–886.

Maki Y, Tominaga D, Okamoto M, Watanabe S and Eguchi Y (2001) Development of a system for the inference of large scale genetic networks. *Pacific Symposium on Biocomputing*, **6**, 446–458.

Moriyama T, Shinohara A, Takeda M, Maruyama O, Goto T, Miyano S and Kuhara S (1999) A system to find genetic networks using weighted network model. *Genome Informatics Series: Proceedings of the ... Workshop on Genome Informatics. Workshop on Genome Informatics*, **10**, 186–195.

Noda K, Shinohara A, Takeda M, Matsumoto S, Miyano S and Kuhara S (1998) Finding Genetic Network from Experiments by Weighted Network Model. *Genome Informatics Series: Proceedings of the ... Workshop on Genome Informatics. Workshop on Genome Informatics*, **9**, 141–150.

Quackenbush J (2001) Computational analysis of microarray data. *Nature Reviews Genetics*, **2**, 418–427.

Somogyi R, Fuhrman S, Askenazi M and Wuensche A (1997) The gene expression matrix: Towards the extraction of genetic network architectures. *Nonlinear Analysis-theory Methods and Applications*, **30**, 1815–1824.

Thomas R, Mehrotra S, Papoutsakis ET and Hatzimanikatis V (2004) A model-based optimization framework for the inference on gene regulatory networks from DNA array data. *Bioinformatics*, **20**, 3221–3235.

Van Someren EP, Wessels LFA and Reinders MJT (2000) Linear modeling of genetic networks from experimental data. *Eighth International Conference on Intelligent Systems for Molecular Biology*, AAAI, La Jolla.

Weaver DC, Workman CT and Stormo GD (1999) Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing*, **4**, 112–123.

Yeung MKS, Tegner J and Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *PNAS*, **99**, 6163–6168.

Short Specialist Review

Data standardization and the HUPO proteomics standards initiative

Henning Hermjakob, Chris Taylor, Sandra Orchard, Weimin Zhu and Rolf Apweiler
European Bioinformatics Institute, Cambridge, UK

An increasing number of large-scale proteomics studies are currently being published or are in progress, providing a wealth of data and biological insights on a new scale. From protein expression and interactions to posttranslational modifications, data becomes available at an amazing rate. However, large-scale proteomics technologies are still far from the reliability levels of gene sequencing; comparative analyses are therefore essential to improve reliability (von Mering *et al*., 2002). The combination of different data types, for example, protein interactions and gene expression, allow new biological insights, but such integrative approaches are currently difficult and in many instances are impossible due to the fragmentary nature of proteomics data. Unlike other domains of molecular biology, for example, DNA (Kulikova *et al*., 2004; Galperin, 2004) and protein sequences (Apweiler *et al*., 2004), protein structures (Bourne *et al*., 2004) and gene expression (Brazma *et al*., 2003), no well-established databases exist, and the proteomics data supporting publications is often not available at all, or is scattered over supplementary tables and authors' and journals' websites in numerous formats. In addition to data collection, this also renders the scientific scrutiny of experimental conclusions based on the underlying data very time consuming if not impossible. The public availability of large microarray data sets has contributed to the improvement of statistical approaches and has led to important insights gained through the combination of published data with new data sets (Gygi *et al*., 1999); such approaches would usually be difficult in the proteomics domain as it currently stands.

The tide is turning though: an increasing awareness of the need to systematically annotate and archive proteomics data is illustrated by the publication of several proposed schemata and public databases for such data (Wilke *et al*., 2003; Carr *et al*., 2004; Prince *et al*., 2004; Xirasagar *et al*., 2004; Pedrioli *et al*., 2004). However, with the exception of the PEDRo schema (Taylor *et al*., 2003), these systems are developed by individual institutions or small groups of institutions and, as a result, suffer from problems with respect to community ownership and acceptance.

The HUPO Proteomics Standards Initiative (PSI) (Orchard *et al*., 2003) is constituted to define standards for data representation in proteomics, with which to

facilitate data comparison, exchange, and verification. PSI standards are developed in an open, community-based process, where participation of all interested parties is invited. The PSI initially focused on two priority areas for data integration; protein interactions and mass spectrometry. Additionally, the development of a representation of a complete proteomics experiment is ongoing; as is the generation of guidelines laying out the level of detail that is expected to adequately describe protocols, data, and analyses.

The HUPO PSI Molecular Interaction standard (MI) has been jointly developed by major interaction data providers from both the academic and commercial sectors, among them Hybrigenics (**http://www.hybrigenics.fr**), DIP (**http://dip.doe-mbi.ucla.edu/**), BIND (**http://bind.ca/**), MINT (**http://cbm.bio.uniroma2.it/mint/**), Cellzome (**http://www.cellzome.com/**), and IntAct (**http://www.ebi.ac.uk/intact/**). The PSI MI standard is developed in a layered approach, with successive levels providing increasing expressive power, obviously at the cost of increasing complexity. PSI MI 1.0 was published in February 2004 (Hermjakob *et al*., 2004), providing a basic exchange format for protein interaction data. PSI MI formatted data is provided by amongst others MINT, IntAct, HPRD (**http://www.hprd.org**), and DIP; a number of compatible tools are also available. Using PSI MI, it is now possible to connect data and tools from independent organizations, without any data conversion scripting, such as dynamically loading interaction data from IntAct into the Cytoscape analysis tool (**http://www.cytoscape.org**). A full list of PSI MI data and tool resources is in the documentation at **http://psidev.sourceforge.net/mi/xml/doc/user/**. PSI MI level 2, released in autumn 2004, adds additional interactor types, in particular, small molecules, DNA, and RNA. Using the PSI MI format, a number of major interaction data providers have established the International Molecular-Interaction Exchange (IMEx) consortium, through which they will regularly exchange user-submitted data; this is similar to the EMBL/GenBank/DDBJ collaboration, and will underpin synchronized, stable, reliable resources for molecular interaction data.

The PSI mass spectrometry work group (PSI MS) has developed the mzData format; a vendor-independent representation of mass spectra, providing a unified format for data archiving and exchange, and for search engine input. It has been jointly developed by academic users, commercial users, and instrument vendors; amongst them Eli Lilly, EBI, Bruker, Matrix Science, Shimadzu (Kratos), Applied Biosystems/MDS Sciex, Agilent, Waters (Micromass), and Thermo Electron. A controlled vocabulary, in particular, for instrument parameters, has been developed for use with mzData, which will be maintained jointly by the American Society for Mass Spectrometry (ASMS) and PSI; this controlled vocabulary builds on work by the American Society for Testing and Materials (**http://www.astm.com/**) on the Analytical Data Interchange standard (**http://andi.sourceforge.net/**).

The implementation phase for mzData is underway at the time of writing; all major mass spectrometer vendors, and other companies such as Matrix Science (peptide/protein identification software vendor), are scheduled to provide mzData import/export this year. In the next step, the mass spectrometry work group will provide mzIdent, a standard search engine output format designed to make peak list search results (i.e., peptide and/or protein identifications) comparable and amenable

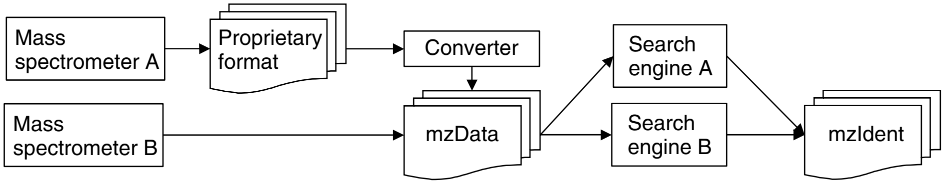


Figure 1 Proposed mass spectrometry data flow based around the formats developed by the PSI-MS workgroup: Mass spectrometers either directly generate mzData files or have their output formats converted to mzData, which itself then serves as a standard input format for search engines. The PSI's standard search engine output format (mzIdent) will then allow an easy comparison of search results from different search engines, and facilitate their postprocessing and submission to public repositories

to standardized postprocessing. Figure 1 shows the complete data flow for a mass-spec-based proteomics experiments.

In addition to PSI MI and MS, there exists a third HUPO PSI workgroup, Global Proteomics Standards (GPS); this workgroup concerns itself with the full description of proteomics experiments, comprising sample preparation, separation and analysis, and the collection and processing of results (a similar scope to that of the PEDRo schema; Taylor *et al*., 2003). The GPS workgroup's efforts fall into three broad areas: First, the Minimum Information About a Proteomics Experiment (MIAPE) guidelines – the analog of the Minimum Information About a Microarray Experiment (MIAME) guidelines for a microarray experiment (Brazma *et al*., 2001) – which lay out the minimum level of detail required for a description of a dataset, and the work that generated it, to a level considered sufficient to both understand and potentially recreate that work. Second, an object model (PSI-OM) and derived XML format (PSI-ML) with which to capture the data and metadata from any proteomics experiment (mzData, mzIdent, and the nascent GelML format for gel electrophoresis all represent components of this overall format). Lastly, the PSI ontology, which shares the same broad scope as the PSI-ML/OM; this ontology incorporates the controlled vocabulary developed for mass spectrometry data described above, and will itself ultimately be fully integrated into the MGED Ontology (**http://www.mged.org/ontology/**; Stoeckert and Parkinson, 2003).

As variously described above, all three PSI workgroups strive to standardize not only data formats but also contents, through the provision of guidelines, documentation, and in particular, through the development or adoption of controlled vocabularies and ontologies. As an example, there are more than 10 different spellings for the "yeast two-hybrid" technology (Legrain and Selig, 2000), for example, Y2H, 2H, two-hybrid; while equivalents are often straightforwardly interpretable by a human, such "families" of terms describing single key concepts make automated searches in large databases very difficult. Where possible, PSI uses existing reference systems, for example, GO (Harris *et al*., 2004) or the MGED ontology, while developing additional controlled vocabularies where necessary, for example, for PSI MI to describe specific methodologies to determine protein interactions. In addition, documentation, and guidelines documents such as MIAPE provide guidance in the use of the PSI formats. Such detailed usage instructions are essential to avoid the fragmentation of standards into dialects, where different data

sets are compliant to a standard, but are not compatible with each other, because flexibilities in the format are differently interpreted.

Beyond the recommendation of standards, the PSI aims to coordinate the implementation of public repositories accepting and providing data in PSI formats, and collaborates with scientific journals and funding agencies to encourage scientists to support their publications with underlying data sets made available through standardized, freely accessible public resources.

The PSI strongly encourages active participation in its work groups. For further information and mailing lists, please refer to the project's web pages at **http://psidev.sf.net/**.

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**(Database issue), D115–D119.

Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, *et al*. (2004) The distribution and query systems of the RCSB protein data bank. *Nucleic Acids Research*, **32**(Database issue), D223–D225.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al*. (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics*, **29**(4), 365–371.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, *et al*. (2003) ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**(1), 68–71.

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K and Nesvizhskii A (2004) The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Molecular & Cellular Proteomics: MCP*, **3**(6), 531–533.

Galperin MY (2004) The molecular biology database collection: 2004 update. *Nucleic Acids Research*, **32**(Database issue), D3–D22.

Gygi SP, Rochon Y, Franza BR and Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, **19**(3), 1720–1730.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*. (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**(Database issue), D258–D261.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al*. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**(2), 177–183.

Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, *et al*. (2004) The EMBL nucleotide sequence database. *Nucleic Acids Research*, **32**(Database issue), D27–D30.

Legrain P and Selig L (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Letters*, **480**(1), 32–36.

Orchard S, Hermjakob H and Apweiler R (2003) The proteomics standards initiative. *Proteomics*, **3**(7), 1374–1376.

Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti R, Apweiler R, *et al*. (2004) A common open representation of mass spectrometry data and its application in a proteomics research environment. *Nature Biotechnology*, **22**(11), 1459–1466.

Prince JT, Carlson MW, Wang R, Lu P and Marcotte EM (2004) The need for a public proteomics repository. *Nature Biotechnology*, **22**(4), 471–472.

Stoeckert CJ and Parkinson H Jr (2003) The MGED ontology: a framework for describing functional genomics experiments. *Comparative and Functional Genomics*, **4**, 127–132.

Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**(3), 247–254.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Wilke A, Ruckert C, Bartels D, Dondrup M, Goesmann A, Huser AT, Kespohl S, Linke B, Mahne M, McHardy A, *et al*. (2003) Bioinformatics support for high-throughput proteomics. *Journal of Biotechnology*, **106**(2–3), 147–156.

Xirasagar S, Gustafson S, Merrick BA, Tomer KB, Stasiewicz S, Chan DD, Yost KJ 3rd, Yates JR 3rd, Sumner S, Xiao N, *et al*. (2004) CEBS object model for systems biology data, SysBio-OM. *Bioinformatics*, **20**(13), 2004–2015.

Large error models for microarray intensities

Wolfgang Huber
European Molecular Laboratory, European Bioinformatics Institutey, Cambridge, UK

Anja von Heydebreck
Department of Bio- and Chemoinformatics, Merck KGaA, Darmstadt, Germany

Martin Vingron
Max-Planck Institute for Molecular Genetics, Berlin, Germany

1. Motivation

An error model is a description of the possible outcomes of a measurement. It depends on the true value of the underlying quantity that is measured and on the measurement apparatus. For microarrays, the quantities that one wants to measure are the abundances of specific molecules in a biological sample. The measurement apparatus consists of a cascade of biochemical reactions and an optical detection system with a laser scanner or a CCD camera. Biochemical reactions and detection are performed in parallel, allowing up to millions of measurements on one array.

What is the purpose of constructing error models for microarrays? There are three aspects:

1. An appreciation of the distribution of all possible outcomes of a measurement is necessary for basing *inference* on one or a limited number of measurements. Consider an experiment in which we want to compare gene expression in the colons of mice that were treated with a certain substance and mice that were not. If we have many measurements, we can simply compare their empirical distributions. For example, if the values from 10 replicate measurements for the *DMBT1* gene in the treated condition are all larger than 10 measurements from the untreated condition, the Wilcoxon test tells us that with a p-value of 10^{-5} the level of the transcript is really elevated in the treated mice. But often it is not possible, too expensive, or unethical, to obtain so many replicate measurements for all genes and for all conditions of interest. Often, it is also not necessary. If we are sufficiently confident in an error model, we are able to draw significant conclusions from fewer replicates.

2. An error model is an efficient tool for the summarization and *reporting* of experimental results. If we have reason to be confident that the measured

outcomes follow a certain distribution, then they are sufficiently described by that distribution's parameters, for example, mean and standard deviation; it may then not be necessary to report all of the individual measurements.

3. An error model is a summary of past experience and of our understanding of the measurement apparatus. It can be used for *quality control*: if the distribution of a new set of data greatly deviates from the model, this may direct our attention to quality issues with these data.

2. The additive-multiplicative error model

Consider the following generic observation equation

$$z = f(x, y) \tag{1}$$

where z is the outcome of the measurement, x is the true underlying quantity that we want to measure, the function f represents the measurement apparatus, and $y = (y_1, \dots, y_n)$ is a vector that contains all other parameters on which the functioning of the apparatus may depend. The functional dependence of f on some of the y_i may be known, on others it may not. Some of the y_i are controlled by the experimenter, some are not. If the measurement apparatus is well constructed, then f is a well-behaved, smooth function, and we can write equation (1) as

$$z = f(0, y) + f'(0, y)x + O(x^2) \tag{2}$$

where $f(0, y)$ is the baseline value that is measured if x is zero, f' is the derivative of f with respect to x, $f'(0, y)$ is a gain factor, and $O(x^2)$ represents nonlinear effects. By proper design of the experiment, the nonlinear terms can be made negligibly small within the relevant range of x. Examples for the parameters y in the case of microarrays are the efficiencies of mRNA extraction, reverse transcription, labeling and hybridization reactions, amount and quality of probe DNA on the array, unspecific hybridization, dye quantum yield, scanner gain, and background fluorescence of the array. All of these have an influence either on the baseline $f(0, y)$ or the gain $f'(0, y)$. Ideally, the parameters y could be fixed exactly at some value $\overline{y} = \left(\overline{y}_1, \dots, \overline{y}_n\right)$. In practice, they will fluctuate around $\overline{y}$ between repeated experiments. If the fluctuations are not too large, we can expand

$$f(0, y) \approx f(0, \overline{y}) + \sum_{i=1}^{n} \frac{\partial f(0, \overline{y}}{\partial y_i} (y_i - \overline{y}_i) \tag{3}$$

$$f'(0, y) \approx f'(0, \overline{y}) + \sum_{i=1}^{n} \frac{\partial f'(0, \overline{y})}{\partial y_i} (y_i - \overline{y}_i) \tag{4}$$

The sums on the right-hand sides of equations (3) and (4) are linear combinations of a large number n of random variables with mean zero. Thus, it is a reasonable approximation to model $f(0, y)$ and $f'(0, y)$ as normally distributed random

variables with means $a = f(0, \overline{y})$ and $b = f'(0, \overline{y})$ and variances σ_a^2 and σ_b^2, respectively.

Thus, omitting the nonlinear term, equation (2) leads to

$$z = a + \varepsilon + bx(1 + \eta) \tag{5}$$

with $\varepsilon \sim N(0, \sigma_a^2)$ and $\eta \sim N(0, \sigma_b^2/b^2)$. This is the *additive-multiplicative error model* for microarray data, which was proposed by Ideker *et al*., (2000). Rocke and Durbin, (2001) proposed it in the form

$$z = a + \varepsilon + bx\exp(\eta) \tag{6}$$

which is equivalent to equation (5) up to first order terms in η. Models (5) and (6) differ significantly only if the coefficient of variation σ_b/b is large. For microarray data, it is typically smaller than 0.2, thus the difference is of little practical relevance.

One of the main predictions of the error model (5) is the form of the dependence of the variance Var(z) of z on its mean E(z):

$$\mathrm{Var}(z) = v_0^2 + \frac{\sigma_b^2}{b^2}\,(\mathrm{E}(z) - z_0)^2 \tag{7}$$

that is, a strictly positive quadratic function. In the following, we will assume that the correlation between ε and η is negligible. Then the parameters of equation (7) are related to those of equation (5) via $v_0^2 = \sigma_a^2$ and $z_0 = a$. If the correlation is not negligible, the relationship is slightly more complicated, but the form of equation (7) remains the same.

3. Nesting

Consider a situation in which the quantity x from equation (6) is itself the result of a process whose outcome is approximately described by an additive-multiplicative error model,

$$x = a' + \varepsilon' + b'x'\exp(\eta') \tag{8}$$

The resulting distribution of z can again be described by an additive-multiplicative model with new parameters. This means that the class of models of the form (6) is closed under such hierarchical nesting, and the range of its applicability can be quite large.

For example, equation (6) could be used as a model for microarray measurements in a study of diseased tissues, with x the true abundance of a certain gene transcript in the tissue from an individual patient, while equation (8) could model the population distribution of this gene's transcript levels in a set of similar tissues from different patients.

4. Detection of differentially expressed genes

Suppose we want to compare two measurements z_1, z_2 that are distributed according to equation (6) with the same parameters a, b, σ_a, and σ_b, but possibly with different values of x_1, x_2. We want to find a function $h(z_1, z_2)$ that fulfills the two conditions:

$$\begin{array}{llcll} \text{(i) antisymmetry:} & h(z_1, z_2) & = & -h(z_2, z_1) & \forall x_1, x_2 \\ \text{(ii) homoskedasticity:} & \mathrm{Var}(h(z_1, z_2)) & = & \text{const} & \text{independent of } x_1, x_2 \end{array} \tag{9}$$

An approximate solution is given by Huber *et al*. (2003)

$$h(z_1, z_2) = \operatorname{arsinh}\left(\frac{z_1 - a}{\beta}\right) - \operatorname{arsinh}\left(\frac{z_2 - a}{\beta}\right) \tag{10}$$

with $\beta = \sigma_a b/\sigma_b$. If both z_1 and z_2 are large, this expression coincides with the log ratio

$$q(z_1, z_2) = \log(z_1 - a) - \log(z_2 - a) \tag{11}$$

However, $q(z_1, z_2)$ has a large, diverging variance for $z_i \to a$, a singularity at $z_i = a$, and is not defined in the range of real numbers for $z_i < a$. These unpleasant properties are important for applications: many genes are not expressed or only weakly expressed in some, but not all conditions of interest. That means, we need to compare conditions in which, for example, x_1 is large and x_2 is small. The log ratio (11) is not a useful quantity for this purpose, since the second term will wildly fluctuate and be sensitive to small errors in the estimation of the parameter a. In contrast, the statistic (10), which is called the *generalized log ratio* (Rocke and Durbin, 2003), is well defined everywhere and robust against small errors in a. It is always smaller in magnitude than the log ratio (see also Figure 1):

$$|h(z_1, z_2)| < |q(z_1, z_2)| \qquad \forall z_1, z_2. \tag{12}$$

The exponentiated value

$$\widehat{FC} = \exp(h(z_1, z_2)) \tag{13}$$

can be interpreted as a shrinkage estimator for the *fold change* x_1/x_2. It is more specific, that is, it leads to fewer false-positives in the detection of differentially expressed genes, than the naive estimator $(z_1 - a)/(z_2 - a)$ (Huber *et al*., 2002; Durbin *et al*., 2002).

5. Normalization and parameter estimation

The explanatory power of the model (6) can be greatly increased if we take into account the systematic dependence of its parameters on known experimental factors.

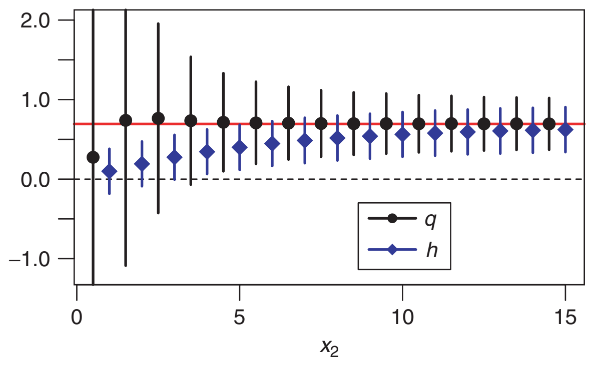


Figure 1 The shrinkage property of the generalized log ratio h. Blue diamonds and error bars correspond to mean and standard deviation of $h(z_1, z_2)$, cf. equation (10), black dots and error bars to $q(z_1, z_2)$, cf. equation (11). Data were generated according to equation (6) with $x_2 = 0.5, \ldots, 15$, $x_1 = 2x_2$, $a = 0$, $\sigma_a = 1$, $b = 1$, $\sigma_b = 0.1$. The horizontal line corresponds to the true log ratio $\log(2) \approx 0.693$. For intensities x_2 that are larger than about 10 times the additive noise level σ_a, h and q coincide. For smaller intensities, we can see a *variance-bias trade-off*: q has no bias but a huge variance, thus an estimate of the fold change based on a limited set of data can be arbitrarily off. In contrast, h keeps a constant variance – for the price of systematically underestimating the true fold change

This is often called *normalization*. A parametrization that captures the main factors that play a role in current experiments is

$$z_{ip} = a_{i,s(p)} + \varepsilon_{ip} + b_{i,s(p)} B_p x_{j(i),k(p)} \exp(\eta)_{ip} \tag{14}$$

Here, p indices the probes on the arrays and $k = k(p)$ the genes. Each probe is intended to represent exactly one gene, but one gene may be represented by several probes. B_p is the probe affinity of the pth probe (Li and Wong, 2001; Irizarry *et al*., 2003). i counts over the arrays and, if applicable, over the different dyes. $j = j(i)$ labels the biological conditions (e.g. normal/diseased). $a_{i,s(p)}$ and $b_{i,s(p)}$ are normalization offsets and scale factors that may be different for each i and possibly for different groups of probes $s = s(p)$. Probes may be grouped according to their physicochemical properties (Wu and Irizarry, 2004) or array manufacturing parameters such as print-tip (Dudoit *et al*., 2002) or spatial location. In the simplest case, $a_{i,s(p)} = a_i$ and $b_{i,s(p)} = b_i$ are the same for all probes on an array (Beißbarth *et al*., 2000). The noise terms ε and η are as above.

A method for the estimation of these parameters that uses the variance stabilizing transformation (10) was described by Huber *et al*. (2002) and Huber *et al*. (2003); software is available as an R package (Huber, 2004).

References

Beißbarth T, Fellenberg K, Brors B, Arribas-Prat R, Boer JM, Hauser NC, Scheideler M, Hoheisel JD, Schütz G, Poustka A, *et al*. (2000) Processing and quality control of DNA array hybridization data. *Bioinformatics*, **16**, 1014–1022.

Dudoit S, Yang YH, Speed TP and Callow MJ (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **97**, 77–87.

Durbin BP, Hardin JS, Hawkins DM and Rocke DM (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(Suppl 1), S105–S110.

Huber W (2004) *Vignette: Robust Calibration and Variance Stabilization with vsn*, The bioconductor project, **http://www.bioconductor.org**.

Huber W, von Heydebreck A, Sültmann H, Poustka A and Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl 1), S96–S104.

Huber W, von Heydebreck A, Sültmann H, Poustka A and Vingron M (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, **2**(1).

Ideker T, Thorsson V, Siegel AF and Hood LE (2000) Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805–818.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31–36.

Rocke DM and Durbin B (2001) A model for measurement error for gene expression arrays. *Journal of Computational Biology*, **8**, 557–569.

Rocke DM and Durbin B (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**(8), 966–972.

Wu Z and Irizarry RA (2004) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Proceedings of RECOMB 2004, San Diego.

Relevance networks

Atul J. Butte
Children's Hospital, Boston, MA, USA

There are several ways one can classify the algorithms used in the analysis of large RNA expression data sets (Article 53, **Statistical methods for gene expression analysis**, Volume 7). One way is to make the distinction between supervised and unsupervised methods. Supervised methods typically involve finding predictors of conditions or labels, such that unknown cases can be accurately predicted or labeled. Supervised methods are typically trained and tested using real cases, so metrics such as sensitivity, specificity, and accuracy can be evaluated. Differing from this, unsupervised methods involve finding characteristics of a data set without a priori labels. These algorithms may find helpful or significant patterns within a dataset, like the commonly applied hierarchical clustering (Eisen *et al*., 1998), as opposed to finding an efficient way to predict a "correct answer". Algorithms that determine a network of associations between genes generally fall in the category of unsupervised methods, and the method of relevance networks falls within this group of algorithms. Other network determination algorithms used in the analysis of genomic measurements include Boolean networks (Liang *et al*., 1998; Wuensche, 1998; Szallasi and Liang, 1998) and Bayesian networks (Heckerman, 1996; Friedman *et al*., 2000; *see also* Article 60, **Extracting networks from expression data**, Volume 7).

Relevance networks offer a method for constructing networks of similarity, with the principal advantages being the ability to (1) include features of more than one data type, (2) represent multiple connections between features, and (3) capture negative as well as positive correlations. Like hierarchical clustering, the algorithm begins by evaluating the similarity of features by comprehensively comparing all features with each other in a pairwise manner over the same cases. Several dissimilarity measures have been used in this methodology, including mutual information (Butte and Kohane, 2000) and the correlation coefficient (Butte *et al*., 2000). Those pairwise associations that do not exceed a threshold measure are then ignored, and the remaining associations are visualized as a graph, with features as nodes and associations as edges.

Formally, relevance networks are defined and implemented as a graph where n nodes $(g_1, g_2, \ldots, g_n)$ are connected by p sets of m edges each holding a score, where $m = \frac{n^2-n}{2}$. Each of the p sets of edges completely connects the n nodes, and each pair of nodes is connected by a single edge with a score. In practice, p is typically one, as only a single dissimilarity measure is typically used at a time. Each of the p sets of edges can represent a different dissimilarity measure

(such as Euclidean distance, correlation coefficient, or mutual information), and several of these may be simultaneously applied to the same set of n nodes. When used with microarray data, genes are represented as nodes, and edges are labeled with a real-valued score, which represents the strength of association between two genes.

Relevance networks are viewed by specifying a subset of nodes and edges specified by the values $t_1 \ldots t_p$, considered as thresholds, and the functions $f_1 \ldots f_p$, applying each threshold to the appropriate set of the p sets of edges. Each of the p sets of edges has a threshold t and a function f for applying that threshold to each of the edges in that set. Only those edges where the function returns true are kept in the subset. Typically, if the set of edges contains values between -1.0 and 1.0, then t is set to a value between 0 and 1, and f returns true for every edge if the absolute value score of the edge is greater than t. The semantics of the edges when the algorithm is used with microarray data is that an edge with a higher-positive or lower-negative score is more likely to represent a hypothesis of a biological relationship. Using thresholds essentially breaks apart the completely connected network graph into sets of smaller subgraphs. Typically, the threshold values are determined by studying the false discovery rates of various potential thresholds from a repeated analysis of permuted data sets.

We will now discuss some advantages of using relevance networks. First, relevance networks naturally handle negative interactions. One example of an important negative interaction in biology is that of *p53,* a tumor suppressor gene that is known to suppress expression of other genes. In such an interaction, increased expression of *p53* could be associated with decreased expression levels of other genes. Negative interactions are clearly different than of no interaction. Hierarchical clustering, as commonly computed with correlation coefficients and Euclidean distances, miss negatively correlated interactions; simultaneously visualizing both positive and negative associations in the same tree is not defined. By ignoring negative associations, the behavior of tumor suppressor genes and other negative transcriptional factors could be missed.

Second, while hierarchical clustering effectively clusters data of a single data type, mixing alternate measurements with expression measurements is not defined. For example, clustering a data set with both phenotypic and expression measurements will result in trees in which the leaves of phenotypic measurements are scattered throughout the tree; it is not clear how this would be useful. Because of this, two-dimensional dendrograms are typically used to mix two data types, where phenotypic and expression measurements are separately clustered (Ross *et al*., 2000), but this adds to the complexity. Compared to this, relevance networks can naturally mix phenotypic measurements with expression measurements, and direct hypotheses and links are provided between different type data (Butte *et al*., 2000). This is important as more experiments are designed integrating data across multiple genome-scale modalities.

There are other minor advantages to using relevance networks. Features can only be positioned in a single place in hierarchical clustering. Each gene is directly connected to the tree with only one stem. In reality, a transcription factor may be responsible for regulating the expression of multiple other genes, but hierarchical clustering may directly link that transcription factor only with the one gene it

most closely resembles in expression pattern. With relevance networks, if a gene is closely linked to few or many other genes, then each link is shown separately. This is also important when a gene is similar to two different groups of genes, or when a pharmaceutical agent displays activity similar to two different classes of compounds (Butte *et al*., 2000).

Relevance networks have appeared in several publications. We initially applied the method to clinical laboratory results, as a proxy for patient phenotypic data, for unsupervised medical knowledge discovery without a prior model or information, and found networks that were easily validated with known relationships in physiology and pathophysiology, as well as novel links (Butte and Kohane, 1999). We then applied the method to gene expression measurements in yeast, using mutual information as a dissimilarity measure (Butte and Kohane, 2000). To find putative functional relationships between pairs of genes and pharmaceuticals, we applied relevance networks to a pharmacogenomic data set. Specifically, we joined a database of baseline RNA expression levels measured from the NCI60 human cancer cell line panel with a database of cell line susceptibility measures against a panel of anticancer agents, to see how baseline gene expression levels across the cell lines correlated with the growth inhibition of these same cell lines from thousands of anticancer agents. At the threshold used, we found only one network containing an association between a gene and an anticancer agent (Butte *et al*., 2000).

Relevance networks were used to study resistance to anticancer drugs and expression of various genes in human hepatoma cells (Moriyama *et al*., 2003). The methods were used to identify a network of coregulated genes in mice in which dysferlin was knocked out, as a model for limb girdle muscular dystrophy (Lennon *et al*., 2003). Relevance networks across a large collection of 60 human microarray data sets helped in elucidating functionally related gene associations (Lee *et al*., 2004).

A Java-based software package to compute and display relevance networks, called RelNet, is publicly available from the Children's Hospital Informatics Program at **http://www.chip.org/relnet**.

References

Butte AJ and Kohane IS (1999) In *Fall Symposium, American Medical Informatics Association*, Lorenzi N (Ed.), Hanley and Belfus: Washington, pp. 711–715.

Butte AJ and Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, pp. 418–429.

Butte AJ, Tamayo P, Slonim D, Golub TR and Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(22), 12182–12186.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(25), 14863–14868.

Friedman N, Linial M, Nachman I and Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3–4), 601–620.

Heckerman D (1996) Bayesian networks for knowledge discovery. In: *Advances in Knowledge Discovery and Data Mining*, Fayyad UM, Piatetsky-Shapiro G, Symth P and Uthurusamy R (Eds.), The MIT Press: Cambridge, pp. 273–305.

Lee HK, Hsu AK, Sajdak J, Qin J and Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Research*, **14**(6), 1085–1094.

Lennon NJ, Kho A, Bacskai BJ, Perlmutter SL, Hyman BT and Brown RH Jr (2003) Dysferlin interacts with annexins A1 and A2 and mediates sarcolemmal wound-healing. *The Journal of Biological Chemistry*, **278**(50), 50466–50473.

Liang S, Fuhrman S and Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, pp. 18–29.

Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N and Omata M (2003) Relevance network between chemosensitivity and transcriptome in human hepatoma cells. *Molecular Cancer Therapeutics*, **2**, 199–205.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**(3), 227–235.

Szallasi Z and Liang S (1998) Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies. *Pacific Symposium on Biocomputing*, pp. 66–76.

Wuensche A (1998) Genomic regulation modeled as a network with basins of attraction. *Pacific Symposium on Biocomputing*, pp. 89–102.

Introductory Review

Protein structure analysis and prediction

William R. Taylor
National Institute for Medical Research, London, UK

1. Introduction

Of all biological macromolecules, proteins exhibit the greatest diversity of structure and function. Virtually all the processes that we associate with life, both basic and esoteric, are carried out or enabled by proteins. It follows that the more we know and understand about proteins, the more we know about Life – not only how it functions but also how it goes wrong in diseases.

Over the past 40 years, our knowledge of protein structure has gradually accumulated, initially through the efforts of X-ray crystallographers and, over the past 20 years, joined by NMR spectroscopists. The number of known structures is now considerable (*see* Article 37, **Signal peptides and protein localization prediction**, Volume 7), and what is apparent is that protein structures often reveal in a very direct way how things work. On the whole, proteins do not function through some abstract diffuse effects but behave more like little machines that trap, chop, bind, and join the substrates on which they act – including sugars, nucleic acids, and lipids but principally each other.

Despite having very direct actions, given just a sequence or even structure, it is almost impossible to figure out from first principles what a protein will do. For example, if we know the structure of an enzyme (at high resolution), to determine what it does would be necessary to locate a binding pocket, analyze its shape to see what might fit, and then see if any catalytic groups fall in the correct juxtaposition (or could be moved into place) to effect the catalysis of some reaction. Like the ab initio folding problem (*see* Article 66, ***Ab initio* structure prediction**, Volume 7), this is an almost impossible task but, again like the structure prediction problem, if we have a collection of structures about which we have a lot of information, and if our mystery protein looks like one of them, we can make some pretty good guesses.

The extent to which we can rely on these "guesses" can be evaluated in two qualitatively different directions. In one, we hope to retain an overall match to the whole protein, and this dimension extends from the trivial case where the two sequences share 100% agreement (the same protein!), through clear but decreasing levels of similarity, to the region (sometimes called the "twilight zone") where the level of similarity has become so little that we can no longer be sure that we have a match between equivalent proteins. The methods based on this approach

fall into the area of homology modeling or, for the more difficult problems, fold recognition and threading (*see* Article 70, **Modeling by homology**, Volume 7). In the other direction, the overall fold of the protein is retained less firmly and instead fragments of similarity are used to suggest local structure. These fragments may come from one or many proteins and can range in size from a whole domain (where they meet homology modeling), through supersecondary structure and motifs down to the smallest fragments of single secondary structure and turns. At this limit, the result is just a secondary structure prediction, with no indication given of the fold of the protein (*see* Article 76, **Secondary structure prediction**, Volume 7).

Underpinning our ability to extrapolate information in this way from well-characterized proteins to novel proteins are comparison methods. Through the comparative analysis of known structures, we have built up some understanding of the degree to which sequence and structure are conserved between related proteins. Depending on the extent of similarity, these can use just sequence data, structural data, or both. Their application is limited by two fundamental, and still poorly characterized, conditions: first, how can we tell that a similarity is significant and not due to chance and, second, given a significant similarity, to what extent can the function and properties of one protein be transferred to the other.

2. Protein sequence and structure analysis

The comparison of protein structures began as soon as there was more than one known structure: specifically, the structures of myoglobin and hemoglobin. From this early point, it was immediately apparent that equivalent protein structures (having the same fold) can result from sequences that are very different. This emphasizes the power of comparing protein structures to see beyond similarities that are not obvious from just sequence data. (For details of the methods used in both these approaches, see Eidhammer *et al*., 2004).

2.1. Protein structure comparison

2.1.1. Residue-based comparison methods

Over the years, many methods have been developed to compare protein structures (*see* Article 75, **Protein structure comparison**, Volume 7). Most of these are based on the idea of optimal superposition of equivalent atoms or residues and the almost universal measure of how close the structures compare is the root-mean-square of the distance between equivalenced pairs (RMS or RMS deviation). This measure is not ideal under all circumstances, and variations have been developed that use the comparison of local subsets to avoid difficulties associated with the relative movement of substructures between the two proteins being compared.

Computational problems also arise in the initial determination of the equivalence between positions. This is an alignment problem, and the methods used in sequence comparison have been adapted to this more difficult domain. Despite the rigor of the RMS measure, the alignment of residues based on structure is not well

determined, and most methods adopt an iterative heuristic approach (*see* Article 75, **Protein structure comparison**, Volume 7). Because of this inherent uncertainty, although less than in sequence comparison, a structure-based alignment is not necessarily unique and a range of alternative alignments of comparable score must be considered.

2.1.2. Fold-based comparison methods

For the most distantly related protein structures, the methods outlined above may be too focused on the details to allow an overall similarity to be recognized in the fold (or part of the fold) of the two proteins. To concentrate on this "bigger picture", the details of the structures can be neglected, either by (literally) smoothing out the chain or by reducing secondary structure elements to simple line segments. Many structure comparison methods have been adapted to use this more reduced representation, which also has the advantage of comparing fewer points, typically an order of magnitude less, giving a corresponding increase in computation time.

When secondary structures are reduced to line segments (or "sticks"), an even greater reduction in the level of detail can be achieved by collapsing each stick to a point: turning the 3D protein representation into a 2D diagram. The resulting 2D representation is not always unambiguous, especially for larger and multidomain proteins, but for a typical domain size of 200–300 residues, the resulting figure is reasonably unique. These projections of structure are often "tidied up" either automatically or manually by placing secondary structure symbols on a regular lattice. The resulting figures are called "topology cartoons", and have proved useful from the earliest investigations of protein folds, especially for abstract concepts like handedness (see Taylor and Aszódi, 2005 for examples).

2.2. Sequence structure correlations

From the viewpoint of modeling and prediction, any correlations between structural features and the sequence are potentially useful. As outlined above, these extend from the overall fit of the sequence onto the structure down to fragments or motifs. As with all protein sequence analysis, the uncertainty of features used in the analysis can be reduced by using a multiple sequence alignment (Eidhammer *et al*., 2004).

Sequence/structure correlations can embody either generic properties such as hydrophobic amino acids with burial in the structure or more specific properties such as a particular configuration of residues that form a motif associated with a particular structure or function. The former tend to be approached using statistical methods, of which one of the most general is the Artificial Neural Net (ANN), while the latter are associated more with sequence alignment or pattern matching methods. Both approaches can, of course, be applied together to provide mutual support.

2.2.1. Secondary structure and exposure prediction

In situations where there is weak sequence similarity between the protein of unknown structure (target) and the protein of known structure (template), it is

possible to use the predicted secondary structure to support the match, both as an overall fit and for specific details of the alignment. The state of the art in secondary structure prediction and exposure is currently obtained through training an ANN on multiple sequence alignments, an approach used both in the PsiPred and PHD programs (*see* Article 76, **Secondary structure prediction**, Volume 7). The accuracy of these methods is around 80+%, however, it must be remembered that the spread of accuracy is wide and any individual prediction may be susceptible to considerable error, including variation in the predicted end points up to the total omission of some secondary structure elements.

Both secondary structure and exposure predictions can be used to constrain the alignment of sequences through the inhibition of gaps in secondary structure and their encouragement in regions with high predicted exposure. An improved alignment can, in turn, lead to a better prediction, and the process can be iterated.

Predicted residue states provide a powerful guide for matching a sequence directly onto a structure. In this type of match, commonly called "threading" (*see* Article 70, **Modeling by homology**, Volume 7), the predicted states can be aligned directly with the observed states in the known structure.

2.2.2. Motif matching

Progressing toward more remote similarity between sequences leads to the situation where the target sequence has no significant overall sequence similarity with any known structure but shares a fragmentary similarity (or motif) with a superfamily or class of protein. Such motifs typically indicate a wide group of proteins ranging in specificity from those that derive from a clearly related but diverse family (such as a protein kinase) to those that are associated with a particular function such as calcium binding (as in the EF-hand motif).

The problems associated with modeling a target sequence based on a motif are different to those utilizing an overall alignment. The regions outside the motif regions may be much more difficult to align or fit to any template, possibly because they indeed have a different structure. This may mean that the resulting model must also be left incomplete or predicted using ab initio methods (*see* Article 66, ***Ab initio* structure prediction**, Volume 7).

There are collections of motifs available on the Web, the oldest of which (PROSITE) has a comprehensive collection of patterns associated with most protein families. Other more specific collections can also be found, which have an association with a specific family or function (such as transcription factors) or protein modifications (such as signal sequences).

3. Protein structure prediction

The ideal of predicting the structure of a protein directly from just its amino acid sequence has long been held as the "Holy Grail" of theoretical calculations on proteins. Despite great increases in computation power over the last 30 years, this

goal seems almost as remote as ever. For small (20–30 residues), fast-folding proteins, the native structure can sometimes be calculated but this is far from the performance required to tackle the typical globular protein of biological interest.

Current approaches to this problem have strayed from the "true path" into a variety of hybrid methods that incorporate components of matching against known structures along the lines outlined above for sequence/structure comparison. Principal among these are fragment-based methods that attempt to assemble model structures from fragments drawn from the collection of known structures. These must be put together in a physically reasonable way and evaluated for their packing contribution. Similar approaches have attempted to "grow" model structures as a constrained (almost) random walk, again being evaluated for packing as parts are added (*see* Article 66, ***Ab initio* structure prediction**, Volume 7).

As evaluated by the CASP assessment (*see* Article 73, **CASP**, Volume 7), some of these methods have had good results on a limited number of proteins, which although small in size, are still much larger than anything that can be attempted by "pure" ab initio methods. This approach could be more widely applied if there was a good method to predict domains (or subdomains) of proteins, allowing a divide-and-conquer approach. Domain prediction from sequence is still, however, a very difficult problem (*see* Article 68, **Protein domains**, Volume 7), to the extent that one domain prediction methods even relies on an ab initio prediction method.

4. Future directions

The above survey summarizes the state of research on computational studies of protein sequence and structure, concentrating on methods that attempt to extract and correlate patterns from these data. As new data are gathered, often at an impressive rate, the balance between the power and use of the different methods will also change.

4.1. Structure and sequence genomics

As genomes are sequenced from many different organisms, we are now almost certain to have a few or several phylogenetically diverse sequences for the same protein from different organisms. This places greater importance on multiple sequence alignment methods and on any method that uses a multiple sequence alignment to derive properties. This includes many of the "threading" (or fold recognition) programs mentioned above and also the secondary structure prediction methods.

Similar data-gathering initiatives are also being pursued on the structural front in an attempt to find, if not every protein structure, at least a representative of each fold class. With such data, the ability to construct good molecular models of specific proteins within the family based on the nearest representative will be important. With increasing coverage of "fold space", the separation between target and template will decrease, placing greater emphasis on standard modeling tools

rather than on the more adventurous modeling based on distant or fragmentary similarity or even ab initio prediction.

4.2. Complexes with membranes and nucleic acid

The bulk of the structures being determined by the structural genomics initiatives are relatively small-to-medium sized globular proteins (that tend to crystallize more easily). This neglects both large partially unstructured multidomain proteins and also integral membrane proteins. While the large proteins might be solved as separate domains, there is still a problem of finding the best points where they can be broken into parts and also the problem of reconstructing a full overall structure of how the parts might function together (perhaps as part of an even larger protein complex). For this, electron microscopy and atomic force microscopy can provide the larger picture.

For membrane proteins, methods of structure determination are improving but there is a vast number of proteins that from their sequence can be clearly identified as membrane associated. This places increased importance on the computational analysis of membrane proteins (*see* Article 65, **Analysis and prediction of membrane protein structure**, Volume 7), and although there are good prediction methods for transmembrane helical segments, the prediction of three-dimensional structure is limited.

As computational work shifts toward larger assemblies, this will inevitably involve nucleic acid components both in the form of protein/RNA particles (including the ribosome) and in the interaction and control of genomic DNA. The latter has been widely studied from the viewpoint of protein/DNA binding interactions (*see* Article 77, **DNA/protein modeling**, Volume 7) and will undoubtedly progress toward larger assemblies such as chromatin and DNA packing.

5. Conclusions

When sufficient sequence data is available to form a link to a protein of known structure, then even if the similarity is weak, there is a good chance that a reasonable molecular model can be constructed. While such results may not be of sufficient accuracy to allow detailed binding analysis of substrates or drugs, they are often useful in rationalizing the effects of known mutations or identifying potential binding sites for other proteins.

In addition to these practical benefits, the comparison and modeling of different proteins also provide insight into how one structure might evolve into another. In extrapolating this evolutionary approach, it is possible to gain an idea of how the whole range of protein structures may have evolved from some basic architectures or Forms. The degree to which the protein structures that we see today are a product of their history or are constrained by "laws of structure" is not yet clear but with increasing numbers of known structures and ever more powerful computational tools for comparison and prediction, some answers may soon emerge.

References

Eidhammer I, Jonassen I and Taylor WR (2004) *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*, John Wiley & Sons: London.

Taylor WR and Aszódi A (2005) *Protein Structure: Geometry, Topology, Classification and Symmetry: A Computational Analysis of Structure*, Institute of Physics.

Specialist Review

Analysis and prediction of membrane protein structure

Tina A. Eyre and Linda Partridge
University College London, London, UK

Janet M. Thornton
European Bioinformatics Institute, Cambridge, UK

1. Introduction

Transmembrane (TM) proteins are those that span the membrane lipid bilayer and they are estimated to comprise 20–50% of most genomes (Arkin *et al.*, 1997; Wallin and von Heijne, 1998). They are of great biological significance since they mediate most of the communication between cells and cellular compartments and contain many potential drug targets. Despite their abundance and importance, difficulties in the practical approaches to TM protein structure determination have led to few crystal structures being solved until recently. As a result, compared to water-soluble proteins, little is known about the structure of membrane proteins and, even today, only approximately 0.5% of the 23 792 structures in the Protein Data Bank (PDB Bernstein *et al.*, 1977) are TM proteins (Figure 1). An understanding of the structure of this class of protein is therefore desirable since it may facilitate structural modeling until high-throughput TM protein structure determination is possible.

As shown in Figure 2, the central region of the membrane consists of a highly hydrophobic 30-Å-thick lipid region formed by the hydrocarbon chains of the phospholipids. This is surrounded on either side by a 15-Å-thick region formed by the highly polar phospholipid head groups. The majority of TM proteins (known collectively as the α-bundle TM proteins) consist of several relatively hydrophobic α-helices, which span the membrane, connected by more polar loops. In contrast, the porin family of TM proteins, and possibly others, span the membrane via a barrel of β-strands and are not considered here. Proteins that span the membrane more than once are referred to as *polytopic*.

2. Characteristics of membrane protein structure

Since TM helices can now be readily identified from protein sequence (von Heijne and Gavel, 1988; Jayasinghe *et al.*, 2001; Jones *et al.*, 1994; Moller *et al.*, 2001),

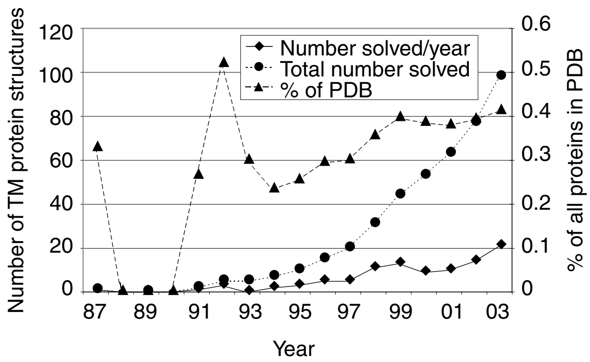


Figure 1 Plot showing the slow increase in the number of TM protein structures being solved each year and the proportion of the PDB made up by TM proteins. All structures are considered, including homologs

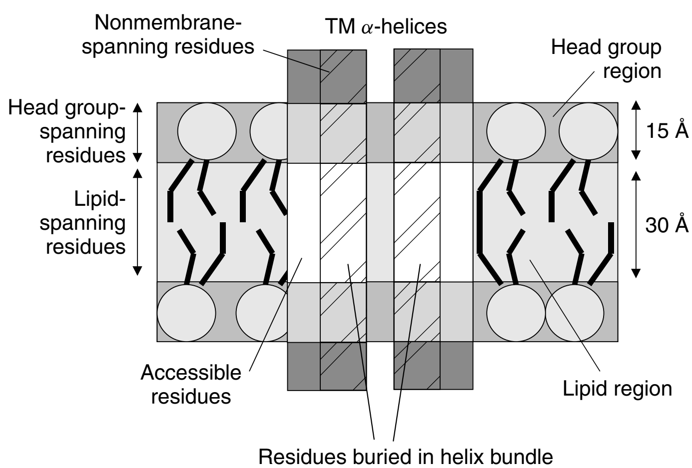


Figure 2 Schematic diagram showing the structure and thickness of a typical membrane and illustrating the different environments in which transmembrane residues may be found

the current challenge is to predict the 3D packing of these helices. Analysis of TM protein structures is therefore necessary to establish whether general rules can be identified to account for the folding of all membrane proteins and for use in model prediction. This is a particularly important goal, given the discrepancy between the number of TM proteins in the proteome and the number of structures known.

There are currently approximately 60 membrane protein structures known. Of these, only around 20 are nonhomologous polytopic α-helical proteins (Eyre *et al.*, 2004). Their structures are shown in Figure 3. This set is of interest since their multiple TM helices allow analysis of packing interactions, unbiased by inclusion of more than one member of each family. As a result, a number of characteristics of TM protein structure have been identified that are of particular relevance to

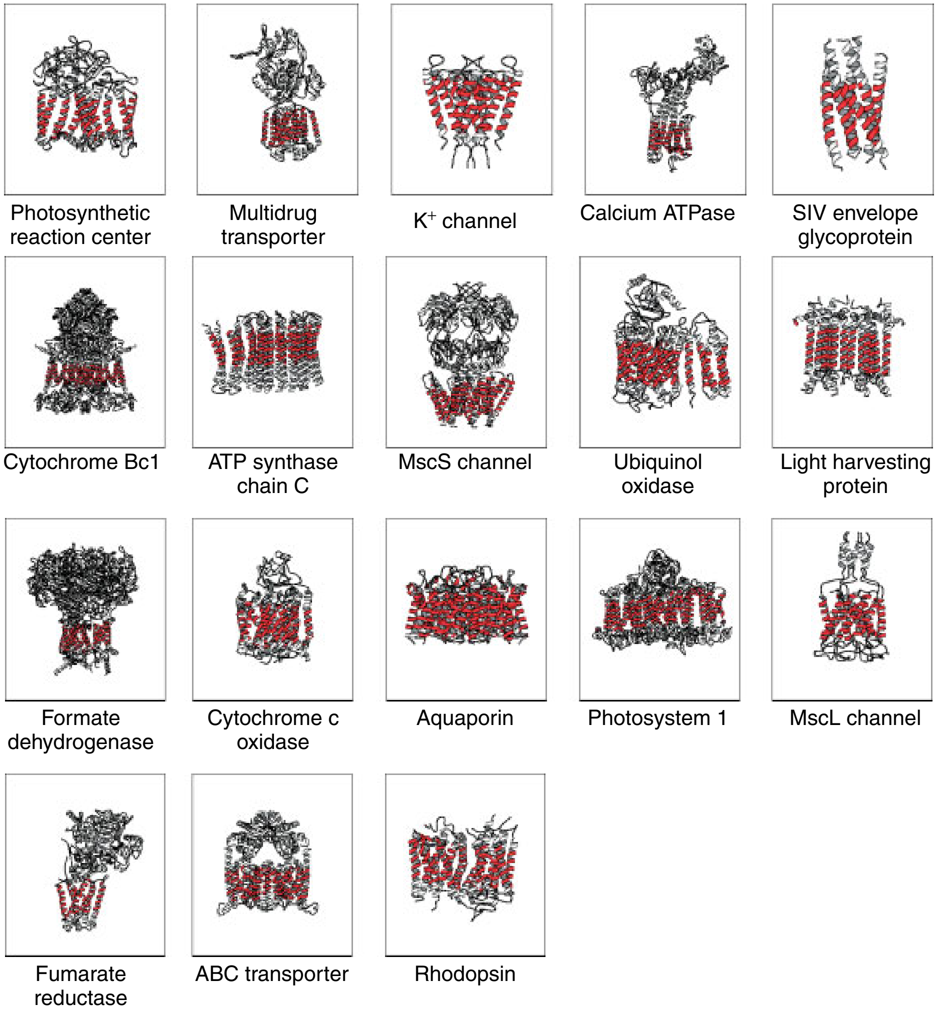


Figure 3 Currently available membrane protein structures showing the membrane lipid-spanning region in red

membrane protein modeling. These characteristics are described in the following sections.

3. Transmembrane helices are long, hydrophobic, and parallel

Figure 4 shows the preferences of each residue type for different regions of the membrane (Eyre *et al*., unpublished data). As would be expected from the hydrophobic nature of the bilayer, the lipid-spanning region of TM helices is enriched in hydrophobic amino acids compared to helices found in soluble proteins (Ulmschneider and Sansom, 2001) and to non-lipid-spanning regions of TM proteins (Eyre *et al*., 2004). This is the basis of many TM helix identification

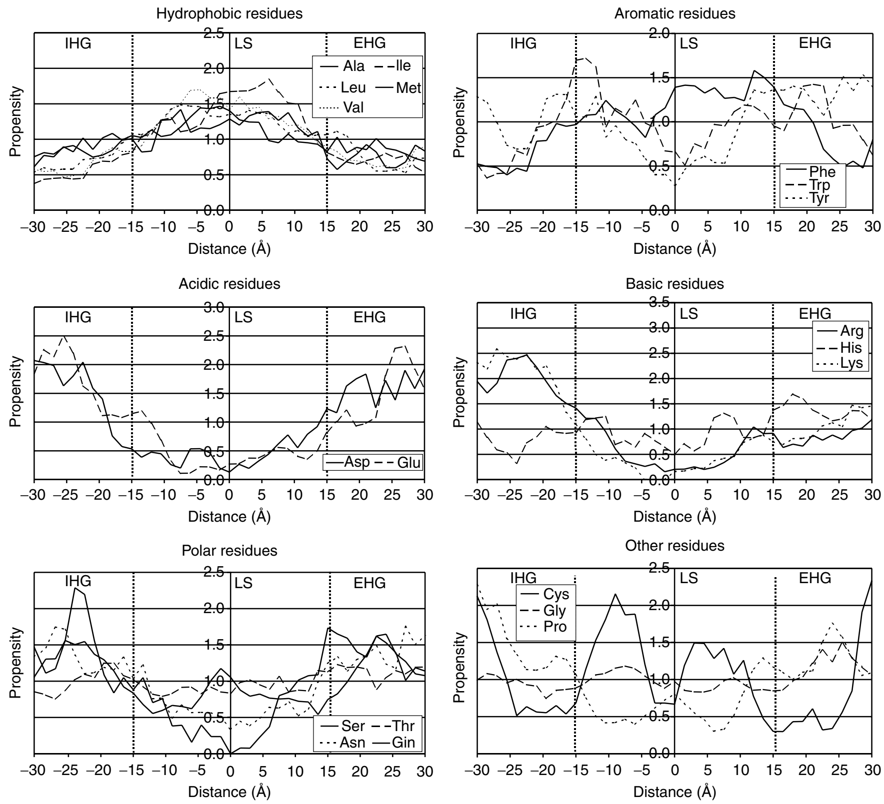


Figure 4 Plots showing the distribution of different residue types across the membrane-spanning region of 18 TM proteins. A propensity of greater than 1 indicates an enrichment in that area relative to other areas. IHG = intracellular head-group region, LS = lipid-spanning region, EHG = extracellular head-group region (Eyre *et al*., unpublished data)

algorithms. In addition, TM helices contain fewer charged and polar amino acids. Interestingly, however, the polar residues serine and threonine show no preference for either TM or non-TM regions. This is likely to be due to their ability to satisfy their hydrogen bonding requirements, even in the hydrophobic environment of the membrane lipid, by interaction with the previous backbone carbonyl oxygen of the same chain (Gray and Matthews, 1984). In addition, Figure 4 shows the 2.5-fold preference of the positively charged residues lysine and arginine for the intracellular head-group region, known as the *Positive Inside Rule*. This phenomenon is thought to be linked to the negative electrical potential that is found inside cells and may provide topological information during membrane insertion (von Heijne and Gavel, 1988).

Eyre *et al*. (2004) calculated various statistics concerning the number, length, and angle of α-helices and β-sheets in TM proteins and compared them to the corresponding values for soluble proteins. The results are broadly consistent with those obtained previously with smaller data sets (Bowie, 1997, Ulmschneider and Sansom, 2001). The average length of a TM helix is 23Å $\pm$ 4 residues (Eyre *et al*.,

2004). In contrast, the average length of both helices in soluble proteins and of non-TM helices in membrane proteins is only 9Å ± 5 residues (Eyre *et al*., 2004). The distributions of the lengths of non-TM α-helices and β-sheets between soluble and TM proteins are very similar. It can therefore be concluded that apart from the addition of several longer helices that span the membrane, TM proteins do not differ in their secondary structure composition from non-TM proteins.

The average angle of the helices to the membrane normal is $17 \pm 11^\circ$ (Eyre *et al*., 2004). This roughly parallel arrangement of TM helices is likely to facilitate the modeling of their packing in comparison to that for the helices in soluble proteins. However, there are large variations so that the degree of helix tilt cannot be ignored during modeling (Eyre *et al*., 2004).

Interestingly, while most TM helices are considerably longer than is required to span the membrane, the structures also contain a number of TM helices that only partially span the bilayer. The role of these helices remains unknown. Unfortunately, there appears to be no significant correlation between angle to the membrane normal and helix length (data not shown), excluding the possibility of estimating the tilt of a TM helix from its length during TM protein modeling.

4. There are differences between lipid-accessible and buried residues

The residues within the TM lipid-spanning regions of the helices can be considered to be found in one of the two main environments: either buried or accessible, as shown in Figure 2. The terms buried and accessible are used here to refer to the residues that are buried against other parts of the TM helix bundle or are exposed to membrane lipid, respectively, rather than the more common usage of buried or accessible relative to water. A small number of residues may also be found lining channels or cavities within the protein.

The different environments of lipid-exposed and buried residues result in different sequence characteristics. Specifically, lipid-accessible transmembrane helix residues are significantly more hydrophobic and less conserved than buried residues. In addition, lipid-accessible regions contain different residue types to buried regions (Eyre *et al*., 2004). It, therefore, is possible that membrane protein structure could be predicted by identifying the class to which different residues in the sequence are likely to belong. The characteristics of buried and lipid-accessible residues should be considered from this perspective.

Early analysis of the preferences of residues for buried or lipid-accessible positions was performed on the photosynthetic reaction center of *Rhodobacter sphaeroides* (Rees *et al*., 1989). Whereas in this TM protein the lipid-accessible residues were more hydrophobic than those buried (Rees *et al*., 1989), in soluble proteins, the buried residues were more hydrophobic (Chothia, 1976). Interestingly, the buried residues in the reaction center had a very similar hydrophobicity to residues buried within soluble proteins, as determined previously by Chothia (1976). Hence, soluble and TM proteins can be broadly considered to be internally stabilized in similar ways, with surface residues modified in order to facilitate solubility in the required medium. Certainly, TM proteins are not simply "inside-out"™

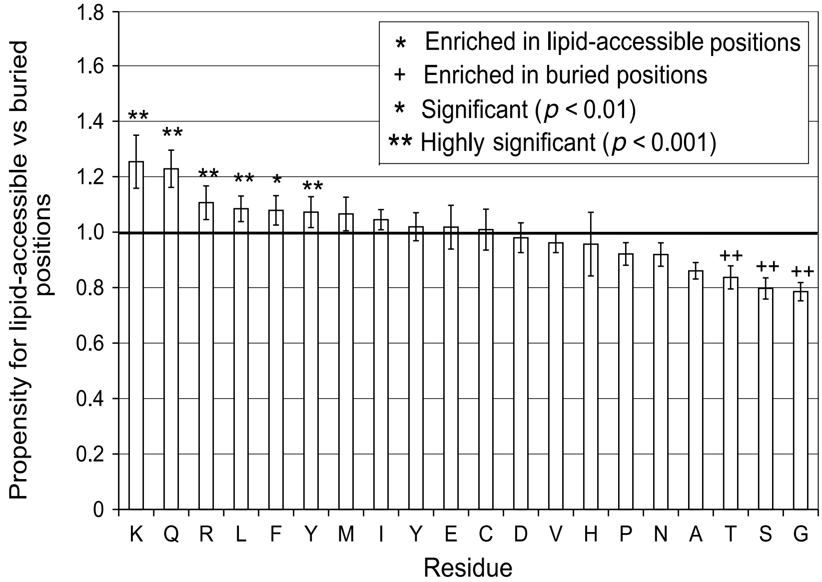


Figure 5 Residue preferences for lipid-accessible or buried positions in 18 TM proteins. Only residues in the lipid-spanning region are included. A value of greater than 50% indicates an enrichment in the lipid-accessible positions relative to buried positions. Similarly, a value of less than 50% indicates an enrichment in the buried positions relative to lipid-accessible positions. For details, see Eyre *et al*. (2004)

soluble proteins, with highly polar internal regions, as has been proposed (Stevens and Arkin, 1999). These results have been confirmed more recently using the 23 non-homologous TM protein structures that are now available (Eyre *et al*., 2004).

The work of Eyre *et al*. (2004) generally confirmed the results of the earlier smaller studies (Javadpour *et al*., 1999, Ulmschneider and Sansom, 2001). As shown in Figure 5, glycine, serine, and threonine are preferred in buried positions in TM proteins, whereas leucine, tryptophan, and phenylalanine prefer lipid-accessible positions. Hence, the general trend followed is that lipid-accessible residues tend to be more hydrophobic than those buried. Despite their hydrophilicity, the charged residues lysine and arginine, and the polar residue glutamine, strongly prefer lipid-accessible positions (Eyre *et al*., 2004). Possible explanations for this unexpected behavior are described below.

In addition to differences in residue content, lipid-accessible residues are significantly less conserved in terms of their sequence than buried residues are (Rees *et al*., 1989, Eyre *et al*., 2004). There is presumably strong selective pressure to conserve buried residues so that they continue to interact favorably with their interaction partners on other helices. There will be less selective pressure for lipid-exposed residues to be conserved since mutations in these residues are less likely to disrupt the folding, and hence the function, of the protein. As a result, throughout evolution, those individuals in which a protein contains mutations in a buried TM residue are likely to be less viable than those in which the protein contains mutations only in TM lipid-accessible residues. Hence, mutations in buried residues are less

likely to be inherited, and sequence variation between homologs will be much less at buried positions than at lipid-accessible positions.

5. Interactions between transmembrane helices differ from those between soluble protein helices

In soluble proteins, the helices of left-handed coiled-coils interact using a mechanism known as *knobs into holes*™ or *leucine zipper packing* (Crick, 1953). This has also been observed to be a common motif in TM helices (Langosch and Heringa 1998). The motif consists of a heptad repeat with hydrophobic residues, particularly leucine and valine, found at the positions buried between the helices. The side chains of these residues interact via extensive hydrophobic interactions, forming "rungs"™ connecting the two helix backbones.

Despite the similar hydrophobicity of internal residues in soluble and TM proteins, and the shared use of leucine zipper packing, there are several differences in the way their helices pack together. TM helices associate more closely than the helices in soluble proteins (Eilers *et al*., 2000). This may be because the residues buried between helices tend to have shorter side chains than those that are lipid accessible (Jiang and Vakser, 2000), allowing closer approach of the helix backbones. In support of this, several groups have identified an inverse relationship between the residue volume and its tendency to be found buried at a helix interface (Javadpour *et al*., 1999, Eilers *et al*., 2000, Adamian and Liang, 2001).

Another major difference between the packing of the two classes of proteins is the presence, in TM proteins alone, of many additional interhelix interactions between two polar residues, or between a polar and a charged residue. Adamian and Liang (2001) described the helix-packing interactions of TM proteins as involving a far more diverse set of polar residues than those of soluble proteins. While, in soluble proteins, they observed three types of interactions between polar groups on interacting helices, in TM proteins, they found 22 different types of polar interaction.

Hence, in addition to the leucine zipper packing, using hydrophobic residues, found in both classes of protein, TM proteins also make use of a wide variety of interactions between polar and charged residues. In TM proteins, the greater importance of polar interactions is to be expected because of their greater strength in the low dilectric lipid environment. Similarly, hydrophobic interactions are likely to be weaker between TM helices, and so they would be expected to play a less important role.

6. Glycine is important in transmembrane helix packing

A striking trend is the strong preference of glycine for buried positions (Javadpour *et al*., 1999, Eilers *et al*., 2000, Adamian and Liang, 2001, Ulmschneider and Sansom, 2001, Eyre *et al*., 2004). The role of glycine has been specifically studied by Javadpour *et al*. (1999). Since its side chain consists of a single H atom, glycine

allows adjacent helices to approach more closely than any other residue. Perhaps as a result of this, in cytochrome c oxidase, it is often found at helix crossings where, as Javadpour *et al*. (1999) have suggested, it may function as a "molecular notch" for orientating one helix against another. This motif also occurs in water-soluble proteins (Richmond and Richards, 1978).

The presence of a glycine residue also exposes the polar backbone atoms of its own chain, facilitating the formation of hydrogen bonds and dipolar interactions (Javadpour *et al*., 1999). These forces are likely to be particularly strong in the hydrophobic environment of the bilayer and hence may play a major role in protein stability. While interactions between the backbone atoms of different helices are rare in both classes of proteins, they seem to be more common in TM proteins than in soluble proteins (Adamian and Liang, 2001). Interactions with the backbone are most common where glycine residues interact.

7. Charged residues show unexpected behavior

Despite the hydrophobic nature of the bilayer, charged residues comprise approximately 8% of all TM lipid-spanning residues (Eyre *et al*., 2004). Surprisingly, charged residues are not necessarily buried, and lysine, arginine, and the polar residue glutamine are more common in lipid-accessible than in buried positions. In addition, as shown in Figure 6(a), when lipid-accessible, charged residues are not necessarily paired with another charge. Hence, the presence of opposing charges within a transmembrane region does not facilitate structure prediction by implying that these residues will be located adjacent to one another. Instead, they form hydrogen bonds with the phospholipid head groups of the membrane or with other residues of all kinds (Eyre *et al*., 2004).

Interestingly, in an investigation of lipid-accessible charged and polar residues, it was found that approximately one third form hydrogen bonds with residues on the same helix (intrahelix) rather than on another one (interhelix) (Eyre *et al*., 2004). These results are summarized in Figure 6(b). The importance of intrahelix hydrogen bonding had not been recognized previously. The role of these intrahelically bonded charged and polar residues is not understood. Since they are so common, there is likely to be some advantage to the presence of charged residues rather than hydrophobic ones at particular positions in the lipid-spanning region. Since they only rarely hydrogen-bond interhelically, this role is unlikely to be associated with maintaining the conformation of the protein by anchoring the TM helices relative to one another. Further work is needed to understand this phenomenon.

8. More structures are needed to fully characterize the properties of transmembrane pore-lining residues

Many TM proteins contain a channel or pore that traverses the membrane through which various substrates are transported. This transport is required for many processes of biological significance, such as the maintenance of cellular ion balance,

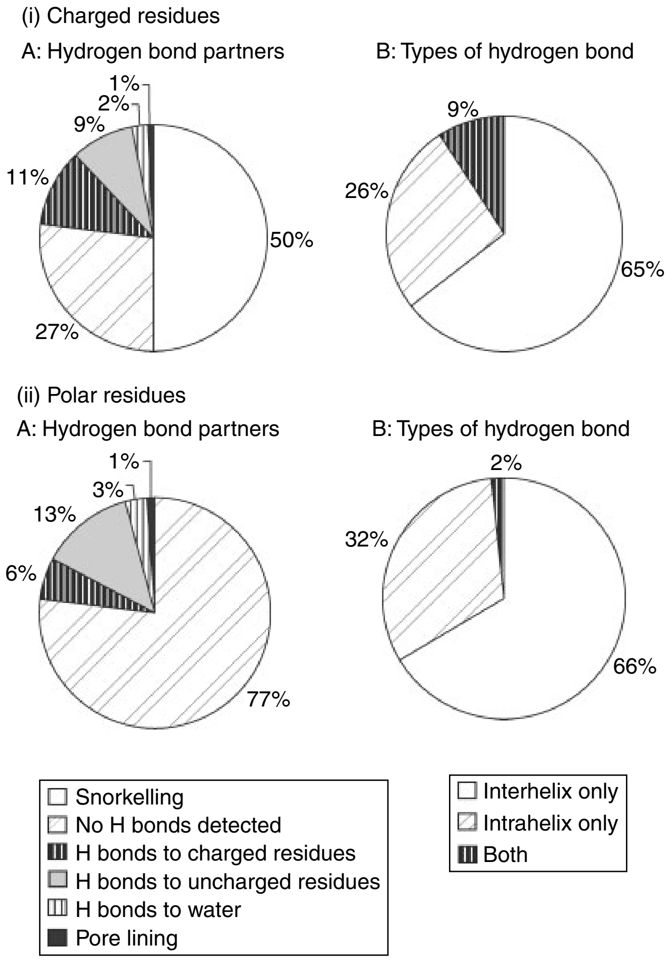


Figure 6 (a) Hydrogen (H) bonding partners and (b) types of hydrogen bonds for (i) the 679 observed lipid-accessible charged residues and (ii) the 2897 lipid-accessible polar residues in 18 TM proteins. For details, see Eyre *et al*. (2004)

nerve conduction, and the uptake of nutrients into cells. Hence, the residues that line these pores are of particular interest since they may give clues about how specificity is achieved.

Pore-lining residues are usually hydrophobic, consisting mainly of isoleucine, alanine, glycine, leucine, and valine (Eyre *et al*., 2004). This is perhaps unexpected, since it is often suggested that, in order to permit the passage of a polar species such as an ion, a channel must be lined with polar residues. It seems likely that hydrophobic residues lining a pore facilitate rapid translocation by minimizing interactions between the transported substrate and the channel walls.

It is difficult to distinguish pore lining from buried residues in terms of either residue type or evolutionary sequence conservation. The major difference between the two groups appears to be an enrichment of pore-lining residues with isoleucine

relative to buried regions (Eyre *et al*., 2004). It seems likely that more pore-containing membrane protein structures will be needed before these residues can be fully characterized and pore-lining helix faces identified from sequence.

9. The number of pore-lining helices and the diameter of the pore can be predicted

In addition to knowledge of pore-lining residues, it is also valuable to understand more general principles such as the size and shape of the pore. Eyre *et al*. (2004) identified linear correlations between

1. the total number of TM helices and the number of pore-lining helices and
2. the number of pore-lining helices and the pore diameter.

As a result, predictions can be made about the proportion of helices that contribute to lining a pore and the likely pore diameter simply from knowledge of the total number of helices. Since the total number of TM helices can now be predicted with relatively high accuracy, this information may facilitate the modeling of membrane proteins.

10. Membrane protein structure prediction

Until recently, there have been few attempts to perform 3D structure prediction on TM proteins. The earliest work in the area analyzed TM helices by comparing their characteristics to those of soluble proteins (Rees *et al*., 1989). This has lead to considerable advances in the automatic prediction of TM helix location and, particularly, through the application of the Positive Inside Rule, of TM protein topology (von Heijne and Gavel, 1988). It is now possible to identify TM helices and predict their topology with an accuracy of greater than 90% (Jayasinghe *et al*., 2001, Jones *et al*., 1994, Moller *et al*., 2001). The next challenge is to develop methods to successfully predict the 3D packing of these helices.

There have been many studies of TM protein structure, as summarized in the previous sections. However, this work has, until recently, been based on few membrane protein structures or simply on their primary sequence. As a consequence, many of the early attempts at TM protein 3D modeling suffered from the same limitations. The importance of the use of data derived from structure and not sequence for prediction is highlighted by the work of Ulmschneider and Sansom (2001), who compared the effectiveness of these two methods in discriminating between TM and non-TM spanning regions. The absence of structural TM data has also led to the use of data derived from soluble protein structures for TM protein modeling.

Despite some limited success (Taylor *et al*., 1994), ab-initio prediction of complete structures has proved difficult. On a more limited problem, Fleishman

and Ben-Tal (2002) used the knowledge of residue environment preferences to predict the likely arrangement of pairs of TM helices with greater than 70% accuracy. Ledesma *et al*. (2002) produced a model for uncoupling protein 1, using computational docking methods. However, in both cases, the pair-potentials used were derived from soluble proteins, shown by many to differ considerably from TM proteins (Rees *et al*., 1989, Eilers *et al*., 2000, Jiang and Vakser, 2000, Javadpour *et al*., 1999, Ulmschneider and Sansom, 2001, Adamian and Liang, 2001, Eyre *et al*., 2004). Pellegrini-Calace *et al*. (2003) used position-specific membrane potentials to perform simulated annealing for the modeling of TM protein structure. However, owing to high computational demands, it is likely to be a few years before the method becomes suitable for the modeling of large TM proteins. In addition, this method also relies upon the docking of protein fragments and residue potentials derived from soluble proteins, which is likely to limit its accuracy. Chen and Chen (2003) have successfully used three-stage Monte Carlo folding methods for small proteins. Other techniques have included homology modeling of proteins belonging to a family in which at least one structure is known (Giorgetti and Carloni, 2003, Nikiforovich *et al*., 2001). Unfortunately, only very few families contain a member with a known structure, so, currently, this method has limited applicability. However, it is likely to become the method of choice as more representative structures are determined.

Now that more TM protein structures have been solved, improved TM structure-based modeling studies can be attempted. There is a need for a simple computationally inexpensive method for the prediction of TM protein structure for all families. Given the considerable differences between the packing of TM and soluble protein helices, it should use data derived from TM proteins alone. Eyre *et al*. (2004b) used a hydrophobic-moment-based approach for prediction, using data derived from the analysis of TM proteins to pack the helices and score possible models. While the method was able to identify the buried face of TM helices with high accuracy, there were major problems with translating this single-helix data into a multiple-helix whole protein model. The work highlighted the importance of helix tilting and kinking in TM protein structure and suggested that more advanced modeling methods are needed that consider these factors. In addition, more channel-containing TM protein structures are needed to enable the characterization of pore-lining residues and their consequent identification from sequence.

11. Conclusions

Over the past few years, there has been an increase in the study of TM protein structure and modeling because of the emergence of a number of new TM protein structures. As with water-soluble proteins, ab initio structure prediction from sequence is challenging. Although we can identify the helices from sequence and predict the TM protein topology, packing pairs of helices by identifying buried and lipid-accessible TM helix faces is difficult. Further work and, crucially, more TM protein structures will be needed before 3D modeling of TM proteins is possible.

Acknowledgments

The authors gratefully acknowledge the financial support of the BBSRC and Wellcome Trust throughout this project.

Further reading

Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EF, Brice MD, Rodgers TR, Kennard O, Shimanouchi T and Tasumi M (1978) The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch Biochem Biophys*, **185**(2), 584–591.

References

Arkin IT, Brunger AT and Engelman DM (1997) Are there dominant membrane protein families with a given number of helices? *Proteins*, **28**, 465–466.

Adamian L and Liang J (2001) Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *Journal of Molecular Biology*, **311**, 891–907.

Bowie JU (1997) Helix packing in membrane proteins. *Journal of Molecular Biology*, **272**(5), 780–789.

Chen CM and Chen CC (2003) Computer simulations of membrane protein folding: structure and dynamics. *Biophysical Journal*, **84**(3), 1902–1908.

Chothia C (1976) The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology*, **105**, 1–12.

Crick FH (1953) The packing of alpha-helices: Simple coiled-coils. *Acta Crystallographica*, **6**, 689–697.

Eyre TA, Partridge L and Thornton JM (2004) Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural. *Protein Eng Des Sel*, in press.

Eilers M, Shekar SC, Shieh T, Smith SO and Fleming PJ (2000) Internal packing of helical membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5796–5801.

Fleishman SJ and Ben-Tal N (2002) A novel scoring function for predicting the alpha-helices. *Journal of Molecular Biology*, **321**, 363–378.

Giorgetti A and Carloni P (2003) Molecular modelling of ion channels: structural predictions. *Current Opinion in Chemical Biology*, **7**(1), 150–156.

Gray TM and Matthews BW (1984) Intrahelical hydrogen bonding of serine, threonine and relevance to membrane-bound proteins. *Journal of Molecular Biology*, **175**(1), 75–81.

Javadpour MM, Eilers M, Groesbeek M and Smith SO (1999) Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophysical Journal*, **77**, 1609–1618.

Jayasinghe S, Hristova K and White SH (2001) Energetics, stability, and prediction of transmembrane helices. *Journal of Molecular Biology*, **312**, 927–934.

Jones DT, Taylor WR and Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Jiang S and Vakser IA (2000) Side chains in transmembrane helices are shorter at helix-helix interfaces. *Proteins*, **40**, 429–435.

Langosch D and Heringa J (1998) Interaction of transmembrane helices by a coiled coils. *Proteins*, **31**(2), 150–159.

Ledesma A, de Lacoba MG, Arechaga I and Rial E (2002) Modelling the transmembrane arrangement of the considerations of the nucleotide-binding site. *Journal of Bioenergetics and Biomembranes*, **34**, 473–486.

Moller S, Croning MD and Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Nikiforovich GV, Galaktionov S, Balodis J and Marshall GR (2001) Novel approach to computer modeling of seven-helical case of bacteriorhodopsin. *Acta Biochimica Polonica*, **48**(1), 53–64.

Pellegrini-Calace M, Carotti A and Jones DT (2003) Folding in lipid membranes (FILM): A novel method structures. *Proteins*, **50**(4), 537–545.

Rees DC, DeAntonio L and Eisenberg D (1989) Hydrophobic organization of membrane proteins. *Science*, **245**(4917), 510–513.

Stevens TJ and Arkin IT (1999) Are membrane proteins 'inside-out'™ proteins? *Proteins*, **36**, 135–143.

Richmond TJ and Richards FM (1978) Packing of alpha-helices: geometrical constraints and contact areas. *J Mol Biol*, **119**(4), 537–555.

Taylor WR, Green NM and Jones DT (1994) A method for alpha-helical integral membrane protein fold prediction. *Proteins*, **18**, 281–294.

Ulmschneider MB and Sansom MS (2001) Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta*, **1512**, 1–14.

Wallin E and von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science*, **7**, 1029–1038.

Von Heijne G and Gavel Y (1988) Topogenic signals in integral membrane proteins. *European Journal of Biochemistry*, **174**, 671–678.

Specialist Review

Ab initio structure prediction

David Baker
University of Washington, Seattle, WA, USA

William R. Taylor
National Institute for Medical Research, London, UK

1. Introduction

The prediction of protein structure from sequence remains one of the most fundamental problems at the interface of physics, chemistry, and biology. Besides its biological importance, its continual fascination derives from the simplicity with which it can be stated: given a protein sequence, what is its three-dimensional structure? This is sometimes called the protein folding problem since one way to approach it is to simulate how a protein would fold within the cell but, as we will see below, this is not the only way to tackle the problem.

The simple statement of the problem hides some assumptions. Specifically, that there will be a well-defined structure to predict and if there is, under what conditions it can be obtained. The latter may include more than the obvious temperature and pH, and could involve posttranslational modifications (cleavage, phosphorylation) or additional cofactors not specified by the sequence (such as a haem group) or the fact that the protein may only be structured in a multimeric state or in combination with another protein. Despite all these caveats, many experiments, beginning with the classic work of Anfinsen (1973), have shown that to a first approximation, the three-dimensional structure of the protein is indeed encoded in its sequence.

The fact that a sequence will generate a well-defined structure does not mean that each sequence will produce a unique structure, and it can be expected that many sequences will fold to similar structures. This property provides an alternative route to structure prediction in which the basic structure of one protein is just modified to accommodate some changes in sequence. For the more obvious similarities, this approach is referred to as "modeling by homology" or for the more difficult problems as the "empirical approach" to structure prediction, in contrast to the more fundamental folding simulations that hope to generate structure ab initio.

Modeling and ab initio prediction effectively lie at extreme ends of a range defined by the size of unit taken as given. In modeling, this is the full-sized protein, whereas in ab initio prediction only the chemical structure of the peptide is assumed. Between these extremes lie many approaches that make use of intermediate-sized fragments, including secondary structure elements or subassemblies of secondary

structure (sometimes called “supersecondary structure”). The latter also range in size from small functional units (called “motifs”) to full protein sized domains. Methods using these intermediate-sized chunks will be referred to generally as “fragment-based” methods.

2. Ab initio structure prediction

The direct simulation approach to protein structure prediction is also the oldest. This is not unexpected since, unlike the empirical or fragment-based methods, it requires no examples to copy. Perhaps the first successful result using this approach was the prediction of the α-helix by Pauling and Corey (1951) before even the first protein structure was known. Pauling, of course, used a physical model but to tackle larger problems an equivalent model encoded in a computer program is preferable.

2.1. *Computational approaches*

2.1.1. Spin bonds or move atoms?

Protein structure can be encoded as a virtual model in two distinct ways. Perhaps the more obvious approach is to have fixed length atomic bonds that can rotate where necessary. The alternative is to have independent atoms that are restrained (In describing molecular calculations, it is usual that the terms “restraint” and “constraint” take specific meanings. “Restraint” implies an imposed tendency toward a specific value, while “constraint” implies that the quantity has a fixed value.) to maintain required bond lengths.

The rotation approach, while more intuitive, is difficult to animate since a rotation around any bond will displace all of the following parts of the protein in an unnatural way. To rotate a bond in a chain such that there are only local perturbations in conformation is quite a difficult computational problem, especially when the effect of nonlocal interactions (involving remote parts of the sequence) must be taken into consideration.

The alternative is to displace each atom individually under the combined influence of all local and nonlocal effects. If the local contributions are dominated by a restraint to maintain connecting bond lengths and the angle between these bonds, then the overall result can replicate a bond rotation.

2.1.2. Slide or jump?

In both the above approaches, each rotation or move is considered in turn and the structure is displaced slightly from its current conformation into a new conformation. A choice must be made for the size of each displacement that equates with the time period between moves and the temperature of the simulation. It might be thought desirable to have the system evolve as smoothly as possible, which requires having a very fine time step. However, this directly increases the computation time of the simulation and may prevent any significant structural changes being observed in the available time.

Two approaches can be adopted to avoid this. One is to make the program run faster using a simplified model (considered in the following section) and the other is to greatly increase the size of the displacements and ignore the continuity between them. This is more like a random sampling of possible conformations and is called "Monte Carlo" (after the spiritual home of random number generation). The random component means that, even if the old conformation had a low energy, the new conformation might have a high energy. To avoid these undesirable states, a cutoff on the energy of the conformation can be applied, but only sometimes. Whether to accept or reject a conformation is determined with a probability that is a function of its energy. So some high-energy conformations are occasionally accepted, just as some low-energy conformations are occasionally rejected.

2.1.3. Simulate or select?

Given a finite computational resource, it would be possible to run a single simulation for a long time or to run multiple simulations for a shorter time. In the former, if the protein folded quickly, time would be wasted: either the correct conformation is obtained, in which case further simulation is not informative or the structure may be trapped in an incorrect conformation from which it cannot escape. Running a larger number of simulations can increase the range of conformations that are generated, giving an idea of how reproducible (and maybe how correct) any one particular solution might be.

This strategy can also be taken to an extreme using no, or minimal dynamics. Given an algorithm to generate compact protein-like structures, then a large number of these can be created and simply evaluated to identify those that have low energy. This is somewhat like a one-jump Monte Carlo approach, and to have any hope of getting a good prediction in one-shot, a very large number of candidate structures need to be generated.

Rather than rely on getting a reasonable structure on the first pass, a number of good structures can be selected and used as a base from which to generate further variations. This is rather like Monte Carlo in parallel and can be implemented using an optimization strategy called the Genetic Algorithm (GA) (Dandekar and Argos, 1994; Petersen and Taylor, 2003).

2.1.4. Out of hyperspace

The methods described above have in common the need to generate an initial starting conformation for the chain. The specification of this can affect either how the molecule folds or, more directly, its actual conformation in the one-shot approaches. Clearly, those that generate most models will be less sensitive to the initial configuration, however, there is an alternative approach based just on distances that avoids it completely.

Not unreasonably, this method is known as Distance Geometry (DG), and is widely used in calculating structures based on NMR data, which come in the form of interatomic distances. When the distances are uncertain (as all predicted distances will be to some extent), then the method requires careful handling. A reasonably

robust implementation has been devised using a combination of DG with refinement in high-dimensional space (Aszódi and Taylor, 1994).

Although the method avoids a specific starting conformation, it can still be run multiple times (with different random displacements on the starting distance set) to give an idea of the uniqueness or spread of the final models. Because DG has no constraint to follow a folding pathway, it is prone to generate tangled or even knotted structures. In folding simulations, there is a bias to form sequentially local contacts, which reflect what is seen in known structures. Without this bias, DG may waste time looking in biologically infertile regions of conformational space.

2.1.5. Statistical model

Like DG, there are some interesting models for proteins do not even have a tangible representation of the structure at all. These tend to be statistical in nature, dealing with the probabilities of multiple options. These include: the self-consistent field (SCF) method (Finkelstein and Reva, 1991), associative memory Hamiltonians (AMH) (Friedrichs *et al*., 1991), and artificial neural nets (ANN) (Holbrook *et al*., 1993). All of these iteratively refine a set of probabilities. The SCF method takes a physical framework and refines the probabilities of residues to lie in particular places, converging to a self-consistent result where each point is occupied by only one residue. AMH and ANN methods are related techniques in which the probabilities are represented by a set of multidimensional vectors and a set of weights in a network, respectively. Both these methods are more typically "trained" on native proteins and are used to evaluate models generated by other means.

2.2. Molecular models

The preceding section has concentrated on the mechanics of manipulating structures with little reference to the molecular models themselves. In this section, some representations of proteins are described, and almost all of them can be used in combination with any of the computational strategy described above.

The geometric specification of the models and the description of their internal interactions range from the very simple – single sphere per residue with only two residue types (hydrophobic/hydrophilic) – to complex quantum mechanical representations. The degree of model complexity directly affects the computation time required, limiting the length of simulation or the number of models that can be calculated or the size of protein/peptide that can be studied.

In this section, we will ignore quantum calculations (which are typically only performed on very short peptides). We will also give scant consideration to the molecular force calculations used in conventional molecular dynamics calculations (*see* Article 74, **Molecular simulations in structure prediction**, Volume 7). Instead, we will concentrate on simplified molecular representations and their simplified "forces" as it is only with these that a large number of calculations can be performed on proteins of even average size.

2.2.1. Lollypop models

One of the earliest attempts to simulate the folding of a small protein used a simplified model of the chain with two points (spheres) per residue. One representing the α-carbon position and the other representing a residue centroid (Levitt, 1976). The resulting model looked rather like a string of lollypops from which the model derived its unofficial name. An even simpler lollypop model can also be based on just a single point per residue if it is assumed that the side-chain centroid can be approximated by an extension of the line bisecting the angle formed by three consecutive α-carbon atoms.

While computational speed can be gained using a simplified model, this is directly traded against accuracy in the packing and other interactions of residues. In the lollypop model, flexibility is lost in the side chain, which may make the difference between packing a large residue in the core and excluding it. In addition, the specificity of exact hydrogen bonding is also lost, which means that only general preferences for residues to interact can be expressed.

It can be argued that the overall structure of proteins is dominated by the formation of a hydrophobic core and if this can be captured using only simple residue properties, then protein-like structures will be generated. However, for any given sequence, there are many ways of creating a good hydrophobic core and without the fine specificity of side-chain interactions, it is probably impossible to discriminate one from the other.

2.2.2. Secondary structure models

Secondary structure elements in proteins fix a number of residues into a relatively rigid substructure. If these elements can be predicted, then the conformational freedom of the chain becomes greatly reduced. Over the years, many methods have tried to exploit this advantage.

One approach to this is to move the secondary structures as rigid bodies and “dock” them together to build up a structure. This is computationally feasible for α-helices, which can behave as independent units. However, it is difficult to apply the approach to β-structure. An alternative approach is to take a predefined general framework over which the secondary structure elements can be allocated in a combinatorial way and the resulting structures filtered.

3. Empirical approach

3.1. Combinatorial secondary structure packing

In the late 1970s, the protein folding problem, tackled using direct folding simulation, was generally considered to be computationally intractable due to the large number of atomic interactions that occur even in a small protein. To find a way around this problem, an approach was developed that did not calculate the interactions from first principles (ab initio) but instead compared the model interactions to what had already been observed in known protein structures. This approach was based on a detailed analysis of secondary structure packing in proteins (Cohen

et al., 1981) and began what was later to be called the empirical approach to protein structure prediction. The approach was applied to simplified models constructed from secondary structure elements. By considering many different secondary structure packing combinations, it was possible to select a limited number of structures that included the true (native) protein structure (Cohen *et al*., 1980). Although initially applied only to a subclass of proteins, the method was later extended to other classes (Cohen *et al*., 1982; Taylor, 1991), including membrane proteins (Taylor *et al*., 1994).

These early applications of the empirical approach relied on finding clusters of hydrophobic residues that packed to form a hydrophobic core; however, the same principle can be applied at other levels from pairs of residues, through clusters of residues to the assessment of an intact structure.

3.2. Fragment-based approach

3.2.1. The ROSETTA program

The most successful approach to ab initio structure prediction over the past 6 years based on the CASP tests (*see* Article 73, **CASP**, Volume 7) is called ROSETTA. Rosetta is based on a model of folding in which short segments of the protein chain flicker between different local structures, consistent with their local sequence, and folding to the native state occurs when these local segments are oriented such that low free-energy interactions are made throughout the protein (Simons *et al*., 1997). In simulating this process, it is assumed that the ensemble of local structures sampled by a given sequence segment during folding is roughly approximated by the distribution of local structures sampled by that sequence segment in native protein structures. A list of possible conformations is extracted from experimental structures for each nine residue segments of the chain, and protein tertiary structures are assembled by searching through the combinations of these short fragments for conformations that have buried hydrophobic residues, paired beta strands, and other low free-energy features of native proteins. This strategy resolves some of the typical problems with both the conformational search and the free energy function: The search is greatly accelerated as switching between different possible local structures can occur in a single Monte Carlo step, and less demands are placed on the free energy function since local interactions are accounted for in the fragment libraries.

To generate structures consistent with both the local and nonlocal interactions responsible for protein stability, 3 and 9 residue fragments of known structures with local sequences similar to the target sequence were assembled into complete tertiary structures using a Monte Carlo simulated annealing procedure (Simons *et al*., 1997). The scoring function used in the simulated annealing procedure consists of sequence-dependent terms representing hydrophobic burial and specific pair interactions such as electrostatics and disulfide bonding and sequence-independent terms representing hard sphere packing, alpha-helix and beta-strand packing, and the collection of beta-strands in beta-sheets (Simons *et al*., 1999).

For each of 21 small, ab initio targets, 1200 final structures were constructed, each the result of 100 000 attempted fragment substitutions. The five structures submitted for the CASP III experiment were chosen from the approximately 25

structures with the lowest scores in the broadest minima assessed through the number of structural neighbors (Shortle *et al*., 1998).

The results were encouraging: highlights of the predictions include a 99-residue segment for MarA with an rmsd of 6.4 A to the native structure, a 95-residue (full length) prediction for the EH2 domain of EPS15 with an rmsd of 6.0 A, a 75-residue segment of DNAB helicase with an rmsd of 4.7 Å, and a 67-residue segment of ribosomal protein L30 with an rmsd of 3.8 Å. Rosetta predictions in the CASP4, CASP5 and CASP6 international blind prediction experiments were even more promising. These results suggest that ab initio methods may soon become useful for low-resolution structure prediction for proteins that lack a close homolog of known structure.

3.2.2. Refining rough structures

The challenge for the past 4 years has been to refine the low-resolution models into accurate high resolution structures using an all-atom representation and a physically realistic potential. This is necessary because of the critical role steric packing interactions play in protein structure and stability; without explicitly modeling these interactions, the best that can be hoped for is that one of a several plausible models with the hydrophobic residues buried, strand paired, and so on. Full-atom refinement is very challenging because the landscape being searched is very large and very bumpy (even 1-Å shifts from the native structure, presumably the global minimum, can produce huge energy spikes if atoms overlap one another).

A protocol for full-atom refinement has been developed with the Rosetta software package that uses a Monte Carlo with minimization (MCM) optimization protocol for the backbone coordinates, a Monte Carlo search through side-chain rotamer possibilities for the side-chain coordinates, and a potential dominated by a Lennard Jones potential, an implicit solvation model, and an explicit orientation-dependent hydrogen bonding term. This full-atom refinement protocol has been used successfully in protein–protein docking calculations (Gray, 2003) (results from the recent CAPRI V challenge are particularly notable), and supplemented with sequence optimization methodology, to design with atomic accuracy a sequence that adopts a novel fold (Kuhlman, 2003).

An exciting result from CASP6 is that for the first time in a blind test, high-resolution refinement significantly improved a low-resolution model, leading to a prediction of almost atomic level accuracy (1.5 A rmsd) in which features of the native core packing arrangement are partially reproduced. Over the next few years, it is hoped that with continued improvements in methodology and increases in computer power ab initio methods can consistently produce atomically accurate models for small proteins, which would greatly increase their application to real-world problems.

References

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Aszódi A and Taylor WR (1994) Folding polypeptide α-carbon backbones by distance geometry methods. *Biopolymers*, **34**, 489–506.

Cohen FE, Sternberg MJE and Taylor WR (1980) Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature*, **285**, 378–382.

Cohen FE, Sternberg MJE and Taylor WR (1981) Analysis of the tertiary structure of protein β-sheet sandwiches. *Journal of Molecular Biology*, **148**, 253–272.

Cohen FE, Sternberg MJE and Taylor WR (1982) Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *Journal of Molecular Biology*, **156**, 821–862.

Dandekar T and Argos P (1994) Folding the main-chain of small proteins with the genetic algorithm. *Journal of Molecular Biology*, **5**, 637–645.

Finkelstein AV and Reva BA (1991) A search for the most stable folds of protein chains. *Nature*, **351**, 497–499.

Friedrichs MS, Goldstein RA and Wolynes PG (1991) Generalised protein tertiary structure recognition using associative memory Hamiltonians. *Journal of Molecular Biology*, **222**, 1013–1034.

Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA and Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, **331**(1), 281–299.

Holbrook SR, Dubchak I and Kim SH (1993) Probe – a computer-program employing an integrated neural-network approach to protein-structure prediction. *Biotechniques*, **14**, 984–989.

Kuhlman B, Dantas G, Ireton G, Stoddard B, Varani G and Baker D (2003) De novo design of a novel protein fold with atomic level accuracy. *Science*, **302**(5649), 1364–1368.

Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, **104**, 59–107.

Pauling L and Corey RB (1951) The structure of synthetic polypeptides. *Proceedings of the National Academy of Sciences of the United States of America*, **37**, 241–250.

Petersen K and Taylor WR (2003) Modelling zinc-binding proteins with GADGET: Genetic Algorithm and Distance Geometry for Exploring Topology. *Journal of Molecular Biology*, **325**, 1039–1059.

Shortle D, Simons KT and Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1158–1162.

Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225.

Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C and Baker D (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Genetics*, **34**, 82–95.

Taylor WR (1991) Towards protein tertiary fold prediction using distance and motif constraints. *Protein Engineering*, **4**, 853–870.

Taylor WR, Jones DT and Green NM (1994) A method for α-helical integral membrane protein fold prediction. *Proteins: Structure, Function, and Genetics*, **18**, 281–294.

Specialist Review

Score functions for structure prediction

Richard A. Goldstein
National Institute for Medical Research, London, UK

1. Introduction

Protein structure prediction is often divided into homology modeling (*see* Article 66, ***Ab initio* structure prediction**, Volume 7, Article 70, **Modeling by homology**, Volume 7, and Article 72, **Threading for protein-fold recognition**, Volume 7), fold recognition, and ab initio prediction, depending upon the availability of proteins of related sequence and structure and the resolution required tasks. For all three, one needs (1) a space of possible structures, (2) a method for searching through this space, and (3) a criterion for selecting one (or more) of those structures as the prediction. The selection criterion is often in the form (or can be put in the form) of a "score function", a method for producing a score quantifying how well the target sequence matches a given structure. The form of the score function depends upon the structure space and the search strategy; particularly important is the manner in which the possible structures are represented. Is the structure represented to atomic-level accuracy, or by a highly simplified model? Are all angles and bonds free to vary, or are there only discrete possibilities? Simplified, discretized representations (e.g., representing the structure as a string of beads, each residue represented by one bead) make the search problem more tractable (we do not have to worry about searching over all conformations of the side chains), but limits the details that can be used in the score function. Simpler score functions, however, tend to be faster to compute, allowing for more exhaustive and far-ranging search methods. In addition, simpler score functions may be easier to create, as described below. In this review, we discuss various approaches to generate score functions. While our emphasis will be on such functions, some discussion of the protein representation cannot be avoided.

The representation and the corresponding score function depends upon the nature of the problem one is trying to solve; homology modeling, where we are interested in atomic-level detail based on the structure of a related protein, obviously requires an atomic-resolution representation, while we might be satisfied for a rougher representation for an ab initio prediction. We do not have to be consistent in our

methods, and can use more detailed representations as the prediction goes through successive stages (Feig *et al*., 2000). As homology modeling is represented in a separate chapter (*see* Article 70, **Modeling by homology**, Volume 7), we focus in this review on scoring functions for fold recognition and ab initio prediction. There are approaches that, while formally "score functions", will not be considered here. Feed-forward neural networks, for instance, represent a particular class of highly flexible functions over an extremely large dimensional space, with a potentially huge number of adjustable parameters generally determined with a data-fitting algorithm. We will restrict our focus to score functions that, given a sequence and a three-dimensional structure, calculate the score of the sequence when it assumes the given structure.

The construction of a score function involves two distinct steps: the choice of a "functional form" for the function, generally involving a (potentially large) set of adjustable parameters, and the determination of those parameters. The choice of functional form is especially tied up with the representation of the protein structure – obviously a score function that depends upon atomic interactions requires an atomic-resolution representation.

The most obvious approach to generating a score function is based on the so-called thermodynamic hypothesis that states that proteins folded into the conformation of lowest free energy (Anfinsen, 1973; Govindarajan and Goldstein, 1998). If this hypothesis is correct, the value of the free energy would be the perfect score function. Such a score function would also provide the most insight into the physicochemical interactions that govern the folding process. Beyond questions about the validity of the thermodynamic hypothesis, there are additional challenges generating sufficiently accurate free-energy values (entropic terms, including the hydrophobic effect, and conformational and vibrational entropy, are especially difficult) sufficiently rapidly to perform an effective conformational search.

Given the problems with generating an accurate free-energy function, alternative strategies have arisen. We have one great resource at our disposal: we know many right answers. Embodied in the protein structure databases are many sequences with their corresponding structures. A number of different "knowledge-based" techniques have been developed in which a representative "training set" of known protein structures is assembled, a score function is produced on the basis of an analysis of this training set, and the method evaluated on a separate "test set" of proteins (Tanaka and Scheraga, 1976). There have been three approaches to constructing score functions based on the set of known protein structures; thermodynamics, statistics, and machine learning. In the following, we describe some of the functional forms that have been used for score functions, and outline how these approaches are used to determine the values of the various parameters.

There is a problem of sign convention. Researchers who think of score functions generally associate high-scoring structures as likely predictions. Those who think in terms of approximate free-energy functions associate structures with low energy as likely predictions. I will use the term "score function" to imply a high score means a better structure. This score function will then be the negative of any energy function.

2. Functional forms

The score function determines a numerical score for how well a target protein (with sequence $\{A_i\}$) is compatible with a structure (with coordinates $\{\vec{x}_i\}$). As described above, the value of the free energy is often used as a basis for the (negative of the) score function. Often, this function takes a form similar to those in use in molecular modeling (*see* Article 74, **Molecular simulations in structure prediction**, Volume 7):

$$\text{Score} = -\Delta G = -\left(\Delta G_{\text{covalent}} + \Delta G_{\text{VdW}} + \Delta G_{\text{electrostatic}} + \Delta G_{\text{hydrophobic}} + \cdots\right) \tag{1}$$

where $\Delta G_{\text{covalent}}$, ΔG_{VdW}, $\Delta G_{\text{electrostatic}}$, and $\Delta G_{\text{hydrophobic}}$ represent the contributions to the free energy from the covalent bond geometry (including bond stretching and both bond and dihedral angles), Van der Waals interactions, electrostatic interactions (including hydrogen bonds), and interactions with the solvent respectively. Other terms can be added as appropriate. Each of these have standard forms, with parameters that have been adjusted based on ab initio quantum mechanical calculations and/or observed properties of various molecules, although the parameters could also be derived using the knowledge-based methods described below. The sequence of the target protein is included by atomic composition and the covalent connections between the atoms making up the residues. Of these terms, interaction with the solvent is potentially the most problematic. Issues of computational resources generally precludes the use of an atomic-resolution description of the solvent, so simpler approximations are made. For instance, calculations can be made of atomic surface area exposed to the solvent, and $\Delta G_{\text{hydrophobic}}$ calculated on the basis of observed linear relationships between hydrophobic free energy and this exposed area. It is also possible to use more complicated models of the interaction with solvent, such as the Generalized Born (GB) continuum solvent model (Still *et al*., 1990). It is possible to add and change terms to include observed properties of proteins, for instance, including terms that make the protein more likely to be "protein-like", that is, to be compact and to have the right proportions of local, secondary, and supersecondary structure.

While we are not able to use these realistic forms for searching over a large range of possible structures, progress has been made in using these potentials for identifying the correct structure from amongst a large set of decoy structures (Dominy and Brooks, 2002; Felts *et al*., 2002), as well as refining more approximate structures (Feig *et al*., 2000). Another advantage is that such a method, if successful, would provide much information about how protein structures are determined. In addition, these potentials do not need a database of known structures, which means that they are immune to problems of limited and biased protein structure databases.

The physicality causes one further limitation. Often, information is available that is not physicochemical in nature, for instance, from experimental observations. While this information can be put in the form of an energy function and in this way added to the prediction scheme, such an approach must be done in an ad hoc way,

as there is no natural approach to include nonphysical interactions in a realistic physical scheme.

As described above, this function would require an all-atom representation of the protein conformation as well as a complicated search strategy. For these reasons, simpler functional forms are often used, especially for fold recognition and ab initio prediction. One of the more common encodes the contacts made between different residues in the protein, with a general form similar to

$$\text{Score} = -\sum_{<ij>} \Delta G_{\text{contact}}\left(A_i, A_j\right) u(\vec{x}_i, \vec{x}_j) \tag{2}$$

where $\Delta G_{\text{contact}}(A_i, A_j)$ represents the effective free energy of interaction that would occur between amino acid A_i and amino acid type A_j, $u(\vec{x}_i, \vec{x}_j)$ is 1 if and only if residues i and j "interact" according to some criterion, and the sum is over all pairs of residues. Numerous criteria have been used for interaction, generally involving fudicial points on the two residues (e.g., α-carbons, β-carbons, center of mass of the side chain, or any heavy atoms) within some specified distance. Similar to this contact potential is a distance-dependent potential, where the contact energy is dependent upon the distance between the interacting residues.

Studies have shown that pairwise contact energies such as described in equation (2) are insufficient to correctly predict protein structures (Vendruscolo and Domany, 1998). Indeed, more complicated terms can be included, such as adding higher-order terms or anisotropic interactions (Buchete *et al*., 2003). For instance, three-body terms would involve sums over all triplets of amino acid with a contact potential depending upon all three residues. The contact potential $\Delta G_{\text{contact}}(A_i, A_j)$ can also be made more complex by including information regarding neighboring residues, predicted or structure-based secondary structure or solvent accessibility, and so on. Finally, care must be taken in situations in which contacts are not pairwise-additive, such as cysteine–cysteine interactions – it is possible for one cysteine to make contact with multiple other cysteines, so that the presence of one disulfide bond must eliminate the possibility of further disulfide bonds with these residues.

The above scoring function, while certainly simpler to calculate than the molecular-modeling score function, still necessitates a complicated strategy for searching over possible conformations. A simpler strategy is to consider the propensity for various residues to be in various structural "contexts" (Bowie *et al*., 1991):

$$\text{Score} = -\sum_{i} \Delta G_{\text{context}}\left(A_i, C_i\right) \tag{3}$$

where $\Delta G_{\text{context}}(A_i, C_i)$ represents the free energy of residue type A_i when in context C_i, where C_i can include secondary structure, solvent accessibility, character of nearby residues, and so on. The advantage of such an expression is that it is much more amenable to fast searches based on dynamic programming algorithms, as is used in sequence–sequence alignments.

There has also been some work on extremely simple functional forms, attempting to capture the most basic observed properties of proteins, such as compactness, the

placement of hydrophobic residues, and the presence of local structures (Huang *et al*., 1995; Yue and Dill, 1996; Srinivasan and Rose, 1995). One advantage of such schemes is that they are often designed in association with a rapid method for exploring conformation space, and have the potential to yield comprehensible insights into the nature of the relationship between sequence and structure.

All of these score functions are to some extent "physical" functions in that they try to model the free energy of the protein in some simplified manner. There are other score functions that owe their construction more to machine learning than to physical chemistry, generating a score function that explicitly includes information about proteins of known structure. An example of this is neural-network-based score functions. For instance, one can construct a score function that seeks to model how well the target protein's structure matches that of "template" proteins through the use of an "Associated Memory Hamiltonian" (Friedrichs *et al*., 1991).

There are many methods of introducing supplementary information, besides the sequence of the target protein, into the score function through the inclusion of additional terms. Experimentally determined distance constraints, for instance, are a staple of molecular modeling techniques, and have found use in structure prediction. Similarly, predictions made from other methods, such as neural-network-predicted secondary structure, can be included. Hybrid methods exist that combine the sequence-structure score functions discussed here with traditional sequence–sequence methods, increasing the score when the target protein has some degree of sequence similarity with the template protein. Finally, the availability of large numbers of evolutionarily related (homologous) proteins, likely forming highly similar structures, can provide much information. This has been used by searching for a single structure that is compatible with the set of homologs (Goldstein *et al*., 1993), encoding the conservation and variation at each location as an element of the target protein sequence (Defay and Cohen, 1996), and looking directly at how compatible the evolutionary substitutions observed at each position are with the local structure at that position (Goldman *et al*., 1996).

3. Determination of parameters

The functional forms described above all involve particular parameters that must be determined. The atomic-resolution formulation can rely on matching calculations and observations for a wide range of molecules. Most of the other forms, however, require an analysis of the known database of protein structures for the determination of these parameters. As described above, three approaches are in common use: thermodynamic, statistical, and machine learning. These approaches will be discussed in turn below.

3.1. *Thermodynamic approach*

The thermodynamic approach seeks to determine an appropriate free-energy function from the databases of known protein structure. The archetypal example is "potentials of mean force". The Boltzmann distribution states that at equilibrium

at a given temperature, the probability $f(S_i)$ that a system will be in state S_i with free energy $G(S_i)$ is given by

$$f\,(S_i) = \frac{\exp\left(-G\,(S_i)\,/kT\right)}{Z} \tag{4}$$

with k equal to Boltzmann's constant, T as the temperature, and Z is the partition function equal to $Z = \sum_{i'} \exp\left(-G(S_{i'})/kT\right)$, where the sum is over all possible states of the system. The above can be written as

$$G\,(S_i) = -kT \log\left[f\,(S_i)\right] - kT \log Z \tag{5}$$

that is, the free energy of a state can be calculated (up to an additive constant) on the basis of the probability of the system being in that state at thermal equilibrium.

Although the Boltzmann distribution refers to fluctuations in a *single system* or ensemble of *identical systems* at thermal equilibrium, we now extrapolate to the distribution of a *particular interaction* s_i in the ensemble of a large set of *different systems*, corresponding to various protein structures in their *ground* state, that is, *ignoring* any fluctuations. We compare the probability $f(s_i)$ that this interaction is found in this set of proteins compared with the probability $f_{\text{ref}}(s_i)$ that it occurs in a prespecified reference state. Computing $\Delta G(s_i)$, the difference between $G(s_i)$ in the set of proteins and $G_{\text{ref}}(s_i)$ in the reference state (both given by equation (5) with the appropriate value of $f(s_i)$) and neglecting the partition function, yields (Sippl, 1990)

$$\Delta G\,(s_i) = RT \log \frac{f\,(s_i)}{f_{\text{ref}}(s_i)} \tag{6}$$

In reality, we do not have the probabilities $f(s_i)$ but only the observed frequencies. We can correct for the limited nature of our data by considering (Sippl, 1990)

$$f\,(s_i) = \frac{1}{1+m\sigma} f_{\text{background}}\,(s_i) + \frac{m\sigma}{1+m\sigma} f_{\text{obs}}\,(s_i) \tag{7}$$

where m is number of observations, $f_{\text{obs}}(s_i)$ is the observed frequency of interaction s_i, $f_{\text{background}}(s_i)$ is a suitably constructed estimated frequency of s_i neglecting the amino acid dependence of the interaction, and σ is a parameter representing our confidence in the observed data. As can be seen, $f(s_i)$ changes from $f_{\text{background}}(s_i)$ to $f_{\text{obs}}(s_i)$ as more data becomes available.

This approach can be used on almost all of the different functional forms, although most uses of this method has involved pairwise interactions, including contact and distance-dependent potentials, at either residue or atomic resolution.

Although the derivation as stated lacks rigor, this approach has been justified using either a "quasi-chemical approximation" (Miyazawa and Jernigan, 1985) or evolutionary considerations (Finkelstein *et al*., 1995). (According to the latter justification, the "temperature" in equation (6) is not the temperature of the organism, or the protein when crystallized, or anything directly related to a physical temperature.) There are, however, still issues that remain. First, it is not clear what

is the reference state used for calculating $f_{\mathrm{ref}}(s_i)$. Should it represent the target protein in its unfolded state, or a set of proteins in random structures, or something else? One of the main differences between different implementations of this approach is in the choice of this reference state. More importantly, in the potential of mean force approach, the strength of each interaction is influenced by all other interactions with which it is correlated. (This is, for instance, how hydrophobicity can be included in the pair contact term; contacts between residues are negatively correlated with contacts between those residues and solvent, and thus the hydrophobicity of a residue will increase its apparent attraction for other residues, even in the absence of interactions between these residues.) Any correlation between the various interactions in the potential lead to double-counting; each interaction will be included explicitly in the potential as well as implicitly, through its effect on the strength of the various other interactions. As a result, this approach requires that the various terms in the potential not only be additive, but that their occurrences be *uncorrelated* in the ensemble of database proteins. This is unlikely to be true. For instance, residues found in β-sheets are more likely to be in contact with each other, simply because β-sheets have a tendency to be clustered in the interior of the protein. More subtly, it has been conjectured that interactions in the folded state are highly correlated as a way of assisting the folding process (Go, 1983; Bryngelson and Wolynes, 1987). Correlations increase as the energy function becomes more complicated. Approaches to examine the validity of the thermodynamic approach have yielded mixed results (Skolnick *et al*., 1997; Thomas and Dill, 1996). Despite the doubts raised, this method has been popular and has yielded success in fold recognition. One positive aspect has been that, by ignoring correlations between terms, the parameters of the model are rather well behaved. This might be less of an advantage as more data becomes available.

3.2. *Statistical approach*

The statistical approach is similar in form to the thermodynamic approach, but differs in underlying basis, resembling the formulations used for sequence–sequence alignments.

The easiest example is with the simple single-body terms described in equation (3), describing the propensity $\Delta G_{\mathrm{context}}(A_i, C_i)$ for any amino acid A_i to be in a given structural context C_i. This is used to evaluate the probability that sequence $\{A_i\}$ would be found in structure $\{\vec{x}_i\}$. This is in fact a conditional probability $P(\{\vec{x}_i\} \mid \{A_i\})$ that a protein would form structure $\{\vec{x}_i\}$ given that its sequence is $\{A_i\}$, where the probability is meant as a measure of our ignorance and uncertainty. The predicted structure can be seen as, in this sense, the structure with the highest value of this probability.

The conditional probability can be inverted using Bayes law to

$$P(\{\vec{x}_i\} \mid \{A_i\}) = \frac{P(\{A_i\} \mid \{\vec{x}_i\})\ P(\vec{x}_i)}{P(\{A_i\})} \tag{8}$$

Often, the "prior" term $P(\{\vec{x}_i\})$ representing the a priori probability of this structure is neglected, although this is not necessary. (In fact, it is possible to use this term to

include additional understanding of the nature of protein structures, such as they are compact and have significant secondary structure (Simons *et al*., 1997).) Neglecting this term yields the result that we want to find the structure $\{\vec{x}_i\}$ that maximizes $P(\{A_i\} \mid \{\vec{x}_i\})$, that is, the conditional probability that the given sequence would result from that structure, divided by $P(\{A_i\})$, the probability that the structure would occur in a random structure. If we now assume that the choice of amino acid at location is independent, that is, the probability of any amino acid at a given location depends only upon the nature (or context) of that location C_i, we can factor $P(\{A_i\} \mid \{\vec{x}_i\})$ over all locations. Doing the same thing with $P(\{A_i\})$ (which is, anyway, independent of structure) yields

$$P(\{A_i\} \mid \{\vec{x}_i\}) \approx \prod_i \frac{P(A_i|C_i)}{P(A_i)} \tag{9}$$

or

$$\log(P(\{A_i\} \mid \{\vec{x}_i\})) \approx \sum_i \log\left(\frac{P(A_i|C_i)}{P(A_i)}\right) \tag{10}$$

which is the same as the functional form of equation (3), where

$$\Delta G_{\text{context}}(A_i, C_i) = \log\left(\frac{P(A_i|C_i)}{P(A_i)}\right) \tag{11}$$

This approach can be applied to a wide variety of functional forms, and in the case of pair potentials yields a result similar to the thermodynamic approach. In contrast to the thermodynamic approach, however, the assumptions are more clearly stated, meaning that they can be more easily evaluated and, if necessary, corrected (Simons *et al*., 1997).

3.3. *Machine-learning approach*

The thermodynamics and statistics approaches described above are based on statistics of correct structures, including random structures as, at most, normalizing parameters. In contrast, one can treat protein structure prediction as a discrimination problem, finding a scoring function that separates correct from incorrect structures. In this manner, the “best” scoring function is one that maximizes this discriminatory power. This leads to a more symmetric handling of correct and random structures.

This approach has proceeded through a number of different directions. One method has been to maximize the minimum “gap” between the worst scoring correct structure and the best scoring incorrect structure, or alternatively, by maximizing the minimum gap adjusted for the difference between the correct and incorrect structure, so that more dissimilar structures have worse scores (Tobi and Elber, 2000; Maiorov and Crippen, 1992; Micheletti *et al*., 2001; Crippen, 1991; Vendruscolo and Domany, 1998).

There is a concern whether the “maximin” method may be overly sensitive to the (possibly extremely small) subset of high-scoring incorrect structures. An

alternative approach has been to look at the distributions of scores generated with correct and incorrect structures, model these distributions with a functional form, and try to optimize the characteristics of this particular functional form. The first approach with this method, based on ideas from spin-glass theory, described the distribution of scores for incorrect structures as a Gaussian distribution, and then maximized the so-called Z-score, defined as the difference between the correct score and the average of the random scores, divided by the standard deviation of the random scores (Goldstein *et al.*, 1992b; Goldstein *et al.*, 1992a). As the potential was derived for a large ensemble of proteins of known structure, both the difference between correct and average and the standard deviation of the random scores were averaged over the training set, to allow a closed-form solution for the optimal score function. Other methods have been described, such as maximizing the harmonic mean of the Z-score (Mirny and Shakhnovich, 1996) or maximizing the probability that the highest-scoring structure is correct (Chiu and Goldstein, 1998a). This optimization procedure has been used for a wide range of different functional forms, from contact potentials (Goldstein *et al.*, 1992b; Mirny and Shakhnovich, 1996) to Associative Memory Hamiltonians (Goldstein *et al.*, 1992a) to more realistic energy force-fields (Liwo *et al.*, 1997). This method can be used on the statistically different problem of inverse folding, finding those sequences most likely to fold into a given structure. There is no a priori reason to believe that the optimal score function for inverse folding would be the same as for fold prediction, and test cases with lattice proteins indicate that such optimized score functions can actually outperform the perfectly accurate free energy function (Chiu and Goldstein, 1998b).

One general problem with such optimization schemes is that they optimize what they are asked to optimize, not necessarily what is desired to be optimized. For instance, in Z-score optimization, increasing the scores of the most unlikely, low-scoring structures can increase the Z-score by shrinking the standard deviation of the scores of random structures, even if this compromises the gap between the correct structures and the best, high-scoring structures (Chiu and Goldstein, 2000). This problem has been approached by iterative methods to more realistically model the higher-ranking (lower energy) states (Koretke *et al.*, 1996; Hao and Scheraga, 1996). In contrast to the statistical and thermodynamic approaches, it is also less obvious how to adjust for the limited statistics available in current sets of protein structures. An additional problem is that such discrimination methods distinguish between a "correct" fold and a diverse set of incorrect folds. This latter set is highly heterogeneous, and may be poorly represented by a single distribution.

A major advantage of the optimization scheme, however, is that it determines the "best" set of parameters, given all of the correlations between the different terms in the set of known structures. As a result, it does not make the same assumptions of independence required in the thermodynamic and statistical approaches. This makes the scheme ideally suited to the mixing of heterogeneous types of information, including data regarding predicted local structure, experimental observations, information of related sequences, and so on. The optimization strategy automatically weights the various physical and nonphysical terms in the appropriate way, accounting for the correlations between them.

4. Future directions

Previous work in the development of these score functions has involved a balance between realism and simplicity. Physics-based potentials provide the ultimate realism, and, in principle, the most accuracy and insight, but the requirement for an atomic-level description of the protein conformation has restricted their use to homology-modeling and limited searches over decoy structures. More complicated knowledge-based functions require more data to fix the various parameters, as well as more computer resources for performing conformational searches. This has meant that much of the progress in fold recognition and ab initio folding has involved highly simplified score functions. This may change rapidly as computational resources become more available and the set of proteins of known structure expands. This may enable more complicated score functions with many more adjustable parameters. The rise in complexity is likely to extend to more than score functions. It will be necessary to deal with the complexity of the data, including various biases and correlations, as well as with the nonadditive nature of the various interactions. This work will not proceed in isolation. In particular, ways of preprocessing the sequence data, including predictions of local structure, analysis of related sequences, and so on, is likely to increasingly provide relevant information for the scoring function.

References

Anfinsen C (1973) Principles that govern the folding of a protein chain. *Science*, **181**, 223–230.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bryngelson JD and Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 7524–7528.

Buchete N-V, Straub J and Thirumalai D (2003) Anisotropic coarse-grained statistical potentials improve the ability to identify native-like protein structures. *The Journal of Chemical Physics*, **118**, 76588–77671.

Chiu TL and Goldstein RA (1998a) Optimizing energy potentials for success in protein tertiary structure prediction. *Folding & Design*, **3**, 223–228.

Chiu TL and Goldstein RA (1998b) Optimizing potentials for the inverse protein folding problem. *Protein Engineering*, **11**, 749–752.

Chiu TL and Goldstein RA (2000) How to generate improved potentials for protein tertiary structure prediction: a lattice model study. *Proteins*, **41**, 157–163.

Crippen GM (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232–4237.

Defay TR and Cohen FE (1996) Multiple sequence information for threading algorithms. *Journal of Molecular Biology*, **262**, 314–323.

Dominy BN and Brooks CLI (2002) Identifying native-like protein structures using physics-based potentials. *Journal of Computational Chemistry*, **23**, 147–160.

Feig M, Rotkiewicz P, Kolinski A, Skolnick J and Brooks CL III (2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins*, **41**, 86–97.

Felts AK, Gallicchio E, Wallqvist A and Levy RM (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized born solvent model. *Proteins*, **48**, 404–422.

Finkelstein AV, Gutin AM and Badretdinov AY (1995) Boltzmann-like statistics of protein architectures. *Sub-cellular Biochemistry*, **24**, 1–26.

Friedrichs MS, Goldstein RA and Wolynes PG (1991) Generalized protein tertiary structure recognition using associative memory Hamiltonians. *Journal of Molecular Biology*, **222**, 1013–1034.

Go N (1983) Theoretical studies of protein folding. *Annual Review of Biophysics and Bioengineering*, **12**, 183–210.

Goldman N, Thorne JL and Jones DT (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Evolution*, **263**, 196–208.

Goldstein RA, Katzenellenbogen JA, Luthey-Schulten ZA, Seielstad DA and Wolynes PG (1993) Three dimensional model of the hormone binding domains of steroid receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 9949–9953.

Goldstein RA, Luthey-Schulten ZA and Wolynes PG (1992a) Optimal protein folding codes from spin glass theory. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 4918–4922.

Goldstein RA, Luthey-Schulten ZA and Wolynes PG (1992b) Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 9029–9033.

Govindarajan S and Goldstein RA (1998) On the thermodynamic hypothesis of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5545–5549.

Hao MH and Scheraga HA (1996) How optimization of potential functions affects protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 4984–4989.

Huang ES, Subbiah S and Levitt M (1995) Recognizing native folds by the arrangement of hydrophobic and polar residues. *Journal of Molecular Biology*, **252**, 709–720.

Koretke KK, Luthey-Schulten Z and Wolynes PG (1996) Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Science*, **5**, 1043–1059.

Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S and Scheraga HA (1997) A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *Journal of Computational Chemistry*, **18**, 874–887.

Maiorov VN and Crippen GM (1992) Contact potential that recognizes the correct folding of globular proteins. *Journal of Molecular Biology*, **227**, 876–888.

Micheletti C, Seno F, Banavar JR and Maritan A (2001) Learning effective amino acid interactions through iterative stochastic techniques. *Proteins*, **42**, 422–431.

Mirny LA and Shakhnovich EI (1996) How to derive a protein folding potential? A new approach to an old problem. *Journal of Molecular Biology*, **264**, 1164–1179.

Miyazawa S and Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225.

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. *Journal of Molecular Biology*, **213**, 859–883.

Skolnick J, Jaroszewski L, Kolinski A and Godzik A (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Science*, **6**, 1–13.

Srinivasan R and Rose GD (1995) LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins*, **22**, 81–99.

Still W, Tempczky A, Hawley R and Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, **112**, 6127–6129.

Tanaka S and Scheraga HA (1976) Medium and long-range interaction parameters between amino acids for predicting three-dimensional structure of proteins. *Macromolecules*, **9**, 945–950.

Thomas PD and Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? *Journal of Molecular Biology*, **257**, 457–469.

Tobi D and Elber R (2000) Distance-dependent pair potential for protein folding: results from linear optimization. *Proteins*, **41**, 40–46.

Vendruscolo M and Domany E (1998) Pairwise contact potentials are unsuitable for protein folding. *The Journal of Chemical Physics*, **109**, 11101–11108.

Yue K and Dill KA (1996) Folding proteins with a simple energy function and extensive conformational searching. *Protein Science*, **5**, 254–261.

Specialist Review

Protein domains

Jaap Heringa

Vrije Universiteit, Centre for Integrative Bioinformatics VU, Amsterdam, The Netherlands

1. Introduction

Protein domains are independent or semi-independent folding units of protein structure that serve as structural building blocks of proteins. Their size varies from 36 amino acids in the protein E-selectin to 692 residues in lipoxgenase-1 (Jones *et al*., 1998). Most domains observed in nature have less than 200 residues, with an average domain size of about 100 residues. While smaller domains of less than 50 residues are often structurally reinforced by disulfide bonds or coordinated metal ions, large domains comprising more than 300 residues are expected to contain multiple hydrophobic cores.

Domains are genetically mobile units, and multidomain families are found throughout the three kingdoms (Archaea, Bacteria, and Eukarya). The majority of genomic proteins, 75% in unicellular organisms and more than 80% in metazoa, are multidomain proteins created as a result of gene duplication events (Apic *et al*., 2001). Genetic mechanisms influencing the layout of multidomain proteins include gross rearrangements such as inversions, translocations, deletions and duplications, homologous recombination, and slippage of DNA polymerase during replication (Bork *et al*., 1992). Although genetically conceivable, the transition from two single-domain proteins to a two-domain protein requires that both domains fold correctly and that they manage to bury a fraction of the previously solvent-exposed surface area in a newly generated interdomain surface, the size of which depends on the constraints imposed by the domain linker region and the scale of domain interaction.

Protein structure is intrinsically hierarchic in its internal organization. It progresses from secondary structure (*see* Article 76, **Secondary structure prediction**, Volume 7) elements at the lowest level, via supersecondary structures and domains to complete proteins and assemblies of proteins at the highest level. At the domain level, the connectivity of the polypeptide backbone between the domains appears to be less important, as the protein retains a stable structure irrespective of the sequential ordering of the domains. Moreover, evolution has led to various forms of domain arrangements across the protein sequence that can be classified by their connectivity (Das and Smith, 2000). These arrangements go beyond sequential permutations of complete domain structures and can affect the internal domain organization. Figure 1 displays the observed modes of connectivity between domains

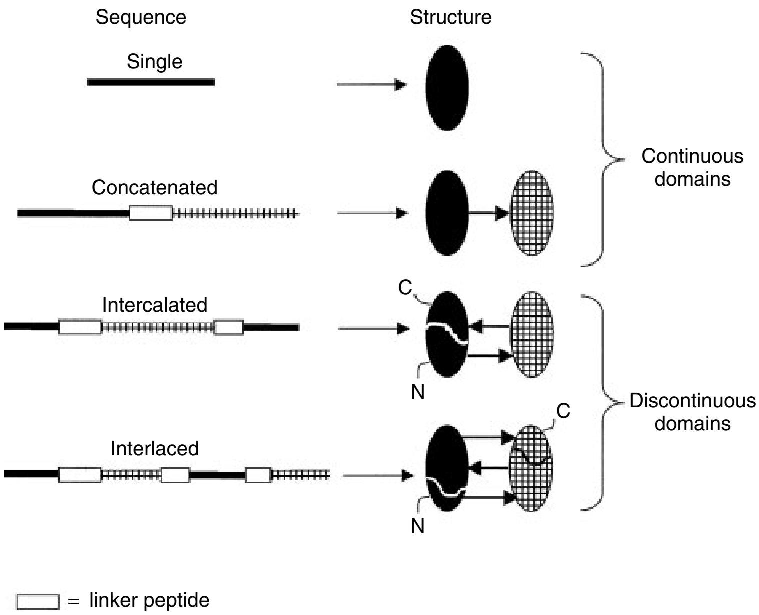


Figure 1 Schematic illustration of different types of backbone connectivity in multidomain protein structures (right) and their coding peptide sequences (left). Note that continuous domains are encoded by uninterrupted sequences, whereas discontinuous domain sequences show inserted sequence fragments

ordered by the degree of complexity. On the basis of their connectivity, domains can be divided into two main classes: continuous domains, where the associated coding sequence is an uninterrupted stretch of amino acids, and discontinuous domains, where the coding sequence is interrupted by subsequences encoding alternative structures, such as inserted domains. An example of a multidomain structure with a combination of continuous and discontinuous domains is the three-domain structure pyruvate kinase, a member of the phosphotransferase family, the structure and connectivity of which are given in Figure 2.

A protein module has been defined as a continuous domain structure that is structurally independent enough to serve as a modular plug-in to many multidomain proteins. Modules are believed to have spread largely by exon shuffling and are observed in many different multidomain proteins with permuted or altogether different domain combinations. The fact that modules often show the N- and C-termini in close proximity appears to facilitate their addition to other protein structures, since their insertion in loop regions would not require gross rearrangements.

A structurally viable mechanism for forming oligomeric domain assemblies is provided by domain swapping (Bennett *et al*., 1995). In domain swapping, a secondary or tertiary element of a monomeric protein is replaced by the same element of another protein. Domain swapping can range from secondary structure elements to whole structural domains in bringing together single-domain or multidomain

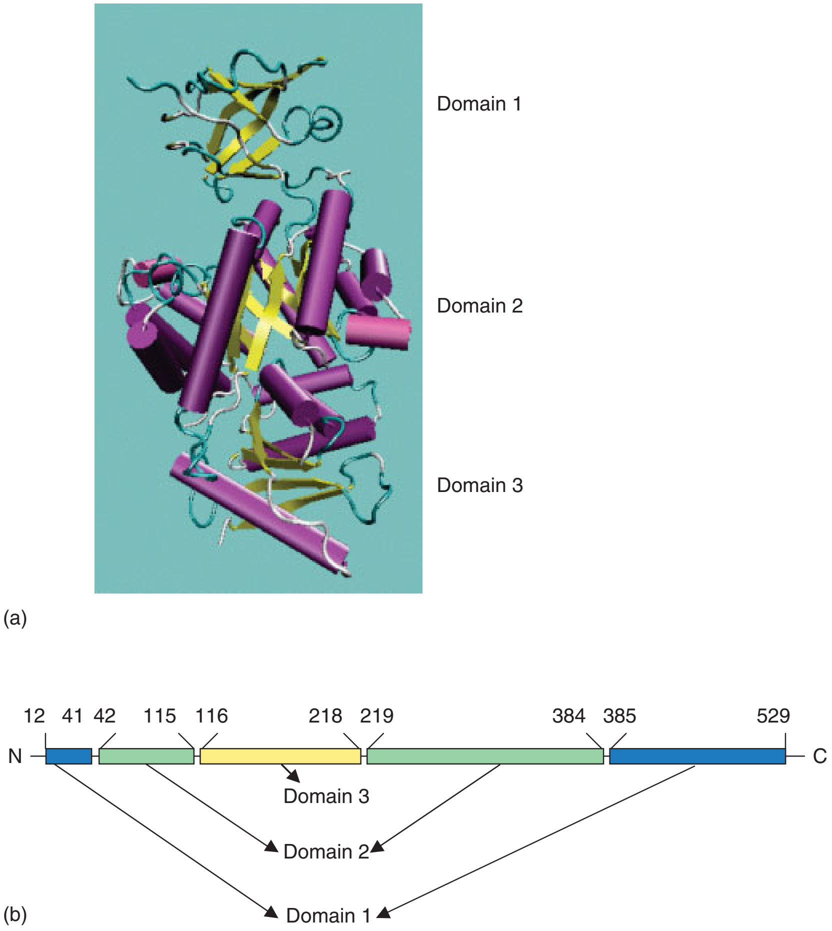


Figure 2 Pyruvate kinase structure and domain connectivity. (a) The structure shows three domains: domain 1 (discontinuous) is a β-barrel regulatory domain; domain 2 is a discontinuous α/β barrel catalytic substrate-binding domain; and domain 3 corresponds to a continuous α/β nucleotide-binding domain. (b) Schematic representation of the corresponding amino acid sequence, showing the segmental arrangement of the three domains. The intercalated and interlaced structures have N- and C-termini indicated (designated "N" and "C", respectively) for clarity

structures. It also represents a model of evolution for functional adaptation by oligomerization, for example, of oligomeric enzymes that have their active site at subunit interfaces (Heringa and Taylor, 1997).

2. Predicting domain boundaries in multidomain proteins

A number of structural classification methods have been developed to group the large number of protein domain folds deposited in the Protein Data Bank (PDB)

(*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7). The most widely used structural databases are CATH (Orengo *et al*., 1997), SCOP (Lo Conte *et al*., 2000), FSSP (Holm and Sander, 1996), and 3Dee (Siddiqui *et al*., 2001). Links to Web servers for these databases are provided in Table 1. For each database, a unique protocol is followed to classify the protein structures at the domain level. Since a common operational definition of a domain is that of a compact globular substructure with more interactions within itself than with the rest of the structure (Janin and Wodak, 1983), two characteristics are commonly exploited in boundary prediction methods: extent of isolation and compactness (Tsai and Nussinov, 1997). Measures of local compactness have been used in a number of recent domain-assignment methods (Holm and Sander, 1994; Islam *et al*., 1995; Sowdhamini and Blundell, 1995; Siddiqui and Barton, 1995; Nicols *et al*., 1995; Zehfus, 1997; Hinsen *et al*., 1999; Taylor, 1999; Xu *et al*., 2000; Siddiqui *et al*., 2001; Guo *et al*., 2003), albeit these approaches generally cannot handle discontinuous or closely interacting domains very well. As a result, Hadley and Jones 1999 observed discrepancies between assignments made by the above structural domain databases.

A large number of profile and HMM (hidden Markov model; *see* Article 98, **Hidden Markov models and neural networks**, Volume 8) methods are available to delineate the domain structure in a query sequence. These methods make use of

Table 1 WWW addresses of domain-related databases of interest

Name	WWW	Comment
Sequence-based classification databases		
InterPro	**http://www.ebi.ac.uk/interpro/**	
PROCLASS	**http://www-nbrf.georgetown.edu/gfserver/proclass.html**	
PIR	**http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml**	
PRINTS	**http://bioinf.man.ac.uk/dbbrowser/PRINTS/**	
PFAM	**http://www.sanger.ac.uk/Pfam/**	
SMART	**http://smart.embl-heidelberg.de/**	
ProtFam List	**http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-family.html**	
MetaFam	**http://metafam.ahc.umn.edu/**	
ProDom	**http://prodes.toulouse.inra.fr/prodom/doc/prodom.html**	
Structural classification databases		
SCOP	**http://scop.mrc-lmb.cam.ac.uk/scop**	
FSSP	**http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html**	
CATH	**http://www.biochem.ucl.ac.uk/bsm/cathnew/index.html**	
3Dee	**http://www.compbio.dundee.ac.uk/3Dee/**	
Protein sequence clustering databases		
SYSTERS	**http://www.dkfz-heidelberg.de/tbi/services/cluster/systersform**	
ProtoMap	**http://www.protomap.cs.huji.ac.il/**	
CluSTr	**http://www.ebi.ac.uk/clustr/**	
COG	**http://www.ncbi.nlm.nih.gov/COG**	
Domain linker databases		
LinkerDB	**http://ibivu.cs.vu.nl/programs/linkerdbwww/**	
Protein domain motions databases		
ProtMotDB	**http://hyper.stanford.edu/~mbg/ftp/ProtMotDB/ProtMotDB.main.html**	

domain classifications in protein domain sequence databases, and use the deposited sequences of each domain family to derive a profile or HMM. Each profile or HMM is then used to scan the query sequence, after which the highest scoring ones are used to delineate the putative domains and their boundaries. Recent developments for these search techniques is the addition of contextual information such as co-occurrence of domains in a sequence (Coin *et al*., 2003) and the taxonomic distribution (Coin *et al*., 2004).

Although a very difficult task, some ab initio computational methods are also available for the prediction of domain boundaries from protein sequence information alone. The method SnapDRAGON (George and Heringa, 2002a) is primarily based on the hydrophobic collapse during folding and folds a 3D protein structure on the basis of the notion of conserved hydrophobicity and information from secondary structure prediction methods. The method takes as input a multiple alignment of the query sequence and a set of homologs, as well as the predicted secondary structure. It then builds 100 models of the protein structure, with each assigned domain boundaries using the aforementioned method of Taylor (1999). A final prediction of the domain boundaries is achieved by assessing the consistency of the domain boundaries over the 100 predicted folds and by determining statistically significant boundaries.

Domain boundary identification by comparative sequence analysis is performed by the method DOMAINATION (George and Heringa, 2002b), which is based on the homology search method PSI-BLAST (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7; Altschul *et al*., 1997) to identify homologs to a query sequence. The distribution of N- and C-terminal positions of the local alignments obtained by PSI-BLAST is used to identify putative domain boundaries by a process of chopping and joining domains and domain segments in an attempt to reconstruct the evolutionary loss and gain of domains. The method follows an iterative protocol of on-the-fly cutting and joining of domain fragments and subjecting newly formed fragments to PSI-BLAST. It is able to recognize continuous and discontinuous domains and increases the accuracy of PSI-BLAST by about 15%.

3. Domain fusion

Domains in multidomain structures are likely to have once existed as independent proteins, and many domains in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes (Davidson *et al*., 1993). As an example of gene fusion, vertebrates have a multienzyme protein (GARs-AIRs-GARt) comprising the enzymes glycinamide ribonucleotide synthetase (GARs), aminoimidazole ribonucleotide synthetase (AIRs), and GAR transformylase (GARt) 1. In insects, the corresponding polypeptide encodes the four-domain structure GARs-(AIRs)2-GARt. However, GARs-AIRs is encoded separately from GARt in yeast, and in bacteria, each domain is encoded separately (Henikoff *et al*., 1997).

Marcotte *et al*. (1999) introduced the "Rosetta Stone" computational method for inferring protein function and interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologs in another

organism fused into a single protein chain. A comparison of sequence homologs from multiple organisms can reveal these fused sequences, which are called Rosetta Stone sequences because the existence of these multidomain proteins implies the putative interactions between the separate protein pairs in other organisms.

Another genomic method to predict domain function is phylogenetic profiling (Pellegrini *et al*., 1999). The method checks for simultaneous presence or absence of pairs of genes across a wide spectrum of genomes, and thereby reveals sets of proteins that are likely to have coevolved, and thus can be expected to act in the same cellular process.

The protein phylogenetic profiling and the Rosetta Stone methods have been very useful in predicting the function of large numbers of genes without experimental functional annotation.

4. Domain evolution and gene duplication

The evolution and spread of domains through different protein families is likely to be the result of *gene duplication*, leading to one or more paralogous genes. After such duplication events, the selection pressure becomes less for the paralogous copies of the gene, so that these are freer to manufacture adaptive functions over time. For more recently evolved protein domains, a correspondence between the sequential boundaries of exons and those of protein domains is observed. It is therefore believed that *exon shuffling* during evolution has enabled organisms to develop new proteins with new functions by the addition of exons from other parts of the genome that form new domains as part of an existing protein.

A multidomain architecture offers the potential of reuse through domain repeats (*see* Article 33, **Protein repeats**, Volume 7). Domain duplication is essential for the formation of the active sites of enzymes such as in the aspartic and serine proteinases, porphobilinogen deaminase or the binding sites of transport proteins such as lactoferrin, transferrin, and sulfate-binding protein. Domain repeats also result in duplication of active sites or binding sites such as the calcium binding sites in calmodulin or the nine DNA-binding zinc fingers of the transcription factor TFIIID.

Recent methods to delineate repeats in protein sequences are RADAR (Heger and Holm, 2000), REPRO (George and Heringa, 2000; Romein *et al*., 2003), and TRUST (Szklarczyk and Heringa, 2004). Taylor *et al*. (2002) devised a method based on Fourier analysis to automatically recognize repeated substructures in protein structures.

5. Nonorthologous displacement

An increasing number of cases are identified for nonorthologous displacement (Koonin *et al*., 1996), where enzymes carrying out an identical function in different organisms belong to entirely different protein families, and, thus, are not expected to show any sequence similarity. Examples of nonorthologous displacement include ornithine decarboxylase in *E. coli* and *S. serevisiae*, where the isozymes speF

and speC are responsible for this function in *E. coli* and share the same structure comprising three domains (ornithine decarboxylase N-terminal “wing” domain, PLP-dependent transferase, and ornithine decarboxylase C-terminal domain). The corresponding enzyme spe1 in *S. cerevisiae* is a two-domain protein with entirely different domain structures (PLP-binding barrel and alanine racemase-like domain). A problem with nonorthologous displacement is that it cannot be recognized by sequence comparison methods, which are aimed at recognizing divergent evolution by mutation and insertions/deletions, including changes of gene structure by gene fusion or fission. However, the techniques are not able to trace evolutionary cases of horizontal gene transfer or functional displacement of one gene by another one within a genome.

Suppose that a set of proteins is studied that correspond to a linear metabolic pathway, and that all proteins are found for a given organism, except for one corresponding to an intermediate step. It can then be expected that the missing activity is performed by some hitherto undetected protein, in which case sequence comparison techniques can lead to the identification of the missing protein by a homology search using a homologous domain sequence in another genome that has the appropriate functional annotation. In the absence of any putative homolog, nonorthologous gene displacement could be suspected, in which case direct sequence analysis is of no use. However, functional ontologies, such as GO (*see* Article 82, **The Gene Ontology project**, Volume 8; Ashburner *et al*., 2000), can aid the process in that various nonorthologous sequences can be identified using a GO functional descriptor. This leads to a direct candidate whenever a protein domain is identified in the target genome. Otherwise, if the ontology identifies a putative nonorthologous gene with the desired function in another genome, a homology search (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7) using the latter gene can reveal putative nonorthologous genes for the intermediate metabolic step in the query organism.

In the case of a prokaryotic organism, the fact that gene sequences in close proximity in the genome tend to be functionally related can be exploited. Given a confirmed protein domain sequence associated with the above metabolic pathway, the neighborhood of its homologs in complete prokaryotic genomes can be analyzed using the STRING server (Snel *et al*., 2000). The latter server has been used, for example, for the reannotation of the *Mycoplasma pneumoniae* genome (Dandekar, *et al*., 2000).

6. Homology regions

As a result of a homology search operation, for example, by using the program PSI-BLAST (Altschul *et al*., 1997; *see also* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7), regions with significant similarity between otherwise unrelated proteins in different species can abound. Such so-called homology regions can be sustained within different sequence contexts, and often add a specific functionality to their constituent proteins. In some cases, the function of regional homology domains can be inferred by comparing the properties of the host proteins, when experimental functional annotation is available.

However, the detection of these domains in uncharacterized sequences can also provide valuable clues for the protein's function. In other cases where the biological role of a local homology domain is not known, the delineation and comparison of these regions might still be useful, for example, for finding conserved residues as targets for mutagenesis or for structural studies.

7. Assessing functional specificity of domain families

Predicting or assessing the functional specificity of a given domain sequence is possible when a set of homologs is available for a query sequence, and the homologs are suspected to have a roughly identical function but with subtle variations (e.g., dehydrogenase domain structures acting on different substrates). Insight can then be gained by aligning the members of the family and deriving a phylogenetic tree. If the family is organized into a tree that correlates well with the functions described (or derived) for the proteins, then one can draw relatively reliable conclusions regarding the query's function. If the query falls into a clear subgroup of the tree, one can assume that as a function of the branch. If, however, the functional annotation of the branch holding the query is heterogeneous, or the query falls in between two branches, then one should assume a less-specific description in accordance with the neighboring branches (e.g., "alcohol dehydrogenase"), or resort to the generic functional description of the whole family (e.g., "dehydrogenase").

The transfer of function from one protein to another should only be done when the protein sequences match over their total lengths in the pairwise sequence alignment, without any large unaligned sequence fragments that might correspond to an additional domain. If some fragments do not align, one has to carefully check whether the function that is transferred really corresponds to the part that is common. The use of domain databases and methods for domain identification (*vide supra*) can be of great help at this point.

If a query sequence contains a domain that is not matched in the pairwise alignment, this should be noted. Conversely, if the protein used as source of the annotation contains an unmatched domain, the annotation should just state the presence of the domain(s) of the query without transferal of the complete functional description of the source.

8. Protein sequence clustering

Protein sequence cluster databases group related proteins together and are derived automatically from protein sequence databases using different clustering algorithms.

Each database uses a different clustering algorithm aimed at grouping the sequences generally in homologous families and superfamilies. Assignment of a query sequence without annotation to any of the clusters can provide valuable clues about the protein's function. Since they are derived automatically from protein sequence databases without manual crafting and validation of family discriminators, these databases are relatively comprehensive, although the biological relevance of clusters can be ambiguous and can sometimes be an artifact of particular thresholds.

Multidomain protein sequences are a particular source of error in that the clustering algorithm can establish a link between two sequences A and B on the basis of a particular domain, while a third sequence C might be linked to sequence B on the basis of another domain sequence. If unchecked, this would lead to an incorrect clustering of sequences A, B, and C. Table 1 provides a number of links to websites of protein sequence clustering databases.

8.1. Interaction domains

A large and growing body of evidence suggests that proteins involved in the regulation of cellular events such as the cell cycle, signal transduction, protein trafficking, targeted proteolysis, cytoskeletal organization, and gene expression are typically multidomain proteins comprising a catalytic domain and one or more interaction domains. Interaction domains induce the interaction of signaling polypeptides and specific multiprotein complexes. For example, they can couple a cell surface receptor to an intracellular biochemical pathway that controls the cellular responses to an external signal. Interaction domains can be involved in a series of protein–protein interactions, designed to localize signaling proteins to appropriate subcellular locations or to determine the enzyme specificity. A well-studied example is the family of protein kinases in relation to their substrates. Typically, protein–protein interaction domains are relatively small and independently folding modules of 35–150 amino acids that are able to fold in isolation from their host proteins while retaining their intrinsic ability to bind their physiological partners. Their N- and C-termini are usually close together in space, while their ligand-binding surface lies on the opposite face of the domain. This arrangement facilitates their evolutionary use as a modular plug-in since it allows the domain to be easily inserted into a host protein while projecting its ligand-binding site to contact other polypeptides.

Protein–protein interaction domains can be clustered into separate families, based either on their sequence or ligand-binding properties. Interaction domains are often used repeatedly by a large number of multidomain proteins that display very different domain structures, where they mediate a particular type of molecular recognition.

1. An important class of interaction domains recognizes peptides that have been posttranslationally modified. These modifications include phosphorylation, acetylation, or methylation. The dynamic control of cellular behavior exerted by covalent protein modifications, therefore, appears largely mediated by interaction domains that regulate the associations of signaling protein cascades.
   (a) Phosphorylation:
      (i) An important member of this class is the SH2 domain, which is anticipated to be encoded in the human genome at least 120 times. A large number of cytoplasmic proteins contain one or two SH2 domains that directly recognize phosphotyrosine-containing motifs that are found, for example, on surfaces of activated receptors for cytokines, growth factors, and antigens. SH2 domains all recognize phosphotyrosines, but have different preferences for amino acids immediately following the

phosphorylated residue, allowing a degree of specificity in signaling by tyrosine kinases.

(ii) Phosphotyrosine-containing motifs are also recognized by the PTB domain, a very different structure than the SH2 domain. PTB domains are found as constituents of binding proteins such as the IRS-1 substrate of the insulin receptor.

(iii) Phosphoserine and phosphothreonine motifs are targeted by a growing family of interaction domains, including FHA domains, 14-3-3 proteins, and WD40-repeat domains. Each of the subfamilies and proteins contained therein bind specific phosphoserine/threonine motifs, and thus mediate the biological activities of protein-serine/threonine kinases.

(b) In addition to phosphorylation, acetylation and methylation have recently been implicated in inducing modular protein–protein interactions. For example, acetylation or methylation of lysine residues on the surfaces of histone structures, respectively, create effective binding sites for Bromo and Chromo domain constituents of multidomain proteins involved in chromatin remodeling.

2. A large group of interaction domains bind proline-rich motifs, including SH3, WW, and EVH1 domains. The complexes that are recognized by these domains are therefore not dependent on posttranslational modifications, and hence are more stable partners than those that, for example, need to be phosporylated to enable binding by SH2 domains. SH3 domains interact with a large number of different domain types and are therefore referred to as “promiscuous” domains.
3. PDZ domains bind the C-termini of their interaction partner domains, ion channels, and receptors in a fashion that appears important for the localization of their targets to particular subcellular sites, as well as for downstream signaling. In addition to binding short C-terminal peptide motifs, PDZ domains can form heterodimers.
4. In addition to binding specific peptide motifs, a growing number of interaction modules are known to recognize a variety of phospholipids, particularly phosphoinositides (PI). For example, PH domains are constituents of lipid kinases and phosphatases, where they affect cellular functioning by binding promiscuously to either the phosphoinositide PI-4,5-P2 or PI-3,4,5-P3. On the basis of their phospholipid-binding, PH domains can localize signaling proteins to specific subregions of the plasma membrane, where they regulate the enzymatic activities of their host proteins.

A predominant feature of interaction domains is their apparent versatility. Though PTB domains were originally discovered on the basis of their ability to bind phosphorylated tyrosines within Asn-Pro-X-Tyr (NPXY) motifs, many PTB domains appear to be able also to recognize unphosphorylated NPXY-related peptides. It, therefore, appears that PTB domains originally evolved to bind unphosphorylated peptides, but have subsequently developed a capacity to recognize phosphotyrosine in relatively few specific cases. Furthermore, some individual PTB domains, for example, those within the FRS-2 and Numb proteins, have evolved to recognize two quite different peptide ligands.

Although many interaction domains bind very different substrates, a number of them have similar tertiary structures, such as, for example, the PTB, EVH1, and PH domains, which, respectively, recognize phosphotyrosine-containing peptide motifs, proline-rich peptides, and phospholipids (*vide supra*). The PTB/PH/EVH1 fold is shared by other types of interaction domains as well. It seems therefore that this domain fold provides a stable framework that can be used for modulating very many different intermolecular interactions. On the other hand, many signaling proteins contain a number of different interaction domains to yield a multidomain protein that can mediate multiple protein–protein and protein–phospholipid interactions. This modular organization of signaling proteins can target proteins to the appropriate site within the cell, and direct binding to cell surface receptors and downstream targets.

A number of bioinformatics approaches have been developed for the prediction of protein–protein interactions based on the correspondence between the phylogenetic trees of interacting proteins (Goh *et al*., 2000; Pazos and Valencia, 2001) and detection of correlated mutations between pairs of proteins (Pollock and Taylor, 1997; Pazos *et al*., 1997; Pollock *et al*., 1999).

9. Spacer domains and domain linkers

A number of protein domains are thought to have a purely structural role, allowing functional domains to be advantageously situated, or providing the overall protein with the necessary flexibility. It must be kept in mind that a domain may be considered to have a purely structural role simply because knowledge about its true function is missing. The immunoglobulin (Ig) family provides examples of this, as the Ig constant heavy 1 and 2 domains and the Ig constant light domain act as "spacer" domains to allow the Ig variable light and heavy domains to function optimally.

Multifunctional enzymes are generally composed of a number of discrete domains that are connected by interdomain linkers. A number of site-directed mutagenesis experiments have shown that linkers play a crucial role in the enzymatic activity of multidomain complexes (George and Heringa, 2003). In general, mutating or altering the length of linkers has been shown to affect protein stability, folding rates, and domain–domain orientation. The linkers provide dynamic structural bridges for the movement of ligands across the modules. Many linkers act as rigid spacers separating two domains, and provide a scaffold to prevent unfavorable interactions between folding domains. George and Heringa (2003) found the two main types of linker in natural proteins: helical linkers and nonhelical linkers. Helical linkers derive their rigidity from their helical conformation, while nonhelical linkers are rich in proline residues, which also leads to structural rigidity. They should be flexible, keeping domains apart while allowing them to move as part of their catalytic function. Linkers should, furthermore, be invulnerable to host proteases, as they are often the targets for degradation.

10. Domain movements

Domain motions are important for a large repertoire of functions including ligand binding, catalysis, regulation, formation of protein assemblies, and transport of

metabolites (Gerstein *et al*., 1994). Often, the binding site is situated at the interface of a pair of domains in the "closed" conformation, corresponding to the inactive form of the associated multidomain enzyme, while the "open" conformation of the domains exposes the binding site and, thus, provides the active form of the enzyme. Domain motions are often responsible for the mechanism of "induced fit" observed in protein–protein recognition.

Protein flexibility often corresponds to large relative movements of domains. In aspartic proteinases, for example, the two main domains forming the lobes of the structure show large relative shifts that are crucial for the activity of the constituent enzyme (Sali *et al*., 1992). Another example is the so-called swiveling domain in pyruvate phosphate dikinase (Herzberg *et al*., 1996), which brings an intermediate enzymatic product over about 45 Å from the active site of one domain to that of another. Such movements have been determined by a conformational analysis of protein structures, where the identification of, for example, hinge axes and their corresponding rotation angles permits a useful representation of protein domain movements (Gerstein *et al*., 1994).

Three main mechanisms of domain movements are discernible (Gerstein *et al*., 1994):

1. *Intrinsic flexibility:* These movements lead to deformations of the domain structure. Several small movements in secondary structures have a cumulative effect. These motions are due to small changes in backbone torsion angles of strands and/or helices, kinks involving prolines in helices as in adenylate kinase (Gerstein *et al*., 1994), or interconversion of helix and extended conformations as in calmodulin (Ikura *et al*., 1992).
2. *Hinged domain movements:* Of special interest are hinge-bending movements, where rigid domains are connected by flexible joints that tether the domains and constrain their movement. Hinge-bending is believed to allow an induced fit of molecular surfaces in protein assembly and ligand docking. These movements are caused by rotations of a domain around a pivotal point, often afforded by a linker region in between domains that changes its conformation to allow the rotation. For example, to attain icosahedral symmetry, the S and P domains of tobacco bushy stunt virus perform hinge movement mainly by dihedral angle changes in the peptide that forms a linker between the two domains (Olsen *et al*., 1983). In lactoferrin, the open and closed forms of the two domains involve a 53° rotation, which is achieved by conformational changes within two strands that connect the protein's two domains (Jameson *et al*., 1998).
3. *Ball-and-socket motion:* The variable (light and heavy chain) domains of antibodies rotate about 50° with respect to the constant (light and heavy chain) domains by means of a combination of hinge and shear motions (Lesk and Chothia, 1988).

The method HINGEFIND (Wriggers and Schulten, 1997) identifies domain movements and characterizes and visualizes the effective rotation axes (hinges). The program compares a pair of known structures (e.g., two different crystal structures of a protein or the results of molecular dynamics or Monte Carlo simulations) and partitions the protein with a prespecified tolerance in preserved subdomains.

It then determines effective rotation axes that characterize the domain movements with respect to the reference domain ("rigid core"). Hayward and Berendsen (1998) have developed a method to analyze comparative domain movements over larger sets of structures.

11. Conclusion

The analysis and prediction methods that elucidate the evolutionary development and function of domains and multidomain complexes have recently been enriched by the integration of genomic and dynamic contextual information. Many successes are to be derived from the ongoing developments in spawning an integration web between genomics, phylogenetic, sequence, structure, structural dynamics, and function data. This will shed more light on the evolution of the structural and functional versatility of protein domains, which will benefit biotechnical and biomedical applications.

Further reading

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, *et al*. (2000) InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, **29**, 37–40.

George RA, Kleinjung J and Heringa J (2003) Predicting protein structural domains from sequence data. In *Bioinformatics and Genomes – Current Perspectives*, Horizon Scientific Press: Norfolk, pp. 1–26, ISBN 1-898486-47-6.

Holm L and Sander C (1998) Dictionary of recurrent domains in protein structures. *Proteins: Structure, Function and Genetics*, **33**, 88–96.

Schultz J, Copley RR, Doerks T, Ponting CP and Bork P (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research*, **28**, 231–234.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Apic G, Gough J and Teichmann SA (2001) An insight into domain combinations. *Bioinformatics*, **17**, S83–S89.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Bennett MJ, Schlunegger MP and Eisenberg D (1995) 3D domain swapping: a mechanism for oligomeric assembly. *Protein Science*, **4**, 2455–2468.

Bork P, Sander C and Valencia A (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 7290–7294.

Coin L, Bateman A and Durbin R (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4516–4520.

Coin L, Bateman A and Durbin R (2004) Enhanced protein domain discovery using taxonomy. *BMC Bioinformatics*, **5**, 56–65.

Dandekar T, Huynen M, Regula JT, Zimmermann CU, Ueberle B, Andrade MA, Doerks T, Sánchez-Pulido L, Snel B, Suyama M, *et al*. (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Research*, **28**, 3278–3288.

Das S and Smith T (2000) Identifying nature's protein LEGO set. *Advances in Protein Chemistry*, **54**, 159–183.

Davidson J, Chen K, Jamison R, Musmanno L and Kern C (1993) The evolutionary history of the first three enzymes in pyrimidine biosynthesis. *Bioessays*, **15**, 157–164.

George RA and Heringa J (2000) The REPRO server: finding protein internal sequence repeats through the Web. *Trends in Biochemical Sciences*, **25**, 515–517.

George RA and Heringa J (2002a) Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins: Structure, Function and Genetics*, **48**, 672–681.

George RA and Heringa J (2002b) SnapDRAGON – a method to delineate protein structural domains from sequence data. *Journal of Molecular Biology*, **316**, 839–851.

George RA and Heringa J (2003) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Engineering*, **15**, 871–879.

Gerstein M, Lesk A and Chothia C (1994) Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739–6749.

Goh CS, Bogan AA, Joachimiak M, Walther D and Cohen FE (2000) Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, **299**, 283–293.

Guo JT, Xu D, Kim D and Xu Y (2003) Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Research*, **31**, 944–952.

Hadley C and Jones D (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure with Folding & Design*, **7**, 1099–1112.

Hayward S and Berendsen HJC (1998) Systematic analysis of domain motions in proteins from conformational change; new results on citrate synthase and T4 lysozyme. *Proteins: Structure, Function and Genetics*, **30**, 144–154.

Heger A and Holm L (2000) Automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function and Genetics*, **41**, 224–237.

Henikoff S, Greene E, Pietrokovski S, Bork P, Attwood T and Hood L (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.

Heringa J and Taylor WR (1997) Three-dimensional domain duplication, swapping and stealing. *Current Opinion in Structural Biology*, **7**, 416–421.

Herzberg O, Chen CCH, Kapadia G, McGuire M, Carrol LJ, Noh SJ and Dunaway-Mariano D (1996) Swiveling-domain mechanism for enzymatic phosphotransfer between remote reaction sites. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 2652–2657

Hinsen K, Thomas A and Field MJ (1999) Analysis of domain motions in large proteins. *Proteins: Structure, Function and Genetics*, **34**, 369–382.

Holm L and Sander C (1994) Parser for protein folding units. *Proteins: Structure, Function and Genetics*, **19**, 231–234.

Holm L and Sander C (1996) Touring protein fold space with Dali/FSSP. *Nucleic Acids Research*, **26**, 316–319.

Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB and Bax A (1992) Solution structure of a Calmodulin-target peptide complex by multidimensional NMR. *Science*, **256**, 632–638.

Islam S, Luo J and Sternberg M (1995) Identification and analysis of domains in proteins. *Protein Engineering*, **8**, 513–525.

Jameson GB, Anderson BF, Norris GE, Thomas DH and Baker BN (1998) Structure of human apolactoferrin at 2.0-A resolution. Refinement and analysis of ligand-induced conformational change. *Acta Crystallographica. Section D*, **54**, 1319–1335.

Janin J and Wodak S (1983) Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology*, **42**, 21–78.

Jones S, Stewart M, Michie A, Swindells M, Orengo C and Thornton J (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science*, **7**, 233–242

Koonin EV, Mushegian AR and Bork P (1996) Non-orthologous gene displacement. *Trends in Genetics*, **12**, 334–336.

Lesk AM and Chothia C (1988) Elbow motion in the immunoglobulins involves a molecular ball and socket joint. *Nature*, **355**, 188–190.

Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG and Chothia C (2000) A structural classification of proteins database. *Nucleic Acids Research*, **28**, 257–259.

Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates T and Eisenberg D (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Nicols WL, Rose GD, Ten Eyck LF and Zimm BH (1995) Rigid domains in proteins: an algorithmic approach to their identification. *Proteins: Structure, Function and Genetics*, **23**, 38–48.

Olsen AJ, Briscogne G and Harrison SC (1983) Structure of tomato bushy stunt virus IV. The virus particle at 2.9 Å resolution. *Journal of Molecular Biology*, **171**, 61.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pazos F and Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**, 609–614.

Pazos F, Helmer-Citterich M, Ausiello G and Valencia A (1997) Correlated mutations contain information about protein–protein interaction. *Journal of Molecular Biology*, **271**, 511–523.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Pollock DD and Taylor WR (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering*, **10**, 647–657.

Pollock DD, Taylor WR and Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, **287**, 187–198.

Romein JW, Heringa J and Bal HE (2003) *A million-fold speed improvement in genomic repeats detection*, *SuperComputing'03*, Phoenix, November, 2003.

Sali A, Veerapandian B, Cooper D, Moss DS, Hofmann T and Blundell TL (1992) Domain flexibility in aspartic proteinases. *Protein: Structural, Functional and Genetics*, **12**, 158–170.

Siddiqui A and Barton G (1995) Continuous and discontinuous domains – an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, **4**, 872–884.

Siddiqui AS, Dengler U and Barton GJ (2001) 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201.

Snel B, Lehmann G, Bork P and Huynen MA (2000) STRING: A web-server to retrieve and display the conserved neighborhood of genes. *Nucleic Acids Research*, **28**, 3442–3444.

Sowdhamini R and Blundell TL (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Science*, **4**, 506–520.

Szklarczyk R and Heringa J (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**(Suppl 1), i311–i317.

Taylor WR (1999) Protein structure domain identification. *Protein Engineering*, **12**, 203–216.

Taylor WR, Heringa J, Baud F and Flores TP (2002) A Fourier analysis of symmetry in proteins. *Protein Engineering*, **15**, 79–89.

Tsai CJ and Nussinov R (1997) Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Science*, **6**, 24–42.

Wriggers W and Schulten K (1997) Protein domain movements: Detection of rigid domains and visualization of effective rotations in comparisons of atomic coordinates. *Proteins: Structure, Function and Genetics*, **29**, 1–14.

Xu Y, Xu D and Gabow HN (2000) Protein domain-decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.

Zehfus M (1997) Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains. *Protein Science*, **6**, 1210–1219.

Specialist Review

Complexity in biological structures and systems

Arthur M. Lesk
The Pennsylvania State University, University Park, PA, USA

1. Introduction

Complexity is a concept that we encounter in our daily lives, but generally we can get along without needing to define it precisely. But, how well do we really understand complexity, and allied concepts including entropy, randomness, predictability, and chaos? What are the relationships among them? And how can they be used to illuminate biology?

No one doubts that life is complex. Until recently, the response of biological scientists has been to carve off pieces that might be simpler to deal with by themselves. The observation by the Buchners of enzymatic activity in yeast cell extracts initiated an exhaustive program of isolation and purification of proteins. The activity of a purified enzyme is not without complexity, but the approach of classical biochemistry was to simplify. Now we hope to face complexity head-on (for additional reading, see Adam, 2002; Cramer and Loewus, 1993; Colosimo, 1997; Colosimo, 2004).

How to proceed? Interrogating our naive and colloquial concepts of words is a starting point, if only to expose our prejudices. However, it ultimately becomes unproductive, and must be abandoned when it does. What formal theory can we call upon as an alternative?

2. Complexity of sequences

Perhaps the simplest complex object in biology is a sequence. We have heard of random sequences, and probably have the general idea that the more random the sequence the more complex it is. Genomic sequences contain "low-complexity" regions – are these regions more or less random than a protein-coding region? How can such properties of sequences be measured?

Given a sequence of characters $x_1, x_2, x_3, \ldots$ with each x_i chosen from an alphabet $\mathcal{A}$, what controls the amount of information needed to identify individual characters? If the alphabet $\mathcal{A}$ is very large, and if the x_i are evenly distributed, more information is needed than if $\mathcal{A}$ is small or the distribution of the x_i is highly skewed.

For example, genomic sequences contain characters A, T, G, and C. To identify each symbol, it suffices to ask two "yes-or-no" questions:

- Question 1: Is it a purine (or a pyrimidine)?
- Question 2: Is it 6-keto (or 6-amino)?

Representing yes = 1 and no = 0, the answer to each "yes-or-no" question provides 1 binary digit, or *1 bit* of information. In principle, we could encode each nucleotide as a two-bit binary string. In practice, we encode each nucleotide as a 1-byte (8-bit) character. This is more than necessary: with 1 byte we can encode the entire English alphabet plus special symbols of the ASCII character set.

It is reasonable that a string containing text plus special symbols is more complex than a genomic sequence of the same length containing only A, T, G, and C. As we shall see, allocation of unnecessarily generous space per character in the 1 byte/nucleotide representation allows genomic sequences to be compressed, and they can, in principle, be compressed more effectively than a string of ordinary text. Here is a hint of a relationship between compressibility and complexity, a theme we shall develop.

Questions about the required length of representations appear in the genetic code itself. How many nucleotides are required to encode 20 amino acids? Two are not enough: an alphabet of 4 nucleotides provides only 16 dinucleotides. If all amino acids are to require the same number of nucleotides, there must be at least three nucleotides per codon, as observed.

How can these informal considerations be made quantitative? In 1948, C. E. Shannon introduced the concept of *entropy* in information theory, in his analysis of signal transmission. Suppose the probability of each symbol x_i is p_i, with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$. Shannon's measure of entropy is:

$$H = -\sum_i p_i \log_2 p_i \tag{1}$$

The entropy can be interpreted as the minimum average number of bits per symbol required to transmit the sequence. For a genomic sequence with equimolar base composition, $p_A = p_T = p_G = p_C = 0.25$, $H = 2$. This recovers our informal result that each symbol requires two bits, or answers to two "yes-or-no" questions. For a sequence limited to two equiprobable characters A and T: $p_A = p_T = 0.5$, $H = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1$. This makes sense because, knowing that the only choices are A and T, we can decide which base it is with *one* 'yes-or-no' question. We could encode A = 1 and T = 0, in one bit.

Suppose instead that a sequence has the skewed nucleotide composition: $p_A = p_T = 0.42$, and $p_G = p_C = 0.08$. Then

$$H = -[0.42 \log_2 0.42 + 0.42 \log_2 0.42 + 0.08 \log_2 0.08 + 0.08 \log_2 0.08] = 1.63 \tag{2}$$

What is the significance of the fact that $H = 1.63$ is less than 2? The uncertainty in each transmitted symbol is not complete – A and T are more likely than G and C. In principle, we can use this to improve the coding efficiency.

The Morse code for telegraphy took advantage of unequal letter distribution frequencies to encode common letters (in English) with short sequences and uncommon letters with longer ones (for instance, E = dot and J = dot-dash-dash-dash). Huffman's (1952) general method assigns length-optimal codes to symbols from their relative probabilities.

It would be difficult to devise a Morse code for nucleotides because the fact that we can easily encode them with no more than two bits does not give us much room to play with: we cannot subdivide a bit. However, consider encoding a genome sequence as a succession of trinucleotides, assuming no bias in trinucleotide frequencies other than that expected from the mononucleotide frequencies. That is, if $p_{\mathrm{A}} = 0.42$ and $p_{\mathrm{C}} = 0.08$, $p_{\mathrm{ACC}} = 0.41 \times 0.08 \times 0.08$. There are 64 trinucleotides to encode. Six bits per triplet obviously suffice, but for the skewed distribution $p_{\mathrm{A}} = p_{\mathrm{T}} = 0.42$, $p_{\mathrm{G}} = p_{\mathrm{C}} = 0.08$, $H = 4.9$. We could encode the sequence using 5 bits per trinucleotide instead of 6 (a good exercise for the reader).

It is possible to calculate entropies of genome sequences, and to use the entropy computed, over a sliding window, a window to detect low-complexity regions. In the human genome, these include simple repeats, or microsatellites, or regions of highly skewed nucleotide composition such as AT-rich or GC-rich regions, or polypurine or polypyrimidine tracts.

Conversely, it is observed that the distribution of oligonucleotides (dinucleotides, triplets, etc.) is uneven. Therefore, an encoding of a genetic sequence based on oligonucleotides could be shorter than an encoding of individual bases. It might also reveal biologically significant patterns; for instance, codon usage patterns in protein-coding regions. Some algorithms for gene identification make use of biases, in coding regions, of hexanucleotide frequencies.

The genetic code does not achieve theoretical coding efficiency, producing instead redundancy in the form of synonymous codons. (There does not seem to be selection for reduction in size of nonviral genomes.) The redundancy in the genetic code is utilized in evolution. Many single-nucleotide polymorphisms are silent, giving the code a certain error tolerance. Conservative mutations allow proteins to evolve with small nonlethal changes that, cumulatively, achieve large changes in structure and function.

In signal processing, extra coding bits can be used for error detection or even correction. In biology, the redundancy in having two copies of the genetic information in two strands of DNA is used to detect and correct errors in replication and translation, and to repair DNA damage.

Shannon entropy is related to the thermodynamic entropy by multiplying by Boltzmann's constant $k = 1.3807 \times 10^{-23}\ \mathrm{J \cdot K^{-1}}$ to give the statistical-mechanical equation:

$$S = -k \sum_{i} p_i \ln p_i \tag{3}$$

where the p_i are the probabilities of states of individual components of a physical system, such as the particles in a gas. The relationship between information theory and statistical mechanics has been explored by physicists, including J. C. Maxwell and L. Szilard, in their discussions of "Maxwell's demon"; and by E. T. Jaynes (see Rosenkrantz, 1983).

3. Randomness of sequences

The Shannon entropy of sequences is related to their randomness, another concept we use in everyday life without worrying too much about exactly what it means. Randomness has been made precise in independent work by A. N. Kolmogorov, R. Solomonoff, and G. Chaitin (Li and Vitanyi, 1997).

Kolmogorov defined the randomness of a sequence of numbers as *the length of the shortest computer program that can reproduce the sequence*. The sequence 0, 0, 0, 0, 0, 0, 0 . . . is far from random, as it is the output of the short program:

Step 1: print 0.
Step 2: go back to step 1.

Periodic sequences, such as:

Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, Monday, . . . are also nonrandom. In contrast, a truly random sequence has no description shorter than the sequence itself.

Kolmogorov's definition distinguishes truly random sequences from pseudorandom sequences. A pseudorandom sequence is a sequence of numbers generated by some program; with a distribution similar to that of a random sequence, with no obvious biases or correlations. Pseudorandom numbers suffice for many applications, such as Monte Carlo calculations. However, the random-number generator is a shorter program to generate the sequence than the sequence itself (provided that we generate a long enough sequence). The Kolmogorov randomness of a pseudorandom sequence is therefore relatively low.

4. The relationship between compressibility and complexity of sequences

One way to shorten the specification of a sequence is to compress it. We all compress our files, using, for instance, gzip (Ziv and Lempel, 1977; Welch, 1984). If a sequence is truly random, in the sense of Kolmogorov, it cannot be compressed, by gzip or any other method. Although nonrandom sequences *can* be compressed, particular algorithms such as gzip do not give optimal compressions of all data sets. Indeed, the Kolmogorov definition of randomness does not specify the compression method. Rather it is defined in terms of *decompression* algorithms, that (re)generate the sequence.

5. The relationship between predictability and compressibility

One basic principle of compression algorithms is that if you can predict what is coming next, you can compress a description. The reason that the sequences 0, 0, 0, . . . and Monday, Tuesday, Wednesday, . . . are so compressible is that it is simple to specify the successor of any element. Even sequences without rules stating explicitly what the next element is can be compressed, if some indications

are available. It is not necessary that the indicators be supplied "up-front", as they can be for sequences such as 0, 0, 0, ... and Monday, Tuesday, Wednesday, ... The rules and statistics for predicting a successor can be probabilistic rather than unambiguous, and can be generated "on the fly" from incoming data.

6. The relationship between predictability and complexity

These considerations suggest a general idea that *the harder it is to predict the contents of a data set from a subset of the data, the more complex the data set is.* This proves to be a most useful guiding principle in analyzing the nature of the complexity of new types of data, and if there is any single take-home message from this article it is the significance of the relationship between predictablility and complexity. In facing a new type of object, and trying to devise a measure of complexity for it, ask yourself: knowing part of the object, how well can I predict the rest of it?

The considerations we have discussed describe the complexity of the *structures* of objects, specifically character strings. But, for biology, we are also interested in complexity of *processes.*

7. Static and dynamic complexity

Complexity has many dimensions, and one of them is time. If we could define and measure the static complexity of a system, that would provide an approach to dynamic complexity: we could ask how the static complexity is changing with time. Such changes appear to be governed by general rules, such as the laws of thermodynamics.

Because living things are dynamic, we must consider complexity of processes. One approach to complexity of dynamics appeals to physics: biological molecules are systems of particles. They follow the laws of dynamics, and usually classical Newtonian mechanics is adequate. We could define and describe the trajectories of the particles, and use the results to analyze the complexity. From the point of view of Kolmogorov, the initial positions and velocities of the particles, the forces between them, and Newton's laws of motion, together provide a concise description of dynamics of the system. However, even within the framework of classical dynamics, this can break down in chaotic states. In these states, very small changes in the initial conditions lead to very large changes in subsequent trajectories. Prediction of the dynamics requires very precise statement of the initial conditions, and very precise knowledge of the forces. Specification of the information required to describe the dynamics cannot, in these cases, be concise. Chaos is an extreme form of dynamic complexity.

Given the difficulty of analyzing the complexity of dynamical systems, how can we approach this topic? Perhaps the best-developed area of analysis of complexity of processes is in studies of complexities of computational problems.

The question of the limits to efficiency of computations is now often presented in terms of the technical problem of writing code that can be pushed to solve larger and larger problems. However, historically it emerged from the attempts by Hilbert

to axiomatize all of mathematics, and by Russell and Whitehead to automate the process of deriving the consequences of axioms; through Gödel's showing the futility of these efforts; and on through von Neumann and Turing's early theories of programs (for history and perspective, see Hartmanis, 1989; Sipser, 1992; Papadimitriou, 1997 and Cook, 2003. For modern textbook treatments see Garey and Johnson, 1979; Papadimitriou, 1994 and Hopcroft *et al*., 2000).

8. Computational complexity

Computational complexity is relevant to our subject for two reasons. One is that it is in computer science that ideas of complexity have been made most precise. The ideas that have emerged give us tools for deciding whether we understand the systems with which we deal. Papadimitriou (1997) noted the belief that "complexity must be the manifestation of mathematical poverty, lack of structure". Many of us work in the hope that this might be modified to: ... *apparent* complexity must be the *remediable* manifestation of poverty of mathematical understanding, lack of *appreciation* of underlying principles of structure. Indeed, some problems even in computer science turned out to be less complex than they originally appeared.

The second reason is that if we, as computational biologists, consider undertaking some calculation, what we propose must be feasible. In this context, Papadimitriou also mentioned "the thesis that 'polynomial worst-case time' is a plausible and productive mathematical surrogate of the empirical concept of 'practically solvable computational problem' ".

An algorithm in computer science defines a process for solving a problem. Some algorithms have closed forms: an example is the algorithm for solving the linear equation $x - a = b$: **Return** $\mathbf{x} = \mathbf{b} + \mathbf{a}$. However, most algorithms are iterative, and compute a sequence of intermediate results during their operation. Usually, the larger the data set to which the algorithm is applied, the longer the execution time required. In computer science, the measure of the complexity of a problem is the relationship between execution time and problem size, as the problem size grows larger and larger, for optimal algorithms.

Here is a simple example:

> *Given a list of N unequal numbers, sorted in increasing order, and a probe number that appears in the list, determine the position at which it appears.*

A simple approach is *sequential search:* Check the elements in order, and stop when you hit the match. (Note that this algorithm would work equally well for unsorted as for sorted lists.) On an average, sequential search will find the answer after $\sim N/2$ comparisons. The execution time of a program based on this algorithm is therefore expected to be proportional to N. This is called the *order* of the algorithm, $\mathcal{O}(N)$ (read "Oh-N"). We do not distinguish $\mathcal{O}(N)$ and $\mathcal{O}(N/2)$. Clever coding of an algorithm might speed up execution time by a factor of 2, but will not change an $\mathcal{O}(N/2)$ algorithm to an $\mathcal{O}(\log N)$ one. We want to characterize the behavior of the algorithm, not the implementation.

Does this *algorithm,* with execution time increasing as $\mathcal{O}(N)$, adequately characterize the *problem?*

No, because there are better algorithms that solve the same problem. A more clever approach is to compare the value of the submitted number with the number in the middle of the list, element $M = N/2$ (in integer arithmetic). If the target number is equal to element M, report M and stop. If the target number is larger than element M, throw away the initial half of the list. If the target number is smaller than the element M, throw away the final half of the list. Then repeat these operations, on the part of the list that remains. This *binary search* algorithm will require no more than $\log_2 N$ steps. The execution time of a program based on it will scale as $\log_2 N$ as N grows large. Just as we did not care to distinguish between $\mathcal{O}(N)$ and $\mathcal{O}(N/2)$ in the case of the sequential search algorithm, we say that the binary search algorithm has order $\mathcal{O}(\log N)$.

For very large N, binary search will be more efficient than sequential search. Of course, in special cases, sequential search would be better. (Suppose that B happens to = A(1).) We must ask for the asymptotic execution time of the algorithm that performs best for the *hardest* or *worst-case* data. That is the true test.

9. Classes P and NP

A problem that can be solved by an algorithm for which the running time is bounded (above) by a polynomial function $P_m(N) = a_m N^m + a_{m-1} N^{m-1} + \cdots a_0$ (with $a_m > 0$) of the problem size N – that is, the algorithm is $\mathcal{O}(P_m(N))$ – is said to be in *class P*. (Because for sufficiently large N the highest-power term will dominate, $\mathcal{O}(P_m(N))$ is equivalent to $\mathcal{O}(N^m)$.) Because $\mathcal{O}(\log N)$ is even *better* than $\mathcal{O}(N^m)$, our table-look-up problem, and other problems with $\mathcal{O}(\log N)$ optimal algorithms, are also in class P.

However, suppose that the optimal algorithm to solve a problem has order worse than polynomial – for instance exponential order $\mathcal{O}(10^N)$ – but that a *proposed* solution can be *checked* in polynomial time. Such a problem is said to be of *class NP*. (NP stands *not* for nonpolynomial, but for nondeterministic polynomial, referring to a different model for the computation. Do not worry about this distinction.)

Consider the problem of sorting a list of numbers. That is, given a series of N numbers: 2, 1, 7, 5, 8, 4, 3; an algorithm must produce as output the numbers rearranged from smallest to largest: 1, 2, 3, 4, 5, 7, 8. Whatever be the order of the optimal algorithm that solves the problem, an algorithm to *verify* that 1, 2, 3, 4, 5, 7, 8 is a solution (or that 1, 8, 7, 2, 4, 5, 3, is *not*) can run in linear time. Therefore, sorting a list of numbers into order is in class NP. (Sorting happens also to be in class P; sorting algorithms are known with order $\mathcal{O}(\log N)$.) Figure 1 shows the topology of complexity space, on the assumption that P is a proper subset of NP. (Whether or not this is true is an outstanding open problem in computer science; see below.)

Now, an obvious difficulty with the definition of class P is this: to define the complexity of a problem in terms of the complexity of an optimal algorithm, we must know the optimal algorithm. Some problems have been proved to be not in class P. But other problems may *appear* not to be in class P only until someone discovers a polynomial-time algorithm that solves them. The best currently known algorithm for any problem gives an *upper bound* to its computational complexity.

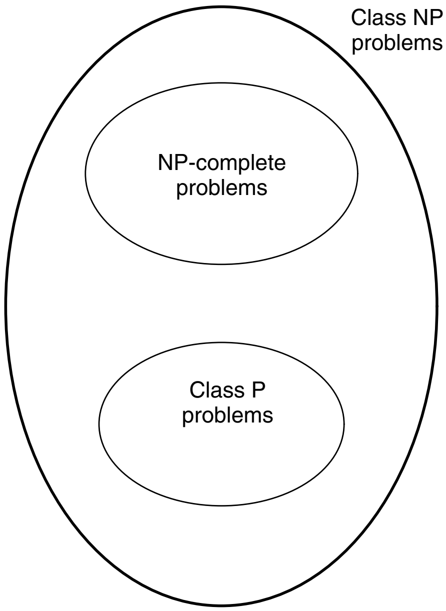


Figure 1 Computational problems can be in class P = problems for which algorithms of polynomial asymptotic order are known; NP-complete problems, a set of problems for which no algorithms of asymptotic polynomial-time order are known but which are reducible to one another in the sense that the discovery of an algorithm of asymptotic polynomial-time order for one of them (proving it to be of class P) would show that all NP-complete problems are of class P, or NP problems, for which all optimal algorithms are provably nonpolynomial. If it is found that P = NP, class P would be expanded to fill the entire class NP set

An example of such a reclassification was the discovery of a polynomial-time algorithm for linear programming (LP), the problem of maximizing or minimizing a linear function subject to linear equality and/or inequality constraints. It has a wide variety of important scientific and commercial applications. The *simplex algorithm* for solving LP problems was developed by G. Dantzig during World War II. This algorithm is quite effective in practice, but does not solve worst-case problems in polynomial time. (However, S. Smale showed that worst-case LP problems are sparse, accounting for the practical success of the simplex algorithm.)

For many years, we simply did not know whether linear programming is in class P or not. Then, in 1979, L. Khachyan presented a polynomial-time algorithm for LP. Although this resolved the theoretical problem – linear programming *is* in class P – it did not in itself provide software that would beat the simplex algorithm in the field. Five years later, N. Karmarkar devised a polynomial-time algorithm that did lead to effective programs.

10. NP-complete problems. Does P = NP?

Many NP problems have equivalent complexities, in the sense that if a polynomial algorithm were discovered for any one of them, it could be applied to solve others.

The set of *NP-complete problems* is the subset of NP problems, such that if we could solve any of them in polynomial time, we would be able to solve all of them in polynomial time. In other words, the discovery of a polynomial-time algorithm for *any* problem known to be NP-complete would cause the classes P and NP-complete in Figure 1 to coalesce. Are there *any* NP problems that are NOT in class P? This is the famous question: Does P = NP? A prize of $1 million is on offer for an answer.

Given that (1) many problems we wish to solve are NP-complete and (2) if we ask computer scientists for help with a problem, they are likely to say – typically with a sniff – "sorryit'sNPcompletegoodbye"; what is a poor biologist to do? In practice:

1. It is not always a vain hope that the worst-case behavior does not happen too often. (The simplex algorithm was successful for that reason.)
2. Often good approximations are available. Algorithms for some optimization problems often find an optimum or near-optimum quickly, and then spend a lot of time verifying optimality.

Theoretical barriers are genuine obstacles, but do not entirely exclude regions of practical feasibility and of genuine interest and utility.

11. Relevance of computational complexity to biological complexity

In applying ideas of computational complexity to biological systems, keep in mind that computational complexity describes the complexity of the *problem,* not the complexity of the device that solves it. Certainly, living systems function as computers. If I add 2 + 2 and get 4, I am doing a relatively simple computation; if I try to predict which horse will finish first in a race, I am doing a more complicated one. The regulatory activities that biological systems carry out are also optimization calculations. An example described by S. Brenner is the growth of bacteria in heavy water. Changing from H_2O to D_2O has the effect of changing the kinetic constants of many enzymatic reactions. After a relatively short period, cells readjust and resume activity and growth.

However, the theory of computational complexity is tied very closely to specific models of computations, which may not correspond to the computational architectures active in biology. The classic literature on computational complexity is based on computers that execute programs *sequentially*. Biological computers do lots of parallel processing. Exploration of generalizations of the classic results are a current research topic in computer science.

12. Computational complexity and sequence complexity

Let us revisit the two algorithms for determining the position of a number in a sorted list: simple sequential search and interval bisection. Given

a fixed finite sorted sequence of integers, each application of either algorithm to a test value generates a sequence of numbers as its search history. The search-history sequences generated by the two algorithms have different characteristics.

The search-history sequences generated by the sequential search method will have a distribution skewed to the earlier numbers in the list. The probability of checking the number in position i of the list is proportional to $N - i + 1$, where N is the length of the list. The search-history sequences generated by the binary search algorithm have a different distribution. We can calculate the entropies of these search-history sequences, and apply that and other tools to measuring their complexity as *sequences*. We find that the sequences generated by binary search are more complex than those generated by sequential search.

This provides an important connection between *complexity of structure* and *complexity of process*. We can collect the historical records of a process, and treat them as a structure. We can apply ideas of predictability and complexity of structures to these historical records, to give insight into the complexity of the process.

This relationship between the predictability of the course of a process and its complexity leads us naturally to reconsider the idea of chaos.

13. Predictability and chaos

The discovery of the laws of mechanics in the seventeenth century – Newton's *Principia* was published in 1687 – gave rise to the hope that the dynamics of the solar system (at least!) was predictable. Laplace wrote:

> *If we can imagine a consciousness great enough to know the exact locations and velocities of all the objects in the universe at the present instant, as well as all forces, then there could be no secrets from this consciousness. It could calculate anything about the past or future from the laws of cause and effect (quoted in Peitgen et al., 2004).*

Indeed, there were important implications about determinism in many-particle systems and even for free will in human behavior.

There were also questions of computability. How much information do we need, and how accurately do we need it, to predict the dynamics of the solar system? the weather? the universe? In some cases, the physical development of a system is relatively insensitive to the initial conditions. But in others, very small changes in initial conditions will lead to very different dynamic development of the system. Weather is an example: the meteorologist Lorenz gave a talk entitled: "Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" Accurate prediction of the dynamics of such systems requires unachievably accurate knowledge of the initial conditions. Such systems exhibit chaotic dynamics. The courses of their development are unpredictable, and would remain unpredictable even if unlimited computational resources were available. Indeed, at the atomic level, Heisenberg's uncertainty principle killed off Laplace's hope of perfect determinism.

We can still make some predictions about the behavior of complicated systems. For instance, we can derive general macroscopic properties of a sample of a perfect

gas, such as $PV/T =$ constant, by averaging, even if we cannot follow the paths of the individual particles. Moreover, even chaotic (classical) systems are subject to Poincaré's recurrence principle: any system of particles held at fixed total energy will return arbitrarily closely to any set of initial molecular positions and velocities. What rescues the second law of thermodynamics is that the closer the reapproach demanded, the longer the time required, and the less probable the fluctuations required to achieve it.

In terms of complexity, chaotic dynamics is associated with unpredictability. However, chaotic behavior is not entirely incompatible with order and even the "spontaneous" generation of order. Many sequences associated with chaotic behavior have a *fractal* structure. This means that if an object is dissected, the parts have a structure similar to that of the whole (as well as to one another). Many beautiful fractal images have appeared (see, for example, Mandelbrot, 1982). This self-similarity at different scales implies that if we know part of such a structure we can predict a larger segment of it. This predictability should permit compressibility, and effectively reduce complexity. Indeed, fractal image compression is an effective tool for reducing the sizes of images (Barnsley, 1993).

Chaotic dynamics even sometimes produces stable states or approximations to stable states – called *attractors.* Some attractors are periodic and/or localized states. Others are *strange attractors,* which have a fractal structure. There have been examples of generation of order in model systems evolving "at the edge of chaos" (Kauffman, 1993).

We have now examined ideas of complexity for sequences, and for processes. Many biological phenomena can be described in terms of networks. Networks have both structure and dynamics. Can we apply the ideas developed to describe and analyze them?

14. Structure and complexity of networks

A network, or graph, is an abstraction of several types of biological phenomena in terms of a set of *nodes,* some or all pairs of which are joined by *edges* (see, *e.g.,* Strogatz, 2001). A *path* in a graph is a sequence of nodes $x, y, \ldots z$ with an edge between each successive pair of nodes. A graph containing a path between any two pairs of nodes is called *connected.*

The idea of a network is very general, including most if not all of the data types of modern biology. Sequences are networks: the nucleotides or amino acids are nodes, and edges connecting successive residues specify the order. The chemical bonding patterns of molecules are also networks, in which atoms are the nodes and bonds the edges. Later we shall see that abstract protein-folding patterns, defined in terms of contact patterns between helices and strands of sheet, are also representable by networks.

Microarray data on protein expression patterns or gene detection give rise to networks of relationships between genes. Although the output of a single chip is usually presented as a two-dimensional image, *topologically* these data are not two-dimensional. If two microarray patterns arise from comparable oligo sets – such as an experiment to determine the effect of a drug, or to determine the difference in

expression patterns between cancerous cells and normal tissue – the correspondence between the spots on the two arrays amounts to an alignment of a pair of sequences. An alignment is also a network, in which the nodes are the individual residues in the sequences, or the individual genes in the expression array, and the edges link both successive elements of individual sequences, and elements in different sequences or different arrays that are associated in the alignment.

Gene expression tables – matrices in which the rows correspond to different genes and the columns to different samples – are often analyzed by principal component analysis. This procedure derives, from the matrix, a sequence of eigenvalues. The extent to which the few largest eigenvalues account for the variability provides a measure of the complexity of the table.

15. Structures of networks

Typical networks in molecular biology include metabolic pathways, in which the nodes are metabolites and the edges are reactions that interconvert them; and gene regulation networks in which the nodes are genes and the edges are relationships between genes such that the expression of one gene activates or represses another.

Other familiar graphs include the World Wide Web, and networks of human acquaintance. "Six degrees of separation", the title of a play by John Guare, made into a film, refers to the assertion (attributed to Marconi) that if we form a graph with all human beings as the nodes, and connect by edges all pairs of people who know each other, then any two nodes are connected by a path of length ≤ 6.

The *density of connections* = the mean number of edges per vertex is an important characteristic of the structure of a network that governs its behavior. Nervous systems of higher animals achieve their power not only through having a large number of neurons but also by high degrees of connectivity.

In some networks, connectivities follow observable regularities. The World Wide Web is a network in which individual documents are nodes and hyperlinks are edges. The distribution of incoming and outgoing links follow power laws: $P(k) =$ probability of k edges $\sim k^{-q}$, where $q = 2.1$ for incoming links (other links that cite the node in question), and $q = 2.45$ for outgoing links (links cited by the node in question) (Albert and Barabási, 2002).

The density of connections is an important determinant of the dynamic behavior of a network. For instance, the interactions that spread disease among humans and/or animals form a network. As the density of connections increases, the system can exhibit a qualitative change in behavior, analogous to a phase change, from a situation in which the disease remains under control, to an epidemic spreading through an entire population.

16. "Small-world" networks

The Web exemplifies many networks that have the characteristics of high clustering and short path lengths. They contain relatively few nodes with large numbers of connections, called "hubs", and many nodes with few connections. These produce

short path lengths between all nodes. From this feature, they are called "small-world networks". Such networks tend to be fairly robust – staying connected after failure of random nodes. Failure of a hub would be disastrous but is unlikely, because there are few hubs. Hubs exist in biological networks also. Irizarry *et al*. (2005) suggest that regulatory networks of genomes may be partitioned into self-contained modules.

Many networks, notably the World Wide Web, are continuously adding nodes. The connectivity distribution tends to remain fairly constant as the network grows. These are called 'scale-free' networks (see Albert and Barabási, 2002; Barabási, 2003).

17. Robustness of networks

Biological systems need to be robust, both for survival of individuals under stress and for the plasticity required for evolution. In yeast, for example, single gene knockouts of over 80% of the ∼6200 open reading frames are survivable injuries.

In principle, networks can achieve robustness through redundancy. Wagner (2004) distinguished:

1. *Substitutional redundancy*: If two proteins are both capable of a job, knock out one and the other takes over. For example, it was found during attempts to develop mouse models for diabetes that mice and rats (but not humans) have two similar but nonallelic insulin genes. Substitutional redundancy requires equivalence not only of function but of control of expression. In the mouse, knocking out either insulin gene leads to compensatory increased expression of the other, and produces a normal phenotype. Such equivalent expression patterns are more probable among duplicated genes than among unrelated ones with equivalent functions.

2. *Distributed redundancy*: The same effect is achieved through different routes. In normal *E. coli*, approximately $\frac{2}{3}$ of the NADPH produced in metabolism arises via the pentose phosphate shunt, requiring the enzyme glucose-6-phosphate dehydrogenase. Knocking out this enzyme leads to increased NADH production by the tricarboxylic acid cycle and conversion of NADP → NADPH. The growth rate of the knockout strain is comparable to that of the parent.

In principle, redundancy – more obviously, substitutional redundancy – should permit compressibility in the description of a network, and thereby a reduction in its complexity relative to a nonredundant network of equal size and comparable topology.

18. Network structure and dynamics

For some networks such as metabolic pathways or patterns of traffic on city streets, the *dynamics* of the system depend on the transmission capacities of the links. In molecular biology, metabolic pathways and signal transduction cascades are networks that lend themselves to pathway and flow analysis. For control and signaling pathways, much is known about the mechanisms of individual elements, and understanding their integration is a subject of current research.

19. Structures of regulatory networks

Regulatory interactions are organized into linear signal transduction cascades, and reticulated into control networks. Any regulatory action requires (1) a stimulus, (2) transmission of a signal to a target, (3) a response, and (4) (in most cases) a "reset" mechanism to restore the resting state. Many regulatory actions are mediated by protein–protein complexes. Transient complexes are common in regulation, as dissociation provides a natural reset mechanism.

Control, or regulatory networks, are assemblies of *activities*. They:

1. *tend to be unidirectional.* A transcription activator may stimulate the expression of a metabolic enzyme, but the enzyme may not be involved directly in regulating the expression of the transcription factor.
2. *have a logical component.* It is not enough to describe the connectivity of a regulatory network. Any regulatory action may stimulate or repress activity of its target. If two interactions both influence a common target, activation may require *both* stimuli (logical "and"), or *either* stimulus may suffice (logical "or").
3. *can produce dynamic patterns.* Signals may produce combinations of effects with specified time courses. Cell-cycle regulation is an example.

20. Common motifs in biological control networks

Certain patterns of connectivity recur frequently in regulatory networks. These building blocks contribute to higher levels of organization on which a quantitative definition of network complexity will ultimately depend. Shen-Orr *et al*. (2002) have described: the *fork*, the *scatter*, and the *"one-two punch"* (a phrase from the boxing ring) (see Figure 2).

The *fork*, or the single-input motif, transmits a single incoming signal to two outputs. Generalizations include more downstream genes under common control (more tines to the fork), and autoregulation of the control node. Forks can achieve general mobilization. Moreover, if the genes regulated have different thresholds for activation, the dynamics of building up of the signal can produce a temporal pattern of successive initiation of expression of different genes.

The *scatter* configuration, also called the multiple input motif, can function as a logical "or" operation: both downstream targets become active if *either* of the input impulses is active. Note that the scatter pattern is a superposition of two forks.

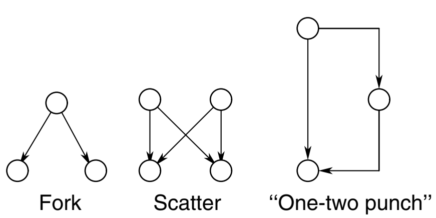


Figure 2 Three common substructures of regulatory networks: the fork, the scatter, and the "one-two punch"

The *"one-two punch"*, also called the "feed-forward loop", affects the output both directly through the vertical link and indirectly and subsequently through the intermediate link. This motif can show interesting temporal behavior if activation of the target requires simultaneous input from direct and indirect paths (logical "and"). Because the buildup of intermediates requires time, the direct signal will arrive before the indirect one. A short pulsed input will not activate the output – by the time the indirect signal builds up, the direct signal is no longer active. The system can thereby filter out transient stimuli in noisy inputs.

To identify common motifs that are similar but not identical, it is necessary to compare different networks (or subnetworks) and measure the differences. The model of such calculations for sequences is based on alignment. Berg and Lässig (2004) have presented an algorithm for alignment of subgraphs, and used it to identify higher-order motifs in networks.

21. Reconfiguration of networks – another level of dynamics

The transcription regulation network in yeast includes 142 genes that encode transcription regulators, and 3420 that encode target proteins exclusive of transcription regulators, corresponding to approximately half the known proteome of *Saccharomyces cerevisiae*. The 7074 known regulatory interactions among these genes include effects of regulators on one another and of regulators on nonregulatory targets.

The high ratio of interactions-to-transcription regulators implies that individual regulatory molecules do not have single, dedicated activities (as most metabolic enzymes do). Instead, the network involves the coordinated activities of many components.

The regulatory network achieves versatility and responsiveness through an ability to reconfigure its connectivity. Babu *et al*. (2004) compared changes in the networks controlling yeast gene expression patterns, in different physiological regimes: cell cycle, sporulation, diauxic shift (the change from anaerobic fermentative metabolism to aerobic respiration as O_2 levels increase), DNA damage, and stress response. Cell cycling and sporulation involve the unfolding of endogenous gene expression programs; the others are responses to environmental changes.

Different states are characterized by similarities and differences in gene expression patterns, and in the components of the regulatory network that are active. Think of the transcription factors as "hardware" and the connections as reprogrammable "software": The molecules do not change but the interactions do. In different states, many transcription regulators change most, or a substantial part, of their interactions. In particular, the set of transcription regulators that form the hubs of the network – those with many outgoing nodes that form foci of control – differs in different states. Some hubs are common to all states, but others step forward to take control in different physiological regimes.

The changes in the active interaction patterns alter the topological characteristics of the network. Under panic conditions – DNA damage and stress – the average number of genes under control of individual transcriptional regulators increases, the average minimal path length between regulator and target decreases, and the

clustering becomes less dense (i.e., there is less interregulation among transcription factors). This corresponds to a need for fast and general mobilization. Normal circumstances – cell-cycle control for instance – allow for a more dignified and precise regulatory state, permitting finer control over the temporal course of expression. In cell-cycle control and sporulation, there is a much denser interregulation among transcription factors, and longer minimal path lengths between transcriptional regulators and target genes.

Different physiological states also differ in their usage of the common motifs – fork, scatter, and "one-two punch". Forks are more used in conditions of stress, diauxic shift, and DNA damage. They satisfy the need for quick action. The "one-two punch" motif is more common in cell-cycle control. This is consistent with the need for a signal from one stage to be stabilized before the cycle proceeds.

Much of evolution proceeds toward greater specialization. Many evolutionary pathways show a trade-off between specialized adaptation and generalized adaptability. Regulatory networks are an exception. Evolution has produced structures that are both specialized *and* versatile. The reconfigurability of regulatory networks allows them to respond robustly to changes in conditions, by creating many different structures, specialized to the conditions that elicit them. This also adds additional dimensions to their complexity.

22. Complexity of protein structures

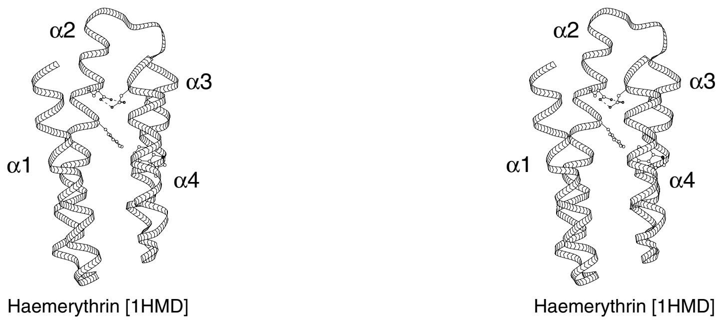
Anyone who has tried to assemble a "do-it-yourself" furniture has encountered structural complexity. Many have suspected that the difficulty lies in the complexity of the description rather than the complexity of the structure (which in a way echoes the ideas of Kolmogorov).

A paradox in attempts to apply Kolmogorov's ideas to protein structures is that the shortest representation of a protein structure is an amino acid or DNA sequence. In this context, the ribosome is a decompression algorithm! (Protein structures are also representable as distance matrices, with the position, orientation, and enantiomorph undetermined (Crippen and Havel, 1988). It is interesting that proteins have one-, two-, *and* three-dimensional representations!)

23. Representation and analysis of protein-folding patterns as networks

If we put aside the representation of protein structures by their sequences, we know that protein structures show a very wide variety of topologies, some apparently more complex than others. Related proteins tend to retain, but unrelated proteins may share, similar folds. Visual examination of structures has led to qualitative classifications of protein-folding patterns (see, e.g., Lesk, 2001).

What is the essence of a protein-folding pattern? If it is defined by the order along the chain of the helices and strands of sheet, and the geometry of interaction of the pairs of secondary structure elements that are in contact, then this information



(a)

Haemerythrin [1hmd]				
	α1	α2	α3	α4
α1	α1	↑↓ **S**		↑↓ **S**
α2	↑↓ **S**	α2	↑↓ **S**	| | **D**
α3		↑↓ **S**	α3	↑↓ **S**
α4	↑↓ **S**	| | **D**	↑↓ **S**	α4

(b)

Figure 3 (a) Structure of haemerythrin, a 4-helix bundle (Protein Data Bank code 1hmd). The chevrons indicate the direction of the polypeptide chain, from N- to C-terminus. (b) Tableau representation of its folding pattern. The rows and columns are labeled by the helices appearing successively in the chain, α_1, α_2, α_3, and α_4, and these labels are repeated along the main diagonal. The off-diagonal elements encode the relative orientation of the helices that are in contact (for details, see Lesk, 1995; Lesk, 2003)

can be encapsulated in a square symmetric matrix or tableau (Lesk, 1995; Lesk, 2003). The rows and columns are labeled by the elements of secondary structure, α-helices and strands of β-sheet, in order of appearance in the amino acid sequence. Each off-diagonal position is either blank if the corresponding pair of secondary structure elements is not in contact or encodes the relative orientation of secondary structure elements that are in contact. It is convenient to list the names of the secondary structure elements along the main diagonal (see Figure 3).

24. Use of the tableau representation to measure complexity of protein-folding patterns

Can the ideas of Kolmogorov be applied to the tableau or other representations of protein-folding patterns, to measure complexities of protein structures? What is the minimal amount of information that specifies a folding pattern?

The tableau representation of protein-folding patterns makes it possible to address these questions. By regarding the tableau as a succession of diagonals, we can ask for the smallest part of the tableau that uniquely determines the protein fold. The main diagonal contains only the *order* of secondary structural elements. This

information alone does not specify the folding pattern; for all α-helical proteins contain the same main diagonal: $\alpha 1$, $\alpha 2$, $\alpha 3$, ... One can then ask whether the main diagonal plus the adjacent diagonal of such a tableau – which together specify the secondary *and* supersecondary structure – are enough to uniquely determine the fold; if not, how many diagonals are necessary? Alternatively, Przytycka *et al*. (1999) have shown that the sequence of secondary structural elements, together with the lengths of the loops between them, generally suffice to determine the protein fold. These approaches, the subjects of current research, involve an application of Kolmogorov's ideas to the analysis of complexity of structures.

25. Complexity of the protein folding process

Another measure of protein structural complexity is *contact order* (Plaxco *et al*., 1998). Contact order measures the average separation in the *sequence* of residues that are close by in *space:*

$$\text{Contact order} = \frac{1}{LN}\sum(i-j) \tag{4}$$

where L is the number of residues in the protein, N is the number of residues in contact, $i-j$ is the distance in the sequence between residues i and j, and the sum is taken over all pairs of residues in contact.

Not only does this sound like a reasonable measure of structural complexity according to naive notions, it may well correspond to the proteins' idea of complexity: for proteins with simple two-state folding behavior, the folding rate is strongly correlated with the contact order.

In more complex cases with multiple transition states or even stable intermediates, the folding process of a peptide or protein is describable in terms of a network. The nodes of the network are the localized free energy minima in conformation space, and the edges describe the dynamics of interconversion among the nodes.

Krivov and Karplus (2004) analyzed the potential energy surface for folding of the β-hairpin from protein G, a 16-residue peptide with a well-defined native structure. They generated a population of 2×10^5 conformations by molecular dynamics at 360 K for 4 µs. By clustering the conformations, Krivov and Karplus showed that the conformational space for folding contains many basins. The topology of fold space can be regarded as broken up into discrete states, each a distinguishable cluster of related conformations. Continuous trajectories between the discrete states of this model must exist, but if the system spends most of its time in states corresponding to these clusters, it is appropriate to regard folding space as a network and the folding process as a succession of transitions within this network (see Figures 4 and 5).

26. Conclusions

The subject of complexity will be a fertile source of research problems in biology for many years. The basic challenge is to define concepts that will support

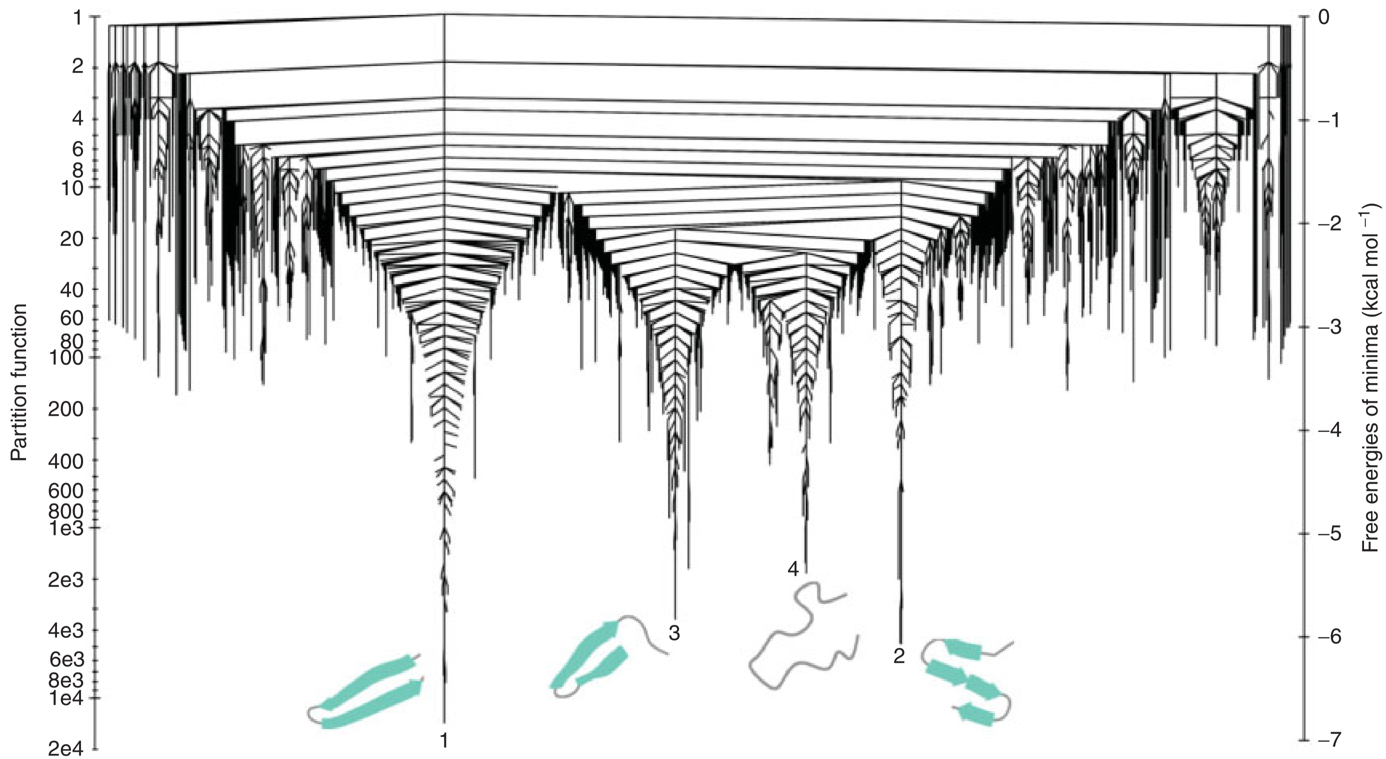


Figure 4 Plot of conformation space for the β-hairpin from protein G. The minimum-energy conformation, 1, corresponds to the native state. Other states appear with different energy distributions and with different barriers between them. The conformations illustrated (green) are representative of the states in the cluster and not necessarily those uniquely at the minimum-energy point of the cluster. Because the topology of conformation space is reduced in this figure to one-dimension (the horizontal axis), the distribution of probabilities of interconversion between states is not evident from this diagram (see Figure 5) (Reproduced from Krivov and Karplus (2004) by permission of National Academy of Sciences U.S.A.)

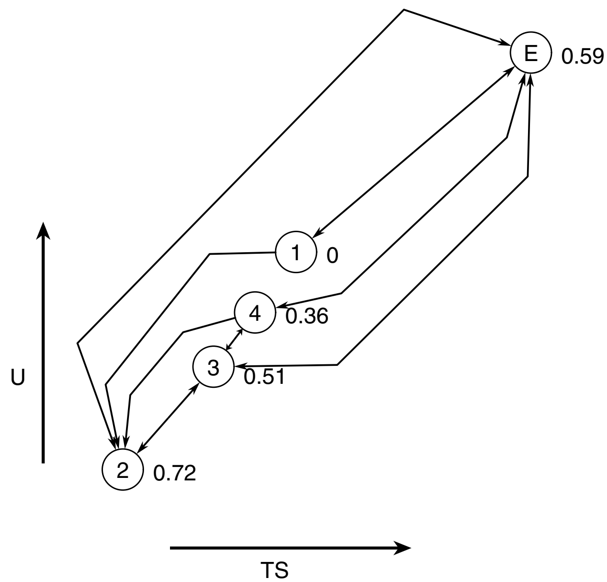


Figure 5 Transition matrix between major nodes in the folding network for the β-hairpin from protein G. (Redrawn from data in Krivov, S. V. and Karplus, *loc. cit.*) The labeled circles are plotted on axes showing the values of average energy U and TS of each state depicted in Figure 4. The numbers next to the circles are the relative Helmholtz free energies of the states, relative to the lowest (state 1). Lines of constant Helmholtz free energy are diagonal lines in the southwest–northeast direction. The fact that such a line almost passes through all the points shows that the differences in free energy are small compared to the individual differences in U and TS. The end states of the folding process are E (denatured) and 1 (hairpin, highest stability – see Figure 4). In understanding the relationship between this and the previous figure, recognize that the values of U and TS are computed as an average over the states in the basin surrounding the structure depicted, and that the transition probabilities are determined by features of the states near the *top* of the barriers, not by the conformations at the minima. The transition probabilities between the states, which determine the folding pathway, are very unevenly distributed. (Note, from the absence of arrows, the very low probability of a state 2 → state 4 and a state 1 → state 2 transition.) In particular, transitions between the E state and states 3 and 4 are much more probable than any others. Nevertheless, the major folding transition is directly from E → 1, and the other states, including states 3 and 4, are "off-pathway". In the language of networks, the E state is the "hub" of the folding process.

analysis of data on sequences, expression patterns, structures and interactions of proteins and nucleic acids, and metabolic and regulatory networks. In this article, I have tried to provide as background, for the general biological researcher, results of analyses of complexity in other fields, in a way applicable to biological research.

The basic ideas for the reader to take away are:

- Complexity is related to the extent to which knowledge of part of an object allows prediction of the whole, or the extent to which the portion of a process already observed determines the part that is yet to come.

- This characterization of complexity is based on the ideas of Kolmogorov that the complexity of a sequence depends on its shortest possible description, and the related idea that the complexity is related to compressibility.
- Computational complexity, in principle, describes characteristics of problems, but is related to the complexity of processes.
- Complexity is a characteristic both of the structure of an object and of its dynamics.
- Many biological phenomena, including metabolic and regulatory networks, and even protein structures and folding pathways, are amenable to treatment by the same basic ideas that have been well worked out for sequences. However, our understanding of the underlying principles of their structures and dynamics is only now emerging.

Related articles

Article 96, **Fundamentals of protein structure and function**, Volume 6; Article 97, **Structural genomics – expanding protein structural universe**, Volume 6; Article 107, **Integrative approaches to biology in the twenty-first century**, Volume 6; Article 10, **Graphs and metrics**, Volume 7; Article 88, **Bioinformatics pathway representations, databases, and algorithms**, Volume 8

Acknowledgments

I thank P. Berman, N. Faux, and M. E. Lesk for helpful discussion, and R. Casadio for the opportunity to present this material at the Sixth Bologna Winter School on Bioinformatics.

Further reading

Bornholdt S and Sneppen K (2000) Robustness as an evolutionary principle. *Proceedings of the Royal Society of London*, **B267**, 2281–2286.

Ideker T (2004) A systems approach to discovering signaling and regulatory pathways – or, how to digest large interaction networks into relevant pieces. *Advances in Experimental Medicine and Biology*, **547**, 21–30.

Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q and Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology*, **4**, 805–809.

References

Adam C (2002) What is complexity? *BioEssays*, **24**, 1085–1094.

Albert R and Barabási A-L (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47–97.

Babu MM, Luscombe NM, Aravind L, Gerstein M and Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, **14**, 283–291.

Barabási A-L (2003) *Linked: How Everything Is Connected to Everything Else and What It Means*, Plume Books: New York.

Barnsley MF and Hurd LP (1993) *Fractal Image Compression*, Peters: Wellesley.

Berg J and Lässig M (2004) Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14689–14694.

Colosimo A (Ed.) (1997) *Complexity in the Living: a Modelistic Approach*, Università di Roma "La Sapienza": Rome.

Colosimo A (Ed.) (2004) *Proceedings of the Complexity in the Living 2004/ A Problem-oriented Approach*, Università di Roma "La Sapienza": Rome.

Cook S (2003) The importance of the P versus NP question. *Journal of the Association for Computing Machinery*, **50**, 27–29.

Cramer F (Translator) and Loewus DI (1993) *Chaos and Order: The Complex Structure of Living Systems*, John Wiley & Sons: New York.

Crippen GM and Havel TF (1988) *Distance Geometry and Molecular Conformation*, John Wiley & Sons: New York.

Garey M and Johnson D (1979) *Computers and Intractability*, W.H. Freeman: New York.

Hartmanis J (1989) Gödel, von Neumann and the P = ?NP problem. *Bulletin of the European Association for Theoretical Computer Science*, **38**, 101–107.

Hopcroft JE, Motwani R and Ullman JD (2000) *Introduction to Automata Theory, Languages, and Computation*, Second Edition, Addison-Wesley: Reading.

Huffman DA (1952) A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, **40**, 1098–1101.

Irizarry KJL, Merriman B, Bahamonde ME, Wong M-L and Licinio J (2005) The evolution of signaling complexity suggests a mechanism for reducing the genomic search space in human association studies. *Molecular Psychiatry*, **10**, 14–26.

Kauffman SA (1993) *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press: Oxford.

Krivov SV and Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proceedings of the National Academy of Sciences*, **101**, 14755–14770.

Lesk AM (1995) Systematic representation of protein folding patterns. *Journal of Molecular Graphics*, **13**, 159–164.

Lesk AM (2001) *Introduction to Protein Architecture: The Structural Biology of Proteins*, Oxford University Press: Oxford.

Lesk AM (2003) From electrons to proteins and back again. *International Journal of Quantum Chemistry*, **95**, 678–682.

Li M and Vitanyi PMB (1997) *An Introduction to Kolmogorov Complexity and Its Applications*, Second Edition, Springer-Verlag: New York.

Mandelbrot B (1982) *The Fractal Geometry of Nature*, W. H. Freeman and Co: New York.

Papadimitriou CH (1994) *Computational Complexity*, Addison-Wesley: Reading.

Papadimitriou CH (1997) NP-completeness: A retrospective, *Proceedings of the 24th International Colloquium on Automata, Languages and Programming*, Victoria, Canada. pp. 2–6.

Peitgen H-O, Jürgens H and Saupe D (2004) *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag: New York.

Plaxco KW, Simons KT and Baker D (1998) Contact order, transition state placement, and the refolding rates of single domain proteins. *Journal of Molecular Biology*, **277**, 985–994.

Przytycka T, Aurora R and Rose GD (1999) A protein taxonomy based on secondary structure. *Nature Structural Biology*, **6**, 672–682.

Rosenkrantz RD (Ed.) (1983) *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, Springer: New York.

Shen-Orr SS, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, **31**, 64–68.

Sipser M (1992) The history and status of the P versus NP question, *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, Victoria, British Columbia, pp. 603–618.

Strogatz SH (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Wagner A (2004) *Distributed Robustness Versus Redundancy as Causes of Mutational Robustness*, SFI Working Paper 04-06-018, Santa Fe Institute: Santa Fe.

Welch T (1984) A technique for high-performance data compression. *Computer*, **17**, 8–19.

Ziv J and Lempel A (1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, **IT-23**, 337–343.

Specialist Review

Modeling by homology

Kenji Mizuguchi
University of Cambridge, Cambridge, UK

1. Introduction

Comparison of the three-dimensional (3D) structures of homologous proteins shows a clear correlation between structural similarity and sequence identity (Figure 1). (Note that homology means common ancestry (Reeck *et al*., 1987).) Chothia and Lesk's analysis (Chothia and Lesk, 1986) was among the first to quantify this relationship; they defined the common core of a homologous protein family to be the superimposable elements of secondary structure (*see* Article 76, **Secondary structure prediction**, Volume 7) and adjacent residues, and obtained a relationship between the root mean square deviation (RMSD) of the core backbone atoms and the percentage sequence identity (PID). Although for distantly related protein pairs (with PID $<$ 25%), the common core may account to as little as 40% of the structure, the equivalent atoms can be still superimposed with a reasonably low RMSD ($<\sim$1.5 Å).

Therefore, if a homologous relationship can be established between a pair of proteins and the 3D structure of one of them has been determined experimentally, it is possible to predict the structure of the other on the basis of the known structure. The prediction accuracy increases as the PID between the two proteins increases. This possibility had been realized long before the formal relationship between sequence and structural similarity, as discussed above, was established. Browne *et al*. (1969) proposed a model of α-lacalbumin based on the crystal structure of lysozyme. Subsequently, this "modeling by homology" approach was more formally established, and fully automated methods for comparative (or homology) modeling have been developed (for example, by Sutcliffe *et al*., 1987a,b; Sali and Blundell, 1993).

In the 1969 paper, Browne *et al*. mentioned, " ... the very many protein molecules may belong to a comparatively small number of families. It may be possible to derive the structures of all the members of a family relatively easily once one or two have been analysed in detail. Certainly this approach seems more promising at the moment than does the prediction of unique protein structures from chemical information alone." Thirty-five years on, comparative modeling remains the most reliable method for predicting protein 3D structures in atomic detail, and there are

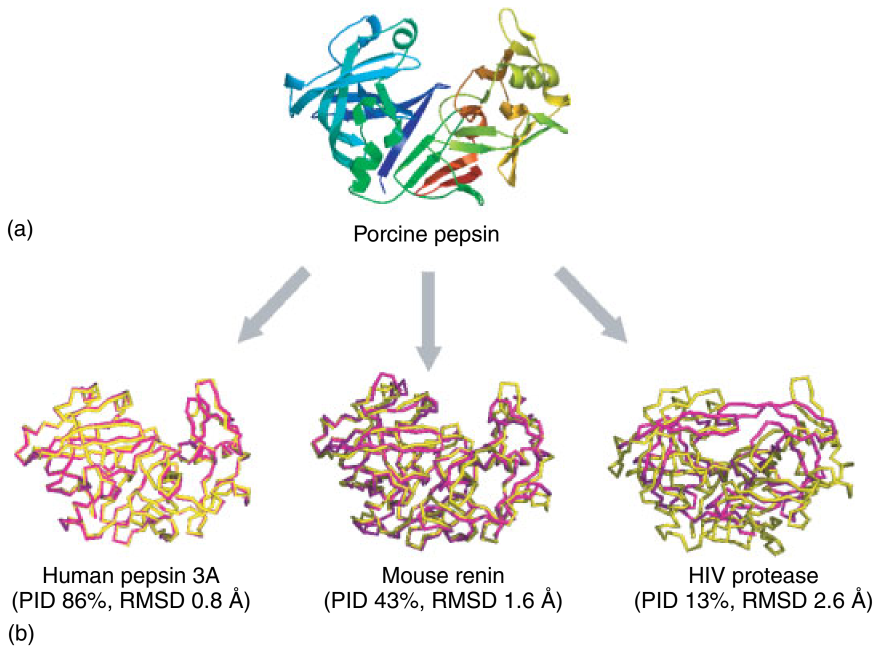


Figure 1 Relationship between sequence and structural similarity among homologous proteins. A cartoon drawing of the structure of porcine pepsin (Protein Data Bank (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7) code 5pep; a) and its superpositions (b) with the structures of human pepsin 3A (1psn), mouse renin (1smr), and HIV protease (1kzk). PID: percentage sequence identity; RMSD: root mean square deviation of the equivalent Cα atoms (see text). Figure generated with PyMOL (**http://www.pymol.org**)

now databases (Kopp and Schwede, 2004; Pieper *et al*., 2004) that provide easy access to a large number of automatically built comparative models.

Key to the success of this approach is to identify homologous protein pairs. Conventionally, this was achieved by standard sequence comparison methods. To apply comparative modeling, BLAST (Altschul *et al*., 1990) or other sequence database searches must identify one or more homologs of known structure (often referred to as "templates" or "parents" in comparative modeling although they are more like cousins in evolution). If no obvious homologs can be obtained in this way, searching for potential templates (remote homologs or even analogous structures with no evolutionary link) is considered a different kind of challenge. A series of new methods (Bowie *et al*., 1991; Jones *et al*., 1992), now collectively known as *fold recognition techniques*, have been developed for this purpose. Comparative modeling and fold recognition have been considered related but separate attempts, as demonstrated in two distinct categories of the Critical Assessment of Techniques for Protein Structure Prediction (CASP (*see* Article 73, **CASP**, Volume 7; Moult *et al*., 2003)) exercises.

In recent years, however, significant advances have been made to identify distant evolutionary relationships between proteins that show very weak sequence similarity. As a result, the conventional distinction between comparative modeling

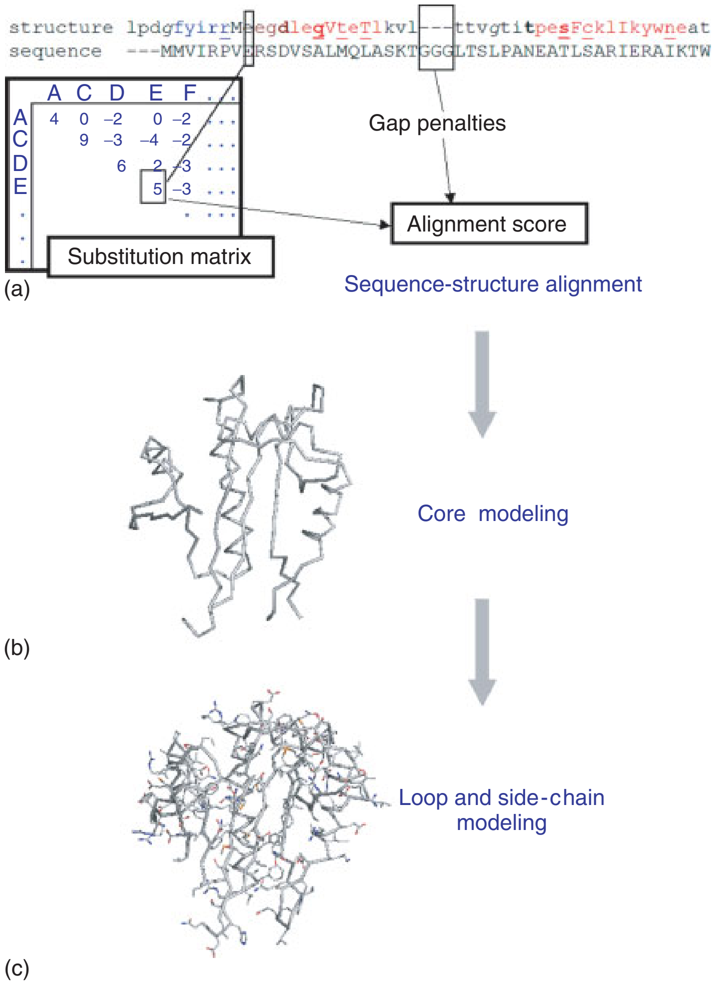


Figure 2 From sequence-structure alignment to full atomic model. (a) Alignment of the sequences of the target protein (in black) and the template (of known 3D structure, where secondary structure elements are shown in color; red, α-helices; blue, β-strands, maroon, 3_{10} helices; coil, black). (b) Cα trace of the core model. (c) Full atomic model with loop and side-chain atoms

and fold recognition has become obscured and the concept of comparative modeling can be now extended to cover any model building attempt based on homology (Kinch *et al*., 2003; Mizuguchi, 2004). Therefore, the following discussion will treat comparative modeling and fold recognition within a common framework, with particular emphasis on the latter.

Traditionally, comparative modeling consists of several distinct steps (Figure 2) (Blundell *et al*., 1987; Tramontano, 1998; Marti-Renom *et al*., 2000):

1. Template (parent) identification and sequence-structure alignment
2. Core building

3. Loop building
4. Side-chain building.

Although some methods, including the popular program MODELLER (Sali and Blundell, 1993; Fiser *et al.*, 2000), rely on finding solutions that optimally satisfy distance restraints and do not take the exact steps above, they can be still characterized by this generic scheme. In the following sections, I will first describe fold recognition techniques as a key component of the initial stage of comparative modeling. Since serious challenges remain in the model-building stages (steps 2–4 above) of conventional comparative modeling methods (Tramontano and Morea, 2003), these will be discussed in the subsequent section. Finally, a brief comment will be made on the accuracy of models.

2. Template identification

In recent years, there has been a change in the way the question of identifying homologs of known structure has been addressed. Conventional sequence database searching methods ask the question, "Is the query sequence A homologous to sequence B?", whereas more modern methods ask, "Does sequence A belong to family B?". The former (such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990)) perform pairwise searches, where the query sequence is compared with each sequence in the database and their similarity quantified. Normally, an amino acid substitution matrix is used to quantify the similarity between the two sequences (Figure 2, top *see* Article 67, **Score functions for structure prediction**, Volume 7). This is necessary because simply counting the number of complete matches (i.e., using PIDs) is known to be inadequate to judge the likelihood of common ancestry for distantly related proteins (Brenner *et al.*, 1998). The use of a 20-by-20 score matrix is based on the observation that during the course of evolution, substitutions between certain amino acid pairs were more frequent than others because they preserved the chemical and structural environments (and thus, these mutations are more likely to be accepted by natural selection) (Dayhoff *et al.*, 1978). In addition, insertions and deletions are dealt with by using a set of parameters called *gap penalties* (Figure 2, top).

Newer methods go further than this and use the information about a group of sequences that are already known to be homologous. In addition to the general tendency to preserve the chemical nature of the amino acid side chain, the degree of amino acid conservation within a family of homologous proteins depends on the structure and function of these proteins. For example, the surface positions are more likely to be variable than the buried positions and catalytic residues are highly conserved. These differences give rise to highly conserved and less conserved regions in a multiple sequence alignment of the family members. Thus, it is possible to derive a family-specific "profile" of patterns of amino acid substitutions and use it to evaluate the likelihood of membership for new sequences.

There are several methods for constructing these profiles. A sequence profile (Gribskov *et al*., 1987) is a natural extension of amino acid substitution matrices; it provides a set of substitution scores for each position in the multiple sequence alignment, from which the profile is made. It adopts the form of a matrix in which the number of rows is equal to the length of the original alignment and the number of columns is 20 to store the substitution scores to each of the 20 possible amino acids. Since a profile contains position-specific information about the sequence conservation within a homologous family, it is also called a *position-specific scoring matrix* (PSSM).

The patterns of amino acid substitutions can be also expressed using mathematical representations known as hidden Markov models (HMMs) *see* Article 98, **Hidden Markov models and neural networks**, Volume 8 (Durbin *et al*., 1998). From a machine-learning point of view, both profiles and profile HMMs are considered a generative probability model; the parameters of a statistical model representing a homologous family are derived from the sequences of known members of the family (positive training examples) and the model assigns a probability to any new sequence to assess the likelihood of its membership to the family of interest. There are also discriminative approaches (Jaakkola *et al*., 2000), in which the parameters are estimated using both positive training examples and negative training examples (proteins that are known to be nonmembers of the family being analyzed) and machine-learning techniques such as support vector machines (*see* Article 110, **Support vector machine software**, Volume 8) are applied (Liao and Noble, 2003).

The ability of these methods to identify distant homologs can be tested by using a set of proteins, the structures of which are already known. These tests have shown (for example, by Lindahl and Elofsson, 2000) that the family-based methods (using information derived from multiple sequence alignments) perform significantly better than conventional pairwise comparison methods.

Some of these methods offer a user-friendly Web interface and have been widely used in genome annotation. For example, the HMM-based method HMMER (Eddy, 1998) is the default search engine for the Pfam (Bateman *et al*., 2002) and Smart (Schultz *et al*., 1998) databases. PSI-BLAST (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7) (Altschul *et al*., 1997) is another popular tool for detecting remote homologs.

Most of the methods discussed so far perform general sequence–sequence comparison and are not restricted to searches for proteins of known 3D structure. Recent benchmark results show that these methods still miss a significant fraction of homologs, which adopt similar structures to that of the query (Fischer *et al*., 2003). Can we improve homolog detection further, particularly in our specific situation where we need to identify proteins of known structure? The answer seems yes and there appear to be two possible directions; (1) "fold recognition methods" to utilize specific information about protein 3D structure and (2) profile–profile methods to extend the family-based recognition approach. In discussing these methods, we need to diverge from pure sequence–sequence comparison and deal with specific issues regarding sequence-structure comparison. Therefore, it is more appropriate to discuss details of these methods in the next section.

3. Sequence-structure alignment

Once one or more homologs of known structure have been identified, it is necessary to align the query sequence against that of a template. Although the template identification (discussed in the previous section) and the alignment generation are intimately related and they are normally performed simultaneously, there are some issues unique to alignment. Theoretically, the optimal choices for some parameters (such as gap penalties) may be different for database searching and for alignment (Vingron and Waterman, 1994). In the first case, the objective is the best contrast between sequences related to the query and all others, whereas in the latter case, the alignment should cover the query sequence as much as possible.

For comparative modeling, this issue is particularly important. Standard sequence–sequence comparison methods normally adopt a local alignment algorithm. These methods are good at capturing the most conserved regions shared between related proteins but these regions may not cover the entire structural domains and are indeed often smaller than the known structures.

More generally, if our aim is to generate sequence-structure alignments, it is possible to extract specific information about conserved features from known structural families and incorporate that into our search and alignment strategies. This approach is adopted by fold recognition methods (the first direction mentioned at the end of the previous section).

As the number of experimentally defined protein structures continues to grow, protein family and classification databases have played important roles in organizing essential structural information. These databases can be manually curated (such as SCOP (Murzin *et al*., 1995)) or constructed using semiautomated (such as CATH (Orengo *et al*., 1997) and HOMSTRAD (Mizuguchi *et al*., 1998b)) and fully automated methods (FSSP (Holm and Sander, 1994)). Family-specific conserved features in sequence and structure can be obtained from these databases and such information is used, directly or indirectly, in many fold recognition methods.

Structural information can be used in several different ways in homolog detection and alignment (Mizuguchi, 2004). There are “threading-type” approaches, in which the query sequence (a “string”) is threaded through the backbone coordinates of a known protein structure and the degree of compatibility between the sequence and the structure is evaluated using knowledge-based potentials (Sippl, 1995). The programs GenThreader (Jones, 1999), PROSPECT (Xu and Xu, 2000), RAPTOR (Xu *et al*., 2003) are a few examples (*see* Article 72, **Threading for protein-fold recognition**, Volume 7). Structural information has also been incorporated directly into an alignment algorithm, which produced some promising results (Zhang *et al*., 2003).

There are “structure-aided sequence comparison” methods, whose operations are analogous to that of family (profile)-based sequence–sequence comparison methods, but with information about the 3D structures of proteins incorporated. The program 3D-PSSM (Kelley *et al*., 2000) uses PSSMs that are derived from multiple sequence alignments as in PSI-BLAST and other profile-based methods. The multiple sequence alignments, however, are based on the 3D structures of the family members and thus, are more reliable (as structure is better conserved than sequence during the course of evolution). The method also uses predicted secondary

structures and some form of knowledge-based potential. In the program FUGUE (Shi *et al*., 2001), the similarity between the query sequence and a potential homolog is quantified using environment-specific substitution tables (ESSTs) (Overington *et al*., 1990; Luthy *et al*., 1991). As mentioned in the previous section, amino acid substitutions are structurally conservative and the local environment of each amino acid affects the likelihood that it will change into other amino acids during the course of evolution. Thus, it is possible to derive amino acid substitution scores depending on the local environment and compile them in ESSTs. If we know the structure of one of the proteins being compared (as in the case of sequence-structure alignment), we can use ESSTs, instead of a generic amino acid substitution matrix, to evaluate an alignment score between the sequence and the structure.

Another feature of the program FUGUE is that a profile is constructed for both a database structure (using the ESSTs, combined with a multiple sequence alignment of homologs) and a query sequence (from its homologous sequences). Thus, the program performs "profile–profile comparison", an example of the second extension (see the previous section), from "sequence-profile" comparison to "profile–profile" comparison. If the change from pairwise methods (comparing a sequence with a sequence) to family-based profile methods (comparing a sequence with a profile) has proven a significant improvement, it is natural to extend this and develop methods for comparing a profile with a profile. There are different ways of defining similarity between profiles (Rychlewski *et al*., 2000; Yona and Levitt, 2002) but when properly parameterized, these methods appear to outperform conventional sequence-profile methods. Combining predicted secondary structures help but profiles can still be constructed without referencing to any known 3D structures. Some profile–profile methods even achieve performances comparable to those of the best fold recognition methods despite using no explicit structure information (Ginalski *et al*., 2003b).

Recent benchmark results show that (Fischer *et al*., 2003) many of these newer methods perform better than PSI-BLAST, a standard high-performance sequence–sequence comparison program. This is in terms of both the sensitivity and specificity of homolog identification (how many homologs can a method detect without producing many false-positives?) and the quality of sequence-structure alignments. The latter can be evaluated by comparing the sequence-structure alignment with the structure–structure alignment (*see* Article 75, **Protein structure comparison**, Volume 7).

Many of these methods are available as fully automated Web servers. There are also special servers called meta servers (Bujnicki *et al*., 2001; Douguet and Labesse, 2001). They collect predictions from individual servers and produce consensus predictions, which are generally more reliable than the results from the individual servers. For example, 3D-jury (Ginalski *et al*., 2003a), Pcons (Lundstrom *et al*., 2001), and shotgun (Fischer, 2003) are popular and successful meta servers.

4. Model building

A conventional approach to obtaining the atomic coordinates of the target protein consists of three steps: (1) constructing a model for the core of the query

protein, (2) building the structurally variable regions, and (3) modeling side-chain conformations.

Despite many efforts, the currently available techniques cannot correct errors introduced in the earlier stage; if the choice of the template(s) is not optimal and/or the alignment between the target protein and the template(s) contains errors, the quality of the final model can never exceed that of a model based on a better alignment. This is because the current technology is essentially the "copy-from-the-template" approach; in many cases, even the best model is no closer to the experimental structure than the closest homologous structure (Tramontano and Morea, 2003).

There are a few new approaches that show the potential to produce models better than the closest homolog (Kolinski *et al*., 2001; Contreras-Moreira *et al*., 2003; Fischer, 2003) and in many applications, it is desirable to produce complete atomic models rather than alignments alone. Therefore, specific issues regarding each of the model building steps will be briefly described below. More detailed descriptions of the techniques used can be found in recent reviews such as (Marti-Renom *et al*., 2000).

4.1. Core building

In its simplest form, core building can be described as follows: given an alignment between the sequence of the protein of interest and that of a template, the coordinates of the template atoms where the conformations are conserved between the template and the target are copied. One obvious choice of the conserved core (also called structurally conserved regions – SCRs) is the backbone atoms of all the nongapped regions in the alignment. By definition, copying the conserved core will provide an accurate prediction of (part of) the structure of the protein of interest and if the PID between the template and the target is high, the nongapped regions will be indeed a good approximation of the conserved core.

As the PID decreases, however, more structural changes will be observed. One of the most important challenges here is to predict the changes in the packing of secondary structure elements between homologous structures. Amino acid replacements of hydrophobic core residues are usually accommodated by small shifts of secondary structure elements (Lesk and Chothia, 1980). Therefore, simply copying the template coordinates for the nongapped regions may not model these changes accurately. This problem is related to a more general issue of how to define the conserved core when multiple template structures are available. The difficulty of quantifying shifts in secondary structures in homologous proteins was clearly documented in the pioneering review of Blundell *et al*. (1987) and it remains a challenge.

4.2. Loop modeling

A core model lacks the atomic coordinates of regions corresponding to insertions and deletions, which occur mostly in exposed loops that connect secondary

structure elements. By definition, the most significant variations in protein structures are observed in these structurally variable regions (SVRs) and predicting their structures is among the most difficult parts of model building. (Note that SVRs and loops are not always the same; active sites in enzymes usually consist of loops but they are conserved among the members of the family. In this article, the commonly adopted term "loop modeling" is used for convenience.)

Several approaches have been adopted. Knowledge-based methods search a database of known structures for a backbone fragment that fits the anchoring region of a loop. This approach is particularly effective when a set of empirical rules are derived from database structures to understand and predict the sequence-structure relationships for certain types of loops (Sibanda and Thornton, 1985; Chothia *et al*., 1989). Ab initio approaches, on the other hand, rely on conformational sampling and selection using potential energy functions or other scoring methods. Numerous methods for searching and scoring have been proposed (see, for example, Fiser *et al*., 2000 and references therein). The latest methods (de Bakker *et al*., 2003; DePristo *et al*., 2003; Jacobson *et al*., 2004) can build loops up to eight residues with reasonable accuracy (<2.3 Å backbone RMSD) when the experimental core structures are supplied. It is still a challenge to build loops (even of moderate lengths) when the remainder of the structure contains errors.

4.3. Side-chain modeling

After a model for the backbone, side-chain conformations have to be predicted to complete the full atomic model. Apart from the "copy-from-the-template" approach, which can be applied throughout all model building steps, side-chain placements normally consist of two elements: a rotamer library and combinatorial optimization.

It has been observed that each side chain adopts a limited number of preferred conformations referred to as *rotamers*. A rotamer library is a collection of rotamers for each residue type, usually derived from statistical analysis of known structures (Ponder and Richards, 1987; Dunbrack and Karplus, 1993; Lovell *et al*., 2000).

Although rotamer libraries can reduce the possible number of side-chain conformations significantly, the optimal conformation of a particular side chain depends on its interactions with other side chains (a combinatorial problem). To obtain the best combination of all side-chain conformations, an efficient search algorithm is required. A dead-end elimination algorithm (Desmet *et al*., 1992) is adopted in the commonly used program SCWRL (Bower *et al*., 1997). Other approaches include a mean-field algorithm (Mendes *et al*., 2001) and Monte Carlo methods.

5. Accuracy of models

As described in the previous section, the accuracy of comparative models is primarily determined by the sequence similarity between the target and the template proteins. Quantitative relationships between PID and structural similarity, such as

the one derived by Chothia and Lesk (1986), can provide general ideas. As a rule of thumb, below are several cutoffs estimated from different types of analysis:

30% – Comparison of models and experimentally determined structures for the CASP targets has shown that models based on templates with >30–35% PID generally have reasonably low alignment and structural errors (Vitkup *et al*., 2001). This 30% cutoff is often quoted as the minimum PID to cover fold space in structural genomics projects.

40% – For pairs of sequences with >40% PID, sequence–sequence comparison methods can produce alignments that are almost identical to those based on the direct comparison of 3D structures (for example, Shi *et al*., 2001). This suggests that there are minimal alignment errors when templates with >40% PID exists and that the model accuracy will be accordingly high.

Systematic analyses of the functional diversity of protein families have also identified a 40% PID threshold, above which precise function, as defined by the first three digits of the Enzyme Commission number, is conserved (Todd *et al*., 2001). Therefore, models based on templates above this threshold can be used to discuss detailed features of protein function.

How about models based on distant homologs? Even though the overall accuracy of these models may be limited, building models can be still useful in several different ways.

First, building models and examining conserved structural features by itself can provide additional support for the homology detection and alignment. Tools such as Verify3D (Luthy *et al*., 1992) and ProsaII (Sippl, 1993) can be used for this purpose. These tools evaluate the compatibility between a sequence and a predicted 3D structure using measures (normally knowledge-based potentials), which are often not used by homolog detection and alignment methods. Therefore, high model scores by these evaluation tools provide an independent line of evidence for the distant relationship identified. These evaluation tools, combined with manual examination of models and alignments (Mizuguchi *et al*., 1998a), can also help identify better templates or correct alignment errors and this iterative approach is an efficient way to improve model building on the basis of remote homologs (Burke *et al*., 1999; Kosinski *et al*., 2003). Note that good stereochemistry, as assessed by programs such as PROCHECK (Laskowski *et al*., 1993) and WHATCHECK (Hooft *et al*., 1996), does not guarantee closeness to the experimental structure.

Second, while functional diversity can be commonly observed for distantly related proteins, they tend to retain common features and furthermore, certain specific features, such as the reaction chemistry of enzymes, can be highly conserved. Thus, it is possible, for example, to discuss detailed mechanism of action on the basis of a model for the region near the catalytic site, even though the model is based on a distantly related homolog (e.g., Shirai *et al*., 2001; Shirai and Mizuguchi, 2003). These analyses typically present hypotheses that can be tested by site-directed mutagenesis and other experiments.

6. Conclusions

Modeling by homology remains the most reliable method for predicting protein 3D structures. With recent advances in methods for detecting distant homologs

and the increasing number of experimentally determined structures, this approach has opened up a wider range of applications. The technique involves several distinct stages, of which the first part, the homolog identification and sequence-structure alignment, is crucial for the success of the following steps. Fundamental problems remain but advances in the techniques at each stage may eventually provide reasonably accurate models for essentially any protein.

Acknowledgments

I thank Simon Lovell for reading the manuscript critically, Mark DePristo and Rick Smith for comments on loop and side-chain modeling and for support from the Wellcome Trust.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.

Blundell TL, Sibanda BL, Sternberg MJ and Thornton JM (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**, 347–352.

Bower MJ, Cohen FE and Dunbrack RL Jr (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *Journal of Molecular Biology*, **267**, 1268–1282.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brenner SE, Chothia C and Hubbard TJ (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6073–6078.

Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC and Hill RL (1969) A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *Journal of Molecular Biology*, **42**, 65–86.

Bujnicki JM, Elofsson A, Fischer D and Rychlewski L (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.

Burke DF, Deane CM, Nagarajaram HA, Campillo N, Martin-Martinez M, Mendes J, Molina F, Perry J, Reddy BV, Soares CM, *et al*. (1999) An iterative structure-assisted approach to sequence alignment and comparative modeling. *Proteins*, **37**(3), 55–60.

Chothia C and Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5**, 823–826.

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, *et al*. (1989) Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.

Contreras-Moreira B, Fitzjohn PW and Bates PA (2003) In silico protein recombination: Enhancing template and sequence alignment selection for comparative protein modelling. *Journal of Molecular Biology*, **328**, 593–608.

Dayhoff MO, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Dayhoff MO (Ed.), National Biomedical Research Foundation: Washington, pp. 345–352.

de Bakker PI, DePristo MA, Burke DF and Blundell TL (2003) Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins*, **51**, 21–40.

DePristo MA, de Bakker PI, Lovell SC and Blundell TL (2003) Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins*, **51**, 41–55.

Desmet J, De Maeyer M, Hazes B and Lasters I (1992) The dead-end elimination theorem and its use in protein side chain positioning. *Nature*, **356**, 539–542.

Douguet D and Labesse G (2001) Easier threading through web-based comparisons and cross-validations. *Bioinformatics*, **17**, 752–753.

Dunbrack RL Jr and Karplus M (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *Journal of Molecular Biology*, **230**, 543–574.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fischer D (2003) 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.

Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR and Elofsson A (2003) CAFASP3: The third critical assessment of fully automated structure prediction methods. *Proteins*, **53**, 503–516.

Fiser A, Do RK and Sali A (2000) Modeling of loops in protein structures. *Protein Science*, **9**, 1753–1773.

Ginalski K, Elofsson A, Fischer D and Rychlewski L (2003a) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM and Rychlewski L (2003b) ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Research*, **31**, 3804–3807.

Gribskov M, McLachlan AD and Eisenberg D (1987) Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 4355–4358.

Holm L and Sander C (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, **22**, 3600–3609.

Hooft RWW, Vriend G, Sander C and Abola EE (1996) Errors in protein structures. *Nature*, **381**, 272.

Jaakkola T, Diekhans M and Haussler D (2000) A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, **7**, 95–114.

Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE and Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*, **55**, 351–367.

Jones DT (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, **287**, 797–815.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kelley LA, MacCallum RM and Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D- PSSM. *Journal of Molecular Biology*, **299**, 499–520.

Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H and Grishin NV (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53**(Suppl 6), 395–409.

Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P and Skolnick J (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins*, **44**, 133–149.

Kopp J and Schwede T (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Research*, **32**, Database issue, D230–D234.

Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM and Bujnicki JM (2003) A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, **53**(Suppl 6), 369–379.

Laskowski RA, MacArthur MW, Moss DS and Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291.

Lesk AM and Chothia C (1980) How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, **136**, 225–270.

Liao L and Noble WS (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Journal of Computational Biology*, **10**, 857–868.

Lindahl E and Elofsson A (2000) Identification of related proteins on family, superfamily and fold level. *Journal of Molecular Biology*, **295**, 613–625.

Lovell SC, Word JM, Richardson JS and Richardson DC (2000) The penultimate rotamer library. *Proteins*, **40**, 389–408.

Lundstrom J, Rychlewski L, Bujnicki J and Elofsson A (2001) Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science*, **10**, 2354–2362.

Luthy R, Bowie JU and Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Luthy R, McLachlan AD and Eisenberg D (1991) Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, **10**, 229–239.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F and Sali A (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, **29**, 291–325.

Mendes J, Nagarajaram HA, Soares CM, Blundell TL and Carrondo MA (2001) Incorporating knowledge-based biases into an energy-based side-chain modeling method: Application to comparative modeling of protein structure. *Biopolymers*, **59**, 72–86.

Mizuguchi K (2004) Fold recognition for drug discovery. *Drug Discovery Today: Targets*, **3**, 18–23.

Mizuguchi K, Deane CM, Blundell TL, Johnson MS and Overington JP (1998a) JOY: Protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

Mizuguchi K, Deane CM, Blundell TL and Overington JP (1998b) HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, **7**, 2469–2471.

Moult J, Fidelis K, Zemla A and Hubbard T (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**(Suppl 6), 334–339.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**, 536–540.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Overington J, Johnson MS, Sali A and Blundell TL (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings of the Royal Society of London. Series B. Biological sciences*, **241**, 132–145.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, *et al*. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, **32**, Database issue, D217–D222.

Ponder JW and Richards FM (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, **193**, 775–791.

Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, *et al*. (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**, 667.

Rychlewski L, Jaroszewski L, Li W and Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, **9**, 232–241.

Sali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815.

Schultz J, Milpetz F, Bork P and Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5857–5864.

Shi J, Blundell TL and Mizuguchi K (2001) FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, **310**, 243–257.

Shirai H, Blundell TL and Mizuguchi K (2001) A novel superfamily of enzymes that catalyze the modification of guanidino groups. *Trends in Biochemical Sciences*, **26**, 465–468.

Shirai H and Mizuguchi K (2003) Prediction of the structure and function of AstA and AstB, the first two enzymes of the arginine succinyltransferase pathway of arginine catabolism. *FEBS Letters*, **555**, 505–510.

Sibanda BL and Thornton JM (1985) Beta-hairpin families in globular proteins. *Nature*, **316**, 170–174.

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.

Sippl MJ (1995) Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, **5**, 229–235.

Sutcliffe MJ, Haneef I, Carney D and Blundell TL (1987a) Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering*, **1**, 377–384.

Sutcliffe MJ, Hayes FR and Blundell TL (1987b) Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Engineering*, **1**, 385–392.

Todd AE, Orengo CA and Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, **307**, 1113–1143.

Tramontano A (1998) Homology modeling with low sequence identity. *Methods*, **14**, 293–300.

Tramontano A and Morea V (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53**(Suppl 6), 352–368.

Vingron M and Waterman MS (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *Journal of Molecular Biology*, **235**, 1–12.

Vitkup D, Melamud E, Moult J and Sander C (2001) Completeness in structural genomics. *Nature Structural Biology*, **8**, 559–566.

Xu J, Li M, Lin G, Kim D and Xu Y (2003) Protein threading by linear programming. *Pacific Symposium on Biocomputing*, 264–275.

Xu Y and Xu D (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343–354.

Yona G and Levitt M (2002) Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology*, **315**, 1257–1275.

Zhang Z, Lindstam M, Unge J, Peterson C and Lu G (2003) Potential for dramatic improvement in sequence alignment against structures of remote homologous proteins by extracting structural information from multiple structure alignment. *Journal of Molecular Biology*, **332**, 127–142.

Short Specialist Review

The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org

Helen Berman
The State University of New Jersey, NJ, USA

Haruki Nakamura
Osaka University, Osaka, Japan

Kim Henrick
MSD-European Bioinformatics Institute, Hinxton, UK

1. Introduction

The Protein Data Bank (PDB) (Berman *et al*., 2000; Bernstein *et al*., 1977) is the repository for macromolecular structures determined by experimental methods, including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. The PDB, founded in 1971 with seven structures, is one of the oldest biological databases and in January 2004 contained over 23 500 entries. The archive's growth has been accompanied by increases in both data content and the structural complexity of individual entries. The PDB data archive is an important tool for all biological researchers and is free to all as a collection of FTP-accessible files. In November 2003, the Worldwide Protein Data Bank (wwPDB) (Berman *et al*., 2003) was established to ensure the continuation of the PDB as a single, uniform, and enduring archive, with an agreement between the Research Collaboratory for Structural Bioinformatics (RCSB, **http://www.rcsb.org/**) (The Research Collaboratory for Structural Bioinformatics (RCSB) consortium members who are responsible for PDB management are Rutgers, the State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST.), European Bioinformatics Institute (MSD, **http://www.ebi.ac.uk/msd**) (Boutselakis *et al*., 2003), and the Protein Data Bank Japan (PDBj, **http://www.pdbj.org/**) (Kinoshita and Nakamura, 2004).

2. PDB information content

In addition to the 3D coordinates, entries contain general information required for all depositions and information specific to the method of structure determination. The format of the PDB entries has evolved over the years. In the early days, the data were loosely structured with free text information describing how the experiment was done and what the results might mean. The current PDB entries are now well structured and computer readable. They contain the sequence and source of the macromolecules, chemistries of small molecule ligands, secondary structural descriptions, and classifications (*see* Article 69, **Complexity in biological structures and systems**, Volume 7) and qualitative descriptions of the structures (*see* Article 76, **Secondary structure prediction**, Volume 7). They also contain experimental information including crystallization conditions, crystal data, data collection and refinement statistics, and literature references. Approximately 50% of the entries contain primary experimental data, including structure factor tables (h,k,l; Fobs; sigma-Fobs) and NMR-specific data.

3. Data uniformity

A significant objective of the PDB is to make the archive as consistent and error-free as possible. Query across the complete PDB has been limited by missing, erroneous, and inconsistently reported experimental data, nomenclature, and functional annotation. The evolution of experimental methods, functional knowledge of proteins, and methods used to process these data over the years has introduced various inconsistencies into the PDB archive and has inspired different versions of the PDB format. In addition to concentrating on making current depositions to the PDB consistent, efforts are being made as part of the RCSB's Data Uniformity Project to enhance the consistency of existing entries (Bhat *et al*., 2001). The goal of these efforts is to provide a higher level of query capability through the higher curation of the database. The legacy data are checked for accuracy, uniformity, and completeness in both individual and global respects. Much progress has been achieved in this "cleanup" process by means of two complementary methods used to update and unify the archive. *File-by-file* uniformity processing brings each entry up to the current PDB format by adding records that were not present in some early entries, correcting unresolved problems, providing standard nomenclature, and evaluating data using current validation software. Additionally, key records within each PDB entry have been targeted for *archive-wide* uniformity processing, wherein data are examined and updated on certain global parameters such as synonyms, common names, and names used by other data centers to enrich the reliability of query results. The MSD group has applied a different strategy by making use of relational database technologies in a comprehensive cleaning procedure to ensure data uniformity across the whole archive. Here, each record type is matched automatically from text manipulation tools. PDBj has used an XML Schema to detect problems in the PDB archive, and efforts are in progress to combine all three sets of corrections into a new set of uniform PDB entries.

4. Deposition to the PDB

One of the reasons the PDB archive offers such a complete record of experimental structural biology is that funding agencies and scientific journals require the deposition of relevant data as a condition of financial support or publication. Two deposition tools are available. ADIT is used for data submitted to RCSB and PDBj. This tool is an mmCIF (macromolecular Crystallographic Information file) editor and contains translators between the PDB and mmCIF formats. AutoDEP is used by the MSD group.

5. Annotation and validation of the PDB

The task of formatting, annotating (*see* Article 29, ***In silico* approaches to functional analysis of proteins**, Volume 7), validating, and releasing PDB entries in a manner that maintains uniformity across the archive has been made possible by adopting well-defined software-accessible data dictionaries and controlled vocabularies (*see* Article 81, **Unified Medical Language System and associated vocabularies**, Volume 8) using mmCIF (Bourne *et al*., 1997).

The validation processes include a range of issues including authentication of source (*see* Article 36, **Large-scale, classification-driven, rule-based functional annotation of proteins**, Volume 7), authentication of versions, validation of correct methodology, conformity to standards, error checks for consistency, and outlier detection. Specific sequence checks are made to ensure that the records are consistent and are correlated with the sequence database records. The chemistry of the ligands is also carefully checked. Nomenclatures are reviewed and corrected. Geometric checks of the small and large molecules are made with special attention to stereochemistry, close contacts, and valence geometry. The database includes goodness of fit, and quality indicators for entries, chains, and residues, allowing the coordinate information to be studied in conjunction with reference data. For those crystallographic entries in which structure factors have been deposited, the map correlation indicators are included at the residue level. Accurate, consistent data along with clear indicators of quality are key aspects that make the PDB useful to a diverse community of researchers.

6. PDB FTP access

The ftp site is maintained at **ftp://ftp.rcsb.org/pub/pdb/**. It is updated weekly and contains all the coordinate and experimental data for NMR and X-ray determinations. In addition, all obsolete structures that have been replaced by updated versions are maintained. As part of the wwPDB agreement, RCSB has sole write access to this site, thus ensuring that there is a single uniform archive of PDB entries. The ftp archive is available from all three wwPDB partner sites.

7. WWW search and retrieval systems

The challenge of presenting the available information in an intuitive way to users from various backgrounds and expertise demands that the data are archived in a meaningful and flexible way that represents the hierarchy and constraints within the data. Technologies of relational database and XML-DB offer both the flexibility and the framework to achieve this goal. The wwPDB partners have applied different database technologies for the extremely complex processes of importing legacy data from the PDB, creating deposition systems with automated annotation procedures, developing molecular graphics viewers, achieving data uniformity, and integrating relevant information from other biological databases. The services are continually undergoing improvements, and users should consult the different sites for the latest search systems.

The RCSB PDB Website is available from the three member sites (Rutgers, UCSD/SDSC, and CARB/NIST) and is mirrored at the Cambridge Crystallographic Data Centre (CCDC, UK), the National University of Singapore (Singapore), Osaka University (Japan), the Universidade Federal de Minas Gerais (Brazil), and the Max Delbrück Center for Molecular Medicine (Germany).

8. Inter-database consistency

To maintain consistency between the structure and sequence databases, it is important to determine the correct sequence database cross-reference. The PDB partners maintain up-to-date cross linkages between the PDB and SwissProt (*see* Article 97, **Ensembl and UniProt (Swiss-Prot)**, Volume 8) (Bairoch and Apweiler, 2000), InterPro (Apweiler *et al*., 2001), SCOP (Murzin *et al*., 1995), CATH (*see* Article 75, **Protein structure comparison**, Volume 7) (Orengo *et al*., 1997), and PFAM (*see* Article 35, **Measuring evolutionary constraints as protein properties reflecting underlying mechanisms**, Volume 7) (Bateman *et al*., 2002). For new depositions, steps are taken to ensure that the SEQRES record in the PDB entry represents the correct amino acid sequence of the sample. Since many of the legacy PDB entries contain only the coordinates of the observed atom positions, it is difficult to obtain the complete sequence of the protein(s) studied.

9. Conclusion

The PDB is a strategic resource in modern biology. After recent successes in genome-sequencing projects, initiatives in structural genomics aim to fully understand the biological role of proteins by determining representative structures for protein families on a genomic scale. The PDB now faces the challenge of managing a massive amount of data being produced (Westbrook *et al*., 2003; Berman *et al*., 2000). Harnessing the data within the PDB and combining it with data from other resources will enable biology to continue to advance at breathtaking speed.

Acknowledgments

The RCSB PDB is supported by funds from the National Science Foundation, the office of science Department of Energy, and the National Institutes of Health. E-MSD gratefully acknowledges the support of the Wellcome Trust (GR062025 MA), the EU (TEMBLOR, NMRQUAL, and IIMS), CCP4, the BBSRC, the MRC, and EMBL. PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, *et al*. (2001) InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, **29**, 37–40.

Bairoch A and Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acid Research*, **28**, 45–48.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer EL (2002) The Pfam families database. *Nucleic Acids Research*, **30**, 276–280.

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H and Westbrook, J (2000) The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, **7**, 957–959.

Berman H, Henrick K and Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, **10**, 980.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T and Tasumi M (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**, 535–542.

Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Schneider VB, Schneider K, Rhanki N, Weissig H, Westbrook J, *et al*. (2001) The PDB data uniformity project. *Nucleic Acids Research*, **29**, 214–218.

Bourne P, Berman HM, Watenpaugh K, Westbrook JD and Fitzgerald PMD (1997) The macromolecular Crystallographic Information File (mmCIF). *Methods in Enzymology*, **277**, 571–590.

Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, *et al*. (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Research*, **31**, 458–462.

Kinoshita K and Nakamura H (2004) *e*F-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**, 536–540.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM (1997) CATH-a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Westbrook J, Feng Z, Chen L, Yang H and Berman HM (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Research*, **31**, 489–491.

Short Specialist Review

Threading for protein-fold recognition

Kuang Lin
National Institute for Medical Research, London, UK

It has been observed that proteins with similar sequences, often closely related homologous, tend to have similar three-dimensional structures and functions. If their sequences and structures can be accurately aligned, corresponding amino acid residues in the alignments will be found in similar positions of the structures with similar conformations. Thus, given a protein sequence with unknown structure (often called *the query or target sequence*), if we can find proteins with similar sequences and known structures (often called *the structure framework or template*), it is possible to predict the structure of the query with confidence using comparative modeling approaches (e.g., Sali and Blundell, 1993; Burke *et al*., 1999).

When no template can be detected via sequence similarity, protein fold–recognition methods are often employed for structure prediction. These are based on the observation that proteins of very different sequences, without recognizable sequence similarities, can have the same folds (Orengo *et al*., 2001). Even when a template is a remote homolog of the query sequence, it is still possible to use fold-recognition tools to generate a more accurate structure model.

Threading is a class of methods for protein fold–recognition. When the word was first coined by Jones *et al*. (1992), the original term was "optimal sequence threading", but soon shortened to "threading". The query sequence will be "threaded" onto backbones of structure templates. Contact potential and other scoring schemes are then applied to evaluate the proposed model.

In this process, it is assumed that the backbones of the structures are rigid and only the side chains of amino acids are different between the query and the template. However, given the low sequence similarities between queries and templates, even with the same fold, the backbone structures will probably have large differences from each other, and side chains of most residues will be different. Threading programs cannot therefore use contact potentials or pseudo-energy functions that require detailed structure descriptions or coordinates of all atoms. Often scoring schemes are applied just on pair interactions between residue centers or residue side chain/backbone centers. Instead of using physical energy functions of atom interaction, empirical energy scores can be derived on the basis of the observed associations of amino acids in known protein structures. Contacts that are more frequently observed in known structures will have higher scores (for reviews, Sippl, 1995; Jones and Thornton, 1996; Lazaridis and Karplus, 2000; Lin *et al*., 2002).

Ungapped threading is the simplest form of threading. Insertion and deletion of residues are disallowed in alignment. The starting point of matching determines the alignment and the quality of the model. The method is obviously the fastest. However, without the gaps the sensitivity of the method is not suitable for the recognition of distant fold relationships.

Standard dynamic programming algorithms for pairwise sequence alignment can be employed for matching structure templates and query sequences. From known structures, it can be estimated which amino acids are preferred in certain structure environments. Alternatively, which amino acids are often aligned to residues with certain structure environments can be observed from available structure alignments. Having these statistics, PSSM (Position Specific Scoring Matrix) profiles can then be derived from structure templates using the structure environments of their residues. Sequence alignment algorithms can be easily adapted for the aligning of profiles and sequences. New profile-based methods often integrate other features, including predicted secondary structure and residue exposure and sequence profile from PSI-Blast searching and multiple sequence alignment. These programs are fast. Their speed is comparable to sequence alignment programs. By including structural information, they achieved good performances, proved in various experiments (Bowie *et al*., 1991; Rice and Eisenberg, 1997; Rost, 1997; Jones, 1999; Kelley *et al*., 2000; Shi *et al*., 2001; Sippl *et al*., 2001).

However, the independent assumption of standard dynamic programming algorithms is incompatible with the contact potentials for threading. When aligning two sequences (or a profile with a query sequence) using dynamic programming algorithm, it is assumed that each aligned residue pair is independent of others. Contact potential scores are decided by the aligning of residues with structural contacts. Dynamic programming algorithms cannot be directly used with these scoring schemes.

Jones *et al*. (1992) used the double dynamic programming algorithm to find the optimal alignment. The algorithm was originally designed for protein-structure alignment (Taylor and Orengo, 1989). A second level of dynamic programming algorithm was introduced for the scoring of nonlocal spatial interactions between residues. The algorithm includes more aligned residue pairs through iterations, till the completion of the alignment. Similar heuristic or exhaustive approaches had been developed later using techniques such as recursive dynamic programming (Thiele *et al*., 1999), Gibbs sampling algorithm (Bryant, 1996), and other heuristic algorithms (e.g., Taylor, 1997; Huber *et al*., 1999; Xu and Xu, 2000). These programs, which fit sequences directly onto templates using contact potentials, are sometime considered the “real” threading methods, compared to the profile-based 3D-1D matching programs, which are essentially augmented sequence alignment methods. The “real” threading programs are much slower and require more computing resources, which make them unfit for genome scale investigations. It is worth noting that few of them use contact potentials alone: secondary structure matching scores, sequence profile, and solvent potentials are often combined into the scoring schemes for better evaluation of models.

The CASP (Critical Assessment of Techniques for Protein Structure Prediction) exercise is a widely recognized examination of protein structure–prediction methods (Moult *et al*., 2001). Participants predict protein structures using sequences

the CASP organizers collected from structure solvers before the native structures were published. More objective assessment can be made since the coordinates of structures are not available in any public databases in the process of prediction. At the first CASP, the original THREADER program by Jones and coworkers correctly recognized 8 out of 11 queries, and was marked as one of the best fold-recognition tools. Later, with advanced techniques and larger databases of structure templates, profile-based programs have gained more success, especially when templates and query sequences share some revolutionary relationship (Sippl *et al*., 2001).

Because of the difficult task of detecting and aligning queries to remotely related templates, alignments produced via threading methods are often of poor quality. The assumption of rigid backbones also limits accuracy. Manual inspection and modification using comparative modeling approaches is often needed to generate a good structure model using threading alignments. However, when there is no obvious evolutionary relation between the query sequence and templates, threading is one of the very few methods that can predict the fold.

Threading is a template-based protein fold–recognition method. With the increasing number of known folds, improved contact potentials will become available, and template libraries will have better coverage. As the CASP exercise continues, it will be interesting to follow the progress of threading techniques and their use in genome annotation.

References

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known 3-dimensional structure. *Science*, **253**(5016), 164–170.

Bryant SH (1996) Evaluation of threading specificity and accuracy. *Proteins*, **26**(2), 172–185.

Burke DF, Deane CM, Nagarajaram HA, Campillo N, Martin-Martinez M, Mendes J, Molina F, Perry J, Reddy BV, Soares CM, *et al*. (1999) An iterative structure-assisted approach to sequence alignment and. *Proteins*, **37**(Suppl 3), 55–60.

Huber T, Russell AJ, Ayers D and Torda AE (1999) Sausage: protein threading with flexible force fields. *Bioinformatics*, **15**(12), 1064–1065.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature*, **358**(6381), 86–89.

Jones DT and Thornton JM (1996) Potential energy functions for threading. *Current Opinion in Structural Biology*, **6**(2), 210–216.

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, **287**(4), 797–815.

Kelley LA, MacCallum RM and Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*, **299**(2), 499–520.

Lazaridis T and Karplus M (2000) Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, **10**(2), 139–145.

Lin K, May AC and Taylor WR (2002) Threading using neural nEtwork (TUNE): the measure of protein. *Bioinformatics*, **18**(10), 1350–1357.

Moult J, Fidelis K, Zemla A and Hubbard T (2001) Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins*, **45**(Suppl 5), 2–7.

Orengo CA, Sillitoe I, Reeves G and Pearl FM (2001) Review: what can structural classifications reveal about protein evolution? *Journal of Structural Biology*, **134**(2–3), 145–165.

Rice DW and Eisenberg D (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology*, **267**(4), 1026–1038.

Rost B, Schneider R and Sander C (1997) Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, **270**(3), 471–480.

Sali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**(3), 779–815.

Shi J, Blundell TL and Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, **310**(1), 243–257.

Sippl MJ (1995) Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, **5**(2), 229–235.

Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A and Wiederstein M (2001) Assessment of the CASP4 fold recognition category. *Proteins*, **45**(Suppl 5), 55–67.

Taylor WR (1997) Multiple sequence threading: an analysis of alignment quality and stability. *Journal of Molecular Biology*, **269**(5), 902–943.

Taylor WR and Orengo CA (1989) Protein structure alignment. *Journal of Molecular Biology*, **208**(1), 1–22.

Thiele R, Zimmer R and Lengauer T (1999) Protein threading by recursive dynamic programming. *Journal of Molecular Biology*, **290**(3), 757–779.

Thornton JM, Orengo CA, Todd AE and Pearl FM (1999) Protein folds, functions and evolution. *Journal of Molecular Biology*, **293**(2), 333–342.

Xu Y and Xu D (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**(3), 343–354.

CASP

Krzysztof Fidelis
Lawrence Livermore National Laboratory, Livermore, CA, USA

1. Main goals

Elucidation of protein structure plays a major role in transforming the biological sciences from the descriptive analysis of organisms to the understanding of many processes at the molecular level. Nevertheless, even with rapid advances in crystallography and NMR spectroscopy, our knowledge of protein structures lags far behind our knowledge of sequences. Therefore, for many identified proteins, no structural data are available. The primary goal of CASP is to address the question of how effectively computational methods can help narrow this information gap. By providing critical assessment, CASP aims at identifying which methods work and which do not, where the barriers to progress remain, and how to best formulate the goals for future methods development.

2. Rigorous testing of prediction methods

Historically, bona fide testing of prediction methods has been difficult to implement because of the small number of unique structures being solved. With the considerably increased number of new structures, it was possible to design a prediction experiment in which the answers are not initially known but become available in a reasonable time. CASP follows the principles of blind prediction and independent evaluation of the results. Initially, essentially only the sequence information on structures about to be solved is released to predictors. These data are made available through the CASP website or sent directly to registered servers. Next, the participating research groups deposit models of target structures prior to the public release of experimental results. Groups operating prediction servers comply with tighter 48-h deadlines. Finally, the models are compared with structures derived experimentally, using numerical evaluation techniques and human assessment, and a meeting is held to discuss the significance of the results. This approach offers several advantages over assessments made previously. In addition to a truly predictive mechanism, CASP allows testing multiple prediction methods simultaneously and side-by-side comparisons of the results obtained on the same set of prediction targets. The selection of targets is quite arbitrary, removing any possibility of bias that could have been introduced if targets were selected by predictors themselves.

And finally, the assessments are made with a uniform set of evaluation methods, and verified by independent assessors.

3. Community involvement

CASP invites independent assessors to evaluate the quality of the received predictions and to assess the state of the art. Assessors are provided with numerical analysis data generated using approved methods but may also use their own evaluation techniques. The experiment itself is shaped by all the predictors getting together at the biannual CASP conferences and by consultancy groups of veteran CASP participants. The issues involve CASP policy and any major changes or extensions of the CASP process. In addition, since CASP5 in 2002, the FORCASP website (**www.FORCASP.org**) provides a forum where members of the prediction community may discuss the various aspects of the CASP experiment.

4. CASP Prediction Center

Since 1996, the Prediction Center at Livermore handles all the data needs of CASP, including (1) information about the experiment and timings on all targets; (2) verification of target and prediction data; (3) evaluation of predictions; and (4) results analysis, presentation, and visualization. With the need for obtaining a comprehensive picture of prediction quality and progress between assessments, CASP provided strong impetus for the development of many new evaluation, analysis, and visualization techniques. The demand for assessment is growing rapidly. In the recently concluded CASP6, approximately 200 research groups from over 20 countries have participated submitting over 40 000 predictions.

5. Categories of prediction

The final result of a structure prediction largely depends on how much information from the already known structures can be used. At one extreme, models competitive with experiment can be produced for proteins with sequences very similar to those of known structures. At the other, models for proteins with no detectable sequence or structure relationship to one of the known structures are still at best very approximate. Historically, reflecting on how extensively models could be based on knowledge of other structures, targets have been divided into three broad categories: comparative modeling, fold recognition, and prediction of new folds. Comparative or homology modeling relies on the fact that for all pairs of natural proteins so far encountered, a clear sequence relationship implies similar structures. Thus, the structure of a homologous protein can to a large degree guide generating a model of a new one. Fold recognition takes advantage of the fact that protein structure is much more strongly conserved than sequence. Increasingly, new structures deposited in the Protein Data Bank turn out to have folds that have been seen before, even though there is no obvious sequence relationship

between the related structures. The goal is to identify these structural relationships in cases in which the sequence signal is either weak or does not exist. Over time, the boundary lines between these two categories have become blurred, and finer grained intracategory distinctions have begun to appear. In the recent CASP6 (2004), prediction targets were divided into two categories, those with detectable sequence homology to known structures, and those characterized by only structural similarity or constituting entirely new folds. In CASP, models are accepted as sets of 3D coordinates and alignments to known structures. Also accepted are predictions of residue–residue contacts, predictions of disordered regions, and most recently structure-based predictions of function.

6. Evaluation of predictions

Any approach to the evaluation of model structures faces several challenges. First, it must provide not only an overall picture of model quality but also information about specific areas, allowing to identify precisely where development is most needed. Second, it should handle well predictions submitted on targets ranging in difficulty and resulting in different levels of accuracy or providing different levels of detail. Finally, as evaluation methods are critical in comparing the performance of different prediction groups, such an approach must be agreed upon and respected by the community. Details of the evaluation techniques developed for CASP can be found in the special issues of Proteins (1995).

7. Progress over CASPs

What kind of progress, if any, can be seen over the six CASP experiments held so far (1994–2004)? To answer this question is not particularly easy as assessments have to be performed over separate sets of prediction targets. Nevertheless, some trends are quite clear (Venclovas *et al*., 2003). First of all, there has been a steady progress over all six CASPs, amounting to a gradual "pushing of the envelope" rather than a breakthrough. Second, progress is most visible for targets of medium difficulty, corresponding to the area of remote homology modeling through (homologous) fold recognition, that is, where sequence signal is weak but still recognizable. Progress in comparative modeling, particularly in the case of relatively high sequence identity to a known template, proved to be rather slow. This is particularly disappointing as more and more of structure space is explored experimentally, both by conventional structural biology and more aggressively by structural genomics, and progressively more modeling targets will fall into this category. Finally, progress in the most difficult area of structure prediction has been visible but the results are quite different for targets with identifiable structure similarity to a known template and for new folds. In the first case, a significant majority of targets can be modeled to a degree where at least the fold is correctly identified; in the second, a similar result is achieved for only very few small targets.

References

Special issues of PROTEINS, Structure, Function, and Genetics, Volume 23, Number 3, 1995; Supplement 1, 1997; Supplement 3, 1999; Supplement 5, 2001; and Volume 53, Number S6, 2003.

Venclovas C, Zemla A, Fidelis K and Moult J (2003) Assessment of progress over the CASP experiments. *PROTEINS: Structure, Function, and Genetics*, **53**(Suppl 6), 585–595; as well as Kryshtafovych A, Venclovas C, Fidelis K and Moult J *PROTEINS: Structure, Function, and Genetics*, in preparation.

Short Specialist Review

Molecular simulations in structure prediction

Franca Fraternali

National Institute for Medical Research, London, UK

Jens Kleinjung

Vrije Universiteit, Amsterdam, The Netherlands

1. Introduction

Molecular simulations allow us to study properties of many-particle systems and, in particular, to extract those properties that are not easily accessible to experimental techniques (Frenkel and Smit, 1996). If we were to use the physicist approach, we have to admit that the basic laws of nature have the unpleasant feature to be expressed in equations that cannot be solved exactly (analytically), except for few cases. Even the description of the motion of a three-body system by means of the simple laws of Newtonian mechanics is analytically intractable. Computer simulations, on the other hand, use a number of algorithms that allow us to calculate properties of many-particles systems (more than 100 000 atoms) at the required degree of accuracy. All molecular simulations to which we will refer are based on the assumption that classical mechanics can be used to describe the motions of atoms and molecules. Physics-based force fields at different levels of accuracy are generally used to describe the topology and geometry of molecules (van Gunsteren and Berendsen, 1990).

On the other hand, Bioinformatics tools allow us to analyze and rationalize the information content of biological systems and, from the acquired knowledge, to extrapolate and predict molecular properties that are not yet experimentally available. Therefore, in terms of logical strategies, molecular simulations use the inductive argument to infer knowledge, while Bioinformatics uses a deductive strategy; they can be seen as working synergically for the aim of filling the gaps of current experimental knowledge. Computational biology embraces both approaches, and experience reveals that both are indeed necessary to shed light on the complexity of biomolecules.

Since the genomic projects have started their proliferation of sequences, the so-called sequence gap (the difference between the number of known sequences and the known protein 3D structures) has become one of the most prominent challenges in biology. In particular, structural genomics initiatives address the

investigation of the native fold of sequenced proteins as a fundamental step to a complete understanding of their biological function. Experimental techniques such as NMR and X-ray crystallography are very effective in providing atomic resolution structures of a large number of proteins, but large-scale initiatives of high-throughput structure determination are still at a different pace compared to genomic sequencing projects.

Until now, all structure prediction methods still suffer in accuracy and are not yet reliable enough to compete even with available experimental low-resolution structure determination methods. If one could predict the topology at a low-resolution level (3–6 Å) and use molecular simulations to refine the atomic details of the structure, the sequence gap could be filled more rapidly and be useful in the assignment of function. The bottleneck of such a procedure is that long simulations times are required with conventional MD simulations in order to refine accurately misfolded structures (Fan and Mark, 2004).

In this article, we review some of the recent efforts in the fields of structure prediction and molecular simulations, with the common aim to efficiently contribute to large-scale structural genomics initiatives.

2. Structure prediction

Structure prediction projects for the modeling of large proteins (Amodeo *et al*., 2001; Fraternali and Pastore, 1999) and entire genomes have been performed (Fischer and Eisenberg, 1997; Sanchez and Sali, 1998), and genome annotation has been shown to benefit from the use of structural homology information (Mayor *et al*., 2004). The qualitative progress in structure prediction methods is regularly assessed in the CASP experiments (Venclovas *et al*., 2001).

The field is traditionally separated into three disciplines, depending upon the level of identity between the query sequence and sequences of template structures: (1) comparative modeling (>30% sequence identity), (2) fold recognition or threading (<30% sequence identity), and (3) ab initio folding (no template). Independent of this classification, benchmarks show that the quality of the predicted structure, given as root mean square deviation from the native structure, decreases exponentially with decreasing sequence identity, that is, from about 2 Å at 95% to >5 Å below 25% identity (Contreras-Moreira and Bates, 2002). Even at relatively high sequence identity of about 50–60%, predicted structures achieve on average only medium resolution (∼3 Å) (Vitkup *et al*., 2001).

The deviation of predicted structures from their native counterpart has chiefly two reasons: alignment errors and the template structure approximation. Matching a sequence onto a template structure is a computationally hard problem (Lathrop, 1994; Kolodny and Linial, 2004), which is generally solved using heuristic approximations that lead frequently to suboptimal solutions, in particular for difficult alignments of distant homologs. The template structure approximation is due to the simple fact that two homologous structures are similar but usually not identical. Taking one structure as the template for the other naturally causes deviations. Moreover, structure prediction programs use simplified potentials and heuristic methods to generate target structures at reasonable computational expense.

Therefore, subsequent refinement is imperative to evolve the modeled structure toward the native conformation.

3. Conformational sampling

The "sampling problem" is a well-known obstacle in molecular simulations of biomolecules, and it is particularly relevant for endeavors to detect the native fold. It is primarily caused by the enormous dimension of the conformational space, which is the collection of all accessible internal coordinate combinations. The backbone of an average size protein has about 10^{100} degrees of freedom, far beyond our systematic searching capabilities. Fortunately, biological evolution has selected for biological macromolecules that adopt a distinct fold, so that in many cases conformational sampling can be restricted to a subspace around a limited number of template folds.

However, the energy profile of a biological macromolecule with respect to internal motion is rugged, with barriers above the background energy kT between neighboring conformers. Thermal fluctuations are rarely high enough to drive the system over these barriers, and energy minimization calculations lead to local minima instead to the global minimum. Therefore, structure prediction methods use Monte Carlo protocols to assemble a large variety of low-resolution conformers, from which the best ones are chosen for further calculations (Simons *et al*., 1997). Approaches to overcome the sampling problem in refinement by molecular simulations are discussed in the last paragraph.

4. Identification of the native state

The central paradigm of structure prediction and refinement by molecular simulations is the identification of the native structure as the conformer at the global free energy minimum, or alternatively at the probability maximum. The underlying thermodynamic principle is the notion that a system (here a macromolecule in solution) in weak contact with a thermal bath evolves spontaneously to the state of minimal free energy (we use here the more common notation of the free enthalpy G). Thus, successful structure prediction and refinement tools need a target function that discriminates effectively between native and nonnative folds (Park *et al*., 1997; Mirny and Shakhnovich, 1998). The free enthalpy of a protein can be decomposed into the intramolecular van der Waals and electrostatic contributions and an additional solvent term:

$$\Delta G = \Delta G_{\text{vdW}} + \Delta G_{\text{ele}} + \Delta G_{\text{sol}} \quad (1)$$

However, in computer simulations the entropic component of the free enthalpy is often neglected, leading to a description in terms of intramolecular enthalpy:

$$\Delta G \simeq \Delta H_{\text{vdW}} + \Delta H_{\text{ele}} + \Delta G_{\text{sol}} \quad (2)$$

Intramolecular enthalpies are calculated as the sum of all pairwise (and sometimes triple-wise) atomic interaction energies. These energies are based on carefully

parametrized interaction functions, most importantly the Lennard–Jones potential for van der Waals interactions and the Coulomb potential for electrostatic interactions. The ensemble of parametrized interaction functions defines a “force field”. We distinguish between two types of force fields: physics-based force fields and knowledge-based force fields.

Physics-based force fields are derived from quantum-molecular or molecular dynamics simulations of small molecules. Prominent examples for physics-based force fields are AMBER (Wang *et al*., 2000), CHARMM (Brooks *et al*., 1983), and GROMOS (van Gunsteren *et al*., 1996). Specific energy parameters have been designed for RNA molecules, owing to their particular intramolecular interactions and secondary structure patterns (Freier *et al*., 1986; Jaeger *et al*., 1989).

Knowledge-based or statistical force fields are derived from probabilities of states of known structures (Bowie *et al*., 1991; Jones *et al*., 1992; Madej *et al*., 1995; Huber and Torda, 1999; Lu *et al*., 2003). The transformation function between the (observed) probability of a state (for example the distance between two residues i and j) and the associated interaction energy is the Boltzmann term:

$$p(X_{ij}(r)) = \mathrm{e}^{-\frac{E(X_{ij}(r))}{kT}} \tag{3}$$

where $X_{ij}(r)$ is the state X as a function of the distance r between i and j, $E(X_{ij}(r))$ is the interaction energy (potential) associated with $X_{ij}(r)$, k is the Boltzmann constant, and T is the temperature. Transformation of equation (3) yields the knowledge-based potential

$$E(X_{ij}(r)) = -kT \log p(X_{ij}(r)) \tag{4}$$

often referred to as “potential of mean force”, because the interaction energy represents a collection of states.

Physics-based force fields model the physical reality starting from first principles, which renders them suitable for simulating accurately a wide range of molecules. However, early force fields were based on relatively small training sets, so that knowledge-based force fields were superior in describing specific classes of molecules at considerably reduced computational cost. On the other hand, knowledge-based potentials are by definition biased toward the training set, yielding questionable results for molecules with features far outside the range of those in the training set. Over the last decades, continuous refinement has improved the performance of physics-based force fields to a level equaling knowledge-based potentials (Lazaridis and Karplus, 1999a).

As an alternative to the conversion of state probabilities to energies as exemplified by equations (3) and (4), evaluation of predicted structures can be performed entirely in probability space, in which case the target function is the total probability maximum instead of the free energy minimum. It is common practice in Bioinformatics not to use the raw observed probabilities of states, but to convert those to a scoring function by normalization and transformation to a logarithmic scale:

$$S(X_{ij}(r)) = \log \frac{p(X_{ij}(r))}{p(X_{ij}^{\mathrm{rand}}(r))} \tag{5}$$

Here, $p(X_{ij}^{\text{rand}}(r))$ is the probability of observing state $X_{ij}(r)$ in a random set of conformers, which normalizes the observation probability with the expectation probability. The score S is often referred to as "log odds" score or "relative entropy", and it has the property to be additive.

If the logarithm is taken as $\log_2$, the score S in equation (5) is equivalent with the amount of information in units of "bit" that is gained when state $X_{ij}(r)$ is observed.

5. Refinement by molecular simulations

Structure refinement has originally been developed for experimental structure determination, that is, in NMR or X-ray resolution of proteins and DNA/RNA molecules. Experimental data are converted into distance restraints and combined with a classical molecular force field. Shortly after the invention of restrained molecular dynamics for (experimental) de novo structure calculation of proteins (Brünger *et al*., 1986), the "distance-geometry" method was developed, in which the classical force field was replaced with a simplified interaction function to increase computational speed, and simulated annealing was performed to enhance conformational sampling (Nilges *et al*., 1988). These developments led to the structure refinement program XPLOR (Brünger, 1992). Another approach was implemented in the program DIANA, where conformational sampling starts from random conformations and converges to the final structure by target function minimization in angular space (Güntert and Wüthrich, 1991).

A different situation emerges with regard to the refinement of predicted structures. The accuracy of generated conformers relies entirely on the precision of the force field and the efficiency in sampling around the native state. The detailed physics-based force fields and the associated extensive interaction calculations in molecular dynamics provide a much finer resolution than structure prediction methods and thus should be superior in defining the native state. Reports about the success of molecular dynamics in structure refinement imply that simplified representations or limited sampling fail to improve the predicted structure (Schonbrun *et al*., 2002), but carefully defined simulations can yield significantly better conformers (Lee *et al*., 2001; Flohil *et al*., 2002; Fan and Mark, 2004).

However, long simulation times are prohibitive for large-scale structure prediction applications. An effective means to shorten simulation time without sacrificing the solvent contribution to the free energy is the usage of implicit solvation (Fraternali and van Gunsteren, 1996; Lazaridis and Karplus, 1999b Ferrara *et al*., 2002). Numerous protocols have been developed to improve the sampling efficiency of molecular simulations. The extension of molecular dynamics to four dimension allows the molecule to bypass 3D barriers (van Schaik *et al*., 1993), and "local elevation" disfavors already visited conformations in order to enhance exploration of conformational space (Huber *et al*., 1994). Leap dynamics is a combination of Monte Carlo and molecular dynamics methods, in which limited conformational changes are induced and the new conformer performs a local search (Kleinjung *et al*., 2000; Kleinjung *et al*., 2003). An example for a purely stochastic protocol is the optimal-bias Monte Carlo method, in which the global energy minimum is searched within the internal coordinate space (Abagyan and Totrov, 1999).

Systematic studies about the predictive power of these refinement methods in combination with large-scale structure prediction projects should be undertaken to provide high-resolution models for the majority of sequenced biomolecules.

References

Abagyan R and Totrov M (1999) *Ab initio* folding of peptides by the optimal-bias Monte Carlo minimization procedure. *Journal of Comparative Physica*, **151**, 402–421.

Amodeo P, Fraternali F, Lesk AM and Pastore A (2001) Modularity and homology: modelling of the type I modules and their interfaces. *Journal of Molecular Biology*, **311**, 283–296.

Bowie JU, Lüthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **12**, 164–170.

Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Comparative Chemistry*, **4**, 187–217.

Brünger AT (1992) *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*, Yale University Press: New Haven.

Brünger AT, Clore GM, Gronenborn AM and Karplus M (1986) Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proceedings of the National Academy of Sciences of the United States*, **83**, 3801–3805.

Contreras-Moreira B, Fitzjohn PW and Bates PA (2002) Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. *Applied Bioinformatics*, **1**, 177–190.

Fan H and Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science*, **13**, 211–220.

Ferrara P, Apostolakis J and Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins*, **46**, 24–33.

Fischer D and Eisenberg D (1997) Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proceedings of the National Academy of Sciences of the United States*, **94**, 11929–11934.

Flohil JA, Vriend G and Berendsen HJ (2002) Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins*, **48**, 593–604.

Fraternali F and Pastore A (1999) Modularity and homology: modelling of the type II module family from titin. *Journal of Molecular Biology*, **290**, 581–593.

Fraternali F and van Gunsteren WF (1996) An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *Journal of Molecular Biology*, **256**, 939–948.

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MM, Neilson T and Turner DH (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States*, **83**, 9373–9377.

Frenkel D and Smit B (1996) *Understanding Molecular Simulations*, Academic Press: London.

Güntert P and Wüthrich K (1991) Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *Journal of Biomolecular NMR*, **1**, 446–456.

Huber T and Torda AE (1999) Protein sequence threading, the alignment problem and a two step strategy. *Journal of Computational Chemistry*, **20**, 1455–1467.

Huber T, Torda AE and van Gunsteren WF (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-aided Molecular Design*, **8**, 695–708.

Jaeger JA, Turner DH and Zuker M (1989) Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences of the United States*, **86**, 7706–7710.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kleinjung J, Bayley PM and Fraternali F (2000) Leap-dynamics: efficient sampling of conformational space of proteins and peptides in solution. *FEBS Letters*, **470**, 257–262.

Kleinjung J, Fraternali F, Martin SR and Bayley PM (2003) Thermal unfolding simulations of apo-calmodulin using Leap-dynamics. *Proteins*, **50**, 648–656.

Kolodny R and Linial R (2004) Approximate protein structural alignment in polynomial time. *Proceedings of the National Academy of Sciences of the United States*, **101**, 12201–12206.

Lathrop RH (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein. Engineering*, **7**, 1059–1068.

Lazaridis T and Karplus M (1999a) Discrimination of the native from mis-folded protein models with an energy function including implicit solvation. *Journal of Molecular Biology*, **288**, 477–487.

Lazaridis T and Karplus M (1999b) Effective energy functions for proteins in solutions. *Proteins*, **35**, 133–152.

Lee MR, Tsai J, Baker D and Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology*, **313**, 417–430.

Lu H, Lu L and Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, **84**, 1895–1901.

Madej T, Gibrat JF and Bryant SH (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Mayor LR, Fleming KP, Muller A, Balding DJ and Sternberg MJ (2004) Clustering of protein domains in the human genome. *Journal of Molecular Biology*, **340**, 991–1004.

Mirny LA and Shakhnovich EI (1998) Protein structure prediction by threading. Why it works and why it does not. *Journal of Molecular Biology*, **283**, 507–526.

Nilges M, Clore GM and Gronenborn AM (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Letters*, **239**, 129–136.

Park BH, Huang ES and Levitt M (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *Journal of Molecular Biology*, **266**, 831–846.

Sanchez R and Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proceedings of the National Academy of Sciences of the United States*, **95**, 13597–13602.

Schonbrun J and Wedemeyer WJ (2002) Protein structure prediction in 2002. *Current Opinion in Structural Biology*, **12**, 348–354.

Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225.

van Gunsteren WF and Berendsen HJC (1990) Computer simulations of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie (International ed. in English)*, **29**, 992–1023.

van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott W and Tironi I (1996) *Biomolecular Simulations: The GROMOS96 Manual and User Guide*. BIOMOS b.v. Laboratory of Physical Chemistry: ETH Zentrum, CH-8092 Zürich vdf Hochschulverlag AG, Zürich, ISBN 3 7281 2422 2.

van Schaik RC, Berendsen HJ, Torda AE and van Gunsteren WF (1993) A structure refinement method based on molecular dynamics in four spatial dimensions. *Journal of Molecular Biology*, **234**, 751–762.

Venclovas C, Zemla A, Fidelis K and Moult J (2001) Comparison of performance in successive CASP experiments. *Proteins*, **45**(Suppl 5), 163–170.

Vitkup D, Melamud E, Moult J and Sander C (2001) Completeness in structural genomics. *Nature Structural Biology*, **8**, 559–566.

Wang J, Cieplak P and Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Comparative Chemistry*, **21**, 1049–1074.

Short Specialist Review

Protein structure comparison

Liisa Holm
University of Helsinki, Helsinki, Finland

1. Introduction

Improved methods of protein engineering, crystallography, and NMR spectroscopy have led to a surge of new protein structures deposited in the Protein Data Bank (PDB) (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7). At the end of 2004, the PDB contained over 28 000 protein structures, and the structural genomics initiative aims to provide a structure for each major protein family within a decade. This wealth of data needs to be organized and correlated using automated methods. Nearly all proteins have structural similarities to other proteins. General similarities arise from principles of physics and chemistry that limit the number of ways in which a polypeptide chain can fold into a compact globule. Evolutionary relationships result in surprising similarities. Because structure tends to diverge more conservatively than sequence during evolution, structure alignment is a more powerful method than pairwise sequence alignment for detecting homology and aligning the sequences of distantly related proteins. In favorable cases, comparing 3D (three-dimensional) structures may reveal biologically interesting similarities that are not detectable by comparing sequences and may help infer functional properties of hypothetical proteins.

2. Overview

Automatic methods enable exhaustive all-against-all structure comparisons. As a result, each structure in the PDB can be represented as a node in a graph in which similar structures are neighbors of each other and structurally unrelated proteins are not neighbors. Clustering the graph at different levels of granularity removes redundancy and aids navigation in protein space. At long range, the overall distribution of folds (*see* Article 69, **Complexity in biological structures and systems**, Volume 7) is dominated by secondary structure composition (for example, all-alpha or alternating alpha/beta). At intermediate range, clusters are related by shape similarity that does not necessarily reflect similarity of biological function (for example, globins and colicin A). At close range, clusters represent protein families related through

strong functional constraints (for example, hemoglobin and myoglobin). Evolutionary relationships can be recovered by searching for continuous neighborhoods, since evolution is a diffusive process in protein space (Maynard-Smith, 1970).

In order to identify natural groupings of any set of objects, one needs a measure of distance or similarity. Structure comparison programs derive a structural alignment, which maximizes similarity or minimizes distance (Brown *et al*., 1996). The alignment defines a one-to-one correspondence of amino acid residues (sequence positions) in two proteins. This is analogous to sequence alignment (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7) except that the notion of (dis)similarity is much more complex between 3D objects than between linear strings (Figure 1). For example, the conformation of a point mutant differs from the wild-type protein only locally and may only be a few tenths of an Angstrom. Much larger deviations are observed in pairs of homologous proteins; with increasing sequence dissimilarity, small shifts in the relative orientations of secondary structure elements accumulate and reach several Angstroms and tens of degrees. At the largest evolutionary distances, only the topology of the fold or folding motif is conserved; topology here means the relative location of helices and strands and the loop connections between these. Deviations can be even larger and qualitatively different when structural similarity is the result of convergent rather than divergent evolution. In particular, convergent evolution may result in similar 3D folds that differ in the topology of loop connections.

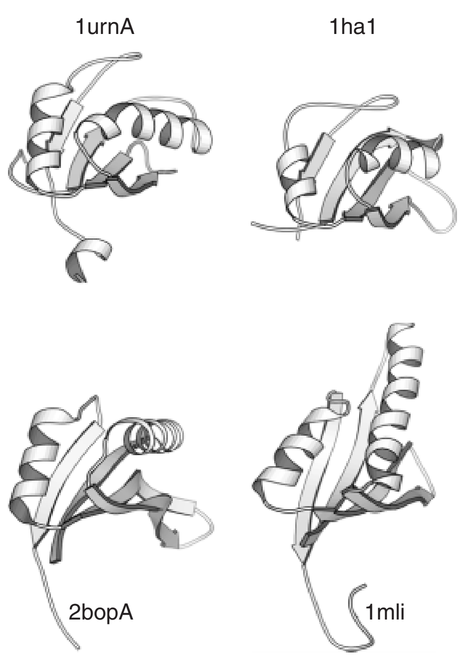


Figure 1 Structural neighbors of u1a spliceosomal protein (top left) from closest to more distant: hnRNP A1 ($Z = 10$, top right), DNA-binding domain of papillomavirus E2 ($Z = 5$, bottom left), muconolactone isomerase ($Z = 2$, bottom right). 1mli has the same topology even though there are shifts in the relative orientation of secondary structure elements

The modular architecture of proteins presents another complication. Large proteins can be decomposed into semiautonomous, globular folding units called domains. Domains are often evolutionarily mobile modules and may carry specific biological functions. Because a common domain may be surrounded by completely unrelated domains, most structure comparison methods search for local similarities.

3. Measures of structural similarity

Most programs for protein structure comparison quantify the quality of the alignment based on geometric properties of the set of points (C-alpha positions) or vectors representing the secondary structure elements. It is noteworthy that there is no consensus on the best measure of similarity. One criterion for a useful measure is that it should yield stronger similarity to presumably homologous than to analogous folds, in agreement with evolutionary classifications based on visual inspection of protein structures.

Similarity measures are usually defined in terms of comparing either inter- or intramolecular distances. The first approach, in essence, superimposes each protein so that corresponding points are as close together as possible. This method, therefore, needs to first uncover the rigid transformation that optimally positions both proteins relative to each other. An early suggestion of a structural distance metric was based on the surface area enclosed by the C-alpha traces of two superimposed structures. The most conventional measure is the root-mean-square deviation (RMSD) of the corresponding atoms:

$$RMSD = \sqrt{1/N\,\Sigma(|\mathbf{x}(i) - \mathbf{y}(i)|^2)} \tag{1}$$

where the summation is over all N corresponding atoms i and $\mathbf{x}$ and $\mathbf{y}$ are the coordinates of the atoms. RMSD is sensitive to outliers and does not have a unique optimum if the set of corresponding atoms can be made smaller or larger. Many variants of RMSD with a smoother penalty for ill-fitting atoms have been proposed. The most common formulation of an objective is that of accommodating the largest possible number of equivalent points within small deviations in position, typically less than 2 to 3 Å.

The structural core that can be superimposed as a rigid body at low RMSD shrinks with increasing evolutionary distance. For example, the (beta/alpha)$_8$-barrel is a frequent topology but the cross section of the barrel varies from circular to elliptical. This leads to high root-mean-square distance deviations in rigid-body 3D superimposition. To account for conformational flexibility, an edit distance like measure has been proposed that selects very well fitting rigid blocks with a penalty for twists around pivot points (hinges) introduced in the reference protein. Distance matrix comparison identifies both local similarities and tolerates plastic deformations between more distant parts of the molecules with the same topology. The distance matrix representation of protein structure is independent of coordinate frame, but contains more than enough information to reconstruct the 3D coordinates, except for overall chirality, by distance geometry methods. Some intramolecular-distance-based methods compare the respective distance matrices

of each structure, trying to match the corresponding intramolecular distances for selected aligned substructures. One definition of similarity has the form

$$\text{Dali-score} = \Sigma\Sigma\left(\frac{\alpha - \Delta}{\langle D\rangle}\right)\exp\left(\frac{\langle D\rangle^2}{\beta^2}\right) \tag{2}$$

where the summation is over all pairs of C-alpha atoms in the common substructure, Δ is the difference between the two intramolecular distances, $\langle D\rangle$ is the average of the two intramolecular distances and α and β are arbitrarily defined constants (Holm and Sander, 1996). A variety of other functional forms of a sum-of-pairs score are in use.

The statistical significance of an observed similarity is typically determined empirically against a background distribution of similarities observed between randomly selected pairs. Statistical models of structural evolution similar to the Dayhoff model for sequences have also been proposed.

4. Search algorithms

Given a measure of similarity or distance, the algorithmic problem is to find the set of corresponding points in two structures that optimize this target function. Just as there is much latitude in the formulation of the structure comparison problem, many different types of optimization algorithm have been employed. Similarity measures of the sum-of-pairs form and subgraph isomorphism formulations of the structure comparison problem belong to the NP-complete class of problems, and one has to resort to heuristics for practical algorithms. Heuristic approaches do not aim for provably correct solutions, gaining computational performance at the potential cost of accuracy or precision. Many programs use a hierarchical approach, where promising seeds for alignment are identified using local criteria based on dynamic programming, distance difference matrices, maximal common subgraph detection, fragment matching, geometric hashing, unit vector comparison, or local geometry matching. The initial set of correspondences is then optimized globally using methods such as double dynamic programming, Monte Carlo algorithms or simulated annealing, a genetic algorithm, or combinatorial searching. Recently, it has been proved that brute-force exhaustive scanning of the six degrees of freedom from rotations and translations in rigid-body superimposition leads to a polynomial-time approximation algorithm for the problem of determining the maximum number of C-alpha atom pairs that can be superimposed within a given RMSD at a given error. However, this solution is too computationally demanding for practical application.

5. Related problems

Biologists are interested in comparing a newly solved structure against all known structures in the PDB. For database searching, fast indexing filters have been developed. These methods compute a signature function from the geometrical properties,

with the idea that very similar structures will obtain similar values of the function. Highly similar structures can thus be identified quickly and efficiently. Database searching can benefit from a hierarchical strategy that uses initially fast filters to detect identity or strong similarity to known structures, and applies slower but more sensitive methods to the nontrivial cases.

Structural alignment algorithms detect similarities between globular folds that are composed of many secondary structure elements. Recurrent structural motifs are even more abundant on a smaller scale than domains. Complete protein structures can be approximately reproduced by concatenating pieces from relatively small fragment libraries. Functional sites may be defined by the geometry of a few essential side chains, and specialized tools have been developed for finding this type of motifs (*see* Article 35, **Measuring evolutionary constraints as protein properties reflecting underlying mechanisms**, Volume 7). Defining protein function through its interactions with other molecules has also generated interest in the comparison of molecular surfaces.

6. Applications

Many methods for pairwise protein structure alignment and database searching have been proposed and are available on the World Wide Web. The most commonly used server is Dali, hosted by the European Bioinformatics Institute (**http://www.ebi.ac.uk/dali**). Precomputed hierarchical classifications are likewise available. They organize folds in structural dendrograms (CATH and Dali database) or a taxonomic classification using a combination of structural and evolution information with a good deal of human expertise (SCOP). The integration of structural alignments and sequence comparisons is useful, for example, to identify conserved functional motifs. Structure comparison also provides benchmarks for assessing the accuracy and sensitivity of sequence comparison methods (*see* Article 73, **CASP**, Volume 7).

References

Brown NP, Orengo CA and Taylor WR (1996) A protein structure comparison methodology. *Computers & Chemistry*, **20**, 359–380.

Holm L and Sander C (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Maynard-Smith J (1970) Natural selection and the concept of protein space. *Nature*, **225**, 563–564.

Short Specialist Review

Secondary structure prediction

Ching Wai Tan and David T. Jones
University College London, London, UK

The prediction of protein secondary structure from amino acid sequence has been attempted since the late 1950s. It is a natural intermediate step to uncovering the tertiary structure of a sequence because discovering the local conformations of regions in the protein sequence precedes any informed guesses of the sequence's 3D fold. Secondary structure prediction is also significantly easier because it is essentially the prediction of 1D structural data from sequence (assignment of local state to each residue), as opposed to tertiary structure prediction, which has the goal of predicting 3D coordinates from sequence.

The secondary structure of a subsequence forms during the folding process because it is energetically favorable for that particular region of the sequence to adopt such a local conformation. Electrostatic charges belonging to C=0 and N−H groups in the backbone chain interact with each other to form hydrogen bonds. As a result of hydrogen bonds forming between the C=0 and N−H groups, two possible regular local conformations exist, namely, the helix and the strand (Pauling and Corey, 1951). A helix forms when the C=0 group hydrogen bonds with an N−H group 3, 4, or 5 positions away along the backbone chain, and a sheet forms when the two participating hydrogen bonding groups are from two distinct subsequences far apart in the sequence. Irregular conformations do form between residues, and these are loosely regarded as loop conformations, which also include residues that do not form any hydrogen bonds at all.

Intuitively, we can visually inspect and assign, say, the helix state to a set of consecutive residues. However, an objective assignment of a conformation to each and every residue in the sequence is necessary to avoid ambiguity when it comes to secondary structure prediction. The Define Secondary Structure of Proteins (DSSP) program (Kabsch and Sander, 1983) gives a systematic and unambiguous definition of these secondary structural elements in terms of the presence and location of hydrogen bonds between C=0 and N−H groups in the protein sequence. A hydrogen bond is assigned when the net electrostatic force between the C=0 and N−H group is below -0.5 kcal mol^{-1}. In assigning secondary structure, DSSP defines two elementary conformations, namely, the n-turn, $n = 3, 4, 5$ T, and the bridge (B), depending on the locations of the two interacting C=0 and N−H groups within the sequence. Helices are built from two or more consecutive n-turns, and these can be the α-helix where $n = 4$ (H), 3_{10} helix where $n = 3$ (G) and π helix where $n = 5$ (I). Bridges can be parallel or antiparallel, depending on the direction of the

two participating subsequences. Continuous stretches of bridges form β-strands (denoted as E in the DSSP alphabet) and one or more β-strands form β-sheets.

Other methods for objective assignment of secondary structural state, such as STRIDE (Frishman and Argos, 1995), have been developed. The main aim of STRIDE is to attempt to model the expert knowledge of crystallographers in terms of secondary structure assignment. It assumes that, barring obvious errors and the usage of DSSP, the authors of PDB files use their expertise to correctly assign secondary structural information to the protein whose structure they have just solved. STRIDE operates on the assumption that hydrogen bonds alone are insufficient criteria for assigning secondary structural states to residues. It defines a formula that incorporates torsion angle and hydrogen bond information, and the parameters of this formula are empirically fitted to match the secondary structure definitions found in the PDB database. One drawback of DSSP is that it tends to split a long helix into two owing to missing hydrogen bonds in the middle in spite of its completely acceptable geometry. STRIDE can overcome this because in addition to looking at hydrogen bonding patterns, it also takes main chain torsion angles into account. Despite these improvements, DSSP still remains the popular method when it comes to assigning secondary structural states.

While the DSSP definitions of the eight types of secondary structural state are unambiguous and used for exact assignment of state to secondary structure prediction datasets, most secondary structure prediction methods are content to deal with only the three most common types, namely, helices, strands, and loops or coil. An exception to this is Baldi's SSPro8 (Pollastri *et al*., 2002) method, which strives to predict the eight possible DSSP states. This leads to the issue of reducing the 8 DSSP states to 3 states, prior to the actual usage of the prediction method and the training, if required, of the prediction method. Of course, if DSSP is not used and the secondary structural states found in the PDB files are taken to represent the correct assignments, there is no need for any reduction method whatsoever. However, if DSSP is used as the standard means for assignment, then the reduction issue is relevant. One 8-to-3 state reduction method is to assign G and H states to the helix state, B and E to the sheet state, and all others to the loop state. Another is to assign H to the helix state, E to the sheet state, and all others to the loop state. Barton and coworkers (Cuff and Barton, 1999) have demonstrated the effect of different 8-to-3 reduction methods on the accuracy of some secondary structural prediction methods to be minimal. Nevertheless, it is important to note that when comparing different secondary structure prediction techniques, if DSSP is used as the "gold standard", it is essential to ensure that the same 8-to-3 reduction method has been applied for all methods being evaluated.

Having discussed the issues of defining secondary structural states, we now discuss the evaluation of the accuracies of secondary structure prediction. The most common accuracy measure is the Q3 score. The Q3 score is the average of each Qi (i = helix, sheet, loop), where Qi is defined as the percentage of correctly predicted residues in state i to the total number of experimentally observed residues in state i. The numbers of residues in helices, strands, and loops in the database (testing datasets) are frequently not evenly distributed, with loops usually comprising the greatest proportion, and because we are usually more interested in the correct

predictions of helices and strands, it is important to be aware that the Q3 score can give an overoptimistic estimate of accuracy than might be expected.

The best prediction methods currently have Q3 scores of somewhere between 75 and 80%. However, the Q3 score does not tell the whole story regarding the accuracy of secondary structure prediction methods. Because the Q3 score focuses on per-residue accuracy, it neglects to evaluate the overall picture of how predicted secondary structure elements are correctly positioned across the sequence. For example, if one were to predict every residue in helix-dominated myglobin to adopt a helical conformation, an accuracy of 80% would be obtained. Of course, such an uninformative prediction would hardly be very useful in understanding the overall topology of myglobin.

Rost and coworkers (Rost *et al*., 1994; Zemla *et al*., 1999) proposed another accuracy measure for secondary structure prediction, known as the Segment OVerlap measure (SOV). SOV_i, for each i (i = helix, strand, loop), measures the extent to which the predicted segment of state i is identical to the experimentally observed structure. The SOV3 score is taken as the average of all SOV_i scores. It is recommended by Rost that both the Q3 and SOV3 scores are used for secondary structure evaluation, and this has been the case in all CASP experiments since CASP2. The CASP (Critical Assessment of Methods of Protein Structure Prediction) experiments are a biannual event since 1994 where participants compare their methods in order to determine which method produces the best results for secondary and tertiary structure prediction of recent experimentally solved proteins. A key aspect of CASP is that the details of the experimentally solved structures are withheld from the predictors until the end of the experiment. CAFASP (Critical Assessment of Fully Automated Structure Prediction) is a similar but independent event, run since 1998, that focuses solely on automatic evaluation of automated servers.

Now that we have looked at how secondary structure is defined and how predictions can be effectively evaluated, we now look at some commonly used approaches for secondary structure prediction. All sorts of algorithms have been proposed for secondary structure prediction over the past 40 plus years, and we limit our brief discussion here to knowledge-based techniques. For historical reasons, it is important to note some early pioneering prediction techniques such as the GOR method (Garnier *et al*., 1978) that made use of information theoretical techniques and the Chou and Fasman method (Chou and Fasman, 1974) that used a qualitative method, in the form of simple rules, to try to predict secondary structure. These techniques, which do not have high accuracies, nevertheless, paved the way for more sophisticated knowledge-based approaches such as the machine-learning-based techniques widely used today. Diverse multivariate methods such as discriminant analysis (Solovyev and Salamov, 1994), nearest neighbors (Salamov and Solovyev, 1995), and linear discriminant functions (King and Sternberg, 1996) have been exploited, but by far, the most popular machine-learning technique used in secondary structure prediction is that of neural networks (Jones, 1999; Rost and Sander, 1993; Rost and Sander, 1994), though recently, support vector machines (SVMs) have been successfully used as well (Ward *et al*., 2003). Perhaps owing to their popularity, neural network–based methods have generally yielded the best results in the literature, and numerous network architectures have been tried. For the widely used PSIPRED (Jones, 1999) and PHD (Rost and Sander, 1993) methods, feedforward

neural networks have proved sufficient, while SSPro (Pollastri *et al*., 2002) uses a recurrent network architecture.

In contrast to the earliest methods, which used just the target sequence itself, modern secondary structure prediction methods make use of multiple sequence alignments based on the given target sequence (Rost *et al*., 1994). Typically, this alignment is produced using PSI-BLAST (Altschul *et al*., 1997), which is considered one of the better methods for generating divergent sequence profiles. The rationale behind the successful use of multiple sequence alignments is that the alignments of homologous families can reveal evolutionary patterns such as highly conserved or highly variable regions. For example, conserved regions in a family of proteins may correspond to functionally important residues that are unlikely to have undergone significant changes over time. Of more relevance to secondary structure prediction is the observation that the most conserved hydrophobic regions in a protein tend to correspond to the hydrophobic core of the protein. In contrast to the core residues, the surface residues have few structural constraints and will typically exhibit a high degree of mutation within the family of proteins. Such information can be used to increase the accuracy of predicting secondary structural state. Current state-of-the-art secondary structure prediction methods such as PSIPRED and PHD all use multiple sequence alignment information as their inputs. PSIPRED uses PSI-BLAST profiles, intermediate outputs of the PSI-BLAST tool, to represent the alignment information, while PHD does its own calculation and alignment of multiple sequence information.

The current state-of-the-art secondary structure prediction methods have Q3 scores ranging from around 74 to 80%, with the exact measured accuracies varying for different test sets used. It is hard to pinpoint a particular method as being a hands-down winner as these methods have several common features, and also use different datasets for training purposes, which makes it hard for a fair comparison to be made. However, an indication of the best performing secondary structure prediction technique can be found in CASP4 held in 2000. Here the PSIPRED method achieved an average Q3 score of 80.6% across all 40 submitted target domains with no obvious sequence similarity to structures already present in PDB.

While CASP results are indicative of the performances of various prediction methods, the CASP experiments are run only once every 2 years, and it would be desirable to have an automated service to compare secondary structure predictions on a regular basis. Fortunately, such a service exists in the form of EVA (Rost and Eyrich, 2001), which is an automated secondary structure assessment server that attempts to evaluate the performances of several secondary structure prediction servers. (EVA actually does more; it evaluates comparative modeling and contact prediction techniques as well.) This implies that EVA can only evaluate prediction techniques that are automated in the form of servers. EVA runs several secondary structure prediction servers on recently solved protein structures and compares and presents the prediction results on-line. It also uses the following 8-to-3 reduction technique: HGI to the helix state, EB to the sheet state, and the others to the loop state. Some of the participating prediction servers are PSIPRED, PHD, JPred (Cuff *et al*., 1998), and SSPro. Apart from JPred, which uses a consensus-based approach, these methods make use of knowledge-based approaches, in the form of neural networks, that learn from input evolutionary information. Such commonality

between these methods is probably the reason why there is not much difference between them in terms of Q3 and SOV3 accuracies, as evaluated by EVA.

The field of secondary structure prediction has reached a level of maturity where consistent Q3 and SOV3 accuracies of beyond 80 to 85% are arguably difficult to achieve. One possible reason is that the formation of secondary structure of a sequence segment is in part due to long-range interactions within the protein sequence and these are extremely difficult to take into account. To completely model these interactions, the full tertiary structure of the protein would need to be predicted, which of course creates a "chicken and egg" situation if secondary structure information is being used to model protein tertiary structure. The remaining accuracy gaps are therefore probably quite difficult to close using sequence information alone. In fact, the CASP community has informally reached a decision to drop the evaluation of secondary structure prediction techniques from the ongoing CASP6.

The challenge now regarding secondary structure is less of improving the accuracy of prediction techniques; rather it is more of how secondary structure prediction techniques can aid in constructing the 3D fold of a target protein sequence. For instance, given a particular prediction result, the question of how secondary structural elements can pack together in a compact manner remains unclear in protein structure prediction.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Chou PY and Fasman GD (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.

Cuff JA and Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–519.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M and Barton GJ (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.

Frishman D and Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Garnier J, Osguthorpe DJ and Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, **120**, 97–120.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, **292**, 195–202.

Kabsch W and Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

King RD and Sternberg MJE (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, **5**, 2298–2310.

Pauling L and Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*, **37**, 729–740.

Pollastri G, Przybylski D, Rost B and Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Rost B and Eyrich VA (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins* **45**(Suppl 5), 192–199.

Rost B and Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, **232**, 584–599.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Rost B, Sander C and Schneider R (1994) Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, **235**, 13–28.

Salamov AA and Solovyev VV (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*, **247**, 11–15.

Solovyev V and Salamov AA (1994) Predictin alpha-helix and beta-strand segments of globular proteins. *Computer Applications in the Biosciences*, **10**, 661–669.

Ward JJ, McGuffin LJ, Buxton BF and Jones DT (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650–1653.

Zemla A, Venclovas C, Fidelis K and Rost B (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

Short Specialist Review

DNA/protein modeling

Richard Lavery
Institut de Biologie Physico-Chimique, Paris, France

1. DNA structure and dynamics

How DNA interacts with proteins and is packed within the cell depends not only on its 1D sequence, but also on its 3D structure and on its mechanical properties (expressed through thermal fluctuations at room temperature). Experimental results have confirmed that both of these properties can vary along the length of the DNA molecule as a function of its base sequence. Computer modeling, using appropriate atomic interaction potentials, can help in predicting DNA structure and in simulating its thermal dynamics.

Atomic scale simulations of DNA have progressed greatly over the last decade, and it is now possible to simulate the behavior of oligomers containing 20–30 bp over nanosecond timescales (allowing for motions such as helix bending and twisting, sugar repuckering, etc.). Improved electrostatics, better parameterization, and faster computers have all played a part in this progress (Orozco *et al*., 2003). In the hope of predicting sequence effects on duplex DNA, extensive molecular dynamic studies are now being carried out on oligomers containing all 136 tetranucleotide sequences. These simulations will hopefully refine current models limited to di- or trinucleotide steps (Beveridge *et al*., 2004a). Simulations have already helped to understand the origins of intrinsic DNA curvature (Beveridge *et al*., 2004b) and how base sequence can change the mechanical properties of the double helix (Lankas, 2004). A number of specific DNA deformations (stretching, twisting, A-B transitions, etc.) have also been modeled, as well as local perturbations such as base-pair opening (Giudice and Lavery, 2003). Simulations are being used to design the modified oligonucleotides needed to block gene expression by binding to single-stranded RNA or double-stranded DNA targets and to study the structure and stability of quadruplex DNA, which has become an important target in cancer therapy (Read *et al*., 2001; *see also* Article 65, **Complexity of cancer as a genetic disease**, Volume 2).

2. Protein-DNA recognition

Identifying protein-binding sites is an increasingly important component of genome analysis and is essential for understanding gene regulation (*see* Article 19, **Promoter**

prediction, Volume 7 and Article 22, **Eukaryotic regulatory sequences**, Volume 7). Although experimental binding assays can yield consensus sequences (or weight matrices) that are sufficient in many cases, some proteins, and notably transcription factors, are difficult to characterize in this way because they bind to a broad range of sequences. Despite experimental progress, see, for example, (Roulet *et al*., 2002), isolating and identifying the DNA binding sites for a rapidly growing catalog of proteins will become increasingly difficult and theoretical approaches are needed.

In recent years, structural and thermodynamic data have both indicated that the formation of specific protein-DNA complexes cannot be understood in terms of a collection of pairwise interactions between amino acid side chains and nucleic acid bases. In many cases, such "direct" recognition is supplemented by "indirect" recognition, which relies on the sequence-dependent properties of the double helix. Indirect recognition remains difficult to characterize experimentally since it can involve structure, mechanics, and dynamics and, as discussed above, our understanding of these properties as a function of base sequence is still limited.

An analysis of complexes of known structure has helped in defining at least the structural component of indirect recognition by deriving the energy necessary to deform the target DNA during complex formation. Such analyses have shown that the total free energy of complexion (typically of the order of 40 kJ mol^{-1}) results from the compensation of many large terms, which, beyond the interacting macromolecules themselves, also involve the surrounding solvent counterions (Jayaram *et al*., 2002; Zakrzewska, 2003).

Despite the size of protein-DNA complexes, it has also been possible to carry out a limited number of molecular dynamics simulations (Zakrzewska and Lavery, 1999). Such simulations are currently too short to allow the complex formation process to be studied, but they can provide an ensemble of conformational "snapshots", which may then be analyzed using simple solvent models to estimate interaction-free energies. This method is applicable to studies of point mutations (within the binding protein or its DNA target), but is hindered when complex formation involves large structural rearrangements (Gohlke and Case, 2004). This problem is also a major hurdle for docking algorithms that are required to identify the interaction geometry if no experimental data is available or if mutations significantly perturb this geometry.

For complexes of known geometry, approaches related to the threading algorithms used in protein fold identification (*see* Article 72, **Threading for protein-fold recognition**, Volume 7 and Article 99, **Threading algorithms**, Volume 8) enable sequence-dependent binding and DNA deformation energies to be estimated very rapidly (Paillard and Lavery, 2004). Results for a range of DNA binding proteins confirm that indirect recognition often plays a major role and also show that the sequence selectivity of neighboring base pairs within the target sequence may be correlated, a problem ignored by conventional weight matrices (Benos *et al*., 2002). Such approaches are a promising route for extending data on known complexes to the study of other (closely) related proteins.

Whatever theoretical approaches are developed, it should not be forgotten that a major challenge will be to go beyond binary complexes and to understand the sequence-recognition processes involved in constructing multiprotein assemblies

such as the nucleosome or the transcriptosome (*see* Article 33, **Transcriptional promoters**, Volume 3).

3. Large DNA fragments

Despite the progress made in atomic scale simulations of DNA, applications to fragments containing more than 100 or so nucleotides are unlikely. Understanding the behavior of fragments many times this length therefore requires coarse-grained approaches (Orozco *et al*., 2003; Schlick, 2002). Elastic rod models are a natural answer to such problems. Such models exist at various levels of sophistication and may treat DNA as a uniform rod or may take into account anisotropy (e.g., easier bending towards the grooves of the double helix). The generic effects of base sequence (e.g., intrinsic or protein-induced curvature) have also been investigated, although, for the reasons discussed above, modeling the properties of arbitrary biological sequences is hindered by insufficient data. Recent results showing that very small DNA fragments (<100 bp) can cyclize with unusually high probabilities (Cloutier and Widom, 2004) also suggest that sharp, possibly sequence dependent, bending or "kinking" needs to be added to present rod models. Nevertheless, the generic studies have already shown that the local properties of DNA can have an impact on a mesoscopic scale, for example, by bringing together two distant sequence fragments as the result of an intrinsically bent segment placing itself at the highest radius of curvature within a supercoiled structure. It should also be noted that local curvature influences not only the pathway of the adjoining segments but also their rotational state and this can be important for the formation of contacts between distantly bound proteins.

Coarse grain models have many possible uses, but a biologically important application will be in probing the physical organization of genomes, whether this involves the relatively simple packing of naked DNA into a phage head (Purohit *et al*., 2003) or the organization of the nucleus within higher organisms (O'Brien *et al*., 2003). Note that large genomes can be represented by simplified bead models, akin to those used in polymer physics, where a single bead represents hundreds or thousands of base pairs. Ideally, progress in this field will allow smooth passages to be created from the most refined all-atom models to very coarse grain approaches with, in addition, the possibility of combining several levels of representation within a single study.

4. Future challenges

Beyond the standard wish list for specialists of molecular simulation (longer dynamic trajectories for larger molecular systems with more refined interaction potentials), DNA modeling should be able to help in translating a variety of experimental information into structural or dynamic data. This is already the case for the increasingly sophisticated refinement procedures used in crystallography or NMR spectroscopy, but other challenging areas include the calculation of chemical shifts, which would greatly extend the applicability of NMR data (Wishart and Case,

2001), and the simulation of circular dichroism or infrared spectra, which would allow a more quantitative use of these inexpensive experimental techniques. In this area, one can also cite the interpretation of the force curves resulting from mechanical experiments on single molecules or single molecular complexes (Lavery *et al*., 2002).

Quantum chemistry calculations of the electronic structure of DNA are still in their infancy, but are undoubtedly important for understanding interactions with radiation (UV, γ-rays), or chemical species such as drugs and mutagens. DNA conductivity (can DNA be used as an electric wire?) is another area where quantum chemistry needs to be coupled with molecular dynamics simulations to resolve experimental contradictions (Endres *et al*., 2004). The results of such calculations will be important for progress in the nanotechnological applications of DNA. Related studies of how electronic excitations move along DNA are also necessary for understanding DNA damage.

Lastly, more accurate and efficient methods of extracting thermodynamic data from molecular simulations would be important in analyzing not only DNA interactions with other molecules, but also in areas such as strand hybridization, in order to refine the interpretation of microarray data (*see* Article 57, **Low-level analysis of oligonucleotide expression arrays**, Volume 7) and to improve the selection of the DNA sequences already used in a wide variety of autoassembly processes (*see* Article 8, **Real-time DNA sequencing**, Volume 3). These goals imply an understanding of the properties of single-stranded DNA, which is currently as problematic as the treatment of unfolded polypeptide chains.

Related articles

Article 33, **Transcriptional promoters**, Volume 3; Article 19, **Promoter prediction**, Volume 7; Article 22, **Eukaryotic regulatory sequences**, Volume 7

References

Benos PV, Lapedes AS and Stormo GD (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays*, **24**, 466–475.

Beveridge DL, Case DA, Cheatham III TE, Dixit S, Giudice E, Lankas F, Lavery R, Maddocks JH, Osman R, Sklenar H, *et al*. (2004a) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design, informatics and results on CpG steps. *Biophysical Journal*, in press.

Beveridge DL, Dixit SB, Barreiro G and Thayer KM (2004b) Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. *Biopolymers*, **73**, 380–403.

Cloutier TE and Widom J (2004) Spontaneous sharp bending of double-stranded DNA. *Molecular Cell*, **14**, 355–362.

Endres RG, Cox DL and Singh RRP (2004) Colloquium: the quest for high-conductance DNA. *Reviews of Modern Physics*, **76**, 195–214.

Giudice E and Lavery R (2003) Nucleic acid base pair dynamics: the impact of sequence and structure using free-energy calculations. *Journal of the American Chemical Society*, **125**, 4998–4999.

Gohlke H and Case DA (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *Journal of Computational Chemistry*, **25**, 238–250.

Jayaram B, McConnell K, Dixit SB, Das A and Beveridge DL (2002) Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *Journal of Computational Chemistry*, **23**, 1–14.

Lankas F (2004) DNA sequence-dependent deformability–insights from computer simulations. *Biopolymers*, **73**, 327–339.

Lavery R, Lebrun A, Allemand J-F, Bensimon D and Croquette V (2002) Structure and mechanics of single biomolecules: experiment and simulation. *The Journal of physiology(Cond. Mat.)*, **14**, R383–R414.

O'Brien TP, Bult CJ, Cremer C, Grunze M, Knowles BB, Langowski J, McNally J, Pederson T, Politz JC, Pombo A, *et al*. (2003) Genome function and nuclear architecture: from gene expression to nanoscience. *Genome Research*, **13**, 1029–1041.

Orozco M, Perez A, Noy A and Luque FJ (2003) Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, **32**, 350–364.

Paillard G and Lavery R (2004) Analyzing protein-DNA recognition mechanisms. *Structure (Camb)*, **12**, 113–122.

Purohit PK, Kondev J and Phillips R (2003) Mechanics of DNA packaging in viruses. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3173–3178.

Read M, Harrison RJ, Romagnoli B, Tanious FA, Gowan SH, Reszka AP, Wilson WD, Kelland LR and Neidle S (2001) Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4844–4849.

Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N and Bucher P (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, **20**, 831–835.

Schlick T (2002) *Molecular Modeling and Simulation. An Interdisciplinary Guide*, Springer: New York.

Wishart DS and Case DA (2001) Use of chemical shifts in macromolecular structure determination. *Methods in Enzymology*, **338**, 3–34.

Zakrzewska K (2003) DNA deformation energetics and protein binding. *Biopolymers*, **70**, 414–423.

Zakrzewska K and Lavery R (1999) Modelling DNA-protein interations. *Computational Molecular Biology*, **8**, 441–483.

Short Specialist Review

Modeling tertiary structure of RNA

Benoît Masquida and Eric Westhof
Université Louis Pasteur, Strasbourg, France

1. The modeling process

The starting point for modeling an RNA molecule is the 2D or secondary structure that ideally represents, on a planar drawing, the Watson–Crick paired helical regions linked by non-Watson–Crick paired helical segments: hairpin loops, internal bulges, and multiple junctions (*see* Article 96, **RNA secondary structure prediction**, Volume 8 and Article 27, **Noncoding RNAs in mammals**, Volume 3). Secondary structure diagrams provide a framework-summarizing sequence and biochemical data without providing a full structural basis. The folded architecture results from the assembly of the helical framework through interactions between the helices and the single-stranded regions. The assembly often appears modular with a recognizable hierarchy and modules, recurring in various RNAs, and that assemble into a fold in a concerted fashion. There are three main categories of tertiary structure interactions: those between two double-stranded helices, those between a helix and a single strand, and those between two single-stranded regions. A modeling process based on this principle of natural folding processes has been developed (Westhof *et al*., 1996). Modules can be either architectural (forming a bend, e.g., kink-turn motif (Klein *et al*., 2001), or a reorientation within a helix or between helices) or anchors for association (e.g., tetraloop–tetraloop receptor interaction; Cate *et al*., 1996). Modules generally contain RNA motifs that we define as an ordered assembly of non-Watson–Crick base pairs (Leontis and Westhof, 2003).

Because of the underlying sequence variations within the motifs (the rules of which are still unknown), it is crucial to perform extensive sequence alignment and secondary structure analysis prior to any 3D (three-dimensional) modeling in order to identify with confidence the nature of the various motifs. The secondary structure can then be parsed into modules based on elementary motifs, the 3D coordinates of which can be generated later using appropriate programs (see below). Afterward, these modules can be assembled to form the RNA architecture. Finally, the geometry of the model is regularized using least-square refinement. In this process, hydrogen bonds between nucleotides are used as explicit constraints.

One of the most difficult tasks is the arrangement of the multiple-way junctions. Usually, a 2D structure is a set of hairpins linked together by various single-stranded

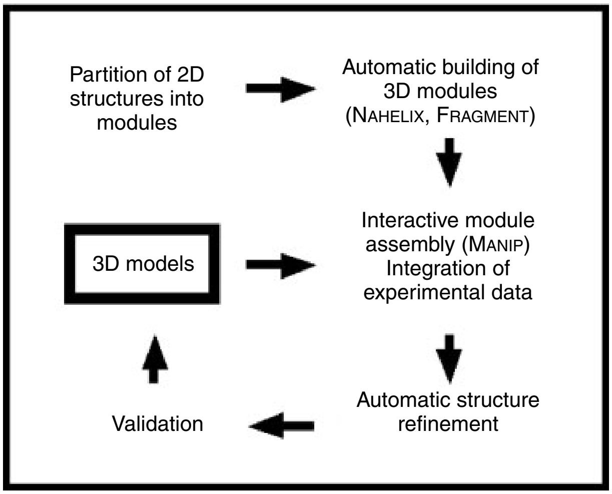


Figure 1 Diagram of the modeling process described (Massire and Westhof, 1998). From an accurate sequence alignment, the secondary (2D) structure can be deduced. Then, the first step consists of partitioning the 2D structure into modules for which 3D coordinates can be generated using adequate computer programs. These modules can then be assembled interactively on a graphic computer. When all the modules are placed next to each other while respecting the covalent linkages of the RNA chain, refinement can be performed. Validation of the model proceeds through collating it to biochemical data. Additional steps of interactive modeling refinement can be performed until the model is fully satisfactory

regions and junctions. The main problem is then the proper choice of helices that stack on each other. If an internal loop lies within a helical stem, does it favor coaxial stacking or does it make a kink that prevents coaxially stacking between the helical parts of the stem? In the case of 3-way or 4-way junctions, which helices do stack, if any? A thorough analysis of the sequence of the junctions should be performed to exhibit features homologous to junctions for which a crystal structure is known. If no convincing homology is found, assumptions should be made that eventually lead to the building of several models (Westhof *et al*., 1996). Then, tertiary interactions coupled to biochemical data may help to discriminate between the various possibilities.

The suite of in-house modeling programs runs on SGI graphical stations under the IRIX 6.2 system or beyond. Figure 1 summarizes the overall process. The MANIP (Massire and Westhof, 1998) package contains programs dedicated to generate 3D coordinates from the sequence information, and is available at the URL **http://www-ibmc.u-strasbg.fr/upr9002/westhof/**.

2. Getting the right secondary structure

In order to build a reasonable secondary structure for a given RNA, one needs to collect and align enough sequences to perform statistical analysis of the sequence variations. Chemical probing techniques can compensate for a lack of sequences

and help discriminate between various possibilities (Powers and Noller, 1991). With experience, secondary structure prediction programs such as MFOLD can be invaluable (Zuker, 2003; Ding and Lawrence, 2003). The latter approach is the only theoretical method that can be used when a single sequence is available. As a first step of the molecular modeling process, sequences are roughly aligned according to fully conserved nucleotide stretches, if present. Then, helices can be identified by checking for sequence covariations obeying Watson–Crick rules. Within this scenario, conserved Watson–Crick base pairs cannot be proven and, thus, should be regarded as nonproven or with extreme caution. Finally, the RNA parts that do not follow the Watson–Crick rules can be suspected to form specific tertiary motifs that should be analyzed for sequence homologies to motifs of known 3D structures.

3. Extrusion of the secondary structure in three dimensions

Two complementary programs help extruding secondary structure elements into three dimensions and have been described elsewhere (Westhof, 1993).

- NAHELIX is dedicated to building nucleic acid helices, and is run using either a graphical interface or interactive mode in a shell window. The program reads the sequence length, numbering, and sequence as input, and outputs a coordinate file for all the RNA atoms of the described sequence.
- FRAGMENT threads any sequence over a known 3D structure. A given structure file can be transformed to a fragment file that contains no sequence information, but only important conformational parameters for maintaining the sugar-phosphate backbone characteristic of the motif. Then, a polynucleotide backbone with any base sequence can be threaded onto that of the starting folded polynucleotide, which considerably facilitates the building of variants of a given motif.

4. Interactive molecular modeling

Once all secondary structure elements have been built in 3D, they can be assembled interactively using the MANIP program (Massire and Westhof, 1998). MANIP is a user-friendly program operated by pop-up menus in the main window. It allows definition of zones that can be independently moved using the `move` command in order to generate a 3D model. The other main modeling function is `tor`, which activates all the dihedral angles of a selected residue. A switch allows the user to quickly activate dihedral angles from previous or next residues all along the RNA chain. RNA fragments can thus be moved like tentacles and efficiently linked to each other.

5. Refinement of the model

Refinement of the model is achieved by geometrical least-squares using the Konnert–Hendrickson algorithm (Konnert and Hendrickson, 1980) implemented

in the program NUCLIN/NUCLSQ (Westhof *et al*., 1985). The algorithm takes into account bond lengths, valence angles, dihedrals, and has antibump restraints. The resulting function is minimized against a dictionary of values that have been observed in high-resolution crystal structures of nucleotides and oligonucleotides. Since the refinement program optimizes the geometry along the steepest descent, the conformation of the starting model should not contain extremely distorted regions to avoid refinement failure. The refined model can then be collated to the data, and the process of interactive modeling and least-square refinement can be iterated until the model is satisfactory. A subset of the data should be used as a blind test during the building of the model so as to help validating the model. Further steps of interactive modeling may then be required until a satisfactory solution is reached.

6. Conclusions

The availability of genome sequences (URL **http://www.tigr.org/**), comparative sequence analysis to elucidate RNA structure, both at the level of secondary and tertiary structures, has already undergone wide use (Lambert *et al*., 2004). Second, major improvement in modeling the tertiary structure of RNA results from the increasing number of available RNA crystal structures that provide a structural basis to understand sequence variations. Third, this surge in sequence and 3D data has led to the development of several computer tools for analysis and comparisons. Nowadays, sequences and crystal structures can, thus, be merged together in order to decipher the rules governing nucleotide interactions in complex RNA motifs, and, hence, provide new tools to unravel RNA structures (Gendron *et al*., 2001; Waugh *et al*., 2002; Yang *et al*., 2003). A powerful tool for building tertiary folds of RNA molecules has been developed by F. Major with the MC-Sym suite of programs (Major *et al*., 1991; Major and Griffey, 2001). In the end, modeling the tertiary structure of RNA molecules ab initio or by assembling homologous modules is still a process fraught with difficulties, especially in the absence of sufficient aligned sequences. Although a coarse-grained architecture can be derived for structured RNAs rather successfully, atomic details and contacts do not lead to reliable predictions for either RNA–ligand interactions or catalytic mechanisms.

References

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Szewczak AA, Kundrot CE, Cech TR and Doudna JA (1996) RNA tertiary structure mediation by adenosine platforms. *Science*, **273**, 1696–1699.

Ding Y and Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, **31**, 7280–7301.

Gendron P, Lemieux S and Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology*, **308**, 919–936.

Klein DJ, Schmeing TM, Moore PB and Steitz TA (2001) The kink-turn: a new RNA secondary structure motif. *The EMBO Journal*, **20**, 4214–4221.

Konnert JH and Hendrickson WA (1980) Restrained parameters thermal factors refinement procedures. *Acta Crystallographica*, **A36**, 344–349.

Lambert A, Fontaine JF, Legendre M, Leclerc F, Permal E, Major F, Putzer H, Delfour O, Michot B and Gautheret D (2004) The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Research*, **32**, W160–W165.

Leontis NB and Westhof E (2003) Analysis of RNA motifs. *Current Opinion in Structural Biology*, **13**, 300–308.

Major F and Griffey R (2001) Computational methods for RNA structure determination. *Current Opinion in Structural Biology*, **11**, 282–286.

Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E and Cedergren R (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, **253**, 1255–1260.

Massire C and Westhof E (1998) MANIP: an interactive tool for modeling RNA. *Journal of Molecular Graphics & Modeling*, **16**, 197–205, 255–197.

Powers T and Noller HF (1991) A functional pseudoknot in 16 S ribosomal RNA. *The EMBO Journal*, **10**, 2203–2214.

Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, *et al*. (2002) RNAML: a standard syntax for exchanging RNA information. *RNA*, **8**, 707–717.

Westhof E (1993) Modeling the three-dimensional structure of ribonucleic acids. *Journal of Molecular Structure and Dynamics*, **286**, 203–210.

Westhof E, Dumas P and Moras D (1985) Crystallographic refinement of yeast aspartic acid transfer RNA. *Journal of Molecular Biology*, **184**, 119–145.

Westhof E, Masquida B and Jaeger L (1996) RNA tectonics: towards RNA design. *Folding & Design*, **1**, R78–R88.

Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H and Westhof E (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, **31**, 3450–3460.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, 3406–3415.

Introductory Review

Introduction to ontologies in biomedicine: from powertools to assistants

Russ B. Altman
Stanford University Medical Centre, Stanford, CA, USA

Ontologies are a "specification of a conceptualization" (Gruber, 1993). They provide a structured way to discuss an area of knowledge by specifying a controlled set of terms with well-defined meanings. They also provide (and this distinguishes them from a simple glossary or dictionary) a formal set of constraints on the relations between terms, so that the ontology can only be used to make statements that are semantically valid. We define terminology, ontology, qualitative represesentations, and quantitative representations in more detail below, and explore these even more deeply within this section of the encyclopedia. The first key point about ontologies, however, is that they allow us to build computational representations of knowledge, to go along with the standard computational representations of data.

What is the difference between data and knowledge? This question touches many deep philosophical issues, but for the purposes of this discussion, we can define data as pieces of information, usually of basic "types" (strings, integers, real numbers, Booleans), and often organized in a table, such as a spreadsheet. Thus, for biology, data may include genome sequences, expression levels, protein levels, protein locations, and measurements of physical interactions between molecules. Knowledge is the principles or rules that are generated by an analysis of data that allows different "columns" of a spreadsheet to be related by rules or equations. There are many ways to express the relationships between data elements, and these include relational calculus, statistical associations, axiomatic rule systems, first-order logic, natural language, and other formal and informal systems. In biology, knowledge may include the rules for turning gene expression on and off, the rules for folding protein molecules, the rules for trafficking proteins to their appropriate cellular location, and the networks of regulatory or metabolic interactions that produce the emergent phenomena associated with cellular life.

The field of Artificial Intelligence emerged in the 1970s and 1980s with the goal of producing high levels of performance in tasks normally associated with the need for human intelligence. A variety of technologies emerged, and one of them was expert systems. One of the ideas behind expert systems such as Dendral (Lindsay *et al*., 1980), Mycin (Buchanan and Shortliffe, 1984), and Molgen (Feigenbaum

and Martin, 1977) was that complex performance could be achieved by organizing expert knowledge (in this case, as a set of rules that could be chained together) and then by applying relatively simple algorithms to the organized knowledge to make inferences. The relative success of these systems led to adoption of the mantra that "Knowledge is Power" and the belief that time was best spent eliciting and organizing complex knowledge, and then writing relatively straightforward algorithms to use this knowledge. In the case of rules, forward and backward chaining algorithms were used to "prove" conclusions on the basis of evidence.

Bioinformatics has blossomed as a field because of the recent development of high-throughput biological data collection methods. These methods include genome sequencing, micorarray expression analysis (DeRisi *et al*., 1997), proteomics (Lowe and Jermutus, 2004; Zhu *et al*., 2003), and technologies for measuring metabolite concentrations (Watkins and German, 2002; Mendes, 2002). It is not surprising that the first generation of bioinformatics tools were simply concerned with managing this deluge of data. The data structures were simple, and the tasks were to collect the information, organize it, create databases to hold it, and to search through the collections using various similarity measures. Thus, for DNA sequences, the formats for exchange and storage (Genbank (Benson *et al*., 2004)) were created, basic tools for comparing sequences (Smith and Waterman, 1981; Needleman and Wunsch, 1970) and retrieving them (FASTA (Pearson, 1994), BLAST (Altschul *et al*., 1990)), and for characterizing alignments (Hidden Markov Models (Krogh *et al*., 1994), and many others) were created. Parallel activities have accompanied the emergence of other high-throughput data sources. These tools can often be characterized as "power tools" because they allowed biologists to manage the new wealth of information by either extending previous capabilities that were inefficient or providing new capabilities for systematically examining the data sets, and pulling out individual data items of interest.

The genomic ethic of high-throughput data collection and analysis has led to the emergence of a new approach to biology that has been called "systems biology" or "integrative biology". The key features of this approach are (1) emphasis on high-throughput data gathering, (2) focus on biological systems as a whole, and less as the sum of reduced parts, and (3) reliance on computation to facilitate progress. As part of this evolution, computers are being asked to perform increasingly complex tasks, including some previously performed only by human experts. These include mining and summary of the published literature, creation of pathways and networks based on the experimental associations between genes or gene products, integration of multiple disparate data sources, and comparison of entire systems (genomes, transcriptomes, proteomes), with the goal of generating hypotheses or even suggesting the next set of experiments. These activities have a distinctly different flavor from the powertools of bioinformatics, and really represent the need for intelligent computational assistants to help biologists in their work. One example of an assistant might be a computer program that looks at clusters of genes (generated by a microarray expression clustering powertool) and determines the degree to which the literature on these genes supports their association in a cluster (Raychaudhuri *et al*., 2003; Raychaudhuri and Altman, 2003; Raychaudhuri *et al*., 2002).

The computational infrastructure required to support a biomedical research enterprise that integrates automatic assistants into the process of science is quite

substantial. In particular, there are four main capabilities that are required, and all of these can be considered "knowledge structures" upon which the assistant algorithms of the future will work. These are:

1. *Terminologies.* The first step in creating a durable knowledge infrastructure is to have a standard set of terms, with agreed-upon meanings for each of the domains of biology. These terminologies must be constantly maintained and updated as knowledge expands, but must retain backward compatibility as much as possible, in order to allow programs upon which they are based to continue functioning. Examples of current terminologies that serve as a useful starting point for the biomedical knowledge infrastructure are the MESH terminology used by the National Library of Medicine (NLM) to index Medline/PubMED articles, the UMLS from the NLM that integrates a number of biomedical vocabularies into a common framework (Bodenreider, 2004), the Human Genome Nomenclature Committee (HGNC) gene-naming conventions (Povey *et al*., 2001), elements of the Gene Ontology for classifying gene function (Ashburner *et al*., 2000), and many others.
2. *Ontologies.* The next step in creating the biomedical knowledge infrastructure is the expression of the allowed semantic relationships between terms in the terminologies. For example, in addition to having a term for "Amino acids", "proteins", "genes", and "translation", we need to be able to state that "amino acids" are part of "proteins" and that "genes" are translated into "proteins". This challenge includes at least two parts: the definition of languages for expressing relations and then the acquisition and coding of the actual relations that are relevant in biomedicine. Just as for terminologies, the need to maintain ontologies as a dynamic structure is critical. Ontologies allow us to encode the semantics of our understanding of biological relationships and, therefore, biological knowledge.
3. *Qualitative representations of biological processes.* Terminologies provide us with the basic entities. Ontologies provide the language for expressing relationships between these entities. We use these tools, then, to create complex computational descriptions of our current understanding of biological models. Thus, we may use an ontology to create a qualitative description of the processes, of transcription, translation, protein trafficking, degradation, or other biological processes. These models will necessarily be incomplete, but at any given time they will capture the current best understanding of how things work. Most biological theories begin as qualitative relationships, and only achieve quantitative precision through the execution of experiments designed to characterize the detailed physical and temporal features of these theories. Thus, a qualitative level of representation provides the freedom to experiment with prototype ideas that are not yet quantitatively expressed. Work in this area has included the use of Petri Nets for modeling biological systems (Peleg *et al*., 2002), and the creation of probabilistic models based on statistical associations found in high-throughput data sets (Friedman, 2004; Segal *et al*., 2003; Stuart *et al*., 2003).
4. *Quantitative representations of biological processes.* Eventually, there is sufficient information in a model of a biological process to allow the creation of a quantitative mathematical model that can be used to perform precise simulations of the modeled domain of biology. The resulting quantities, concentrations, and

interactions can be independently validated by comparison with experiments, and the resulting inconsistencies can feed back on the terminologies, ontologies, and qualitative models upon which the simulation was based. In many ways, a faithful mathematical representation of biological systems is an ultimate goal of bioinformatics and computational biology. There have been precious few examples of successful simulations (and always in systems of limited scope, including the lambda-phage lysis or lysogeny decision (McAdams and Shapiro, 1995) or intracellular calcium flux (Slepchenko *et al*., 2002)), but these are the ultimate indication of our understanding of a biological system.

As biological theories become more complex, and the data supporting them becomes more voluminous, computational storage of the theories and the data will be mandatory. Unfortunately, although data can be scored fairly efficiently using current storage technology, the available technology for representing and storing knowledge is still immature. There is still resistance to the idea that computers can be "trusted" with knowledge, and some believe that manipulating knowledge is still strictly the domain of the human mind. Although the current supremacy of the human intellect at managing complex ideas and theories cannot be denied, the value in capturing some or all of this capability in digital computers seems clear. The ability of computers to systematically generate and test hypothesis and mine large volumes of data for associations suggests that they will be able to assist humans in the scientific enterprise when they can represent human knowledge in an effective manner.

This section of the encyclopedia is devoted to the discussion of the second element of the biomedical knowledge infrastructure: ontologies. There is active and important work on all the elements, some of which is reviewed in other parts of this volume. The work on ontologies is particularly important because it promises to provide the basis for the qualitative models of the future, which in turn will evolve into mathematical representations. The elements of this section include first a brief review of the key terminologies in the UMLS (UMLS and associated vocabularies), reviews of the different ontology conventions and representation languages that have been proposed (Frame-based systems, Description Logics, The Gene Ontology Project), basic technologies for managing ontologies (Automatic Concept Identification, Merging and Comparing Ontologies), important general applications of ontology technology (Ontologies for data integration, Ontologies for information retrieval, Ontologies for natural language processing, and a review of TAMBIS), and finally, important emerging application areas in which ontology design and adoption is a major limiting factor for progress (Representing and reasoning about pathways, Ontologies for 3D molecular structure, Ontologies for Proteomics). This is a dynamic and exciting time for the applied science of building and using ontologies, because biomedicine has a clear and present need for these technologies, which are needed to build the next generation of application algorithms.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS and Eppig JT (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (2004) Genbank: update. *Nucleic Acids Research*, **32**(Database issue), D23–D26.

Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–D270.

Buchanan B and Shortliffe E (Eds.) (1984) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley: Reading, MA.

DeRisi JL, Iyer VR and Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Feigenbaum E and Martin N (1977) *Heuristic Programming Project*, Stanford University, Computer Science Department.

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Gruber T (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**, 199–220.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology: application to protein modelling. *Journal of Molecular Biology*, **235**, 1501–1531.

Lindsay R, Buchanan B, Feigenbaum E and Lederberg J (1980) *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*, McGraw-Hill: New York.

Lowe D and Jermutus L (2004) Combinatorial protein biochemistry for therapeutics and proteomics. *Current Pharmaceutical Biotechnology*, **5**, 17–27.

McAdams HH and Shapiro L (1995) Circuit simulation of genetic networks. *Science*, **269**, 650–656.

Mendes P (2002) Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics*, **3**, 134–145.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Pearson WR (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods in Molecular Biology*, **24**, 307–331.

Peleg M, Yeh I and Altman RB (2002) Modeling biological processes using workflow and Petri Net models. *Bioinformatics*, **18**, 825–837.

Povey S, Lovering R, Bruford E, Wright M, Lush M and Wain H (2001) The HUGO gene nomenclature committee (HGNC). *Human Genetics*, **109**, 678–680.

Raychaudhuri S and Altman RB (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401.

Raychaudhuri S, Chang JT, Imam F and Altman RB (2003) The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Research*, **31**, 4553–4560.

Raychaudhuri S, Schutze H and Altman RB (2002) Using text analysis to identify functionally coherent gene groups. *Genome Research*, **12**, 1582–1590.

Schulze-Kremer S (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biology*, **2**, 179–193.

Segal E, Yelensky R and Koller D (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**(Suppl 1), i273–i282.

Slepchenko BM, Schaff JC, Carson JH and Loew LM (2002) Computational cell biology: spatiotemporal simulation of cellular events. *Annual Review of Biophysics and Biomolecular Structure*, **31**, 423–441.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Stuart JM, Segal E, Koller D and Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Watkins SM and German JB (2002) Metabolomics and biochemical profiling in drug discovery and development. *Current Opinion in Molecular Therapeutics*, **4**, 224–228.

Zhu H, Bilgin M and Snyder M (2003) Proteomics. *Annual Review of Biochemistry*, **72**, 783–812.

Introductory Review

BioPAX – biological pathway data exchange format

Gary D. Bader, Michael P. Cary and Chris Sander
Memorial Sloan-Kettering Cancer Center, New York, NY, US

1. Introduction

To understand biological processes, we must integrate new observations with existing knowledge to create testable models that can be iteratively refined. This will be successful only if the vast amounts of data gathered by large-scale profiling of biological features, such as mRNA transcripts and proteins, can be efficiently integrated with data from literature and databases for visualization and analysis. Collecting and integrating biological pathway data is vital for this process. As the number of pathway databases increases, pathway data integration becomes more difficult. At the start of 2006, there were over 215 databases containing pathway information, widely varying in form and content (**www.pathguide.org**). A standard exchange format – a common language – for pathway data, supported by major pathway databases, will significantly reduce the amount of time and energy spent by computational biologists on data integration and lead to increased use of pathway data for biological process modeling, visualization, and analysis (Cary *et al*., 2005).

2. What is a pathway?

One abstraction that biologists have found extremely useful in their efforts to describe and understand the inner workings of the cell is the notion of a biomolecular network, often called a *pathway*. A pathway is a set of interactions, or functional relationships, between the physical and/or genetic (Tong *et al*., 2004) components of the cell which operate in concert to carry out a biological process. Despite tremendous variety in the cellular processes described as pathways, several pathway representation patterns are prevalent in current practice: metabolic, signaling, gene regulation or genetic. These distinctions are mainly organizational, since the cell map represents a fully connected molecular network. Metabolic pathways are generally composed of a series of biochemical reactions geared toward metabolite conversions, whereas signaling pathways generally involve protein–protein interactions and posttranslational modifications. Gene regulation

networks capture transcription factors and the genes they regulate. Genetic pathways are composed of genetic interactions, such as epistasis or synthetic lethality, which occur when two mutations have a combined phenotypic effect not caused by either mutation alone. Additionally, proteomics and functional genomics studies detect molecular interactions, such as protein–protein and protein–DNA, without much detail, but with unparalleled and unbiased coverage (Bader *et al*., 2003).

3. BioPAX – a common language for pathways

BioPAX (**www.biopax.org**) is a community-based effort to develop a biological pathway data exchange format. All development is performed by the BioPAX Workgroup, composed of interested members of pathway database groups and the user community. BioPAX Level 1, which focuses on metabolic pathway information, was released in July 2004. BioPAX Level 2, which adds support for molecular interactions via inclusion of the Proteomics Standards Initiative-Molecular Interaction (PSI-MI) data model (Hermjakob *et al*., 2004), was released in late 2005. Level 3, currently underdevelopment, will add improved support for signal transduction, which often requires the ability to define molecular states, and genetic regulatory networks. Future levels will be able to represent genetic interactions and generic molecules and processes.

BioPAX is being developed in a practical, leveled approach in which each level supports a greater variety of pathway data. BioPAX is implemented in Web Ontology Language (OWL), a standard Extensible Markup Language XML-based language designed for representing ontologies, composed of such things as class hierarchies, properties, and controlled vocabularies. The BioPAX format and any associated software developed by the BioPAX workgroup are open source and freely available to all under the GNU LGPL license (**www.gnu.org/copyleft/lesser.html**). The BioPAX group is coordinating work with other pathway related standards initiatives, such as Systems Biology Markup Language (SBML) (Hucka *et al*., 2003), CellML (Lloyd *et al*., 2004), and PSI-MI (Hermjakob *et al*., 2004), to minimize duplication of work and to ensure compatibility with these standards in areas of overlapping coverage.

4. The BioPAX ontology

Pathways, interactions, and physical entities form the basis of the BioPAX ontology. Four basic classes are defined: the root level entity class and its three subclasses: pathway, interaction, and physicalEntity. *Entity* is a discrete biological unit used when describing pathways and contains a number of basic properties, such as name, database links, and data source, inherited by all subclasses (most of the classes in BioPAX). *Pathway* is a set of interactions or pathway steps with associated evidence. *Interaction* is a relationship between two or more entity participants. *PhysicalEntity* is a tangible physical object, such as a molecule. Physical entities are frequent building blocks of interactions. Interactions at this level are very

general, as they can occur between pathways, interactions, or physical entities. For instance, protein–protein, small molecule–pathway, and pathway–pathway interactions are all valid at this abstract level of the class hierarchy. This allows BioPAX to be flexible when confronted with new types of pathway information. Specialized versions of the interaction class are defined by creating subclasses, which limit the types of entities that can interact and further define properties specific to each interaction type. In this way, classes become less abstract as the hierarchy is traversed from class to subclass, though the number of levels and classes is limited. In this way, the BioPAX ontology remains expressive enough to capture important levels of abstraction that exist in biology, while remaining simple and accessible.

BioPAX Level 2 defines seven main types of interaction and four types of physical entity. Interactions types are: conversion (and three subtypes: complex assembly, transport, biochemical reaction) and control (and two subtypes: catalysis and modulation); physical entities type are: complex, protein, RNA, DNA, and small molecule.

Pathway information in BioPAX is represented by creating instances of these classes. For example, defining a typical enzyme-catalyze biochemical reaction requires physical entity instances for the substrates, products, and enzyme, a biochemical reaction instance to describe the conversion of the substrates to the products, and a catalysis instance to define the relationship between the enzyme and the reaction. For more detail, see the full documentation at **www.biopax.org**.

5. How can BioPAX be used?

5.1. Pathway data warehouse

A pathway data warehouse collects all available pathway information in one place so that it can be conveniently accessed. BioPAX could make creation of pathway data warehouses easier if many databases provided access to their data in the BioPAX format as only the BioPAX format would have to be converted instead of converting a different format for each available dataset.

5.2. Molecular profiling analysis in a pathway context

Molecular profiling experiments, using such technologies as gene expression microarrays and mass spectrometers, are often compared across two or more conditions (e.g., normal tissue and cancerous tissue). The result of this comparison is often a large list of genes that are differentially present in the tissue of interest. It is interesting and useful to analyze these lists of genes in the context of pathways. For instance, one could look for pathways that are statistically overrepresented in the list of differentially expressed genes. The result is a list of pathways that are active or inactive in the condition of interest compared to a control. The list of pathways is often much shorter than the list of differentially expressed genes and thus easier to comprehend.

BioPAX could facilitate this and other kinds of pathway-based analyzes by giving analysis tools easy access to a large body of pathway data in a common format.

5.3. Visualizing pathway diagrams

Pathway diagrams are useful for examining pathway data, as they provide an efficient way to present large amounts of data via the human eye. A number of formats are available for pathway images, but only a few available viewing tools link components such as proteins in the image, to underlying data. A mapping of BioPAX to a symbol library for pathway diagrams (such as Kohn maps (Kohn, 1999) could be used as a basis for a general pathway diagram generation tool.

5.4. Pathway modeling

Mathematical modeling to describe and capture the dynamics of a pathway system is a frequent use of pathway information. Many of the tools available for pathway modeling support the SBML (**sbml.org**) and CellML (**www.cellml.org**) standards, which describe models in sufficient mathematical detail to allow model sharing between tools (actual mathematical equations describing the rates of component reactions in the pathway are captured). While BioPAX is not designed to represent pathway models in high mathematical detail like SBML and CellML, it contains a number of biological concepts not present in these standards, such as "protein" that are required for linking different data sets in a meaningful way. Use of BioPAX to annotate SBML and CellML models could allow linking models to pathway databases and their functional annotations.

6. Available data and tools

Major pathway databases and a growing number of tools support BioPAX. Three large academic metabolic focused pathway databases support BioPAX. The pathway and reaction subset of a significant number of the BioCyc collection of pathway genome databases, some carefully curated such as EcoCyc (Karp *et al*., 2002b), MetaCyc (Karp *et al*., 2002a) and some automatically reconstructed by computer with some curation, such as HumanCyc (Romero *et al*., 2005), are available in BioPAX Level 1 format at **www.biocyc.org.** The pathway database of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*., 2004) is also available in BioPAX Level 1 format from the KEGG FTP site. The WIT database (Overbeek *et al*., 2000), makes all reconstructions of metabolic networks derived from public DNA sequence data available in BioPAX Level 1 format. Major academic signal transduction pathway focused databases are starting to make their data available in BioPAX Level 2. This includes the Reactome database (Joshi-Tope *et al*., 2005) of human-focused pathways, including metabolic and signaling pathways, the aMAZE (Lemer *et al*., 2004) pathway

database and the Integrating Network Objects with Hierarchies (INOH) pathway database (Fukuda and Takagi, 2001), all of which contain hand collected pathway information.

BioPAX formatted data can be visualized by the VisANT (Hu *et al*., 2004) and Cytoscape (Shannon *et al*., 2003) network analysis and visualization tools. Pathways in the PATIKA pathway visualizer and editor can be exported in BioPAX format. Since BioPAX is written in OWL, a number of generic OWL tools are available for storing, querying, browsing, creating, and editing BioPAX data. Protégé is a powerful ontology editor that is freely available and has been used with the Protégé OWL plug-in (Knublauch *et al*., 2004) to create the BioPAX ontology. SWOOP (Kalyanpur *et al*., 2004) is an OWL file browser that can be used with BioPAX pathways. Further a number of software libraries, such as Jena and OWL-API (Applications Programming Interface), written in Java, are useful for developing software that can read, process, and write BioPAX OWL files. OWL/RDF (Resource Description Framework) database storage tools, such as the Oracle RDF database and InstanceStore (Horrocks *et al*., 2004) are also available. Simple Protocol and RDF Query Language (SPARQL) and RDF Data Query Language (RDQL) languages act like Structured Query Language (SQL) for OWL/RDF data. More advanced querying tools, such as the Pellet reasoner, can take advantage of logic and the structure of an OWL ontology to allow more powerful queries. These generic OWL technologies are still new, so expect them to require maturation before they fulfill their promise of easy storage and manipulation of large amounts of BioPAX data. Finally, since OWL is XML based, any generic XML software development tools are compatible with BioPAX files.

7. Conclusion

The ultimate aim of projects like BioPAX is to enable effortless collection of pathway data so that it may be efficiently applied to answer biological questions. Ideally, biologists should never need to perform time-consuming data collection tasks in order to perform a particular analysis. Instead, they should be able to locate, retrieve, and apply data of interest without worrying about data models, exchange formats, or integration methods.

To achieve this goal, data standards must become broadly adopted by pathway databases. This would enable a variety of large-scale data integration approaches, such as a centralized or distributed pathway data warehouse or a query engine able to retrieve data from multiple standards-compliant primary databases. Importantly, pathway data analysis tools must be built to interface with these integration systems to make pathway data retrieval painless.

With public data sharing infrastructure, we can build software platforms that allow high-level and effective pathway data manipulation and analysis. When combined with the trend toward cheap, high-throughput cellular profiling technology, we can imagine a swift convergence on biological process understanding through iterative systems biology modeling methods.

Acknowledgments

BioPAX is a community effort and work has involved the following people: Level 1: Gary D. Bader, Erik Brauner, Michael P. Cary, Robert Goldberg, Chris Hogue, Peter Karp, Joanne Luciano, Debbie Marks, Natalia Maltsev, Eric Neumann, Suzanne Paley, John Pick, Aviv Regev, Andrey Rzhetsky, Chris Sander, Vincent Schachter, Imran Shah, Jeremy Zucker.

Level 2: Mirit Aladjem, Gary D. Bader, Michael P. Cary, Kam Dahlquist, Emek Demir, Peter D'Eustachio, Ken Fukuda, Frank Gibbons, Marc Gillespie, Chris Hogue, Michael Hucka, Geeta Joshi-Tope, David Kane, Peter Karp, Christian Lemer, Joanne Luciano, Natalia Maltsev, Eric Neumann, Suzanne Paley, Elgar Pichler, Jonathan Rees, Alan Ruttenberg, Andrey Rzhetsky, Chris Sander, Vincent Schachter, Andrea Splendiani, Mustafa Syed, Edgar Wingender, Guanming Wu, Jeremy Zucker.

The BioPAX workgroup has been coordinated by Mike Cary and Gary Bader in Chris Sander's group with help from many of the above mentioned people.

References

Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T and Boone C (2003) Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends in Cell Biology*, **13**(7), 344–356.

Cary MP, Bader GD and Sander C (2005) Pathway information for systems biology. *FEBS Letters*, **579**(8), 1815–1820.

Fukuda K and Takagi T (2001) Knowledge representation of signal transduction pathways. *Bioinformatics*, **17**(9), 829–837.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, van Mering C, *et al.* (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**(2), 177–183.

Horrocks I, Turi D, Li L, Bechhofer S (2004) *The Instance Store: Description Logic Reasoning with Large Numbers of Individuals*. Proc. of the 2004 Description Logic Workshop (DL 2004).

Hu Z, Mellor J, Wu J and DeLisi C (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, **5**(1), 17.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**(4), 524–531.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**, D428–D432, Database Issue.

Kalyanpur A, Sirin E, Parsia B and Hendler J (2004) *Hypermedia inspired ontology engineering environment: SWOOP. International Semantic Web Conference (ISWC)*, Hiroshima, Japan.

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**, D277–D280, Database issue.

Karp PD, Riley M, Paley SM and Pellegrini-Toole A (2002a) The MetaCyc database. **30**(1): 59–61.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C and Gama-Castro S (2002b) The EcoCyc database. *Nucleic Acids Research*, **30**(1), 56–58.

Knublauch H, Fergerson RW, Noy NF and Musen MA (2004). *The protégé OWL plugin: an open development environment for semantic web applications. Third International Semantic Web Conference - ISWC 2004*, Hiroshima, Japan.

Kohn KW (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. **10**(8), 2703–2734.

Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J and Wodak SJ (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Research*, **32**, D443–D448, Database issue.

Lloyd CM, Halstead MD and Nielsen PF (2004) CellML: its future, present and past. *Progress in Biophysics and Molecular Biology*, **85**(2-3), 433–450.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr, Kyrpides N, Fonstein M, Maltsev N and Selkov E (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. **28**(1): 123–125.

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M and Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, **6**(1), R2.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11), 2498–2504.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**(5659), 808–813.

Specialist Review

Ontologies for the life sciences

Steffen Schulze-Kremer
Universtät Hannover, Hannover, Germany

Barry Smith
Universität Leipzig, Leipzig, Germany
University at Buffalo, Buffalo, NY, USA

1. Introduction

The multitude of heterogeneous and autonomous data resources available to life scientists today includes genomic (Fasman *et al*., 1996), cellular (Jacobson and Anagnostopoulos, 1996), structural (beginning with the seminal Bernstein *et al*., 1977), phenotype (McKusick, 1994), and a range of other types of biologically relevant information (Bairoch, 1993). Even for one type of information, for example, DNA sequence data, there exist several databases of different scope and organization (Fasman *et al*., 1996; Keen *et al*., 1996; Benson *et al*., 1997 and their successors).

There exist terminological differences (alternative synonyms, aliases), syntactic differences (in file structure, separators, spelling), and semantic differences (the same words are used to mean different things in different sources). Conventions for naming data objects, object identifier codes, and record labels differ between databases and do not follow a unified scheme. Even terms for important high-level concepts and relations that are fundamental to life sciences are often used in conflicting or ambiguous ways (Smith and Rosse, 2004).

One prominent example is the concept *gene*. For GDB (Fasman *et al*., 1996), a gene is a "DNA fragment that can be transcribed and translated into a protein". For GenBank (Benson *et al*., 1997) and GSDB (Keen *et al*., 1996), a gene is a "DNA region of biological interest with a name and that carries a genetic trait or phenotype". The latter definition is problematic, not only because it makes the answer to the question of which genes exist depend on the vagaries of human naming acts but also because it comprehends nonstructural coding DNA regions such as intron, promoter, and enhancer. There is a clear distinction between the two underlying notions of *gene,* but both continue to be used, thereby adding another level of complexity to data integration. Another term with multiple meanings is *protein function* (biochemical function, e.g., enzyme catalysis; genetic function, e.g., transcription repressor; cellular function, e.g., scaffold; physiological function, e.g., signal transducer).

If a user queries a database using an ambiguous term, she must herself take responsibility for verifying the congruence in meaning between her use of the term and what the database returns. Those semantic incompatibilities that are known must be resolved with each new search result, while those that are unknown propagate errors behind the scenes. Ontologies can help resolve such incompatibilities in a global fashion, for example, by flagging cases where a single term is used for entities in distinct ontological categories and enforcing manual disambiguation.

The advent of microarray technology for mRNA expression analysis requires additional standardization in terminology for characterizing not only genes, tissues, and samples but also experimental setups and the factors involved in mathematical postprocessing of raw measurements. A comparison between different experiments is only feasible if consistent terminology and standardized input forms are used. The development of ontologies suitable for this purpose is pursued by the MGED consortium (Brazma *et al*., 2001).

Standardized nomenclatures are required also where the new more integrated approach to biology has led to the merging of subfields with historically independent origins. This applies, for example, to genetics, protein chemistry, and pharmacology. Pharmaceutical companies have expressed an urgent need to harmonize the technical languages of these fields so that the knowledge derived from each can be stored in a unified way. The fast growth of sequence, structure, expression, metabolic, and regulatory data pertaining to many organisms adds additional pressure to utilize standardized and compatible nomenclature in molecular biology.

Text mining and natural language understanding in biology can also profit from ontologies. Currently, it is mostly techniques based on string-based statistical and proximity approaches that are applied to text analysis. Ontologies can, however, support parsing and disambiguating sentences by enforcing grammatically compatible uses of terms via rules for ontologically compatible combinations of referents (Jackson and Ceusters, 2002; Nirenburg and Raskin, 2001).

To support consistent reporting of results in molecular biology, it will be necessary to develop controlled vocabularies comprehending the most important and frequently used terms with coherent definitions in such a way as to allow database managers, curators, and annotators to create new and more coherent software and database schemata, to provide exact, semantically precise specification of the concepts used in existing schemata, and to curate and annotate existing database entries in a consistent way.

It is important to understand that terminological ambiguity affects not only interoperability on the level of computers but also communication between human beings. But where humans have the facility to resolve ambiguities in an efficient way, computer programs and databases have, for the moment at least, no analogs of the capabilities that allow such concurrent disambiguation. Ontologies are one important tool designed to make up for this shortfall.

Data integration, too, must overcome the problem of syntactic and semantic heterogeneity. While some syntactic incompatibilities, for example, prefix versus postfix operators, can be easily aligned automatically, incompatibilities that arise from differences in meaning require manual resolution within a common unifying framework. Thus, for example, each table, object, and so on, of one database must be manually aligned with the corresponding components of each other database that

is to be integrated. If we begin with *n* databases, and if the process of integration is carried out in pairwise fashion, then this requires $n*n$ attempts at resolution of differences in meaning. If, however, a single ontology exists that can serve as a central switchboard for those *n* databases, then the integration effort is reduced from $n*n$ to *n*, since each database has to be mapped to the ontology only once in order to become interoperable with any other database (Köhler and Schulze-Kremer, 2002).

2. Overview of ontologies

Work in ontology can be classified along a number of distinct dimensions. Most important is the distinction between (1) ontology as the study of beings or entities, the study of what exists at the highest level of abstraction – ontology as a branch of philosophy – and (2) domain ontologies, which result from the analysis of particular domains of reality and correspond broadly to separate areas of scientific inquiry.

Ontologies in either of these senses are, in principle, language independent. Thus, there can be a German equivalent to an English domain ontology, even if the actual translation process need not be trivial. A domain ontology may also be formalizable in some artificial language such as the language of first-order predicate logic, Description Logic (Rector *et al*., 2003), or some other form of representation formalism.

Ontology as a branch of philosophy can be important in bringing clarity to the life science field even before we enter the specific territory of biological domain ontologies. Thus, consider the fact that "DNA" can be used to designate quite different entities. First, there is DNA as physical stuff, which can be measured with a spectrophotometer. Second, there is the class of all of the chemical substances that share the general features common to DNA molecules. Third, there is the family of specific types of sequences or strings in the sense of abstract structures that can be subject to mathematical operations but cannot be measured or detected in reality. Fourth, "DNA" is often used in the lab to refer to a particular instance of a sequence, for example, the DNA sequence of *E. coli* K12 that can be stored in a database and needs carrier (memory chip, paper) to survive. These and similar distinctions – between classes and instances, between sequences and stuffs – are distinctions between philosophical categories, which have analogs in a number of different domains, and which are unfortunately often the source of confusions, not least when attempts are made at systematic representation of empirical knowledge for purposes of automatic information retrieval.

Among domain ontologies, we can draw distinctions between ontologies of varying scope and content. Thus, we can distinguish between

- upper-level ontologies, which are primarily concerned with the small number of general categories (such as *cell*, or *gene*, or *molecule*) that serve as the basis of our understanding of a particular domain;
- terminology-based ontologies, which are centered around the many highly specific terms used in the formulation of the results of scientific inquiry (such as *enzyme active site formation, postsynaptic membrane,* or *receptor signaling protein activity*).

Since the world around us in general and molecular biology and bioinformatics in particular are such as to manifest an enormous multidimensional complexity, no single ontology can suffice for every purpose. Rather, we must content ourselves with ontologies representing different views of reality created in connection with practical goals – often views of reality at different levels of granularity (Bittner and Smith, 2003). Thus, before building an ontology, it is important to understand what is its intended use, since otherwise there is a risk of being overwhelmed by the multitude of facets by which we are confronted. This aspect is acknowledged by the use of the term “situated ontologies” (Mahesh and Nirenburg, 1995) to emphasize the fact that a domain ontology should be evaluated with respect to its intended use.

The larger terminology-based ontologies clearly must be constantly updated in light of new experimental evidence and developments in language usage. At the highest levels, however, an ontology is designed to be much more stable than, for example, a database schema. The latter is dependent on specific choices concerning a database representation formalism, database management system, and requirements from the applications that access the data. Since an upper-level ontology is by its nature designed to be easily translatable from one knowledge representation formalism to another (given equivalent expressive capability), it can also be converted into a database schema. But a domain ontology addressing the fundamental categories and relations of an application domain is designed to be independent of given software implementations. When new knowledge classes are discovered, the ontology should be extendible in relatively straightforward ways, along lines to be described below.

The interplay between ontologies, biology, computer science, linguistics, and philosophy is depicted in Figure 1.

2.1. *Upper-level ontologies*

The first important ontologist was Aristotle (384–322 BC) who, among many other things, pursued the question of how reality is organized into classes or universals. His solution is presented in his *Categories*, which can be seen as the first upper-level ontology (Barnes, 1984). From Aristotle’s point of view, 10 categories suffice to express anything that can be known about something:

- Substance
- Quantity
- Quality
- Relation
- Place
- Time
- Situation
- Condition
- Action
- Affection.

Of course, from the point of view of the annotation of entities in molecular biology, the categories here distinguished will not suffice. However, if one subscribes

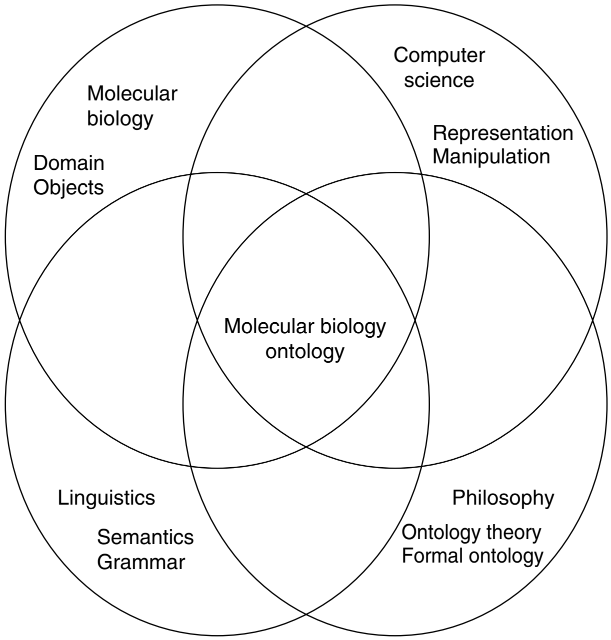


Figure 1 Molecular biologists discover facts that need to be organized and stored in databases. Computer scientists provide techniques for data representation and manipulation. Linguists help organize the meanings underlying database labels. Philosophers provide formal theories of basic ontological relations and principles governing the best practice in definition and classification

to the view that Aristotle provides a still serviceable account of the most fundamental set of categories, then one could see molecular biology and other special life sciences as the results of further subclassifications of Aristotle's categories into ever more specific kinds.

Another feature of Aristotle's ontology is the paucity of interconnections between his 10 categories, each of which is assumed to be an atomic category in the sense that it cannot be meaningfully decomposed into smaller units. Aristotle does allow that substance is the primary category, so that instances of all other categories are *dependent on* instances of substance. As concerns the interrelations between the nine "accidental" categories, however, he tells us too little. Later ontologists added further interrelations between their basic categories, including a taxonomy of different kinds of dependence relations provided by Husserl in his *Logical Investigations* (Husserl, 1970) and the related taxonomy offered in our own day by the DOLCE ontology and depicted in Figure 2 (Guarino, 1997; Masolo *et al*., 2003).

A contemporary philosophically motivated upper-level ontology for Molecular Biology (MBO) is advanced in Schulze-Kremer (1997). Like DOLCE, this starts from a single node, and it also extends to incorporate those physical and abstract entities that are relevant for biology and bioinformatics.

The upper level of the Molecular Biology Ontology (MBO) is shown in Figure 3. Starting from the root node *Being,* which includes all entities of any sort, it distinguishes two disjoint classes of *Object* and *Event,* which are discriminated on

- Entity
  - Particular (e.g., "large molecule", "green spot")
    - Concrete particular
      - Location
      - Object
    - Abstract particular
      - Set
      - Structure
  - Universal (e.g., "largeness", "color")
    - Property
      - Property kinds...
    - Relation

Figure 2 DOLCE top-level ontology

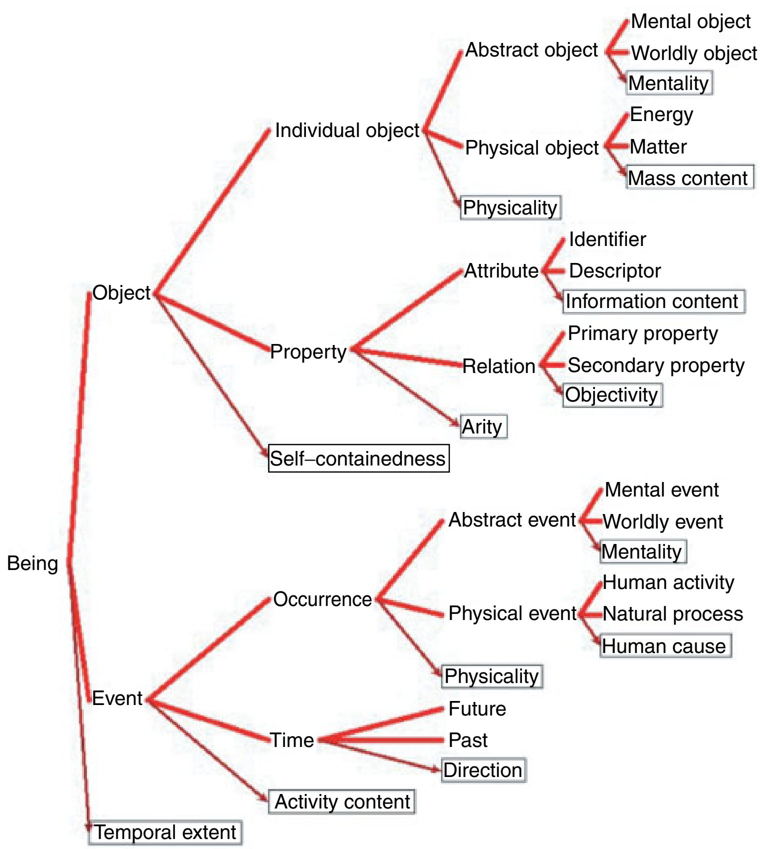


Figure 3 Upper level of the Molecular Biology Ontology of Schulze-Kremer (1997). Links represent the SUBCLASS-OF relation. Discriminating criteria are marked by arrows and boxes; thick lines denote disjoint subclasses (Reproduced by permission of The American Association for Artificial Intelligence)

the basis of their mode of existence in time. An *Object* retains its identity from one moment to the next; an *Event,* in contrast, is divided into temporal parts or phases and unfolds itself through these phases in successive moments of time. This distinction is passed on to all subclasses of *Object* and *Event*. The class *Object* is further subclassified into *Individual Object* and *Property*. Both preserve their identity from one moment to the next. They are discriminated on the basis of their ability to exist in a self-contained way. An *Individual Object* can stand alone, whereas a *Property* always needs another *Object* or *Event* of which it is the property. *Property* is further subclassified on the basis of distinctions in arity, into *Attribute*, a property with only one argument, and *Relation*, a property relating two or more *Beings*.

Attribute is subclassified into *Identifier* (e.g., "ID-2394873") and *Descriptor* (e.g., "*E. coli* R12 DNA") on the basis of whether an item simply labels an entity or carries additional information about it. *Relation* can be subclassified into *Secondary Property,* each instance of which involves some relation to a cognitive subject, and *Primary Property*, whose instances are objective and measurable entities such as mass or charge (Locke, 1975).

Individual Object is subclassified via the criterion of physicality into *Abstract Object*, which has no physical equivalent per se (except the capacity for being represented in writing, etc.), and *Physical Object*, which must have a defined spatial extent and/or energy content. *Physical Object* is further subclassified on the basis of mass content into *Energy* and *Matter*.

Abstract Object is further subclassified via the criterion of mentality – that is, according to whether it refers to an object within the mind or to an object in the outside world – into *Mental Object* (e.g., thought, love) and *Worldly Object* (e.g., circle, sequence).

The category *Event* is subclassified via the criterion of activity into *Occurrence*, in the instances of which at least one object participates, and *Time*, where this is not the case.

Time is further subclassified according to direction into *Past* and *Future*. Because the present moment is, strictly speaking, instantaneous, it does not appear in this branch. Analogous to abstract and physical objects, *Occurrence* is subclassified via the criterion of physicality into *Abstract Event* and *Physical Event.* The former is further classified on the basis of the criterion of mentality into *Mental Event* (e.g., thinking, feeling) and *Worldly Event* (e.g., binding, transport). *Physical Event* is subclassified on the basis of whether it is initiated by human intention into *Human Activity* and *Natural Process*. The operations of man-made devices in laboratory experimentation fall under the first of these two headings, the natural processes subject to molecular biological analysis under the second.

Another closely related more general purpose ontology is the Basic Formal Ontology (BFO) (Grenon *et al*., 2003; Grenon and Smith, 2004), which draws a still more radical distinction than that between *Objects* and *Events* by distinguishing two separate ontologies, called SNAP, for enduring entities (such as organisms, cells, enduring attributes, functions, and dispositions) and SPAN, for processes or events (Figure 4).

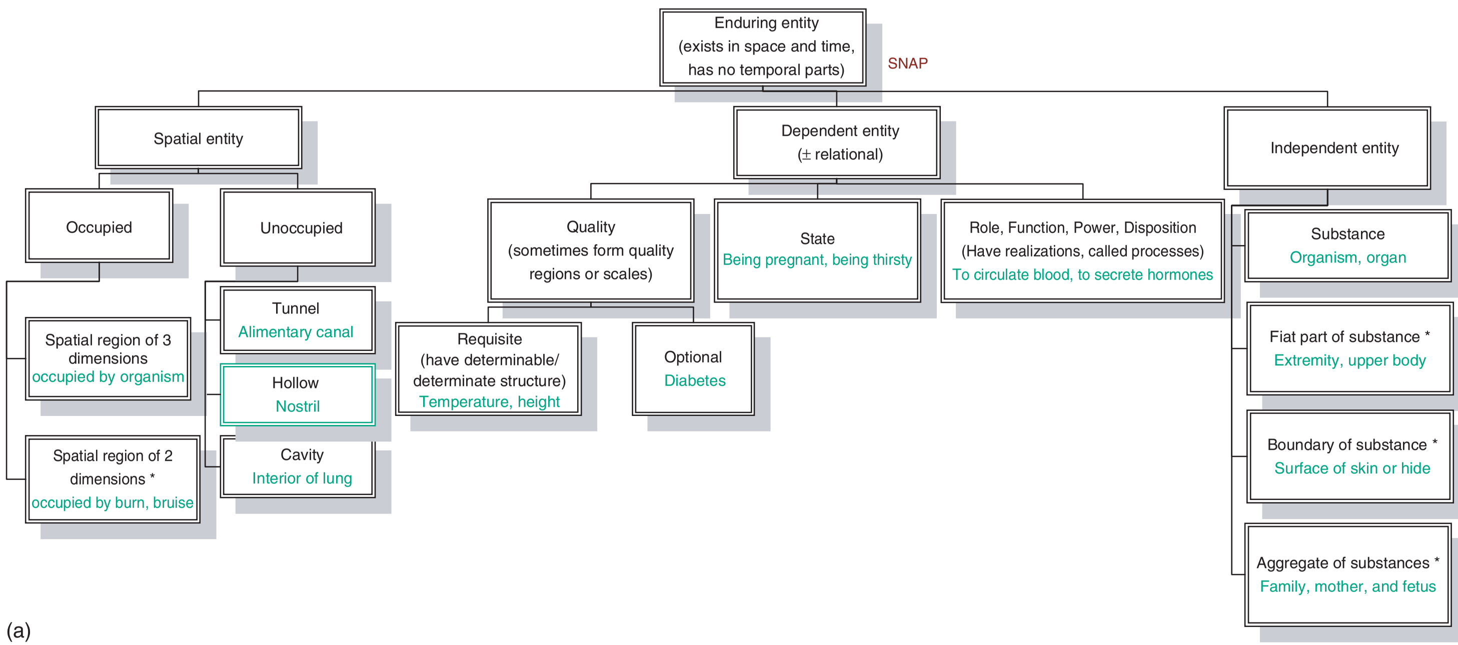


Figure 4 Top-level categories of the (a) SNAP and (b) SPAN ontologies of Basic Formal Ontology (BFO)

(b)

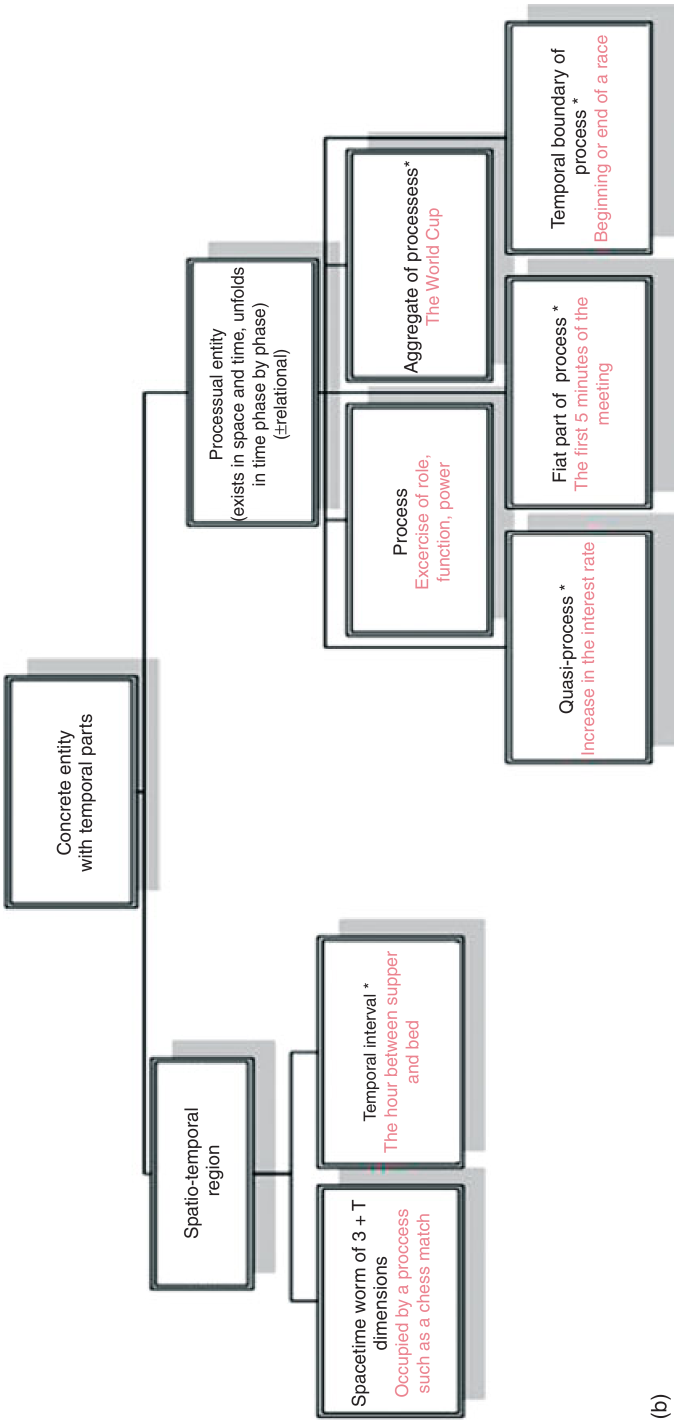


Figure 4 (*continued*)

3. What is an ontology?

Three conceptions of domain ontologies can be distinguished:

1. A concise and unambiguous description of principal relevant entities with their potential, valid relations to each other (Schulze-Kremer, 1998).
2. A system of categories accounting for a particular vision of the world (Guarino, 1998).
3. A specification of a conceptualization (Gruber, 1993).

(1) and (2) represent a view of ontology broadly in the spirit of Aristotle's ontology and of traditional philosophical ontology, and realized in the top-level category systems illustrated above. (3) rests on a view of ontologies rooted in logic-based knowledge representation; it tells us that to build an ontology we must analyze our domain of interest and represent the basic concepts that are exemplified therein in some formal language. Although this describes in broad terms some of what is involved in ontology development, the definition itself does not yet go far enough, and we here specify some further requirements that a domain ontology should satisfy.

1. Each term in an ontology should be defined as precisely as possible. Definitions are the basis for establishing the relations between terms in an unambiguous way and are indispensable when laying down the foundation of an ontology. But writing good definitions is often very hard (not least in the domains of the life sciences). How detailed does one have to be in specifying the concept at hand to make it distinguishable from others already present and those that are to be added in the future? Since this question cannot be answered a priori, definitions frequently need to be updated and such updating should be supported by ontology editing software.
2. The set of concepts covered by an ontology must comprehend all the categories instantiated by the entities in the application domain at the pertinent level of granularity.
3. There should be a specification of the structure or organization of an ontology. This can be by means of an ontology representation language such as KIF (Genesereth and Fikes, 1992) or DAML + OIL (Joint United States/European Union ad hoc Agent Markup Language Committee, 2001; *see also* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8), or it can be a semiformal or informal specification, for example, utilizing UML (Unified Modeling Language) diagrams or natural language.
4. An ontology should be associated with documentation specifying its expressive capabilities, as well as its scope and the levels of granularity of the entities with which it deals.
5. Standardized procedures should be defined specifying how to add, modify, or remove categories.
6. Indications should be given specifying how the ontology and its categories can be used, for example, what kind of inference is supported, how it can be applied, for example, to support information retrieval.

4. Components

The building blocks of an ontology are as follows:

1. *Terms* are the atomic building blocks of an ontology conceived as a syntactic structure.
2. *Links between terms*, for instance, the link between “mammal” and “animal”, representing the fact that the class designated by the former term is included as subclass within the class designated by the latter.
3. *Predicates*, for instance, TRANSCRIBED-BY, CATALYZED-BY, IS 2-DIMENSION.
4. *Propositions* are definite statements about (parts of) the world, sometimes encoded in an ontology representation language and typically representing relations between classes. They include as a special distinguished group: *axioms*, which are fundamental statements that are assumed to be true and that are given without proof. The set of axioms must be logically consistent. Some propositions can be derived from axioms via logical reasoning.
5. A *logical formalism*, including logical constants such as *and* and *or*, variables, and quantifiers.
6. *Definitions*, which can be divided into real, nominal, and ostensive. Real definitions capture the essence or nature of the entities referred to by the term to be defined: they specify the universal marks, which all instances of a defined category share (Michael *et al*., 2001). Nominal definitions reflect the way a given term is used. They may be analytic, which means that they decompose the concept to be defined into its necessary and sufficient conditions (e.g., bachelor is an unmarried man). Or they may be stipulative, which means that they serve to introduce a new concept (e.g., an alpha-helix peptide is a polypeptide molecule with the following geometry . . . , nonlytic viral exocytosis is the exit of the virion particle from the host cell by exocytosis, without causing cell lysis). Ostensive definitions define concepts by pointing to or by enumerating examples, as when we define “yellow” by pointing to yellow things.

Some principal rules governing the formulation of good definitions are as follows:

- Definitions should not be negative in form, for example: *protein is not made of DNA*.
- Definitions should not be too broad. Thus, *proteins are chemicals* is to be rejected.
- Definitions should not be too narrow. *Proteins are covalent strings of amino acids* is to be rejected because it does not embrace posttranslational modifications and quaternary structure.
- Definitions should not be circular (*A protein is made of a protein chain*). Definitions should not convey extra or redundant information. Consider, for example, *regulation of protein catabolism is any process that modulates the frequency, rate, or extent of the breakdown into simpler components of a protein by the destruction of the native, active configuration, with or without the hydrolysis of peptide bonds*.

In addition, we have that in the world toward which an ontology is directed. This includes:

1. The *classes of entities* in reality to which terms refer. These are generalizations of instances, for example, *gene, protein*, and are connected with each other inter alia by the SUPERCLASS-OF and SUBCLASS-OF relations. Often, SUBCLASS-OF is also called IS-A. If one class stands in an immediate SUBCLASS-OF relation to a second class, we say that they stand in the relation of *child* to *parent*; two classes with the same parent are called *siblings*; classes with no children are called *leaves*; a class with no parents is called a *root*.
2. These are *entities* themselves, that is, the instances of classes that are individuals (such as this organism or this sample of protein) connected by the IS-OF-TYPE relation to at least one class. Normally instances play a role only generically – thus, we define the PART-OF relation between classes as follows:
3. *A* PART-OF *B* = def. All instances of *A* are such that there is some instance of *B* of which they are a part (Smith and Rosse, 2004).
   However, specific experiments or specific research findings may enjoy a biological significance and may be referred to explicitly in a corresponding domain ontology (e.g., Mutated-HeLa-Cell-Sample, *E. coli*-R12-Batch-001-DNA-Sequence).
4. *Attributes* and *Relations* to which the predicates refer.

5. Distinguishing marks of ontologies

An ontology is to be distinguished from *a knowledge base* conceived as a *collection of statements of fact.* Rather, an ontology is a specification of the principal categories instantiated by the entities in a given domain of reality and at a given level of granularity. Thus, one might say that an ontology is a certain highly general type of knowledge base filled with knowledge about categories and their ontological relations.

An ontology is to be distinguished from *a model of an application domain*; rather it is a compendium of the building blocks of such domains together with their valid modes of combination, the whole formulated as a theory.

An ontology is *not a database schema*, that is, it does not describe the categories, data types, and organization in a database. Rather, it is a specification of the classes and relations among entities in the real world. Such datatypes may represent such classes and relations also, which means that a database schema can be derived from an ontology by adding data type information and translating the knowledge representation formalism into a database management format. A database schema can also be used as a starting point for ontology building. The corresponding entity types and attributes can then be taken as an initial set of categories to populate an ontology.

An ontology is *not a taxonomy* that knows only about superclass and subclass relations; an ontology is open to other types of fundamental relations, including temporal, mereological (part-whole), topological, compositional, and casual relations, as well as dependence relations, for example, between qualities and functions and the objects which they are the qualities and functions of.

An ontology is *not a vocabulary* or *dictionary* or *thesaurus,* since all these lack the logical organization that an ontology demands in order to support computational inference. Moreover, they standardly do not describe the hierarchy and relations between the entities designated by the terms they include. In an ontology, one can follow a path from any term to any other along the edges of some IS-A hierarchy or by following other relations.

An ontology is *not a semantic net*. Rather, a semantic net is one sort of formalism that can be used to represent ontologies, but this formalism can be used for other purposes also.

6. How to build an ontology?

There are several ways to build an ontology, some of which are surveyed in Fernandez *et al*. (1997). A currently popular methodology, especially in the biological and medical domains, is text mining (Maedche and Staab, 2003; Kashyap *et al*., 2003). Here we describe the method used in Schulze-Kremer (1998).

In order to assemble the components described above (terms, propositions, axioms, formalism, subclass relations, etc.), we apply the following steps in succession:

- Collect an initial list of domain-relevant terms. These can be taken, for example, from database tables or textbooks. This list will have to cover the main central objects and processes in that application domain and will further be extended when populating the ontology.
- Provide a unique and explicit definition for each high-level term. This definition must be precise enough to discriminate the reference of that term from all other entities referred to in the ontology and it should be detailed enough to provide a clear representation of the term's meaning. Experts often have only a tacit understanding of the technical terms they use; thus, it is often difficult to provide an explicit formal definition, not least given the ambiguities by which many terms are affected. Ontology management software should therefore be capable of disambiguating terms with multiple meanings, for example, by imposing unique identifiers.

With the move to lower-level terms as the details of an ontology are filled in, we need to add subclasses to classes already recognized. Here, it is important to fix upon and to use consistently and explicitly one and only one discriminating criterion for each superclass. When this design principle is followed, the ontology automatically manifests the properly hierarchical structure of a tree of subclasses that can also be used as a decision tree when adding or searching for terms. To use the hierarchical tree of subclasses, one starts at the single top-most node and applies the discriminating criterion at each level to the new term. Depending on the characterization of the new term with respect to the discriminating criterion, a decision is reached as to which subclass to follow. For full expressivity, this will require a choice among a number of inheritance modes (e.g., multiple distinct inheritance, where the subclass can be distinctly interpreted depending on its parent

types, e.g., *queen*, seen as a monarch or as a piece on the chess board; or combined cumulative inheritance, where all properties of all parents are inherited by the child term, e.g., a protein-DNA complex inherits features from both DNA and protein). Ideally one should use the same classification criterion throughout the ontology, as is done in the Foundational Model of Anatomy, where the single criterion of *structure* is used (Rosse and Mejino, 2003).

- Be explicit about the disjointness of subclasses, that is, state where subclasses of a single class can or cannot overlap. For example, the distinction of molecules into pure protein and DNA is disjoint, since no molecule can be both at the same time. This greatly helps to focus searches through the subclass hierarchy, since if it is known in advance that a subclassification is disjoint, then only one of its subclasses need be followed when proceeding further down the hierarchy. Using only disjoint subclasses is also called the "single inheritance" mode and implies the creation of a true hierarchy with no fusion between branches as one moves down the tree to more specialized classes.
- Obtain complete connectivity via SUBCLASS-OF relations (or their inverses) from any one term in the ontology to any other term. Thus, at least one SUBCLASS-OF relation (or its inverse) must exist for each term. In this way, we ensure that all terms are defined consistently in such a way as to form a single ontology with no separate ontological islands, which would require integration later on.
- Use one root node only. This root must be general enough to embrace the entirety of the domain-relevant categories, since otherwise different conflicting lineages could emerge.
- Add background knowledge for each term to express domain-relevant properties, but keep this strictly separate from the definitions. The attributes and relations should themselves be reified first (by adding corresponding terms to the ontology) for maximal inference capability.
- Add links from terms in the ontology to other ontologies, natural language dictionaries, database keywords, and so on, thereby interfacing the ontology with applications of various types and supporting its integration with other information sources.

6.1. Guidelines on syntax

The following have emerged as rules of good syntactic practice that serve to make an ontology more manageable for human users.

- Use singular rather than plural forms in a term name.
- Use lower case letters only for terms for classes.
  - Names of instances should begin with a capital letter, for example, *E. coli*-Strain-K12-Sequence.
  - Acronyms should be upper case throughout.
- Observe syntax requirements of the selected representation formalism.
  - Quotes, hyphens, and so on, may be required or forbidden.

  - Unique names may be required by the representation formalism.
- When naming a subclass, start by specializing the name of the superclass.
  - The specializing text should be appended, not prepended.
  - This makes the term easier to be recognized.
- Always provide aliases where known, and keep records of equivalences.

If these rules are followed, this means that when adding a new term one can use the discriminating criteria of the ontology as a decision tree to travel down from the root and at each branch deterministically decide where each new term should belong. One then either finds that the term is already there (possibly under another name), and the insertion process consists merely in the addition of another alias to the existing term, or one ends at some point in the hierarchy where no appropriate superclass can be found. This is then the place where the new term should be added, either directly or by introducing intermediary terms designed to separate the already existing terms and branches from the branch to be newly created. This also guarantees that the ontology remains consistent after a new term has been inserted. Searching for a known or even unknown term can be done in the same way, that is, by traversing the decision tree of discriminating criteria.

There are several difficulties to be overcome when building an ontology. Some difficulties are inherent to the ontology building process, others reflect specific application areas. First is the problem of determining the best (e.g., most informative) criterion of subclassification for a given class. Here one faces, to some degree, an arbitrary decision as to how to proceed in creating subclasses, and this implies that there will, in general, not be one single optimal ontology for a given domain at a given level of granularity but rather only ontologies more or less well integrated with other information resources, having greater or lesser reasoning power, and so on. Also, since the information content of the terms that will need to be added to an ontology in the future cannot be known in advance, the choice of subclassifying criteria may lead to a more complex inheritance structure than necessary, and may thus itself have to be revised.

Other difficulties arising in the ontology building process are the following:

- For many application domains, it is unrealistic to aim for exhaustiveness of an ontology. However, each domain ontology must cover all entities that are of practical relevance for its application domain.
- The arity of relations may be a source of confusion.
  - Relations may be 1:1 (e.g., each person has a social security number that is unique to that person) or 1:Many (e.g., each single person has one weight under standard conditions but several people may have the same weight).
  - There are also 1:Many relations (e.g., a fountain pen writes in a single color but one color may be used by several pens) and Many:Many relations (e.g., a shirt may have several colors and each color may be present in several shirts).
- There is the danger of overelaborating an ontology by getting lost in those branches that are already well understood and thus face few representational difficulties, but which are of little relevance to applications.
  - Therefore, avoid superfluous ontological elements.
  - Check whether all details are really relevant to the intended purpose.

- Storing important data as free text or in comment fields rather than as defined terms can lead to confusion since free text fields are not well accessible for automatic reasoning. Thus, wherever possible one should encode the quality with which to annotate another term as a term itself in the ontology, thereby making its scope explicit and enabling links to its inverse and other relations.
- Multiple inheritance should be carefully applied to make sure that the resulting subclasses really exist as genuine classes. Single inheritance is generally safer and easier to understand.

Of the domain-specific difficulties in ontology building, ill-defined technical terms, controversial technical terms, difficulty of analyzing and separating homonyms, imprecise or lacking documentation of database categories are the most common.

The degree of abstraction and detail one chooses to adopt in building an ontology of a given domain at a given level of granularity determines the practical quality of the ontology, which results in a range from useless (too abstract, only upper-level terms defined that do not give sufficiently detailed information) to impossible to complete (ultimate granularity by going to the finest level of detail irrespective of application needs).

7. Ontology integration

Ontologies can be distinguished according to choice of axioms, which reflect those highly general background beliefs that are taken for granted by those working in the corresponding field. They can also be distinguished by the level of detail in the terms and definitions used and by the choice of subclassifying criterion. All these decisions should be stated explicitly. Important, too, is choice of *domain* – which can extend from single cell to whole populations of organisms – and of *granularity* (from molecule to whole organism).

Given these distinctions, the goal of constructing one comprehensive ontology for the life sciences begins to seem like an unreachable goal. Many groups have thus concluded that they must rest content with several smaller task-oriented ontologies, although the question is still being extensively debated in the bio-ontologies community (Bio-Ontologies Workshop, 2004). However, the approach of building smaller ontologies must eventually come to terms with the goal of combining ontologies together, for example, via techniques for ontology integration of the sort outlined in Ceusters *et al*. (2004). Such integration is by no means a simple matter, for given the heterogeneity of the domain ontologies contained in a system like the UMLS (Lindberg, 1990; *see also* Article 81, **Unified Medical Language System and associated vocabularies**, Volume 8), the relevant integrating steps can hardly be carried out automatically. Each concept must be located and identified in the various subdomain ontologies on the basis of manual search and comparison of definitions, and decisions must be made whether concepts are similar enough to be merged or if several similar concepts need to be defined and cross-related. The corresponding concepts must then be added to a new ontology that will incorporate all subdomain terms within a single consistent framework.

In the special case where the top-level terms of one ontology exactly match with those of another ontology, the corresponding branches can be merged. However, in this case, the data format (syntax, representation formalism) and the relations between terms of the two ontologies still need to be verified and if necessary, manually cross-calibrated.

Since this process of manual ontology integration is quite cumbersome, it might be more sensible to start of with an ontology that has a rather general upper level and can accommodate all of the diverse ontological types that are to be expected from the application domain. This was exactly the motivation for starting the MBO ontology described in Schulze-Kremer (1997).

8. Applications of bio-ontologies

Ontologies can provide computer programs with a counterpart of much of the commonsense background knowledge that human experts bring to bear in processing information. The range of applicability of ontologies is thus rather broad, and two examples, database integration and data annotation, will be discussed here briefly.

Data annotation is the process of linking data records, for example, in a gene product database to other knowledge resources, for example pertaining to cellular locations, in a process comparable to the process of indexing or cataloging books or other literature items. It is not a full-fledged ontology as described above that is required for this purpose, but rather only a controlled vocabulary, whose main purpose is to provide a fixed and unambiguous terminology for communication of research results. A controlled vocabulary of this sort is developed in the Gene Ontology (GO) project (Ashburner *et al*., 2000; *see also* Article 82, **The Gene Ontology project**, Volume 8), which attempts to ensure consistency in gene product annotations by means of the so-called GO identifiers (GO ID). This means that new concepts get new GO IDs, old concepts keep their GO IDs even if they are moved to another location within the hierarchy, and GO IDs of deleted concepts are not reused. As an ontology, GO has a rather simple, informal structure, which rests on the use of only two kinds of links: IS-A and PART-OF.

The GO approach has brought considerable benefits:

1. Work on populating GO could start immediately, without its authors needing to solve some of the intricate problems that face ontologies formalized as logical theories.
2. Extending GO does not require the completion of complex protocols of formally determined steps but can be done intuitively by the expert biologist.
3. There are few formal constraints standing in the way of easy incorporation of existing biological terms into the GO vocabulary.
4. The principle of unique identifiers allows GO terms to be used for database annotation without the consideration of their place in the GO hierarchy.

Focusing on the rapid population of GO has, however, a number of drawbacks (Smith *et al*., 2004):

1. It is unclear what kinds of reasoning are permissible on the basis of GO's hierarchies.
2. The rationale of GO's subclassifications is unclear. The reasoning that went into current choices has not been preserved and thus cannot be explained to or reexamined by a third party.
3. No procedures are offered by which GO can be validated.
4. There are insufficient rules for determining how to recognize whether a given concept is or is not present in GO. The use of a mere string search presupposes that all concepts already have a single standardized representation, which is not the case.

9. Open biomedical ontologies

GO is a part of the Open Biomedical Ontologies (OBO) project (**http://sourceforge.net/projects/obo**), which offers a framework for the development of well-structured controlled vocabularies for shared use across different biological domains. Contributions to OBO obey the following guidelines:

1. The ontologies must be open source, which means that they may be used by all without any constraints other than that their origin is acknowledged and they not be redistributed in altered form. OBO ontologies are intended to be resources for the entire biological community.
2. The ontologies employ, or can be instantiated in, or can be easily converted into, a common shared syntax. This may be either the GO syntax, extensions of this syntax, or OWL. This criterion is not met in all of the OBO ontologies currently listed.
3. The ontologies are orthogonal to other ontologies already lodged within OBO. Thus, different ontologies, for example, ontologies for anatomy and process, can be combined through additional relationships, and the latter can then be used to constrain when terms from different ontologies can be jointly applied to describe one and the same biological entity from distinct perspectives.
4. The ontologies share a unique identifier space. The source of concepts from any ontology can be immediately identified by the prefix of the identifier of each concept. It is, therefore, important that this prefix be unique.
5. The ontologies include textual definitions of their terms. Many biological terms are ambiguous; thus, each term should be defined in such a way that its precise meaning within the context of a particular ontology is clear to a human user.

10. Resources on (bio-)ontologies

The following include information relevant to work on bio-ontologies.

- Protégé 2000, an ontology editing software from Stanford Medical Informatics (*see* Article 91, **Frame-based systems: Protégé**, Volume 8), is at **http://smi.stanford.edu/projects/protege**.

- GKB Editor, the Generic Knowledge Base Editor of Peter Karp and SRI can be found at **http://www.ai.sri.com/~gkb**.
- OilEd, a simple ontology editor, resides at **http://www.ontoknowledge.org/oil/tool.shtml**.
- The Semantic Web Community Portal at **http://www.semanticweb.org** has lots of ontology-related information and pointers.
- Ongoing KBS/Ontology Projects and Groups are listed at **http://www.cs.utexas.edu/users/mfkb/related.html**.
- On-To-Knowledge: Content-driven Knowledge-Management through Evolving Ontologies is a European funded research project at **http://www.ontoknowledge.org**.
- The previous Bio-ontologies Workshops and other material on ontologies is compiled by Robert Stevens at **http://img.cs.man.ac.uk/stevens**.
- Cycorp has its own webpage at **http://www.cyc.com**.
- Formal Ontology in Information Systems is an international conference series on ontologies with a webpage at **http://www.fois.org**.
- Barry Smith has an extensive collection of works on ontology development in general and biomedical ontologies in particular at **http://ontology.buffalo.edu/smith**.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS and Eppig JT (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Bairoch A (1993) The ENZYME data bank. *Nucleic Acids Research*, **21**, 3155–3156.

Barnes J (Ed.) (1984) *The Complete Works of Aristotle*, Vol. 2, Princeton University Press.

Benson DA, Boguski MS, Lipman DJ and Ostell J (1997) GenBank. *Nucleic Acids Research*, **25**, 1–6.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Shimanouchi OKT and Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**, 535–542.

Bio-Ontologies Workshop (2004) **http://bio-ontologies.man.ac.uk**.

Bittner T and Smith B (2003) A theory of granular partitions. In *Foundations of Geographic Information Science*, Duckham MFG, Worboys MF (Eds.), Taylor & Francis: London, pp. 117–151.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA and Causton HC (2001) Minimum information about a microarray experiment (MIAME): Toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

Ceusters W, Smith B and Matthew Fielding J (2004) LinkSuite: formally robust software tools for ontology-based data and information integration. *Proceedings of DILS 2004 Data Integration in the Life Sciences (Lecture Notes in Computer Science 2994)*, Springer: Berlin, pp. 79–94.

Fasman KH, Letovsky SI, Cottingham RW and Kingsbury DT (1996) Improvements to the GDB human genome data base. *Nucleic Acids Research*, **24**, 57–63.

Fernandez M, Gomez-Perez A and Juristo N (1997) METHONTOLOGY: from ontological arts towards ontological engineering. *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, Stanford, USA, March 1997, pp. 33–40.

Genesereth MR and Fikes RE (1992) *Knowledge Interchange Format Reference Manual, Version 3.0*, Technical Report Logic-92-1, Computer Science Department, Stanford University.

Grenon P and Smith B (2004) SNAP and SPAN: Prolegomenon to geodynamic ontology. *Spatial Cognition and Computation*, **4**(1), 69–103. **http://ontology.buffalo.edu/bfo/SNAP.pdf**, **http://ontology.buffalo.edu/bfo/SPAN.pdf**.

Grenon P, Smith B and Goldberg L (2003) Biodynamic ontology: applying BFO in the biomedical domain. In *Ontologies in Medicine: Proceedings of the Workshop on Medical Ontologies*, Rome, October 2003, Pisanelli DM (Ed.), IOS Press: Amsterdam, pp. 20–38.

Gruber TR (1993) A translation approach to portable ontology specification. *Knowledge Acquisition*, **5**(2), 199–220.

Guarino N (1997) Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In: *Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology. International Summer School SCIE-97*, Pazienza MT (Ed.), Springer, pp. 139–170.

Guarino N (1998) Some ontological principles for designing upper level lexical resources. *Proceedings of First International Conference on Language Resources and Evaluation*, Granada: Spain.

Husserl E (1970) *Logical Investigations*, 2 vols, translated by Findlay J, London: Roufledge and Kegan Paul.

Jackson B and Ceusters W (2002) A novel approach to semantic indexing combining ontology-based semantic weights and in-document concept co-occurrences. In *EFMI Workshop on Natural Language Processing in Biomedical Applications*, Baud R and Ruch P (Eds.), Cyprus, pp. 75–80.

Jacobson D and Anagnostopoulos A (1996) Internet resources for transgenic or targeted mutation research. *Trends in Genetics*, **12**, 117–118.

Kashyap V, Ramakrishnan C and Rindflesch TC (2003) Towards (Semi-)automatic generation of Bio-medical ontologies. *AMIA 2003 Annual Symposium on Biomedical and Health Informatics*.

Keen G, Burton J, Crowley D, Dickinson E, Espinosa-Lujan A, Franks E, Harger C, Manning M, March S and McLeod M (1996) The genome sequence database (GSDB): meeting the challenge of genomic sequencing. *Nucleic Acids Research*, **24**, 13–16.

Köhler J and Schulze-Kremer S (2002) The Semantic Metadatabase (SEMEDA): Ontology Based Integration of Federated Molecular Biological Data Sources. In *Silico Biology* 2, 0021.

Lindberg C (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. *Journal of American Medical Association*, **61**(5), 40–42.

Locke J (1975) In *An Essay Concerning Human Understanding (originally published 1690)*, Nidditch PH (Ed.), Oxford University Press: Oxford.

Maedche A and Staab S (2003) Ontology learning. In *Handbook on Ontologies in Information Systems*, Staab S and Studer R (Eds.), Springer, pp. 173–190.

Mahesh K and Nirenburg S (1995) A situated ontology for practical NLP. *Proceedings Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, August 19–20, Montreal, Canada.

Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A (2003) Ontology library WonderWeb deliverable D18, IST Project 2001-33052 WonderWeb, Laboratory For Applied Ontology - ISTC-CNR, Via Solteri, 38, 38100 Trento, Italy **http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf**.

McKusick VA (1994) Mendelian inheritance in man. *Catalogs of Human Genes and Genetic Disorders*, Eleventh Edition, Johns Hopkins University Press: Baltimore.

Michael J, Mejino JLV and Rosse C (2001) The role of definitions in biomedical concept representation. *Proceedings, American Medical Informatics Association Fall Symposium*, Washington DC, pp. 463–467.

Nirenburg S and Raskin V (2001) Ontological semantics, formal ontology, and ambiguity. In *FOIS 2001 (Formal Ontology and Information Systems)*, Guarino N (Ed.), ACM Press, pp. 151–161.

Rector AL, Rogers JE, Zanstra PE and Van Der Haring E (2003) OpenGALEN: Open Source Medical Terminology and Tools. *Proceedings of AMIA Symposium*, p. 982.

Rosse C and Mejino JLV (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, **36**, 478–500.

Schulze-Kremer S (1997) Adding semantics to genome databases: towards an ontology for molecular biology. *Proceedings International Conference on Intelligent Systems, Molecular Biology*, **5**, 272–275.

Schulze-Kremer S (1998) Ontologies for molecular biology. *Pacific Symposium Biocomputing*, **3**, 693–704.

Smith, B and Rosse C (2004) The role of foundational relations in the alignment of biomedical ontologies. *Proceedings of MedInfo 2004*, San Francisco.

Smith B, Köhler J and Kumar A (2004) On the application of formal principles to life science data: a case study in the gene ontology. *Proceedings of DILS 2004: Data Integration in the Life Sciences (Lecture Notes in Computer Science 2994)*, Springer: Berlin, pp. 124–139.

The Joint United States / European Union ad hoc Agent Markup Language Committee (2001) **http://www.daml.org/**.

Specialist Review

Unified Medical Language System and associated vocabularies

Christopher G. Chute
Mayo Clinic College of Medicine, Rochester, MN, USA

1. Introduction

1.1. Terminologies, vocabularies, and ontologies

Biology and medicine, like most sciences, generate new knowledge and understanding by careful and systematic analyses of observations, outcomes, and experimental evidence. In the case of clinical medicine, these "experiments" are often cast in the form of epidemiologic outcomes research, which comprises a kind of "natural experiment". However, to represent biological and clinical observations in a fashion that can sustain analyses, these observations must conform to a system of weights and measures suitable to inferencing. Unlike absent metrics akin to meters, seconds, and grams, biology and medicine invoke well-controlled terminologies to provide consistent and comparable representations of biological functions, abnormal functions, and genomic variation. Clinical medicine in particular extends these characterizations to include pathologies, disease states, and syndromes.

The representation of biomedical data in the form of comparable and consistent concepts poses all the logistical and philosophical problems confronted by the knowledge representation community within the artificial intelligence specialty of computer science. It should not be surprising, then, that the sophistication of biomedical terminologies ranges from enumeration of concepts intended to populate "pick lists" for data entry to highly formal, description-logic-based ontologies. Indeed, within the ontology and knowledge representation communities, practical examples of sophisticated ontologies used by a broader community are most frequently found within biology and medicine.

The terms vocabulary, terminology, and nomenclature are often used interchangeably. While technical differences discriminate these terms, the more relevant distinction is between these three terms and ontology. Ontology has evolved to mean a catalog of well-formed, internally consistent, and explicitly interrelated concepts that cover a specific domain, such as anatomy or gene functions. Virtually all formal ontologies today are constructed using description logics, a family of first-order predicate logic subsets that are computationally tractable using reasoning algorithms. These description logics are used not only to define concepts in

terms of other concepts and their relationships, such as "part of" or "kind of," but also to assert knowledge-class relationships, such as

"Myocardial infarction" has_location "wall of the heart".

The fundamental dependence of biology and medicine on well-controlled terminologies, simple or complex, has caused a proliferation of terminology efforts. These terminologies have greater or lesser degrees of overlap, redundancy, consistency, and rigor. Not infrequently a researcher may wish to invoke a preexisting "catalog of terms" or even a more formal ontology in the conduct of large-scale experiments, clinical protocols, or randomized trials. Any resource that can transform a search for an appropriate terminology or ontology from a hit-or-miss exercise into an efficient and effective process would obviously be welcome. The Unified Medical Language System (UMLS) of the National Library of Medicine (NLM) was designed and has evolved to be the resource of choice in finding and working with biomedical terminologies.

2. History and origins of the UMLS

In 1984, a collaborative project among academic contractors to the National Library of Medicine began to consider how to contain and interrelate major biomedical terminologies. Mark Tuttle, a protégé of the late Scott Blois at the University of California, San Francisco, is widely credited with promoting the entity-relationship design of ASCII data tables that became ultimately adopted as the first information model for the industry in 1988. The first version of the Metathesaurus contained a few hundred thousand English language terms drawn from little more than a dozen terminologies. By 2005, the UMLS contained more than 1 million concepts and 5.6 million unique concept names from more than 100 different source vocabularies with representations in 16 languages.

The original Metathesaurus, the core collection of terminologies within the UMLS, was often referred to as the "Rosetta stone" of biology and medicine. An early goal of UMLS was to facilitate mapping of concepts between and among the many terminologies, such as the MeSH (Medical subjects heading) codes used in PubMed and the ICD (International Classification of Disease) codes used throughout clinical medicine. Over time, this ambitious goal was replaced by the UMLS providing the basis for medical informatics research about terminologies, a primary source for ontological relationships among concepts, and a major reference source for natural language processing, with a focus on identifying biomedical concepts and text.

Today, the NLM publishes that "The purpose of . . . [the] UMLS is to facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health". While this is a broad and abstract goal, the NLM has enjoyed substantial success in having the UMLS broadly adopted throughout the medical informatics research community as a basis for medical concept and natural language processing research. Increasingly, the basic biology, electronic medical record, and clinical systems communities are recognizing the substantial inherent value that the UMLS has accumulated.

2.1. Methods and formats

In the nearly two decades that the UMLS has been published, its presentation, data formats, information model, and syntax have undergone numerous changes. Ranging from experiments with unwieldy and unscalable ASCII flat files of fully denormalized data to the elegant but obscure ASN.1 (Abstract Syntax Notation) formatting, the UMLS has always had as its core a family of vertical-bar (|) delimited ASCII tables of terms, concepts, and relationships.

In 2004, the NLM introduced a dramatic overhaul of UMLS representation called the Rich Release Format (RRF). The RRF is intended to capture the more complex annotations and relationships present in sophisticated ontologies, specifically SNOMED-CT, which had become available in the United States under an NLM-brokered national site license. The primary motivation for introducing the RRF syntax was to achieve "source transparency" between a publisher's version of an ontology and that represented within the UMLS. This implies that an extraction of a terminology from the UMLS should be computationally identical to that distributed by the publisher. NLM sought, in their words, "to represent the detailed semantics of each source vocabulary exactly". The NLM and the College of American Pathologists (publisher of SNOMED-CT) concur that the UMLS has achieved source transparency for SNOMED-CT.

3. UMLS resources

The NLM's UMLS represents a suite of resources, ranging from a core catalog of concepts drawn from many terminologies (the Metathesaurus) to highly specialized natural language-processing tools within the Specialist Lexicon component. The major components are outlined here, though the number and persistence of minor components have varied over the lifetime of the UMLS project. Many references to the UMLS confuse the entire suite of resources with the Metathesaurus. This confusion is likely to persist, since many authors continue to refer to the Metathesaurus as the UMLS – implying that the Metathesaurus is the entirety.

3.1. Metathesaurus

The Metathesaurus of the UMLS has remained the core component of the system since its inception. Comprising a common representation of more than 100 terminologies, the Metathesaurus has emerged as an invaluable resource to anyone with a need for biomedical ontologies. The Metathesaurus can be best understood by outlining how concepts are represented in a common format, and then how relationships and attributes about these concepts are represented.

It is important to realize that with few exceptions, concepts do not exist in the Metathesaurus unless they are imported from one of the source terminologies. As such, the Metathesaurus does not comprise an ontology created and maintained by the NLM, but rather a synthesis of many code systems, terminologies, nomenclatures, and ontologies in a common format. Furthermore, the NLM

faithfully represents the semantic hierarchies and relationships contained in these source terminologies; such fidelity, not infrequently, results in ambiguous and occasionally contradictory assertions about concepts. As a practical matter, the NLM has established good relationships with the editors and maintainers of the major source vocabularies. Over time, such relationships have led to far fewer contradictions and less ambiguity. However, the Metathesaurus is a general-purpose resource, which typically must be adapted or constrained for specific applications.

3.1.1. Concepts

The representation of concepts in the Metathesaurus centers around several related but purpose-specific identifiers:

Concept unique identifier (*CUI*): The Metathesaurus defines as its core semantic unit the notion of a concept. Many different terms can mean the same thing, and thus be regarded as the same concept. The UMLS enforces a rigorous definition of synonymy; however, synonymous concepts across different source terminologies may not have identical ancestries.

Lexical unique identifier (*LUI*): The Metathesaurus distinguishes terms with similar appearance from synonyms to a concept. The NLM collapses text strings that differ in inflection, plural form, capitalization, and occasionally word order into a common lexical unit. However, this collapse is based on syntactic similarity alone; specifically, terms that share a lexical form may be asserted to have different meanings (CUIs) in their source terminologies.

String unique identifier (*SUI*): The string identifiers in the Metathesaurus preserve literal differences in text strings as they occur in the source vocabularies. Thus, many SUIs will typically map to one LUI. SUIs, as with LUIs, may be mapped to more than one CUI (concept) owing to assertions about their meanings that may differ across source vocabularies.

Atom unique identifier (*AUI*): The atom identifiers are a new innovation within the RRF representation of the Metathesaurus and correspond to a single instance of a string within a source vocabulary. Identical strings, that is, the ones with an equivalent SUI, may exist within a single source because a string may play different contextual roles for which it would be assigned unique atom identifiers. Obviously, many atoms might exist across different terminologies with identical string (SUI) forms. However, atoms are assigned one and only one CUI, since an atom may have only one semantic interpretation within a given source.

3.1.2. Relationships

Relationships among entries in the Metathesaurus are of two fundamental types: (1) the hierarchical relationships typically of the "is_a" variety (though these can occasionally include partonomy or "part_of" relations) and (2) asserted relationships. Hierarchical relationships, including sibling relations, are explicitly represented in the Metathesaurus using relational (in the database sense) tables.

Finally, most relationships about entries within a terminology derive from the source terminology itself. Relationships across terminologies, with the exception of synonymy, are not uniformly asserted by the NLM, though many such relationships

do exist. However, a special class of relationships that captures information about well-curated "mappings" between concepts across vocabularies is emerging. An example would be the mapping of SNOMED-CT terms to the International Classification of Disease (ICD).

Relationship unique identifiers (RUI): All relationships in the Metathesaurus are assigned a unique identifier, in part to facilitate version management over time.

Attribute unique identifier (ATUI): An attribute about a Metathesaurus entry provides additional information about that entry. Attributes may apply to concepts, atoms, or relationships. They may include highly generalizable information about a concept, such as a source-asserted definition, or fairly parochial meta-data relevant only to a source vocabulary, such as the date an atom was last edited. Attributes form a large class of information about Metathesaurus entries and can include computable characteristics, such as Semantic type (drawn from the Semantic Network, below), or the human language in which a string has linguistic meaning.

3.1.3. Major source terminologies

The relevance and appropriateness of a source are invariably dependent upon the user. However, some generalizations can be drawn by looking across the spectrum of terminologies that comprise the Metathesaurus, particularly with respect to clinical versus basic science emphasis. The heritage of the UMLS arose from a dominantly clinical legacy, as measured by the nature of source terminologies throughout its history. However, in the recent past, this is changing with the inclusion of sources aligned with the interests of biologists and basic scientists, such as the Gene Ontology (GO) or the Foundational Model of Anatomy (not yet included in the Metathesaurus).

Medical subjects heading (MeSH): The core indexing vocabulary underpinning Index Medicus, Medline, and PubMed, MeSH has reflected the contents of the published literature in biology and medicine for decades. Sometimes underrated and often undervalued, MeSH is a well-formed ontology whose original structure and elegance predates description logics and even the relational data model. MeSH uniquely traverses the content areas of cutting-edge science and standard medical practice. MeSH contains about 22 500 concepts and is fully represented in the Metathesaurus in about 560 000 row entries.

International classification of disease, 9th revision, clinically modified (ICD-9-CM): The International Classification of Disease, revised by the World Health Organization (WHO) on a schedule of decades, has been extensively modified from its original mortality application for clinical description and morbidity in the United States. Its primary use is for insurance and Medicare reimbursement in the United States, and it has limitations associated with being a code for financial billing. Nevertheless, ICD-9-CM remains ubiquitous in clinical settings, the primary coding system for hospital-based disease profiling, and widely reused for disease registries and patient databases. Domain concept headings for diseases and surgical procedures are enumerated in the Metathesaurus and comprise about 20 000 row entries.

Systematized nomenclature of medicine – clinical terms (SNOMED-CT): SNOMED is a comprehensive vocabulary of diseases, pathologies, clinically relevant organisms, symptoms, and related terms, with a 40-year heritage of development by the College of American Pathologists. In 1999, SNOMED merged with the broadly based Clinical Terms Version 3 (formerly the Read Codes) of the UK National Health Service; the merged effort became SNOMED-CT. In the mid-1990s, SNOMED began to adopt description logic underpinnings, which have expanded and become more robust over time. While not yet a formal ontology in the computer science sense, it is widely regarded as the largest and most complete body of clinical concepts available. SNOMED contains approximately 360 000 concepts with 1 million associated terms and 1.5 million relationships.

Gene ontology (GO): The Gene Ontology is a compilation of concepts about molecular function, biological process, and cellular components created by biologists for biologists. It began as a consortium among the fly (Drosophila), worm (Saccharomyces), and mouse model communities in 1998, and is now a collaborative among 16 organizations coalescing around the Open Biomedical Ontology (OBO) projects. Significant efforts are in progress to evolve the Gene Ontology into a formal ontology. GO contains approximately 17 000 concepts among 22 000 terms.

National drug files – reference terminology (NDF-RT): The US Department of Veterans Affairs (VA) has maintained a comprehensive catalog of drugs and pharmaceuticals used throughout the VA hospital system. Beginning in 2001, efforts were made to formally link drugs and drug classes in an ontology structure. The NDF-RT data model comprises enumerations and relationships among active ingredients, drug components, clinical drugs, finished dose form, drug products, packaged drugs, and associated taxonomies. These taxonomies include mechanism of action, physiologic effect, therapeutic intent, drug class, chemical structure, pharmacokinetics, clinical indications, and dosage form. The terminology contains approximately 38 000 drugs descriptions and 4000 drug classes.

Orderable drugs, normal form (RxNorm): RxNorm is part of the collaboration among the Food and Drug Administration (FDA), the Department of Veterans Affairs, and the National Library of Medicine to work on a common drug model and suite of terminologies. The model is outlined in the NDF-RT description above; RxNorm comprises the clinical drugs, finished dose form, and drug products component of that model. The NLM maintains RxNorm, which presently contains nearly 140 000 terms about 110 000 distinguishable drug forms.

3.2. Semantic Network

The Semantic Network is the core component of UMLS, which enumerates the underpinning semantic types and type relationships of the Metathesaurus. Unlike the Metathesaurus, the Semantic Network is entirely authored and maintained by the NLM; it depends on no source terminologies for its contents. As a consequence, the Semantic Network is highly generalizable. Every concept from every source in the UMLS Metathesaurus has at least one semantic type drawn from the Semantic Network. These assignments are asserted by NLM content editors.

Compared to the Metathesaurus, the Semantic Network is quite small. It contains only 135 concepts or semantic types, which are interrelated with 54 different semantic relationships. However, the sophistication of this resource should not be estimated from its size. The inheritance hierarchy and interrelationships are highly specified. The assignment of semantic type to Metathesaurus concepts enable inferences about concepts from different source terminologies, by reference to inheritance and relationships in the semantic network.

Examples of semantic types within the network include physiologic function, disease or syndrome, and genetic function. Semantic relationships, which have their own hierarchy, include "affects", which in turn has its own children, for example, manages, treats, disrupts, complicates, interacts with, and prevents.

An enhanced feature of the Semantic Network is the explicit sanctioning of appropriate inheritances through a mechanism the NLM describes as "blocking". For example, the semantic type "mental process" has a theoretical ancestor in the form of "Plant". Since plants are not capable of mental processes, given the definition of those concepts in the semantic network, this theoretical inheritance is explicitly disallowed or blocked.

To the extent that the UMLS Metathesaurus can be considered to assert a cohesive ontology about biology and medicine, the root nodes of such a metaontology are the semantic types within the semantic network. Similarly, the ontological relationships among the hundreds of thousands of concepts in the Metathesaurus can be viewed at the metaconcept level by the relationships among semantic types asserted in the Semantic Network.

3.3. Specialist lexicon

The content-oriented resources of the UMLS, including the Metathesaurus in the semantic network, have enormous potential for application in data entry systems, electronic health records, research databases, and enterprise vocabulary systems. A classical barrier between the effective application of UMLS resources and the community of potential users is the human user interface problem. Accurately typing long and complex biomedical terms, or attempting to sort among thousands of items on a pick list is not an attractive prospect to most humans. The NLM recognizes that the effective access and use of UMLS knowledge sources is contingent upon underpinning tools and lexical resources that can effectively map human natural language to terminological concepts. The Specialist Lexicon is an English-language-centric effort to provide precisely such a resource.

The specialist lexicon is divided into two fundamental components: (1) a comprehensive lexicon of biomedical terms and their functional variance and (2) suites of open-source, Java-based tools for manipulating words, language, text, and spelling.

3.3.1. Lexicon

The comprehensive lexicon draws unique words and terms from the Metathesaurus, and extends this word list with linguistic, syntactic, and inflectional information about these words. For example, verbs are annotated as to whether or not they

conform to regular inflections, and if they are irregular, their conjugation is specified. Similar data corresponds to case forms for nouns. Additional examples include the modality of pluralization, ranging from Greco-Latin form, to variants on group, regular, fixed, uncountable, and metalinguistic forms.

The Specialist Lexicon is quite rich and sizable, given its focus on biomedical concepts and words. It includes about 470 000 unique words, 500 000 agreement and inflection entries, 310 000 inflection instances, 17 000 abbreviations and acronyms, 34 000 modifiers, 22 000 spelling variants, 33 000 word properties, 81 pronouns, 8000 normalizations, and 2000 trademarks.

3.3.2. Tools

As a component of the Specialist Lexicon, the NLM provides open-source tools that harness the richness and content of the word-oriented lexicons of the UMLS. These tools are widely used throughout the medical natural language processing community, and have become embedded in applications such as terminology servers and electronic health records.

Lexical variant generator (LVG): Lexical variant generator is the core tool for the application of Specialist Lexicon knowledge. The basic tool takes as an argument the kind of variance a user might seek to generate from an input phrase, such as a diagnostic description. Presently there are 26 "flow" parameters that can be specified when using LVG. They include such options as canonicalize, uninflect, tokenized, lowercase, proper name filter, remove plurals, remove genitive, process punctuation, transform to Unicode, spelling variants, synonymy, and sorting. These options can be used to inflate terms to all specified variants or reformat input into a standard normal form.

Normalize word form (Norm): Norm is a callable form of LVG that prespecifies a convenient and standard way of lexically normalizing biomedical terms. The current version strips possessives, punctuation, and stop words, makes all characters lowercase, converts words into their inflection base form using the Specialist Lexicon data tables, and alphabetically sorts all words in the string. A version of the norm program, LUInorm, is the basis on which the Metathesaurus defines lexical identity among term variants. The Norm entry point to LVG is by far the most frequently used application tool in the Specialist Lexicon by application developers, researchers, and those responsible for emerging standards for terminology services in the biomedical community.

3.4. Access, use, and intellectual property

3.4.1. Distribution and formats

Early versions of UMLS were distributed exclusively by CD-ROM. In recent years, physical distribution has required migration to higher-capacity DVD formats. However, most users today – once licensed – obtain UMLS resources via the Internet. The standard URL for the UMLS starting point is **http://www.nlm.nih.gov/research/umls/**.

The NLM provides a sophisticated installation tool and customizing resource called Metamorphosys, which allows users to specify language-specific filters such as English only, which terminologies a user wishes to include in a local installation, a rank order priority of precedence for terminology sources within a local installation, and a rich variety of default subsets and installation scripts. Users may also specify that detailed content extraneous to their application use be suppressed during installation. Metamorphosys is delivered as an open-source tool and precompiled to run in standard computing environments, including Linux, Windows, Mac, and Solaris.

3.4.2. Knowledge source service

The NLM provides and continuously improves a service layer interface to UMLS resources, including the Metathesaurus, Semantic Network, and Specialist Lexicon. The Knowledge Source Server (KSS) permits live browsing of UMLS contents with a highly configurable interface. Additionally, the KSS supports a broad spectrum of API (application programming interface) calls that return concepts, synonyms, conceptual children or ancestors, definitions, and normalized forms.

3.4.3. Licensing issues

Many of the sources within the UMLS are contributed under restrictive license. The NLM therefore requires all users of the UMLS to register and agree to the license terms of the corpus. As a practical matter, more than 65 percent of UMLS concepts in the Metathesaurus have no intellectual property restrictions for general use. The rest of the source terminologies are classified into four categories of license.

Category 1: This is a liberal license that merely prohibits users of the source terminology from producing derivative works based on the source terminology. Users may employ the terms and concepts in research for commercial applications.

Category 2: This license category restricts the use of UMLS-provided concepts that derive from a Category 2 license source from being used for business applications on a production basis. For example, if users wish to use such source information for diagnoses within a hospital electronic record, additional licensing from the terminology provider would be required. However, use of the source material for research or product development is permitted.

Category 3: This is the most restrictive license category, in which the use of UMLS source-provided concepts is restricted solely to internal use for terminological research or product development. All other uses require a licensing arrangement with the source provider.

Category 4: This is a special category of use created to accommodate the circumstances for the US site license of SNOMED-CT.

Specialist Review

The Gene Ontology project

Midori A. Harris, Jane Lomax, Amelia Ireland and Jennifer I. Clark

The European Bioinformatics Institute, Cambridge, UK

1. Introduction

As noted in the introductory chapter (*see* Article 79, **Introduction to ontologies in biomedicine: from powertools to assistants**, Volume 8), ontologies represent domains of knowledge by defining terms used within the domain as well as relationships between them. The organization of knowledge about a given topic into an ontology allows databases to standardize annotations of their contents, improves computational queries, and can support the construction of inference statements on the basis of available information.

In the case of genetic and genomic databases, ontologies can be used to rationalize the storage and querying of information about sequence features and functional attributes of gene products. For example, both biologists and bioinformaticians need to pose meaningful questions such as:

- What is the function (at the molecular level) of a gene product?
- In what larger process(es) does a gene product participate?
- What is the subcellular localization of a gene product?
- Which gene products in one species have the same function as a given gene product in a different species?

These questions, particularly the last, can only be answered easily if both species share a scheme for the functional classification of gene products, with the necessary links between the databases. (*See also* Article 29, ***In silico* approaches to functional analysis of proteins**, Volume 7 and Article 36, **Large-scale, classification-driven, rule-based functional annotation of proteins**, Volume 7 for more information on annotation of protein function.)

The Gene Ontology (GO) project provides a set of ontologies that address the need for consistent, computable descriptions of biological entities, such as genes and gene products, in different databases (Blake and Harris, 2003; The Gene Ontology Consortium, 2000, 2001, 2004). The project is the work of the Gene Ontology Consortium, which includes most prominent repositories for plant, animal, and microbial genome information as well as other, non-species-specific, sequence databases (a complete list of GO Consortium members is available on

the GO website, **http://www.geneontology.org/**). All of the materials produced by the GO project – including the ontologies, gene product annotation sets, and software – are freely available to the public. Because the GO project continues to develop and is far from completion, the GO Consortium welcomes suggestions and contributions from anyone in the biology and bioinformatics communities.

2. The GO vocabularies

The ontologies developed by the GO Consortium cover four nonoverlapping domains of molecular biology. Three of the vocabularies describe attributes of gene products and the fourth describes sequence features.

The GO is designed to be applicable to all species, so each of the ontologies includes all terms falling into its domain, regardless of whether the biological attribute is restricted to certain taxonomic groups. Therefore, biological processes that occur only in plants (e.g., flower development) or mammals (e.g., lactation) are included. GO terms and relationships do not attempt to reflect the structure of gene products, nor represent evolutionary relationships.

Molecular function describes activities at the molecular level, such as catalysis or binding; molecular function terms do not specify where, when, or in what context an action takes place. Examples of individual molecular function terms are the broad concept "kinase activity" and the more specific "6-phosphofructokinase activity", which represents a subtype of kinase activity. A gene product may have a number of molecular functions; for example, myosin would be assigned the molecular function terms "actin binding", "ATPase activity", and "microfilament motor activity". This gene product would therefore be associated with all three molecular function terms. The annotation of gene products to more than one term in an ontology is supported by the GO system.

Within the context of GO, it is important to distinguish activities from the entities (molecules or complexes) that perform the actions. Molecular function terms represent activities, and should not be confused with gene products that have been named after activities they possess. For example, in many cases, a protein whose function is a catalytic activity has been named by that activity ("alcohol dehydrogenase activity" is a molecular function, with a corresponding GO term; "alcohol dehydrogenase" is also the name of a specific protein).

Biological process describes biological goals accomplished by ordered assemblies of molecular functions. High-level processes such as "cell death" can have both subtypes, such as "apoptosis", and subprocesses, such as "apoptotic chromosome condensation". Terms representing the regulation of biological processes are also included in this ontology. Note that a biological process in GO is not equivalent to a pathway; GO does not attempt to represent the dynamics or dependencies that would be required to describe a pathway (*see* Article 88, **Bioinformatics pathway representations, databases, and algorithms**, Volume 8).

Cellular component describes locations, at the levels of subcellular structures and macromolecular complexes. Examples of cellular components include "nuclear inner membrane", with the synonym "inner envelope", and the "ubiquitin ligase complex", which has several subtypes.

More recently, the Sequence Ontology (SO, **http://song.sourceforge.net/**) has been developed to permit the classification and standard representation of sequence features. Defined sequence features include terms such as “exon”, whose meaning is widely accepted, and the more problematic term “pseudogene”, for which several different usages have yet to be resolved. Although the SO is a relatively new vocabulary, and is still undergoing refinement, it is already being used for genome annotation projects in *Drosophila* and *Caenorhabditis elegans*.

3. Vocabulary structure

The ontologies in GO are structured as directed acyclic graphs (DAGs), wherein any term may have one or more parents and zero, one, or more children. Within each vocabulary, terms are defined, and parent–child relationships between terms are specified. A child term is a subset of its parent(s). The ontologies thus capture, for example, the fact that the nucleolus is part of the nucleus, which in turn is part of the (eukaryotic) cell; further, the DAG structure permits GO to represent “endoribonuclease activity” as a subcategory of both “endonuclease activity” and “ribonuclease activity”.

At present, the GO vocabularies define two semantic relationships between parent and child terms: is_a and part_of. In GO, an is_a relationship means that the term is a subclass of its parent. (Note that “is_a” should not be confused with “instance of”; the latter is a specific example of a member of the class; see Odell, 1998; Stevens *et al*., 2000.) For example, “terminal glycosylation” is a subclass of “protein glycosylation”. The is_a relationship is transitive, which means that if “GO term A” is a subclass of “GO term B” and “GO term B” is a subclass of “GO term C”, “GO term A” is also a subclass of “GO term C”. For example,

- glucose metabolism is a subclass of hexose metabolism;
- hexose metabolism is a subclass of monosaccharide metabolism;
- therefore, glucose metabolism is a subclass of monosaccharide metabolism.

The second relationship type in GO, part_of, is semantically complex as employed in the GO system. A discussion of meronomy, the study of the relationship between wholes and parts, is beyond the scope of this article (see Simons, 1987). That said, within GO, part_of may mean “physically part of” (as in the cellular component ontology) or “subprocess of” (as in the biological process ontology). Furthermore, the part_of relationship used in GO is “necessarily is_part”: wherever the child exists, it is as part of the parent. GO does not impose the “necessarily has_part” condition, which would require that wherever the parent exists, it has the child as a part. A GO term may be a part_of child of one parent and an is_a child of one or more additional parents.

The use of only two relationship types limits the expressive power of GO; the GO Consortium anticipates that in the future its ontologies will incorporate more relationship types, which will allow GO to be used more effectively for logical reasoning.

4. Gene product annotation

The GO vocabularies were originally developed to facilitate the functional annotation of gene products. Accordingly, collaborating databases provide data sets comprising links between database objects (which may represent genes, transcripts, proteins, or macromolecular complexes) and GO terms, with supporting documentation. Because a single gene may encode very different products with very different attributes, GO protocols recommend associating GO terms with database objects representing gene products rather than genes. At present, however, many participating databases are unable to associate GO terms to gene products, and therefore use genes instead. If the database object is a gene, it is associated with all GO terms applicable to any of its products, including RNAs and proteins, with all possible splice variants.

GO annotation conventions require that every annotation be attributed to a source. The source may be a literature reference, another database, or a computational analysis. Furthermore, the annotation must indicate the type of evidence the cited source provides to support the association between the gene product and the GO term. A standard set of evidence codes qualifies annotations with respect to different types of experimental determinations. The codes are available on-line (**http://www.geneontology.org/GO.evidence.html**), together with documentation that describes their use in functional annotations. For example, the "inferred from mutant phenotype" (IMP) evidence code indicates that the evidence supporting the association between a GO term and the gene product is derived from the examination of the phenotypic results of a genetic alteration.

Annotations may result from automated or curated operations. The most reliable annotations are those made manually by database curators; these are normally based on curatorial review of published literature and supported by experimental evidence. Curated annotations provide depth and accuracy; curators not only can apply any term from the ontologies, but can also request new terms as needed. Several evidence codes represent types of experimental data, such as "inferred from mutant phenotype" (IMP) or "inferred from direct assay" (IDA), and are extensively used in the manual annotation process. Manually curated annotation sets are now available for gene products in many model organisms (see below for link).

In addition to the curated annotation sets, large sets of electronic annotations made using computational methods have been generated both for model organisms and for many less experimentally tractable organisms, including human. A number of different automated methods have been applied (e.g., Camon *et al*., 2003; Mi *et al*., 2003; Pouliot *et al*., 2001; Xie *et al*., 2002). The evidence code "inferred from electronic annotation" (IEA) is used for all of these automated procedures. One example of this type of strategy is that based on the use of existing properties of the entries, such as the presence of keywords and Enzyme Commission (EC) numbers (Webb, 1992). Mapping of InterPro (Mulder *et al*., 2003) entries to GO terms (described below) also allows further associations of GO terms to entries to be made, on the basis of the presence of InterPro cross-references in other databases. Sequence similarity to well-studied gene products can also be used to generate annotations for large sets of genes or gene products.

A table summarizing annotation status for contributing databases is maintained on the GO website, from which annotation sets can also be downloaded (**http://www.geneontology.org/doc/GO.current.annotations.shtml**). Additional information on GO annotations can be found in references listed in the on-line GO Bibliography (**http://www.geneontology.org/GO.biblio.html**).

5. Other uses of GO

In addition to gene product annotation, GO has been put to a wide variety of uses, many of which were not envisaged when it was first established. One of the most common examples is the use of GO terms and gene product annotations to help interpret the results of large-scale experiments such as microarray studies (*see* Article 90, **Microarrays: an overview**, Volume 4), where GO annotations can reveal which broad biological processes are affected under the conditions tested (e.g., Bonner *et al*., 2003; Hawse *et al*., 2003; Meyer *et al*., 2004; additional references can be found in the on-line GO Bibliography at **http://www.geneontology.org/GO.biblio.html**). Within the biological and bioinformatics communities, other applications of the ontologies include gene function prediction (King *et al*., 2003), collaborative construction and analysis of cellular pathways (Demir *et al*., 2004), and association of genes to genetically inherited diseases (Perez-Iratxeta *et al*., 2002). GO terms have also been incorporated into the Unified Medical Language System (UMLS; *see* Article 81, **Unified Medical Language System and associated vocabularies**, Volume 8) maintained by the US National Library of Medicine (Lomax and McCray, 2004), and have been used to guide the development of a vocabulary for molecular imaging research (Tulipano *et al*., 2003).

GO is also attracting interest in the computer science community. For example, GO has been used as a test of applying description logic (*see* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8) approaches to building sound, complete, and logically consistent ontologies (Stevens *et al*., 2003; Wroe *et al*., 2003), and has featured in research into machine-processable ontologies (Williams and Andersen, 2003) and into the automated checking of ontological consistency (Yeh *et al*., 2003). GO terms also offer a valuable standardized terminological resource to natural language processing researchers, facilitating information extraction from texts, knowledge discovery, and ontology building from large collections of documents (*see* Article 84, **Ontologies for natural language processing**, Volume 8).

6. GO resources on the web

The GO website (**http://www.geneontology.org/**) is the comprehensive resource for users seeking information on the Gene Ontology. It provides the ontologies and annotations, and offers GO-related tools, a full suite of documentation, frequently asked questions (FAQ), mailing lists, and more. The Gene Ontology is a public resource and all data are freely available. The GO ontologies are available in a

number of formats, including the OBO flat file format, XML, and as a MySQL dump, all of which can be downloaded from the GO FTP archive or an anonymous CVS repository. Most of the database groups annotating gene products using GO release their files to the public, and these can be accessed from the website. Other useful resources include archived monthly GO releases, a set of GO slims (described below), and GO monthly reports, which detail what changes have been made to the ontologies. All are available from the Downloads section of the GO website (**http://www.geneontology.org/index.shtml#downloads**).

6.1. GO tools

As noted above, GO has been used for a wide variety of different purposes, and this is reflected in the diversity of tools available. Software is available to navigate, use, and manipulate the GO vocabularies and annotations, and for using GO for specific applications such as microarray analysis and text mining. Several tools are developed and maintained by the GO Consortium itself, of which two are of particular interest: the web-based GO browser AmiGO and the ontology editor DAG-Edit. A full list of tools is available on the GO website (**http://www.geneontology.org/GO.tools.html**).

6.2. DAG-Edit

DAG-Edit, shown in Figure 1, is a graphical tool developed to browse, query, and edit the GO. It can also be used to create and edit any ontology that has a DAG structure.

DAG-Edit is written in Java, and therefore runs on any platform that supports Java applications; several platform-specific installers are available. A user guide is included in the package and is also available on-line (**http://www.godatabase.org/dev/java/dagedit/docs/index.html**).

A separate package containing a selection of DAG-Edit plug-ins specific for GO, such as tools for managing obsolete terms and for viewing gene product annotations, is available. DAG-Edit is open source and can be downloaded freely (**http://sourceforge.net/project/showfiles.php?group_id=36855**).

6.3. AmiGO

The AmiGO tool (**http://www.godatabase.org/cgi-bin/go.cgi**), shown in Figure 2, provides an HTML interface for querying and browsing the ontologies, term definitions, and associated annotated gene products. AmiGO displays GO terms as an expandable tree view, with a summary view for displaying the list of gene products associated with each term. The results can be filtered by evidence code or by the organization that submitted the association.

Amino acid sequences are available for most gene products, and these can be selected and downloaded from AmiGO as FASTA files. The AmiGO feature GOst,

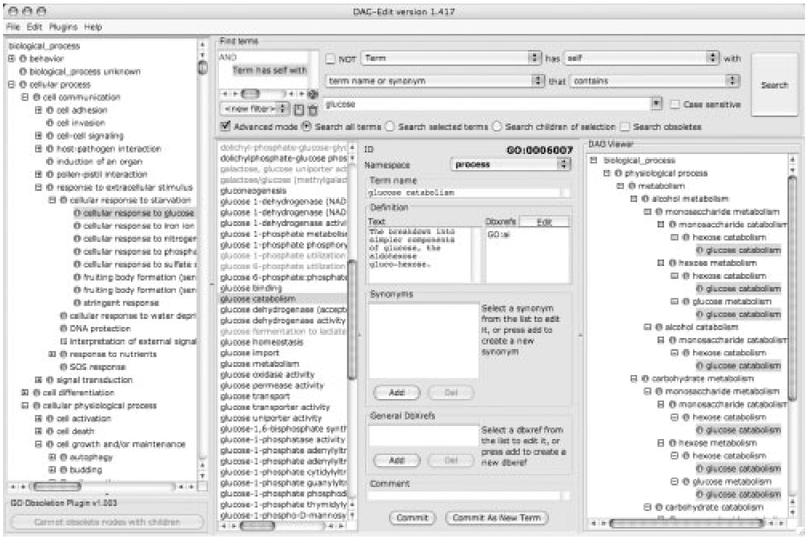


Figure 1 The open source ontology editor DAG-Edit, showing the results of a text search for “glucose”. Tree view, search result, term editing, and DAG view panes are visible

a GO BLAST server, allows users to submit query sequences and retrieve the sequences and GO annotations of all similar gene products in the GO database.

6.4. Documentation

The GO website provides documentation (**http://www.geneontology.org/GO.contents.doc.html**) on all aspects of the project, from the aims and scope of GO to practical advice for curators editing the ontologies or annotating gene products. Links to other resources such as the GO dictionary, GO cross-references, and the GO site at SourceForge are also in this section. GO provides mappings (**http://www.geneontology.org/GO.indices.html**) of its terms to those in other databases to allow preexisting annotations using other classification systems to be “translated” into GO terms. Mappings are maintained to a number of other databases, including the Enzyme Commission (Webb, 1992), MetaCyc (Krieger *et al*., 2004), UniProt Knowledgebase keywords (Apweiler *et al*., 2004), and InterPro entries (Mulder *et al*., 2003).

GO also offers four mailing lists on various aspects of the project, all of which are archived and can be searched. See the Contacting GO web page for details (**http://www.geneontology.org/GO.contacts.html**).

6.5. GO slims

The name “GO slim” refers to condensed versions of the ontologies that consist of a set of high-level GO terms (**ftp://ftp.geneontology.org/pub/go/GO_slims/**). To

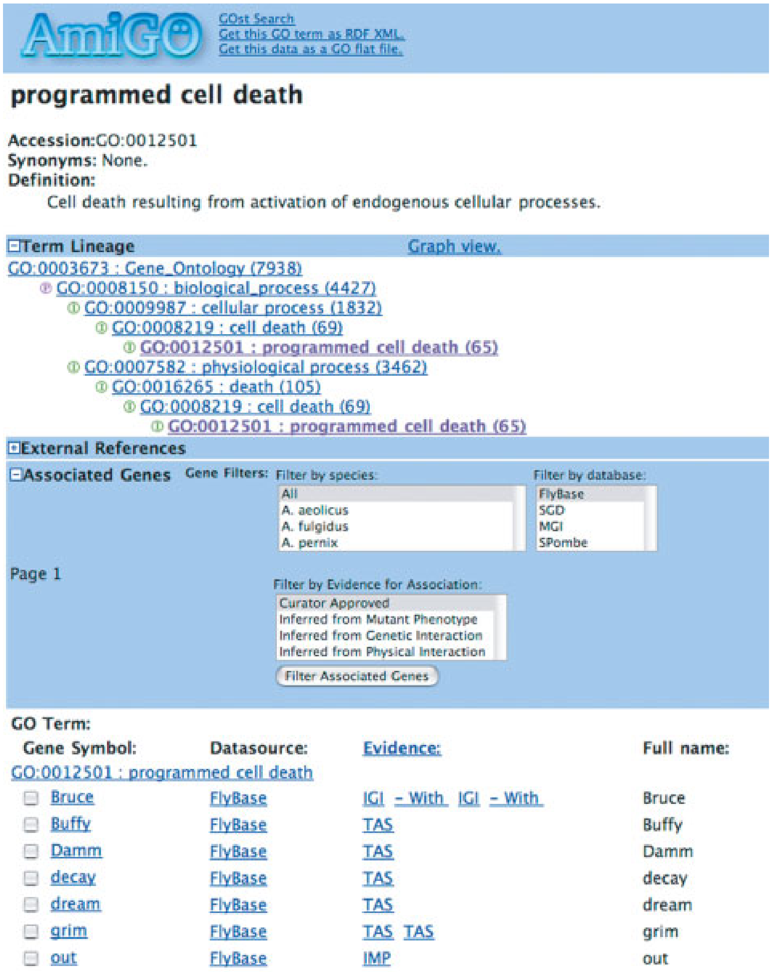


Figure 2 The AmiGO browser (**http: //www.godatabase.org/cgi-bin/go.cgi**), showing details of the GO term "programmed cell death"

gain an overview of the distribution of annotations or just as an introduction to the terms within GO, GO slims are a useful tool. Stored scripts can be used to map a set of GO annotations to the GO slim terms, making comparisons between and within datasets simple. Since GO slims have a number of possible applications, different GO slims have been created using terms relevant to a certain organism or database group. These specific GO slims, along with a generic version, are all stored on the GO web site.

7. OBO

GO is one of the ontologies held in the Open Biological Ontologies (OBO) collection (**http://obo.sourceforge.net/**). By providing controlled vocabularies that

are freely available to be downloaded and used, OBO aims to extend the principles that have guided GO development to cover many additional different biological domains.

There are currently over 40 ontologies lodged in OBO, covering domains such as anatomy, development, and phenotype, genomic, and proteomic information and taxonomic classification. The OBO website provides information on the OBO project in general, a description of each individual ontology, links to related projects, and a mailing list.

8. Conclusion

The Gene Ontology project provides and maintains structured, controlled vocabularies that cover specific domains of biology. An ever-increasing number of model organism and protein databases contribute GO annotations to a central repository. The ontologies and annotation sets are freely available for download from **http://www.geneontology.org/**. The GO web resource, which is continually being updated and expanded, provides documentation about the GO project and links to GO tools.

In the future, the GO project aims to improve consistency, both within the ontologies themselves and within and between gene product annotation sets. GO curators and developers meet regularly to focus on topics such as outstanding issues in representing biological content, or standards for annotations. Data representation improvements and new software features will also help improve the integrity of GO data. The GO Consortium will thus increase the utility of its ontologies even as they become more widely used by the biology, bioinformatics, and computer science communities.

Acknowledgments

This work represents the efforts of all members of the Gene Ontology Consortium. In particular, we thank Dr. Judith Blake for critical reading of the manuscript. The Gene Ontology Consortium is supported by NIH/NHGRI grant HG02273, and by grants from the European Union RTD Programme "Quality of Life and Management of Living Resources" [QLRI-CT-2001-00981 and QLRI-CT-2001-00015].

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, 115–119, **http://www.ebi.ac.uk/uniprot/**.

Blake JA and Harris MA (2003) The Gene Ontology project: structured vocabularies for molecular biology and their application to genome and expression analysis. In *Current Protocols in*

Bioinformatics, Baxevanis AD, Davison DB, Page R, Stormo G and Stein L (Eds.), John Wiley & Sons: New York.

Bonner AE, Lemon WJ and You M (2003) Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis. *Journal of Medical Genetics*, **40**, 408–417.

Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, *et al*. (2003) The Gene Ontology Annotation (GOA) Project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, **13**, 662–672.

Demir E, Babur O, Dogrusoz U, Gursoy A, Ayaz A, Gulesir G, Nisanci G and Cetin-Atalay R (2004) An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics*, **20**, 349–356.

Hawse JR, Hejtmancik JF, Huang Q, Sheets NL, Hosack DA, Lempicki RA, Horwitz J and Kantorow M (2003) Identification and functional clustering of global gene expression differences between human age-related cataract and clear lenses. *Molecular Vision*, **9**, 515–537.

King OD, Foulger RE, Dwight SS, White JV and Roth FP (2003) Predicting gene function from patterns of annotation. *Genome Research*, **13**, 896–904.

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY and Karp PD (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, **32**, D438–D442. **http://metacyc.org/**.

Lomax J and McCray A (2004) Mapping the Gene Ontology into the unified medical language system. *Comparative and Functional Genomics*. **5**(4), 354–361.

Meyer MH, Dulde E and Meyer RA Jr (2004) The genomic response of the mouse kidney to low phosphate diet is altered in X-linked hypophosphatemia. *Physiological Genomics*, **18**(1), 4–11, online publication ahead of print.

Mi H, Vandergriff J, Campbell M, Narechania A, Majoros W, Lewis S, Thomas PD and Ashburner M (2003) Assessment of genome-wide protein function classification for Drosophila melanogaster. *Genome Research*, **13**, 2118–2128.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, **31**, 315–318. **http://www.ebi.ac.uk/interpro/**.

Odell J (1998) Six different kinds of aggregation. *Advanced Object-oriented Analysis and Design Using UML*, Cambridge University Press: Cambridge, pp. 139–149.

Perez-Iratxeta C, Bork P and Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, **31**, 316–319.

Pouliot Y, Gao J, Su QJ, Liu GG and Ling XB (2001) DIAN: a novel algorithm for genome ontological classification. *Genome Research*, **11**, 1766–1779.

Simons P (1987) *Parts. A Study in Ontology*, Clarendon Press: Oxford.

Stevens R, Goble C and Bechhofer S (2000) Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, **1**, 398–416.

Stevens R, Wroe C, Bechhofer S, Lord P, Rector A and Goble C (2003) Building ontologies in DAML + OIL. *Comparative and Functional Genomics*, **4**, 133–141.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

The Gene Ontology Consortium (2001) Creating the Gene Ontology resource: design and implementation. *Genome Research*, **11**, 1425–1433.

The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D256–D261.

Tulipano PK, Millar W and Cimino JJ (2003) Linking molecular imaging terminology to the Gene Ontology (GO). *Pacific Symposium of Biocomputing*, **8**, 613–623.

Webb EC (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and*

Classification of Enzymes, Academic Press: San Diego. **http://www.chem.qmul.ac.uk/iubmb/enzyme/**.

Williams J and Andersen W (2003) Bringing ontology to the Gene Ontology. *Comparative and Functional Genomics*, **4**, 90–93.

Wroe CJ, Stevens R, Goble CA and Ashburner M (2003) A methodology to migrate the Gene Ontology to a description logic environment using DAML + OIL. *Pacific Symposium of Biocomputing*, **8**, 624–635.

Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A and Mintz L (2002) Large scale protein annotation through Gene Ontology. *Genome Research*, **12**, 785–794.

Yeh I, Karp PD, Noy NF and Altman RB (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, **19**, 241–248.

Specialist Review

Ontologies for information retrieval

William Richard Hersh
Oregon Health & Science University, Portland, OR, USA

1. Definition and overview of IR

To some, the term "information retrieval" may imply the retrieval of all types of information. But for those who work in the field and publish its literature, information retrieval (IR) implies the more focused aspect of document retrieval. However, even those in the field have come to take a broader view of IR with the growth of multimedia content and the World Wide Web. IR applications were among the earliest computer applications, with initial systems available in the 1950s and 1960s. This author has written a textbook on IR from the biomedical perspective (Hersh, 2003).

Nonetheless, IR concerns itself with the indexing and retrieval of information from heterogeneous and mostly textual information resources. The term itself was coined by Mooers in 1951, who advocated that it be applied to the "intellectual aspects" of the description of information and systems for its searching (Mooers, 1951). The advancement of computer technology, however, has altered the nature of IR. As recently as the 1970s, Lancaster (1978) stated that an IR system does not inform the user about a subject; it merely indicates the existence (or nonexistence) and whereabouts of documents related to an information request. At that time, of course, computers had considerably less power and storage than today's personal computers. As a result, systems were sufficient to handle only bibliographic databases, which contained just the title, source, and a few indexing terms for documents. Furthermore, the high cost of computer hardware and telecommunications usually made it prohibitively expensive for end users to directly access such systems, so they had to submit requests that were run in batches and returned hours to days later.

In the twenty-first century, however, the state of computers and IR systems is much different. End-user access to massive amounts of information in databases and on the World Wide Web is routine. Not only can IR databases contain the full text of resources but they may also contain images, sounds, and even video sequences. Indeed, there is continued development of the *digital library*, where books and journals are replaced by powerful file servers that allow high-resolution viewing and printing, and library buildings are augmented by far-reaching computer networks (Lesk, 1997; Arms, 2000; Witten and Bainbridge, 2003).

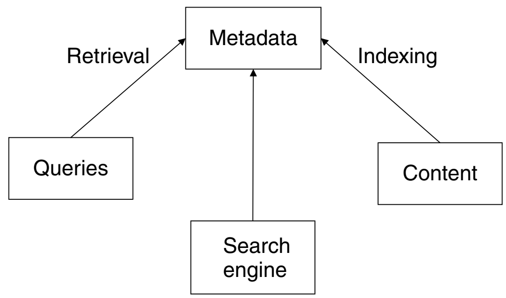


Figure 1 Components of the information retrieval system

One of the early motivations for IR systems was the ability to improve access to information. Noting that the work of early geneticist Gregor Mendel was undiscovered for nearly 30 years, Vannevar Bush started calling in the 1940s for science to create better means of accessing scientific information (Bush, 1967). In current times, there is equal if not more concern with “information overload” and how to find the more pertinent information within “data smog” (Shenk, 1998). However, important information is still missed. For example, a patient who died in a clinical trial in 2000 might have survived if information about the toxicity of the agent being studied from the 1950s (before the advent of MEDLINE) had been more readily accessible (McLellan, 2001). Indeed, a major challenge in IR is helping users find “what they don’t know” (Belkin, 2000).

There are a number of terms commonly used in IR. An *IR system* consists of *content*, computer hardware to store and access that content, and computer software to process user input in order to retrieve it. Collections of content go by a variety of terms, including *database*, *collection*, or – in modern Web parlance – *site*. In conventional database terminology, the items in a database are called *records*. In IR, however, records are also called *documents*, and an IR database may be called a *document database*. In modern parlance, items in a Web-based collection may also be called *pages*, and the collection of pages called a *website*.

A view of the IR system is depicted in Figure 1. The goal of the system is to enable the user to access the content. Content consists of units of information, which may themselves be an article, a section of a book, or a page on a website. Users seek content by the input of queries to the system. Content is retrieved by matching *metadata*, which is meta-information about the information in the content collection, common to the user’s query and the document. Metadata consist of both indexing *terms* and *attributes*. Indexing terms represent the subject of content, that is, what it is about. They vary from words that appear in the document to specialized terms assigned by professional indexers. Indexing attributes can likewise be diverse. They can, for example, consist of information about the type of a document or page (e.g., a journal article reporting a randomized controlled trial) or its source (e.g., citation to its location in a journal and/or its Web page address). A *search engine* is the software application that matches indexing terms and attributes from the query and document to return to the user.

There are two major intellectual or content-related processes in building and accessing IR systems: indexing and retrieval. *Indexing* is the process of assigning metadata to items in the database to facilitate and make their retrieval efficient. The term *indexing language* refers to the sum of possible terms that can be used in the indexing of terms. There may be, and typically are, more than one set of indexing terms, hence indexing languages, for a collection.

In most bibliographic databases, for example, there are usually two indexing procedures and languages. The first indexing procedure is the assignment of indexing terms from a *controlled vocabulary* or *thesaurus* by human indexers. In this case, the indexing language is the controlled vocabulary itself, which contains a list of terms that describe the important concepts in a subject domain. The controlled vocabulary may also contain nonsubject attributes, such as the document publication type. The second indexing procedure is the extraction of all words that occur (as identified by the computer) in the entire database. Although one tends typically not to think of word extraction as indexing, the words in each document can be viewed as descriptors of the document content, and the sum of all words that occur in all the documents is an indexing language.

Retrieval is the process of interaction with the IR system to obtain documents. The user approaches the system with an *information need*. Belkin *et al*. (1978) have described this as an *anomalous state of knowledge* (or ASK). The user (or a specialized intermediary) formulates the information need into a *query*, which most often consists of terms from one or more of the indexing vocabularies, sometimes (if supported by the system) connected by the Boolean operators AND, OR, or NOT. The query may also contain specified metadata attributes. The search engine then matches the query and returns documents or pages to the user.

Researchers in the IR field have a strong history of evaluating their systems using large-scale, widely available test collections, which usually consist of content (documents), statements of information needs (queries), and relevance judgments designating which documents are relevant for given queries. Many researchers participate in the Text Retrieval Conference (TREC, **trec.nist.gov**), an annual forum devoted to system evaluation. TREC is subdivided into various tracks representing common interests of researchers, such as question-answering and document filtering. While most tracks use common newswire and related content, one track that has been recently organized is the Genomics Track (**http://ir.ohsu.edu/genomics/**), which is devoted to evaluating IR systems using genomics documents (Hersh *et al*., 2004).

2. IR ontologies and thesauri

Many who are purists in their definition of ontologies would probably not describe the thesauri used for indexing in IR systems as ontologies. These resources are best described as controlled terminologies that describe the content likely to appear in the database (*see* Article 79, **Introduction to ontologies in biomedicine: from powertools to assistants**, Volume 8 and Article 87, **Merging and comparing ontologies**, Volume 8). However, thesauri serve an important role in indexing

content for retrieval, and they can also be the core of concepts that could be derived to generate a purer ontology.

The purpose of thesauri in IR is to provide a controlled set of terminology that both describes the content and provides concepts for user searching. Indexing of documents long preceded the computer age. The most famous early cataloger of medical documents was John Shaw Billings, who avidly pursued and cataloged medical reference works at the Library of the Surgeon General's Office (Miles, 1982). In 1879, he produced the first index to the medical literature, *Index Medicus*, which classified journal articles by topic. For over a century, *Index Medicus* was the predominant method for accessing the medical literature. By the middle of the twentieth century, the Medical Literature Access and Retrieval System (MEDLARS) was put into place to facilitate the cataloging and indexing of medical literature. This gave rise to MEDLARS on-line, or MEDLINE, the database of literature formerly accessed in *Index Medicus*.

Even though the medium has changed, the human side of indexing the medical literature has not. The main difference in the computer age is that a second type of indexing, automated indexing, has become available. Thus, most modern commercial content is indexed in two ways:

1. Manual indexing – where human indexers, usually using standardized terminology, assign indexing terms and attributes to documents, often following a specific protocol.
2. Automated indexing – where computers make the indexing assignments, usually limited to breaking out each word in the document (or part of the document) as an indexing term. Ongoing research is attempting to identify terms in standardized terminology for automated indexing (*see* Article 86, **Automatic concept identification in biomedical literature**, Volume 8 and Article 84, **Ontologies for natural language processing**, Volume 8)

Manual indexing has mostly been done with bibliographic databases. In the age of proliferating electronic databases, such as on-line textbooks, practice guidelines, and multimedia collections, manual indexing has become either too expensive or outright infeasible for the quantity and diversity of content now available. Thus, there are increasing numbers of databases that are indexed only by automated means.

3. Characteristics of IR ontologies/thesauri

Before discussing specific vocabularies, it is useful to define some terms, since different writers attach different definitions to the various components of thesauri. A *concept* is an idea or object that occurs in the world, such as the condition under which human blood pressure is elevated. A *term* is the actual string of one or more words that represent a concept, such as `Hypertension` or `High Blood Pressure`. One of these string forms is the preferred or *canonical* form, such as `Hypertension` in the present example. When one or more terms can represent a concept, the different terms are called *synonyms*.

A controlled vocabulary usually contains a list of certified terms that are the canonical representations of the concepts. Most thesauri also contain relationships between terms, which typically fall into three categories:

1. Hierarchical – terms that are broader or narrower. The hierarchical organization not only provides an overview of the structure of a thesaurus but can also be used to enhance searching.
2. Synonymous – terms that are synonyms, allowing the indexer or searcher to express a concept in different words.
3. Related – terms that are not synonymous or hierarchical but are somehow otherwise related. These usually remind the searcher of different but related terms that may enhance a search.

4. Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) vocabulary was created by the NLM for indexing *Index Medicus*; the MeSH vocabulary was and is now used to index most of the databases produced by the NLM (Coletti and Bleich, 2001). Other systems use MeSH to index their bibliographic content as well, such as the Web catalog CliniWeb (Hersh *et al*., 1996) and National Guidelines Clearinghouse (**www.guideline.gov**). The latest version of MeSH contains over 19 000 *headings* (the word MeSH is used for the canonical representation of its concepts). It also contains over 100 000 supplementary concept records in a separate chemical thesaurus. In addition, MeSH contains the three types of relationships described in the previous section:

1. Hierarchical – MeSH is organized hierarchically into 15 *trees* (see Table 1).
2. Synonymous – MeSH contains a vast number of *entry terms*, which are synonyms of the headings. The majority of these entry terms are “nonprint entry terms”, which are mostly variations of the headings and entry terms in plurality, word order, hyphenation, and apostrophes. There are a smaller number of “print entry terms”, which appear in the printed MeSH catalog. These are often called *see references* because they point the indexer or searcher back to the canonical form of the term.
3. Related – terms that may be useful for searchers to add to their searches when appropriate are suggested for many headings.

The MeSH vocabulary files, their associated data, and their supporting documentation are available on the NLM’s MeSH website (**www.nlm.nih.gov/mesh/**). There is also a browser that facilitates exploration of the vocabulary (**www.nlm.nih.gov/mesh/MBrowser.html**). A paper-based catalog with documentation is available as well (*Medical Subject Headings*, 2005).

Figure 2 shows a partially pruned version of some of the terms in hierarchical proximity to various genes.

There are additional features of MeSH designed to assist indexers in making documents more retrievable (*Features of the MeSH Vocabulary*, 2000). One of these

Table 1 The top-level trees of MeSH

Anatomy [A]
Organisms [B]
Diseases [C]
Chemicals and Drugs [D]
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
Psychiatry and Psychology [F]
Biological Sciences [G]
Physical Sciences [H]
Anthropology, Education, Sociology and Social Phenomena [I]
Technology and Food and Beverages [J]
Humanities [K]
Information Science [L]
Persons [M]
Health Care [N]
Geographic Locations [Z]

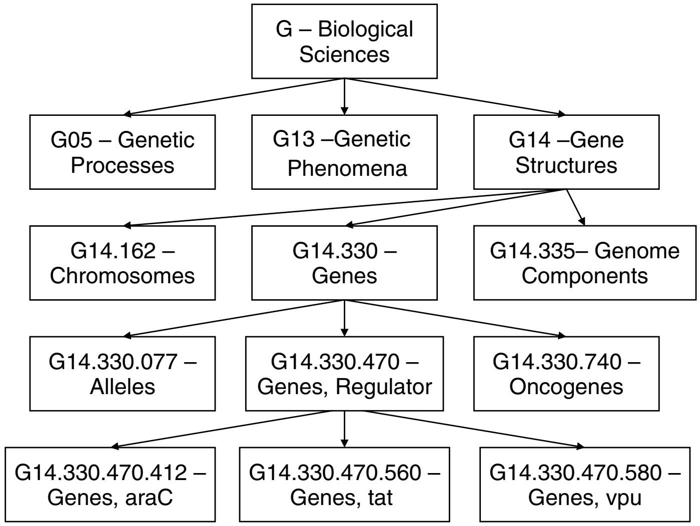


Figure 2 Partially pruned depiction of MeSH hierarchy

is *subheadings*, which are qualifiers to headings that can be attached to narrow the focus of a term. In the `Hypertension`, for example, the focus of an article may be on the diagnosis, epidemiology, or treatment of the condition. Assigning the appropriate subheading will designate the restricted focus of the article, potentially enhancing precision for the searcher. There are also rules for each tree restricting the attachment of certain subheadings. For example, the subheading `drug therapy` cannot be attached to an anatomic site, such as the `femur`.

Another feature of MeSH that helps retrieval is *check tags*. These are MeSH terms that represent certain facets of medical studies, such as age, gender, human or nonhuman, and type of grant support. They are called check tags because the indexer is required to use them when they describe an attribute of the study. For

example, all studies with human subjects must have the check tag `Human` assigned. Related to check tags are the geographical locations in the Z tree. Indexers must also include these, like check tags, since the location of a study (e.g., *Oregon*) must be indicated.

One feature that is gaining increasing importance for clinical searching is the *publication type*, which describes the type of publication or the type of study. A searcher who wants a review of a topic will choose the publication type `Review` or `Review Literature`. Or, to find studies that provide the best evidence for a therapy, the publication type `Meta-Analysis`, `Randomized Controlled Trial`, or `Controlled Clinical Trial` would be used.

While not necessarily being helpful to a searcher using MeSH, the *tree address* is an important component of the MeSH record. The tree address shows the position of a MeSH term relative to others. At each level, a term is given a unique number that becomes part of the tree address. All children terms of a higher level term will have the same tree address up to the address of the parent. It should be noted that a MeSH term can have more than one tree address. `Pneumonia`, for example, is a child term of both `Lung Diseases` (C08.381) and `Respiratory Tract Infections` (C08.730). It thus has two tree addresses, C08.381.677 and C08.730.610.

Another feature of MeSH is *related concepts*. Most well-designed thesauri used for IR have related terms, and MeSH is no exception. Related concepts are grouped into three types. The first is the *see related* references. The MeSH documentation provides examples of how concepts are related (*Features of the MeSH Vocabulary*, 2000). Examples include concepts related between a physiological process and a related disease (e.g., `Blood Pressure` and `Hypertension`) or between a physiological process and a drug acting on it (e.g., `Blood Coagulation` and `Anticoagulants`).

An additional type of related concept is the *consider also* reference, which is usually used for anatomical terms and indicates terms that are related linguistically (e.g., by having a common word stem). For example, the record for the term `Brain` suggests considering terms `Cerebr-` and `Encephal-`. A final category of related concepts consists of main heading/subheading combination notations. In these instances, unallowed heading/subheading combinations are referred to a preferred precoordinated heading. For example, instead of the combination `Accidents/Prevention & Control`, the heading `Accident Prevention` is suggested.

There are two other important features of MeSH terms. The first is the *Annotation*, which provides tips on the use of the term for searchers. For example, under `Congestive Heart Failure`, the searcher is instructed not to confuse the term with `Congestive Cardiomyopathy`, a related but distinctly different clinical syndrome. Likewise, under `Cryptococcus`, the searcher is reminded that this term represents the fungal organism, while the term `Cryptococcosis` should be used to designate diseases caused by Cryptococcus. The second feature is the *Scope Note*, which gives a definition for the term. These can be helpful to the searcher when the use of a term by indexers has changed over time, for example, the term was derived from a broader term used previously or had its name changed at some point in the past.

5. Other ontologies for IR

While MeSH is the most widely used ontology for biomedical IR, it is not the only one used for indexing biomedical documents. Another large biomedical bibliographic database is EMBASE, which is part of *Excerpta Medica*. EMBASE has a vocabulary called EMTREE, which has many features similar to those of MeSH (**www.elsevier.nl/homepage/sah/spd/site/locate_embase.html**). EMTREE is also hierarchically related, with all terms organized under 16 *facets*, which are similar but not identical to MeSH trees. Concepts can also be qualified by *link terms*, which are similar to MeSH subheadings. EMTREE includes synonyms for terms, which include the corresponding MeSH term.

A growing number of resources are linking annotations of gene function in the form of GO codes to the biomedical literature (*see* Article 82, **The Gene Ontology project**, Volume 8). For example, the Mouse Genome Informatics (MGI, **www.informatics.jax.org**) resource links gene names and GO annotations to references in the literature classified by the type of evidence provided (Bult *et al*., 2004). Likewise, the NLM Entrez Gene resource (**http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene**) which catalogs literature, data, and other information about genes, provides GO annotations with linkages to the literature (Maglott *et al*., 2005). Indeed, the growing number of model organism databases and other linked resources will ensure that the ontologies described in other chapters will provide access, even if only directly, to literature and other content present in IR systems.

No discussion of ontologies for biomedical IR would be complete without the mention of the Unified Medical Language System (UMLS) (Humphreys *et al*., 1998; *see also* Article 81, **Unified Medical Language System and associated vocabularies**, Volume 8). Although described in another chapter, this resource is important for IR. The UMLS Metathesaurus provides linkages across different controlled terminologies as well as tools to facilitate the use of IR systems. Indeed, as genome-related discoveries make their way into clinical medicine, the linkage to clinical findings (SNOMED), tests (LOINC), and diagnoses (ICD-9) will become more important.

6. Future directions for ontologies in IR

The growing amount of data from sequencing, microarray, and other experimental modalities provide a case for the use of IR systems. As knowledge is gained and biological processes annotated, the importance of rich ontologies will continue to grow. This situation will motivate both the understanding and improvement of IR systems.

References

Features of the MeSH Vocabulary (2000) National Library of Medicine. **http://www.nlm.nih.gov/mesh/features.html**, accessed, July 1, 2002.

Medical Subject Headings (2005) Government Publishing Office, Washington, DC.

Arms W (2000) *Digital Libraries*, MIT Press: Cambridge.

Belkin N (2000) Helping people find what they don't know. *Communications of the ACM*, **43**, 58–61.

Belkin N, Oddy R and Brooks H (1978) ASK for information retrieval: I. Background and theory. *Journal of Documentation*, **38**, 61–71.

Bult C, Blake J, Richardson JE, Kadin JA, Eppig JT, Baldarelli RM, Barsanti K, Baya M, Beal JS and Boddy WJ (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Research*, **32**, D476–D481.

Bush V (1967) *Science is Not Enough*, Morrow: New York.

Coletti M and Bleich H (2001) Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, **8**, 317–323.

Hersh W (2003) *Information Retrieval: A Health and Biomedical Perspective*, *Second Edition*, Springer-Verlag: New York.

Hersh W and Bhuptiraju RT, Ross L, Johnson P, Cohen AM and Kraemer D (2004) TREC 2004 genomics track overview. The Thirteenth Text Retrieval Conference: TREC 2004, Gaithersburg, MD. National Institute of Standards and Technology. **http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf**.

Hersh W, Brown K, Donohoe LC, Campbell EM and Horacek AE (1996) CliniWeb: managing clinical information on the World Wide Web. *Journal of the American Medical Informatics Association*, **3**, 273–280.

Humphreys B, Lindberg D, Schoolman H and Barnett G (1998) The Unified Medical Language System: an informatics research collaboration. *Journal of the American Medical Informatics Association*, **5**, 1–11.

Lancaster F (1978) *Information Retrieval Systems: Characteristics, Testing, and Evaluation*, Wiley: New York.

Lesk M (1997) *Practical Digital Libraries – Books, Bytes, & Bucks*, Morgan Kaufmann: San Francisco.

Maglott D, Ostell J, Pruitt KD and Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**, D54–D58.

McLellan F (2001) 1966 and all that – when is a literature search done? *Lancet*, **358**, 646.

Miles W (1982) *A History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge*, U.S. Department of Health and Human Services: Bethesda.

Mooers C (1951) Zatocoding applied to mechanical organisation of knowledge. *American Documentation*, **2**, 20–32.

Shenk D (1998) *Data Smog: Surviving the Information Glut*, Harper Collins: San Francisco.

Witten I and Bainbridge D (2003) *How to Build a Digital Library*, Morgan Kaufmann: San Francisco.

Specialist Review

Ontologies for natural language processing

Yves A. Lussier
Columbia University, New York, NY, USA

1. Introduction

Rapid technological improvements of biomedical computational semantics and natural language processing (NLP) are leading to a profound transformation in the reuse of scientific journal articles in the form of curated and highly computational knowledge stored in specialized biomedical databases, which are proliferating and increasingly structured with ontologies (e.g., Gene Ontology). New computational semantic methods use these ontologies as intermediaries to bridge the gap among research communities and their databases by providing cross-disciplinary and high-throughput results. Moreover, technological standardization of *declarative knowledge*, a highly computable data structure using logic, and communication standards (e.g., XML, RDF, OWL) have profoundly accelerated the development cycles in computational semantics, resulting in ontology-anchored databases that can be automatically transformed. However, scientific biomedical literature remains the pinnacle of knowledge in breadth and depth. Though current generation of *literary knowledge* trumps declarative knowledge in quantity, its noticeable value is compromised at a *retrieval cost*, owing to literary knowledge being buried in an overwhelming growth of scientific articles. Similarly, data entry of biomedical databases derived from scientific journals, generally accomplished via manual annotation, is a rate-limiting and costly process. Automated data structures derived from NLP portend instantaneous and comprehensive *linguistic knowledge* across boundless scientific articles and research communities. However, while linguistic knowledge is computable, it does not allow for the same quality of inference as declarative knowledge. Thus, increasing efforts have been invested to translate linguistic data structures generated by NLP into ontology-anchored declarative data sets to obtain otherwise unattainable large-scale or cross-disciplinary inferences. As the gap between linguistic and declarative knowledge is bridged, we are witnessing a paradigm shift. Biomedical knowledge amassed in unprecedented volume is likely to enable large-scale computational analysis that, in turn, could facilitate reductionist and synthetic hypotheses.

Our working definition of the term “ontology” will encompass formal logical terminological systems of logic erected on axioms (Sowa, 2000) with terminologies

based on semiexplicit subsumptions from which inferences can be derived (Guarino, 1998).

This article focuses on the convergence of biomedical ontologies, originally developed to structure databases, and linguistic data structures produced by NLP operating on unstructured or semistructured corpora. Theories upon which the convergence of NLP, biomedical informatics and ontologies are being conducted will first be addressed in Section 1, followed by a description of the properties of ontologies that make them suitable for NLP in Section 3. Section 4 will follow with concrete examples of ontologies addressing representation of biological knowledge and a succinct analysis of their readiness for use by NLP systems. Section 5 will provide examples of NLP systems mapping to ontologies. In conclusion, Section 6 will summarize past accomplishments and their limitations, and point out future directions.

2. Theoretical background

As biomedical ontologies, NLP systems and computational semantics converge, their respective theories and principles are bound to challenge or enrich one another. As described elsewhere, Aristotle's theories of syllogism, knowledge representation and first-order description logic are widely used in the field of computational semantics (Sowa, 2000). Here, we will discuss three additional contemporary theories, which are germane to the use of ontologies for NLP in biomedicine: (1) Ogden and Richards' semiotic triangle; (2) Marsden Blois' conjecture on information in biomedicine; and (3) the sublanguage theories of Zellig Harris.

2.1. Ogden-Richards' semiotic triangle

As shown in the bottom of Figure 1, a subject matter of an investigation or science (object) can be represented by a symbol or *term*; a term is a text string consisting of one or many words that together have a precise meaning. This theory provides a basis for defining symbol *ambiguity* and *redundancy*, which respectively refer to the clashing of related concepts sharing similar symbols superficially confounded as being the same object, and to unique objects having distinct and apparently unrelated symbol representations, leading to duplication of the concept in the metathesauri. While these principles are suitable for domains of science where relevant knowledge is relatively stable over time, they are sometimes oversimplified for representing the rapid evolving field of biomedicine, in which many discoveries are cross-disciplinary and likely to extend or modify the meaning of some terms. Concurrently, there has been a growth of specialized biomedical ontologies and metathesauri of ontologies, such as the Unified Medial Language System (UMLS) (Lindberg, 1990) or the National Cancer Institute (NCI) Metathesaurus (Covitz *et al*., 2003), that are aimed at bridging the gap between different standards. The proposed approach of systematically analyzing the three semiotic facets, "symbol, object, and concept", provided a model-theoretic foundation for evaluating metathesauri (Campbell *et al*., 1998) that could also be used to evaluate the

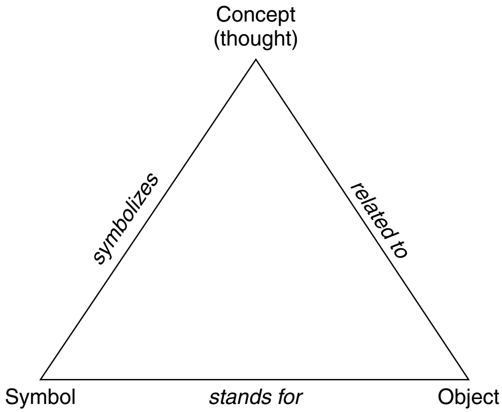


Figure 1 The adapted Ogden and Richards semiotic triangle

meanings of linguistic-semantic data structures composed by an NLP system where the original narrative is considered as the object or its proxy (Figure 1).

2.2. *Marsden Blois' conjectures on information in biomedicine*

Blois (1984) proposed a visionary conjecture of biomedical information processing derived from the Hartley-Shannon-Weaver view of information. He hypothesized that structures, functions, and processes observed at different scales of biology could be represented and assembled in a formal system by following a few simple rules. (1) At any given "descriptive level" (scale of biology), a "nominal" structure, function, or process can be described as a function of "attributes". (2) These nominals cannot be attributes at the same descriptive level. (3) Moving toward the macroscopic level, nominals at one descriptive level respectively become attributes, while attributes of a previous level eventually disappear and become "embedded" in nominals and attributes of higher levels. (4) "Emerging attributes come into existence with each increased level" and contribute to the behavior of nominals in all levels above. (5) Moving up the descriptive levels, nominals and attributes which are assumed to have crisp and unambiguous definitions (e.g., nucleotide sequences, etc.) and behaviors at molecular levels become increasingly fuzzy to describe and "particular" (used by Blois to address somewhat dissimilar phenotypes from one organism to another), thus giving rise to the complexity of presentation of nonmolecular phenotypes, diseases, and individuality. While some have rightly argued that this theory requires confirmation with falsifiable hypotheses, it is generally regarded as providing a foundation for the development of second-generation knowledge-based systems and reusable domain ontologies in biomedicine (Musen, 2002; Rector *et al*., 2002). Besides ontologies, Blois' theory can help postulate hypotheses on current manually annotated biological databases that span different scales of biology such as the molecular networks of genes and gene products. Indeed, Blois' theory is implicitly or explicitly used in biomedical object-oriented software and databases erected on foundational ontologies (e.g.,

Table 1 Content coverage of NLP and distinct ontologies according to the scale of biology and scientific field
legend: ● = biological scale covered, ○ = biological scale partially covered, "empty box" = biological scale not covered

Biological scale		Example of distinct structure types		Content coverage ontologies											NLP	
Size (m)	Weight (kg)	Count (H. Sapiens in Bold)	Structure type	Gene ontology[a]	Gene nomenclature[b]	ICTVdb[c]	NCBI taxonomy[d]	Cell ontology[e]	Mammalian phenotype[f]	ICD[g]	MeSH[h]	SNOMED[i]	UMLS[j]	NCI Metathesaurus[k]	Various systems	Scientific fields
10^{-10}	10^{-27}–10^{-26}	**10^1**	Atoms H,C,N,O													Biochemical Pathways Chemoinformatics, lead compound libraries[l]
10^{-9}	10^{-25}	**10^9**	Base pair sequence													
10^{-8}	10^{-24}–10^{-21}	**10^5**	cDNA, microRNA, sRNA, ncRNA													Transcriptome, hybridome, Peptidomics
	10^{-23}–10^{-22}	**10^5**	Proteins								○	○	○	○	●	Proteome, interactome, Structural Genomics, Gene product pathways
		10^3	Gene product functions	●							○	○	●	●	○	
	10^{-21}	**10^4**	Genes[m]		●										●	Genomics, PharmacoGenomics, Toxicogenomics,
10^{-8}–10^{-5}	10^{-19}–10^{-17}	10^4–10^5	Viruses[n,o]			●	○				○	○	●	●	○	Metabolome, Virology, Bacteriology, Host-Pathogen Interactions
10^{-7}–10^{-5}	10^{-15}	10^4–10^6	Bacteria				●				○	○	●	●	○	
		10^2	Organelles	●							○	○	●	●	○	Cellulome, Cellular Interactions, Cytogenetics, Cellular Markers,
10^{-5}	10^{-12}	10^4–10^6	Fungi				●				○	○	●	●	○	
		10^4	Cell lines[p,q]													In vitro, in vivo studies
		10^3	Cell morphologies						○		●	●	●	●	○	Histology, Pathology
		10^2	Cell types					●	○		○	●	●	●	○	
10^{-4}–10^0	10^{-11}–10^0	**10^3**	Tissues, morphologies						●		○	●	●	●	●	
		10^2	Organs						●		●	●	●	●	●	Clinical Genomics, Pharmacogenomics,
		10^1	Systems						●		●	●	●	●	●	

10^0	10^0	$\mathbf{10^6}$	Macroscopic organisms	●		○	○	○	●	●	○	Phylogenetic, Systematics
$>10^0$	$>10^0$	$\mathbf{10^5}$	Diseases, syndromes, populations		○	○	○	●	●	●	●	Medicine, Nursing, Public Health

[a] The Gene Ontology Consortium. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261. **http://www.geneontology.org/**

[b] Wain *et al*. (2002) Guidelines for human gene nomenclature. *Genomics*, **79**(4), 464–470.

[c] Buchen–Osmond (1997) Further progress in ICTVdB, a universal virus database. *Archives of Virology*, **142**(8), 1734–1739. PMID: 9672633. **http://www.ncbi.nlm.nih.gov/ICTVdb/**

[d] Wheeler *et al*. (2001) Database resources of the national center for biotechnology information. *Nucleic Acids Research*, **28**(1), 10–14.

[e] Open Biological Ontologies, *Cell Ontology*, **http://cvs.sourceforge.net/viewcvs.py/obo/obo/ontology/anatomy/cell_type/cell.ontology**

[f] Blake *et al*. (2001) The Mouse Genome Database (MGD): Integration nexus for the laboratory mouse. *Nucleic Acids Research*, **29**(1), 91–94.

[g] International Statistical Classification of Diseases and Related Health Problems, 1989; Revision, Geneva, World Health Organization, 1992.

[h] Humphrey (1984) File maintenance of MeSH headings in MEDLINE. *Journal of the American Society for Information Science*, **35**(1), 34–44.

[i] Spackman *et al*. (1997) A reference terminology for health care. *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, 640–644.

[j] Lindberg C (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. *Journal (American Medical Record Association)*, **61**(5), 40–42.

[k] Covitz *et al*. (2003) caCORE: A common infrastructure for cancer informatics. *Bioinformatics*, **19**(18), 2404–2412.

[l] **http://www.combichem.net/files/complib2.htm**

[m] Claverie (2001) What if there are only 30,000 human genes? *Science*, 1255–1257.

[n] NCBI Taxonomy statistics. **http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi**

[o] Balaram (2003) Extinction. *Current Science*, **85**(1): 1–3.

[p] American Cell Type Collection. **http://www.atcc.org/**

[q] European Collection of Cell Cultures (ECACC), **http://ecacc.org.uk/**

KEGG; *see* Article 101, **Design of KEGG and GO**, Volume 8) and the model organism database systems (e.g., Flybase – Ashburner *et al*., 1994, etc.). In summary, Blois' theory has been pivotal in improving ontologies and is likely to play a role in conceiving NLP systems capable of reusing, improving, or coding in these ontologies.

2.3. *Zellig Harris sublanguage theory*

In 1968, Harris observed that the grammar of a sublanguage is quite different from that of standard language. For example, well-formed patient histories in medicine are written telegraphically, omitting verbs and articles. Harris proceeded to hypothesize that scientific subspecialties use specialized sublanguages, which can be represented as information content and structure. The grammar of a sublanguage can be determined by observing its combination of elements. Harris' theory has become well established as it has been comprehensively evaluated and corroborated. Friedman provides an overview of the theory with examples in biology and medicine (Harris, 2002; Friedman *et al*., 1995). Fundamentally, Harris' theory has helped determine a number of semantic properties that result in additional restrictions on linguistic structure. Thus, accurate NLP systems based on this theory have been shown to provide semantically rich data structures that can be straightforwardly associated with ontologies and predicate logic (Hripcsak *et al*., 1995; Friedman *et al*., 2001; Rzhetsky *et al*., 2004).

3. Properties of ontologies

In this section, we will describe the properties of ontologies, which may contribute to discerning their usefulness for NLP. In Section 3.1, we will address the scale of biology covered by contemporary ontologies, followed by criteria for ontological control (Section 3.2), relationships (Section 3.3), and suitability for NLP (Section 3.4).

3.1. *Scales of biology and of phenotypes*

The dogma of molecular biology postulates that three distinct components, genes, phenotypes, and environment, intervene in the development and function of an organism. Table 1 analyzes the content coverage of several ontological tools according to the conceptualization of different structure and attribute scales, such as size (meter), weight (kilogram), and scientific field. A recently published review of biomedical ontologies provides an in-depth analysis of their design (Bard *et al*., 2004). Owing to the wide span of phenotypic information and the large number of related ontologies, we will focus this section on the heterogeneous "conceptual" properties of phenotypes.

As described in Blois' theory, genes and gene products generally have a "crisp" definition owing to their quantized physical nature. In contrast, nonmolecular

phenotypes possess fuzzy physical boundaries that are highly compositional in nature (Rector *et al*., 2002; Elkin *et al*., 1998a,b Mays *et al*., 1996; Stuart *et al*., 1995). For example, one can refine a phenotypic description with additional modifiers as shown below in pseudocode:

```
[ {Abnormal Anatomy: "right humeral fracture"}]≡
    [{Regional Anatomy: Laterality: "right"}
    {Systemic Anatomy: Bone: "humerus"}
    {Abnormal Anatomical Structure: Morphology: "fracture"} ]
```

When components of a composite phenotypic concept are implemented in a database schema, instead of the terminology model, related implicit knowledge may be buried in the software architectures. Some ontologies propose a postcoordination strategy analogous to the schema of the databases they serve. In such architectures, both meta-data and terminological information are required to reconstruct a complete composite concept. For example, one ontology could represent separately the anatomy and the morphology of the previous example, while another ontology could represent them as a single composite concept. In addition, differences in *granularity of conceptual* representations are variant representations of the same phenotypic concept in an ontology focusing on the class concept, where in an ontology capturing their specific nuances they would be considered distinct phenotypes. Highly granular ontologies (nomenclature-like) maximize the content of necessary concepts and minimize the contingent ones, but it is difficult to define crisp criteria to distinguish ambiguous cases. While ontological data structures are challenged by the compositionality, granularity, and expressiveness of phenotypes, this is not necessarily the case of linguistic data structures, which may support nested and recursive relationships.

3.2. Principles of controlled terminology applied to ontologies

The terms of ontologies, not their relationships, can be considered a subset of controlled terminologies. In Table 2, we have adapted established desiderata for controlled terminologies (Cimino, 1998) to characterize the contemporary ontologies. The “ontological control” subclasses are defined as follows:

- *Concept-oriented*: unit of symbolic processing is the concept.
- *Formal semantic definition of concepts*: use formal descriptive language incorporating other concepts within the definition.
- *Concept permanence*: meaning of a concept should not change over time and the representative symbol should not be deleted from the ontology, simply “retired” if need be.
- *Nonredundancy*: definition of a concept should be unique and explicit.
- *Nonambiguity*: distinct concepts should not share the same term or synonym unless explicitly stated.

Table 2 Properties of biomedical ontologies including suitability for NLP thesauri

legend: ● = property provided, ○ = property partially provided, "empty box" = property not provided

Properties		Biomedical ontologies											
Class	Subclass	Gene ontology	Gene nomenclature	ICTVdb	NCBI taxonomy	Cell ontology	Mammalian phenotype	ICD-9, ICD-10	ICD-9-CM	MeSH	SNOMED CT	UMLS	NCI Metathesaurus
Content Coverage	**Concepts** (count)	**10^4**	**10^4**	**10^3**	**10^5**	**10^3**	**10^3**	**10^4**	**10^4**	**10^4**	**10^5**	**10^5**	**10^5**
	Synonymy (count)	n/a	**10^4**	**10^3**	**10^5**	n/a	**10^3**	**10^4**	**10^4**	**10^5**	**10^5**	**10^6**	**10^6**
Ontological Control (Change Management)	**Concept-Oriented**	●	●	●	●	●	●	●	○	○	●	●	●
	Formal Semantic Definition	●				●					●		
	Concept Permanence	●	●	●	●	●	●	●		●	●	●	●
	Concept Nonredundancy	●	●	●	●	●	●	●	○	●	●	●	●
	Concept Nonambuiguity	●	●	●	●	●	●	●	○	●	●	●	●
	Nonsemantic Concept Identifiers	●	●	●	●	●	●			○	●	●	●
	No Residual Category	●	●	●	●	●				●	●		
Relationships	**Monohierarchy (Tree)**			●	●		●			●			
	Polyhierarchy	●	●			●		●	●		●	●	●
	DAG (Cycle-free)	●	●	●	●	●	●	●	●	●	●		
	Internal Semantic relationships (in addition to subsumption)	●				●					●	●	●
	Semantic Relations to External Ontologies	●	●	●	●	n/a	●				●	●	●
Suitability Criteria for NLP	**Semantic Disambiguation of Synonyms**	n/a	○	○	○	n/a	○	○	○	○	●	●	●
	Abbreviations / Symbols	n/a	●	○		n/a				○	○	○	○
	Historical usage	○	●	○	●	○	○	○	○	●	●	●	●
	Lexical Variants (e.g. adjectives)										○	●	○
	Lexicon Adapted NLP									○	○	●	○
	Lexical Software (Open source)											●[a]	
	Terminology Server (Open source)	○								○	○	○	●
	Proof-of-Concept: Ontology used by NLP (for coding or as lexicon)	○	●				○	●	●	○	○	●	

[a]McCray (1994) Lexical methods for managing variation in biomedical terminologies. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 235–239

- *Nonsemantic concept identifier*: codes representing concepts should be meaningless (not representative of relationships to other concepts), as concepts are sometimes relocated in hierarchies, and it has been shown that most semantic identifiers become a barrier to concept permanence.

3.3. Relationships

Relationships between concepts differentiate the expressiveness of ontologies. For example, property-derived ontologies usually only maintain subsumption types of relationships ("is a" or "part of"). The subclasses are defined as follows:

- *Monohierarchy (tree)*: each concept has only one parent, which is critical for current phylogenetic analyses.
- *Polyhierarchy*: concepts may have more than one parent, which is useful for complex phenotypesthat can be classified in many complementary, yet distinct, ways.
- *Directed Acyclic Graph* (*DAG*): no cycles are created in the semantic network of relationships.
- *Internal semantic relationships*, in addition to subsumption.
- *Semantic relations to external ontologies*: some new ontologies point to external ontologies using a semantic relationship to generate a transontology network (e.g., translation tables).

3.4. Suitability criteria for NLP

The "NLP Suitability" subclasses that provide criteria to discriminate among the values of distinct terminologies are defined as:

- *Semantic Disambiguation of Synonyms*: Though controlled terminologies used for database queries should be unambiguous, those used for NLP need to reflect their original knowledge sources. While merging different ontologies from cross-disciplinary domains, it is not unusual to have terminological overlaps. By design, some ontologies make such overlaps explicit by leaving existing relationships in the source terminologies unchanged. Such explicit semantic disambiguation of synonyms may provide clues on contextual usage and be advantageous to the NLP community.
- *Abbreviations/Symbols*: must be recognized for the disambiguation of narratives.
- *Historical usage of concepts/terms/relationships*.
- *Lexical Variants*.
- *Lexicon-Adapted NLP*.
- *Lexical Software*.
- *Terminology Server*: reduces the total cost of operation of a distributed component-based system.
- *Proof-of-Concept*: ontology has been proven to work with an NLP system.

4. Biomedical ontologies for NLP

Biomedical ontologies are constrained data structures favoring computations with formal hierarchies and semantics. Formal ontologies and their associated systems (Musen *et al*., 1995; Rector *et al*., 1997, 1998; Campbell *et al*., 1994) are robust architectures designed for knowledge representation of concepts in a formal language (e.g., predicate logic) and have been used in biology and medicine (Rubin *et al*., 2002; Embley *et al*., 1998; Snoussi *et al*., 2001; Silvescu *et al*., 2001; Friedman *et al*., 1995). Obstacles in modeling knowledge in a formal ontology involve the difficulties and costs of (1) achieving consensus regarding the definition of knowledge representations and (2) enumerating the *context features* and the background knowledge required to ascribe meaning to a specific representation (Musen, 1992; Rector *et al*., 2002; Pole *et al*., 1996). In addition, ontological systems are not designed to operate over semistructured narratives that are not formally represented.

Biomedical ontologies, such as those presented in Tables 1 and 2, differ significantly in their focus (e.g., number of biological scales), granularity (e.g., classifications or nomenclatures), and formalisms of representation. In order to summarize their capabilities for NLP, we have organized them in three categories: (4.1) focused and full-depth ontologies, (4.2) focused classifications, and (4.3) broad-scope and full-depth ontologies.

4.1. Focused and full-depth ontologies

Gene Ontology (GO), Gene Nomenclature (GN), the Universal Virus Database (ICTVdb),National Center for Biotechnology Information NCBI Taxonomy (Taxon) and Cell Ontology (CO) are examples of tightly focused ontologies that have narrow breadth, full depth, and share a number of properties (Table 1). In addition, all these ontologies share a strong formalism of representation distinguishing them from data dictionaries (Table 2). However, with the exception of GN, they do not provide features to accelerate their reuse by an NLP system.

Since the inception of GO in 2000, over 25 distinct model organism database groups and other phenotypic databases have been manually annotated, totaling over five million distinct genes and proteins with a relationship to one of the ten thousand distinct GO concepts (The Gene Ontology Consortium, 2004). This fundamental work has set the stage for increasing the productivity of phenotypic research across different communities using genome-wide integrative discovery tools (Jenssen *et al*., 2001; Perez-Iratxeta *et al*., 2002; Raychaudhuri *et al*., 2003, 2002; Sarkar *et al*., 2003; Jensen *et al*., 2003; McCray *et al*., 2002; Tulipano *et al*., 2003; Bodenreider *et al*., 2002).

4.2. Focused classifications

Various versions of the International Classification of Diseases (ICD-9, ICD-10, and ICD9-CM) are focused *terminologies* that are implicitly constructed, but are not as

straightforward to use as the other ontologies. For example, during annual updates, ICD-9-CM allows unique non-concept-oriented identifiers to be reused for diseases. While these classifications satisfy the previously described ontology definitions, they are generally considered significantly less formal than the other classes. Although methods to circumvent these limitations and use these terminologies as ontologies have been proposed by Cimino (Cimino, 1998), none achieve the depth and scope of the Systemized Nomenclature of Medicine (SNOMED), the UMLS, or the NCI Metathesaurus in the diseases area, and, thus, are probably not the appropriate terminology choices for NLP.

4.3. Broad-scope and full-depth ontologies

These are definitively the most readily available for NLP. SNOMED Clinical Terms (CT), developed and maintained by the College of American Pathologists, provides a broad coverage of clinical and medical concepts. Recently, the content has been significantly increased by merging with the READ codes, used extensively in UK clinical practices. In contrast, the UMLS is a "pure metathesaurus". Since it integrates, using a full graph with cycles, versions of GO, NCBI Taxonomy, ICD-9 CM SNOMED, and Medical Subject Headings (MeSH), the UMLS offers broader breadth and depth than all other ontologies, with the exception of the NCI Metathesaurus. However, it does not offer the most recent translations to its source terminologies since the update process is semiautomated and rate-limiting. Nevertheless, emerging new properties offered by the UMLS, composed of derivative databases and a software toolbox specialized for NLP, make it by far the most adapted terminology for immediate use in NLP (McCray *et al*., 2001; Johnson, 1999; Browne *et al*., 2003).

The NCI Metathesaurus is a hybrid ontology, and likely the largest biomedical ontology available, containing additional source contents and external ontologies. The software services offered by the NCI Center for Bioinformatics (NCICB) are distinct from those of the UMLS, offering direct high-throughput access to their Metathesaurus via the internet and a molecular bioinformatics toolbox called caBIO (Covitz *et al*., 2003). In partnership with the research community, the NCI is creating a common and extensible bioinformatics platform that integrates diverse data types and supports interoperable tools (Cancer Biomedical Informatics Grid - caBIG **http://cabig.nci.nih.gov**).

5. NLP and ontologies

NLP produces data structures that are significantly more flexible than ontologies, but come for a cost. Their linguistic structures, which can be nested and compositional, are more expressive and complicated to analyze than those of ontologies, and the semantic networks produced by NLP systems often comprise cycles that make the reuse of the information challenging and often problematic. A number of approaches have been used to determine the usage of ontologies for NLP. Proof-of-concept projects demonstrated that, though the knowledge found in narratives is

highly nested and densely embedded, an NLP system based on sublanguage grammar can be coordinated with a large scale ontology for coding purposes (Friedman *et al*., 2004; Lussier *et al*., 2001). Other studies aimed at understanding the ambiguity sources of combined ontologies, showed them to be both internal and external ambiguity (Tuason *et al*., 2004). Additionally, the semantic context afforded by some ontologies has been shown to be useful for NLP (Rindflesch *et al*., 2003), extending its accuracy and expressiveness (Baclawski *et al*., 2000). Other NLP researchers have conceived their systems, relying intensively on a body of domain knowledge derived from the conceptual formalism of ontologies, and common methodological and theoretical principles to enforce this convergence (Zweigenbaum *et al*., 1995; Friedman *et al*., 2001). As shown in Table 1, under "NLP Content Coverage", while contemporary NLP systems provide adequate coverage of the gene and gene product (Friedman *et al*., 2001), and clinical scales (Sager *et al*., 1995; Friedman *et al*., 1997; Zweigenbaum *et al*., 1997, 1998; Wagner *et al*., 1998; Lovis *et al*., 1998; Rassinoux *et al*., 1997), the cellular and subcellular levels are inadequately represented. Significantly, an NLP method coding narratives in Gene Ontology has also been developed (Raychaudhuri *et al*., 2002). The National Library of Medicine (NLM) provides the most comprehensive services and products designed to facilitate the integration of and focus scientific research at the intersection of NLP and ontologies, such as: (1) the UMLS database, comprising a semantic network connecting several ontologies and providing a comparative analysis of term ambiguity, (2) the "Specialist Lexicon", designed specifically for NLP technology, (3) an open-source computational toolbox for reuse of the lexical and ontological UMLS databases, and (4) Lexical Properties of GO (McCray *et al*., 2002).

6. Conclusions and future directions

Three contemporary theories providing the framework for the integration of their respective fields were covered: semantics (Semiotic Triangle), informatics (Information Flow in Scales of Biomedicine), and NLP (Sublanguage Theory). In addition, the potential of ontologies covering a large variety of genotypic and phenotypic information in biomedical knowledge representation have also been described. Metathesauri, such as the UMLS and NCI Metathesaurus, combine the best ontologies in an increasingly computable form, as well as providing lexicon and open-source software specialized for NLP, flexible linguistic models of language and formal ontological models of concepts. However, at least two important domains are not well covered by classification terminologies: (1) strains of viruses, bacteria, fungi, and cell lines, and (2) environmental factors from experimental designs and conditions.

While reusing ontologies as lexicons for NLP systems is reasonably straightforward, coding narratives in a formal ontology remains challenging. In practice, coding genes from narratives remains difficult for NLP due to ambiguities in the English language, as well as the problems related to expressiveness, compositionality, and granularity. In order to query or infer using such large-scale knowledge, one has to simplify the nested or complicated forms to a standardized format via inefficient and rate-limiting transformations. However, recent studies have identified

methods to incrementally annotate narratives in XML format, first with linguistic constructs and second with declarative constructs (Friedman *et al*., 1999). In such form, data (narratives), information (linguistic structures), and knowledge (declarative structures, ontological coding) coexist, and queries can be conducted efficiently. Lexicons of NLP techniques have been enriched with ontological terms to produce coded output, but the fundamental problem of expressiveness remains: data structures produced by an NLP system are more informative than ontologies, while the latter are currently more computable. By extension of the Harris Sublanguage Theory, linguistic systems may eventually augment or even challenge the current curation of ontologies with automated and dynamic annotation of biomedical ontologies driven by the corpora of the scientific communications (Mollah *et al*., 2003; Campbell *et al*., 2002).

In conclusion, hybrid technologies using ontologies and NLP are poised to produce a paradigm shift in the annotations of biological databases, currently accomplished manually.

Related articles

Article 79, **Introduction to ontologies in biomedicine: from powertools to assistants**, Volume 8; Article 80, **Ontologies for the life sciences**, Volume 8; Article 83, **Ontologies for information retrieval**, Volume 8; Article 86, **Automatic concept identification in biomedical literature**, Volume 8; Article 87, **Merging and comparing ontologies**, Volume 8

Further reading

Bono H, Kasukawa T, Furuno M, Hayashizaki Y and Okazaki Y (2002) FANTOM DB: Database of functional annotation of RIKEN mouse cDNA clones. *Nucleic Acids Research*, **30**, 116–118.

Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J Cox A, *et al*. (2003) The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, **13**, 662–672.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, **419**, 512–519.

Dellaire G, Farrall R and Bickmore WA (2003) Vertebrate nuclear proteins: The Nuclear Protein Database (NPD): Sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Research*, **31**, 328–330.

Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M Sherlock G, *et al*. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research*, **30**, 69–72.

Friedman C, Kra P and Rzhetsky A (2002) Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, **35**(4), 222–235.

Ingenerf J and Giere W (1998) Concept-oriented standardization and statistics-oriented classification: Continuing the classification versus nomenclature controversy. *Methods of Information in Medicine*, **37**(4–5), 527–539.

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**, D277–D280. Database issue.

Pletcher SD, Macdonald SJ, Marguerie R, Certa U, Stearns SC, Goldstein DB and Partridge L (2002) Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. *Current Biology*, **12**, 712–723.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, *et al*. (2003) The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, **31**, 224–228.

Simin K, Scuderi A, Reamey J, Dunn D, Weiss R, Metherall JE and Letsou A (2002) Profiling patterned transcripts in Drosophila embryos. *Genome Research*, **12**, 1040–1047.

Sonstegard TS, Capuco AV, White J, Van Tassell CP, Connor EE, Cho J, Sultana R, Shade L, Wray JE, Wells KD, *et al*. (2002) Analysis of bovine mammary gland EST and functional annotation of the Bos taurus gene index. *Mammalian genome: Official Journal of the International Mammalian Genome Society*, **13**, 373–379.

Toppo S, Cannata N, Fontana P, Romualdi C, Laveder P, Bertocco E, Lanfranchi G and Valle G (2003) TRAIT (TRAnscript Integrated Table): A knowledge base of human skeletal muscle transcripts. *Bioinformatics*, **19**, 661–662.

Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB and Robinson GE (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Research*, **12**, 555–566.

Yu H, Friedman C, Rhzetsky A and Kra P (1999) Representing genomic knowledge in the UMLS semantic network. *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, 181–185.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, *et al*. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science*, **298**, 149–159.

References

Ashburner M and Drysdale R (1994) FlyBase – the Drosophila genetic database. *Development*, **120**(7), 2077–2079.

Baclawski K, Cigna J, Kokar MM, Mager P and Indurkhya B (2000) Knowledge representation and indexing using the unified medical language system. *Pacific Symposium on Biocomputing*, Hawaii, 4–9 January, 2000 pp. 493–504.

Balaram P (2003) Extinction. *Current Science*, **85**(1), 1–3. July 2003.

Bard JB and Rhee SY (2004) Ontologies in biology: Design, applications and future challenges. *Nature Reviews. Genetics*, **5**(3), 213–222. Review.

Blake JA, Eppig JT, Richardson JE, Bult CJ and Kadin JA (2001) The Mouse Genome Database (MGD): Integration nexus for the laboratory mouse. *Nucleic Acids Research*, **29**(1), 91–94.

Blois MS (1984) *Information and Medicine: The Nature of Medical Description*, University of California Press: Berkeley.

Bodenreider O, Mitchell JA and McCray AT (2002) Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proceedings of the Annual Symposium of the American Medical Informatics Association*, San Antonio, 9–13 November, 2002, pp. 61–65.

Browne AC, Divita G, Aronson AR and McCray AT (2003) UMLS language and vocabulary tools. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, Washington 8–12 November, 2003, p. 798.

Buchen-Osmond C (1997) Further progress in ICTVdB, a universal virus database. *Archives of Virology*, **142**(8), 1734–1739. PMID: 9672633. **http://www.ncbi.nlm.nih.gov/ICTVdb/**

Campbell DA and Johnson SB (2002) A transformational-based learner for dependency grammars in discharge summaries. Association for computational linguistics. *Proceedings of the*

Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, pp. 37–44.

Campbell KE, Das AK and Musen MA (1994) A logical foundation for representation of clinical data. *Journal of the American Medical Informatics Association*, **1**(3), 218–232.

Campbell KE, Oliver DE, Spackman KA and Shortliffe EH (1998) Representing thoughts, words, and things in the UMLS. *Journal of the American Medical Informatics Association*, **5**(5), 421–431.

Cimino JJ (1998) Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, **37**(4–5), 394–403.

Claverie JM (2001) Gene number. What if there are only 30,000 human genes? *Science*, **291**(5507), 1255–1257.

Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S and Buetow KH (2003) caCORE: A common infrastructure for cancer informatics. *Bioinformatics*, **19**(18), 2404–2412.

Elkin PL, Bailey KR and Chute C (1998a) A randomized controlled trial of automated term composition. In *Proceedings of the Fall AMIA 1998 Annual Symposium*, Chute CG (Ed.), Hanley & Belfus: Philadelphia, pp. 765–774.

Elkin PL, Tuttle MS, Keck K, Campbell K, Atkin G and Chute C (1998b) The role of compositionality in standardized problem list generation. In *Proceedings of MEDINFO 98*, Cesnik B (Ed.), IOS Press: Amsterdam, pp. 660–664.

Embley DW, Campbell DM, Liddle SW and Smith RD (1998) Ontology-based extraction and structuring of information from data-rich unstructured documents. *Proceedings of International Conference on Information and Knowledge Management*, 7, Bethesda.

Friedman C (1997) Towards a comprehensive medical language processing system: Methods and issues. *Proceedings: A Conference of the American Medical Informatics Association / . . . AMIA Annual Fall Symposium. AMIA Fall Symposium*, Nashville, Tenesse, 25–29, October, 1997, pp. 595–599.

Friedman C, Hripcsak G, Shagina L and Liu H (1999) Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, **6**(1), 76–87.

Friedman C, Huff SM, Hersh WR, Pattison-Gordon E and Cimino JJ (1995) The Canon group's effort: Working toward a merged model. *Journal of the American Medical Informatics Association*, **2**(1), 4–18.

Friedman C, Kra P, Yu H, Krauthammer M and Rzhetsky A (2001) GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(Suppl 1), S74–S82.

Friedman C, Shagina L, Lussier Y and Hripcsak G (2004) Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*. **11**(5), 392–402. Epub 2004.

Guarino N (1998) *"Formal Ontology and Information Systems" Proceedings of FOIS '98*, Trento, 1998. IOS Press: Amsterdam, pp. 3–15.

Harris ZS (2002) The structure of science information. *Journal of Biomedical Informatics*, **35**(4), 215–221.

Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB and Clayton PD (1995) Unlocking clinical data from narrative reports: A study of natural language processing. *Annals of Internal Medicine*, **122**(9), 681–688.

Humphrey SM (1984) File maintenance of MeSH headings in MEDLINE. *Journal of the American Society for Information Science American Society for Information Science*, **35**(1), 34–44.

ICD-10 (1992) *International Statistical Classification of Diseases and Related Health Problems, 1989 Revision*, World Health Organization: Geneva.

Jensen LJ, Gupta R, Staerfeldt HH and Brunak S (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, **19**, 635–642.

Jenssen TK, Laegreid A, Komorowski J and Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**, 21–28.

Johnson SB (1999) A semantic lexicon for medical language processing. *Journal of the American Medical Informatics Association*, **6**(3), 205–218.

Lindberg C (1990) The Unified Medical Language System (UMLS) of the national library of medicine. *Journal (American Medical Record Association)*, **61**(5), 40–42.

Lovis C, Baud R, Rassinoux AM, Michel PA and Scherrer JR (1998) Medical dictionaries for patient encoding systems: A methodology. *Artificial Intelligence in Medicine*, **14**(1–2), 201–214.

Lussier YA, Shagina L and Friedman C (2001) Automating SNOMED coding using medical language understanding: A feasibility study. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, Washington, 3–7, November, 2001, pp. 418–422.

Mays E, Weida R, Dionne R, Laker M, White B, Liang C and Oles FJ (1996) Scalable and expressive medical terminologies. In *Proceedings of 1996 AMIA Fall Annual Symposium*, Cimino JJ (Ed.), Hanley & Belfus: Philadelphia, 259–263.

McCray AT, Bodenreider O, Malley JD and Browne AC (2001) Evaluating UMLS strings for natural language processing. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, Washington, 3–7 November, 2001, pp. 448–452.

McCray AT, Browne AC and Bodenreider O (2002) The lexical properties of the gene ontology. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium 2002*, San Antonio, 9–13 November, 2002, pp. 504–508.

McCray AT, Srinivasan S and Browne AC (1994) Lexical methods for managing variation in biomedical terminologies. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 1994, 235–239.

Mollah SA and Johnson SB (2003) Automatic learning of the morphology of medical language using information compression. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, Washington, 8–12 November, 2003, p. 938.

Musen MA (1992) Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research, an International Journal*, **25**(5), 435–467.

Musen MA (2002) Medical informatics: searching for underlying components. *Methods of Information in Medicine*, **41**(1), 12–19.

Musen MA, Gennari JH, Eriksson H, Tu SW and Puerta AR (1995) PROTEGE-II: Computer support for development of intelligent systems from libraries of components. *Medinfo*, **8**(Pt 1), 766–770.

Perez-Iratxeta C, Bork P and Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, **31**, 316–319.

Pole PM and Rector AL (1996) Mapping the GALEN CORE model to SNOMED-III: Initial experiments. *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, Washington, 26–30 October, 1996, pp. 100–104.

Rassinoux AM, Miller RA, Baud RH and Scherrer JR (1997) Compositional and enumerative designs for medical language representation. *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, Nashville, Tenesse, 25–29 October, 1997, pp. 620–624.

Raychaudhuri S and Altman RB (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401.

Raychaudhuri S, Chang JT, Sutphin PD and Altman RB (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, **12**(1), 203–214.

Rector A, Bechhofer S, Goble CA, Horrocks I, Nowlan WA and Solomon WD (1997) The GALEN modeling language for medical terminology. *Artificial Intelligence in Medicine*, **9**, 139–171.

Rector AL, Rogers J, Roberts A and Wroe C (2002) Scale and context: issues in ontologies to link health- and bio-informatics. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, San Antonio, 9–13 November, 2002, pp. 642–646.

Rector A, Rossi A, Consorti MF and Zanstra P (1998) Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN organisation. *International Journal of Medical Informatics*, **48**(1-3), 71–84.

Rindflesch TC and Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, **36**(6), 462–477.

Rubin DL, Hewitt M, Oliver DE, Klein TE and Altman RB (2002) Automating data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML. In *Proceedings of the Pacific Symposium on Biology*, Lihue.

Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W Wilbur WJ, *et al*. (2004) GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, **37**(1), 43–53.

Sager N, Lyman M, Nhan NT and Tick LJ (1995) Medical language processing: Applications. *Methods of Information in Medicine*, **34**(1-2), 140–146.

Sarkar IN, Cantor MN, Gelman R, Hartel F and Lussier YA (2003) Linking biomedical language information and knowledge resources in the 21st century: GO and UMLS. *Pacific Symposium on Biocomputing*, **8**, 427–450.

Silvescu A, Reinoso-Castillo J and Honavar V (2001) Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous biological data sources. In *IJCAI-01 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Dynamic, Autonomous Data and Knowledge Sources. International Joint Conference on Artificial Intelligence*.

Snoussi H, Magnin L and Nie J-Y (2001) Heterogeneous web data extraction using ontologies. In third international bi-conference workshop on Agent-Oriented Information Systems (AOIS-2001) In *Proceedings of the 5th International Conference on Autonomous Agents*, Montréal, 2001.

Sowa JF (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co: CA, p. 593.

Spackman KA, Campbell KE, Cote RA and Snomed RT (1997) A reference terminology for health care. *Proceedings: A conference of the American Medical Informatics Association / . . . AMIA Annual Fall Symposium. AMIA Fall Symposium*, Nashville, Tenesse, 25–29 October, 1997, pp. 640–644.

Stuart NJ, Nels OE, Lloyd F, Tuttle MS, William CG and Sherertz DD (1995) Identifying concepts in medical knowledge. In *MEDINFO 95*, Greenes RAea (Ed.), IOS Press: Amsterdam, pp. 33–36.

The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261. **http://www.geneontology.org/**.

Tuason O, Chen L, Liu H, Blake JA and Friedman C (2004) Biological nomenclatures: A source of lexical knowledge and ambiguity. *Pacific Symposium on Biocomputing*, Hawaii, 6–10 January, 2004, pp. 238–249.

Tulipano PK, Millar W and Cimino JJ (2003) Linking molecular imaging terminology to the Gene Ontology (GO). *Pacific Symposium on Biocomputing*, **8**, 613–623.

Wagner JC, Rogers JE, Baud RH and Scherrer JR (1998) Natural language generation of surgical procedures. *Medinfo*, **9**(Pt 1), 591–595.

Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW and Povey S (2002) Guidelines for human gene nomenclature? *Genomics*, **79**(4), 464–470.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA and Rapp BA (2000) Database resources of the national center for biotechnology information. *Nucleic Acids Research*, **28**(1), 10–14.

Zweigenbaum P, Bachimont B, Bouaud J, Charlet J and Boisvieux JF (1995) Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of Information in Medicine*, **34**(1-2), 15–24.

Zweigenbaum P, Bouaud J, Bachimont B, Charlet J and Boisvieux JF (1997) Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. *Proceedings of the AMIA Annual Fall Symposium*, Nashville, Tenesse, 25–29 October, 1997, pp. 590–594.

Zweigenbaum P and Courtois P (1998) Acquisition of lexical resources from SNOMED for medical language processing. *Medinfo*, **9**(Pt 1), 586–590.

Specialist Review

TAMBIS: transparent access to multiple bioinformatics services

Robert Stevens, Norman W. Paton, Sean Bechhofer, Gary Ng, Martin Peim, Patricia Baker, Carole Goble and Andy Brass
University of Manchester, Manchester, UK

1. Introduction

In this article, we describe Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) (Goble *et al.*, 2001), an ontological approach to solve a perennial problem in bioinformatics – that of integrating or interoperating between multiple diverse bioinformatics data and analytical resources (Karp, 1995). Bioinformatics resources present a classic example of autonomous and heterogeneous resources that need to be integrated to accomplish many bioinformatics analyses. For instance, a bioinformatician may select a sequence from swiss-prot, perform a similarity search with BLAST, and then perform a search against PROSITE with each of the resulting top hits from the similarity search. As well as the biological question prompting this analysis, the bioinformatician needs to know:

- the resources to be used in the analysis;
- where those resources are located (usually on the Web);
- in which order to use those resources;
- how to use the individual resources; and
- how to transfer data between those resources.

All of these requirements place a large burden of knowledge on a bioinformatician or biologist. The last point, in particular, is the crux of the computational problem in the integration of multiple bioinformatics information sources (Karp, 1995; Davidson *et al.*, 1995). The reality of integrating these resources is that they were rarely designed to be used together and that constructing requests that refer to several resources is time consuming and error-prone.

This scenario makes work in such an area a central topic in bioinformatics. The aim is to let a biologist ask the biology question and have the application deal with the mechanics of how to answer that question. There is a general need to *integrate* bioinformatics resources – that is, make these diverse resources work together

(Karp, 1995). Thus, within bioinformatics, there is a need to create suitable query-building facilities over diverse, heterogeneous, and distributed resources as well as suitable middleware technology for answering such questions (Davidson *et al*., 1995). In essence, there is a need to capture the knowledge of how to perform a task within bioinformatics software and to enable a scientist to use his or her domain knowledge to pose the question.

The difficulty in integrating these diverse resources to enable common queries are the differences or heterogeneities between those resources. These heterogeneities exist at two levels in bioinformatics resources – first, at the syntactic or structural level; and second, at the semantic level. Databases and services operate with different communication protocols and through applications running on different platforms, and yield results in many proprietary formats. All of these points and more make even simple integration between resources difficult because of syntactic heterogeneity. On top of this problem is that of the concepts within bioinformatics being represented differently in many resources. The same concept can have different representations and different concepts can have the same representation. This heterogeneity can exist both at the schema level and at the level of the values held within a schema (Karp, 1995). This semantic heterogeneity makes asking questions over multiple resources difficult, as the user or the query machinery needs to accommodate the differing representations of biological knowledge within diverse resources.

A biologist should be free to ask the biological questions he or she wishes to ask – without necessarily having to know the answers to all those points listed above. Our proposal in TAMBIS was to make all these issues *transparent* to a biologist by holding that knowledge within the query application in the form of an ontology (*see* Article 79, **Introduction to ontologies in biomedicine: from powertools to assistants**, Volume 8 and Article 80, **Ontologies for the life sciences**, Volume 8) (Goble *et al*., 2001). This *transparency* gives the name to the project: Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) (**http://img.cs.man.ac.uk/tambis**).

2. Integrating bioinformatics resources

In general, the strategy in overcoming syntactic heterogeneity is to provide a middleware layer into which heterogeneous resources can be *included* and *transformed* to provide a common programming or querying interface for an application. BioKleisli, later to become K1 (Davidson *et al*., 2001), is a middleware layer, data exchange format, and query language directed toward the bioinformatics domain. It includes a wrapping facility that *includes* third-party resources into the middleware layer, and the Collection Programming Language (CPL) allows these wrapped resources to be further *transformed* to provide a set of query facilities of the same syntactic nature over these wrapped resources. In BioKleisli, the declarative, composition-based CPL is provided to orchestrate multiple resources via their wrappers and collect results. Figure 1 shows the CPL for a typical multi source bioinformatics query. In the BioKleisli system, the joining of resources is delegated

```
{m |
\p <getspentrybyspecies("guppy"),
\p2 <doblastpsearchonspentry(p),
\m <doprositescanonspentry(p1)};
```

Figure 1 CPL expression for the query "Find all the motifs on all proteins that are significantly similar to any guppy protein"

to the middleware, but the user still has to manage the semantic heterogeneity (e.g., which resource to use and in which order to use those resources).

There are several integration efforts within the bioinformatics arena, but none of them afford the level of transparency needed by some non expert bioinformaticians. The most successful and widely used is the Sequence Retrieval System (SRS) (Etzold *et al*., 1996). SRS has a language for describing, parsing, and indexing flatfile databanks. Combined with a query language and a Web interface, users are able to both formulate queries and query by navigation. The user is, however, still presented with the files with their original semantics; the user has to perform any semantic reconciliation required. For an integration solution, middleware such as BioKleisli is essential, but that only provides a syntactic or structural integration, as seen in SRS. Much of the knowledge burden still lies on the biologist user.

Conforming to a common terminology or ontology can remove these remaining knowledge burdens. By committing to a common terminology or controlled vocabulary that captures knowledge about the domain, a homogenizing layer, acting as a global schema over the autonomous, distributed and heterogeneous bioinformatics resources, can be made. The difficulty is to then map from this global schema, against which queries can be made, down to the underlying resources. That is, there is a need to *mediate* between the global schema and the underlying resources (Wiederhold, 1992). In TAMBIS, we use an ontology to provide this global view or schema over the various bioinformatics resources. This ontology, together with a query processor, provides the mediation between the ontology and the underlying resources.

3. TAMBIS architecture

The following are the kinds of queries that TAMBIS can answer.

- Find any motif indicating a receptor function found on a protein similar to any nucleic acid sequenced for the aardvark.
- Find DNA-binding motifs from eukaryotic proteins involved in apoptosis.
- Find chimpanzee homologs of human proteins containing seven-bladed propeller domains.
- Find homologs to apoptosis-receptor proteins that also have ATPase activity.
- Find motifs in enzymes using thymine as a substrate and iron as a cofactor.

All such queries are complex and multisource. To ask a question such as "Find all antigenic human apoptosis receptor protein homologs and their phosphorylation

site motifs", TAMBIS uses the three sources swiss-prot, BLAST, and PROSITE. In TAMBIS, the user does not have to choose the sources or the keywords with which to filter the proteins or perform any filtering, transformation, or mapping; all these tasks are transparent to the user. The knowledge of how to accomplish these tasks is captured in the TAMBIS ontology and the rest of the TAMBIS application.

Figure 2 shows the architecture of TAMBIS. The geneticXchange middleware layer, K1 or discoveryHub, can be seen at the base of the system; it provides a Java API by which queries formulated in discoveryHub's query language, CPL, (Davidson *et al*., 2001) are passed to and executed against the wrappers. The bioservices currently wrapped within TAMBIS are swiss-prot, ENZYME, BLAST, PROSITE, and CATH. This layer of TAMBIS is the part that actually answers the queries posed by biologists.

The rest of the system is concerned with formulating conceptual, declarative queries; mapping query elements to sources and values used within sources; optimizing the order of calls to methods on the wrapped sources; and finally encoding the query in discoveryHub's query language.

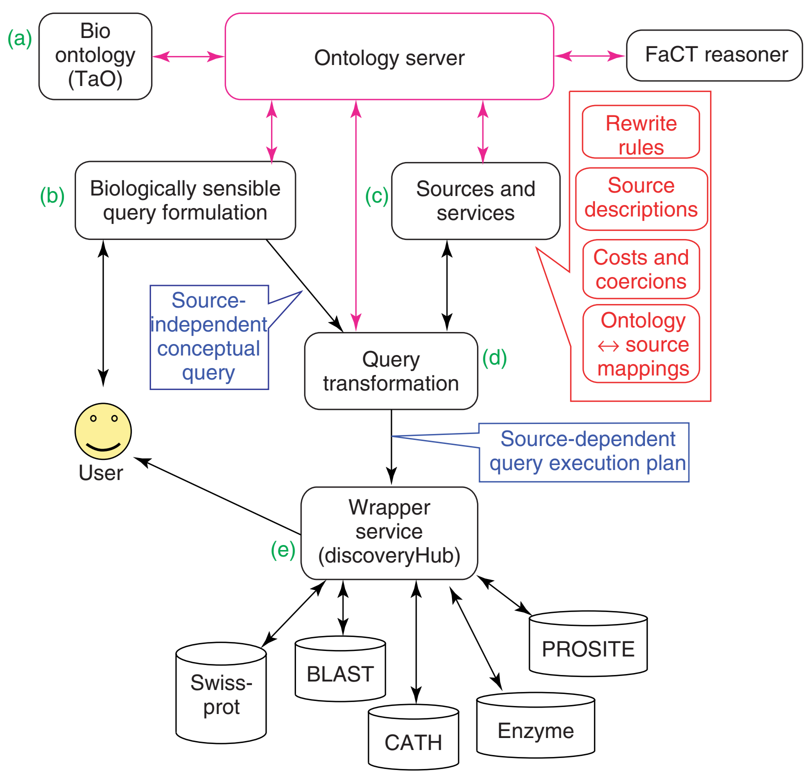


Figure 2 The TAMBIS architecture: (a) The TaO, an ontology of biological and bioinformatics terms; (b) A knowledge-driven query-formulation interface; (c) A services model linking the TaO with the source services; (d) Transformation from high-level, source-independent queries into source-dependent, ordered queries; (e) A wrapper service dealing with external sources provided by discoveryHub

3.1. The TAMBIS ontology

The TAMBIS ontology (TaO) captures knowledge about molecular biology and bioinformatics. Thus, in one ontology, we describe the major entities and relationships within the resources over which we wish to be able to form queries. Table 1 describes the major categories of classes within the TaO, together with examples of the more specialized classes.

As well as the "is a" relationship, these concepts are linked by many other kinds of relationships that describe the biological and bioinformatics domain. These include, but are not limited to, *hasFunction, functionsInProcess, hasAccessionNumber, hasSequence*, and *hasComponent*.

The TaO is encoded in a Description Logic (DL) called GRAIL (Rector *et al*., 1997; *see also* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8). The TaO makes use of the compositional approach to describing concepts that is possible in DLs. For example, `Enzyme` was not simply asserted to exist as a child of `Protein`. `Enzyme` was described (in a logical form) to be a kind of `Protein` that *catalyzes* a `Reaction`. This description "defines" what it is to be an enzyme (*see* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8), and the DL reasoner is able to take such descriptions and infer a subsumption (or "is a" lattice), as well as check for logical consistency within those descriptions (*see* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8).

The TaO acts as a global schema, so users need to be able to form queries against that schema. TAMBIS takes advantage of the DL's dynamic ability to compose queries while the application using the ontology is running. The query-formulation interface (*see* Section 3.2) allows users to build new concepts that act as queries. Thus, the TaO only describes the potential queries a user may make against the resources – the ontology is dynamic and a user can build new queries in the same manner as the concept Enzyme was built.

It is not, however, sensible to be able to compose queries out of concepts without restriction. Naturally, the relationships existing within the ontology provide some restriction, but still some combinations would be biologically nonsensical. Take, for instance, the TaO description `bioMolecule` *hasComponent* `Motif`. In general this is true, but the concept of motif subsumes those that are parts of `Protein` and those that are parts of `NucleicAcid`. The TaO needs a further level of restriction

Table 1 The *domain-general* concepts, which describe the range of concepts covered in the TaO, and *domain-specific* concepts, which describe the depth of coverage within the TaO

General biology	Specific biology
BioMolecule	Protein, enzyme, nucleic acid, DNA, RNA
SequenceComponent	Motif, active site, modification site, amidation site, alpha-helix
BiologicalFunction	Enzyme function, motor activity, hormone function
BiologicalProcess	Cellular process, apoptosis, lactation
CellularComponent	Organelle, chloroplast, ribosome, nucleus, mitochondrion
Species	Eukaryote, yeast, human, aardvark
ProteinStructure	Mainly alpha, mainly beta, seven-bladed propeller, TIM barrel

limiting the general case of `BioMolecule` *hasComponent* `Motif` to be also true at the specific case, for example, of `AlphaHelix` being part of `Protein`. Thus, the TaO contains another level of restriction on the existence of relationships called sanctioning (Rector *et al*., 1997; Baker *et al*., 1999). This sanctioning layer only allows biologically sensible concept compositions to take place and helps guide the user in formulating queries. Such sanctioning is also a form of knowledge about the domain captured by the TaO.

So, we have a mechanism, via the TaO, that offers the chance to form concepts dynamically that describe the instances a biologist wishes to retrieve from bioinformatics resources. The TaO contains knowledge of what can be asked of these resources, but to offer these facilities to biologists, we need to be able to:

1. present a query-formulation user interface;
2. take the conceptual descriptions of instances and transform them to a concrete query plan against the real resources; and
3. execute that query plan against those resources.

The role of the TaO in the TAMBIS architecture is shown in Figure 2. The TaO is made accessible to the rest of the TAMBIS application via a terminology server (TeS). The TeS API allows the TAMBIS client to present a query formulation interface to a biologist. This client can ask questions of the TaO such as "What are the parents/children of concept x?"; "What are the other relationships held by concept x?"; "Is concept x satisfiable, and where does it lie in the lattice?"; and "What relationships are sanctioned for this concept?". Through these questions, the TAMBIS application can use the knowledge within the TaO to promote transparency in the process of building and executing bioinformatics queries.

3.2. The TAMBIS user interface

Figure 3 shows the TAMBIS query-formulation interface. The TaO is used to generate the content of the screen in both the query-formulation interface and an ontology browser. A concept is a query and, in TAMBIS, a user builds a new concept that describes what is to be retrieved. In query formulation, a TaO concept is presented as a button upon the user interface. Clicking the button yields a menu describing what actions may be performed with that concept in building a new concept or query. For instance, the user is offered the relationships to other concepts held by the concept in question. The user may accept a relationship to another concept and thus form a new concept in the TaO. This new concept is progressively refined until the user has described what instances he or she wishes to retrieve.

3.3. The TAMBIS query processor

The Sources and Services Model (SSM) (Figure 2) provides metadata describing the relationship between the resource-independent knowledge captured in the TaO

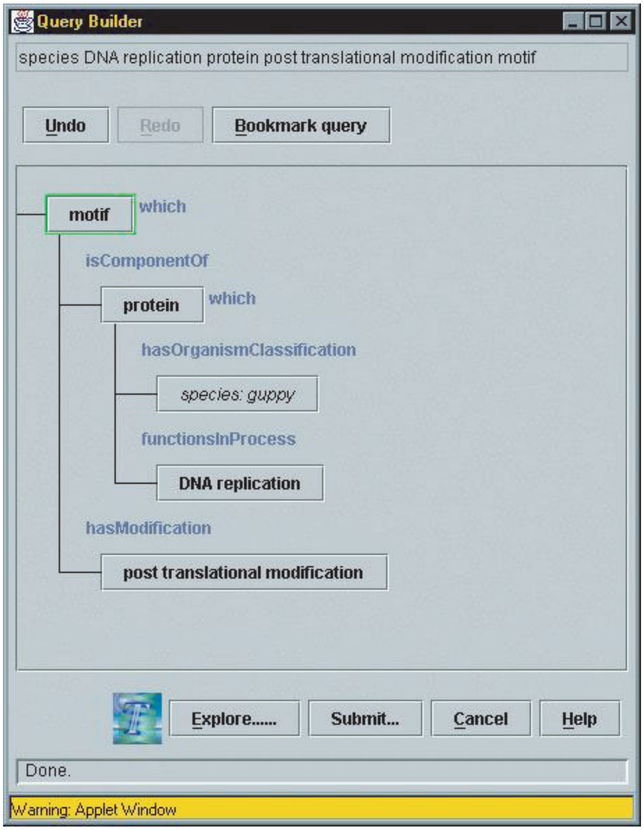


Figure 3 The TAMBIS query-formulation interface

and the wrapped resources in the discoveryHub middleware layer. In broad terms, the SSM maps concepts within the TaO to values representing these concepts held in the resources themselves, and also maps concepts and relationships to functions acting over those resources to retrieve instances. For example, the concept `Kinase` in the TaO maps to the value "kinase" in swiss-prot and "2.27" in ENZYME. In this manner, one concept in the TaO can accommodate the heterogeneous representations of those concepts in the resources themselves. Similarly, the composite concept `Protein hasAccessionNumber AccessionNumber` can be mapped via the SSM to `get-sp-entry-by-ac-no(accession-number)`. The SSM also records resource, parameter, and return types for these functions. Importantly, it also records the cost of executing these functions over the resources, in terms of number of instances returned, of executing these functions over the resources.

The resources swiss-prot, ENZYME, CATH, PROSITE, PROSITE SCAN, and BLAST can be queried using TAMBIS. They are accessed by TAMBIS using discoveryHub's wrapper service. Most bioinformatics resources have no query-language facilities, so requests directed at these resources are represented as a

collection of functions such as `get-sp-entries-by-species` or `get-sp-entries-by-function`. The discoveryHub middleware removes the syntactic or structural heterogeneity pervasive in bioinformatics resources, and the SSM helps remove or manage the semantic heterogeneities of these resources. The TAMBIS query processor (Paton *et al*., 1999) (see Figure 2) takes the high-level conceptual query formulated in the user interface and transforms it to a source dependent, ordered query plan against the wrapped bioinformatics services in discoveryHub. This concrete query plan is expressed in CPL (see Figure 1). The query processor uses a relatively simple algorithm to do this transformation (Paton *et al*., 1999), which relies heavily on the SSM for the transformation:

1. The conceptual query is broken down into a series of components.
2. Each component is matched against the SSM to find the cheapest (in terms of cost) component to fetch.
3. This continues until all query components have been matched against the SSM.
4. This ordered set of components is used to generate CPL code.

This query plan is then executed by the discoveryHub layer and the results presented to the user via HTML markup in a browser.

4. Discussion

TAMBIS tackles the perennial bioinformatics problems of distribution and heterogeneity of resources, both data and tools. The queries possible in TAMBIS are multisource and over distributed resources with heterogeneous APIs, query mechanisms, and semantics. These are typical bioinformatics tasks that require layers of knowledge over the simple computational facilities underlying the query itself. Through its ontology, TaO, TAMBIS provides a conceptual, knowledge-driven query interface over a middleware that allows distributed resources to be included and transformed to a common syntactic interface. The knowledge captured in the TaO enables transparent access to multiple bioinformatics information sources and frees a biologist who is not an expert in bioinformatics to concentrate upon the biological content of his or her query. TAMBIS, linking to the discoveryHub, can be found as an applet via **http://img.cs.man.ac.uk/tambis**. TAMBIS gives the illusion of a common query interface to distributed and heterogeneous bioinformatics resources. Using the TaO, TAMBIS frees biological scientists from the need to hold knowledge of how to answer bioinformatics queries and allows them to capitalize on their knowledge of what questions to ask.

Acknowledgments

This work was funded by AstraZeneca Pharmaceuticals, the BBSRC/EPSRC Bioinformatics Initiative, and recent work has been undertaken in collaboration with geneticXchange under the auspices of ESNW.

References

Baker P, Goble C, Bechhofer S, Paton N, Stevens R and Brass A (1999) An ontology for bioinformatics applications. *Bioinformatics*, **15**(6), 510–520.

Davidson SB, Crabtree J, Brunk BP, Schug J, Tannen V, Overton GC and Stoeckert Jr CJ (2001) K2/Kleisli and GUS Experiments in integrated access to genomic data sources. *IBM Systems Journal*, **40**(2), 512–531.

Davidson SB, Overton C and Buneman P (1995) Challenges in integrating biological data sources. *Journal of Computational Biology*, **2**(4), 557–572.

Etzold T, Ulyanov A and Argos P (1996) SRS: Information retrieval system for molecular biology data banks. *Methods in Enzymology*, **266**, 114–128.

Goble C, Stevens R, Ng G, Bechhofer S, Paton N, Baker P, Peim M and Brass A (2001) Transparent access to multiple bioinformatics information sources. *IBM Systems Journal Special issue on deep computing for the life sciences*, **40**(2), 532–552.

Karp P (1995) A strategy for database interoperation. *Journal of Computational Biology*, **2**(4), 573–586.

Paton N, Stevens R, Baker P, Goble C, Bechhofer S and Brass A (1999) Query processing in the TAMBIS bioinformatics source integration system. In *Proceedings of the 11th International Conference on Scientific and Statistical Database Management (SSDBM)*, Ozsoyoglo ZM (Ed.), IEEE Press: Los Alamitos, pp. 138–147.

Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W and Solomon W (1997) The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, **9**, 139–171.

Wiederhold G (1992) Mediators in the Architecture of Future Information Systems. *IEEE Computer*, **25**(3), 38–49.

Short Specialist Review

Automatic concept identification in biomedical literature

William H. Majoros
The Institute for Genomic Research, Rockville, MD, USA

1. Introduction

While the potential benefits of automatically extracting useful knowledge from biomedical literature are seemingly many and great, nearly all of the techniques that have been considered for doing this rely on a central, nontrivial problem: that of unambiguously identifying the named entities that occur in a particular text. This problem has two parts: (1) identifying the precise boundaries of phrases in a text which refer to biomedical entities and (2) resolving those entity names to known concepts in a knowledge base such as an ontology (Cohen *et al*., 2002; *see also* Article 84, **Ontologies for natural language processing**, Volume 8). These two tasks represent the *syntactic* and *semantic* aspects of the problem, respectively, and though these are not entirely independent in theory, they are often treated as such in practice.

2. Syntactic considerations

One of the computationally simplest approaches to identifying biomedical entities in text is to maintain a lexicon of all the entities of interest and to systematically search through that lexicon for all phrases of any length occurring in a given text. This can be done efficiently by utilizing a data structure based on a *hash table* or a *suffix tree*, with a single pass over the text sufficing to detect all exact matches to phrases in the lexicon. With a sufficiently large lexicon, much can be achieved with this simple procedure, such as primitive *concept-based document retrieval*, or statistical analysis of *co-occurrence patterns* in order to infer possible relationships between entities (Jenssen *et al*., 2001).

But there is also much room for improvement, since many of the named entities in the text will not be found, because either they do not occur in the lexicon or their lexicon entry differs in spelling, word order, hyphenation, or in some other insignificant way from the form that appears in the text. In theory, this shortcoming could be addressed through extensive manual curation of the lexicon so that exhaustive coverage of both concepts and their many lexical variants is

achieved. Unfortunately, this is not practical in many fields both due to the vast number of entities and lexical variants that would have to be curated, and also due to the rapid pace of advance in biomedicine and the constant generation of new concepts that results.

A more sophisticated and promising approach involves the use of *shallow parsing* to identify all noun phrases in a given text. The advantage of this approach is that entities not occurring in the lexicon can at least still be found (to the extent that an accurate parse is obtained), allowing their interpretation to be deferred until semantic information can be curated, inferred, or synthesized for them.

Unfortunately, the accurate identification of noun phrases is itself a difficult problem, especially in the field of biomedicine. A common approach is to apply a *part-of-speech tagger* to mark the nouns, adjectives, and other parts of speech, and then to apply some *chunking* procedure to coalesce appropriate sequences of tagged words into noun phrases. Various techniques have been applied with some success to these two tasks, including the use of rule-based and Markovian systems (Manning and Schütze, 2000).

Although several part-of-speech taggers have been made publicly available, these tend to perform poorly out-of-the-box on biomedical text. Because these are generally trained on generic text such as newswire articles, they have vocabularies and language models that are inappropriate for most specialized domains. As an example, many taggers upon seeing *nucleated erythrocyte* will tag *nucleated* as a verb rather than as a participle, making it very unlikely that this entity will be properly identified as a noun phrase.

Normally, this would be rectified by retraining the tagger on a large corpus of manually tagged text, but unfortunately, such a corpus does not yet exist for biomedicine, and is not likely to exist for some time. The largest publicly available annotated training corpus to date is the GENIA corpus version 3.02 (Kim *et al*., 2003), which contains a mere 2000 of the 12 million abstracts available in MEDLINE. While 2000 full-length documents might be adequate for some domains, these are abstracts only, and were not selected at random. Building a manually annotated corpus of sufficient size to adequately capture the vocabulary and statistical properties of MEDLINE would likely require an inordinate amount of labor.

For this reason, some recent work has explored alternative ways of obtaining an accurate tagger. For example, by utilizing an existing tagger together with the known noun phrases in UMLS (*see* Article 81, **Unified Medical Language System and associated vocabularies**, Volume 8), a training corpus can be automatically synthesized and then used to retrain the original tagger (Majoros *et al*., 2003). In this way, it is conceivable that other automatic or semiautomatic methods might produce useful training data for other parsing strategies.

3. Semantic considerations

Once the named entities have been located in a text, they should, ideally, be mapped then to an ontology in order to link information newly gleaned from the text with what is already known about those entities. Finding exact matches is easy enough,

given the appropriate data structures, but it is the effective detection of inexact matches that is crucial to achieving high sensitivity. This is arguably the most difficult aspect of concept identification.

An example will serve to illustrate some of the difficulties. The gene Apo3 is also variously known as LARD, DR3, TRAMP, wsl, TnfRSF12, and "lymphocyte-associated receptor of death" (Yu and Agichtein, 2003). If any of these are seen in a new text, the occurrences should be resolved to the Apo3 entry in our ontology. Moreover, if our software encounters any of the phrases "lymphocyte-associated receptor of death", "lymphocyte-associated death receptor", or TnfRSF-12, it would be desirable if these also were resolved to Apo3. Yet, to do so requires recognizing that these differences in orthography and word order are insignificant, even though the same types of differences (particularly word order) are certainly significant in other cases.

Thus, a certain amount of domain knowledge must be embedded in a procedure if it is to achieve high accuracy. Among the many features that such a heuristic needs to consider are capitalization, punctuation, spelling, word order, inflection, and the use of acronyms and abbreviations. Any one of these sources of variation might be handled with relative ease, but a complete solution needs to address all of these phenomena. As a result, many of the solutions that have been devised seem somewhat ad hoc, and though a few of them appear to work well on small test sets, it is reasonable to wonder how they will perform on other corpora and whether perhaps a more elegant solution should be sought.

Among the more promising attempts in this direction are those that utilize *machine learning* or statistical techniques (e.g., Leonard *et al*., 2002; Tanabe and Wilbur, 2002). These typically involve the identification of features, such as word morphology or surrounding context, which may be predictive of a given phrase's intended meaning. These approaches seem especially promising for attacking the semantic problem of *concept disambiguation*, in which a single term may conceivably refer to multiple concepts. For example, PSA may refer to *prostate specific antigen*, *psoriasis arthritis*, or *poultry science administration* (Weeber *et al*., 2003), yet it might be reasonably expected that nearby words (such as prostate, psoriasis, or chicken) might facilitate correct disambiguation. Unfortunately, any approach that relies on training is made difficult by the dearth of large, high-quality training sets.

One publicly available tool that appears to perform well (Pratt and Yetisgen-Yildiz, 2003) is MetaMap. This program identifies the noun phrases in a text, generates all lexical variants of all words and subphrases of the noun phrase on the basis of entries in the SPECIALIST lexicon, and then searches through the UMLS ontology for any of the generated variants. Resulting matches are numerically scored and ranked on the basis of match length and other syntactic considerations (Aronson, 2001).

Though semantic resolution of a phrase is not always possible using current methods and ontologies, the unresolved phrases may themselves be of considerable use. Such novel phrases can be appended to lists of terms awaiting manual curation, or they can also be used simply as *uncurated concepts* (perhaps after some clustering or normalization) and subjected to similar types of analyses as curated concepts.

Along a similar vein, automatic means of ontology construction and extension using uncurated resources are receiving considerable attention (e.g., Bodenreider *et al*., 2002; Yu and Agichtein, 2003) as it becomes increasingly clear that truly exhaustive, de novo manual curation will be achieved only at great expense. Most promising, at present, are semiautomatic methods that suggest additions to an existing ontology, which are then subjected to a final manual filtering step. Automated methods for identifying potentially novel synonyms of known concepts include clustering and co-occurrence analyses of known and unknown concepts within a corpus (Blaschke and Valencia, 2002) as well as similarities in phrase morphology (Pustejovsky *et al*., 2002). In the case of the latter, variant generation procedures similar to those employed by MetaMap can be used to suggest not only synonyms but also prospective hyponyms in the case of known concepts which are observed with novel modifiers in a corpus, such as *erythrocyte* and its hyponym *nucleated erythrocyte.* More sophisticated approaches are possible which utilize existing knowledge to infer novel relationships. For example, if *adipocyte differentiation* occurs in a corpus but is absent from the ontology, then we might infer that it is a hyponym of the known concept *cell differentiation*, given that *adipocyte* is a known hyponym of *cell*.

4. Outlook for the future

Much room remains for progress on virtually all fronts of the concept-identification problem. In order to enable further advances, more effort needs to be invested in producing large, diverse, high-quality annotated corpora and ontologies, both for training concept-identification programs, and to provide standard test sets for comparisons. The availability of open-source software will likewise accelerate advancement in this field as researchers are enabled to perform more controlled experiments comparing different algorithms within the same software configuration.

References

Aronson AR (2001) Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings of the American Medical Informatics Association Symposium 2001*, Washington, DC, 17–21.

Blaschke C and Valencia A (2002) Automatic ontology construction from the literature. *Genome Informatics*, **13**, 201–213.

Bodenreider O, Rindflesch TC and Burgun A (2002) Unsupervised, Corpus-based Method for Extending a Biomedical Terminology. *Proceedings of the Association of Computational Linguistics 2002 Workshop on Natural Language Processing in the Biomedical Domain*, 53–60.

Cohen KB, Dolbey AE, Acquaah-Mensah GK and Hunter L (2002) Contrast and Variability in Gene Names. *Proceedings of the Association of Computational Linguistics 2002 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, PA, 14–20.

Jenssen TK, Laegreid A, Komorowski J and Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**, 21–28.

Kim J-D, Ohta T, Tateisi Y and Tsujii J (2003) GENIA corpus – a semantically annotated corpus of bio-textmining. *Bioinformatics*, **19**, 180–182.

Leonard JE, Colombe JB and Levy JL (2002) Finding relevant references to genes and proteins in MEDLINE using a Bayesian approach. *Bioinformatics*, **18**, 1515–1522.

Majoros WH, Subramanian GM and Yandell MD (2003) Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics*, **19**, 402–407.

Manning D and Schütze H (2000) *Statistical Natural Language Processing*, MIT Press: Cambridge, MA.

Pratt W and Yetisgen-Yildiz M (2003) A study of Biomedical Concept Identification: MetaMap vs. People. *Proceedings of the American Medical Informatics Association Symposium 2003*, Washington, DC, 529–533.

Pustejovsky J, Rumshisky A and Castaño J (2002) Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics. *Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases 2002*, Las Palmas, Canary Islands, Spain, 60–68.

Tanabe L and Wilbur WJ (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**, 1124–1132.

Weeber M, Schijvenaars JA, van Mulligen EM, Mons B, Jelier R, van der Eijk C and Kors JA (2003) Ambiguity of human gene symbols in LocusLink and MEDLINE: Creating an inventory and a disambiguation test collection. *Proceedings of the American Medical Informatics Association Symposium 2003*, 704–708.

Yu H and Agichtein E (2003) Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, **19**, 340–349.

Short Specialist Review

Merging and comparing ontologies

Natalya F. Noy
Stanford University, Stanford, CA, USA

The use of formal systems to define biomedical concepts and to represent and store biomedical knowledge has never been more important to the informatics community. The past decade of research has seen a growing consensus regarding the critical nature of controlled terminologies and well-defined ontologies in the construction of all information systems. From the implementation of hospital information systems to the organization of experimental data for research in bioinformatics, developers now identify the key issue to be the manner in which salient concepts are labeled and defined, and ultimately used computationally.

In the past few years, ontologies have become the cornerstone of the information-systems research. While there is no single agreed-upon definition of an ontology, the one that is cited most frequently is the definition by T. Gruber (1993): "ontology is an explicit formal specification of conceptualization." For the purposes of this article, we define an ontology as a set of concepts that categorize an application area (classes), their attributes, relations between concepts, and constraints on these relations. There is also often instance data associated with an ontology, which are individual examples of these concepts with some or all of the attributes and relations specified.

With the wider use of ontologies, the problem of *ontology management* began to arise: ontology developers and users must be able to find and compare existing ontologies or their parts, merge and align ontologies, compare different versions, and translate ontologies from one formalism to another. Only recently have researchers started to address these problems, in large part because until ontologies came into widespread use, their management was not a serious issue.

It is impractical to assume that eventually there will be one single set of standard ontologies that everyone will conform to. In fact, experiences, even in such a mature field as industrial databases, show that having a small set of standard schemas and ontologies is still unattainable: it is not uncommon for a single large enterprise to use more than a dozen database schemas for purchase orders, for example. There are many reasons for someone to develop his own ontology or schema rather than to reuse an existing one. These reasons range from proprietary and institutional reasons (e.g., an institutional requirement to use only components developed in-house) to more practical reasons: if the task is not a standard task, it may be easier

to develop a custom-tailored ontology just for that task than to reuse and adapt an existing one.

We start by considering various ontology-management tasks and the corresponding tools separately (Klein, 2001). Systems such as PROMPT (Noy and Musen, 2000) and Chimaera (McGuinness *et al*., 2000) support *ontology merging* – creating a single coherent ontology from source ontologies that covered similar domains. Systems such as ONION (Mitra *et al*., 2000) and GLUE (Doan *et al*., 2002) look at the issue of finding *mappings* between concepts in different ontologies. Researchers in the SHOE group (Heflin and Hendler, 2000) worked on approaches to *ontology versioning*, defining what it means for a new version of ontology to be backward compatible with the old one.

This research has produced a plethora of interesting approaches and heuristics for finding differences and similarities between ontologies. In fact, as we discussed in our earlier review of the field (Noy and Musen, 2002b), approaches are so different and rely on such different assumptions about the source ontologies that it is hard to compare them. The user must choose the appropriate set of tools on the basis of the specific type of ontology and data that he has at hand. For instance, some approaches relied on instance data to find correspondences between concepts in different ontologies. They use the values in the instance data to infer correspondences between classes of these instances. Thus, these approaches can be used only when there was instance data corresponding to the ontologies. However, since ontologies are often developed to describe the domain and used directly in an application or used as a controlled terminology to annotate other ontologies and data, instance data is not always available. If instance data is available, however, approaches that use it to find correspondences are very beneficial.

Other approaches consider class definitions to find which concepts in one ontology correspond to concepts in another. First, one can lexically compare class names and infer that classes with similar names probably correspond to similar concepts. This approach, however, is not without peril since even if ontologies covered the same domain of discourse, classes with similar names may actually describe different concepts. Consider a class University for example. It could mean a university campus, or university as an organization with its departments, faculty, and so on. Furthermore, ontology designers often use different words or combinations of words to mean the same concept. This – approaches that looked up dictionaries and thesauri – proved useful in finding these types of similarities.

The next generation of tools that had to compare two ontologies used more than just names of classes, and used the structure of class definitions to find parallels. For instance, if two classes had similar names but also a similar set of properties and superclasses, then they were more likely to be similar than two classes, where only the names correlated. Further, tools used the input provided by the users of interactive merging and mapping tools in their analysis as well. For instance, if our algorithm already suggested that two classes were similar and then the user explicitly said that their superclasses are the same, our confidence in the suggestion is much stronger.

Finally, a number of algorithms looked at ontologies as graphs and compared the graphs looking for similar subgraphs indicating similar concepts (see e.g.,

ANCHORPROMPT (Noy and Musen, 2001) or Similarity Flooding (Melnik *et al*., 2002)).

There have been fewer tools for ontology versioning than one might expect, given the collaborative nature of ontology development. Tools for managing versions of software code, such as CVS (**http://www.cvshome.org**), have become indispensable for software engineers participating in dynamic collaborative projects. These tools provide a uniform storage mechanism for versions, the ability to check out a particular piece of code for editing, an archive of earlier versions, and mechanisms for comparing versions and merging changes and updates.

Ontologies change just as a software code does. These changes result from changes in the domain itself (our knowledge about the domain changes or the domain itself changes) or in the conceptualization of the domain (we may introduce new distinctions or eliminate old ones). Furthermore, ontology development in large projects is a dynamic process in which multiple developers participate, generating different versions of an ontology. Naturally, collaborative development of dynamic ontologies requires tools that are similar to software-versioning tools. In fact, ontology developers can use the storage, archival, and checkout mechanisms of tools such as CVS with very little change.

There is one crucial difference, however: Comparison of versions of software code entails a comparison of text files. Program code is a set of text documents and the result of comparing the documents – the process is called a *diff* – is a list of lines that differ in the two versions. This approach does not work for comparing ontologies: two ontologies can be exactly the same conceptually, but can have very different text representations. For example, their storage syntax may be different. The order in which definitions are introduced in the text file may be different. A representation language may have several mechanisms to express the same thing. Therefore, text-file comparison is largely useless in comparing versions of ontologies. As a result, tools such as OntoVIew (Klein *et al*., 2002) and PromptDiff (Noy and Musen, 2002a) were developed. The former enables ontology developers to specify conceptual details of how exactly the concept definitions have changed (e.g., a concept became more general).

The latter uses a set of heuristics to find which parts of an ontology remained unchanged and which were added, deleted, or redefined. The crucial observation on which it relies is that, unlike with different ontologies, in versions of the same ontology, terms with similar names are much more likely to mean the same thing. For instance, a class University in two versions of the same ontology is unlikely to refer to the campus in one and organization on the other. The heuristics that PromptDiff uses enable it to achieve remarkable accuracy in finding differences between the two versions.

Interestingly, each of the tools that we have described addressed only one of the ontology-management problems: ontology merging, ontology mapping, or comparison. As more of these tools and algorithms appeared, however, it became clear that most ontology-management tasks share the same problems and if we could solve one of them, we could have great insights for solving another. Essentially, all of these tasks require finding similarities and differences between two ontologies. When we are merging two ontologies, for example, we need to merge the parts that are similar and just bring along the remaining, different parts (making sure our

merged ontology is still logically consistent, of course). When we need to translate data from one ontology to another, we again need to know what the similarities are and translate instance data corresponding to a class in one ontology to the instance data to the class that is similar to it in another ontology. The same is true for answering queries posed in terms of one ontology with data from another ontology: we need to look for similar concepts in the other ontology. If we are comparing two versions of the same ontology, we are interested in examining the differences rather than the similarities, but this problem can be treated as the complement of finding similarities, and it is certainly amenable to the same techniques.

P. Bernstein and colleagues capitalized on this observation, developing a general theory of model management (Bernstein *et al*., 2000). At the center of the theory is a notion of a *mapping*, which can be applied to structured models of different kinds (not only ontologies but also XML schemas, database schemas, etc.). The mappings are first-class objects that can then be used for tasks such as merging, data transformation, or query answering. Similarly, Noy and colleagues (Noy and Musen, 2004) suggested a generic framework for ontology-management tasks. The same algorithms are used for ontology merging, mapping, and comparison, but with different thresholds for confidence values.

The cross-fertilization between techniques used for different ontology-management tasks proved very productive, and we believe that in the future this general approach to various tasks involving several ontologies will become the foundation of a new generation of sophisticated tools.

References

Bernstein PA, Halevy AY and Pottinger RA (2000) A Vision for Management of Complex Models. *SIGMOD Record*, **29**, 55–63.

Doan A, Madhavan J, Domingos P and Halevy A (2002) *The Eleventh International World-Wide Web Conference, Hawaii*.

Gruber TR (1993) A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, **5**, 199–220.

Heflin J and Hendler J (2000) *Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin*.

Klein M (2001) *IJCAI-2001 Workshop on Ontologies and Information Sharing, Seattle*, pp. 53–62.

Klein M, Kiryakov A, Ognyanov D and Fensel D (2002) *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), Sigüenza*.

McGuinness DL, Fikes R, Rice J and Wilder S (2000) In *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000)*, Cohn AG, Giunchiglia F and Selman B (Eds.), Morgan Kaufmann Publishers: San Francisco, pp. 483–493.

Melnik S, Garcia-Molina H and Rahm E (2002) *18th International Conference on Data Engineering (ICDE-2002), IEEE Computing Society: San Jose*.

Mitra P, Wiederhold G and Kersten M (2000) *Proceedings Conference on Extending Database Technology 2000 (EDBT'2000), Konstanz*.

Noy NF and Musen MA (2000) *Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin*.

Noy NF and Musen MA (2001) *Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle*, pp. 63–70.

Noy NF and Musen MA (2002a) *Eighteenth National Conference on Artificial Intelligence (AAAI-2002), Edmonton*.

Noy NF and Musen MA (2002b) *Workshop on Evaluation of Ontology Tools at EKAW'02 (EON2002), Siguenza*.

Noy NF and Musen MA (2004) Ontology Versioning in an Ontology-Management Framework. *IEEE Intelligent Systems*, **19**(4), 6–13.

Short Specialist Review

Bioinformatics pathway representations, databases, and algorithms

Peter D. Karp

Bioinformatics Research Group, Menlo Park, CA, USA

1. Introduction

Gene products affect living systems through carefully orchestrated cascades of molecular interactions known as pathways. For example, metabolic pathways transform small molecules through a sequence of enzyme-catalyzed reactions that produce energy and that synthesize the chemical building blocks of the cell. Signaling pathways are sequences of interactions among proteins that enable a cell to respond to changes in its environment, such as by altering gene expression.

Management of pathway information is of great importance in bioinformatics because pathways place gene products in a causal, mechanistic context showing how those gene products contribute to a larger biological process, and allow us to reason about perturbations to a living system. In the postgenomic era, pathways are also an important organizational and information-reduction device that decreases the cognitive complexity of a biological system by allowing us to reason about a system at the level of pathways instead of at the level of proteins.

Management of pathway data is a challenging area of bioinformatics because it involves a host of representational and algorithmic problems.

2. Pathway representations

We define a pathway as a cascade of biochemical interactions. Thus, interactions (biochemical reactions) are the building blocks of pathways, in the sense that pathways are built by linking interactions together in sequence. Molecules are the building blocks of interactions, in the sense that interactions are built from the molecules that comprise their reactants and products. For metabolic pathways, those molecules are small-molecular-weight chemical compounds, whereas for signaling pathways, those molecules are often proteins.

The pathway representation developed for the EcoCyc database (DB) (Karp, 2000) takes literally this three-tiered layering in the sense that pathways, interactions, and molecules are all represented as separate DB objects in the EcoCyc

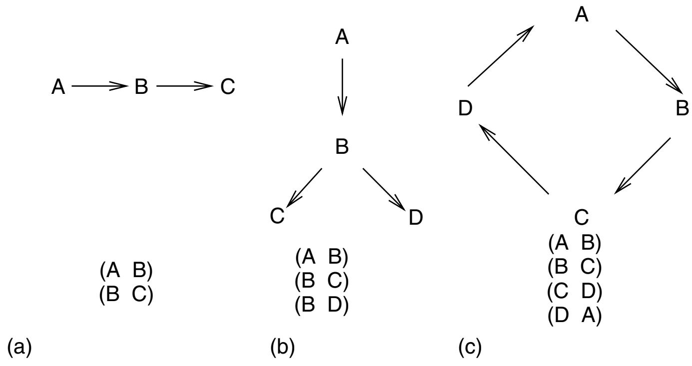


Figure 1 This figure shows three simple pathways, and the corresponding predecessor lists used to encode them (below)

representation, and pathways are built from interactions, which in turn are built from molecules. In addition to facilitating the three-tiered representation, this separation produces a nonredundant representation in which the data about a single chemical compound is represented only once in the DB, even though that chemical compound might be a reactant or a product of many different chemical reactions. Similarly, a given reaction is represented once in the DB even though it might be a component of several different pathways. Nonredundancy is a central principle of DB design because it minimizes data entry and ensures that updates to information about a given object need occur in only one place.

The simplest form of a pathway is a linear cascade – a sequence of biochemical events in which the output of the first event becomes the input of the next event. In a linear metabolic pathway, for example, consecutive steps in a pathway share a common small-molecule substrate, such as molecule B shown in Figure 1(a). Pathways come in a variety of topologies including linear pathways, tree-structured pathways, circular pathways, and combinations thereof.

Karp and Paley developed a representation called the *predecessor list* that allows us to represent the topology in which the interactions that comprise a pathway are linked. The predecessor list specifies the ordering between all adjacent pairs of interactions within a pathway. It is a list of tuples, where each tuple (A B) specifies that reaction A is the predecessor of reaction *B* in the pathway, that is, reaction B follows reaction A in the pathway, and that the two reactions share a common substrate. Figure 1 shows the predecessor lists used to represent linear, circular, and tree-structured pathways.

Each reaction within a pathway is represented as a separate object whose properties encode lists of reactants and substrates for the reaction, and properties such as the EC number and equilibrium constant of the reaction. Each substrate is represented as an object whose properties can include the chemical formula and chemical structure of the substrate.

Pathway representations should capture the many-to-many relationship between enzymes and the reactions they catalyze. Reactions and enzymes must be separate objects within the DB to ensure that the properties of the reaction are captured separately from the properties of the protein (such as its subunit structure and

physicochemical properties). Because a reaction can be catalyzed by multiple isozymes, and because an enzyme can catalyze multiple reactions, there must be an explicit mapping between each enzyme and the reaction(s) it catalyzes.

The definition of pathways is a useful mechanism for dividing a complex network of metabolic reactions into more understandable units. Pathways have biological meaning because they are often regulated as a unit, their genes are often grouped together on the chromosome as a unit in bacteria, and they are conserved through evolution as units. However, in some computational contexts it can be useful to ignore the pathways defined in an organism and compute across the metabolic network as simply a set of interconnected reactions. The preceding representation completely supports this viewpoint since the reaction set of an organism can be queried independently of the pathways to which those reactions have been assigned.

Petri nets provide another formalism for representing networks of biochemical reactions (Peterson, 1981; Peleg *et al*., 2002; Mandel *et al*., 2004). Petri nets are a directed graph–based representation in which two types of nodes are defined: places (corresponding to reaction substrates) and transitions (corresponding to reactions). Vertices of the graph connect places and transitions to define the reactions that operate on each substrate. Each place contains a state descriptor consisting of a number of tokens, which can represent the concentration of the corresponding substrate.

A number of computational operations have been defined on Petri nets including modeling the flow of material through a network, and determining properties of a network such as liveness (all transitions can occur), boundedness (the number of tokens at every place will be bounded by some ceiling), and reachability (can some final state of the network be reached from a starting state of the network?).

Note that there is a straightforward mapping from the first representation described above to Petri nets – each substrate maps to a place and each reaction maps to a transition – allowing the algorithms developed for Petri nets to be applied to databases structured using the preceding representation.

3. Pathway databases

SRI's BioCyc collection of pathway DBs includes 16 DBs that describe the metabolic pathways of specific organisms, and the MetaCyc DB of 500 experimentally determined metabolic pathways from more than 200 organisms.

The BioCyc philosophy is that we should model the metabolic network of each distinct organism of interest in a separate organism-specific DB to provide maximum accuracy. These organism-specific DBs are created in two phases: in Phase 1, we predict the metabolic network using the PathoLogic algorithm. In Phase 2, this computational prediction is refined through a literature-based curation process in which new pathways are manually added to the DB, and predicted pathways are edited to reflect organism-specific variations in the pathways.

The computational pathway prediction process depends on the MetaCyc DB, which provides a reliable reference set of experimentally defined metabolic pathways. MetaCyc enzymes and pathways are all derived from the experimental literature, and are curated with copious comments and citations to 4550 literature references.

Other pathway DBs include KEGG (Kanehisa *et al*., 2002; *see also* Article 101, **Design of KEGG and GO**, Volume 8), WIT (Overbeek *et al*., 2000), and Amaze (van Helden *et al*., 2001).

BioPAX is a standard for encoding and interchange of pathway data. It describes a common ontology for pathway data, and supports encoding of pathway data within the XML-based OWL language, which is a basis for the Semantic Web. See **http://www.biopax.org/**. The Systems Biology Markup Language (SBML) is a standard for representing models of biochemical reaction networks (see **http://sbml.org/**). SBML is applicable to metabolic networks, cell-signaling pathways, regulatory networks, and many others.

The BioPathways Consortium (see **http://BioPathways.org**) is dedicated to catalyzing the development of pathway bioinformatics. The consortium organizes workshops and presentations at a number of bioinformatics and biology conferences.

4. Pathway algorithms

Pathway algorithms are a core component of Systems Biology. Systems Biology seeks to understand the workings of a biological system from the interactions of its components, and many of those interactions occur through pathways. It is instructive to differentiate Symbolic Systems Biology from Quantitative Systems Biology. The latter is concerned with qualitative questions that can be answered through symbolic computations, such as: What biochemical reactions and pathways are present in an organism? How can we fill holes identified in that network? What end products can be produced by the network given a specified set of input compounds?

Quantitative Systems Biology is concerned with quantitative computations such as predicting equilibrium fluxes through a metabolic network, or predicting the time course of behavior of a biological network.

4.1. Pathway prediction

Pathway prediction algorithms predict the existence of pathways within an organism. Prediction by similarity predicts the presence in one organism of pathways with the same reaction topology as pathways previously known in other organisms. For example, the PathoLogic algorithm predicts the presence of pathways from the MetaCyc DB in an organism whose genome has been sequenced (Paley and Karp, 2002). The essence of the algorithm is to match enzyme activities identified in the genome (matching via both enzyme name and EC number) to enzymes in MetaCyc pathways. Pathways for which a significant number of enzymes are present are predicted to be present in the organism being analyzed.

De novo prediction algorithms predict the existence of pathways that were not previously known in other organisms. The algorithm of Shah *et al*. (McShan *et al*., 2003) predicts pathways that synthesize a user-specified product metabolite from a user-specified ending metabolite. The algorithm considers known reactions from a supplied reaction DB, and uses an A^* search to find optimal pathways. The method defines a search space whose nodes are chemical compounds and whose edges are

known reaction transformations. The cost function used to evaluate pathways is based on chemical differences between the intermediate compounds.

4.2. Pathway hole filling

Pathway networks predicted from genome information are typically incomplete in the sense that some of the reactions in some pathways have no enzyme identified in the genome that catalyzes that reaction. The algorithm of Green *et al*. generates hypotheses as to which genes in the genome code for enzymes that catalyze these "pathway hole" reactions (Green and Karp, 2004). The algorithm works by retrieving protein sequences for other enzymes that catalyze that same biochemical reaction in another species, and then BLASTing those sequences against the genome to generate candidate genes. The candidates are evaluated with a Bayesian network that generates a probability score for the correctness of each candidate.

4.3. Flux-balance analysis

Once a metabolic network has been reconstructed for a genome, it is possible to compute a number of aggregate properties of the network.

Computational flux-balance analysis (FBA) (Kauffman *et al*., 2003; Segre *et al*., 2003) can predict the equilibrium fluxes through each reaction in the network. The method works by defining a set of constraints across the whole metabolic network and then solving an optimization problem across those constraints. The constraints typically specified in FBA models include mass balance across the entire metabolic network (i.e., the sum of all reaction fluxes cannot yield a net change in chemical mass), optimization of cell growth rate given the available biomass, and constraints on the chemical composition of the cell based on past measurements of that composition. FBA can not only produce accurate predictions of reactions fluxes, but in many cases can predict which enzyme-coding genes are essential for cell growth.

4.4. Forward propagation through a metabolic network

An important type of qualitative question regarding a metabolic network is: given a specified set of compounds as inputs to the network, what outputs can the network produce? Note that we have not asked about the rates at which these outputs are produced (which requires extensive knowledge of quantitative network parameters), but simply: what outputs *can be produced?* This calculation can serve as a validation of the reconstructed network if the given input is a known growth medium for the organism, in which case the reconstructed network should be able to produce all small-molecule compounds essential for life (including the building blocks of protein, DNA, RNA, etc.).

Intuitively, an algorithm can solve this problem by determining whether a path can be traced through the metabolic network from the starting compounds to the output compounds. If a compound is not reachable through the web of metabolic

reactions, then the cell will be unable to produce that compound. But strictly speaking, reachability through the network is not a sufficient condition for the algorithm to evaluate because a biochemical reaction requires *all* of its reactants to be present for that reaction to be active. Therefore, the algorithm devised by Romero and Karp (2001) converts a set of biochemical reactions to a *production-rule system* and converts the set of input metabolites to the initial axiom set for that production system. A reaction of the form $A + B \rightarrow C$ is converted to a production rule of the form $A \wedge B \supset C$. The algorithm repeatedly searches for rules whose left sides are satisfied and fires a rule by adding the species on the right side of the rule to the current pool of metabolites. For example, the preceding rule would fire when A and B are present in the metabolite pool, adding C to the metabolite pool. A and B would be present in the metabolite pool if they were either in the initial axiom set, or they were created by the firing of a production rule. This algorithm was able to verify that the *E. coli* metabolic network modeled by the EcoCyc DB can produce a set of 40 metabolic building blocks, given a set of input metabolites that constitutes a known growth medium for *E. coli*.

4.5. *Simulation of biochemical systems*

In recent years, a large number of software tools have been developed for simulation of biochemical systems (Pettinen *et al*., 2004; *see also* Article 102, **Simulation of biochemical networks**, Volume 8). The Bio-SPICE (Garvey *et al*., 2003) and Systems Biology Workbench (SBW) (Sauro *et al*., 2003) projects have developed software architectures for integrating these tools into a smoothly interoperating collection that supports construction of simulation models, execution of those models by solver programs, and interpretation of results. For example, the Bio-SPICE Dashboard uses a distributed services discovery and invocation mechanism to permit data sources, models, simulation engines, and output displays provided by different investigators and running on different machines to work together across a network.

Tools supporting graphical interactive construction of simulation models allow users to define the reactions and pathways within a differential-equation based model, and the substrate concentrations and enzyme parameters for the model. Example tools include JDesigner (Sauro, 2001) and Kinetikit (Bhalla, 2002).

Model-execution systems simulate the time-series behavior of a model using a numeric integration package. Example simulators include Jarnac (Sauro, 2000), Gepasi (Segre *et al*., 1997), E-CELL (Tomita *et al*., 1999), and the JigCell Run Manager (Allen *et al*., 2003). Some systems include tools for estimating values of unknown parameters – Gepasi has a particularly comprehensive collection.

Interpreting the results of a simulation and comparing those results with other simulations or with experimental observations is a subject for other tools, such as the JigCell Comparator (Allen *et al*., 2003) and the E-CELL user interface.

Acknowledgments

This work was supported by grants GM70065 from the NIH National Institute for General Medical Sciences and HG02729 from the NIH National Human Genome

Research Institute. The contents of this article are solely the responsibility of the author and do not necessarily represent the official views of the National Institutes of Health.

References

Allen N, Calzone L, Chen K, Ciliberto A, Ramakrishnan N, Shjaffer C, Sible J, Tyson J, Vass J, Watson L, *et al*. (2003) Modeling regulatory networks at Virginia Tech. *OMICS A Journal of Integrative Biology*, **7**(3), 285–300.

Bhalla U (2002) Use of kinetikit and genesis for modeling signaling pathways. *Methods in Enzymology*, Academic Press: Newyork, pp. 3–23.

Garvey T, Lincoln P, Pedersen C, Martin D and Johnson M (2003) BioSPICE: Access to the most current computational tools for biologists. *OMICS A Journal of Integrative Biology*, **7**(4), 411–420.

Green M and Karp P (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**(1), 76. **http://www.biomedcentral.com/1471-2105/5/76**.

Kanehisa M, Goto S, Kawashima S and Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Research*, **30**, 42–46.

Karp PD (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**(3), 269–285.

Kauffman K, Prakash P and Edwards J (2003) Advances in flux balance analysis. *Current Opinion in Biotechnology*, **14**, 491–496.

Mandel J, Palfreyman N, Lopez J and Dubitzky W (2004) Representing bioinformatics causality. *Briefings in Bioinformatics*, **5**, 270–283.

McShan D, Rao S and Shah I (2003) Pathminer: Predicting metabolic pathways by heuristic search. *Bioinformatics*, **19**(13), 1692–1698.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, Kyrpides N, Fonstein M, Maltsev N and Selkov E (2000) WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, **28**, 123–125.

Paley S and Karp P (2002) Evaluation of computational metabolic-pathway predictions for *H. pylori*. *Bioinformatics*, **18**(5), 715–724.

Peleg M, Yeh I and Altman R (2002) Modelling biological processes using workflow and petri net models. *Bioinformatics*, **18**, 825–837.

Peterson J (1981) *Petri Net Theory and the Modelling of Systems*, Prentice-Hall.

Pettinen A, Aho T, Smolander O, Manninen T, Saarinen A, Taattola K, Yli-Harja O and Linne M (2004) Simulation tools for biochemical networks: Evaluation of performance and usability. *Bioinformatics*, **21**, 357–363.

Romero P and Karp P (2001) Nutrient-related analysis of pathway/genome databases. In *Proceedings of the Pacific Symposium on Biocomputing*, Altman R and Klein T (Eds.), World Scientific: Singapore, pp. 471–482.

Sauro H (2000) Jarnac: A system for interactive metabolic analysis. *Animating the Cellular Map: Proceedings of the 9th International Meeting on BioThermoKinetics*, Stellenbosch University Press, pp. 221–228.

Sauro H (2001) *JDesigner: A Simple Biochemical Network Designer*, Technical report, **http://members.tripod.co.uk/sauro/biotech.htm**.

Sauro H, Hucka M, Finney A, Wellock C, Bolouri H, Doyle J and Kitano H (2003) Next generation simulation tools: The systems biology workbench and BioSPICE integration. *OMICS A Journal of Integrative Biology*, **7**(4), 355–372.

Segre D, Zucker J, Katz J, Lin X, D'Haeseleer P, Rindone W, Kharchenko P, Nguyen D, Wright J and Church G (1997) Biochemistry by numbers: Simulation of biochemical pathways with Gepasi 3. *Trends in Biochemical Sciences*, **9**, 361–363.

Segre D, Zucker J, Katz J, Lin X, D'Haeseleer P, Rindone W, Kharchenko P, Nguyen D, Wright J and Church G (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS A Journal of Integrative Biology*, **7**(3), 301–316.

Tomita M, Hashimoto K, Takahashi K, Shimizu T, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter J, *et al*. (1999) E-CELL: Software environment for whole-cell simulation. *Bioinformatics*, **15**(1), 72–84.

van Helden J, Naim A, Lemer C, Mancuso R, Eldridge M and Wodak S (2001) From molecular activities and processes to biological function. *Briefings in Bioinformatics*, **2**, 81–93.

Short Specialist Review

Proteomics data representation and management

Norman W. Paton, Andrew Jones and Stephen G. Oliver
University of Manchester, Manchester, UK

1. Introduction

Bioinformatics tools and techniques depend directly or indirectly upon experimental data. However, interpretation of experimental data often requires access to significant amounts of additional information about the sample used, the conditions in which measurements were taken and the laboratory equipment used. Recent proposals for models that capture such experimental descriptions alongside results include MAGE-ML for transcriptome data (Spellman *et al*., 2002) and PEDRo for proteome data (Taylor *et al*., 2003). The description of a proteomics investigation is challenging, as the range of available experimental techniques is both extremely broad and rapidly evolving. Therefore, although there are well-established data resources for reference maps (Hoogland *et al*., 2004), such resources tend not to contain enough information to enable scientists to validate the techniques used to produce a result, or to allow for the comparison of results from different laboratories. In addition, the cost of producing such large data sets is very high because the experiments rely upon complex instruments that require substantial technical expertise. Therefore, it is imperative that experimental data are available for reanalysis and querying in various ways. While experimental results can be analyzed in-house, often in a labor-intensive manner, the development of bioinformatics techniques for archiving, sharing, and wider exploitation of proteomics results is still in its infancy.

2. Proteomics data sources and users

Establishing the most appropriate kinds of data to include in a proteomics database is not straightforward, as this depends on the use that is to be made of the data. In a large data repository, users may want to search for results based on widely varying criteria – for example, the proteins identified, the change in the level of a protein over time, the mechanism by which a sample was studied, and so on. Furthermore, the users of a proteome data repository may themselves be diverse,

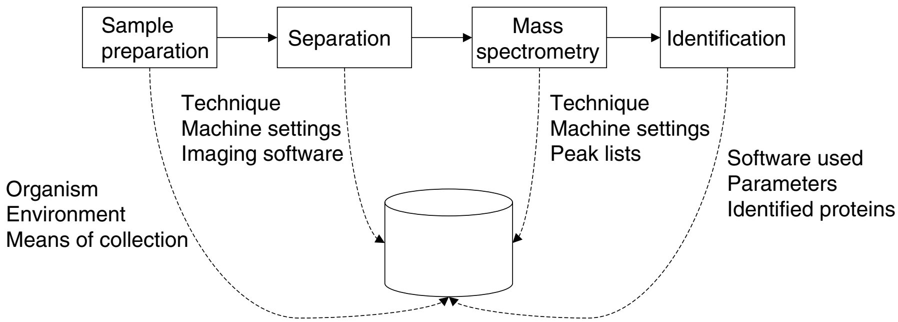


Figure 1 The stages in a proteomics experiment, plus the associated data

and include experimentalists with minimal direct experience of proteomics, but who are interested in proteins or organisms for which proteome studies have been conducted; proteome scientists who want to identify how successful specific techniques have been in different contexts; or mass-spectrometric analysts who want to compare their results with those of others.

Figure 1 illustrates the components of a typical proteomics experiment. The different components of such a pipeline give rise to data that must be captured in a systematic manner if it is to be readily interpreted, shared, or analyzed. The kinds of information emerging from each aspect of the experiment are:

1. *Sample preparation*: The starting sample is any kind of biological material, from a single organelle, up to cell cultures, tissues, organs, or whole organisms (*see* Article 2, **Sample preparation for proteomics**, Volume 5). The sample description must be captured because the proteome is highly dynamic and dependent upon the environmental conditions and the genotype of the organism. A complex mixture of proteins is extracted from the sample and dissolved in solution. The experimental procedure for protein extraction is highly variable and depends upon the type of sample.
2. *Separation*: The principle of separation is to make use of the differences of a particular property between individual protein species, such as the molecular weight (MW), charge, or hydrophobicity. Details of this process affect which proteins may be detected in a sample, and thus are crucial to the interpretation of experimental results. The techniques of one-dimensional or two-dimensional gel electrophoresis (1-DE/2-DE) utilize an electric current to resolve individual proteins into single bands by MW (1-DE) or into discrete spots by MW and charge (2-DE) on a gel (*see* Article 29, **Two-dimensional gel electrophoresis (2-DE)**, Volume 5). The separation stage can produce quantitative measures of protein volume. For example, a gel can be scanned, and image-processing software can estimate the relative volume of a particular protein between different conditions by comparing the corresponding spots across multiple gels. Liquid chromatography (LC) separates proteins in a column filled with a solution through which proteins pass in a variable length of time. Two sequential LC stages can be coupled to resolve proteins at a high definition. Gel spots or column fractions can be treated with a protease that digests proteins into peptides, which

can be analyzed by mass spectrometry for protein identification. An alternative approach in proteomics is to cleave the entire protein mixture into peptides as a first stage. The peptides are then separated by two-dimensional liquid chromatography and applied directly to tandem mass spectrometry (described below). This approach can be used for high-throughput analysis because there are very few manual handling stages.

3. *Mass spectrometry*: The output from mass spectrometry (MS) can be used to identify proteins. The pattern of molecular weights of a set of peptides can often uniquely identify a protein following its digestion with a protease: *the peptide mass fingerprint* (*PMF*). In other cases, individual peptides are collected and cleaved into fragments that allow their amino acid sequence to be determined (tandem MS or MS/MS). A mass spectrometer can have a number of input parameters that must be captured. MS produces raw data, which is a list of X/Y coordinates, also known as a peak list. The peak list can be processed automatically or manually edited to remove noise from the data.
4. *Identification*: There are a number of database search applications that identify proteins on the basis of MS data. The peak list is entered, and the software has a number of optional parameters that affect the results. The output includes the peptides within each sequence that have been matched by their molecular weight (PMF) or by their partial or complete amino acid sequence (MS/MS). There is usually a statistical score to indicate the likelihood of correct protein identification but the score is not standardized across different applications. Therefore, the software used, the input parameters and the version of the database searched must be captured to allow the identification to be externally verified.

Several of these individual sources of data are challenging to model because of the diversity or complexity of the techniques used. For example, there are a number of software applications used for database searches and for gel image analysis, which produce nonstandard output values. The phenotype of a sample is particularly difficult to model because the source of a sample and its environmental history is effectively unlimited. There are also frequent developments in technology for protein separation (Washburn *et al*., 2001) and for quantifying the volumes of proteins on a large scale (Gygi *et al*., 1999).

3. Proteomics data management

Given an understanding of the data that must be captured, there is still a question as to the nature of the data management systems and tools that can most effectively be used to capture, store, and manipulate the experimental data. In some cases, a representation can be developed to address a specific task. For example, the molecular interaction format (Hermjakob *et al*., 2004) developed by the Proteomics Standards Initiative (PSI) (*see* Article 61, **Data standardization and the HUPO proteomics standards initiative**, Volume 7) of the Human Proteome Organization (HUPO) was designed principally to support the exchange of data between different protein-interaction databases (*see* Article 44, **Protein interaction**

databases, Volume 5). As such, an implementation of the representation using XML (**http://www.w3.org/XML**) provides direct support for sharing of interaction data, while leaving open the question as to how the data should be stored for analysis purposes. Another example of an XML representation for proteomics data is mzXML (Pedrioli *et al*., 2004), which is used as an intermediate representation for data from mass spectrometers, which has traditionally been represented using proprietary formats. The mzXML representation is associated with a library of format conversion and analysis programs, which tend to assume that the XML data is stored as files in a file system. However, such representations support less flexible access than database systems, and PEDRoDB (Garwood *et al*., 2004) provides an early example of a repository in which an XML database is used to support storage and searching of proteomics data directly in XML. An advantage of using an XML database system is that there is no need for complex mappings to be developed and maintained from XML into alternative data models, such as relational databases. A corresponding disadvantage is that there is less community experience relating to the management of large XML repositories, and current XML databases may not be able to match relational databases in terms of data security and query performance.

As such, there is a need to identify the role that a proteomics repository is to fulfill, and to select a model and information management infrastructure to match these needs. Objectives for a proteomics data repository could be:

1. To allow the results of different experiments to be analyzed and compared.
2. To enable queries at the level of individual proteins, biochemical pathways, tissues, or whole organisms.
3. To allow the suitability of experiment design and implementation decisions to be assessed.
4. To allow protein identifications to be rerun in the future with new databases or software.
5. To include sufficient detail about the equipment and settings used to enable replication of an experiment.

Different proteome data management systems can be expected to support different levels in the above, depending on their anticipated user requirements and the resources available for data capture and storage. For example, SWISS-2DPAGE (Hoogland *et al*., 2004) currently focuses on (1) and (2), and PEDRoDB (Garwood *et al*., 2004) seeks to support (1) to (4). As integrative studies that produce data at the levels of transcriptome, proteome, and metabolome become more common, it will become essential that the data management systems for all three types of data incorporate a uniform system to describe sample origins.

Acknowledgments

This work is supported by the BBSRC Proteomics and Cell Function Initiative, whose support we are pleased to acknowledge.

References

Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, *et al*. (2004) PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, **51**, 68.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **17**, 994–999.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al*. (2004) The HUPO PSI's molecular interaction format-a community standard for the representation of protein interaction data. *Nature Biotechnology*, **22**, 177–183.

Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF and Appel RD (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, *et al*. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, **22**, 1459–1466.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, **3**, research0046.1-research0046.9.

Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin ZK, Deutsch EW, Selway L, Walker J, Riba-Garcia I, *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**, 247–254.

Washburn MP, Wolters D and Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.

Short Specialist Review

Ontologies for three-dimensional molecular structure

John Westbrook and Helen M. Berman
Rutgers, The State University of New Jersey, Piscataway, NJ, USA

1. Introduction

Structural biology has witnessed an enormous increase in the amount and types of data collected by both experimental and computational methods. Data ontologies provide an important underpinning for the informatics infrastructure supporting structure and biological data. These ontologies provide precise definitions and examples, controlled vocabularies, allowed ranges and boundary conditions, data types, and data relationships (e.g., parent–child/foreign key). Data ontologies are required to support key informatics tasks such as the accurate acquisition information, data exchange including format conversion with semantic precision, database schema definition, application program interface definition, and data documentation.

Data ontologies can be expressed in many concrete formats and are not bound to a particular technology. In creating an appropriate ontology, the underlying data model must be adequate to represent the scientific content. It must also be widely accessible to users in that it must be easy to understand and extend, and it needs to be supported with reusable software tools. Three-dimensional (3 D) molecular structure data presents a variety of challenges in data representation. Approaches that have been used to address these challenges for small and macromolecular structure data are described.

2. Representation of small molecule structure data

2.1. CIF

Comprehensive descriptions of experimental details and 3 D structural results from X-ray diffraction studies of small molecules have been created as part of the Crystallographic Information File (CIF) project (Hall *et al.*, 1991). CIF data dictionaries describing experimental and structural details of single-crystal diffraction, powder diffraction, and incommensurately modulated structures are maintained by a standing committee of the International Union of Crystallography (IUCr) (International Union of Crystallography, 1998). These dictionaries contain data definitions

representing experimental data collection and reduction, phasing, refinement, 2 D (two-dimensional) chemical structure, 3 D coordinates, geometrical features of the resulting chemical structure, and a description of stereochemical quality.

The CIF data dictionaries are organized according to a Dictionary Definition Language (DDL) (Hall, 1998), which defines metadata attributes such as semantic definitions and examples, data type, controlled vocabularies, range restrictions, physical units, and functional relationships (e.g., mean and associated uncertainty). This dictionary metadata can be exploited by software to validate the logical consistency of a CIF data file.

The CIF data representation is used by the archival repositories for organic and inorganic small molecules – the Cambridge Crystallographic Data Center (CCDC) and the Inorganic Crystal Structure Database (ICSD) respectively. The CIF format also serves as an electronic publication format for the IUCr journal *Acta Crystallographica Section C*. Over 95% of the manuscripts describing small molecule structure determinations are submitted to this journal as CIF data files.

2.2. *CML*

The Chemical Markup Language (CML) project (Murray-Rust and Rzepa, 1999) provides an XML-based representation of small molecule chemical structure, and has announced plans to represent both 2 D and 3 D chemical structure, results of electronic structure calculations, chemical reactions, some physiochemical properties, and spectra. CML is described as an XML DTD or schema; however, the current core schema (v 2.1.1) documents only a portion of the proposed content.

2.3. *The SDF format*

Perhaps, the most pervasive representation of 2 D and 3 D small molecule chemical structure is that used within the SDF file format (Dalby *et al*., 1992). The SDF representation contains 3 D coordinates, a connection table describing bonding information, and stereochemistry. Complex chemical reactions and reaction intermediates can also be represented within an SDF file. This file format is widely used to exchange small molecule structure data between cheminformatics databases.

2.4. *Molecular identifiers*

An important issue in the representation of small molecule structure data is the assignment of a unique identifier to each molecule. Owing to the complexity in systematic nomenclature and the many common molecular naming systems, the use of molecular names as unique identifiers is problematic. Unique signatures based on chemical structure including stereochemistry have been developed to address this issue. The Simplified Molecular Input Line Entry Specification (SMILES) (Weininger, 1988) is one approach that has been used to create a canonical string

representation of chemical structure. Typically, a SMILES string is produced by a software tool using input from an SDF format file or a similar chemical description. IUPAC has recently started a project to create a standard unique identifier for small molecules called the *IUPAC Chemical Identifier* (*IChI*) (IUPAC, 2001). A preliminary implementation of an SDF to IChI conversion tool is available.

3. Representation of macromolecular structure data

3.1. The PDB format

The most widely used representation of macromolecular structure is that provided by the Protein Data Bank (PDB) (Berman *et al*., 2000) file format. This format was developed in 1972, and shares the column-oriented punch card style, typical of data formats of this period. In addition to the 3 D coordinates for each macromolecular structure, a PDB data file includes functional assignment, molecular names, biological source, chemical sequence, secondary structure, transformations required to construct a biologically functional assembly, references to related sequences, experimental details of crystallization or sample preparation, data collection, and refinement. The latter experimental details are reported in the form of unstructured or semi-structured text remarks. The content of the PDB format is documented in the PDB Content Guide: Atomic Coordinate Entry Format Description (Callaway *et al*., 1996).

3.2. mmCIF

In 1990, the IUCr formed a working group (Fitzgerald *et al*., 1992) to extend the small molecule CIF dictionary with data items relevant for the macromolecular crystallographic experiment. Initially, the hierarchical nature of macromolecular structure and experiment was difficult to represent using the DDL employed in the small molecule core CIF dictionary. To address this issue, a more structured DDL was adopted for the macromolecular dictionary, DDL2 (Westbrook *et al*., in press-b). DDL2 added language extensions to include all of the expressivity of the relational data model. This permits the required specification of uniqueness and for correspondences between primary and foreign keys. With this extension, the parent–child relationships between structure identifiers (such as chain identifier, residue name, and atom name) that are reused throughout the dictionary could be properly described, and these relationships could then be rigorously validated in software, using only the data dictionary.

After much community input and review, the first version of the macromolecular Crystallographic Information File (mmCIF) (Bourne *et al*., 1997; Westbrook and Bourne, 2000) dictionary was released in 1997. This version incorporated community-vetted definitions for experimental X-ray crystallographic concepts and definitions describing macromolecular structure and crystallographic experimental details included in the PDB at that time. A standing committee of the IUCr oversees the maintenance and extension of the mmCIF dictionary.

3.3. The PDB exchange data dictionary

The PDB adopted mmCIF and its supporting metadata architecture to represent the data that it collects and distributes. The PDB Exchange data dictionary (Westbrook *et al*., in press-a) provides a comprehensive and precise specification of the data content of the PDB. This data dictionary incorporates the full content of the mmCIF dictionary. Extensions are included for NMR, cryo-electron microscopy, structural genomics, protein production, and data extraction. The latter extensions define quantities that may appear in the outputs of structure determination applications. Careful definition of these data permits the automated software extraction of this information and inclusion of this information into the PDB deposition stream.

The PDB Exchange dictionary is also provided in the form of an XML schema (wwPDB, 2004).

3.4. NMR

The BioMagResBank (BMRB) (Ulrich *et al*., 1989) is the archival repository for NMR experimental data. The focus of this data resource is the collection of experimental NMR data (e.g., chemical shift data). The BMRB also includes a description of 3 D macromolecular structure in their data representation. Using a related but slightly different syntax, BMRB developed a data dictionary named NMRSTAR (BioMagResBank, 2004). NMRSTAR has recently been translated to a form that is compatible with the mmCIF style data dictionary (NMRIF). This has facilitated the exchange of data and dictionaries between crystallographic and NMR users and between the PDB and BMRB databases.

3.5. Cryo-electron microscopy

Cryo-electron microscopy is an experimental technique used in the determination of large molecular assemblies. The mmCIF dictionary extensions developed for cryo-electron microscopy (MSD-EBI and RCSB, 2000) include a description of sample preparation, raw volume data (Henrick *et al*., 2003), structure solution, and refinement.

3.6. Protein production

The detailed recommendations of a Task Force of the International Structural Genomics Organization (ISGO) (International Structural Genomics Organization, 2003) provide the foundation for a set of mmCIF dictionary extensions describing protein production. Data items describing protein expression, cloning, purification, NMR screening, and crystallization are included in this extension dictionary. These data extensions have been used as the foundation for the Protein Expression Purification and Crystallization database (RCSB, 2004) and for the protein production

process model developed to support the Structural Proteomics in Europe initiative (SPINE, 2004).

3.7. Protein structure classification

Taxonomies describing common protein shape are important for understanding protein function and evolutionary relationships. The Structural Classification of Proteins (SCOP) database (Murzin *et al*., 1995) has developed a structural taxonomy created by manual inspection and automated methods based on evolutionary-related families, superfamilies, and folds. Another taxonomy, CATH (Orengo *et al*., 1999), is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class (C), Architecture (A), Topology (T), and Homologous superfamily (H).

3.8. RNAML

RNAML (Waugh *et al*., 2002) is an XML representation of RNA sequence and structural features. This representation was developed within the RNA community to provide a standard means of exchanging information about RNA. The Nucleic Acid Database (NDB) (Berman *et al*., 1992) distributes these data files for known RNA structures. Software tools that use RNAML to compute and display RNA secondary and tertiary structure are also available (Yang *et al*., 2003).

3.9. Homology models

Homology models provide important computational insights into understanding 3 D macromolecular structure. An mmCIF extension dictionary describing methodology and outcomes of homology models has been developed (Adzhubei and Migliavacca, 1999). The details of template selection, template alignment, minimization or molecular dynamics treatments, and model quality assessments are included in this dictionary extension.

Acknowledgments

The RCSB PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology.

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, and Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research

Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of Neurological Disorders and Stroke (NINDS).

Further reading

Westbrook J, Henrick K, Ulrich EL and Berman HM In *International Tables for Crystallography*, Vol. G. Definition and exchange of crystallographic data, Hall SR and McMahon B (Eds.), Kluwer Academic Publishers, Dordrecht (in press-a).

Westbrook JD, Berman HM and Hall SR In *International Tables for Crystallography*, Vol. G. Definition and exchange of crystallographic data, Hall SR, McMahon B (Eds.) Kluwer Academic Publishers, Dordrecht (in press-b).

References

Adzhubei A and Migliavacca E (1999) *Model Database Dictionary*. GlaxoSmithKline, Geneva. **http://deposit.pdb.org/mmcif/dictionaries/ascii/mmcif_mdb_supplement_1.dic**.

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR and Schneider B (1992) The Nucleic Acid Database - a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal*, **63**, 751–759.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

BioMagResBank (2004) NMRSTAR. **http://www.bmrb.wisc.edu/dictionary/htmldocs/nmr_star/dictionary.html**.

Bourne PE, Berman HM, Watenpaugh K, Westbrook JD and Fitzgerald PMD (1997) The macromolecular Crystallographic Information File (mmCIF). *Methods in Enzymology*, **277**, 571–590.

Callaway J, Cummings M, Deroski B, Esposito P, Forman A, Langdon P, Libeson M, McCarthy J, Sikora J, Xue D, *et al*. (1996) *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, Brookhaven National Laboratory: Brookhaven, NY.

Dalby A, Nourse J, Hounshell W, Gushurst A, Grier D, Leland B and Laufer J (1992) Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Science*, **32**, 244–255.

Fitzgerald PMD, Berman HM, Bourne PE and Watenpaugh K (1992) Macromolecular CIF working group. International Union of Crystallography.

Hall SR (1998) Data languages and dictionaries for crystallography. *Acta Crystallographica*, **A54**, 820–832.

Hall SR, Allen AH and Brown ID (1991) The Crystallographic Information File (CIF): A new standard archive file for crystallography. *Acta Crystallographica*, **A47**, 655–685.

Henrick K, Newman R, Tagari M and Chagoyen M (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *Journal of Structural Biology*, **144**, 228–237.

International Structural Genomics Organization (2003) *Task Force on the Deposition, Archiving, and Curation of the Primary Information. Proposed data items describing protein production and crystallization*. **http://deposit.pdb.org/mmcif/sg-data/protprod.html**.

International Union of Crystallography (1998) Committee for the Maintenance of the CIF Standard. **http://journals.iucr.org/iucr-top/cif/comcifs/terms.html**.

IUPAC (2001) IUPAC Chemical Identifier (IChI) *Chemistry International*, **23**, 85.

MSD-EBI and RCSB (2000) *3D EM Extension Dictionary*. **http://pdb.rutgers.edu/mmcif/dictionaries/ascii/mmcif_iims.dic**.

Murray-Rust P and Rzepa HS (1999) Chemical Markup Language and XML. Part I. Basic principles. *Journal of Chemical Information and Computer Science*, **39**, 928.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**, 536–540.

Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L and Thornton JM (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Research*, **27**, 275–279.

RCSB (2004) PEPCdb. **http://pepcdb.pdb.org/**.

SPINE (2004). **http://www.spineurope.org/**.

Ulrich EL, Markley JL and Kyogoku Y (1989) Creation of a Nuclear Magnetic Resonance Data Repository and Literature Database. *Protein Sequences and Data Analysis*, **2**, 23–37.

Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, *et al*. (2002) A standard syntax for exchanging RNA information. *RNA*, **8**, 707–717.

Weininger D (1988) SMILES 1. Introduction and encoding rules. *Journal of Chemical Information and Computer Science*, **28**, 31.

Westbrook J and Bourne PE (2000) STAR/mmCIF: An extensive ontology for macromolecular structure and beyond. *Bioinformatics*, **16**, 159–168.

wwPDB (2004) RCSB, PDBj, MSD-EBI. **http://deposit.pdb.org/pdbML/pdbx-v0.906.xsd**.

Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman HM and Westhof E (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, **31**, 3450–3460.

Short Specialist Review

MAGE-OM: An object model for the communication of microarray data

Catherine A. Ball
Stanford University School of Medicine, Department of Biochemistry, Stanford, CA, US

Paul T. Spellman
Lawrence Berkeley National Laboratory, Life Sciences Division, Berkeley, CA, US

Michael Miller
Rosetta Inpharmatics, Seattle, WA, US

1. Background

All attempts to take advantage of the data being generated by large-scale biomedical technologies share two major challenges. First, the scale of data being produced requires computational infrastructure that is often immature and in the process of being built. Second, the considerable investment required to generate large-scale datasets necessitates additional measures to preserve and disseminate date for uses other than the primary experiment. These issues are very much intertwined: computational access to such data helps preserve and disseminate them, although providing the data in a format that is computer-friendly requires special measures that are often beyond the scope of the project that generates the data. Data storage for a single project can usually be provided in a reasonably straightforward manner, but sharing data with collaborators, sending data to repositories, exporting data to software packages, and providing data in a format that can be adequately reanalyzed by others can provide nearly insurmountable challenges. In addition, it is crucial to adequately express the biological context for the experiments or the data are meaningless. Agreeing which details are crucial, what to call those details and the format in which to communicate those details are easily overlooked, often ignored, and unspeakably dull, but is nonetheless critical. Without a format that is predictable and adequately structured, it is impossible to communicate large-scale datasets to others without loss of information.

Microarray experiments are one example of large-scale biological technologies that are being widely used by biomedical researchers. The amount of data

being produced usually exceeds the abilities of the laboratories generating the data. When microarrays were new technology, no standards existed to specify which critical details were needed to preserve the value of the data and no data formats existed with which to exchange data with others. In 1999, the grassroots Microarray Gene Expression Data Society (MGED; **www.mged.org**) was formed to construct such standards. The Minimum Information Annotating a Microarray Experiment (MIAME, Brazma *et al*., 2001) specifies which types of data and annotation ought to be communicated for a microarray study to stand on its own. Later, it was proposed that the MicroArray Gene Expression Markup Language (MAGE-ML) be used to communicate data (Spellman *et al*., 2002). Like MIAME, MAGE was created by the collaboration of individuals at public, academic, and corporate institutions who were united by the need for a common format to promote sharing of microarray data. When complete, the European Bioinformatics Institute and Rosetta Inpharmatics, Limited Liability Company (LLC), in a collaboration between nonprofit and industry, brought MAGE before the Object Management Group where it became the Gene Expression Standard.

The term 'MAGE' refers to a collection of related specifications and technologies that make it possible to encode MIAME-compliant data describing experiments using microarray technology. The Microarray Gene Expression Object Model (MAGE-OM) is a UML (Unified Modeling Language) object model to assist in the conceptualization and representation of a microarray investigation, while MAGE-ML is an XML- (EXtensible Markup Language) based implementation of MAGE-OM. MAGE-OM contains 132 classes that are organized into 17 packages. Classes in MAGE-OM represent discrete things or events associated with a microarray experiment because packages group related objects together. For example, the *BioSequence* package contains six classes that describe a given biological sequence, its features, its relationships to other biological sequences and its location, and use in individual microarrays. Importantly, the MAGE object model closely follows the flow of a microarray investigation and is thereby able to capture the core requirements for MIAME compliance. MAGE provides ways to describe an experiment's goals and design, the biological materials manipulated and used in the execution of an experiment, the design and manufacture of microarrays, the protocols used in the course of a microarray experiment, and the resulting data generated from a microarray experiment.

MAGE relies on external ontologies to provide concrete and specific details that are beyond the scope of the MAGE-OM. For example, the National Center for Biotechnology Information taxonomy (Wheeler *et al*., 2000; Benson *et al*., 2000) can be referenced to describe the species from which DNA sequences on an array (DesignElements in the MAGE-OM) are generated. One could also use the Jackson Laboratories mouse anatomy ontology (Begley and Ringwald, 2002; Ringwald *et al*., 2001) to describe an experiment where gene expression differences between mouse tissues were identified. The MGED Ontology (**http://mged.sourceforge.net/ontologies/MGEDontology.php**) encodes

many MIAME ideas and is one of the first biological ontologies to provide annotations that describe general experimental details.

2. Who is using MAGE?

Several commercial microarray vendors (e.g., Agilent and Affymetrix) provide descriptions of their microarrays using MAGE-ML. The MAGE-ML files are used by experimenters' research software applications and by microarray data repositories (such as ArrayExpress (Parkinson *et al.*, 2005) and the Gene Expression Omnibus (GEO, Barrett *et al.*, 2005)) to identify the biological reporters associated with each microarray spot. In addition, many microarray research databases (e.g., Stanford Microarray Database (SMD, Ball *et al.*, 2005), BioArray Software Environment (BASE, Saal *et al.*, 2002), RNA Abundance Database (RAD, Manduchi *et al.*, 2004), Microarrray Data Manager (MADAM Saeed *et al.*, 2003)) and software packages used by experimenters for the prepublication analysis of microarray data have MAGE-ML export ability. This allows researchers to submit MAGE-ML records of their published experiments directly to ArrayExpress or GEO without loss of information. The number of software applications and microarray vendors using MAGE increases every year, as community demand for more seamless data sharing increases.

3. MAGEstk

MAGEstk (pronounced "majestic") is the MAGE software toolkit that provides code for Java, Perl, and C++ programmers. It is freely available at **http://mged.sourceforge.net/** and is an open source based on the Massachusetts Institute of Technology license. This toolkit, actively developed by contributors all over the world, is targeted to bioinformaticians, database and application programmers, and groups or individuals who use microarrays. MAGEstk provides many basic functionalities for creating and handling MAGE files and using the MGED ontology. MAGEstk, like MAGE, is a product of the bioinformatics community and regular programming jamborees (usually twice a year) are held during which software developers contribute their time and expertise to writing new tools for MAGEstk. The first component of MAGEstk is the UML compiler that produces the second component – MAGE-ML parsers and writers. The third component is a MAGE-ML validator that ensures that individual or sets of documents conform to critical rules.

The MAGEstk parsers process MAGE-ML into in-memory object structures (either Perl or Java) and, conversely, serialize object structures to MAGE-ML. Both sets of modules are available for use from SourceForge. These modules allow a consistent application programming interface (API) based on the names of the objects in the model and the names of the associations between objects. The Perl and Java modules for MAGE-ML parsing are generated almost entirely using a set of rule-based mappings that allow automatic code production, with only a few hand-coded

exceptions from the UML MAGE-OM model. The MAGE validator tool is freely available from **ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/MAGEvalidator-DISTRIB/**. The tool has been written in Java and can be installed locally on Windows or Unix platforms. The tool reads a MAGE-ML file, checks its validity based on criteria such as the nonduplication of object identifiers and the presence of required MIAME fields, and produces the validation log (for more see **http://www.ebi.ac.uk/∼ele/ext/submitter.html#val**).

4. Barriers to using MAGE

Unless an experimenter has access to one of the research databases or software packages that read MAGE-ML and automate MAGE-ML generation, there are significant hurdles to using MAGE. Using the MAGE-OM still requires a high level of programming expertise and familiarity with the object model that, unfortunately, is beyond the abilities of the average bench biologist. Although the MAGEstk provides a certain level of common tools and approaches, any given project is likely to have to write some programming code in order to import or export MAGE-ML files. This is increasingly mitigated by a growth in the number of laboratories with informatics resources as well as the bioinformatics community's embrace of open software. Many high-quality software packages are freely available that are able to produce microarray data in MAGE-ML.

Importing MAGE-ML files produced by others presents additional challenges because there are occasional redundancies in the object model that allow different applications to record the same information in different places in the object model. This ambiguity leads to some difficulties using MAGE-ML from different sources. However, this can usually be addressed with communication, and resolving this difficulty is a major goal in the next version of MAGE (see Section (5) below). A more obstinate problem is that MAGE-ML files are very large, making reading by humans essentially impossible. Keeping in mind that some commercial microarrays have more than a million features whose sequences and biological identity must be recorded, it is clear that the amount of data associated with microarray experiments is by definition large, and the structure of XML results in a very significant increase in file size. Methods to handle these large files are available in the MAGEstk software toolkit, but simply storing, moving or processing such large files provides challenges.

5. Future plans

MAGE developers are now embarking on two new projects. The first is to develop and disseminate a second version of MAGE-OM (MAGE-OM v2) that will require less interpretation than the current version. The second goal is to incorporate MAGE into a greater object model for a superset of large-scale biomedical technologies, called Functional Genomics (FuGE).

MAGE-OM v2 will address five areas that are currently recognized as limiting. The first is that MAGE-OM is difficult to extend to other high-throughput

technologies. This goal is tied in closely with our efforts with FuGE. Second, the model structure needs to be simplified so that it is easier to use – in some cases there are unnecessary levels of detail that is difficult to interpret and apply correctly. Third, ambiguities in the model need to be eliminated so that MAGE-ML documents can be more easily interpreted by different software packages. The fourth task is to completely incorporate the MGED Ontology (which did not exist when MAGE-OM was first developed). Lastly, the model needs to be expanded to fully specify higher-level data analysis procedures and quality assessment measurements.

The goal of FuGE is to share resources between different large-scale data generating groups for similar types of objects (such as descriptions of experimental procedures, biological samples or clinical data) and to have specialized objects be designed by those with expertise in the domain (such as proteomics). In this manner, bioinformaticists will not have to reinvent the wheel for every high-throughput technology that is being applied to biomedical questions. The only new modeling required will be that needed to record technology-specific details.

MAGE development will continue to be a community-driven project. Our experience has taught us that those most closely involved in generating, storing, and analyzing microarray data have the most to gain by the development of resources and software such as MAGE-OM and MAGEstk. Conversely, those same bioinformaticians have the most to lose by having standards foisted upon them that are unwieldy and of little practical value. Accordingly, community involvement and contributions to MAGE-OM and MAGEstk have been remarkable. A far-flung group from many sectors of the biomedical and bioinformatics research communities helped to create MAGE-OM, to contribute code to MAGEstk, and to provide the constructive criticism needed to pinpoint the areas for future research. Without the active participation of individuals from many organizations, there would be no MAGE-OM and no hope of releasing an improved second version.

References

Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, *et al*. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research*, **33**, D580–D582.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W and Edgar R (2005) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Research*, **33**, D562–D566.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 373.

Begley DA and Ringwald M (2002) Electronic tools to manage gene expression data. *Trends in Genetics*, **18**, 108–110.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DL (2000) GenBank. *Nucleic Acids Research*, **28**, 15–18.

Manduchi E, Grant GR, He H, Liu J, Mailman MD, Pizarro AD, Whetzel PL and Stoeckert CJ Jr. (2004) RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics*, **20**, 452–459.

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, *et al*. (2005) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **33**, D553–D555.

Ringwald M, Eppig JT, Begley DA, Corradi JP, McCright IJ, Hayamizu TF, Hill DP, Kadin JA and Richardson JE (2001) The mouse gene expression database. *Nucleic Acids Research*, **29**, 98–101.

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A and Peterson C (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biology*, **3**, SOFTWARE0003.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, **3**, RESEARCH0046.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA and Rapp BA (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **28**, 10–14.

Frame-based systems: Protégé

Iwei Yeh
Stanford University, Stanford, CA, USA

A conceptualization is an abstract, simplified view of a domain of interest (Gruber, 1993). Concepts represent classes of objects in a domain. For example, the concept gene represents the set of genes that occur in nature. These concepts have describable attributes to them, such as the name of a particular gene, as well as relationships between them, such as the relationship between a gene and its protein product. These attributes and relationships define a class of objects in the real world, and their values define the individuals. An ontology is a formal specification of a conceptualization. Well-structured ontologies allow querying of complex relationships within large data sets and facilitate automatic inference of new knowledge (Russell and Norvig, 1995; Bada and Altman, 2000). Modeling data sources to closely reflect the reality of the world they are trying to describe is the key to general reusability and intuitive querying.

An important method for representing ontologies is one that is based on frames (Fikes and Kehler, 1985; Minsky, 1987; Karp, 1992). Minsky popularized the frame as a data structure by drawing an analogy with the mental structures people use when dealing with stereotyped situations. Minsky argued that people use "frames", which can be used to represent generic concepts, such as a birthday party, and these concepts have specific characteristics, such as the number of people attending or the kind of birthday cake. These characteristics of a frame may have default values but can be filled in for the specific occasion (for example, the cake is always chocolate unless otherwise specified). In frame-based systems, a class frame represents each concept and contains slots describing its attributes. Facets describe properties of slots, such as their allowed values and cardinality. A relation is an attribute that has another frame as the allowed value type. An instance frame is one that represents a particular object from the set of objects described by the class; it is a copy of the class frame with default values with slot values filled in for the particular instance.

A special relation is the *is-a* relation that organizes frames into a taxonomic hierarchy and allows for inheritance of slot values. Inheritance allows subconcepts (classes or instances) to inherit the parent concept's slot values, decreasing redundancy of stored information, thereby reducing the chance of inconsistencies and increasing the efficiency of updates.

Frames are conceptually simple and easy to understand, and the information in frames is directly stored or easy to compute via *is-a* relationships. A disadvantage of frames is that they provide a static representation, where each *is-a* relation must be

asserted and cannot be inferred by the system (Baker *et al*., 1999). The meaning of the different relations must be carefully defined so that the semantics of the relation are clear and not misinterpreted. Even for the basic *is-a* link, there are many possible interpretations that must be distinguished (Brachman, 1983). Flier *et al*. describe three types of *is-a* relationships: "A is by definition B", "A is probably a B", and "A is in theory necessarily a B" (Flier and Zanstra, 1998). Each frame is viewed as an assertion, or statement of fact. Therefore, expression of incomplete knowledge in a frame-based system can be very difficult (Brachman *et al*., 1987). For example, it is difficult to state that the slot of a specific instance is one of two out of many possible values. Querying in frame-based systems consists of following *is-a* links for determining subsumption and other links for other types of inference.

A knowledge base (KB) is an ontology with the addition of instances. A knowledge base management system (KBMS) facilitates knowledge modeling, knowledge acquisition, consistency checking, and concurrency control of ontologies and KBs. *Knowledge modeling* is the creation of classes, slots (with default values), and facets. *Knowledge acquisition* involves extracting knowledge about the domain from various sources. *Consistency checking* ensures that classes and instances in the KB have attributes that conform to the knowledge model. *Concurrency control* maintains appropriate behavior of the KB during and after simultaneous operations.

Protégé is an open-source KBMS environment (Gennari *et al*., 2003), which is written in Java and available for several platforms (**http://protege.stanford.edu**). Protégé is based on frames and conforms to the OKBC interface (**http://www.ksl.stanford.edu/software/OKBC/**), which integrates KBs across systems. Many biological ontologies are available in Protégé format, including a biological processes ontology, anatomy ontology, and Gene Ontology (Hahn *et al*., 1999; Peleg *et al*., 2002; Yeh *et al*., 2003). Ontologies in Protégé are easily sharable due to the availability of the software, and the various formats of storing the ontology (including text and database formats). Protégé provides a graphical user interface for creating and editing ontologies and KBs.

In Figure 1, the classes tab of Protégé is activated, and the *is-a* hierarchy for a cellular architecture ontology is shown in the pane at the left (unpublished work). The concept *endosome membrane* is highlighted, and the attributes of the frame representing the concept are shown in the pane on the right. Template Slots are attributes that apply to the instances of endosome membrane, while the other attributes are for the concept or class itself. The hierarchy is clickable, and children concepts can be either hidden or shown.

Protégé assists in the task of knowledge acquisition by automatically generating forms for creating instances based on class definitions. The instance tab is shown in Figure 2.

This allows the tasks of knowledge modeling and knowledge acquisition to be easily separated between users. The automatically generated form enforces the constraints of attribute type and cardinality that are defined during class creation. Users can easily browse instances through the class hierarchy in the left pane.

There have been several tools developed within the Protégé framework to assist users in developing, maintaining, and querying KBs (**http://protege.stanford.edu/plugins.html**). The queries tab allows users to search for instances based on class and slot values. For more complex queries and constraints, the Protégé Axiom

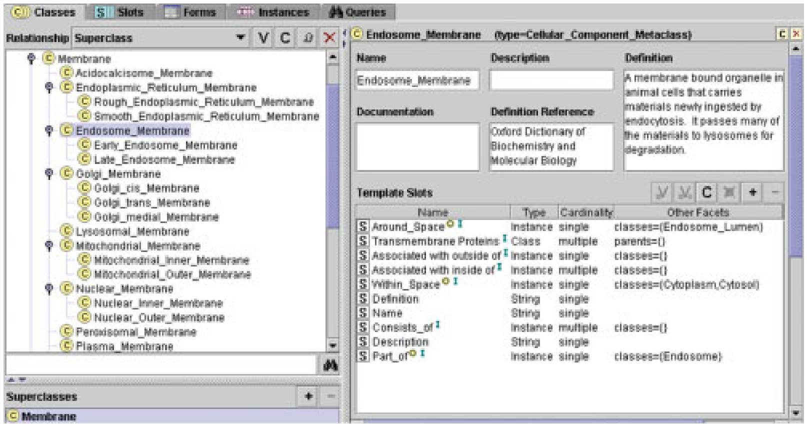


Figure 1 Screenshot of Protégé with classes tab activated, and the class Endosome Membrane highlighted in the pane on the right

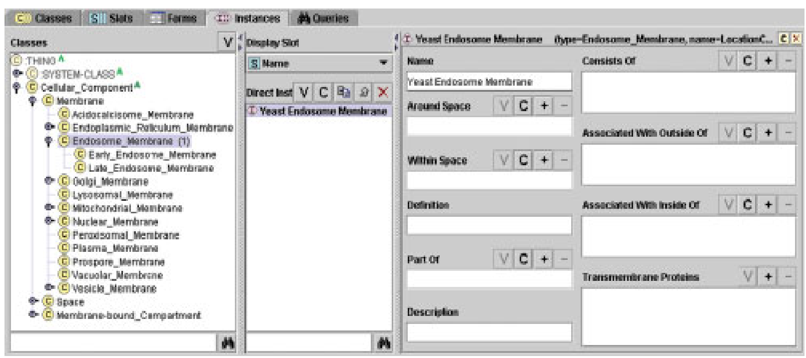


Figure 2 Screenshot of Protégé with instances tab activated. The class Endosome Membrane is highlighted in the classes pane on the far left. The central pane shows the instances of Endosome Membrane and the Yeast Endosome Membrane is highlighted. At the pane on the far right, a user can edit slot values for this instance

Language (PAL), which uses a superset of first-order logic, allows users to logically query and create constraints over a KB based on classes, instances, and their attributes. PROMPT is a suite of tools for managing multiple ontologies, and automatically generates a list of differences and similarities between two input ontologies (Noy and Musen, in press). This allows users to develop the same ontology in parallel and integrate changes at a later time. Protégé has also been extended to enable usage of non-frame-based representations, including XML schema, and OWL (*see* Article 92, **Description logics: OWL and DAML + OIL**, Volume 8) (Gennari *et al*., 2001; Rector, 2004). This allows for import and export of Protégé projects to and from other systems and integration of Protégé projects with ontologies generated in other platforms.

These additional tools developed for Protégé were created using the Protégé Java API. The API is well documented (**http://protege.stanford.edu/doc/pdk/index.**

html) and allows users to develop plug-in extensions for Protégé-2000 or separate programs that interface with Protégé projects. This allows for more individualized and sophisticated querying algorithms as well as automatic KB population from high-throughput data sources.

References

Bada MA and Altman RB (2000) Computational modeling of structural experimental data. *Methods Enzymology*, **317**, 470–491.

Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R and Brass A (1999) An ontology for bioinformatics applications. *Bioinformatics*, **15**(6), 510–520.

Brachman RJ (1983) What ISA and Isn't: an analysis of taxonomic links in semantic networks. *IEEE Computer*, **16**(10), 30–36.

Brachman RJ, Fikes RE and Levesque HJ (1987). KRYPTON: A functional approach to knowledge representation. In *Readings in Knowledge Representation*, Brachman RJ and Levesque HJ (Eds.), Morgan Kaufmann: Los Altos, CA, pp. 412–429.

Fikes R and Kehler T (1985) The role of frame-based representations in reasoning. *Communications of the ACM*, **28**(9), 904–920.

Flier FJFdVRP and Zanstra PE (1998) Three types of IS-A statement in diagnostic classifications: three types of knowledge needed for development and maintenance. *Methods of Information in Medicine*, **37**, 453–459.

Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, Noy NF and Tu SW (2003) The evolution of protege: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, **58**(1), 89–123.

Gennari JH, Sklar D and Silva J (2001) Cross-tool communication: from protocol authoring to eligibility determination. *Proceedings of the AMIA Annual Symposium*, 199–203.

Gruber T (1993) *Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Knowledge Systems Laboratory*, Stanford University.

Hahn JS, Burnside E, Brinkley JF, Rosse C and Musen MA (1999) Representing the digital anatomist foundational model as a protege ontology. *American Medical Informatics Association Fall Symposium*, Washington DC.

Karp PD (1992) *The Design Space of Frame Knowledge Representation Systems*, SRI International Artificial Intelligence Center.

Minsky M (1987) A framework for representing knowledge. In *Readings in Knowledge Representation*, Brachman RJ and Levesque RJ (Eds.), Morgan Kaufmann: Los Altos, CA, pp. 246–262.

Noy NF and Musen MA (2004) Ontology versioning in an ontology management framework. *IEEE Intelligent Systems*, **19**(4), 6–13.

Peleg M, Yeh I and Altman RB (2002) Modelling biological processes using workflow and petri net models. *Bioinformatics*, **18**(6), 825–837.

Rector A (2004) Defaults, context, and knowledge: alternatives for OWL-indexed knowledge bases. *Pacific Symposium on Biocomputing*, **9**, 226–237.

Russell SJ and Norvig P (1995) *Artificial Intelligence A Modern Approach*, Prentice Hall: Upper Saddle River, NJ.

Yeh I, Karp PD, Noy NF and Altman RB (2003) Knowledge acquisition, consistency checking and concurrency control for gene ontology (GO). *Bioinformatics*, **19**(2), 241–248.

Basic Techniques and Approaches

Description logics: OWL and DAML + OIL

Phillip Lord, Robert D. Stevens, Carole A. Goble and Ian Horrocks
University of Manchester, Manchester, UK

OWL and its predecessor, DAML + OIL, are ontology languages developed for the *Semantic Web*. As such, they support its aim of increasing the amount of information on the web that is computationally accessible (i.e., that can be unambiguously interpreted and processed by software as well as humans). With the acceptance of OWL as a recommendation by the W3C (World Wide Web Consortium, the standards body for web technologies), this language is moving from research into mainstream technology with increasing use and availability of tools such as the editors Protégé (*see* Article 91, **Frame-based systems: Protégé**, Volume 8) and OilEd. Underlying a fragment of OWL called *OWL-DL* is a Description Logic (DL), which supports the definition and description of concepts, relationships, individuals and axioms (constraints), and the organization of concepts and relationships into hierarchies.

Description Logics have several key features that make them attractive as ontology languages:

Expressivity: DLs are highly expressive, enabling rich, and complex descriptions of domain concepts. Concepts can be defined in terms of their properties and their relationships to other concepts. It is not necessary to use all of the expressive power in OWL, however. Some or all of the ontology can be represented as a simple taxonomy.

Automated Reasoning: DLs are logics which means there is a clear understanding of the language's formal properties. This has enabled the development of reasoners – software that is capable of checking ontologies for consistency and inferring that one concept is a kind of another concept. This latter characteristic means that the concept hierarchy can be inferred on the basis of the contents of the ontology instead of being handcrafted by the ontologist.

Compositionality: The previous two properties enable the building of ontologies in a compositional way – making new concepts by combining previously defined concepts and properties. This means that it is unnecessary to predetermine and enumerate all the concepts of the ontology beforehand, making the process of building large ontologies more manageable and flexible.

```
Protein that is associated with chromosomal DNA, including
histones, protamines, and high-mobility group proteins
```

Statement 1 Chromosomal Protein: as defined by Swissprot

```
ObjectProperty(isPartOf)
Class(Protein)
Class(Chromosome)
Class(ChromosomalProtein partial
    Protein
    restriction(isPartOf someValuesFrom(Chromosome)))
```

Statement 2 Chromosomal Protein: These statements define a property "isPartOf" and two classes, Protein and Chromosome. The third class, ChromosomalProtein, is defined as a subclass of protein and a part of chromosome. The syntax used here is a slight variation on the OWL Abstract Syntax, which is described at **http://www.w3.org/TR/owl-semantics/**

DLs differ from frame-based ontology languages (*see* Article 91, **Frame-based systems: Protégé**, Volume 8) primarily because of their amenability to automated reasoning. They share a common heritage, however, having evolved from early frame-based systems, but with the addition of a more formally defined semantics.

To illustrate the features of DLs, we show how OWL might be used to describe aspects of chromosome biology. In particular, we aim to show how automated reasoning can help in the construction of these descriptions.

Statement 1 shows Swissprot's definition of the Chromosomal Protein keyword – part of a controlled vocabulary, which avoids problems of synonyms and context. The definition of this term is in English and is therefore not accessible computationally, which leaves no explicit statement of the relationship between this and other similar keywords. The Gene Ontology (*see* Article 82, **The Gene Ontology project**, Volume 8) also recognizes this difficulty and provides a means for defining relationships between concepts.

DL systems not only support but also encourage the definition of relationships between different concepts. So in Statement 2, "chromosomal protein" is defined in terms of "chromosome" and "protein".

Two different kinds of relationship have been used: the subclass, or subsumption, relationship, and isPartOf. OWL enables the definition of new properties, which can then be used in descriptions. OWL properties can themselves have characteristics. So isPartOf might be defined as the inverse of hasPart. In Statement 3, we provide definitions for Chromosome and SegregatingUnit. As well as defining a new property (involvedIn), some of the more expressive features of OWL are used.

DLs are logics and as such have *well-defined* semantics, which enables a precise machine interpretation of the definitions of concepts in an ontology. The most important practical outcome of this has been the development of DL reasoners – software components that are capable of determining the *satisfiability* of ontological concepts. For example, the descriptions in Statement 4 state that an AcentricChromosome is a Chromosome and that it has no Centromere. A DL reasoner will discover that this concept is *unsatisfiable* – the definition is contradictory. In this

```
ObjectProperty(involvedIn)
ObjectProperty(hasPart)

Class(SegregatingUnit complete
    restriction(involvedIn someValuesFrom (Segregation)))
Class(Chromosome complete
    intersectionOf(SegregatingUnit
        restriction(hasPart someValuesFrom (Telomere))
        restriction(hasPart someValuesFrom (Centromere))))
```

Statement 3 Chromosome: Two more properties and two more classes are defined. These class descriptions are "complete" (that is, the definition is equivalent to the class and vice versa) rather than "partial"; all Segregating Units are involved in Segregation, *and* anything involved in Segregation is a Segregating Unit. A Chromosome is defined as a Segregating Unit with a telomere and a centromere

```
Class(AcentricChromosome partial Chromosome)
DisjointClasses(restriction(hasPart someValuesFrom (Centromere))
   AcentricChromosome)
```

Statement 4 Acentric Chromsome: Here, an acentric chromosome is defined. It is also described as "disjoint" from "things having a part centromere", or there can be nothing that both has a centromere and be an acentric chromsome

case, a Chromosome is described as having a Centromere, so either an Acentric-Chromosome must have a Centromere or not be a Chromosome.

The biologist may agree or disagree as to whether an acentric chromosome is really a chromosome or not. DLs can give no guarantee that the ontology being produced is a good model of reality or otherwise. However, the reasoner can at least check that the model is internally consistent. The ontology developer is forced to consider which one of their definitions should be modeled differently. We can draw an analogy to a type system in a programming language in that the explicit definitions enable errors to be picked up and resolved earlier in the process.

In addition to satisfiability, DL reasoners are capable of determining *subsumption* relationships between concepts based on their descriptions. We might choose to define YAC (Yeast Artificial Chromosome) as shown in Statement 5. In this case, the DL reasoner is capable of telling us that a YAC as well as being a CloningVector is also a kind of Chromosome; a YAC is a SegregatingUnit with a Telomere and a Centromere that Statement 3 defines to be equivalent to a Chromosome.

```
Class(CloningVector partial
    SegregatingUnit)

Class(YAC partial CloningVector
   restriction(hasPart someValuesFrom (Telomere))
   restriction(hasPart someValuesFrom (Centromere)))
```

Statement 5 Yeast Artifical Chromosome: A cloning vector with a telomere and a centromere as part. A Cloning Vector is defined as a kind of SegregatingUnit

The use of automated reasoning technologies also enables the ontological engineer to adopt a different style of modeling, one which is highly compositional. Enabling the definition of concepts in terms of their properties reduces the task of building a large hierarchy of concepts to that of building a set of much smaller hierarchies over which the reasoner can operate. In practice, this can enable the generation of large, internally consistent ontologies (*see* Article 85, **TAMBIS: transparent access to multiple bioinformatics services**, Volume 8).

Statement 5 gives an example. The ontology developer need not consider the subsumption relationship between a chromosome and a YAC as the reasoner is capable of determining this. Nor need they worry about the order of definition of concepts. If YAC is defined first and at a later date the ontologist adds a definition for Chromosome, the reasoner will ensure that this concept has all appropriate subclasses including YAC.

In Conclusion, Description Logics and OWL offer the ontology builder a rich and expressive formalism for expressing concepts in terms of their properties. The provision of a formal semantics for the language ensures that the definitions of these concepts are computationally amenable, which in turn facilitates the building of large ontologies. Bioinformatics as a discipline has historically been an earlier adopter and contributor to web technologies. Hopefully, the advent of the widespread use of ontologies in bioinformatics in the shape of the Gene Ontology and arrival of Semantic Web technologies such as OWL will continue this trend.

Further reading

Franz B, Diego C, Deborah M, Daniele N and Patel-Schneider PF (Eds) (2003) *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press: Cambridge, ISBN 0-521-78176-0.

Stefan S and Rudi S (Eds) (2003) *Handbook on Ontologies*, Springer: Berlin and Heidelberg, ISBN: 3-540-40834-7.

Introductory Review

Detecting protein homology using NCBI tools

Wayne T. Matten and Scott D. McGinnis

National Center for Biotechnology Information, Bethesda, MD, USA

1. Introduction

Scientists around the world use a public web page to submit more than 100 000 sequence searches a day to the National Center for Biotechnology Information (NCBI). The NCBI computers run a program called *BLAST*® (Basic Local Alignment Search Tool), which provides an indication of how similar the query sequence is to other known sequences (Altschul *et al*., 1997). What drives this craving for sequence similarity? Common uses are as follows:

- To infer homology. Two sequences from different species are said to be homologous when they are derived from a common ancestor sequence. Sequence similarity provides a measure of this relatedness.
- To infer function. The potential function of a poorly characterized protein can be assigned on the basis of its sequence, and possibly structural, similarity to well-studied proteins in other organisms. This often includes identification of known conserved domains in the protein (*see* Article 91, **Classification of proteins by sequence signatures**, Volume 6).
- To identify unknown sequences, such as newly sequenced DNA, or peptide fragments from high-throughput studies of protein mixtures (*see* Article 17, **Tutorial on tandem mass spectrometry database searching**, Volume 5).
- To study evolution by following changes in protein homologs and protein families over time (*see* Article 9, **Modeling protein evolution**, Volume 1).

This review focuses on protein sequence comparison, which is more sensitive – can detect similarity in more evolutionarily distant organisms – than nucleotide sequence comparison. One must keep in mind that programs such as BLAST do not calculate homology; they only provide data that may support an inference of homology. These data include percent sequence identity, scores to help rank the comparisons, statistics to help judge relevance, and the alignments themselves. A common error is to equate the percent identity between two sequences with the degree of homology; two sequences are either homologous or not; they cannot share low or high degrees of homology. Furthermore, BLAST and other popular

Global alignment

```
               10        20             30             40
cort   MWKAAAG--QSVSMSVDEGADDWET-DPD---FEN-DV----SEKEQRWGAKTVQG-SGH
       : :      .  ...: ..   .:   ::   ::. :.    . ..  :   : .: .:
SH3    MSKPPPKPVKPGQVKVFRALYTFEPRTPDELYFEEGDIIYITDMSDTSWWKGTCKGRTGL
               10        20        30        40        50        60

       50        60         70        80        90       100
cort   QEHINIHKLRQTVSTE-HSDLKQKELATMPKASHGYGGKFGLQQHRMDKSAVGHEYQSKL
            . .  .....  :   :. .:. . .   .  :  ::     ::..    : .
SH3    IPSNYVAEQAESIDNPLHEAAKRGNLSWLRECLDNRVGVNGL-----DKAGSTALYWAC-
               70        80        90       100            110

       110       120       130         140       150       160
cort   SKHCSQTDTSKGFGGKYGLQEDLYQNTEPDTG--AESYEYAEDVGITAVALYDYQAAGDD
         : .. :  . .  . ... .  ::   ::.  : ...   :.    .:        ..
SH3    --HGGHKDIVEVLFTQPNVELN-QQNKLGDTALHAAAWKGYADIVQLLLAKGARTDLRNN
            120       130        140       150       160       170

         170       180         190        200
cort   EISFDPDDIITNIEMID--EGWWRGVCRGAYGLFPA-NYV--EVRQ
       : ..   :. ::    .  .   .:.  ::  :  : .:.  :  .
SH3    EKKL-ALDMATNAACASLLKKKQQGT-DGARTLSNAEDYLDDEDSD
              180       190        200       210
```

Identities (:)   = 19% (42/226) Gaps (-) = 9% (20/226)
Positives (: + .) = 44% (99/226)

Local alignment

```
cort    155 ALYDYQAAGDDEISFDPDDIITNIEMIDEGWWRGVCRGAYGLFPANYV 202
            ALY ++    DE+ F+  DII   +M D  WW+G C+G  GL P+NYV
SH3     19  ALYTFEPRTPDELYFEEGDIIYITDMSDTSWWKGTCKGRTGLIPSNYV 66
```

Identities = 45% (22/48) Gaps = 0% (0/48)
Positives   = 62% (30/48)



Figure 1 Global alignment programs are constrained to find the best match along the full length of the sequences. Here, the Align program (Lipman and Pearson, 1985) was used to compare the *Danio rerio* “similar to cortactin” protein (cort; NCBI Reference Sequence, accession number NP_957335.1; 206 amino acids) and the *Mus musculus* SH3 (src homology 3) domain protein 3 (SH3; NCBI Reference Sequence, accession number NP_059071.1; 215 amino acids). Local alignment programs seek the best *segments* of alignment. The 48 amino acid region found by pairwise BLAST (Tatusova and Madden, 1999) corresponds to the SH3 domain of each protein, highlighted in the global alignment and shown as the thick line in the graphic

tools, such as FASTA (Pearson and Lipman, 1988), generate *local* rather than global alignments, so the percent identity only refers to the aligned regions of the sequences. These local regions of similarity are useful because they often represent conserved domains (Figure 1).

It is easy to understand that the more similar two sequences are, the more likely they are to be related, but at what percent identity overall are two sequences considered homologous? There is no magic degree of similarity that defines homology. Other helpful information includes potential biological function, comparison of many similar sequences (multiple sequence alignment), and the statistics of sequence similarity algorithms.

To help interpret BLAST results, a statistical value called the *Expect value* is calculated for each alignment. The Expect value indicates the number of hits (sequences) in the database searched that are expected to align to the query simply by chance. So an Expect value of 1 means that this alignment could have occurred by chance, that is, may not be biologically meaningful. As the Expect value decreases, the likelihood that the two sequences are related increases (Altschul, 1998).

This review outlines just a few of the possibilities for analyzing protein homology using tools found on the NCBI web site. By design, BLAST trades a bit of sensitivity for the speed that is needed to search large databases via a public search interface. In some cases, more sensitive searches can be performed with algorithms such as Smith–Waterman (Smith and Waterman, 1981) and others (Persson, 2000; Sansom, 2000).

2. Finding homologous sequences

Any search for homologs is dependent both on the sequence comparison program and the database. A very brief review of NCBI's protein databases follows (see also Wheeler *et al*., 2004).

2.1. Entrez protein database

The majority of sequences in NCBI's protein databases are translated coding regions from nucleotide submissions to GenBank, with the exception that high-throughput nucleotide sequences, such as Expressed Sequence Tags (ESTs), are not translated. Text-based access to these records is through Entrez Protein, which includes sequences submitted to outside sources: SWISSPROT, Protein Information Resource (PIR), Protein Research Foundation (PRF), and Protein Data Bank (PDB); the latter contains sequences from solved structures (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7). Sequence-based access, via a BLAST search against the default protein database nr (nonredundant), includes all of these sequences, except for the Patent division of GenBank.

2.2. Conserved domain database

NCBI's Conserved Domain Database (CDD) contains known, conserved regions of sequence that are represented as position-specific scoring matrices (Marchler-Bauer *et al*., 2003). You can think of domains as modules of function, and often three-dimensional (3D) structures, that are maintained during evolution. CDD's primary use is to identify such domains in a protein sequence to help infer protein function or to place a sequence in a known protein family.

NCBI compiles CDD from both outside databases – Pfam (Bateman *et al*., 2002) and SMART (Letunic *et al*., 2002) – and internal sources, including the database

COGs (Clusters of orthologous groups); (Tatusov *et al*., 2001). A related search tool, CDART, the Conserved Domain Architecture Retrieval Tool, is used to find proteins with a similar set of domains to the query protein (Geer *et al*., 2002).

2.3. BLAST implementations

All of the different BLAST programs can be accessed from the main BLAST web page (**http://www.ncbi.nlm.nih.gov/BLAST/**). This page also provides links to the downloadable BLAST executables, databases, and documentation, **ftp://ftp.ncbi.nlm.nih.gov/blast/**, and to the source code, **ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/ncbi.tar.gz**.

2.4. BLASTP

A BLASTP search against the nr (protein) database is a comprehensive search and often yields satisfactory results. The default scoring matrix, BLOSUM62, is remarkably nimble when comparing sequences from either closely or distantly related species (Henikoff and Henikoff, 1992). However, a good principle in homology searching is to use more than one approach. To be exhaustive, try different scoring matrices, databases, search parameters, and even algorithms.

One common use of BLASTP is to find matches to a short amino acid motif, say, 10–20 residues. A separate BLAST web page called *Search for short, nearly exact matches* tailors the search parameters to the use of short peptide queries: PAM30 scoring matrix (Dayhoff *et al*., 1978); minimum word size; no query filter for low-complexity sequence (Wootton and Federhen, 1993); and increased Expect value cutoff. The Expect value needs to be increased from the default value of 10 because matches greater, that is, worse, than the Expect threshold are not reported and Expect values in the 1000's are frequent with these short query sequences. Note that because of the nature of the BLAST algorithm and the PAM30 scoring matrix, only nearly exact matches to the query will be found. If a motif with ambiguous residues is sought, then a pattern-matching algorithm, not BLAST, is a better choice (Mulder and Apweiler, 2001).

2.5. PSI-BLAST

Position-specific-iterated BLAST (PSI-BLAST) is more sensitive than BLASTP. Greater sensitivity is achieved by constructing a position-specific scoring matrix, referred to as a *PSSM* or profile, from a multiple sequence alignment of the results of a standard BLASTP search. One axis of a PSSM is the aligned region of the query sequence, while the other axis is made up of the standard amino acids (Figure 2). The position-*independent* scoring matrices, such as BLOSUM62 or PAM30, have a defined score for each amino acid substitution, regardless of the residue's position in the query. In effect, a PSSM gives greater weight to those amino acids that are conserved among the sequences in the multiple sequence alignment. By then using the PSSM as the query in subsequent searches or iterations,

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T
435	K	-1	0	0	-1	-2	3	0	3	0	-2	-2	1	-1	-1	-1	-1	-1
436	E	0	1	0	2	-1	0	2	-1	0	-1	-1	0	0	0	-1	0	0
437	S	0	0	-1	0	1	1	0	1	1	0	-1	0	0	0	2	0	-1
438	N	-1	0	-1	-1	1	0	-1	3	3	-1	-1	1	-1	0	0	-1	-1
439	K	-2	1	1	-1	-2	0	-1	-2	-2	-1	-2	5	1	-2	-2	-1	-1
440	P	-2	-2	-2	-2	-3	-2	-2	-2	-2	-1	-2	-1	0	-3	7	-1	-2
441	A	3	-2	1	-2	0	-1	0	1	-2	-2	-2	0	-1	-2	3	1	0
442	M	-3	-4	-4	-4	-3	-4	-4	-5	-4	7	0	-4	1	0	-4	-4	-2
443	A	4	-4	-4	-4	0	-4	-4	-3	-4	4	-1	-4	-2	-3	-4	-1	-2
444	H	-4	-2	-1	-3	-5	-2	-2	-4	10	-6	-5	-3	-4	-3	-2	-3	-4
445	R	-4	8	-3	-4	0	-1	-2	-3	-2	-5	-4	0	-3	-2	-4	-3	-3
446	D	-4	-4	-1	8	-6	-2	0	-3	-3	-5	-6	-3	-5	-6	-4	-2	-3
447	I	-4	-5	-6	-6	-3	-4	-5	-6	-5	3	5	-5	1	1	-5	-5	-3
448	K	0	0	1	-3	-5	-1	-1	-3	-3	-5	-5	7	-4	-5	-3	-1	-2
449	S	0	-3	-2	-3	0	-2	-2	-3	-3	-4	-4	-2	-4	-5	2	6	2
450	K	0	3	0	1	-5	0	0	-4	-1	-4	-3	4	-3	-2	2	1	-1
451	N	-4	-3	8	-1	-5	-2	-2	-3	-1	-6	-6	-2	-4	-5	-4	-1	-2
452	I	-3	-5	-5	-6	0	-5	-5	-6	-5	6	2	-5	2	-2	-5	-4	-3
453	M	-4	-4	-6	-6	-3	-4	-5	-6	-5	0	6	-5	1	0	-5	-4	-3
454	V	-3	-3	-5	-6	-3	-4	-5	-6	-5	3	3	-4	2	-2	-5	-4	-3
455	K	-2	1	1	4	-5	0	-1	-2	1	-4	-2	4	-3	-2	-3	0	-1
456	N	1	1	3	0	-4	-1	1	0	-3	-4	-4	3	-2	-5	-2	2	-2
457	D	-3	-2	5	5	-1	-1	1	-1	0	-5	-4	0	-2	-5	-1	0	-2

Catalytic loop

Figure 2 This incomplete position-specific scoring matrix (PSSM) shows the scores for residues surrounding the catalytic loop of the DAF-1 serine/threonine protein kinase (GenBank accession NP_499868.2), which was used as the query in a command line PSI-BLAST search against the protein nr database. Scores for the lysines (K) in this region range from 7 for the perfectly conserved K448 to 1 for the poorly conserved K435

database sequences that share those conserved residues can achieve higher scores (lower Expect values) than they would with a position-independent scoring matrix.

Optimally, as the PSSM is further refined in each round, the sensitivity increases. Eventually, no new sequences will be found and the search is said to have converged. In most cases, 10 or fewer iterations are sufficient. However, careful choice of the original query sequence is important for a successful search (Aravind and Koonin, 1999).

Another consideration when running PSI-BLAST is where to set a second Expect value threshold, the inclusion threshold. This value, 0.005 by default, determines which of the database sequences to include in the PSSM. The lower this value, the more stringent the selection of sequences for the PSSM. Since this inclusion value is set somewhat arbitrarily, sequences that fall below it – have a higher Expect value – can be manually included in the next PSSM by using a checkbox on the web page.

2.6. PHI-BLAST

Pattern-hit-initiated BLAST (PHI-BLAST) performs a similarity search like BLASTP, but restricts the search to a subset of the database sequences that contain

a supplied pattern (Zhang *et al*., 1998). As such, it is not an exhaustive search for all proteins that contain a given pattern; a pattern-matching algorithm might be more useful if that is the objective. The advantage of PHI-BLAST is its incorporation of the BLAST algorithm and associated statistics to enhance the likelihood of finding biologically relevant database sequences and to reduce the number of sequences that contain the pattern simply by chance.

A PHI-BLAST search requires both the pattern, in the syntax of PROSITE (Hulo *et al*., 2004), and a query sequence that contains at least one instance of the pattern. On the PHI-BLAST web page, the results of a PHI-BLAST search are formatted for use in one or more rounds of PSI-BLAST searching.

2.7. Translating BLAST

Protein–protein comparisons using the BLASTP algorithm are also possible when a nucleotide query or database is required. The BLASTX program translates a nucleotide query into all six reading frames – three from the forward DNA strand and three from the reverse – then performs six BLASTP searches against a protein database. TBLASTX similarly translates both the nucleotide query and all sequences in the selected nucleotide database. As you can imagine, TBLASTX absorbs immense computational resources when large nucleotide databases are searched; hence, it is not available on all BLAST pages at NCBI. Finally, the TBLASTN program searches with a protein query against a nucleotide database.

2.8. RPS-BLAST

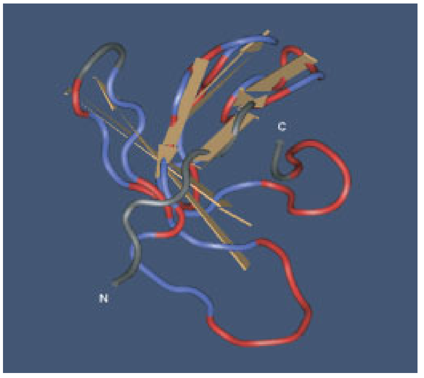
Reverse position-specific BLAST (RPS-BLAST) is so named because a query sequence is searched against a database made up of PSSMs (recall that in PSI-BLAST a PSSM is searched against a database of sequences). The PSSM database is NCBI's CDD mentioned above. This is the program used to identify known, conserved domains in a query sequence.

3. Finding homologs without running BLAST

In many cases, submitting a BLAST search is unnecessary because NCBI has already done the work for you. As new sequences become available in Entrez, they are compared to all others in the database by BLAST. In other words, it is possible to learn a great deal about a sequence, including identification of homologs, simply by following the links in an Entrez record. The *Related Sequences* link displays a list of precomputed BLAST comparisons sorted from best to worst BLAST score. An analogous link, *BLink* (BLAST link), presents the best 200 hits among the same set of comparisons, but graphically displays the extent of the alignment relative to the full length of the query.

BLink also provides a link to the precomputed RPS-BLAST results (CDD-Search). In Entrez, this link is called *Domains*. Another link, *Domain Relatives*,

1NEB in Cn3D



BLAST alignment

```
cort  148 dvgitavALYDYQAAGDDEISFDPDDIITNIEMIDEGWWRGVCR--GAYGLFPANYVEvr 205
1NEB    1 tagkifrAMYDYMAADADEVSFKDGDAIINVQAIDEGWMYGTVQrtGRTGMLPANYVEai 60
Identities =  52% (27/51)
```

Figure 3 BLAST alignment of the solved SH3 domain from human nebulin (1NEB; 60 amino acids) with *Danio rerio* "similar to cortactin" (cort; NP_957335.1; 206 amino acids). The 1NEB structure and the BLAST alignment are color coded by sequence: red = identical residues; gray = no BLAST alignment; arrows denote regions of β-strand. N = amino terminus; C = carboxy terminus

shows the CDART output, immediately revealing other proteins with similar domain architecture.

Another useful link in the BLink output produces the subset of BLAST hits that are in NCBI's structure database, MMDB (Molecular Modeling Database; Chen, 2003). In combination with the structure viewer, Cn3D (Wang *et al*., 2000), you can display the structure for a related protein from the BLink output so that the BLAST alignment – with the original BLink query – is color coded in the Cn3D view. For example, there are currently no structures for a *Danio rerio* SH3 domain. From the BLink output of NP_957335 (Figure 1), you can navigate to the 3D structure of the human nebulin Sh3 domain, and display the BLAST alignment in both the sequence/alignment viewer and the 3D viewer (Figure 3).

4. Homology by structural similarity

Although protein homology has relied heavily on sequence comparison, the underlying assumption is that sequence similarity portends structural similarity, and ultimately, shared function. As more protein structures representing a greater range of protein families are solved, comparisons at the level of 3D structure become more useful (*see* Article 35, **Measuring evolutionary constraints as protein properties reflecting underlying mechanisms**, Volume 7). This is particularly helpful where tools such as BLAST fail to find related proteins because of low sequence similarity (large evolutionary distance). In a general sense, the Vector Alignment Search Tool (VAST) is to structural similarity what BLAST is to sequence similarity (Gibrat

et al., 1996). VAST considers only the secondary structural elements, α-helices and ß-strands, when comparing solved structures, ignoring the sequence. But just as BLAST is used to compute related sequences, new structures entering MMBD get compared to all others by VAST. These precomputed "structure neighbors" are available through links on numerous pages, including the summary page of a structure in MMDB.

Related articles

Article 78, **Classification of proteins into families**, Volume 6; Article 90, **COGs: an evolutionary classification of genes and proteins from sequenced genomes**, Volume 6; Article 93, **Getting the most out of protein family classification resources**, Volume 6; Article 31, **Protein domains in eukaryotic signal transduction systems**, Volume 7; Article 32, **Sequence complexity of proteins and its significance in annotation**, Volume 7; Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7; Article 47, **Phylogenetic analysis of BLAST results**, Volume 7; Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7; Article 94, **Integrated bioinformatics software at NCBI**, Volume 8

References

Altschul SF (1998) *The Statistics of Sequence Similarity Scores*. Available on the World Wide Web: **http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html**.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Aravind L and Koonin EV (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *Journal of Molecular Biology*, **287**, 1023–1040.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.

Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al*. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Research*, **31**, 474–477.

Dayhoff MO, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3, Dayhoff MO (Ed.), National Biomedical Research Foundation: Washington, pp. 345–352.

Geer LY, Domrachev M, Lipman DJ and Bryant SH (2002) CDART: protein homology by domain architecture. *Genome Research*, **10**, 1619–1623.

Gibrat JF, Madej T and Bryant SH (1996) Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**, 377–385.

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915–10919.

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004) Recent improvements to the PROSITE database. *Nucleic Acids Research*, **32**, D134–D137.

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP and Bork P (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research*, **30**, 242–244.

Lipman DJ and Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Research*, **31**, 383–387.

Mulder NJ and Apweiler R (2001) Tools and resources for identifying protein families, domains and motifs. *Genome Biology*, **3**, reviews2001.1–reviews2001.8.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

Persson B (2000) Bioinformatics in protein analysis. *EXS*, **88**, 215–231.

Sansom C (2000) Database searching with DNA and protein sequences: an introduction. *Briefings in Bioinformatics*, **1**, 22–32.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, **1**, 22–28.

Tatusova TA and Madden TL (1999) Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, **174**, 247–250.

Wang Y, Geer LY, Chappey C, Kans JA and Bryant SH (2000) Cn3D: sequence and structure views for Entrez. *Trends in Biochemical Sciences*, **25**, 300–302.

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Research*, **32**, D35–D40.

Wootton JC and Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, **17**, 149–163.

Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV and Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research*, **26**, 3986–3990.

Introductory Review

Integrated bioinformatics software at NCBI

Fabrizio Ferre
Boston College, Chestnut Hill, MA, USA

1. Introduction

The NCBI provides several computational resources that can be accessed on-line from **http://www.ncbi.nlm.nih.gov**, or, in many cases, downloaded and installed locally. A recent review of NCBI services and databases can be found in Wheeler (2004). We divided these tools into three categories on the basis of the type of biological data analyzed: nucleotides, proteins, and three-dimensional structure analysis tools. The user is allowed to start from an entire genome, locate regions of interest (through Map Viewer), retrieve a gene or build a transcript using putative exons (by means of Model Maker), and check whether the putative transcript may be translated into a protein (using ORFfinder); orthologs of the gene can be found in the COGs database. UniGene database stores ESTs that can be analyzed by means of ProtEST, while different tissue-specific EST expression levels can be found using DDD. Functional inference of protein functions through three-dimensional structural comparison may be achieved by means of VAST: the user can compare a protein structure with each domain of all currently known three-dimensional protein structures in the MMDB (Molecular Modeling DataBase), thus finding all structural neighbors of the query protein. These neighbors can be used to identify distant homologs whose sequences have diverged to the point that no significant sequence similarity can be detected. Structural superposition can be visualized using the Cn3D software that can be run as a browser plug-in. When comparing proteins, special care should be given to the fact that proteins are modular molecules: the CDD database and the CDD-search tool provide a way to identify the domain composition of a query sequence. Finally, the CDART web tool allows the user to find proteins in the CDD database sharing the same domain architecture of the protein of interest.

2. Nucleotide sequence analysis

2.1. Map viewer

Complete genome maps, from cytogenetic and physical maps down to the sequence level, are accessible for 23 organisms (at the time of this writing) through the Map

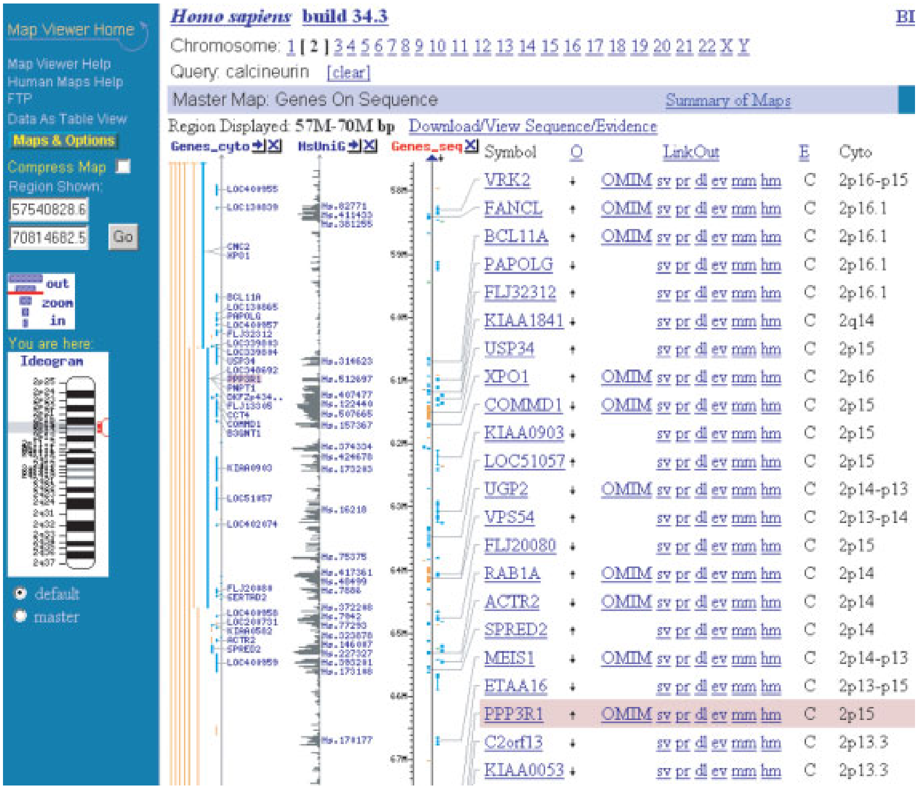


Figure 1 Human calcineurin gene spotted on the chromosome 2 using Map Viewer

Viewer at **http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html**. The maps can be sequence-based or not (e.g., cytogenetic maps or radiation-hybrid maps). Maps showing gene models, EST, and mRNA alignments used to construct gene models are also available (for some organisms). Different levels of detail allow the user to perform searches (keywords, gene names, SNP identifiers, accession numbers), against an entire genome or a specific chromosome, whose hits are shown on chromosome ideograms (an example can be seen in Figure 1). From this genome view, it is possible to access a map view and zoom into progressively more detailed views. Maps are linked to several resources, such as UniGene clusters, LocusLink, Evidence Viewer, and Model Maker.

2.2. *Model maker*

Model Maker can be used for the construction of transcript models by the assembly of putative exons. The available exons may be derived from predictions or from alignments of ESTs or mRNAs to the genomic sequence. Once the transcript is created, potential ORFs (open reading frames) and their translation are shown. mRNA models can be saved locally for use in other programs. Model Maker is accessible from genomic maps displayed with Map Viewer.

2.3. SAGEmap

SAGEmap (**http://www.ncbi.nlm.nih.gov/SAGE/**) is an on-line resource to store, retrieve, and compare Serial Analysis of Gene Expression (SAGE) gene expression profiles (Lash, 2000). SAGE libraries are derived from the Cancer Genome Anatomy Project (CGAP) as well as from GenBank SAGE tags; moreover, SAGEmap accepts user-submitted libraries. The user can search through SAGEmap submitting a sequence or accessing a UniGene cluster or a specific library. The expression level of the user-defined sequence is shown in different SAGE libraries. Finally, different libraries can be compared.

2.4. UniGene, ProtEST, and DDD

UniGene (**http://www.ncbi.nlm.nih.gov/UniGene**) is a system for the automatic clustering of GenBank sequences and ESTs into nonredundant groups; each cluster represents a unique gene. ESTs (Expressed Sequence Tags) are short (up to 600 bases) DNA subsequences of expressed genes, derived from the automated sequencing of randomly selected cDNA clones. For technical reasons, ESTs are prone to errors. Nonetheless, ESTs are extremely useful for the discovery of new genes and as tags for the identification of which genes are expressed in a given cell type or tissue under specific condition, or in particular phases of the cell cycle. The GenBank database hosts a very large section dedicated to ESTs (dbEST). The UniGene project tries to identify all ESTs generated from the same genes, overcoming problems due to the EST sequence errors. UniGene then stores, for a given organism, tissue, organ or pathological condition, libraries of clustered ESTs. ProtEST is a tool that uses BLASTX (*see* Article 93, **Detecting protein homology using NCBI tools**, Volume 8) to search through sequence databases (Swissprot, PIR, PDB, PRF) with possible translations of UniGene clusters. Proteomes from eight organisms (human, mice, rat, *Drosophila*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Escherichia coli*) are used for the comparison, and the best match in each organism is presented to the user. The UniGene cluster containing the human calcineurin B gene, and an example of ProtEST results, are shown in Figure 2. DDD (Digital Differential Display) is a tool for comparing EST-based expression profiles among different UniGene libraries, with the aim of finding genes related to tissue-specific or organ-specific processes, or related to specific pathologies (identifying genes for which the expression levels differ between normal and disease conditions), or related to different development stages. The user must select at least two pools of libraries, then the software outputs those genes whose expression levels are significantly different between each pair of libraries (computed by means of the Fisher Exact test) among the pools.

2.5. ORFfinder

The ORF (Open Reading Frame) Finder (**http://www.ncbi.nlm.nih.gov/gorf/gorf.html**) is a tool for the identification of all ORFs in a user-submitted sequence or



Figure 2 UniGene cluster containing the human calcineurin gene. In the pop-up panel, the ProtEST search result for the first protein in the UniGene cluster is shown

in a sequence in the GenBank database, by locating stop and start codons in all the reading frames. The user can select the minimum and maximum size of the ORF, and can choose among different genetic codes (standard, mitochondrial, bacterial, ciliate, etc.). If an open reading frame is found, the amino acid translation can be used for similarity search by means of BLAST or in the COGs database.

2.6. Electronic PCR

Sequence Tagged Sites (STS) are defined as pairs of oligonucleotide primers that can be used in a PCR assay to identify unique sites in a genome, and have been extremely useful in the construction of genomic maps. The electronic PCR software (e-PCR; Schuler, 1997; URL: **http://www.ncbi.nlm.nih.gov/sutils/e-pcr/**) looks for potential STSs given a pair of PCR primers and a DNA sequence; the tool looks for DNA subsequences that are closely similar to the primers, and checks if order, orientation, and spacing are correct. e-PCR can be used in two ways: forward (searching a STS database with a sequence) or reverse (searching a sequence database with a STS). The forward mode is useful to map a sequence on a genome using a large database of known STSs (UniSTS). The reverse mode is a recent addition to the web server that is used for the prediction of PCR products in a selected genome given one or more pairs of primers.

2.7. VecScreen

VecScreen (**http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html**) is a system for the identification of segments of a nucleic acid sequence that may be the result of a contamination, of vector origin (plasmid, phage, cosmid, YAC DNA) as well as linkers, adapters, and primers. This tool is intended to minimize the incidence and impact of such contaminations in public sequence databases. A specialized nonredundant vector database (UniVec) has been compiled and is used to scan the input sequence. VecScreen performs a BLASTN search using stringent parameters; fragments of vector origin are displayed and classified in four groups (strong, moderate, weak, suspect).

2.8. Spidey

Spidey (**http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/**) is a tool for the alignment of one or more mRNA (FASTA format sequences or accession numbers) on a single eukaryotic genomic sequence, determining the exon/intron structure of the query messenger. This software uses BLAST searches to identify a genome window that covers the entire mRNA length, then refines the alignment to align each exon, taking into account predicted splice sites (four splice-site matrices can be used, i.e., vertebrate, *Drosophila*, *C. elegans*, and plant). The Spidey output is an alignment for each exon, each one evaluated for its quality.

3. Protein sequence analysis

3.1. COG

Orthologs are homologous genes of different species with a common ancestor; finding orthologs is extremely useful for functional predictions in poorly characterized genomes. Tatusov (1997) developed a system to find, cluster, and store

sets of orthologs (Clusters of Orthologous Groups, COGs; URL: **http://www.ncbi.nlm.nih.gov/COG/**). The latest release (Tatusov, 2003) contains information derived from seven complete eukaryotic genomes and 66 prokaryotic genomes. The current eukaryotic section (called KOG) contains 4852 clusters of orthologs, including 54% of known eukaryotic proteins, while the coverage for prokaryotic proteins is higher (around 75%), probably due to higher numbers of analyzed genomes. Each COG consists of individual orthologous proteins (or orthologous sets of paralogs) from three or more genomes. COGs are constructed by means of a mixture of automatic (for finding the orthologs) and manual steps (to split multidomain proteins and to annotate and curate each COG). The integrated COGnitor software allows the comparison of a user-submitted sequence against the database, reporting those COGs to which the input protein may belong. The user can choose the minimum number of organisms represented in the output COGs.

3.2. CDD and CD-Search

The Conserved Domain Database (CDD; Marchler-Bauer, 2003; URL **http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml**) is a collection of protein domain multiple alignments; the database has been built using information derived from Pfam (Bateman, 2002) and SMART (Letunic, 2004), whose multiple alignments have been converted into Position Specific Score Matrices (PSSMs), as well as multiple sequence alignments derived from the COG database that are created and curated by NCBI researchers. CDD is linked to the Entrez retrieval system, and the three-dimensional structural alignments can be visualized by means of Cn3D (see in 3D, **http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml**), a free structure and sequence alignment viewer for NCBI databases, available for Windows, Macintosh, and Unix operating systems. CD-Search (**http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi**) is a web tool that allows the identification of protein domain composition of a query sequence. The query submitted by the user is compared to the CDD PSSMs, by means of reverse-position-specific BLAST (RPS-BLAST; Marchler-Bauer, 2002; *see also* Article 93, **Detecting protein homology using NCBI tools**, Volume 8 for details). See Figure 3 for an example. Hits can be displayed as pairwise alignments with a sequence representative of the domain, or as multiple alignments. The Cn3D viewer can be used to visualize the alignments, and a color coding allows the identification the degree of conservation.

3.3. CDART

The Conserved Domain Architecture Retrieval Tool (CDART; Geer, 2002) is a web-based service that performs similarity searches of the NCBI Entrez Protein Database based on domain architecture (i.e., the sequential order of conserved domains). CDART uses a database of domain definitions and annotations based on a nonredundant protein list; RPS-BLAST is used to assign the domain composition to each protein using the domain definitions from the previously described

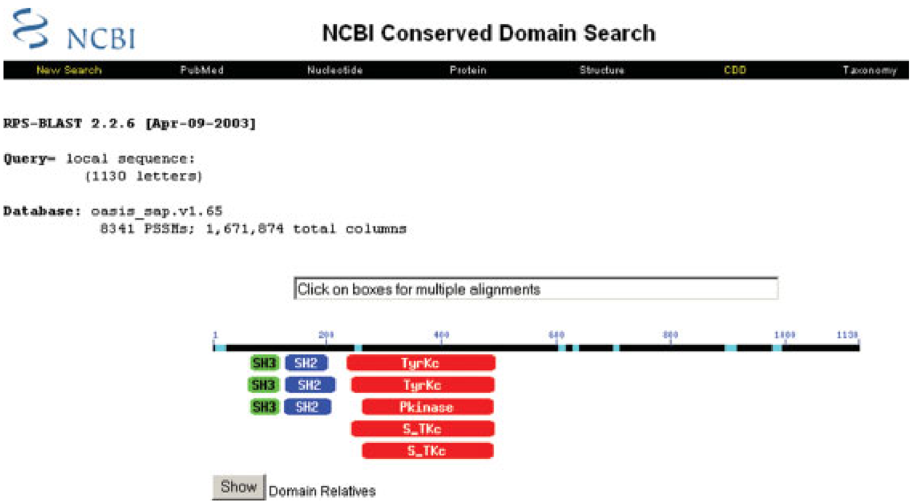


Figure 3 Domains composition of the human C-Abl, determined using CDD-search. Three domains identified (SH3, SH2, kinase catalytic domain)

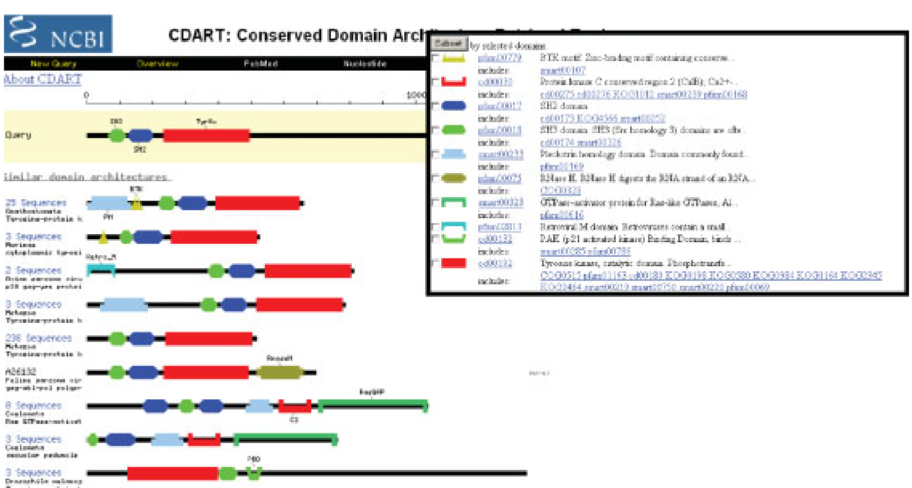


Figure 4 CDART search output. CDART has been used to search for proteins sharing a similar domain architecture as that of the human c-Abl. Each domain is drawn using a different box shape; in the upper right panel the domain legend is shown

CDD (Figure 4). The architectures are recorded in the CDART database along with taxonomic information taken from the NCBI Entrez Taxonomy Database (Wheeler *et al*., 2002). RPS-BLAST is applied to the sequence input by the user, finding domain hits; then, all domain architectures that contain any of the domains in the query are retrieved, and the architectures are compared to that of the query.

CDART is available at **http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi**.

4. Three-dimensional structure analysis

4.1. VAST

Protein functions are strictly dependent on the three-dimensional arrangement of the amino acid residues; the analysis of protein structures is therefore of crucial relevance in the understanding of the molecular mechanisms of cellular functions. The Vector Alignment Search Tool (VAST; Madej, 1995; Gibrat, 1996) algorithm has been developed to compare protein three-dimensional structures. VAST uses a high-level representation of protein structure as an ensemble of secondary structure elements (SSEs): each SSE is represented as a vector. A seed match of structural similarity is found looking for a pair of SSEs (aligned helices or β-strands) whose type and relative orientation is alike; the similarity search is, for that reason, sequence-independent. Therefore, each comparison requires three steps: (1) from the analysis of the three-dimensional coordinates, the SSEs are identified and converted to linear vectors; (2) the SSEs are optimally aligned looking for statistically significant conserved substructures; finally, (3) a Monte Carlo optimization is applied to refine the conformation of each residue. The residue-by-residue alignment refinement step is applied only to the matches of statistical significance, in order to save computation time by focusing only on those similarities that do not occur by chance. The VAST algorithm has been used to compare all-versus-all protein domains stored in the MMDB database (Chen, 2003). Results of precompiled VAST searches, accessible at **http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml**, are ranked by a similarity measure. This measure may be the VAST score (dependent on the number of superimposed SSEs and on the quality of that superposition; higher scores correlate with higher similarity), the p-value (the probability that the match has occurred by chance) or the rmsd (root mean square deviation, a measure of the distance between two sets of points in space, calculated as the square root of the mean square distances between equivalent C-α atoms). VAST search results may be viewed graphically. For each protein in the database, structural neighbors can be retrieved and superpositions can be visualized by means of Cn3D, which allows simultaneous viewing of both the structural and sequence alignments. Figure 5 shows the structural neighbor for the human c-H-ras p21 protein (PDB/MMDB code 5p21). The user can choose the redundancy level, retrieve the ranked list of neighbors, and then select a subset of matching chains; by clicking on the "View 3D structure" button, Cn3D is launched and the structural superposition is visualized; a lower panel shows the sequence alignment of the selected protein chains.

VAST search (URL **http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html**) is the web implementation of the VAST algorithm. It can compare the three-dimensional structure of a newly determined protein structure to those in the MMDB database; the user can choose to compare the query structure only to a

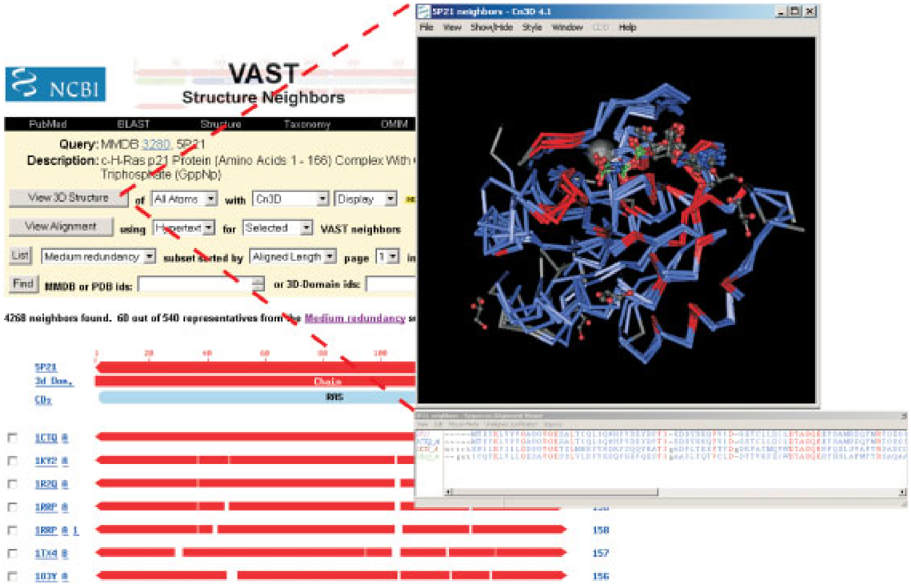


Figure 5 Structural neighbors of the human c-H-ras p21 protein (PDB/MMDB code 5p21). The user can select a subset of neighbors, click on the View 3D Alignment button launching the Cn3D application. The sequence alignment will be shown in a separate window

nonredundant list of structures. VAST Search output is a list of structure neighbors in the VAST format; the user can interactively select one or more structural matches and view the superpositions with Cn3D.

Related articles

Article 29, ***In silico* approaches to functional analysis of proteins**, Volume 7; Article 31, **Protein domains in eukaryotic signal transduction systems**, Volume 7; Article 67, **Score functions for structure prediction**, Volume 7; Article 68, **Protein domains**, Volume 7; Article 70, **Modeling by homology**, Volume 7; Article 93, **Detecting protein homology using NCBI tools**, Volume 8.

Further reading

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

References

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.

Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al*. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Research*, **31**(1), 474–477.

Geer LY, Domrachev M, Lipman DJ and Bryant SH (2002) CDART: protein homology by domain architecture. *Genome Research*, **12**(10), 1619–1623.

Gibrat JF, Madej T and Bryant SH (1996) Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**(3), 377–385.

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ and Altschul SF (2000) SAGEmap: a public gene expression resource. *Genome Research*, **10**(7), 1051–1060.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**, Database issue, D142–D144.

Madej T, Gibrat JF and Bryant SH (1995) Threading a database of protein cores. *Proteins*, **23**(3), 356–369.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Research*, **31**(1), 383–387.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY and Bryant SH (2002) CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, **30**, 281–283.

Schuler GD (1997) Sequence mapping by electronic PCR. *Genome Research*, **7**(5), 541–550.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**(5338), 631–637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(1), 41.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA and Wagner L (2002) Database resources of the national center for biotechnology information: 2002 update. *Nucleic Acids Research*, **30**, 13–16.

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, *et al*. (2004) Database resources of the national center for biotechnology information: update. *Nucleic Acids Research*, **32**, Database issue, D35–D40.

Specialist Review

Bioconductor: software and development strategies for statistical genomics

Robert Gentleman
Dana-Farber Cancer Institute, Boston, MA, USA

Vincent Carey
Brigham and Women's Hospital, Boston, MA, USA

1. Introduction

Recent technological breakthroughs have allowed biologists to investigate the fundamental mechanisms of cellular activity at an extremely fine level of detail. Numerous high-throughput technologies are available for measuring mRNA abundance, DNA copy number, and protein–protein interactions, to name but a few. Additionally, our understanding of the fundamental processes and relationships that underlie these data is growing rapidly. Appropriate analysis of the experimental data relies heavily on the ability to relate it to the existing corpus of biological knowledge as well as on sophisticated statistical analyses and machine learning techniques.

Developing software solutions that are capable of addressing even a small selection of the existent problems is a daunting task. Among the technological challenges are dealing with data size and complexity, fitting appropriate models to the data, reliably ascertaining features that differ between phenotypes, and accurately assessing predictive capabilities of the data and model. The Bioconductor project was devised to provide software solutions for specific data analytic problems and infrastructure to support the development of software solutions by other groups and individuals. The complex and dynamic nature of the problem domain suggests that viable approaches will be based on a large number of interoperable software modules or components. Among the many specific challenges that we have faced are ensuring interoperability, designing and delivering software, and providing substantial and accurate documentation. In this article, we describe and discuss some of the relevant design criteria and our experiences in the first few years of this project.

Our view is very similar to that proposed by Stein (2002), who emphasized web services as the basic computational resource for bioinformatics. We place greater emphasis on the creation and distribution of open source software modules that can

be used independently of the web. With R as a foundation for bioinformatic software development, the Bioconductor project allows, but does not require, developers to deal with issues of web-service design and deployment. R can be embedded in server side software so that services can use R's numerical, statistical, and graphics capabilities, including packages developed in Bioconductor, as needed. Given its excellent XML, SOAP, and general connection-handling facilities, interactive R can function as a richly endowed client to general external web services. An important point is that our modules are open source and hence can be extended by anyone. The same would be true for an open source bioinformatic web service.

There are many competing requirements for any software development project in this area. Substantive requirements include reduction of barriers to entry for cross-disciplinary research. Technical requirements include encouraging extensibility, modularity, transparency, and interoperability. These requirements can conflict, and a successful project must carefully maneuver amongst them.

Perl, Python, and R have all found widespread use by different bioinformatic practitioners. See articles Article 103, **Using the Python programming language for bioinformatics**, Volume 8 and Article 112, **A brief Perl tutorial for bioinformatics**, Volume 8 for background on bioinformatics scripting languages. We chose to use R, partly due to our long history with the S language (Becker *et al*., 1988; Ihaka and Gentleman, 1996) and also because we believe that it has a number of compelling advantages over both Perl and Python. In particular, R has a rich set of statistical algorithms and machine learning algorithms that are well tested. R provides very high level visualization capabilities that are easily extended to provide new tools.

2. Some details

R is an interactive interpreted language generally used from a command-line interface. It is oriented toward statistical and mathematical programming. R has pass-by-value semantics, is vector oriented, has an object-oriented paradigm (inherited from the Lisp family of languages), and the functions are first-class objects. R has lexical scope (Gentleman and Ihaka, 2000) that supports notions of encapsulation.

Like Perl and Python, R is capable of calling routines written in other languages, supports direct database connections, and has support for connections made over `http` and other protocols. Pseudorandom number generation is also important in many aspects of genomic data analysis, and R has a number of well-tested and well-implemented routines that can be used. Concurrent computation, in which one R session or program governs processing on multiple heterogeneous computers, is supported under various parallel computing protocols (e.g., packages such as *snow*, *rpvm*, *rsprng*).

Data structures in R are self-describing and these data structures are used both to store and manipulate the data that is being analyzed as well as the language itself. Because data structures are self-describing, there is no need for type declarations.

R is vector oriented and many computations are *vectorized*, which means that certain operations apply to all elements of a vector or array. Arrays and matrices are simply vectors with appropriate dimension *attributes* set.

```
> x = 1:10
> x

 [1]  1  2  3  4  5  6  7  8  9 10

> x + 1

 [1]  2  3  4  5  6  7  8  9 10 11

> dim(x) = c(2, 5)
> x + 2

     [,1] [,2] [,3] [,4] [,5]
[1,]    3    5    7    9   11
[2,]    4    6    8   10   12
```

Amongst the more powerful of the R language constructs are the subsetting capabilities. Subscripting a vector or matrix is indicated by the use of square brackets (`[]`). Many different types of subscripts are used, positive, negative, and logical subscripts are all allowed. Readers are referred to one of the many texts on the language such as Maindonald and Braun (2004) or Becker *et al*. (1988) for more details.

Nonhomogeneous data are typically stored in the *list* data structure. In many statistical applications, data are of the form samples by variables where the types of the different variables may be different; some may be numerical, others dates or character. The *data.frame* is a data type that was created to accommodate this commonly observed data format. It is essentially implemented as a list of vectors (one for each variable). List subsetting has two forms. The single square bracket operator, `[]`, returns a list of the selected elements. The double square bracket operator, `[[]]`, extracts an element of a list and returns it directly.

Functions are a basic language type, and users can readily examine system functions and easily create their own. As noted above, the semantics are those of pass-by-value so one can act on the arguments to a function as if they were a copy of the supplied value. We note that our simple function below is automatically vectorized by virtue of the fact that the multiplication operation (`*`) is. There is, in essence, no difference between the user-defined function `mysq` and any system function.

```
> mysq = function(x) x * x
> mysq(10)

[1] 100
```

```
> mysq(c(1, 3, 5))
```

```
[1]  1  9 25
```

Of some interest are the set of functions that *apply* a user-supplied function to some particular data structure. These are listed in Table 1. Both `lapply` and `sapply` can be applied to lists and the main difference is that the latter will return a vector, if possible. For `apply`, any matrix or array can be used as the object to apply to, but the user must specify which dimension to collapse over.

```
> le = list(a = c(1, 3, 5), b = 2, c = mean)
> sapply(le, length)
```

```
a b c
3 1 1
```

```
> me = matrix(1:9, nc = 3)
> apply(me, 1, mean)
```

```
[1] 4 5 6
```

3. The anatomy of an R package

The primary method of delivering software tools in R is as packages. A *package* is a collection of software, data, manual pages, and other documentation and resources that have been created and combined, typically to address a particular statistical problem.

R provides substantial infrastructure resources for package creation, maintenance, and unit-testing procedures. Users can create distributable software modules that include documentation, on a variety of levels, and which can be extended by others. The unit-testing features are well developed and greatly aid in quality control. Given our belief that appropriate solutions will be developed in a wide community, any language, or system, that supports that development will need to have substantial package creation and distribution tools.

Table 1 The apply functions

Name	Type of argument
apply	matrix
tapply	ragged array
lapply	list
sapply	list

An R package is a set of files and folders, some of which have specific names. Many more details are given in the R Extension Manual (R Development Core Team, 1999), which is the definitive reference for packages and many aspects of R. We list the most important set of these below:

- `DESCRIPTION`: A file in the main folder that provides various declarative statements about the package. These include its name, the version number, the maintainer, and any dependencies on other packages (as well as several other things).
- `R`: A folder that contains all of the R code for the package.
- `man`: A folder that contains the manual pages for the functions in the package.
- `src`: A folder that contains the source code for any foreign languages (such as C or FORTRAN) that will be used.

The development cycle begins by creating the structure of a package, often using the R function `package.skeleton`. The specific details are then filled in, code added and tested, and manual pages written. Manual pages are stored in a special R language markup that is also described in the R Extension Manual and their creation is often facilitated by the use of the function `prompt`. Among the most important aspects is that the developer provide appropriate examples, demonstrating proper usage of code, on per function basis.

Once the developer is comfortable that their package is ready for distribution, they should engage in unit testing and checking. The R system has a sophisticated, although sometimes cryptic, software verification system. Issues such as consistency between the code (in the `R` folder) and the documentation (in the `man` folder) are evaluated. At the same time, all examples that have been provided are run. Any errors or inconsistencies are reported and the developer should fix the defects and again run the checking system.

Packages that have passed the checking process are ready for others to use. The developer can choose to distribute it or they can upload it to either the CRAN (**www.cran.r-project.org**) or the Bioconductor packages repository (**www.bioconductor.org**), where users can easily obtain a copy. Packages are typically distributed in the form of compressed OS-specific archives.

There are a number of tools provided with R that allow users to automati cally download and install R packages from any of the standard repositories. When a developer updates their package, the version number should increase. That change will alert users to the existence of a new version and will greatly facilitate tracking down errors and inconsistencies. In addition, functionality provided by `update.packages` allows users to detect out-of-date packages and to update all R packages on their computer to a current and mutually compatible set.

Computations programmed in one Bioconductor package routinely depend upon code provided in other packages. As packages evolve independently, it transpires that particular sets of versions of packages may be required for successful interoperation. Version management is a complex task for both developers and users, and software can reduce the complexity and risk of error in creating or maintaining a consistent set of versions. We have developed a new set of routines that simplify a number of aspects of package management. The package *reposTools* includes a definition of an external repository, which includes a table indexing package

Table 2 Some R and Bioconductor packages

Preprocessing	Differential expression	Visualization	Machine learning
affy	limma	geneplotter	class
marray	genefilter	hexbin	cluster
vsn	multtest	Rgraphviz	e1071
PROcess	siggenes	rgl	randomForest

versions and interdependencies, functions that harvest the repository for software addressing certain tasks, functions that collect sets of packages that have mutually consistent versions for interoperation, and tools that allow developers to define and manage their own public repositories.

In Table 2, we list a small number of the very many available R packages that will be of interest to researchers and data analysts working on computational biology and bioinformatics problems. We provide only a small sample of the several hundred packages that are now available. This is not an endorsement of any, but rather a sample selected to have potential interest to many readers. Readers should consult either CRAN or Bioconductor web pages for complete descriptions of the packages and their contents.

4. Documentation

One cannot overemphasize the importance of documentation. Well-written and comprehensive documentation allows the users to make full use of the supplied software. However, early in the Bioconductor project we identified a problem with the usual software documentation paradigm. It is centered on describing the inputs and outputs of specific functions. And while this is essential for using the functions, it does not provide the users with any view of how the different functions can be *used* to carry out a particular analysis.

We developed the vignette concept to deal with the problem of describing complete tasks or analyses. Vignettes use the Sweave system (Leisch, 2002) to provide documents that integrate code and text in a seamless fashion. These documents can be transformed to provide different outputs. In one transformation, the code is evaluated in R and it is replaced by the output, and from this a PDF document is created and users may read this. A second transformation is to extract all of the code chunks into a separate file. Users can examine this file to see what R code was used to create each of the figures, tables, and other outputs reported. To guide users through the sequential evaluation of code chunks, we developed a widget called `vExplorer` in the *tkWidgets* package. Users can open any vignette in the explorer and they are presented with each code chunk in turn; they can examine it, replace some of the code if they desire, and then evaluate it.

All vignettes provided with a package are also checked during the package-checking phase. These vignettes provide a different aspect of quality control. They help to ensure that inputs and outputs of the different functions being employed are kept synchronous and can be used to test more complex interactions.

5. Meta-data packages

Microarray experiments provide information on a specific subset of genes (those with probes on the microarrays that were used). In order to facilitate the use of biological meta-data such as chromosomal location, gene name, gene symbol, Gene Ontology annotations, pathway involvement, and so on, we developed the meta-data package. These are valid R packages that contain the meta-data for a specific microarray chip. The data items are stored in hash tables (called *environments* in R), with the primary key being the manufacturers identifier, in most cases.

Users need only identify the chip that they are using to locate and download the appropriate meta-data packages. The chip-based meta-data packages are auto-generated using tools in *AnnBuilder* (Zhang *et al*., 2003) and are homogeneous with respect to the data that are contained in them. If a user moves from one chip to another (even from oligo-arrays to spotted arrays), the data format remains constant.

6. Design and development issues

6.1. Software reuse

In many cases, one can locate software tools that provide the necessary functionality but which have not been written in R (or whichever language is being used for development). While one can always reimplement the software, this is inefficient and error prone. Wherever possible, software reuse should be the guiding principle. One of the tools that makes this possible is the ability to access code written in other languages (or what might be termed *a foreign function interface*). This interface is very systematized in R, and is documented in the R Extensions manual. Similar functionality is found in most high-level languages.

In creating software for graphs and networks, we relied on two other projects. The BOOST Project (**www.boost.org**) that provides a C++ implementation of many different graph algorithms and Graphviz (**www.graphviz.org**) that provides a C implementation of several different graph layout algorithms. The two R packages, *RBGL* and *Rgraphviz*, provide interfaces to these two substantial pieces of software. We could not have reimplemented these in the time frame available, nor if we had tried would we have been likely to write software that was of the same quality.

By creating interfaces between R and the C and C++ code from these two projects, we have been able to provide the necessary functionality to users, within R, with relatively little effort on our part. Further, with regard to Graphviz, the interactions have sparked some interesting collaborations, as telecommunications scientists have become interested in the specialized layout problems that exist in biological applications.

6.2. Encouraging interoperability

One of the most difficult tasks in distributed software development is promoting and maintaining interoperability. Interoperability is fostered when specific data

structures and software obligations are made explicit and public. Such explicit and public specifications allow different individuals to contribute modules specific to their domains, which can then be reused (or modestly transformed according to precise specifications) for analyses relevant to other domains.

A shared data model is one of the essential ingredients of interoperability. If all modules are written in the same language, then a *class* definition can be made to describe the structure of each data object that is of interest. When different languages are involved, the requirement is that each language be able to query the other languages to determine their view of the data objects. This is generally termed *reflectance* and is most commonly provided by the language's class definitions. However, there are alternatives such as standardized XML markups that are shared.

6.3. Speed and efficiency

In any software project, speed and efficiency can have a myriad of meanings. Scripting languages (*see* Article 103, **Using the Python programming language for bioinformatics**, Volume 8 and Article 112, **A brief Perl tutorial for bioinformatics**, Volume 8) and R have the property that it is relatively easy to create a working version of an idea very quickly. The running time of that early version might not be especially fast, but the development itself can be exceedingly rapid. In a field like bioinformatics and computational biology where the correct way, or even a good way, to carry out specific tasks is generally not known, we believe that this form of efficiency, that is, *efficiency of development*, should take priority.

Once the idea has been proven to work and to be of substantive value, efforts can be made to speed up the software. At this point, we can draw on some of the software engineering code provided by R and carry out substantial profiling experiments. These typically identify specific bottlenecks that can be alleviated either by an altered design or by making use of the foreign function interface and simply calling a compiled language such as C.

7. Examples

We now provide two examples where R/Bioconductor is used to solve some practical and relevant bioinformatic problems.

7.1. Querying the literature

Biological meta-data is constantly evolving. New initiatives are constantly emerging for improved categorization and distribution of data on genomes, SNPs, transcription factors, and so forth. In the subsection on meta-data packages above, we described one method for dealing with meta-data, organizing snapshots of them into the equivalent of local databases. For many of the popular microarrays, these have been precomputed and are available through the Bioconductor project. A different, and complementary, strategy is to provide on-line access to any available data source that provides a web-services interface.

Different initiatives through the NCBI provide well-defined interfaces to a variety of data sources. As part of the *annotate* package, we have developed a set of tools that allow users to access these data, in real time, from within the R session. There are several options, and the interested reader is referred to the documentation in that package for more explicit details. Here we provide code to demonstrate some of the simpler operations that one might want to carry out.

In the code below, we presume that the LocusLink identifiers for the genes we are interested in are contained in the variable named `ids`. We generate the query and store the results in a variable named `x`. This object is of class `XMLDocument` and to manipulate it we will use functions provided by the XML package.

```
> x <- pubmed(ids)
```

```
Loading required package: XML
```

```
> a <- xmlRoot(x)
> numAbst <- length(xmlChildren(a))
> numAbst
```

```
[1] 59
```

Our search of the 59 PubMed IDs (from the 11 Affymetrix IDs) has resulted in 59 abstracts from PubMed (stored in R using XML format). The *annotate* package also provides a `pubMedAbst` class, which will take the raw XML format from a call to `pubmed` and extract the interesting sections for easy review.

```
> arts <- vector("list", length = numAbst)
> absts <- rep(NA, numAbst)
> for (i in 1:numAbst) {
+ arts[[i]] <- buildPubMedAbst(a[[i]])
+ absts[i] <- abstText(arts[[i]])
+ }
> arts[[7]]
```

```
An object of class pubMedAbs
Slot "pmid":
[1] "8121496"

Slot "authors":
[1] "F Kashanchi" "G Piras"    "MF Radonovich" "JF Duvall"
[5] "A Fattaey"   "CM Chiang" "RG Roeder"     "JN Brady"

Slot "abstText":
 The tat gene of the human immunodeficiency virus (HIV)
plays a central...
Slot "articleTitle":
 Direct interaction of human TFIID with the HIV-1 transactivator
```

```
tat....

Slot "journal":
[1] "Nature"

Slot "pubDate":
[1] "Jan 1994"

Slot "abstUrl":
[1] "No URL Provided"
```

Once the abstracts have been assembled, they can be searched by standard search techniques. Suppose, for example, we wanted to know which abstracts include the term `cDNA`. The following code chunk shows how to identify these abstracts.

```
> found <- grep("cDNA", absts)
> goodAbsts <- arts[found]
> length(goodAbsts)


[1] 19
```

So, 19 of the articles relating to our genes of interest mention the term `cDNA` in their abstracts.

Many users find it useful to have a web page created with links for all of their abstracts, leading to the actual PubMed page on-line. These pages can then be distributed to other people who have an interest in the abstracts that you have found. See Figure 1 for an example of the software-generated hypertext abstract listing achievable with the pmAbst2HTML function of the annotate package.

7.2. Networks

New challenges to be faced include the statistical analysis of network/pathway data as well as meta-analyses of several different data sources, possibly derived from different technologies. Through the implementation of packages to handle graphs and graph algorithms within R, we have at our fingertips the tools needed to begin more complex modeling processes.

We will illustrate how simply new tools can be implemented and distributed using the software infrastructure we have developed. We consider the method of Zhou *et al*. (2002) for discovering "transitively coexpressed" genes. This methodology, implemented using data from yeast microarrays, defines a concept of coexpression based on shortest paths in a graph. The nodes of the graph are given by genes and edge lengths are defined by a transformation of correlation in expression values over a collection of diverse experiments. Declarations of coexpression are made when functional relationships between genes on shortest paths are established using Gene Ontology.

Zhou *et al*. (2002) present the procedure in two phases: a validation phase confines attention to genes with known annotation and measures the extent to which transitive coexpression agrees with functional annotation; an application phase uses

BioConductor Abstract List

Article Title	Publication Date
Cloning and characterization of hTAFII18, hTAFII20 and hTAFII28: three subunits of the human transcription factor TFIID.	Apr 1995
Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.	Dec 2002
[Expression, purification and characterization of human prorelaxin-like-protein H2 in Escherichia coli]	Sep 2002
Isolation and analysis of the 3'-untranslated regions of the human relaxin H1 and H2 genes.	Apr 2000
Two human relaxin genes are on chromosome 9.	Oct 1984
Relaxin gene expression in human ovaries and the predicted structure of a human preprorelaxin by analysis of cDNA clones.	Oct 1984
Expression of the human relaxin H1 gene in the decidua, trophoblast, and prostate.	Apr 1991
Molecular characterization and pharmacological properties of the human P2X3 purinoceptor.	Jul 1997
Characterization of recombinant human P2X4 receptor reveals pharmacological differences to the rat homologue.	Jan 1997
The HOXC11 homeodomain protein interacts with the lactase-phlorizin hydrolase promoter and stimulates HNF1alpha-dependent transcription.	May 1998
The human HOX gene family.	Dec 1989
Identification of the alternative splice products encoded by the human protein phosphatase inhibitor-1 gene.	Mar 2002
Identification and characterization of a novel protein inhibitor of type 1 protein phosphatase.	Nov 2000
Isolation and characterization of full-length cDNA clones for human alpha-, beta-, and gamma-actin mRNAs: skeletal but not cytoplasmic actins have an amino-terminal cysteine that is subsequently removed.	May 1983
Isolation and characterization of cDNA clones for human skeletal muscle alpha actin.	Jun 1983
Large-scale concatenation cDNA sequencing.	Apr 1997
Cloning and sequencing of the cDNA encoding human P5.	Oct 1995

Figure 1 The browser displaying the linked PubMed citations

transitive coexpression to predict functions of relatively unstudied genes. In this summary, we focus on the validation phase, which includes a number of tuning parameters. The basic data objects of the analysis are

- $M = [m_{ij}]$, a matrix of expression experiment results with i indexing genes and j indexing experiments,
- $C(a, b)$, a function that computes a robust correlation coefficient between two vectors a and b as the minimum in absolute value of the Pearson correlations,
- K_c, the complete graphs on all yeast genes in cellular component $C \in$ {mitochondria, cytoplasm, nucleus}, with edge between genes g_p and g_q initially assigned length $C(m_{p\cdot}, m_{q\cdot})$,
- the GO DAGs for biological process and cellular component.

We will focus attention on a single cellular component and thus drop the subscript for K; the same process is used for each component. The algorithm is:

1. Form K' by removing edges in K with length less than $\tau = 0.6$.
2. Recompute K' edge length $w_{new} = [1 - w_{old}]^k$, where $k = 6$.

3. Stratify GO annotations into informative and uninformative, with an informative term satisfying the conditions that (a) it is used for annotation at least $\gamma = 30$ times and (b) all immediate refinements of the term are used less than γ times.
4. Enumerate all shortest paths between all pairs of genes (denoted *termini*) annotated to informative GO terms.
5. Discard paths of length greater than $\sigma = 0.008$.
6. Characterize genes along the path between the termini as L0 matches if they are annotated to the same term as the termini, or as L1 matches if they are annotated to a term that shares the same immediate ancestor with the term to which the termini are annotated.
7. Compute permutation-based null distributions of "match ratios" (ratio of number of L0 or L1 matches to total number of transitive genes) for significance appraisal.

This procedure can be implemented fairly directly using basic resources in Bioconductor (see the *GOstats* package). The expression matrix M is best represented as an *exprSet*. The robust correlation function can be straightforwardly coded in R or C; note that `cov.rob` implements high-breakdown correlation estimation, which may be more effective than the leave-one-out procedure adopted by the authors. Given a distance structure (readily computable by any correlation function), the *graph* package *distGraph* class computes K and the `threshold` method computes K' from K. Edge weight transformation is performed by `apply`. Shortest-path computations are carried out through the interface to the Boost graph library, using either `dijkstra.sp` or `sp.between`; the enumeration of pairs of termini is efficiently computed by the *combinat* package. Operations on the GO graph and its association to yeast genes can be carried out directly using package *graph*, or by using an object-ontology complex as defined in *ontoTools*. *ontoTools* supports computation of semantic similarity of ontology terms, facilitating more fine-grained measurement of degree of "functional match" among transitive genes than the L0/L1 dichotomy. Computational inference via node relabeling is straightforwardly carried out using aspects of the *graph* representation of the *graph* package.

A key point to keep in mind when considering the Bioconductor approach to bioinformatic data analysis is the commitment to transparency and extensibility of the analytic tool. We have described separately implemented structures or modules that are plugged together to perform the analysis for a generically characterized input matrix M. Distance computations and thresholding procedures should be kept separate and all tuning parameters should be exposed to facilitate sensitivity analysis. The only special-purpose programming required is the "glue", the top-level function that accepts user parameters and coordinates the application of the different components. The glue serves as a precise specification of the high-level computation, and the individual components are separately documented and verified according to Bioconductor principles.

Through the application of well-designed modular software, this interesting idea can be turned into a functional program in an afternoon. The end result is more general than the original description and can be shared as the author desires.

8. Discussion

Our proposal, and the direction of the Bioconductor project, is for the creation of highly interoperable software modules that are easy to obtain, platform neutral, and self-documenting. A web-services model provides a different abstraction of essentially the same idea. Web services are conceptually quite similar to software packages or modules, and virtually the same set of problems needs to be addressed by either approach. When web services are open source, then they can be extended and modified to suit local purposes. But, in some ways, they are more complex than packages. Web services are language-neutral black boxes for the users but they are not for developers; each one typically has its own idiosyncrasies. By choosing to use a common language and an open source development model, we hope to make it easier for others to adopt, adapt, and extend our work.

Data analysis software components that are transparent, that are sufficiently generic to warrant reuse in multiple contexts, and that allow analysts to maintain fidelity to the biological concepts of interest are at a premium in furthering objective scientific progress at this crucial juncture. While we are in some ways agnostic to the delivery mechanism (different investigators will have good reasons to use virtually every language that is at all relevant), we do believe that the S language is well suited to this domain of application.

Different developers will always have different ideas on how to solve a given problem. We are convinced, however, that modular and interoperable designs and deployments are achievable at acceptable cost by most researchers and developers in this field, regardless of the details of tools and methods chosen. The standards of transparency and explicit reproducibility that are well accepted in wet-lab science have not propagated to the in silico domain of computational biology. The Bioconductor project is an attempt to demonstrate that these standards can be met in effective and widely used analytic software for bioinformatics. Bioconductor would not be possible but for the extraordinary dedication and ingenuity of its core members and its active user community. To them we offer heartfelt thanks.

Acknowledgments

The Bioconductor project is supported by the High Tech Industry Multidisciplinary Research Fund at the Dana-Farber Cancer Institute, the Catt Family Foundation's gift to the Dana-Farber Cancer Institute, the Women's Cancer's Program at the Dana-Farber Cancer Institute, and US NIH Grant #1R33 HG002708: "A Statistical Computing Framework for Genomic Data".

References

Becker RA, Chambers JM and Wilks AR (1988) *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole: Pacific Grove.

Gentleman R and Ihaka R (2000) Lexical scope and statistical computing. *Journal of Computational and Graphical Statistics*, **9**, 491–508.

Ihaka R and Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Leisch F (2002) Sweave: dynamic generation of statistical reports using literate data analysis. In *Compstat 2002 – Proceedings in Computational Statistics*, Härdle W and Rönz B (Eds.), Physika Verlag: Heidelberg, pp. 575–580. URL **http://www.ci.tuwien.ac.at/~leisch/Sweave**. ISBN 3-7908-1517-9.

Maindonald J and Braun J (2004) *Data Analysis and Graphics Using R*, Cambridge University Press.

R Development Core Team (1999) *Writing R Extensions*, R Foundation: Vienna.

Stein L (2002) A web-services model will allow biological data to be fully exploited. *Nature*, **417**, 119–120.

Zhang J, Carey V and Gentleman R (2003) An extensible application for assembling annotation for genomic data. *Bioinformatics*, **19**, 155–156.

Zhou X, Kao M-CJ and Wong WH (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12783–12788.

Specialist Review

RNA secondary structure prediction

David H. Mathews

University of Rochester Medical Center, Rochester, NY, USA

Michael Zuker

Rensselaer Polytechnic Institute, Troy, NY, USA

1. Introduction

RNA is increasingly being discovered to play many roles in Biology aside from its two functions in the Central Dogma. RNA is now known to have roles in peptide bond catalysis (Hansen *et al*., 2002), RNA editing and intron splicing (Doudna and Cech, 2002), protein shuttling (Walter and Blobel, 1982), immunity (Cullen, 2002; Lau *et al*., 2001), development (Lagos-Quintana *et al*., 2001), dosage compensation (Panning and Jaenisch, 1998), posttranscriptional gene regulation (Wellbacher *et al*., 2003), and RNA modification (Bachellerie *et al*., 2002). Novel catalytic RNA sequences can be evolved in vitro to extend the functionality of nucleic acid sequences (Bittker *et al*., 2002). RNA is also a target for pharmaceutical design either through antibiotics that target rRNA, antisense oligonucleotides (Dias and Stein, 2002), mis-splicing redirection (Sazani and Kole, 2003), or catalytic ribozymes (Long *et al*., 2003). Because of the relationship between structure and function, a detailed understanding of the structure of an RNA is needed for understanding the mechanism of action of that RNA.

RNA structure is commonly divided into levels of complexity. The primary structure is the sequence of nucleotides; secondary structure is the sum of canonical pairs, that is, Watson–Crick and G·U pairs; and tertiary structure is the three-dimensional arrangement of atoms. In general, the base pairs of secondary structure are significantly more stable than tertiary contacts (Brion *et al*., 1999); therefore, in practice, the secondary structure can largely be determined independently of tertiary structure.

The accepted standard for determining RNA secondary structure, in the absence of a crystal structure, is comparative sequence analysis. A recent study demonstrated that 97% of base pairs predicted on the basis of comparative sequence analysis are found in the high-resolution crystal structures of ribosomal RNA (Gutell *et al*., 2002). Accurate comparative sequence analysis, however, requires a large number of homologous sequences often obtained from multiple species or

from in vitro evolution. Even when a large number of homologous sequences is available, the alignment and deduction of secondary structure can be exceedingly difficult, as in the case of eukaryotic RNase P (Frank *et al*., 2000).

In the absence of a large number of related sequences, methods have been developed for predicting RNA secondary structures. This article reviews many of the current methods for secondary structure prediction, including methods for finding common structures in related sequences.

2. Nearest-neighbor parameters

Most secondary structure prediction algorithms use the principle of free energy minimization to predict base pairs, where the free energies of RNA secondary structures are predicted using a set of nearest-neighbor parameters based on optical melting experiments (Mathews *et al*., 2004; Mathews *et al*., 1999; Xia *et al*., 1998). Base pairs contribute favorable free energy increments. Loop regions contribute unfavorable free energy increments for initiation, largely because of the entropic penalty associated with restricting nucleotides into loop regions, and favorable increments that are sequence specific.

A recent revision to the free energy rules (Mathews *et al*., 2004) accounts for data from several recent experimental studies for bulge loops of a single nucleotide (Znosko *et al*., 2002), hairpin loops (Dale *et al*., 2000), multibranch loops (Diamond *et al*., 2001; Mathews and Turner, 2002b), and internal loops (Schroeder and Turner, 2000; Schroeder and Turner, 2001). The recent hairpin loop optical melting data generally supports the accepted model for predicting hairpin loop stability (Dale *et al*., 2000). The bulge loop study is the first systematic study of the effect of sequence for the stability of single nucleotide bulges by optical melting (Znosko *et al*., 2002). The multibranch loop studies present the first quantitative data for multibranch loop free energies in RNA, and are used as a basis for the revision of the equation used to predict multibranch loop initiation free energies (Diamond *et al*., 2001; Mathews and Turner, 2002b). Finally, the internal loop studies demonstrate idiosyncrasies in the sequence dependence for internal loop stability (Schroeder *et al*., 1999; Schroeder and Turner, 2000; Schroeder and Turner, 2001). For example, the G·G mismatch has a favorable free energy increment in 2×3 internal loops, but is worth no extra stability in $1 \times n$ internal loops, where n is greater than 2.

3. Free energy minimization by dynamic programming

Given that the free energy for an RNA secondary structure can be estimated for any sequence, several algorithms have been devised to predict a secondary structure from sequence. The naïve approach would be to explicitly generate all possible secondary structures for the sequence, evaluate their free energies, and then find the lowest energy structure. But, the number of possible structures increases exponentially with the length of the sequence, making this approach infeasible. Dynamic

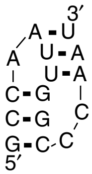


Figure 1 A simple pseudoknot. This secondary structure illustrates the simplest pseudoknot, formed when the nucleotides in a hairpin loop pair to nucleotides either 3′ or 5′ to the hairpin loop stem

programming algorithms were the first practical algorithms used to solve this problem (Nussinov and Jacobson, 1980; Zuker, 1989; Zuker and Stiegler, 1981).

The important feature of dynamic programming is that recursion is used to determine folding free energies. For the most popular algorithms, the recursions result in a calculation that is O(N^3), where N is the number of nucleotides (Hofacker *et al*., 1994; Mathews *et al*., 2004). This means that a doubling in the length of the sequence results in an eightfold increase in calculation time. Intermediate energies are calculated for structures restricted to subsequences of the entire sequence, resulting in O(N^2) storage. To maintain this scaling, these algorithms are unable to predict structures containing pseudoknots, illustrated in Figure 1. The currently available dynamic programming algorithms for RNA secondary structure prediction also limit the size of internal loops. Lyngsø *et al*. (1999) developed a method for predicting internal loops in O(N^3) time without having to limit the size of internal loops.

Modern implementations of energy minimization dynamic programming algorithms predict not only the lowest free energy structure but also an ensemble of structures with free energy close to the minimum, called *suboptimal structures* (Mathews *et al*., 2004; Zuker, 1989; Zuker, 2003). These structures, sampled heuristically, demonstrate other base pairs that are alternatives to the pairs in the lowest energy structure. In addition, the information contained in suboptimal structures can be summarized with a two-dimensional energy dot plot that shows, for each possible i–j pair, the lowest free energy possible for a structure containing that pair (Jaeger *et al*., 1989; Zuker, 1989). Suboptimal structures are important because some RNA sequences populate more than one structure in solution (Schultes and Bartel, 2000) and because of the errors inherent in the nearest-neighbor parameters.

Currently, predicted RNA secondary structures contain, on average, 73% of known base pairs for RNA sequences divided into domains of less than 700 nucleotides (Mathews *et al*., 2004). On average, the best of 750 suboptimal secondary structures contains 86% of known base pairs. This accuracy can be improved by using data derived from experiments (Mathews *et al*., 2004; Mathews *et al*., 1999) using single-stranded or double-stranded specific enzymes, flavin mononucleotide cleavage that can reveal Us in G·U base pairs (Burgstaller *et al*., 1997), or chemical modification data from either in vivo or in vitro probing.

Dynamic programming algorithms for RNA secondary structure prediction by free energy minimization are available from three sources. Mfold is a program for on-line folding or compilation on Unix or Linux (Zuker *et al*., 1999) and can be accessed at **http://www.bioinfo.rpi.edu/applications/mfold** (Zuker, 2003).

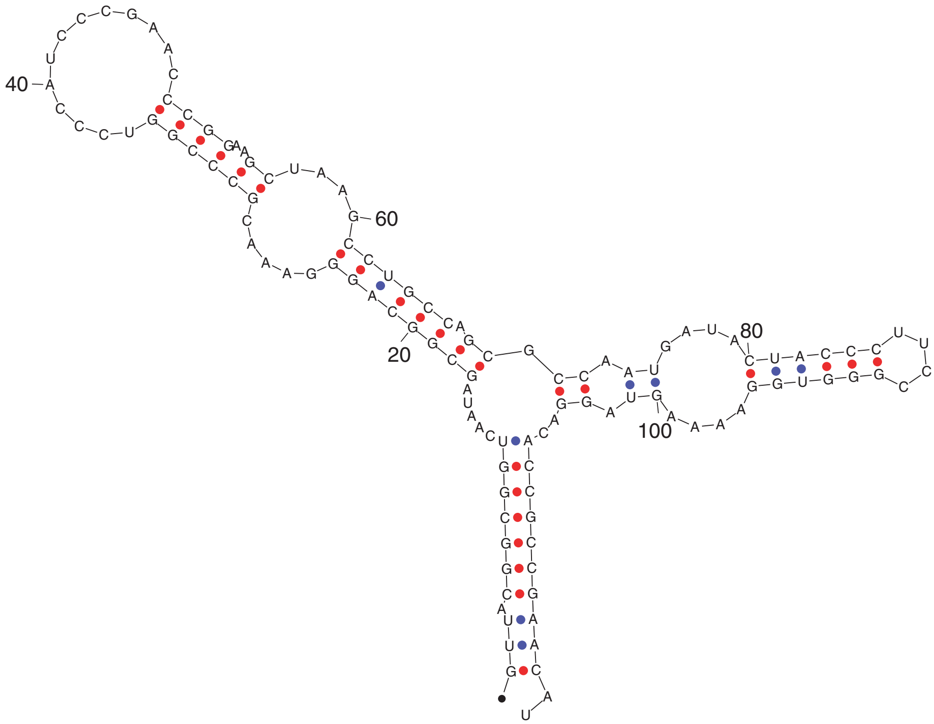


Figure 2 Secondary structure prediction of *M. phlei* 5S rRNA. This is sample output for secondary structure prediction by the mfold Web Server. The sequence was taken from the 5S rRNA database (Szymanski *et al*., 2000)

Figure 2 illustrates the secondary structure prediction output from the Mfold server. The Vienna RNA Package includes RNAfold for RNA secondary structure prediction. It is available as an on-line server, as a code for local compilation, or as Microsoft Windows binaries at **http://www.tbi.univie.ac.at/~ivo/RNA/** (Hofacker, 2003). RNAstructure is a program for the Microsoft Windows environment that uses a graphical user interface and includes RNA secondary structure prediction (Mathews *et al*., 1999). It is available for download from the Turner lab website, **http://rna.urmc.rochester.edu/RNAstructure.html**. The three RNA folding packages differ slightly in their implementation of the nearest-neighbor energies in multibranch loops and exterior loops; therefore, different structures can sometimes be predicted by each program.

Variations on the standard dynamic programming algorithm for free energy minimization are available for specialized calculations. The Vienna RNA Package includes a program, RNAsubopt, for generating all suboptimal secondary structures for a sequence within a user-defined energy increment from the lowest free energy (Wuchty *et al*., 1999). Although the number of suboptimal structures increases exponentially with increasing size of the energy increment, this algorithm can reasonably sample, for example, all 295 722 suboptimal structures within 12 kcal mol^{-1} of the minimum free energy conformation for a sequence of 75 nucleotides (Wuchty *et al*., 1999).

An interesting dynamic programming algorithm was introduced by Rivas and Eddy (1999) demonstrating that a broad class of structures containing pseudoknots can be predicted by dynamic programming in polynomial time. Their algorithm, called PKNOTS, can predict a subset of pseudoknotted structures that sufficiently covers all topologies of practical interest, although the general problem of finding any possible pseudoknot has been shown to be NP-complete (Lyngsø and Pederson, 2000). The drawback is that the algorithm is O(N^6) in time and O(N^4) in storage, making it intractable for sequences much longer than a few hundred nucleotides. The code is available for download at **http://www.genetics.wustl.edu/eddy/software/**.

4. Other secondary structure prediction algorithms

The problem of finding the lowest free energy structure for an RNA sequence has also been approached by the use of a genetic algorithm. Secondary structures are randomly mutated, and those mutations that are most “fit”, on the basis of free energy, are kept for future rounds of mutation. “Unfit” mutations are discarded. At some steps, solutions are crossed from multiple structures to find new structures and are evaluated on the basis of free energy. The advantage of one such program, called STAR (Gultyaev *et al*., 1995), is that folding pathways can be modeled using a sequence that lengthens from the 5′ end in the same way a sequence elongates from the 5′ end during transcription. Predicted helices with low free energy are less likely to rearrange as the sequence elongates. Computational experiments show that STAR is more accurate when a sequence is folded as it is elongated from 5′ to 3′ than when the entire sequence is folded at once. This suggests that kinetic traps of stable structure are important for at least some RNA sequences. STAR is available at **http://wwwbio.leidenuniv.nl/∼Batenburg/STROrder.html**. The drawback of these calculations is that multiple calculations on the same sequence can converge to different structures with different energies.

A two-dimensional polymer theory model that can construct an energy landscape for RNA folding was developed (Chen and Dill, 2000). This method uses the helical nearest-neighbor parameters (Xia *et al*., 1998), but extrapolates energies for loop regions from chain entropies. The resulting landscapes suggest that RNA folding may be fundamentally different than protein folding because the RNA landscape may be more rugged.

Stochastic context-free grammars have also been adapted to the prediction of RNA secondary structure.

5. Partition function calculation

McCaskill (1990) presented a novel dynamic programming algorithm that calculates the partition function for RNA secondary structure and that can predict the probability of base pairing for all possible base pairs. This algorithm is O(N^3) in time and O(N^2) in storage. It has been implemented with recent sets of thermodynamic parameters in the Vienna RNA package (Hofacker, 2003; Hofacker *et al*., 1994)

and in RNAstructure (Mathews, 2004). Figure 3 illustrates a probability dot plot, used for displaying all base pair probabilities, as predicted by the Vienna Web Server.

Recently, the partition function method was extended to explicitly include pseudoknots using the dynamic programming (Dirks and Pierce, 2003). Rivas and Eddy (1999) determine the minimum free energy for any secondary structure on subsequence from nucleotides *i* to *j* (including pseudoknots) as the minimum over several cases. These cases are exhaustive, but not exclusive, that is, there are secondary structures S that are considered by more than a single case. When extending

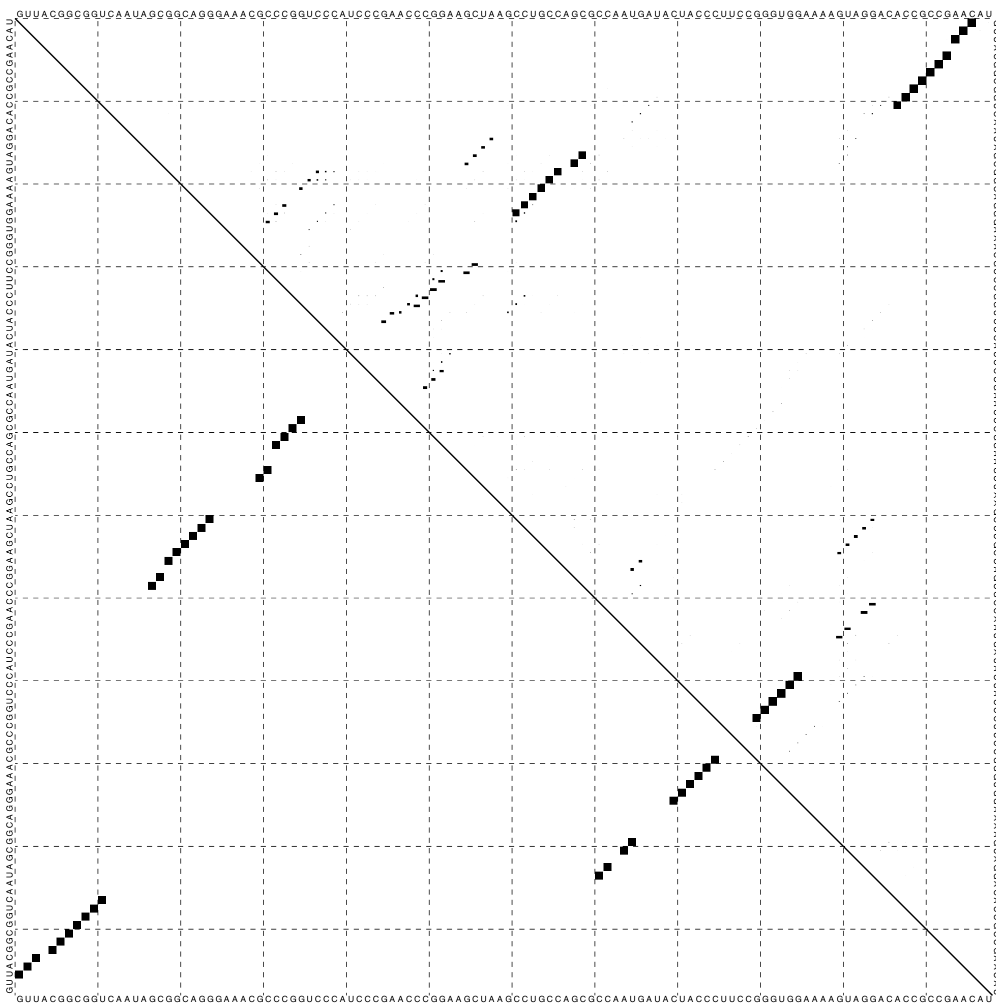


Figure 3 Probability dot plot for *M. phlei* 5 S rRNA as predicted by the Vienna Package Web Server. The lower triangle contains dots representing base pairs in the predicted lowest free energy structure. The upper triangle illustrates the probability of the ensemble of possible canonical base pairs with the area of the dot in proportion to the base pairing probability

energy minimization to the computation of the partition function, it is necessary to first devise recursions where cases are exhaustive and exclusive (or nonredundant). Dirks and Pierce (2003) do this and subsequently extend the algorithm to a computation of the partition function. This algorithm nicely incorporates the work of Lyngsø *et al*. (1999) to predict internal loops of any size using $O(N^3)$ steps.

The partition function predicts base pair probabilities, but does not, in itself, provide a method for the prediction of secondary structures. Ding and Lawrence (2003) have developed an elegant method for sampling secondary structures from the partition function to derive a statistically valid ensemble of structures. They have demonstrated, using a statistically valid sample of structures, that features dependent on secondary structures, such as binding accessibility, can be calculated that correlate with experimental measures (Ding and Lawrence, 2001).

6. Predicting a secondary structure common to two sequences

When two or more homologous RNA sequences are available, it is assumed that both sequences will adopt the same (or very similar) secondary structure. Many algorithms have become available in the last 10 years to take advantage of the auxiliary data provided by more than a single sequence. In general, these algorithms are divided into those that use a fixed alignment and those that do not require a fixed alignment. The algorithms that do not assume a fixed alignment are more robust, but take significantly more calculation time.

6.1. *Methods assuming a fixed alignment*

Two distinct approaches exist for finding a structure common to a fixed multiple sequence alignment. The first approach is to predict the structures for each sequence individually and then postprocess these data based on a sequence alignment (Lück *et al*., 1999). The ConStruct algorithm, by Lück *et al*. (1999) calculates the base pairing probabilities for each sequence using the partition function algorithm from the Vienna Package (Hofacker *et al*., 1994). The base pairing probabilities are assembled into a single consensus probability matrix according to a multiple sequence alignment and then a graphical user interface allows the user to adjust the alignment manually to optimize the base pairing probabilities. This algorithm is $O(SN^3)$, where S is the number of sequences and N is roughly the average sequence length. The source code for ConStruct is available at **http://www.biophys.uni-duesseldorf.de/local/ConStruct.html**.

The second approach to finding a common structure in aligned RNA sequences is to use the sequence alignment to constrain a single secondary structure prediction for multiple sequences (Hofacker *et al*., 2002). The Vienna RNA package includes a program called RNAalifold that uses the free energy nearest-neighbor parameters to predict the base pairing probabilities and the minimum free energy structure for a consensus sequence derived from a multiple alignment. This algorithm is $O(N^3)$ in time where N is the total length of the sequence alignment, including nucleotides and gaps.

A specialized stochastic context-free grammar has been developed to predict a "most probable" alignment together with a common fold for a group of homologous RNAs with reasonably similar sequences (Knudsen and Hein, 1999). It has been implemented in a software package named PFOLD.

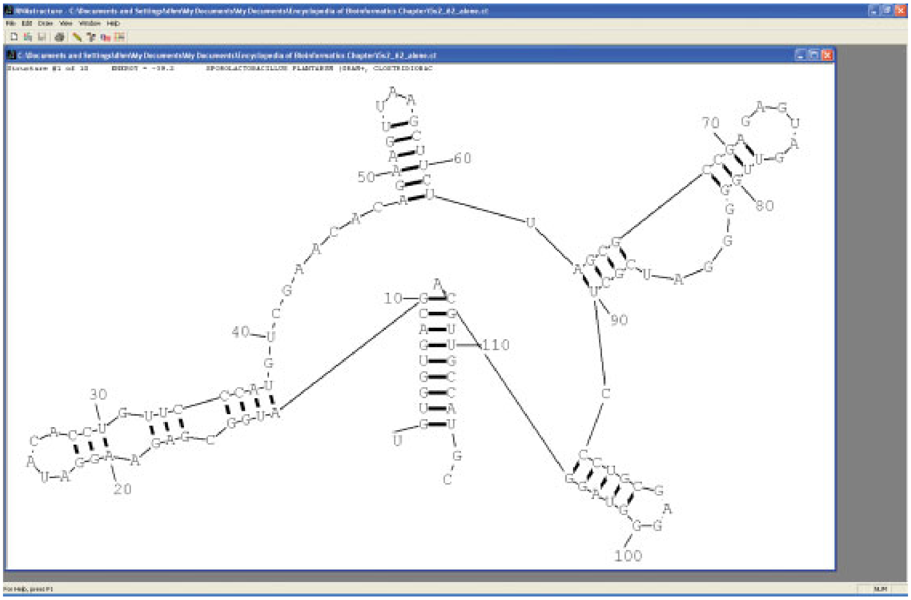
6.2. *Methods that simultaneously predict secondary structure and alignment*

Two distinct computational approaches exist to simultaneously find the common structure and sequence alignment of more than one sequence. The first proposed solution is a dynamic programming algorithm (Sankoff, 1985), but this method is in general $O(N_1^3 \times N_2^3 \times N_3^3 \ldots)$ where N_x refers to the length of sequence x. Therefore, practical solutions are currently limited to two sequences at most.

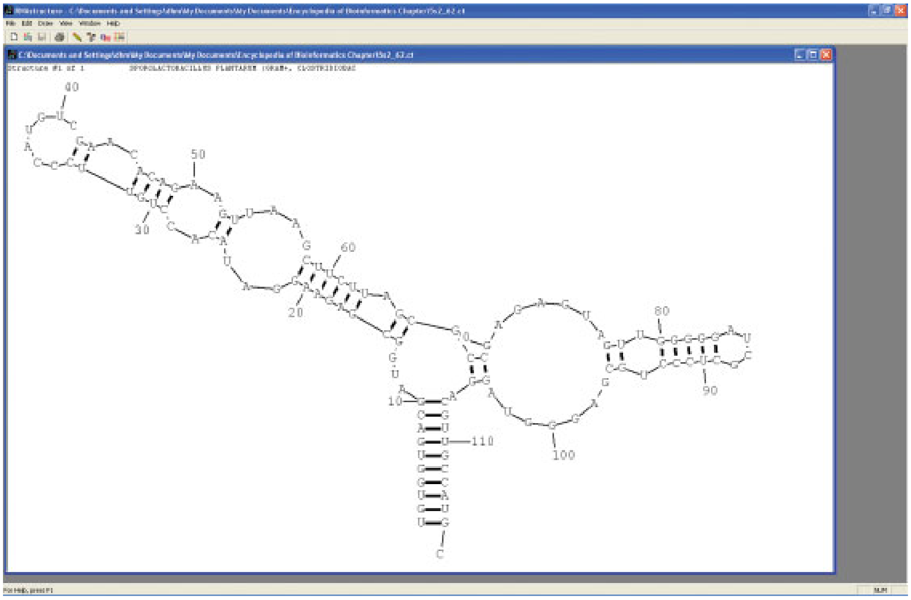
FOLDALIGN is a dynamic programming algorithm that finds local structure and alignment in two RNA sequences by maximizing a score that favors sequence similarity, base pairs, and compensating base changes (Gorodkin *et al*., 1997). The algorithm does not allow multibranch loops in order to limit the scaling in time to $O(N_1^2 N_2^2)$. Multiple pairwise predictions can be assembled into a multiple alignment using a greedy algorithm, a heuristic method common in computer science. An on-line server and the source code for FOLDALIGN are available at **http://www.bioinf.au.dk/FOLDALIGN/**.

A second dynamic programming algorithm, Dynalign, finds the lowest free energy structure common to two sequences and the sequence alignment, using a complete set of nearest-neighbor parameters (Mathews and Turner, 2002a). It allows multibranch loops, but remains $O(N^3 M^3)$ with $M \ll N$, where M restricts the set of possible alignments and N is the shorter of the two sequences. For example, for nucleotide a_i from sequence one to align with b_j in sequence two, $|i - j| \leq M$. The algorithm was tested using tRNA and 5 S rRNA sequences that are, on average, poorly predicted using a single sequence. For 156 pairwise structure predictions of 13 tRNA sequences, Dynalign predicted, on average, 86.1% of known base pairs as compared to an average of 59.7% accuracy for free energy minimization of a single sequence. Similarly, for seven 5 S rRNA sequences, the accuracy of structure prediction improved from 47.8 to 86.4% by using pairwise structure prediction with Dynalign. Dynalign is available as part of the RNAstructure package for Microsoft Windows or for local compilation by download at **http://rna.urmc.rochester.edu**. Figure 4 demonstrates the utility of Dynalign with an example drawn from 5 S rRNA.

A second solution to the problem of simultaneous folding and alignment is a genetic algorithm (Chen *et al*., 2000). This algorithm is capable of handling multiple sequences, and is $O(n^2 m^2 S^2)$, where n is the maximum number of stems for each structure, m is the maximum number of structures considered for each sequence, and S is the number of sequences. On average, 87.7% of known base pairs were correctly predicted for a set of 20 tRNA sequences and 95.3% of known base pairs were correctly predicted for 25 5 S rRNA sequences. This compares favorably to an average accuracy of 83.0% and 77.7% for tRNA and 5 S rRNA, respectively, for free energy minimization of single sequences. The



(a)



(b)

Figure 4 Dynalign improves the accuracy of secondary structure prediction by using mutual information in two sequences. (a) shows the secondary structure predicted for the *S. plantarum* 5 S rRNA sequence by standard free energy minimization. Note that this structure is poorly predicted. For example, it has five branches from the central multibranch loop instead of the known 3-way multibranch loop. (b) shows the same 5 S rRNA sequence, with the secondary structure predicted simultaneously with the *M. phlei* secondary structure using Dynalign. This predicted structure is identical to the structure determined by comparative sequence analysis (Szymanski *et al*., 2000)

Fortran 77 code for this algorithm is available for download from **ftp://ftp.ncifcrf.gov/pub/users/chen/rnaga.tar.Z**.

7. Overview of methods for multiple sequences

Each of the algorithms described here has distinct advantages. For short sequences, Dynalign provides a robust method for predicting a sequence alignment and common structure for two sequences at a time. It does not require any sequence similarity in the two sequences. Dynalign, however, does not scale well to sequences of over a couple of hundred nucleotides long. For longer sequences with low sequence similarity that are difficult to align on the basis of sequence only, FOLDALIGN can be used to find common pairs and a coarse alignment for multiple sequences. Alternatively, the genetic algorithm also provides a computationally tractable means of finding a common structure and the associated alignment for multiple sequences. ConStruct provides a graphical user interface to adjust sequence alignments manually and could, for example, take the predicted alignment from the genetic algorithm or FOLDALIGN as a basis for further manual refinement. Finally, RNAalifold is the fastest method for finding a common structure in aligned sequences. It is the best choice for finding the structure common to multiple sequences that share enough similarity to be well aligned or to multiple sequences too long to be efficiently handled by other methods.

8. Summary

Many tools are currently available for the prediction of RNA secondary structure. The last 10 years have seen both an improvement in the free energy nearest-neighbor parameters and in the breadth of secondary structure prediction tools available. On average, it is reasonable to expect about 73% average accuracy at predicting base pairs for sequences less than 700 nucleotides. Furthermore, given the completion of many whole-genome projects, it should become increasingly easy to find homologous RNA sequences. The recent work in developing algorithms to predict structures common to multiple sequences promises to improve the accuracy of secondary structure prediction.

Further reading

An excellent review of predicting the conformational free energy of an RNA secondary structure at 37° with nearest-neighbor parameters is contained in:

Turner DH (2000) Conformational changes. In *Nucleic Acids*, Bloomfield V, Crothers D and Tinoco I (Eds.), University Science Books: Sausalito, pp. 259–334.

Step-by-step instructions for secondary structure prediction using *mfold* or RNAstructure are available in:

Mathews DH, Turner DH and Zuker M (2000) RNA secondary structure prediction. In *Current Protocols in Nucleic Acid Chemistry*, Vol. 11, Beaucage SL, Bergstrum DE, Glick GD and Jones RA (Eds.), John Wiley & Sons: New York, pp. 2.1–2.10.

References

Bachellerie JP, Cavaille J and Huttenhofer A (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Bittker J, Phillips K and Liu D (2002) Recent advances in the *in vitro* evolution of nucleic acids. *Current Opinion in Chemical Biology*, **6**, 367–374.

Brion P, Michel F, Schroeder R and Westhof E (1999) Analysis of the cooperative thermal unfolding of the td intron of bacteriophage T4. *Nucleic Acids Research*, **27**, 2494–2502.

Burgstaller P, Hermann T, Huber C, Westhof E and Famulok M (1997) Isoalloxazine derivatives promote photocleavage of natural RNAs at G·U base pairs embedded within helices. *Nucleic Acids Research*, **25**, 4018–4027.

Chen J, Le S and Maizel JV (2000) Prediction of common secondary structures of RNAs: A genetic algorithm approach. *Nucleic Acids Research*, **28**, 991–999.

Chen S and Dill KA (2000) RNA folding energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 646–651.

Cullen BR (2002) RNA interference: Antiviral defense and genetic tool. *Nature Immunology*, **3**, 597–599.

Dale T, Smith R and Serra M (2000) A test of the model to predict unusually stable RNA hairpin loop stability. *RNA*, **6**, 608–615.

Diamond JM, Turner DH and Mathews DH (2001) Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, **40**, 6971–6981.

Dias N and Stein CA (2002) Antisense oligonucleotides: Basic concepts and mechanisms. *Molecular Cancer Therapeutics*, **1**, 347–355.

Ding Y and Lawrence C (2001) Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Research*, **29**, 1034–1046.

Ding Y and Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, **31**, 7280–7301.

Dirks R and Pierce N (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, **24**, 1664–1677.

Doudna J and Cech T (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.

Frank DN, Adamidi C, Ehringer MA, Pitulle C and Pace NR (2000) Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. *RNA*, **6**, 1895–1904.

Gorodkin J, Heyer LJ and Stormo GD (1997) Finding the most significant common sequence and structure in a set of RNA sequences. *Nucleic Acids Research*, **25**, 3724–3732.

Gultyaev AP, van Batenburg FHD and Pleij CWA (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, **250**, 37–51.

Gutell RR, Lee JC and Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, **12**, 301–310.

Hansen JL, Schmeing TM, Moore PB and Steitz TA (2002) Structural insights into peptide bond formation. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11670–11675.

Hofacker IL (2003) Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**, 3429–3431.

Hofacker IL, Fekete M and Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, **319**, 1059–1066.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M and Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–168.

Jaeger JA, Turner DH and Zuker M (1989) Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 7706–7710.

Knudsen B and Hein JJ (1999) Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics*, **15**, 446–454.

Lagos-Quintana M, Rauhut R, Lendeckel W and Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lau NC, Lim LP, Weinstein EG and Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.

Long MB, Jones JP, Sullenger BA and Byun J (2003) Ribozyme-mediated revision of RNA and DNA. *The Journal of Clinical Investigation*, **112**, 312–486.

Lück R, Gräf S and Steger G (1999) ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Research*, **27**, 4208–4217.

Lyngsø R and Pederson C (2000) RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, **7**, 409–427.

Lyngsø R, Zuker M and Pederson C (1999) Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, **15**, 440–445.

Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M and Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7287–7292.

Mathews DH, Sabina J, Zuker M and Turner DH (1999) Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA Secondary Structure. *Journal of Molecular Biology*, **288**, 911–940.

Mathews DH and Turner DH (2002a) Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, **317**, 191–203.

Mathews DH and Turner DH (2002b) Experimentally derived nearest neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869–880.

McCaskill JS (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Nussinov R and Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 6309–6313.

Panning B and Jaenisch R (1998) RNA and the epigenetic regulation of X chromosome inactivation. *Cell*, **93**, 305–308.

Rivas E and Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, **285**, 2053–2068.

Sankoff D (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, **45**, 810–825.

Sazani P and Kole R (2003) Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *The Journal of Clinical Investigation*, **112**, 481–486.

Schroeder SJ, Burkard ME and Turner DH (1999) The energetics of small internal loops in RNA. *Biopolymers*, **52**, 157–167.

Schroeder SJ and Turner DH (2000) Factors affecting the thermodynamic stability of small asymmetric internal loops in RNA. *Biochemistry*, **39**, 9257–9274.

Schroeder SJ and Turner DH (2001) Thermodynamic stabilities of internal loops with G·U closing pairs in RNA. *Biochemistry*, **40**, 11509–11517.

Schultes EA and Bartel DP (2000) One sequence, two ribozymes: Implications for emergence of new ribozyme folds. *Science*, **289**, 448–452.

Szymanski M, Barciszewska MZ, Barciszewski J and Erdmann VA (2000) 5S ribosomal RNA database Y2K. *Nucleic Acids Research*, **28**, 166–167.

Walter P and Blobel G (1982) Signal recognition particle contains a 7 S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**, 691–698.

Wellbacher T, Suzuki K, Dubey AK, Wang X, Gudapaty S, Morozov I, Baker CS, Georgellis D, Babitzke P and Romeo T (2003) A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Molecular Microbiology*, **48**, 657–670.

Wuchty S, Fontana W, Hofacker IL and Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.

Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C and Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, **37**, 14719–14735.

Znosko BM, Silvestri SB, Volkman H, Boswell B and Serra MJ (2002) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*, **41**, 10406–10417.

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, 3406–3415.

Zuker M, Mathews DH and Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *RNA Biochemistry and Biotechnology*, Barciszewski J and Clark BFC (Eds.), Kluwer Academic Publishers: Boston, pp. 11–43.

Zuker M and Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, **9**, 133–148.

Specialist Review

Ensembl and UniProt (Swiss-Prot)

David Jackson
Xapien GmbH, Heidelberg, Germany

Reinhard Schneider
European Molecular Biology Laboratory, Heidelberg, Germany

1. Ensembl

The sequencing of the human genome has been heralded as a technological milestone of similar magnitude to the first moon landing. In reality, it has taken numerous giant leaps on the part of bioinformatics to make this a truly valuable source of information. Foremost among these efforts is the Ensembl database project. Capitalizing on the natural knowledge scaffold afforded by an organism's genome, this European-based project (counterpart of the American Locuslink project) focuses on the "geno-centric" integration and analysis of genomic data. By providing researchers with an elegant bioinformatics framework upon which to visualize, analyze, and administer genomic information, Ensembl has established itself as a key player in genomic-based discovery.

1.1. Access and availability

Ensembl is primarily accessed as an interactive web service, available at **http://www.ensembl.org** (Birney *et al*., 2004a). The system is installed globally in both industry and academia, on numerous platforms. Such wide acceptance has been facilitated by its development as an open-source portable software system that can be freely downloaded in its entirety or as individual components, including the website and analysis pipeline. The production code is available via FTP while live development code is available via a CVS server. The system is based on a MySQL relational database (*see* Article 109, **Relational databases in bioinformatics**, Volume 8) and a series of reusable bioperl modules. While primarily coded in Perl (*see* Article 104, **Perl in bioinformatics**, Volume 8 and Article 112, **A brief Perl tutorial for bioinformatics**, Volume 8), some extensions exist in C, as do alternative graphical user interfaces coded in Java.

Upon accessing the Ensembl website, biologists are greeted with the Ensembl genome browser. This page provides introductory information, a download section,

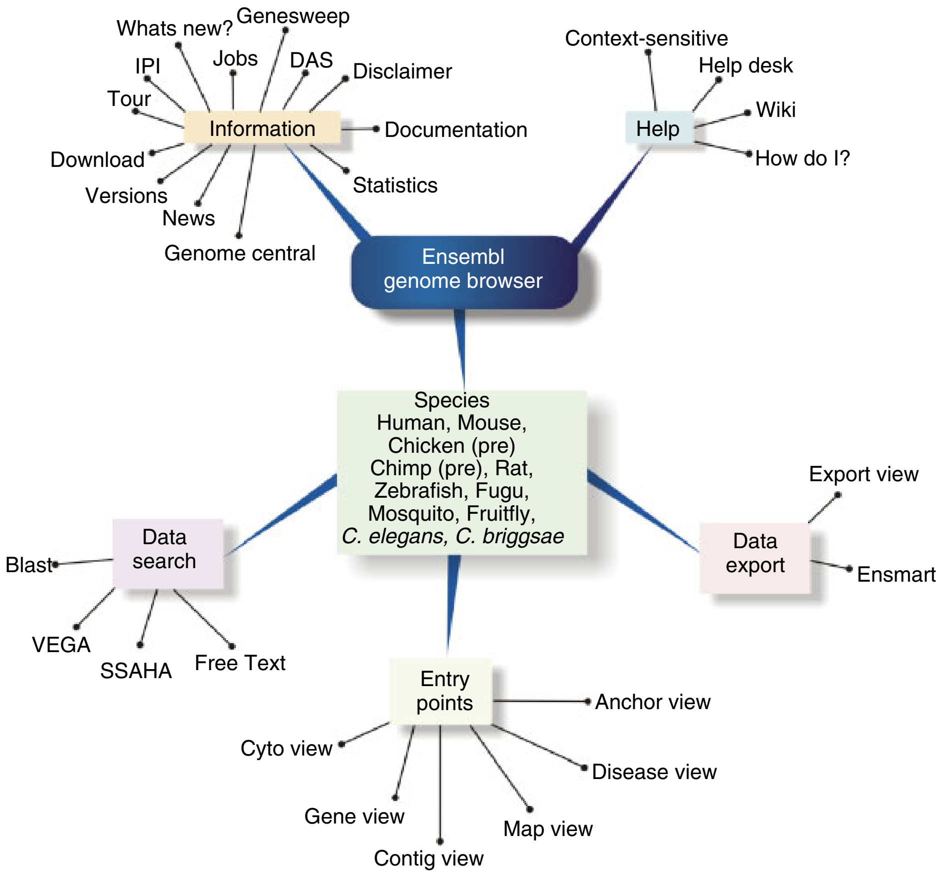


Figure 1 Map of the Ensembl website, available at **http://www.ensembl.org**

and most importantly, access to the individual species sections (see Figure 1 for a more detailed map). The latter acts as the most direct point of entry to a given organism's genomic information. At the time of writing (Ensembl v26), some 12 genomes are readily available for analysis, though the Cow and Dog genomes exist in a so-called Pre-Ensembl state. As full annotation of a genome takes some weeks to complete, Pre-Ensembl was conceived to provide users with immediate access to newly released genome assemblies. Pre-Ensembl is also available via a separate website (**http://pre.ensembl.org/**) with limited functionality.

1.2. Genome annotation

The annotation of both known and novel predicted genes in genomic DNA is the primary forte of the Ensembl system. The gene build method works on the premise that the incorporation of all available evidence is necessary for gene prediction. As such, the process uses a wide range of methods, including ab initio gene predictions, homology, and HMMs (hidden Markov models) (*see* Article 98, **Hidden**

Markov models and neural networks, Volume 8) and proceeds in a series of three steps. First, the "best in genome" positions for all known human proteins from SPTREMBL (Bairoch and Apweiler, 2000) are identified using *pmatch*, a fast but empirical protein to DNA matcher developed by Richard Durbin (unpublished). Subsequent refinement of the gene model using genewise (Birney *et al*., 2004b) is complimented by alignment of the UTRs from full-length cDNAs. In the second step, paralogous and orthologous proteins are aligned to the genome to develop a set of novel human genes. This is complimented in the final step by the HMM-based gene prediction program genscan (Burge and Karlin, 1997) that is used to generate a set of peptides, which if confirmed by homology to proteins, mRNA, and UniGene clusters, are assembled into genes. The resulting human genes have stable identifiers beginning with ENSG, while transcripts begin ENST, exons begin with ENSE, and translations begin ENSP.

Accessing information about a particular organism's genome is as simple as selecting that organism's name from the entry page of the Ensembl genome browser. A variety of alternate views for each organism are available. Perhaps the simplest is the *mapview*, which provides a cytogenetic map and associated markers for a particular chromosome, and also displays feature distribution plots such as known genes v's total genes, SNP's, and %GC content. Selection of a particular band leads one into the more detailed *contigview*. This view is perhaps the most useful, allowing one to scroll along entire chromosomes, while providing detailed views of the features within a selected region. These features are derived from external data sources such as genetic markers, OMIM genes, HUGO gene names with links to the source databases. The user can also control which features are displayed and dynamically integrate external annotation sources (e.g., DAS; see below) including their own. A so-called basepair level panel has also been added to the *contigview*, showing nucleotide, six-frame amino acid translation, and restriction enzyme site features. SNP information is also available, as is the labeling of syntenic blocks.

Other more specific views include *geneview*, which displays information about individual genes with transcripts and gene models and *proteinview*, which shows information about individual gene translations and includes functional annotation from InterPro. At this level, the ability to search for similar sequences is a useful feature, and Ensembl developers have thus provided users with the ability to search against the entire human genome sequence, predicted gene datasets, trace data, and whole genome assembly datasets, using the BLAST (*see* Article 39, **IMPALA/RPS-BLAST/PSI-BLAST in protein sequence analysis**, Volume 7) and SSAHA algorithms. SSAHA is a software tool for very fast matching and alignment of DNA sequences.

Ensembl also caters for comparative genome analysis. Three types of comparative analyses are routinely performed together with gene identification for each new assembly. These include fine-grained DNA-DNA alignments (performed using the exonerate algorithm followed by BLASTN), orthologous protein information (using BLASTP), and large-scale synteny analysis. Synteny information is made available via the *syntenyview* interface. An alternative JAVA-based genomic viewer called *Apollo* (Lewis *et al*., 2002), developed jointly by Ensembl and the Berkeley *Drosophila* genome project, can be used to view both synteny and contig information alike.

1.2.1. DAS

Given that genome annotation is a task far too demanding for any single initiative, Ensembl developers have embraced the so-called distributed annotation system (DAS) to help bridge the gap (Dowell *et al*., 2001, **http://www.biodas.org**). DAS is a client-server system in which a single client integrates information from multiple servers, enabling a single computer to gather, collate, and display annotations in a single view. The Ensembl *contigview* can be configured to act as a DAS client for users who want to view annotation from human genome DAS servers without setting up third-party clients. Ensembl also makes its annotation data available (**http://servlet.sanger.ac.uk:8080/das/**) using the biojava DAS server DAZZLE (**http://www.biojava.org/**) for users with third-party DAS clients. The Apollo genome browser has also been extended to display data from DAS servers.

1.3. Other noteworthy features

1.3.1. EnsMart

EnsMart is Ensembl's data-mining toolset, developed as a generic data warehousing solution for efficient querying of large biological datasets and integration with third-party data and tools. EnsMart contains a "query builder" interface that allows users to specify genomic regions and then refine results using a variety of parameters including expression, disease association, and allele frequency data. Results can then be retrieved in a variety of output formats including HTML, Excel, and plain text.

1.3.2. Vega

The Vega website (**http://vega.sanger.ac.uk/**) provides access to a single database of curated annotation of vertebrate genomes collected from a number of annotation groups. Vega strides to provide consistent high-quality curation of finished vertebrate genomic sequence.

2. UniProt (Swiss-Prot, TrEMBL, PIR-PSD)

2.1. History

The Swiss-Prot Protein Knowledgebase is an annotated protein sequence database established in 1986 by Amos Bairoch at the Department of Medical Biochemistry of the University of Geneva in Switzerland. Since 1987, this has been a collaborative effort with the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Today, Bairoch's team is affiliated with the Swiss Institute of Bioinformatics (SIB) and the EMBL activities are carried out at the European Bioinformatics Institute (EBI) in Rolf Apweiler's group.

The goal of Swiss-Prot was always that each entry should be thoroughly analyzed and annotated by biologists to ensure a high standard of annotation and to maintain

the quality of the database. Because of the vast amount of sequences coming in from genome sequencing projects, the time-consuming task of manually curating sequences became a real challenge in the mid to the end of the 90s. As a solution to this problem, the EBI group created a complement to the Swiss-Prot database: the TrEMBL (Translation of EMBL nucleotide sequence database) protein sequence database. TrEMBL initially consisted of computer-annotated entries derived from the translation of all coding sequences (CDS) in the DDBJ/EMBL-Bank/GenBank nucleotide sequence database, except for those already included in Swiss-Prot. Because of the automatic procedure, it was possible to deliver sequences available as quickly as possible and to maintain the high-quality standard in the Swiss-Prot database.

In 1996, the Swiss-Prot databases went through a financing crisis (**http://www.expasy.org/sprot/crisis96/**) and it became clear that the only feasible solution to this problem was to obtain additional funds through the payment of yearly license fees by nonacademic users. Both SIB and EBI founded a new company, Geneva Bioinformatics (GeneBio), that acts as their representative for the purpose of concluding the necessary license agreements and levying the fees. This licensing setup was in place till January 2005, when the data again became freely available to all entities. This was possible because of the funding for the UniProt consortium by the U.S. National Institutes of Health, totaling US$ 15 million over three years.

In 2002, the SIB, the EBI, and the Protein Information Resource (PIR) group at the Georgetown University Medical Center and the National Biomedical Research Foundation joined forces as a consortium to merge the three coexisting protein databases Swiss-Prot, TrEMBL, and PIR-PSD. The resulting UniProt (Universal Protein Resource) (Apweiler *et al*., 2004) database is now the world's most comprehensive catalog of information on proteins and is the central repository of protein sequence and function.

The consortium's primary mission is to maintain a "high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community" (Apweiler *et al*., 2004).

2.2. *Access and availability*

UniProt is primarily accessed as an interactive web service, available at **http://www.uniprot.org**. The server is mirrored at the EBI, the Swiss Institute of Bioinformatics, and the Georgetown University and allows searching of the databases using a simple text query as well as using a protein sequence as a query. Additional tools such as multiple sequence alignment algorithms, batch retrieval features, and other useful links are provided. The databases can be downloaded via the comprehensive download center as flat files, FASTA-formatted files or XML files.

2.3. *Database structure*

The UniProt databases are designed as a three-database layer, each optimized for different uses as follows.

- The UniProt Archive (UniParc) provides the stable nonredundant sequence collection by storing the complete information of publicly available protein sequence data. It collects sequences from many different sources, including Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, PDB, and others. UniParc stores each unique sequence only once and assigns a unique UniParc identifier and provides cross-references to the source databases. One important feature is the sequence versioning, which is incremented each time the underlying sequence changes. In that way, one can keep track of sequence changes in all source databases.
- The UniProt Knowledgebase (UniProt) provides the central database of protein sequences. It is a curated, added value protein sequence database that provides a high level of annotation, a minimal level of redundancy, and high level of integration with other databases. The UniProt Knowledgebase consists of two sections: Swiss-Prot – a section containing manually annotated records with information extracted from literature and curator-evaluated computational analysis, and TrEMBL – a section with computationally analyzed records that await full manual annotation.
- The UniProt NREF databases (UniRef) provide nonredundant data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions. It combines closely related sequences into a single record to speed searches.

3. Swiss-Prot

The Swiss-Prot database always distinguished itself from other protein databases by four distinct criteria as follows.

3.1. Annotation

Typically, in sequence databases people distinguish between the core data and the annotation of the sequence. The core data consists of the sequence data itself, the citation information, and the taxonomic data. The annotation of a sequence typically describes many different aspects of the protein. Typically, this includes the function description, posttranslational modifications, domain structure, and other secondary and tertiary structure features as well as homology relationships to other proteins and associations to known diseases. This annotation is periodically updated, and external experts have been recruited to send comments and updates.

3.2. Minimal redundancy

In Swiss-Prot, data belonging to one sequence are merged into one entry. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry. In this way, the redundancy is kept to a minimum already at the creation of a new entry.

3.3. Integration with other databases

One of the predominant features in Swiss-Prot is the comprehensive cross-linking to other databases. This includes the basic three different sequence-related databases (DNA, protein, and protein 3D structure) but extends to a broad range of specialist databases. Currently, Swiss-Prot cross-links to more than 50 different other databases (Table 1).

3.4. Documentation

Swiss-Prot is distributed with a large number of documentation files, and numerous additional information files are provided and continuously added. The most important files are the user manual, the release notes, and the forthcoming changes. The documentation provides important information about the use and interpretation of the content, which becomes especially important if one tries to extract information for statistics or data mining purposes. The references Karp *et al*. (2001), Apweiler *et al*. (2001), and Junker *et al*. (1999) are especially useful in this respect.

3.5. Structure of a typical entry

The entries in the UniProt knowledgebase are readable by human readers as well as by computer programs. Each entry is composed of lines, which start with a two-character line code followed by their specific content line(s) (Figure 2).

Table 1 The relationships (cross-references) between Swiss-Prot and some biomolecular databases

Topic	Cross-references to other databases
Sequence	EMBL, PIR
Structure	HSSP, PDB
Domains, families, sites	HAMAP, InterPro, PIRSF, Pfam, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs
Posttranslational modification	GlycoSuiteDB, PhosSite
Organism-specific	DbSNP, DictyBase, EcoGene, EchoBase, FlyBase, GeneDB_Spombe, GeneFarm, Genew, Gramene, HIV, Leproma, ListList, MaizeDB, MGD, MypuList, OMIM, PhotoList, Reactome, RGD, SagaLIst, SGD, StyGene, SubtiList, TubercuList, WormPep, ZFIN
2D-gel-electrophoresis	ANU-2DPAGE, Aarhus/Ghent-2DPAGE, COMPLUYEAST-2DPAGE, ECO2DBASE, HSC-2DPAGE, MAIZE-2DPAGE, OGP, PHCI-2DPAGE, PMMA-2DPAGE, Rat-heart-2DPAGE, Siena-2DPAGE, SWISS-2DPAGE
Miscellaneous	GermOnline, GO, MEROPS, REBASE, TRANSFAC, IntAct

```
ID   CRAM_CRAAB     STANDARD;      PRT;    46 AA.
AC   P01542;
DT   21-JUL-1986 (Rel. 01, Created)
DT   25-OCT-2004 (Rel. 45, Last annotation update)
DE   Crambin.
GN   Name=THI2;
OS   Crambe abyssinica (Abyssinian crambe).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids;
OC   eurosids II; Brassicales; Brassicaceae; Crambe.
OX   NCBI_TaxID=3721;
RN   [1]
RP   SEQUENCE.
RX   MEDLINE=82046542; PubMed=6895315 [NCBI, ExPASy, EBI, Israel, Japan];
RA   Teeter M.M., Mazer J.A., L'Italien J.J.;
RT   "Primary structure of the hydrophobic plant protein crambin.";
RL   Biochemistry 20:5437-5443(1981).
CC   -!- FUNCTION: The function of this hydrophobic plant seed protein is not known.
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- SIMILARITY: Belongs to the plant thionin (TC 1.C.44) family.
DR   PIR; A01805; KECX.
DR   PDB; 1CRN; X-ray; @=1-46. [ExPASy / RCSB]
DR   SWISS-3DIMAGE; CRAM_CRAAB.
DR   InterPro; IPR001010; Thionin.
DR   Pfam; PF00321; Thionin; 1.
DR   PRINTS; PR00287; THIONIN.
DR   PROSITE; PS00271; THIONIN; 1.
DR   ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR   BLOCKS; P01542.
DR   ProtoNet; P01542.
DR   ProtoMap; P01542.
DR   PRESAGE; P01542.
DR   DIP; P01542.
DR   ModBase; P01542.
DR   SMR; P01542.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   3D-structure; Direct protein sequencing; Plant defense; Thionin.
FT   DISULFID      3     40
FT   DISULFID      4     32
FT   DISULFID     16     26
FT   VARIANT      22     22       P -> S (in isoform SI).
FT   VARIANT      25     25       L -> I (in isoform SI).
FT   STRAND        2      3
FT   HELIX         7     17
FT   TURN         18     20
FT   HELIX        23     30
FT   STRAND       33     34
FT   TURN         42     43
SQ   SEQUENCE   46 AA;  4736 MW;  919E68AF159EF722 CRC64;
     TTCCPSIVAR SNFNVCRLPG TPEALCATYT GCIIIPGATC PGDYAN
//
```

Figure 2 A typical entry in the Swiss-Prot database

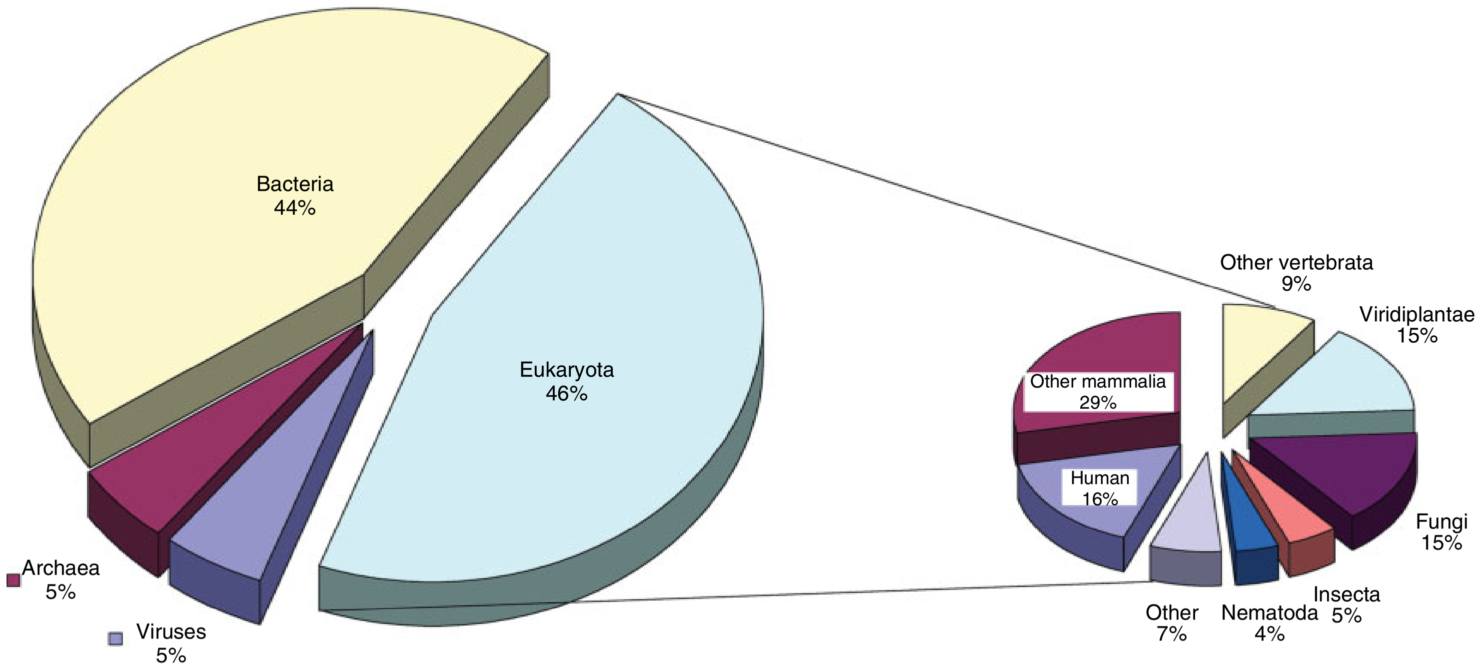


Figure 3 Release statistic for Release 45.2 of Swiss-Prot according to Kingdom and Category

Table 2 Line code and their content and occurrence in a Swiss-Prot entry

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference(s)	Once or more
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RG	Reference group	Once or more (Optional if RA line)
RA	Reference authors	Once or more (Optional if RG line)
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header	Once
(blanks)	Sequence data	Once or more
//	Termination line	Once; ends the entry

Table 2 shows the current line types and line codes and the order in which they appear in an entry.

4. Statistics

Release 45.2 of November 2004 of Swiss-Prot contains 164201 sequence entries, comprising 59974054 amino acids abstracted from 121599 references. Figure 3 shows the distribution of sequences according to the kingdoms and further detailed for the Eukaryota.

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Apweiler R, Kersey P, Junker V and Bairoch A (2001) Technical comment to "Database verification studies of Swiss-Prot and GenBank" by Karp *et al*. *Bioinformatics*, **1**, 533–534.

Bairoch A and Apweiler R (2000) The Swiss-Prot protein sequence database and its supplement TrEMBL. *Nucleic Acids Research*, **28**, 45–48.

Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, *et al*. (2004a) An overview of Ensembl. *Genome Research*, **14**, 925–928.

Birney E, Clamp M and Durbin R (2004b) Genewise and genomewise. *Genome Research*, **14**(5), 988–995.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Dowell RD, Jokerst RM, Day A, Eddy SR and Stein L (2001) The distributed annotation system. *BMC Bioinformatics*, **2**(1), 7.

Junker VL, Apweiler R and Bairoch A (1999) Representation of functional information in the Swiss-Prot Data Bank. *Bioinformatics*, **15**(12), 1066–1067.

Karp PD, Paley S and Zhu J (2001) Database verification studies of Swiss-Prot and Genebank. *Bioinformatics*, **17**(6), 526–532.

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Ricter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biology*, **3**(12). Research0082.1–Research0082.14.

Specialist Review

Hidden Markov models and neural networks

Stefan C. Kremer
University of Guelph, Guelph, ON, Canada

Pierre Baldi
University of California, Irvine, CA, USA

1. Introduction

The development of high-throughput technology (*see* Article 24, **The Human Genome Project**, Volume 3) in the life sciences over the past 15 years has led to an unprecedented expansion in the amount of biological data that has been collected across a broad range of fields and organisms. This explosion in the availability of data has created a need for computer analysis techniques that can be used to keep up with the tasks of labeling, identifying, categorizing, postprocessing, and making predictions about this new corpus of information. Some of the challenges of developing tools to analyze the data are exacerbated by the noise and variability often present in biological data, our incomplete understanding of biological processes and systems, and insufficient computing power that prevents large-scale simulations of atomic and molecular interactions.

One potential approach to this dilemma lies in the application of machine learning (Baldi and Brunak, 2001). Machine learning approaches use example data to automatically extract statistically relevant information that can be used for predictions. The advantages of machine learning approaches include: that it is less necessary to understand the underlying principles, that higher-level approximations can be derived in place of explicit derivations, and that new problems can sometimes be approached without having to build a brand-new system from scratch.

Two popular adaptive systems in the machine learning community that have been applied to bioinformatics are hidden Markov models (HMMs) and artificial neural networks (ANNs). HMMs, originally developed for other applications such as speech recognition, are generative, probabilistic models of sequential information. In an HMM, an observed sequence is modeled as being the stochastic result of an underlying unobserved random walk through the hidden states of the model. The parameters of an HMM are the transition probabilities between the hidden states and the symbol emission probabilities from each hidden state (*see* Article 17, **Pair hidden Markov models**, Volume 7).

ANNs, originally inspired by high-level models of biological neuronal networks, consist of networks of simple processing units, where the output of a typical unit is computed by applying a nonlinear (sigmoidal) function to the weighted average of its inputs. The parameters of an ANN are the "synaptic" weights associated with each connection. Optimization algorithms, such as gradient descent in error space, are used to adapt the parameters of HMMs or ANNs over a set of training examples to exhibit a desired behavior.

Both HMMs and ANNs have been extensively applied to problems in the domain of bioinformatics (Baldi and Brunak, 2001). This review will summarize a few examples of such applications and thus provide the reader with some insights into the types of problems to which these computational methods can be applied, their relative strengths and weaknesses, and an understanding of the issues involved in putting them into practice.

2. Hidden Markov models

HMMs have a long tradition in automated speech processing. Their ability to deal with sequences that can be stretched or compressed, that contain noise, and that vary over time makes them uniquely suited to applications in speech processing. Some of these properties also apply to genomic sequences.

To be precise, a first-order HMM consists of a set of states, {0,1,2,3 ... }, a set of symbols X, an $(L + 2) \times (L + 2)$ matrix of transition probabilities, A, and a matrix of emission probabilities, E. For simplicity, here we label both the start and end state as 0. Each element in the matrix of transition probabilities, a_{kl}, represents the probability that a given state follows a given previous one along a path π:

$$a_{kl} = P(\pi_i = l|\pi_{i-1} = k) \tag{1}$$

The probability of emitting a particular symbol, b, while in a particular state, k, is

$$e_k(b) = P(x_i = b|\pi_i = k) \tag{2}$$

Together, these definitions allow us to compute the probability of a given path occurring and a given set of symbols being emitted:

$$\mathrm{P}(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi i}(x_i) a_{\pi_i \pi_{i+1}} \tag{3}$$

Finally, the probability of a given sequence x is the sum of the probabilities $P(x,\pi)$ over all the paths π that are consistent with that sequence. It is often useful to try to infer the most probable cause of an observed sequence. That is, given a sequence, to compute the most probable state path.

$$\pi^* = \underset{\pi}{\arg\max} \quad P(x, \pi) \tag{4}$$

This can be accomplished by dynamic programming, using the well-known Viterbi algorithm. This algorithm operates by iteratively computing the maximum probability of being in each possible state after each symbol in the sequence of interest. For the first step, there is only one way to get to each state (from the start state), so the maxima are computed over one previous state. In subsequent steps, however, each of the states in the HMM can be a potential previous state, so the maximum probability of being in any given state must be computed over all possible prior states. By recording the prior state that gave the maximum likelihood at each step, it is possible to compute the most probable sequence. This is also called aligning a sequence to a model. A similar dynamic programming recursion can be used to compute the probability $P(x)$ of a sequence, without having to sum over an exponentially large number of possible paths. Similar dynamic programming principles (e.g., the Baum–Welch or EM algorithm) can be applied during training to iteratively modify the matrices A and E to maximize the likelihood of the sequences in the training set. A very simple idea for training, which works quite well in practice, consists in computing the most likely path for each training sequence using the Viterbi algorithm and increasing all the transition and emission probabilities along each optimal path. Details of the algorithms are found in the references.

HMMs are used extensively in bioinformatics applications in tasks ranging from multiple alignments, to protein family modeling and classification (*see* Article 78, **Classification of proteins into families**, Volume 6), to pattern discovery, and to gene finding (Baldi *et al*., 1994; Krogh *et al*., 1994; Durbin *et al*., 1998; Baldi and Brunak, 2001).

3. Artificial neural networks

In ANN models, the output out_i of neuron i is described by

$$out_i = f_i\left(\sum_j w_{ij} out_j\right) \tag{5}$$

where f_i is the transfer function of the neuron and w_{ij} is the synaptic weight of the connection from neuron j to neuron i. It is clear today that the neurons used in ANN models are orders of magnitude simpler than their distant biological cousins. In spite of this apparent simplicity, ANNs have been quite successful as a machine learning approach to pattern recognition, and function approximation (e.g., classification or regression). A typical neural network processes inputs in the form of vectors and computes outputs, also in vector form. Many applications are based on feedforward architectures containing no directed cycles, where the neurons are organized into input, output, and one or more hidden layers. Recurrent or recursive networks are obtained when directed cycles of connections are allowed, for instance, by feedback of output layer onto input layer. One important theoretical property of neural networks is their universal approximation properties – basically,

any reasonable function can be approximated to any degree of precision by a neural network (Hornik *et al*., 1989). While this existence theorem is reassuring, the real problem is to find a reasonable architecture for a given problem in a reasonable time. This is the issue addressed by machine learning methods.

In addition to the ability to compute almost arbitrary output vectors in response to their inputs, ANNs include an automated method for updating the parameter arrays. This is typically done by some optimization algorithm, the most popular being gradient descent, also known as back-propagation and the generalized delta rule (Werbos, 1974; Rumelhart *et al*., 1986). In this approach, an error measure is used to compute the error of the network in response to one or more example patterns (consisting of an input and a target output). Then, the gradient of that error is computed to determine the direction of greatest descent in weight space. By making repeated changes in the weights in the direction of this gradient, a local minimum can be found. While it is possible that this local minimum does not represent a global minimum, often, this local minimum represents an adequate approximation to the function that one wants to learn.

A feedforward ANN generates one output for every input and contains as its only (long-term) memory the parameter matrices that remain static after the training process is completed. For some problems, however, it is desirable to process a sequence of inputs, or produce a sequence of outputs. This is valuable when the input data is distributed in time (i.e., the inputs in the sequence arrive over a period of time) or, for instance, when there is a shift invariance in the relevant input pattern. In this context, shift invariance refers to the notion that an input pattern that occurs within a sequence should have the same effect regardless of its position within the said sequence. Recurrent networks address this problem by including a different kind of memory, short-term memory. This short-term memory is implemented by introducing feedback in the circuits (Kolen and Kremer, 2001). While a feedforward network can be built to simply take a very long sequence as its input, this type of processing cannot produce the kinds of parsimonious solutions that recurrent networks can. More generally, a special kind of recurrent network, called a *recursive ANN architecture*, can be built by combining ANNs with the theory of graphical models (Baldi and Pollastri, 2003).

ANNs have been used extensively in a variety of bioinformatics applications ranging from the detection of signal peptides (Nielsen *et al*., 1997) to the prediction of protein structural features such as secondary structure, relative solvent accessibility, and contact maps (Rost and Sander, 1994; Baldi and Brunak, 2001; Baldi and Pollastri, 2003).

4. Hidden Markov models for identifying sequence families

A standard application of HMMs to bioinformatics is to model protein families (*see* Article 78, **Classification of proteins into families**, Volume 6). The idea is to build an HMM for a given family of sequences that captures the commonalities and differences of the sequence in a probabilistic fashion. Such a model can then be used to produce multiple alignments, to evaluate a new sequence, and determine whether the new sequence belongs to the family represented by the model.

As an example, we consider the case of globins, as proposed in Durbin *et al.* (1998). Globins, of course, are oxygen-binding proteins and include hemoglobins, myoglobins, leghemoglobins, and flavohemoproteins. This commonality of function exhibited across a huge range of organisms (including plants, bacteria, and both vertebrate and invertebrate animals) makes them a particularly interesting and diverse family of molecules.

To model the globin family with an HMM, we can first start from a multiple alignment of a set of known globin sequences. From this alignment, it is possible to construct an HMM with a very specific arrangement of states, as shown in Figure 1.

This HMM has the following properties. S represents the start state of the HMM and has no associated emissions. E represents the end state of the HMM and also has no associated emissions. M_1 represents the first column in the alignment that does not contain mostly gaps. Its emission probabilities are set to match the frequency of the various residues in the column of the alignment. Similarly, M_i represents the ith column in the alignment that does not contain mostly gaps and has emission probabilities corresponding to the ith nongap column's residue frequencies. Thus, the central chain of M_i's represents the most typical residue sequences.

The states labeled I_i represent extra residues inserted into atypical sequences and can be visited (and revisited via the recurrent arrows) in addition to the sequence of M_i's. Each state, I_i, has associated emission probabilities for the "extra" residues that do not occur in the more typical sequences. The states labeled D_j represent deletions of residues from the typical sequences. There are no emissions associated with the D_j states, and since these states are visited instead of M_i states, they effectively delete a residue from the typical sequence. Thus, in short, the M chain represents the match states and it is flanked by a chain of insert and delete states to allow insertions and deletions at each position in the sequence. For this reason, unusually long sequences will tend to pass through more I states, while unusually short sequences will tend to pass through more D states. This provides an elegant solution to the problem that not all sequences have the same length and that even sequences with the same length can be aligned quite differently.

The task of assigning emission and transition probabilities is quite straightforward, based on the frequencies of the residues in the alignment and the frequency of gaps. It should be noted, however, that if raw frequency values were used, new

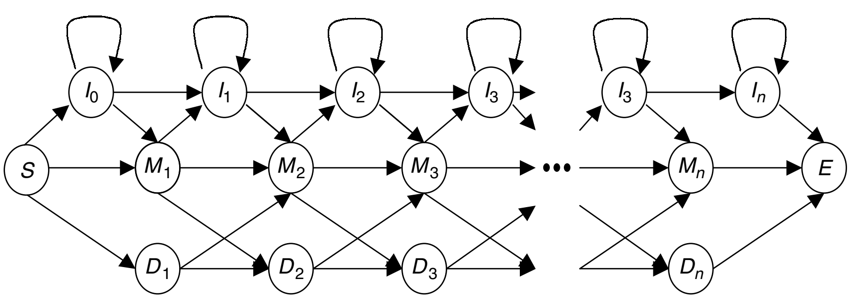


Figure 1 Profile HMM consisting of a start state (S), an end state (E), and sequences of main or match states (M), insert states (I), and delete states (D). (Optionally, there can be D-I and I-D connections, but these are omitted in this diagram for simplicity.)

globins that happened to contain a residue not found in any of the sequences in the training set would automatically be associated with a probability of zero. To avoid this problem, pseudocounts can be added to the frequencies to allow for some tolerance of deviance.

If an initial multiple alignment is not available, the HMM parameters can still be estimated using the dynamic programming learning algorithms discussed above.

A trained HMM can be used in several ways. First, we can align any sequence to the HMM by computing its Viterbi path and also its likelihood. A multiple alignment of globin sequences is immediately derived by aligning their corresponding Viterbi paths.

Second, we can use the likelihood score of a sequence to discriminate between globin and nonglobin sequences and search large databases. Finally, conserved patterns of residues associated, for instance, with structural or functional motifs can be detected from the HMM parameters or from the corresponding multiple alignment.

The use of HMMs for protein classification can be further enhanced by building libraries of HMM models associated with different protein families. There are now fairly comprehensive libraries of preconstructed HMMs for many protein families, available for download on a few websites. The most widely used is the Pfam database (Bateman *et al*., 2004; *see also* Article 86, **Pfam: the protein families database**, Volume 6). As of January 2004, this database contained more than 7200 families and covered an estimated 75% of all protein sequences. Rfam is a similar database for RNA (Griffiths-Jones *et al*., 2003).

5. Protein prediction with recursive networks

Many ANN applications use a simple feedforward network with a fixed window size input. These architectures are suitable, for instance, for pattern recognition problems in which the scale of the patterns to be detected falls within the size of the window. Many problems, however, are characterized by patterns occurring at multiple length scales that do not fall naturally inside a single window size. In this sense, HMMs are more elastic than simple feedforward ANNs because they can accommodate input sequences of variable length. However, this is not an intrinsic limitation of ANNs since recursive architectures can be built that can accommodate inputs of variable structure, size, and dimensions. Here, we give an example of a recursive network for the prediction of 2D protein contact maps taken from Baldi and Pollastri (2003).

Given a protein sequence, the goal is to produce a topological representation of its structure in the form of a contact map. The contact map is a two-dimensional, symmetric, binary matrix, representing the 3D proximity of objects (atoms, amino acids, secondary structure elements) associated with a sequence. Typical thresholds used to assess proximity at the amino acid level are in the 6–12-Å range.

A recursive architecture to tackle this task consists of five feedforward ANNs. A feedforward network N_O that computes the probability of contact $O_{i,j}$ between amino acids i and j as a function of the input vector $I_{i,j}$, and four, hidden, contextual vectors, $NW_{i,j}$, $NE_{i,j}$, $SW_{i,j}$, $SE_{i,j}$ associated with each one of the cardinal corners (Figure 2), and four feedforward networks associated with each one of the four

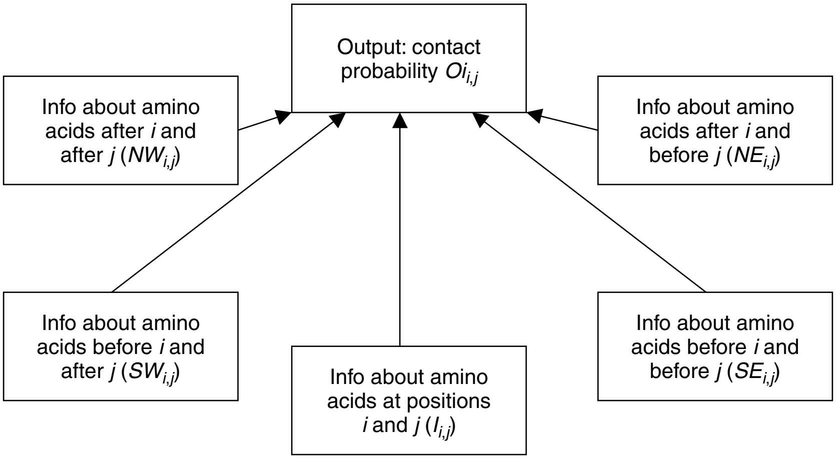


Figure 2 Organization of contact map output prediction

cardinal directions to compute each of the four hidden context vectors. The N_{NW} network, for instance, computes the $NW_{i,j}$ vector as a function of the input vector, $I_{i,j}$, and the neighboring vectors, $NW_{i+1,j}$ and $NW_{i,j-1}$, in the NW lattice. Thus,

$$\begin{aligned} O_{i,j} &= \mathrm{N}_O(I_{i,j}, NW_{i,j}, NE_{i,j}, SW_{i,j}, SE_{i,j}) \\ NW_{i,j} &= \mathrm{N}_{NW}(I_{i,j}, NW_{i+1,j}, NW_{i,j-1}) \end{aligned} \tag{6}$$

and *mutatis mutandis* for the other cardinal directions. The key assumption is that of stationarity, also called weight sharing in the ANN literature, whereby the N_O network is the same across all *i,j* positions, and similarly for the networks that compute lateral contextual information. In the simplest formulation of this approach, the input vector $I_{i,j}$ encodes the identity of the two amino acids at positions i and j in the primary sequence using unary notation, that is, an amino acid is encoded by a 20-dimensional sparse vector containing 19 zeros and a single one. In practice, significantly more complex input vectors are used that contain information about a window of amino acids around i and j, plus information about homologous sequence (profiles) as well as correlated mutations and structural features, such as secondary structure and relative solvent accessibility.

This system takes an amino acid sequence of an arbitrary length, N, as its input and produces an $N \times N$ matrix of contact/noncontact probability values as outputs. The stationarity approach results in a recursive ANN architecture, whereby contact inference is made in the same way at each 2D location. In principle, the decision at position i can be affected by information coming from any other position j through the lateral propagation in the cardinal planes. Note that each one of the five feedforward ANNs can itself comprise its own hidden layers and variable numbers of artificial neurons and connections. The weight-sharing assumption, however, allows us to keep the total number of parameters of the model under reasonable control. Typical models used in contact map applications have several thousand parameters.

The parameters of these recursive architectures can be adjusted using a gradient descent approach, which generalizes the back-propagation algorithm for simple feedforward network. Using this algorithm, a dataset of proteins from the Protein Data Bank (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7) with known structures can be used to adapt the parameters of the networks. Trained networks can be used to predict contact maps for novel proteins and, in turn, contact maps can be used to predict 3D structures. While training can take several days, once trained, the recursive ANNs can predict contact maps very rapidly on a genomics scale. The resulting contact map predictor is a component of the SCRATCH protein structure prediction Web server, which is publicly available through **http://www.igb.uci.edu/servers/psss.html**

6. Conclusion

In this contribution, we have surveyed the application of ANNs and HMMs to problems in bioinformatics. We have described two protein applications in some detail: HMMs to model families and recursive ANNs to predict contact maps. We have obviously only scratched the surface of what defines these techniques, their strengths and weaknesses, how they are applied, and the results that can be obtained. Many more important details and statistical considerations (e.g., ensembles, cross-validation) can be found in the references.

It is important to stress that successful machine learning applications in bioinformatics must go hand in hand with a good understanding of the underlying biological principles and should not start from a *tabula rasa*. This is true not only for the design of the architecture of the learning system and the incorporation of any prior knowledge but also for the selection and processing of the training data. In most cases, training data must be preprocessed to clean up errors and noise and remove important biases. In the Protein Data Bank, for instance, HIV protease structures are highly overrepresented because of the emphasis on AIDS in medical research and funding. Such biases must be removed from any training set aiming at sampling the universe of protein structures as uniformly as possible. This is done, for instance, by aligning all sequences to each other and removing those that are redundant. It is also critical to match the number of free parameters in the models to the problem complexity and the data set size. Too many free parameters can lead to overfitting. Luckily, this is typically not the major problem in bioinformatics applications, given the overabundance of data.

As the volume of biological data collected and accessible on-line continues its exponential growth, machine learning will play an ever-greater role in the extraction and mining of biological knowledge.

Acknowledgments

Dr. Kremer is supported by the NSERC, the CFI, the OIT, and ORDCF. Dr. Baldi is supported by a Laurel Wilkening Faculty Award, a Sun Microsystems Award, and grants from the NIH and NSF.

Further reading

Haykin S (1999) *Neural Networks: A Comprehensive Foundation*, Second Edition, Prentice Hall: Upper Saddle River, NJ.

References

Baldi P and Brunak S (2001) *Bioinformatics: the Machine Learning Approach*, Second Edition, MIT Press: Cambridge, MA.

Baldi P, Chauvin Y, Hunkapiller T and McClure MA (1994) Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, **91**(3), 1059–1063.

Baldi P and Pollastri G (2003) The principled design of large-scale recursive neural network architectures – dag-rnns and the protein structure prediction problem. *Journal of Machine Learning Research*, **4**, 575–602.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, *et al*. (2004) The Pfam protein family database. *Nucleic Acids Research*, Database Issue **32**, D138–D141.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press: Cambridge.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A and Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Research*, **31**(1), 439–441.

Hornik K, Stinchcombe M and White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Kolen J and Kremer SC (2001) *A Field Guide to Dynamical Recurrent Networks*, Wiley/IEEE Press: New York.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, **235**, 1501–1531.

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, **10**, 1–6.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Rumelhart D, Hinton G and Williams R (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing*, Vol. 1, Rumelhart D, McClelland J and the PDP Research Group (Eds), MIT Press: Cambridge, MA, 318–362.

Werbos PJ (1974) *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, PhD thesis, Department of Applied Mathematics, Harvard University.

Specialist Review

Threading algorithms

Jadwiga Bienkowska

Serono Reproductive Biology Institute, Rockland, MA, USA

Boston University, Boston, MA, USA

Rick Lathrop

University of California, Irvine, CA, USA

1. Background

The goal of protein structure prediction by threading is to align a protein sequence correctly to a structural model. This requires choosing both the correct structural model from a library of models and the correct alignment from the space of possible sequence-structure alignments. Once chosen, the alignment establishes a correspondence between amino acids in the sequence and spatial positions in the model. Assigning each aligned amino acid to its corresponding spatial position places the sequence into the three-dimensional (3D) protein fold represented by the model. Typically, the model represents only the spatially conserved positions of the fold, often the protein core, so producing a full-atom protein model would require further steps of loop placement and side-chain packing. Protein threading has a role in protein structure prediction that is intermediate between homology modeling (*see* Article 70, **Modeling by homology**, Volume 7) and ab initio prediction (*see* Article 66, ***Ab initio*** **structure prediction**, Volume 7). Like homology modeling, it uses known protein structures as templates for sequences of unknown structure. Like ab initio prediction, it seeks to optimize a potential function (an objective or score function) measuring goodness of fit of the sequence in a particular spatial configuration. Threading is the protein structure prediction method of choice when (1) the sequence has little or no primary sequence similarity to any sequence with a known structure and (2) some model from the structure library represents the true fold of the sequence.

Protein threading requires (1) a representation of the sequence, (2) a library of structural models, (3) an objective function that scores sequence-structure alignments, (4) a method of aligning the sequence to the model, and (5) a method of selecting a model from the library. Following the initial conception of the threading approach to protein structure prediction (Bowie *et al*., 1991; Jones *et al*., 1992),

there have been very many different approaches to these problems, of which this chapter can present only a few general themes.

2. Representation of the query sequence

It is widely accepted that significantly similar protein sequences also adopt a similar 3D structure. The Paracelsus Challenge demonstrated the design of a protein sequence with 50% sequence identity to a known protein but a different 3D structure (Jones *et al*., 1996), but when natural evolution produces similar protein sequences their protein structures generally are similar as well. Thus, in naturally occurring proteins, sequences that are similar to the query sequence carry useful information about its 3D structure. A multiple-sequence alignment centered on the query sequence reflects sequence variability within the protein family to which the query sequence belongs. Most modern threading algorithms exploit this fact (Jones, 1999; Fischer, 2000; Kelley *et al*., 2000; Panchenko *et al*., 2000; Rychlewski *et al*., 2000; Karplus and Hu, 2001; Skolnick *et al*., 2003).

The query sequence is often represented by a sequence profile, **P**, where the element $P_j = P(A|j)$ is a vector giving a probability distribution over the 20 amino acids at sequence position j. In this notation, a single query sequence has a profile with 1 for the original amino acids and 0 otherwise. The sequence profile is typically constructed from the search of nonredundant databases of proteins (e.g., at NCBI) and sequences are aligned using multiple-sequence alignment programs such as CLUSTAL (Higgins *et al*., 1996) or PSI-BLAST (Altschul *et al*., 1997). Some threading methods also include an independent prediction of the secondary structure (SS) (*see* Article 76, **Secondary structure prediction**, Volume 7) or other derived information as part of the sequence representation. In such cases, the query is represented as two independent vectors $P_j = \{P(A|j), P(SS|j)\}$, where SS might be helix, strand, or coil, a more detailed set of secondary structure assignments, or other information.

3. Representation of protein structure models

What is a model of protein structure? Protein structure is fully determined by the 3D coordinates of all non-hydrogen atoms. For threading, the 3D coordinates are reduced to more abstract representations of protein structure. Typically, structural core elements are defined by the secondary structure elements, α-helices and β-strands, usually with side chains removed. Among proteins with similar structures, large variations occur in the loop regions connecting the structural elements. In consequence, loop lengths, loop conformations, and loop residue interactions are rarely conserved, and often the loop residues are not represented explicitly in the structural models.

The main distinction among threading approaches is the choice of the structure model representation. Threading algorithms fall into two main categories that depend on the protein structure representation they use:

1. In the first category, a protein structure is represented as a linear model.
2. In the second category, a protein structure is represented as a higher-order model.

In a linear representation, the protein structure is modeled as a chain of residue positions that do not interact. In a second-order representation, the model also includes interacting pairs of residue positions, for example, to account for hydrophobic packing, salt bridges, or hydrogen bonding. Still higher order models have been considered to represent triples and higher multiples of interacting residue positions, but are less common.

Approaches that represent protein structure as a linear model consider each structural position in the model independently, neglecting spatial interactions between amino acids in the sequence. This allows very fast alignment algorithms, but loses whatever structural information may be present in amino acid interactions. Approaches that use higher-order models explicitly consider spatial interactions between amino acids that are distant in the sequence but brought into close proximity in the model. This potentially allows for more realistic and informative structural models, but results in an NP-complete alignment problem (Lathrop, 1994). It is known that the information content in higher-order amino acid interactions is modest, but nonzero (Cline *et al*., 2002). What effect this has in practice, and whether the increased information content compensates for the increased complexity, is a subject of some debate within the protein threading community.

3.1. 1D models of the protein structure

A 1D model of a protein structure is a sequence of states representing the residue as if embedded in a 3D structural environment. There are two distinct types of features frequently used to characterize a state, structural features, and amino acid sequence features. The structural features include the solvent exposure of a given residue, the secondary structure of the residue, and so on. The structural features may be representations of a single specific structure or (weighted) averages of structural features from multiple structures in the same family (*see* Article 75, **Protein structure comparison**, Volume 7). The sequence features may include the original amino acids observed in the structure or a sequence profile representing the multiple alignment of sequences from the protein family of the structure's native sequence.

If we denote by s a residue position in the structure (or a position from the alignment of multiple structures), then a vector of features $\mathbf{F}(s)$ describes each position. Thus, a structure model is an ordered chain of feature vectors $\{\mathbf{F}(s)\}$. The dimensionality of the feature vector depends on the specific threading approach.

The original 1D threading papers represented the feature vector as solvent exposure states, where the solvent exposure was calculated from the exposure of amino acids present in the native structure. Since then it has been recognized that, due to variations in the amino acid's size, one must use a measure of exposure that is independent of the native amino acid size. Most recent threading methods use the polyalanine representation of a structure. Solvent exposure state is determined by the solvent exposure of an alanine placed at each residue position. Some approaches vary the radii of the solvent molecule and the β-carbon.

3.2. 2D models of the protein structure

Two-dimensional models attempt to capture the contribution of interactions between pairs of residues. They begin with a 1D representation of a protein structure, and

then overlay representations of pairs of residues that are neighbors in the folded structure. In many threading methods, the pairs are represented as a contact map, where the contact can be defined by any of several methods:

Dependent on the native amino acid side-chain orientation:

1. Residues are in physical contact in the native structure, for example, if the distance between any of their atoms is smaller than a given cutoff, say 5 Å.
2. The distance between the centroids or C_β atoms of the residue side chains is below a certain cutoff.
3. The neighbors are determined by additional geometric constraints imposed by the 3D structure, for example, the C_β atoms may have to be in line-of-sight of each other. This excludes from the neighbor set pairs that can never interact, like residues on the opposite sides of an α-helix.

Independent of the native amino acid side-chain orientation:

1. Any pair separated by a given number of residues, for example, neighbors every 1, 3, or 4 residues in an α-helix, or every 2 residues in a β-sheet.
2. Any pair that has C_α closer than a cutoff value, say 7–10 Å.

Similar to the 1D representation, a pair of residue positions s and r is represented by a feature vector $\mathbf{FF}(s,r)$. The pair associated features fall into three categories:

1. 3D distance–derived features; distance between the β-carbons, distances among all other backbone atoms, distances between the centroids of side-chain positions, and so on.
2. 1D residue separation along the amino acid sequence of the native protein.
3. Structural environments of each residue in the pair like solvent exposure or secondary structure.

Definitions of the various environmental variables differ dramatically among threading approaches. The most commonly used feature for the 2D environments is the 3D distance. Typically, the distance between two atoms is partitioned into bins that are defined by a lower and upper distance threshold. In general, a similar approach can be applied to any feature that is associated with a real or integer variable. Most feature variables require binning, such as distance, solvent exposure, and 1D sequence separation.

3.3. Higher-order structural models

Third-order and higher models attempt to capture regularities of protein structure that cannot be represented by considering amino acid pairs only. For example, adjacent pairs of cysteines may form disulfide bonds, but only one disulphide bond can form among three adjacent cysteines; Godzik *et al*. (1992) used amino acid triples to represent this and related properties. The hydrophobic contact potential of Huang *et al*. (1996) is equivalent to amino acid triples, in this case used to represent the hydrophobic core. A fourth-order representation is the Delaunay tessellation, based on the vertices of irregular tetrahedral lattice (Singh *et al*., 1996; Munson and Singh, 1997; Zheng *et al*., 1997). Higher-order models suffer from the statistician's

"curse of dimensionality"; an Nth-order model must represent 20^N N-tuples. It can be difficult to parameterize the model and the objective function (below) unless reduced amino acid alphabets are used.

4. Objective function (potential or score function)

Most threading approaches do not use the physical full-atom free energy functions commonly used by macromolecular modeling software (*see* Article 74, **Molecular simulations in structure prediction**, Volume 7). Instead, most threading objective functions are determined empirically by statistical analysis of the 3D data deposited in the Protein Data Bank (PDB) (*see* Article 71, **The Protein Data Bank (PDB) and the Worldwide PDB http://www.wwpdb.org**, Volume 7). Thus, they are often referred to as *empirical potentials* or *knowledge-based potentials*. In the case of nonlinear structural models, another common name is *contact potentials*, reflecting their origin in analysis of contacts between atoms or residues in crystal structures. Many approaches augment empirical potentials with other terms thought to be important, for example, contributions from loop regions if the structural model contains only the protein core.

A great many different approaches have been explored. As examples, the hydrophobic contact potential of Huang *et al*. (1996) reflects packing in the hydrophobic core using only two residue classes, hydrophobic and polar, and is remarkable for its explanatory power given its simplicity and near absence of adjustable parameters. Maiorov and Crippen (1994) used linear programming to enforce a constraint that the native threading scores lower than others, but such approaches tend to be brittle. Bryant and Lawrence (1993) used logistic regression, based on multidimensional statistics. Boltzmann statistics is the foundation of many threading methods (Sippl, 1995). White *et al*. (1994) derived a formal probability model based on Markov Random Fields. Many other approaches have been investigated.

The most popular approach involves a negative log odds ratio between the observed and expected amino acid frequencies in a given structural environment. This yields a measure that is analogous to a physical free energy, and gives good results. Given 1D or 2D structural features F or FF defined by a specific structural library, the objective function is determined by counting amino acids with specific features in known 3D structures. The score for observing amino acid A with feature F is determined by $P(A,F)$, the probability of observing the amino acid A with the feature F in the protein structure database. Different methods apply different normalizations to the probability $P(A,F)$. The general motivation is to remove variations that do not contribute to specific sequence-structure recognition, for example, to control for the fact that some amino acids are more common than others. The score (*see* Article 67, **Score functions for structure prediction**, Volume 7) for amino acid A when found in feature F is

$$S(A, F) = -\log \frac{P(A, F)}{N(A, F)} \tag{1}$$

Here $N(A,F)$ is a normalization constant for A and F, typically derived from some assumed reference state or from assumptions of conditional independence.

Various choices of $N(A,F)$ have been explored, one of the simplest being $N(A, F) = P(A)P(F)$.

The same logic applies to pairs of residues, or pairs of atoms associated with residue positions. Some methods use only the amino acids while others use their respective backbone atoms and/or generalized side-chain atoms to define the 2D features of a pair of positions. For two amino acids A and B, where the feature FF is associated with the pair of positions, the score is given by:

$$S(A, B, FF) = -\log \frac{P(A, B, FF)}{N(A, B, FF)} \quad (2)$$

Owing to high redundancy of the PDB, typically a database of nonredundant or representative structures is used for calculation of probabilities. An objective function is often implemented by a number of 20×1 and 20×20 matrices associated with each 1D and 2D structural feature from the feature sets $\{F\}$ and $\{FF\}$ respectively.

Once the objective function is defined and its values estimated from the current database, it is fast and straightforward to calculate the score of a given sequence-structure alignment. The alignment is a placement of amino acids from a sequence $\mathbf{A} = \{A_1, \ldots, A_i, \ldots, A_L\}$ into positions $\mathbf{s} = \{q, \ldots, r, \ldots, s\}$ from the structure model, where the model is a collection of positions and pairs of positions as discussed above. A threading (alignment) is a (possibly partial) map t of sequence indexes i to model indexes, $t(i)\epsilon\mathbf{s}$. Most threading algorithms impose an ordering constraint of mapping increasing sequence indexes to increasing model indexes: if $i < j$, then $t(i) < t(j)$. In principle, relaxing such a constraint would allow threading to recognize/predict a structural topology that is not yet present in a database of known structures, but this is not usual in practice. The score S of the sequence-structure alignment $t(\mathbf{A}) = \{A_1^{t(1)}, \ldots, A_i^{t(i)}, \ldots, A_L^{t(L)}\}$ is given by:

$$\begin{aligned} S(t(\mathbf{A})) = &\sum_{F\in\{F\}} w_F \sum_i S\left(A_i^{t(i)}, F(t(i))\right) \\ &+ \sum_{FF\in\{FF\}} w_{FF} \sum_{\{i,j\},i<j} S\left(A_i^{t(i)}, A_j^{t(j)}, FF(t(i), t(j))\right) \end{aligned} \quad (3)$$

where the sum over different feature categories $\{F\}$, $\{FF\}$ often includes different weights. $F(t(i))$ is the feature of the model position $t(i)$ and $FF(t(i),t(j))$ is the feature associated with a pair of model positions $t(i)$ and $t(j)$. The weights w_{F}, w_{FF}, among different feature terms are subjected to optimization in many threading approaches. One of the 1D features often included in an objective function is a gap opening and extension penalty. The gap opening and gap extension penalties are special features that can depend on the length k of the gap $w(k|F(t(i)))$ and can also depend on the features of the structural model. Many threading approaches do not allow gaps inside the core structural elements, for example, by setting the gap opening penalty to $+\infty$ for positions in the core structural elements. The gap opening and extension penalty terms imply that there are even more weights to optimize.

4.1. Methods of refining the pairwise objective function

The pairwise interaction between residues depends on both the geometric features of positions close in the 3D structure, and on the specific amino acids that are aligned to those positions. There are two methods that attempt to capture these complicated dependencies.

In the Filtered Neighbors Threading approach, the objective function is constructed specifically for each structural model (Bienkowska *et al*., 1999). Each pair of positions has its unique objective function calculated by taking a Hadamard product of two 20 × 20 matrices $S(\cdot,\cdot,FF(s,r))$ and $V(\cdot,\cdot,s,r)$. The $S(\cdot,\cdot,FF(s,r))$ matrix is the usual pairwise objective function as defined by the statistics of amino acid pairs with a given structural feature $FF(s,r)$. The $V(\cdot,\cdot,s,r)$ is a binary matrix of 0's and 1's specific to the pair of structural positions from the template. A "0" for amino acids A and B indicates that physical contact between these amino acids is not plausible when placed at respective positions in the template structure and "1" indicates that a physical contact is plausible. The plausibility of a physical contact is calculated independently of the objective function S. A number of geometrical descriptors of neighboring structural positions describe neighbor pairs. The algorithm first analyzes a set of all 3D neighbor pairs from the database of proteins and partitions the space of geometrical descriptors into regions occupied by most of the observed physical contacts and the rest of the space. Thus, any pair of neighboring positions is defined as a plausible physical contact or not depending where it is in the space of geometrical descriptors.

The approach implemented by PROSPECTOR2 iteratively constructs a pairwise objective function (Skolnick and Kihara, 2001). This algorithm uses an iterative dynamic programming approach with the pairwise objective function redefined at each iteration (see frozen approximation below). During the iteration, a sequence is threaded through the entire library of structural models. Top-scoring structures (within an empirically set score cutoff) are selected and the number of predicted contacts between sequence residues A_i and A_j is calculated as q_{ij}. If q_{ij} is greater than 3, then a "filtering" objective function is given by $V(i,j) = -\ln(q_{ij}/q^0)$, where the normalization factor is $q^0 = \sum_i \sum_j q_{ij}/L^2$. The objective function in the next iteration step is the arithmetic average of the "filtering" objective function and the original objective function $S(A_i, A_j, F(t(i), t(j)))$, if the filtering score is defined for the respective residues.

5. Aligning a sequence to a model

The goal of a threading alignment algorithm is to find an optimal match between the query sequence and a structural model among all possible sequence-structure alignments. The optimality of the match is defined by the objective function. If the objective function includes the quadratic term describing residue pairwise interactions, the general problem of finding the optimal alignment is NP-complete (Lathrop, 1994). Thus, one principal distinction among threading algorithms is determined by the objective function. The algorithms fall into three broad categories.

1. 1D algorithms that use only the information that is associated with the 1D features of the structure.
2. 1D/2D algorithms that apply the 1D search logic (objective function representation) for the optimal alignment but at various steps use the information inferred from the 2D features to redefine the objective function.
3. 2D algorithms that use the full 2D representation of the problem and deal with the higher complexity of the search space.

5.1. 1D algorithms

For one-dimensional models the sequence-structure alignment problem is analogous at an algorithmic level to sequence–sequence alignment. Given an objective function, an optimal alignment can be found using a dynamic programming alignment algorithm (Needleman and Wunsch, 1970; Smith and Waterman, 1981). This results in very fast sequence-structure alignment.

5.2. 1D/2D algorithms

The motivation in this type of algorithm is to include 2D information for its presumed greater information content, but use 1D alignment algorithms for speed. Thus, they look for an approximately optimal score, rather than the global optimum. In addition to 1D amino acid preferences, the score includes the 2D pairwise amino acid preferences. The expectation is that the inclusion of such preferences will improve the recognition of the best structural template. The algorithm of GenThreader (Jones, 1999) uses the quadratic terms of the objective function to evaluate the alignment generated by a 1D Smith–Waterman algorithm. These 2D scores together with sequence-profile similarity scores and solvation scores are the input to a neural network that assesses the confidence of the prediction. The neural network is trained on all known pairs of sequences that are known to have similar structures. The frozen approximation (Godzik *et al.*, 1992; Skolnick and Kihara, 2001) iteratively performs a 1D alignment using 2D information fixed by the previous 1D alignment step.

5.3. 2D search algorithms

All 2D threading algorithms face an alignment problem that is formally intractable, so they differ on the basis of whether they return a good approximate alignment quickly or the global optimum alignment more slowly. Gaps are typically restricted to loops or the ends of secondary structure elements in order to reduce the alignment space and because deletions and insertions are less likely in the tightly packed protein core.

The Gibbs Sampling Algorithm (Bryant, 1996) begins with a random alignment. At each step it randomly chooses a core secondary structure element C, generates all possible alternative alignments for it, calculates each new alignment score S,

chooses a new alignment with probability proportional to $\exp(-S/kT)$, and fixes C at the new location. The procedure iterates using an annealing schedule to reduce the nominal energy units kT, then picks a new random alignment and repeats. The method does not guarantee a global optimum alignment, but is very fast and gives good performance.

The divide and conquer threading algorithm (Xu *et al*., 1998) repeatedly divides the structure model into submodels, solves the alignment problem for submodels, and combines the subsolutions to find a globally optimal alignment. Dividing the model into smaller pieces means that some pairs of model positions that contribute to the pairwise score will be split between different submodels, so the cutting of a pair link is recorded and accounted for when the subsolutions are recombined.

The branch-and-bound search algorithm (Lathrop and Smith, 1996) repeatedly divides the threading search space into smaller subsets and always chooses the most promising subset to split next. Eventually, the most promising subset contains only one alignment, which is a global optimum. It relies on a lower bound on the best score achievable within each subset. An anytime version (Lathrop, 1999) returns a good approximation quickly, then iteratively improves the approximation until finally returning a global optimum alignment.

Protein threading by linear programming (Xu *et al*., 2003) formulates the threading problem as a large scale integer programming problem, relaxes it to a linear programming problem, and solves the integer program by a branch-and-bound method. It is also an optimal method.

6. Selecting a model from the library

Finding the optimal score and alignment of a sequence to a structure leaves open the question what is a likely structure of the query sequence. There are broadly two approaches to this question. The first chooses the structure on the basis of the best alignment score, usually after normalizing scores in some way so that scores from different models are comparable. The second integrates the total probability of a model across all alignments of the sequence to the model. The first approach is more popular and intuitive, though the second is better grounded in probability theory.

Unlike sequence–sequence alignment, there is no straightforward method to determine the statistical significance of the optimal score (Bryant and Altschul, 1995). The statistical significance of the score tells how likely it is to obtain a given optimal score by chance. The distribution of scores for a given query sequence and structural model depends on both the length of the sequence and the size of the model. Currently, there is no generic analytical description of the shape of the distribution of threading scores across different models and sequences, though it is well understood that the distribution of optimal scores is not normal. For gapped local alignment of two sequences, or a sequence and a sequence profile, the distribution of optimal alignment scores can be approximated by an extreme value distribution. Fitting the observed distribution of scores to the extreme value distribution function has been applied by profile threading method FFAS (Rychlewski *et al*., 2000). Some profile-based methods approximate the distribution of scores by a normal

distribution and calculate Z-scores. The Z-scores are calculated with the mean and standard deviation of the scores of a query sequence with the library of all structural models. Similarly, many threading approaches with the quadratic pairwise objective function use the optimal raw score as the primary measure of structure and sequence compatibility and estimate the statistical significance of the score assuming a normal distribution of the sequence scores threaded to a library of available models.

In the approach implemented by GenThreader (Jones, 1999), the neural network score determines the compatibility of the structure with the model. The input to the neural network is a set of values of different scores: sequence similarity scores, the solvent accessibility score, and the pairwise interaction scores. The Gibbs-sampling threading approach (Bryant, 1996) estimates the significance of the optimal score by comparison to the distribution of scores generated by threading a shuffled query sequence to the same structural model. The distribution of shuffled scores is assumed to be normal.

Lathrop *et al*. (1998) formulated a model-selection approach that does not rely on the optimal sequence-structure alignments. The compatibility of the sequence with a structural model is measured by the total alignment probability, where the probability is summed over all possible sequence-to-model alignments. This approach has been applied in the Bayesian fold recognition method (Bienkowska *et al*., 2000). The total probability is calculated using the filtering algorithm proposed and developed by White (1988). This method assumes that different structural folds (*see* Article 69, **Complexity in biological structures and systems**, Volume 7) are independent and equally probable structural hypotheses. The most probable model within the same fold category determines the prior probability of observing the sequence given the fold. The posterior probability of the fold given the sequence is calculated according to Bayes formula.

7. Performance of threading methods

The evaluation of different fold recognition (threading) methods takes place every two years during the Critical Assessment of Structure Prediction (CASP) (*see* Article 74, **Molecular simulations in structure prediction**, Volume 7) meeting. The prediction period is 3 months when different groups submit their predictions to a common depository. Targets for prediction are proteins that are about to have their structures solved. The CASP contest provides an opportunity to evaluate methods on the same sample set and in the context of truly "blind" predictions. This setting allows for performance comparison across methods using varied self-evaluation criteria applied by authors. Many threading approaches have been the subject of evaluation by past CASP contests. Examples of prediction methods that have been successful are: GenThreader (Jones, 1999), 3D_PSSM, (Kelley *et al*., 2000), FFAS (Rychlewski *et al*., 2000), SAM-2000 (Kelley *et al*., 2000), and PROSPECTOR (Skolnick and Kihara, 2001). Many other methods can be accessed at the CASP6 website **http://predictioncenter.llnl.gov/casp6/** (CASP, 2004) including 60 registered prediction servers. Over past CASP experiments, fold recognition methods have proved capable of recognizing distant sequence and structure similarities that are undetectable by sequence comparison methods. The evaluation of the last

CASP5 contest and description of several methods with best overall performance is described in Kinch *et al*. (2003).

Independently of the method, the accuracy of the prediction varies dramatically among proteins. Some proteins are much easier to predict than others and this can be recognized by consensus methods that automatically evaluate predictions generated by different algorithms. Consensus methods are generally more reliable than any single method alone. An example of a meta-server generating consensus predictions is Pcons **http://www.sbc.su.se/~arne/pcons/**. Figure 1 shows an example of a partially successful prediction for protein HI0817, *Haemophilus influenzae* (target T129 in CASP5 contest). The overall topology of three C-terminal helices of the structure is correctly recognized, while the rest of the molecule is mispredicted. Even with the correctly recognized topology, the alignment of the residues in those helices is shifted by four positions. Figure 2 shows an example of a successful prediction for the single-strand binding protein (SSB), *Mycobacterium tuberculosis* H37Rv, (target T151 in CASP5 contest). All secondary structure elements are correctly aligned as well as most loops. Exceptions are one internal loop that is missing a glycine residue and the C-terminal tail. The missing residue was absent from the original target sequence submitted for predictions. This last example demonstrates that threading can be very successful in predicting protein structure. However, much effort is still required from the research community to provide a true solution to the protein structure prediction.



(a) (b)

```
T0129_str    93/92     NVFTQADSLSDWANQFLLGIGLAQPELAKEKGEIGEAVDDLQDICQLGYDEDDNEEELAE
T0129_pred   95/96     QADSLSDWANQFLLGIGLAQPELAKEKGEIGEAVDDLQDICQLG--YDED------DNEE

T0129_str    153/152   ALEEIIEYVRTIAXLFYSHFN
T0129_pred   147/148   ELAEALEEIIEYVRTIAMLFY
```

(c)

Figure 1 Example of a partially correct structure prediction. (a) shows the predicted structure of the HI0817 protein. (b) shows the solved structure of that protein PDB code 1izmA. The left side of the picture corresponds to the C-terminal portion of the molecule. (c) shows the CE algorithm structural alignment of the C-terminal regions of the 1izmA and predicted structure of T0129. The topology of the three C-terminal helices is predicted correctly but the structural alignment shifts residues by four positions. Such shifts are typical misprediction of threading algorithms due to the periodicity of the helix structure. The predicted coordinates are the best structure prediction submitted to the CASP5 contest and are available from the CASP5 website



(a) (b)

```
T0151_str    5/6     TTITIVGNLTADPELRFTPSGAAVANFTVASTPRIYDRQTGEWKDGEALFLRCNIWREAA
T0151_pred   5/6     TTITIVGNLTADPELRFTPSGAAVANFTVAS-TPRIYDRQTGEWKDEALFLRCNIWREAA

T0151_str    65/66   ENVAESLTRGARVIVSGRLKQRSFETREGEKRTVIEVEV---DEIGPS
T0151_pred   64/66   ENVAESLTRGARVIVSGRLKQRSFETREGEKRTVIEVEVDEIGPSLRY
```

(c)

Figure 2 Example of a correct structure prediction. (a) shows the predicted structure of the HI0817 protein. (b) shows the solved structure of that protein PDB code 1izmA. (c) shows the CE algorithm structural alignment of the 1ue6A structure and the predicted structure of T0151. The structural elements are shown in the same colors as in (a) and (b), β-strands in yellow and α-helices in magenta. The structural alignment correctly aligns most residues between the model and the structure. The only misalignment is introduced by the omission of the GLY residue (indicated in red) from the predicted structure and the C-terminal tail. The predicted coordinates are the best structure prediction submitted to the CASP5 contest and are available from the CASP5 website

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Bienkowska JR, Rogers RG Jr and Smith TF (1999) Filtered neighbors threading. *Proteins*, **37**(3), 346–359.

Bienkowska JR, Yu L, Zarakhovich S, Rogers RG Jr and Smith TF (2000) Protein fold recognition by total alignment probability. *Proteins*, **40**(3), 451–462.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**(5016), 164–170.

Bryant SH (1996) Evaluation of threading specificity and accuracy. *Proteins*, **26**(2), 172–185.

Bryant SH and Altschul SF (1995) Statistics of sequence-structure threading. *Current Opinion in Structural Biology*, **5**(2), 236–244.

Bryant SH and Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**(1), 92–112.

CASP (2004) CASP6, **http://predictioncenter.llnl.gov/casp6/**.

Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG Jr and Haussler D (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins*, **49**(1), 7–14.

Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pacific Symposium on Biocomputing*, Hawaii, USA, 119–130.

Godzik A, Kolinski A and Skolnick J (1992) Topology fingerprint approach to the inverse protein folding problem. *Journal of Molecular Biology*, **227**(1), 227–238.

Higgins DG, Thompson JD and Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology*, **266**, 383–402.

Huang ES, Subbiah S, Tsai J and Levitt M (1996) Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *Journal of Molecular Biology*, **257**(3), 716–725.

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, **287**(4), 797–815.

Jones DT, Moody CM, Uppenbrink J, Viles JH, Doyle PM and Harris CJ (1996) Towards meeting the Paracelsus challenge: the design, synthesis, and characterization of paracelsin-43, an alpha-helical protein with over 50% sequence identity to an all-beta protein. *Proteins*, **24**(4), 502–513.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature*, **358**(6381), 86–89.

Karplus K and Hu B (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**(8), 713–720.

Kelley LA, MacCallum RM and Sternderg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*, **299**(2), 499–520.

Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H and Grishia NV (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53**(Suppl 6), 395–409.

Lathrop RH (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering*, **7**(9), 1059–1068.

Lathrop RH (1999) An anytime local-to-global optimization algorithm for protein threading in theta (m2n2) space. *Journal of Computational Biology*, **6**(3-4), 405–418.

Lathrop RH, Rogers RG Jr, Smith TF and White JV (1998). A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology*, **60**(6), 1039–1071.

Lathrop RH and Smith TF (1996) Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology*, **255**(4), 641–665.

Maiorov VN and Crippen GM (1994) Learning about protein folding via potential functions. *Proteins*, **20**(2), 167–173.

Munson PJ and Singh RK (1997) Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein Science*, **6**(7), 1467–1481.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Panchenko AR, Marchler-Bauer A and Bryant SH (2000) Combination of threading potentials and sequence profiles improves fold recognition. *Journal of Molecular Biology*, **296**(5), 1319–1331.

Rychlewski L, Jaroszewski L, Li W and Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, **9**(2), 232–241.

Singh RK, Tropsha A and Vaisman II (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *Journal of Computational Biology*, **3**(2), 213–221.

Sippl MJ (1995) Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, **5**(2), 229–235.

Skolnick J and Kihara D (2001) Defrosting the frozen approximation: PROSPECTOR–a new approach to threading. *Proteins*, **42**(3), 319–331.

Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A and Kihara D (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **53**(Suppl 6), 469–479.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

White JV (1988) Modelling and filtering for discretely valued time series. In *Bayesian Analysis of Time Series and Dynamic Models*, Spall JC (Ed.), Marcel Dekker: New York, pp. 255–283.

White JV, Muchnik I and Smith TF (1994) Modeling protein cores with Markov random fields. *Mathematical Biosciences*, **124**(2), 149–179.

Xu J, Li M, Lin G, Kim D and Xu Y (2003) Protein threading by linear programming. *Pacific Symposium on Biocomputing*, Hawaii, USA 264–275.

Xu Y, Xu D and Uberbacher EC (1998) An efficient computational method for globally optimal threading. *Journal of Computational Biology*, **5**(3), 597–614.

Zheng W, Cho SJ, Vaisman II and Tropsha A (1997) A new approach to protein fold recognition based on Delaunay tessellation of protein structure. *Pacific Symposium on Biocomputing*, Hawaii, USA 486–497.

Short Specialist Review

Acedb genome database

Richard Durbin and Edward Griffiths
Wellcome Trust Sanger Institute, Cambridge, UK

1. Introduction

Acedb (A C Elegans DataBase) was first developed by Jean Thierry-Mieg and Richard Durbin in the context of the *Caenorhabditis elegans* sequencing project in the early 1990s, with the aim of integrating the new sequence with existing genetic and physical map information, and research data from publications. In this era, prior to the World Wide Web, the software and data were made publicly available for researchers to download to provide a graphical interactive interface to the genome data.

It was adopted by many other genome projects and research communities, including those for the yeast, *Saccharomyces cerevisiae* (SGD), the fruit fly *Drosophila* (BDGP), and the model plant *Arabidopsis* (AAtDB) in addition to many smaller scale applications, and played a key role in the early stages of the human genome project in managing physical map construction and genome annotation. Many of these have now moved on to use dedicated relational database systems, but WormBase, the model organism database for *C. elegans*, and a large part of the curated sequence annotation for vertebrate genomes continue to use Acedb, as well as a continuing number of smaller groups.

With the adoption of server–client architectures and scripting languages such as Perl by the bioinformatics community in the 1990s, a server version of Acedb was provided and the AcePerl interface was developed by Lincoln Stein. These support World Wide Web displays and distributed use in larger scale environments. Since its inception, Acedb has been ported to DOS, Unix, MacOS, and Windows operating systems, and most related graphics environments. The scale of databases supported has increased from a few megabytes to tens of gigabytes, with Acedb able to manage annotation on the complete human genome.

2. Architecture

The data model underlying Acedb is neither a traditional relational database design, where data are stored in tables, nor a standard object-oriented design. One of the features that has facilitated the adoption of Acedb by biological researchers is that data are stored in objects that are relatively complex and can conform naturally to standard things that people name and refer to, such as clones, genes, alleles, and so on.

More formally, Acedb supports semistructured data arranged in hierarchical objects, in the same sort of way as ASN.1 (which underlies the NCBI's data design) or XML, which has become widely used since 2000 for data exchange (see Abiteboul *et al*., 2000 for a discussion of Acedb in the context of modern semistructured data theory). However, it is strongly typed, with the schema organized into a number of classes, with each class consisting of a hierarchical tree structure of user-definable tags and data fields. All permissible fields in the hierarchy for a class, known as *the model*, have strongly typed attributes, which can include references to objects in another specified class. Objects strictly contain only the tags and data necessary to represent the available data for that object. Objects, and indeed the schema, can be easily extended dynamically to add more detailed information in an object-specific fashion. Automatically maintained bidirectional references are also supported.

There have been a series of query languages developed, the first of which returned sets of objects satisfying a query, and the more recent AQL (Ace Query Language) giving tables specified using a more powerful declarative grammar. From the practical point of view, Acedb provides two-level caching with a shadow page system to support rollback and protection from hardware or software failure. It also manages indexes to speed up queries.

3. Design and use

The graphical interface to Acedb provides a multiple window hyperlink style environment, like modern Web browsers (see Figure 1). There are generic object displays that can show any object however specified. In addition, there are a number of powerful specific displays for genomic DNA sequence, genetic maps, microarray data, and so on, each of which has associated search and analysis tools. These include built-in interfaces to the gene-structure prediction program Genefinder (Phil Green), genetic map likelihood calculators, and sequence alignment and dot-plot viewers (Erik Sonnhammer). The specialized displays are highly configurable, relying on the values of certain object attributes for key information, but also able to read and display other attributes on the basis of configuration objects stored in the database.

Data can be entered, edited, and exported in a standard structured text file format (.ace files) on the basis of simple paragraphs per object with tag-value lines for attributes. This format is extremely compact as only the data and its associated tag need be specified; intermediate tags can be intuited from the class model. However, there are also interactive editing interfaces, both in the generic object display, and more powerfully in the specialized displays. One of the key continuing uses of Acedb is to support curation of gene exon–intron structures in genomic DNA. All edits are tagged with the time and user, so that a record of data origin is automatically maintained. Similarly, specialized output formats are available for relevant data types, including FASTA and EMBL/Genbank format for sequence, and GFF format for sequence features.

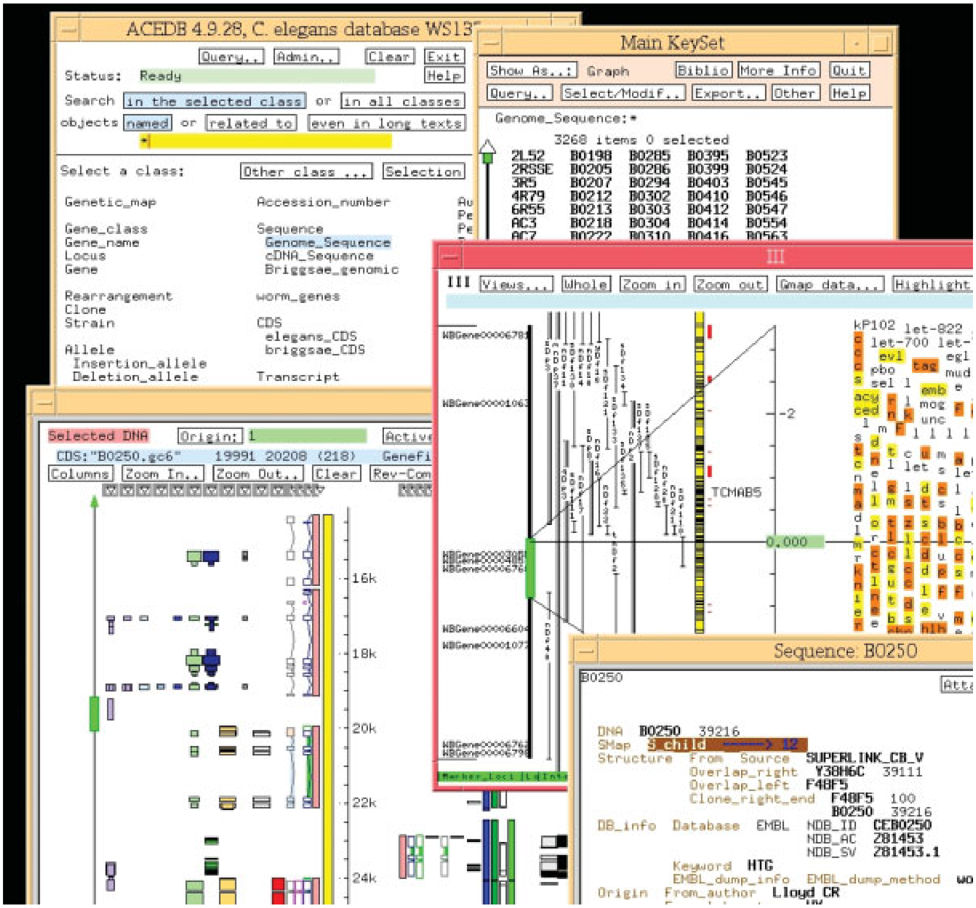


Figure 1 A view of the graphical interface to Acedb, showing examples of specific interfaces that are available, including a control window (top left) with examples of classes, a listing of objects in a class, a genomic DNA display showing genes and related evidence and features, a genetic map display with chromosomal rearrangements and gene locations, and a generic object display

4. Conclusions

Genome projects generate data as their primary output. These data are most valuable in the context of other biological information. Acedb developed within this environment to meet these needs. From the start, it was apparent that it would have to support a very wide range of depth of information, where a lot of detail is available about a few items, whereas very little if anything is known about most objects. Such is the nature of science. In addition, new types of information are always becoming available, so the design itself had to be flexible. And finally there needed to be an easy and intuitive way for researchers to find and interact with specific data.

Out of these requirements arose a novel general database management system and a complex domain-specific user interface for genome data annotation, both substantial software goals. In time, other solutions have been provided for most of the problems that Acedb addressed, many of them influenced to some degree by Acedb. However, Acedb performance and features remain competitive with these newer systems in the areas in which it is actively used and supported, and it is likely to continue to be used for the foreseeable future.

Acedb is freely available for download under the Gnu Public Licence (GPL) from the website **www.acedb.org**, which also contains much additional information.

Further reading

Durbin R and Thierry-Mieg J (1994) The ACEDB genome database. In *Computational Methods in Genome Research*, Suhai S (Ed.) Plenum Press, pp. 45–56.

Stein LD and Thierry-Mieg J (1998) Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Research*, **8**, 1308–1315.

Reference

Abiteboul S, Buneman P and Suciu D (2000) *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann: San Francisco, pp. 22–24.

Short Specialist Review

Design of KEGG and GO

Minoru Kanehisa
Kyoto University, Kyoto, Japan

1. Introduction

An early attempt to classify gene products into categories of cellular functions was done by Monica Riley for *Escherichia coli* (Riley, 1993). Her classification was largely biased toward metabolism, what was already known at that time, but the basic idea was adopted and extended by a number of groups doing complete sequencing of prokaryotic genomes, a notable example being the role category in the TIGR's Comprehensive Microbial Resource (Peterson *et al*., 2001). Obviously, such a classification system is useful for quickly uncovering metabolic and other capabilities of an organism by checking whether necessary components are present in the genome. GO (Gene Ontology) is a far more elaborate system, but its activities started from similar efforts for eukaryotic genome sequencing and annotation (*see* Article 82, **The Gene Ontology project**, Volume 8).

Another important classification scheme for protein functions is the EC (Enzyme Commission) numbering system for enzymes assigned by the Nomenclature Committee of the IUBMB and IUPAC (Barrett *et al*., 1992). The EC number represents a hierarchical classification of enzymatic reactions, six classes, subclasses and sub-subclasses, as well as serial numbering for substrate specificity. The EC numbers are used both as identifiers of reactions and as identifiers of enzymes that catalyze reactions. This duality makes it possible to superimpose the genomic content of enzyme genes (EC numbers) onto known metabolic pathways represented as networks of reactions (also EC numbers), thus to infer metabolic capabilities from the genome (Kanehisa, 1997). This feature was first successfully implemented in KEGG (Kyoto Encyclopedia of Genes and Genomes), providing more detailed pictures than the role category approach. For the purpose of common identifiers, the EC numbers were subsequently replaced by the KO (KEGG Orthology) identifiers in KEGG in order to extend to signaling and other types of regulatory pathways.

2. Design of GO

GO (**http://www.geneontology.org/**) is a database consisting of three structured, controlled vocabularies (ontologies) for describing how gene products behave in a cellular context (Ashburner *et al*., 2000). Its main objective is to compile and

share expanding knowledge about genes and gene products in different organisms. Despite its naming, GO lacks the basic properties for ontology in information science (Smith *et al*., 2003), but it has become a widely accepted standard for genome annotation in biological communities. The three ontologies are named molecular function, biological process, and cellular component, representing respectively molecular level functions of individual gene products, higher-level functions in the interaction network with other molecules, and locations within various structural components of the cell.

The three ontologies are separately defined structures of controlled vocabularies. More precisely, each structure is represented by the directed acyclic graph (DAG), which is somewhat different from the hierarchy, allowing a child to have multiple parents. Although no relations are defined across the three ontologies, intrinsic relations become apparent when GO terms are assigned to individual gene products. For example, dapB gene in *E. coli* codes for an enzyme, dihydrodipicolinate reductase (EC 1.3.1.26), which appears in the lysine biosynthesis pathway in bacteria. Thus, the gene product DapB would be associated with the GO terms:

GO:0008839: dihydrodipicolinate reductase activity
in the molecular_function ontology and
GO:0009089: lysine biosynthesis via diaminopimelate

in the biological_process ontology. They are part of separate DAGs, which include the EC number hierarchy for GO:0008839 and the generic to more specific descriptions of the biosynthesis pathways for GO:0009089. The cellular_component ontology contains terms for the anatomical structure of the cell, as well as for complexes of gene products, which are more often utilized in eukaryotes.

3. Design of KEGG

One of the most basic concepts in the design of KEGG (**http://www.genome.jp/kegg/**) is the level of abstraction illustrated in Figure 1. It also relates to the concept of nested graph. A graph is a set of nodes and edges, and a nested graph is a graph whose nodes can themselves be graphs. For example, the bottom item in Figure 1 is the three-dimensional (3D) structure of *E. coli* dihydrodipicolinate reductase (DapB), which is a graph with atoms as its nodes and atomic bonds as its edges. When the subgraph corresponding to each amino acid is converted to one of the 20 letters, such as C for cystein, a nested graph is obtained, emphasizing the information at the amino acid sequence level. In order to capture higher features at the protein network level, the whole graph object of the amino acid sequence or the 3D structure is better converted to a symbol, like DapB. Here, a higher-level graph represents the lysine biosynthesis pathway in *E. coli*, in which nodes are enzymes and edges are relations between enzymes.

Furthermore, in order to understand generic pathways (called reference pathways in KEGG) that are conserved among organisms, a huge graph object called the *gene universe* is considered in which nodes are genes (or gene products) and edges are, among others, ortholog (or sequence similarity) relations. The ortholog cluster of

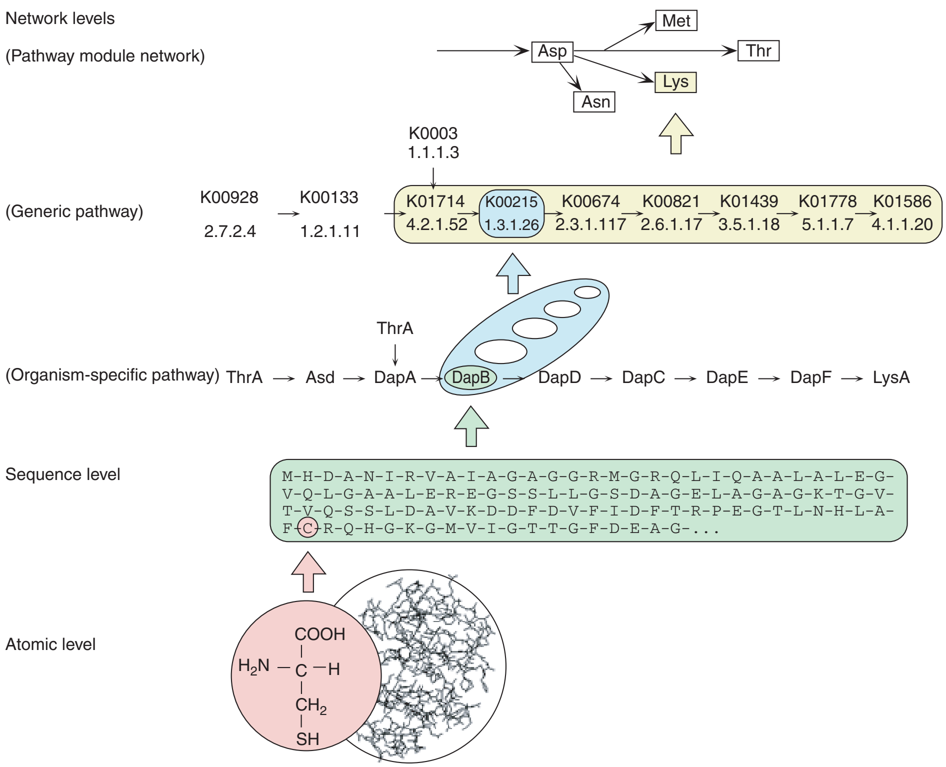


Figure 1 The level of abstraction represented by nested graphs

DapB proteins from various organisms would be a clique-like subgraph and it is converted to a node, with the KO (KEGG Orthology) identifier K00215. There may still be other types of modularity and hierarchy in still higher-level networks. For example, a lysine pathway module may be defined for the KO nodes K01714 to K01586 because of the continuous reaction steps and the coexpression (operon-like structure) information.

Thus, the *E. coli* gene dapB (accession number eco:b0031) is placed in the KEGG hierarchy as follows.

01100 Metabolism
  01150 Amino acid metabolism
    00300 Lysine biosynthesis
      K00215 E1.3.1.26, DAPB; dihydrodipicolinate reductase
        eco:b0031 dapB; dihydrodipicolinate reductase

Here, the third level corresponds to a KEGG pathway map (Lysine biosynthesis can be seen at **http://www.genome.jp/dbget-bin/get_pathway?org_name=ot&mapno=00300**). When compared to GO, this hierarchy is much simpler, but it roughly corresponds to the biological_process ontology and also includes some aspects (molecular complexes) of the cellular_component ontology. The information about the molecular_function ontology is organized in KEGG as many different

structural/functional hierarchies of proteins, including enzyme nomenclature (EC numbers) and subfamily classifications of G-protein-coupled receptors, ion-channel receptors, and other protein families. Each member of these groups or families is assigned a KO identifier, so the additional hierarchies are orthogonal to the fourth level in the hierarchy shown above, not the fifth level (individual genes) as in GO.

In contrast to GO where various types of functions are specified by controlled vocabulary, the specification of functions in natural language is deliberately left out in KEGG. Instead, functions are considered to be encoded, thus implicitly defined, in structural patterns of graph objects. In a similar way that sequence motifs and 3D structural motifs are identified, specific connection patterns of specific gene products may be related to specific cellular functions, such as lysine biosynthesis. In essence, the structural hierarchy of nested graphs also represents the functional hierarchy, enabling logical reasoning of higher-level functions from molecular level information.

4. Integration of genomics and chemistry

Both KEGG and GO can be used to analyze not only the information in the genome for understanding potential capabilities but also the information in the transcriptome or the proteome for understanding actual capabilities. For example, gene expression profiles in microarray experiments can be mapped to KEGG pathways or GO terms to uncover dynamic changes of cellular processes. KEGG can further be used to analyze the metabolome and the glycome, which represent cellular repertoires of small chemical compounds and glycans respectively. Such chemical substances are highly correlated with genomic contents, because they are encoded in the genome in terms of the biosynthetic and biodegradation pathways. In other words, knowledge about chemistry in the cell would greatly facilitate the process of deciphering the genome.

In KEGG, the integration of genomics and chemistry is based on the concept of line graphs (also called interchange graphs), in which nodes and edges are interchanged. The metabolic pathway is a network of chemical compounds in biochemistry textbooks, representing their conversions by enzymatic reactions. In contrast, the metabolic pathway is a network of enzymes (reactions) in KEGG, which can be linked to enzyme genes in the genome. These two networks are complementary; one can be related to the other by a line graph transformation, interchanging nodes and edges between compounds and enzymes. This means, for example, that if we know a set of chemical structures of secondary metabolites, we can infer reaction networks and then uncover a set of enzyme genes in the genome. More generally, the line graph analysis would uncover principles of how chemical or environmental perturbations cause genomic changes and vice versa. As more and more genomes are sequenced, we are beginning to understand the universality and diversity of the genomic space. However, we still have little knowledge about the universality and diversity of the chemical space that is relevant to life. KEGG is designed to explore both the genomic space and the chemical space by integrating genomes, networks, and chemistry in the three main databases named GENES, PATHWAY, and LIGAND.

Further reading

Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27–30.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Barrett AJ, Canter CR, Liebecq C, Moss GP, Saenger W, Sharon N, Tipton KF, Vnetianer P and Vliegenthart VFG (1992) *Enzyme Nomenclature*, Academic Press: San Diego.

Kanehisa M (1997) A database for post-genome analysis. *Trends in Genetics*, **13**, 375–376.

Peterson JD, Umayam LA, Dickinson T, Hickey EK and White O (2001) The comprehensive microbial resource. *Nucleic Acids Research*, **29**, 123–125.

Riley M (1993) Functions of the gene products of *Escherichia coli*. *Microbiological Research*, **57**, 862–952.

Smith B, Williams J and Schulze-Kremer S (2003) The ontology of the gene ontology. *Proceedings AMIA Symposium*, **2003**, 609–613.

Short Specialist Review

Simulation of biochemical networks

Andrzej M. Kierzek
Institute of Biochemistry and Biophysics, Warsaw, Poland

1. Introduction

Recent advances in experimental techniques of molecular biology have resulted in the unprecedented ability to measure and alter the properties of basic molecular components of complex biochemical processes occurring within the living cell. This has resulted in the creation of Systems Biology (Kitano, 2002), an emerging field aimed toward understanding systemic properties of living systems in terms of the knowledge of molecular components. One of the major scientific activities within this field is the mathematical modeling of cellular processes. Models of cellular processes most commonly take the shape of the networks of molecular interactions. Frequently, the complexity of biochemical systems makes impossible any analytical solution of the mathematical equations representing these models, thus creating the need to perform computer simulations. This chapter focuses on the computational approaches used in the simulation of biochemical network dynamics.

2. System representation

A biochemical reaction network may be described as a set of coupled chemical reactions. In this representation, the system is modeled as a set of N molecular species $(S_1, \ldots, S_N)$ interacting via M chemical reaction channels $(R_1, \ldots, R_M)$ in the reaction environment with volume V. The state of the system at time t is defined by the vector $\mathbf{X}(t) = (X_1(t), \ldots, X_N(t))$, where $X_i(t)$ is the number of molecules of substance S_i present in reaction environment. Amounts may be also expressed as concentrations, that is, ratios of the numbers of molecules to volume: $\mathbf{c}(t) = (c_1(t), \ldots, c_N(t))$, where $c_i(t) = X_i(t)/V$. Chemical reactions are described by the stoichiometric matrix $\mathbf{n}_{N,M}$, where $\mathbf{n}_{i,j}$ is the change of X_i resulting from one reaction R_j. A simple reaction network, representing the Michaelis–Menten mechanism of enzymatic reaction, is shown in Figure 1(a), as well as the corresponding stoichiometric matrix in Figure 1(b).

The notion of "reaction" is arbitrary in the modeling of biochemical networks. Elementary reactions in biochemical networks usually represent effective processes

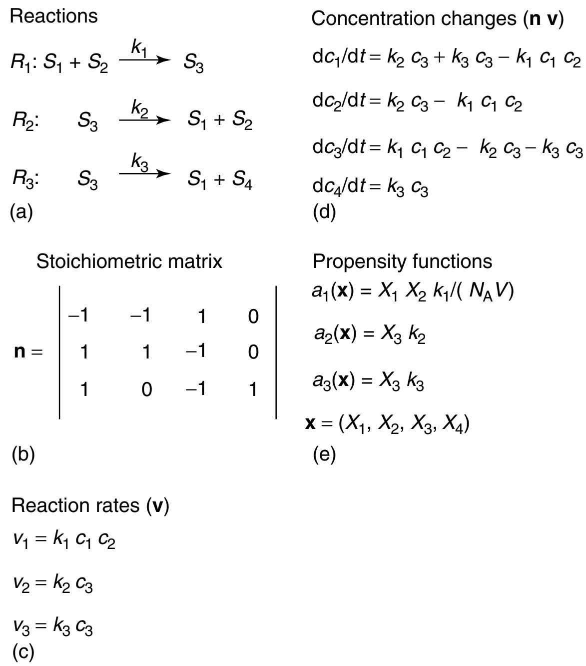


Figure 1 Example of a simple biochemical reaction network representing Michaelis–Menten mechanism of enzymatic reaction. (a) A set of three coupled chemical reactions. S_1, S_2, S_3, S_4 represent enzyme, substrate, enzyme substrate complex, and product respectively. Rate constants of three reactions are denoted by k_1, k_2, and k_3. (b) Stoichiometric matrix of the example reaction network shown in (a). (c) Reaction rates. (d) A set of simultaneous differential equations used to simulate the time evolution of the example reaction network. Note that the vector of concentration changes ($\mathrm{d}c_1/\mathrm{d}t, \ldots, \mathrm{d}c_4/\mathrm{d}t$) is the product of the stoichiometric matrix and the vector of reaction rates. (d) Propensity functions for the example network. Second-order rate constant k_1 is multiplied by $1/N_A V$ (V: volume of the reaction environment, $\mathrm{N_A}$: Avogadro's number) to represent the number of reactive molecular collisions occurring in the reaction environment in the unit of time. First-order rate constants k_2, k_3 represent the number of molecular conversions per unit time in both deterministic and stochastic chemical kinetics

in which many molecular processes are "lumped" together. For example, it is common to model the complex mechanism of transcription initiation by a single reaction, with the effective rate constant representing transcription initiation frequency. The level of approximation and detail depends on the knowledge available for the particular system and the scientific questions or engineering problems to be addressed.

3. Numerical integration of differential equations

Once the system is specified, the time evolution of $\mathbf{X}(t)$ may be studied. Most commonly, it is done by numerical solution of simultaneous differential equations

expressing the rates of concentration change for every substance in the system:

$$\begin{aligned} \frac{\mathrm{d}c_1}{\mathrm{d}t} &= f_1(c_1, \ldots, c_N) \\ &\ldots \\ \frac{\mathrm{d}c_N}{\mathrm{d}_t} &= f_N(c_1, \ldots, c_N) \end{aligned} \tag{1}$$

The functions $f_1, \ldots, f_N$ describe the dependence of the rate of concentration change of a particular substance on the concentrations of other substances in the system and are determined by the stoichiometric matrix and mechanism of reactions $R_1, \ldots, R_M$. They also contain quantitative parameters, referred to as rate constants; these values must be known or estimated in order to obtain quantitative results. An example of a system of differential equations describing a simple reaction network is shown in Figure 1(d).

The time evolution of biochemical reaction networks modeled by the approach described above may be simulated using standard algorithms for the numerical solution of simultaneous differential equations (Figure 2). Therefore, popular software packages such as MATHEMATHICA and MATLAB are frequently used for the simulation. There are also software environments dedicated to the modeling of biochemical reaction networks that implement the numerical solution of differential equations (e.g., GEPASI (Mendes, 1993), E-cell (Tomita *et al*., 1999); see Table 1 for comprehensive list). Examples of applications of this methodology are too numerous to be listed here. They include modeling of metabolic pathways (Wang *et al*., 2001), cell cycle (Tyson *et al*., 2001), and signal transduction cascades (Bhalla and Iyengar, 1999). Such simulations are relatively cheap computationally. Simulations of metabolic networks performed on the whole-cell scale were done in real time on a standard PC computer (Tomita *et al*., 1999).

4. Rule-based approach

The major obstacle in computer simulations of biochemical systems is the lack of knowledge of precise parameters (rate constants) of the reactions. One of the solutions to this problem is to use qualitative rather than quantitative models of biochemical reaction networks. In the qualitative approach, frequently referred to as "logical formalism" (Mendoza *et al*., 1999) or "rule-based simulation", the amounts of substances are replaced by discrete states (most commonly binary states), which denote "activation" levels of the molecular components of the system. This methodology is most frequently applied to so-called genetic networks in which "substances" represent genes and "reactions" represent regulatory influences of the particular gene product on other genes in the system. The binary states of the molecular components represent two activation levels of the gene ("on", "off" or "active", "inactive"). The state of a particular gene in the network is determined by a boolean (i.e., logical) function, which depends on the states of other genes in the system and network connectivity. Logical functions correspond to the rate expressions in equation 1, where it may be argued that they represent the sign of the rate. Here,

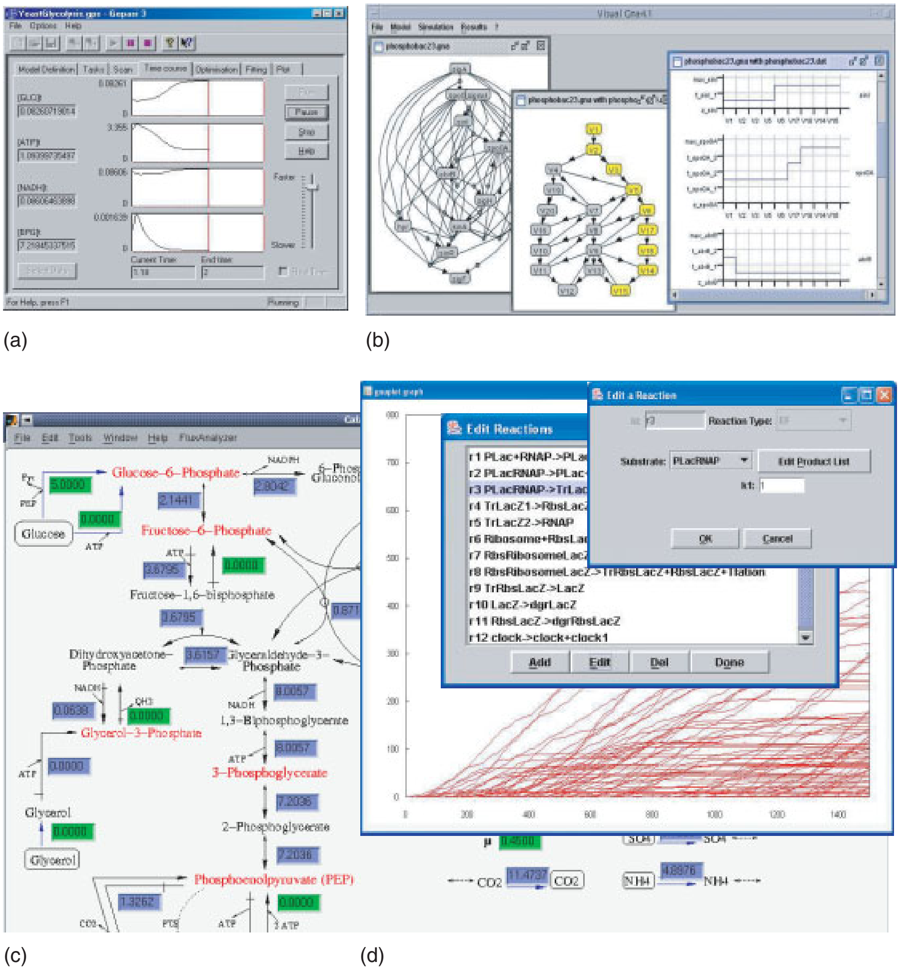


Figure 2 Screenshots of selected simulation programs. (a) Time courses computed by numerical integration of differential equations with Gepasi. (b) Qualitative model building with GNA. (c) Computation and visualization of metabolic flux distribution with FluxAnalyzer. (d) STOCKS software: kinetic model building and individual time courses computed with Gillespie algorithm

a positive sign means activation, while a negative sign denotes repression. In spite of such drastic simplifications, discrete models have provided useful predictions in studies of gene networks *Drosophila* embryos (Sánchez and Thieffry, 2003).

5. Stationary state analysis

Another way of solving the problem of sparse information on quantitative parameters is to study the flux distribution in the system at stationary state. The system at stationary state satisfies the following condition:

$$\mathbf{nv} = 0 \tag{2}$$

Table 1 A selection of software packages for simulations of biochemical reaction networks

Name	Methodology[a]	WWW page
BioSPICE	I, IV	**http://www.biospice.org/**
Dizzy	I, IV	**http://labs.systemsbiology.net/bolouri/software/Dizzy/**
E-cell	I, IV, V	**http://www.e-cell.org/**
Gepasi	I	**http://www.gepasi.org/**
GNA	II	**http://www-helix.inrialpes.fr/gna**
FluxAnalyser	III	**http://www.mpi-magdeburg.mpg.de/research/projects/1010/1014/1020/mfaeng/fluxanaly.html**
Jarnac	I	**http://www.cds.caltech.edu/~hsauro/Jarnac.htm**
Kinetikit	I, V	**http://www.ncbs.res.in/~bhalla/kkit/**
Kinsolver	I	**http://lsdis.cs.uga.edu/~aleman/kinsolver/**
Metabologica	I, III	**http://www.metabologica.com/**
MetaFluxNet	III	**http://mbel.kaist.ac.kr/mfn/**
NetBuilder	I, II	**http://strc.herts.ac.uk/bio/maria/NetBuilder/**
SimPheny	III	**http://www.genomatica.com/**
StochSim	VI	**http://info.anat.cam.ac.uk/groups/comp-cell/StochSim.html**
STOCKS	IV, V	**http://www.sysbio.pl/stocks/**
Promot/DIVA	I	**http://www.mpi-magdeburg.mpg.de/research/projects/1002/comp_bio/promot**

[a]Methodologies: (I) numerical integration of differential equations; (II) qualitative simulation; (III) metabolic flux analysis; (IV) exact stochastic simulation; (V) multiscale, approximated stochastic simulation; (VI) StochSim algorithm.

where **n** is the stoichiometric matrix and **v** is the flux vector, that is, the vector of all reaction rates (Figure 1c) in the system. Equation 2 may be analyzed using methods of linear algebra to study alternative flux distributions available for the system (Schuster *et al*., 2000) or to find the flux distribution that optimizes certain criteria, for example, rate of cell growth (Edwards *et al*., 2001). These methods require only information on the stoichiometric matrix or limited information on some of the fluxes. They were used to find alternative metabolic pathways, to indicate missing annotations of enzymes in bacterial genomes, and to predict mutant phenotypes by using only partial experimental knowledge of certain metabolic fluxes to calculate the flux distribution in cellular metabolism. There was also an interesting attempt to combine stationary state analysis of metabolic pathways with the rule-based simulation of underlying gene regulatory network (Covert and Palsson, 2002).

6. Stochastic simulations

The simulation methods described above are deterministic in a sense that they do not take into account random fluctuations in the outcomes of elementary reactions in the system. They consider reactions as continuous fluxes of matter. This approach is valid if the numbers of molecules are "effectively infinite", as happens in the case of chemical processes occurring in bulk. However, biochemical processes occur in very small volumes (on the order of 10^{-15} L) and involve very small numbers of molecules. There are examples of gene regulatory mechanisms in which as

few as 10 molecules of transcription factor interact with a single "molecule" of gene regulatory region. In such systems, stochastic fluctuations in the outcomes of individual reactions, resulting from random motion of the molecules, become significant and the deterministic simulation approaches are not valid. Moreover, it has been shown, both by theoretical studies and experiments, that stochastic fluctuations in elementary reactions may alter the course of important cellular processes such as gene regulation and metabolism and lead to heterogeneity in populations of genetically identical cells (Puchalka and Kierzek, 2004; Kierzek *et al*., 2001; Rao *et al*., 2002).

To account for the role of random effects in the dynamics of biochemical reaction networks, the stochastic formulation of chemical kinetics and Monte Carlo computer simulation methods need to be employed. In the stochastic formulation of chemical kinetics, the amounts of substances are expressed as the numbers of molecules ($\mathbf{X}(t) = (X_1(t), \ldots, X_N(t))$) rather than concentrations. Every reaction R_μ can be described by its propensity function a_μ such that:

$$a_\mu(\mathbf{x})\mathrm{d}t \equiv \text{probability, given the current state of the system } \mathbf{X}(t) = \mathbf{x}, \text{ that reaction } R_\mu \text{ will occur somewhere inside the volume } V \text{ in next infinitesimal time interval } (t, t + \mathrm{d}t). \tag{3}$$

Propensity functions for the simple example system are shown on Figure 1(e).

Stochastic kinetic simulation algorithms use the reaction probability density function $P(\tau, \mu|\mathbf{x}, t)$ (Gillespie, 1977) to generate a statistical sample of system time courses. By definition, $P(\tau, \mu|\mathbf{x}, t)\,\mathrm{d}\tau$ is the probability, given the state of the system $\mathbf{X}(t) = \mathbf{x}$, that the next reaction in the system will occur in the infinitesimal time interval $(t + \tau, t + \tau + \mathrm{d}\tau)$ and will be an R_μ reaction. The reaction probability density function has the form (Gillespie, 1977):

$$P(\tau, \mu|\mathbf{x}, t) = a_\mu(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau), \text{ where } a_0(\mathbf{x}) = \Sigma_{1,\ldots M}^{a_\mu(\mathbf{x})} \tag{4}$$

In stochastic simulations, a random pair (τ, μ) is generated according to the joint probability density function $P(\tau, \mu|\mathbf{x}, t)$, and simulation variables are updated in the following way: (1) the state of the system is updated by adding the row of stoichiometric matrix describing reaction R_μ; (2) the time of the simulation t is increased by τ; and (3) the propensity functions of all the reactions are recomputed. Iteration of these steps, until the preset timescale is covered, results in a single time course of the system. An appropriate number of independent simulations generates a sample of time courses that are used to compute statistical properties of the system. This approach is frequently referred to as the exact stochastic simulation because it considers the individual reactive collisions occurring at exact time. No arbitrary time step is used and no assumptions other than the uniform distribution of molecules in the reaction environment are invoked. The most commonly used exact stochastic simulation algorithm is Gillespie's direct method (Gillespie, 1977; see Figure 3) that requires generation of two pseudorandom numbers in every iteration. Recently, Gibson and Bruck (2000) proposed the simulation algorithm that remains exact but requires only one call for random number generator per

Let the state of the system at the exact time t be:
$x(t) = (X_1, X_2, X_3, X_4)$

For reactions R_1, R_2,R_3 compute propensity functions a_1, a_2, a_3 (see Figure1). Compute the sum of propensity functions $a_0 = a_1 + a_2 + a_3$.

Generate two random numbers r_1,r_2 uniformly distributed over unit interval (0,1).

Generate the waiting time for the next reaction:
$\tau = (1/a_0)\ \ln(1/r_1)$

Choose next reaction μ such that
$a_1 + ,..., + a_{\mu-1} < r_2 a_0 \leq a_1 + ,..., + a_\mu$

Let us assume that reaction R_1 has been selected.

Update the system. Set simulation time to $t + \tau$ and change the numbers of molecules according to one elementary reaction R_1:
$x(t + \tau) = (X_1 - 1, X_2 - 1, X_3 + 1, X_4)$

Figure 3 Example of exact stochastic simulation. The figure shows a single iteration of Gillespie direct method (Gillespie, 1977) applied to the example network presented in Figure 1

iteration and updates the system in time proportional to the logarithm of the number of reactions, not to the number of reactions itself.

In spite of the availability of the efficient Gibson and Bruck algorithm, accuracy in exact stochastic simulation is achieved at the price of a significant computational burden since for every reactive collision happening in the system at least one random number must be generated. Owing to this fact, it is not practical to use exact stochastic simulation algorithms for biochemical networks involving intensive enzymatic reactions in which as many as 10^8 simulation events must be executed to simulate one second of the time course. Intensive reactions alone could be simulated, in the stochastic framework, by Gillespie's τ-leap method (Gillespie, 2001) in which the number of reactive collisions k_μ happening within the specified time step τ is computed for every reaction R_μ as $P(a_\mu \tau)$, a Poisson random variable with parameter $a_\mu \tau$. However, if the system contains even single reaction involving small numbers of molecules, accuracy of the τ-leap method requires a very small time step and performance of the simulation drops to the level of exact methods. Therefore, the major challenge in stochastic simulations of biochemical reaction networks is to design algorithms capable of efficient simulation of the systems including simultaneous reactions involving small numbers of molecules (e.g., gene expression) and intensive metabolic reactions.

The problem of efficient multiscale stochastic simulation has been addressed by the hybrid simulation algorithms that partition the system into "slow" and "fast" reaction subsets and subsequently apply exact stochastic simulation methods to "slow" reactions and the deterministic method or the τ-leap method to "fast" reactions (Haseltine and Rawlings, 2002; Puchalka and Kierzek, 2004). Recently,

Puchalka and Kierzek (2004) demonstrated that a novel hybrid simulation algorithm, dubbed "maximal time-step method", could be used to study stochastic effects in the biochemical reaction network involving gene expression, enzymatic activity, signal transduction, and transport processes simultaneously. The maximal time-step method is implemented in the recent version of STOCKS simulation software (Kierzek, 2002).

The simulation methods described above do not consider the spatial distribution of the molecules. The StochSim algorithm represents every molecule in the system as an individual software object that makes explicit treatment of positional information possible. In a single simulation time step, the program randomly selects two objects and executes one elementary reaction if the selected molecules can react and the unit-interval random number exceeds the reaction probability determined according to the rate constant. Unimolecular reactions are simulated by selecting a molecule and a pseudomolecule. StochSim has been applied to the simulation of signal transduction in bacterial chemotaxis. In recent simulations, the spatial clustering of the receptors in the cell membrane is explicitly simulated (Shimizu *et al*., 2003). Ability of incorporation of positional information comes at the price of computational cost; StochSim algorithm is slower than exact simulation algorithms of Gillespie and Gibson and Bruck.

Advances in Systems Biology motivate experimental work dedicated to in vivo determination of quantitative parameters that may be used to build quantitative models of cellular processes. As the knowledge of quantitative parameters of elementary interactions within the cell will accumulate, the methods described in this chapter are likely to gain importance and become standard tools of computational biology.

References

Bhalla US and Iyengar R (1999) Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381–387.

Covert MW and Palsson BO (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *The Journal of Biological Chemistry*, **277**, 28058–28064.

Edwards JS, Ibarra RU and Palsson BO (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, **19**, 125–130.

Gibson MA and Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry*, **104**, 1877–1889.

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, **81**, 2340–2361.

Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, **115**, 1716–1733.

Haseltine EL and Rawlings JB (2002) Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *Journal of Chemical Physics*, **117**, 6959–6969.

Kierzek AM (2002) STOCKS: stochastic kinetic simulations of biochemical systems with Gillespie algorithm. *Bioinformatics*, **18**, 470–481.

Kierzek AM, Zaim J and Zielenkiewicz P (2001) The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *The Journal of Biological Chemistry*, **276**, 8165–8172.

Kitano H (2002) Computational systems biology. *Nature*, **420**, 206–210.

Mendes P (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Computer Applications in the Biosciences*, **9**, 563–571.

Mendoza L, Thieffry D and Alvarez-Buylla ER (1999) Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics*, **15**, 593–606.

Puchalka J and Kierzek AM (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical Journal*, **86**, 1357–1372.

Rao CV, Wolf DM and Arkin AP (2002) Control, exploitation and tolerance of intracellular noise. *Nature*, **420**, 231–237.

Sánchez L and Thieffry D (2003) Segmenting the fly embryo: a logical analysis of the pair-rule cross-regulatory module. *Journal of Theoretical Biology*, **224**, 517–537.

Schuster S, Fell DA and Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, **18**, 326–332.

Shimizu TS, Aksenov SV and Bray D (2003) A spatially extended stochastic model of the bacterial chemotaxis signalling pathway. *Journal of Molecular Biology*, **329**, 291–309.

Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, *et al*. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics*, **15**, 72–84.

Tyson JJ, Chen K and Novak B (2001) Network dynamics and cell physiology. *Nature Reviews. Molecular Cell Biology*, **2**, 908–916.

Wang J, Gilles ED, Lengeler JW and Jahreis K (2001) Modeling of inducer exclusion and catabolite repression based on a PTS-dependent sucrose and non-PTS-dependent glycerol transport systems in *Escherichia coli* K-12 and its experimental verification. *Journal of Biotechnology*, **92**, 133–158.

Short Specialist Review

Using the Python programming language for bioinformatics

Michel F. Sanner

The Scripps Research Institute, La Jolla, CA, USA

1. Introduction

The next generation of bioinformatics programs will have to (1) support the interoperation of rapidly evolving and sometimes brittle, computational software developed in a variety of scientific fields and programming languages; (2) support the integration of data across biological experiments and scales; (3) adapt to rapidly evolving hardware environments; and (4) empower users by allowing them to carry out operations unanticipated by the programmers such as trying new combinations of algorithms or new visualizations. These combined requirements of flexibility, portability, resilience to failure, programmability by the user, and responsiveness to changes in the models and the hardware pose a challenging software engineering problem. Below, we will briefly describe the key differences between compiled and interpretive languages, rapidly compare the Python programming language with other interpretive languages, and finally provide a couple of examples of large Python-based software development project in the field of bioinformatics.

2. Programming languages

In compiled languages (FORTRAN, C, C++, etc.), an executable has to be built (i.e., compiled and linked) from the source code before it can be executed. Extending or modifying these programs requires rebuilding the application after the source code has been modified. This cannot be done while the program is running. Moreover, these languages require complete and syntactically correct source code before compilation and testing become possible. Interpretive languages (Perl (Tisdall, 2003; *see also* Article 104, **Perl in bioinformatics**, Volume 8), Python (Lutz *et al.*, 1999), Tcl (Ousterhout, 1994), Ruby (Thomas, 2001), Scheme (Dybvig, 1996), Lisp (Steele, 1990), Basic (Lien, 1986), etc.) use a program called an *interpreter* to evaluate source code as it is read from a file or typed in interactively by a user. These languages are often referred to as *scripting* or *interpreted languages*. Applications written in these languages can be extended and modified while the program is running without having to quit, recompile, or restart the application.

Table 1 Simple classification of some programming languages

	Low-level ⟶ High-level
Compiled	FORTRAN, C, C++, Java
Interpretive	Unix-Shell, awk, Basic, Perl, Tcl, Scheme, Ruby, Python

Code can be executed and tested as it is written, even if only a small fraction of the complete application has been implemented, thus allowing the rapid exploration of novel hypothesis and the early determination of design flaws.

Programming languages can also be classified and described by placing them on a scale ranging from low- to high-level languages (Table 1). We will deem a language "high-level" if it provides: high-level data structures such as lists and associative arrays; mechanisms for preventing errors from aborting the application (exceptions); array boundary checking; and support for automatic memory management as part of the language. The more such features any particular language provides, the higher it will be ranked. Interpretive languages usually are higher level than compiled languages. Java is often mistaken for an interpreted language because the Java Virtual Machine interprets the Java bytecode. However, Java has no support for the dynamic evaluation of source code and hence it is more accurately considered a high-level compiled language.

3. Advantages of interpretive languages

Rapid development: The high-level and dynamic nature of interpretive language offers several advantages for scientific computing. The ability to quickly prototype software is particularly interesting in a research environment where the software requirements are constantly shifting as the understanding of the underlying science evolves.

Easy to learn: Their simple syntax and high-level data structures make it easier for nonprofessional programmers such as computational biologists to develop programming skills, enabling them to interact with data programmatically and eventually develop code on their own. For instance, Python is becoming an increasingly popular language for teaching computational skills (Schuerer *et al*., 2004).

Open-ended application extensibility: Software environments enabling end users to interact programmatically with their data and with the application using a simple yet fully fledged programming language provides the highest level of extensibility. While Graphical User Interfaces (GUIs) are an excellent way to abstract programming syntax and data structures, GUIs can only be implemented for anticipated ways of using a program. This lack of programmability associated with GUI-based interfaces greatly limits the application's range of capabilities and its extensibility. Modifying an application's source code enables unrestricted modifications; however, it is usually a difficult task and always presents the danger of corrupting the application. General-purpose scripting languages provide a safer and easier way to extend applications, while

providing the open-ended extensibility that is missing in GUI-based interfaces. In fact, scripting languages blur the line between users and developers.

Glue languages: Interpretive language can be used to integrate heterogeneous pieces of software written in a variety of compiled languages and let them interoperate.

Interpretive languages provide new solutions to the challenges mentioned in the introduction and are becoming increasingly popular for developing bioinformatics applications (BioPerl, 2005)(BioPerl), BioPython, 2005(BioPython), Chimera (Huang *et al*., 1996), PMV (Coon *et al*., 2001), Vision (Sanner *et al*., 2002), PyMol (Delano, 2002), WhatIf (Vriend, 1990), PHENIX (Adams *et al*., 2002), MMTK (Hinsen, 1997), VMD (Humphrey *et al*., 1996)), and for teaching programming concepts to biologists (Schuerer *et al*., 2004).

4. Performance

The main drawback of interpretive languages is that they are slower than compiled languages, typically by a factor of 15–40 (Cowell-Shah, 2004). When compiled languages appeared, they rapidly replaced assembly languages. The substantial increase in readability and portability they offered outweighed the associated potential decrease in runtime performance. Today, this pattern is repeating itself: the sustained increase of computational power in typical desktop computers makes interpretive languages a practical alternative to compiled languages for a variety of tasks. Moreover, we often have a wrong intuition about how much and where performance is needed. A key observation is that the bulk of the source code of programs used in computational biology deals with input and output, handling events, and bookkeeping of data structures. The computationally intensive parts are often confined to a few functions that amount to a small percentage of the total number of lines of source code, typically less than 10%. Hence, large amounts of code can be implemented in high-level interpretive languages without affecting the overall performance of the application.

5. Which scripting language is right for my application?

Scripting languages such as the various UNIX shell scripting languages and awk (Aho *et al*., 1988) have been available for a long time. However, these languages are arcane and limited in generality. They have been reserved to computer-savvy users. A number of high-level scripting languages are available today. Perl was one of the first to be widely used in scientific computing. It is a powerful language for writing highly concise scripts. However, the way its syntax and notation tends to promote obfuscation is not a desirable feature for reusing and sharing code. Perl is mainly used for writing relatively short, “use-once”-type scripts, but is ill suited for developing large and complex software applications (Raymond, 2001). It is also lacking a good interpreter shell for interactive code development. The Tcl

language is another example of a widely used scripting language. Unfortunately, its one strength – all data are represented as strings – is also its main weakness for scientific computing. This makes Tcl cumbersome and inefficient for numerical computations. The Tcl-shell is also relatively simple and underdeveloped.

Python appeared in 1991, and was designed from the outset as an object-oriented language while allowing for simple scripting. It supports multiple inheritance, introspection, self-documenting code, high-level data structures (i.e., lists and associative arrays called *dictionaries*), and exceptions and warning mechanisms. Its syntax was designed to be free of arcane symbols and to look like English text (i.e., pseudo-code). Python is open source and runs on virtually any computer from super computers to PDAs. New functionality can be added to the interpreter by loading extensions to the language at runtime. Such extensions can be organized into sets called *packages*. The standard Python distribution comes with a comprehensive library of extensions covering needs in areas as diverse as regular expression matching, GUI toolkits, databases, network protocols, numerical calculations, and XML processing to name a few. In addition, an active community of developers provides a variety of packages spanning all areas of computing, reflecting the diversity of Python's user community and application areas. Python extensions can be written both in the Python programming language and in compiled languages such as FORTRAN, C, or C++, thus allowing the incorporation of legacy code. Code written in compiled languages must be "wrapped" (i.e., turned into a Python extension) before it can be called from a Python interpreter. Semi-automatic tools (SWIG 2005) facilitate the process of wrapping compiled code. These extensions execute faster than those written using the Python language. However, they are platform-dependent (i.e., they need to be compiled for every hardware platform and operating system). With Python, it is possible to have "the best of both worlds" by implementing those functions that require the utmost in performance using a compiled language, while the remainder of the application can be written in a high-level and platform-independent language.

Python has been used as a "glue language" to integrate monolithic programs. With the wrapping mechanism described above, a scripting layer can be added to such programs, allowing them to interoperate within a Python interpreter (Delano, 2002); (Humphrey *et al*., 1996); (Huang *et al*., 1996). Using Python for scripting applications should be contrasted with using a custom scripting language (i.e., SVL (Santavy *et al*. 2005), BCL (Daelen 2005), SPL (Tripos 2005), batchmin (Schrodinger 2005), MATLAB (Hanselman *et al*., 2001), etc.). Custom languages often lack extensibility and tend to be domain specific since they are created by developers whose strengths are in a specific application domain. Because of their specificity, these languages will not benefit from the input of a large community. Designing a language is better left to people whose primary occupation is designing languages.

While an excellent scripting and glue language, Python is also a general-purpose, object-oriented programming language. It can be used as the primary language for the implementation of complete packages and applications that have the great advantage of platform independence (i.e., the same source code runs on all platforms). Unfortunately, the often wrong intuition about where and how much performance is needed tends to deter programmers from this approach. This

misconception about performance is difficult to overcome. However, designing and implementing components in the Python programming language first, provides several advantages: (1) The development cycle of a prototype is greatly accelerated through the use of a high-level language and the reuse of software components. It is not uncommon for a Python program to have 3–10 times fewer lines of source code than the same program written in C. (2) The design of software components can be validated rapidly. Python code can be tested as it is written. The ability to run code after only a fraction of an application or component has been implemented helps identify design problems early on. (3) Performance bottlenecks can be identified using Python's profiler and resolved by optimizing the Python code, or by implementing such parts in C or C++. Finally, (4) a potentially slow but working Python implementation is always available in the case that a C or C++ implementation cannot be loaded.

6. Examples of large python-based software developments

BioPython: (**http://biopython.org**) is an example of a joint international effort to create a modular and reusable set of tools in Python. Like its sibling projects, Bioperl, Biojava, and the other so-called Bio* (say: "bio-star") projects, it is developed under an open-source programming license. Contributions are made by dozens of people worldwide. Biopython "core" contributors regularly discuss and enforce a well-defined and consistent style to create a cohesive and integrated set of modules rather than a disjoint collection of programs. Modules are submitted concurrently with documentation and tests. In addition, on-line tutorials, "cookbooks", mailing lists, and bug-reporting software helps support the user community. Biopython provides support for parsing common bioinformatics file formats such as Fasta, GenBank, UniGene, SwissProt, FSSP, SCOP, and PDB; interfaces to common on-line destinations such NBCI and Expasy, and to locally run programs such as local NCBI BLAST, or ClustalW. It also provides pairwise and multiple alignment methods, a standard sequence class with methods such as "reverse complement" or "translate" for DNA, and "charge" and "pI" for proteins, and so on.

MglTools: (**http://www.scripps.edu/∼sanner/software**) is a collection of Python extensions developed in the author's laboratory, and dealing with various common tasks in structural biology, ranging from 3D visualization to the representation and manipulation of molecular data (Coon *et al*., 2001). Eleven packages comprising over 220 000 lines of code and over 1370 classes are implemented in the Python programming language. Ten additional packages wrap C and C++ libraries and provide an additional 200 classes. These components have been combined in various applications, including: PMV (Sanner, 1999), a general-purpose 3D molecule viewer (Figure 1a); ADT, a PMV-based graphical front end to the molecular docking program AutoDock (Morris *et al*., 1998); ARTK (Gillet *et al*., 2004a,b), an augmented reality program enabling the real-time blending of computer-generated graphics with live video of physical models of molecules; and Vision (Figure 1c) (Sanner

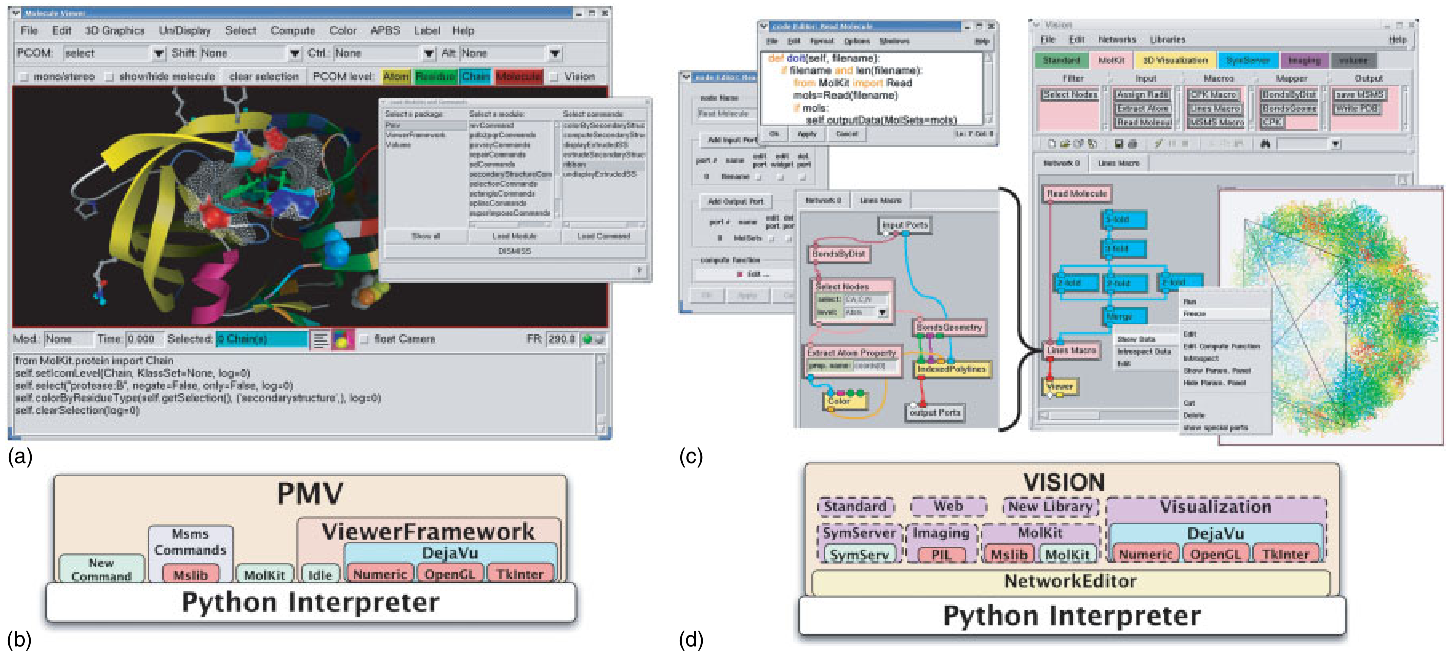


Figure 1 Two applications, PMV and Vision, built from, and sharing many software components. (a) PMV: a general purpose molecular visualization application. (b) Architectural layout of PMV. Nested boxes denote dependencies between Python packages. Packages with dark pink background are platform-dependent. (c) A molecular visualization application built using the Vision visual programming environment. A network used to display a viral capsid is shown. The subnetwork embedded in the “Lines Macro” is shown as an inset. The “node Editor” allowing inspecting and modifying nodes interactively has been started on the “Read Molecule” node. (d) Architectural layout of Vision. Libraries of Vision nodes are shown with dashed outlines. Note the number of software components shared with PMV

et al., 2002), a visual programming environment. PMV and ADT have been downloaded by over 9000 scientists (NCBR 2005).

These reusable software components and the applications built from them demonstrate the feasibility of the programming paradigm we have explored (Sanner, 1999) in which the Python interpreter is at the heart of all applications we create (see Figure 1b and d). In this approach, the software development strategy shifts from writing programs to developing software components. We found that Python's modular nature promotes the development of such components and we have experienced unprecedented levels of code reuse in the various applications mentioned above. Moreover, some of our software components have been reused by others. For instance, the ViewerFramework component underlying PMV has been reused in A. McCulloch's laboratory at UCSD to create a viewer for cardiac modeling and simulations. Similarly, Vision has been used and extended by various other laboratories for tasks as diverse as image analysis, astronomical calculations, and grid-based computing. These examples demonstrate the reusability of our software components beyond the boundaries of our laboratory, and scientific domain.

7. Conclusion

The concepts of modularity and compartmentalization of computational tasks are not new; however, time has shown that these concepts are poorly promoted by compiled languages. When programming in the latter languages, the goal is to write an application that produces the right result. There is little incentive to implement independent, and therefore reusable, software components for solving each particular subtask. In fact, even if the software is properly compartmentalized, special "main" functions have to be written to ascertain the independence and correctness of the various components. The intrinsic modular nature of Python, on the other hand, promotes the creation of self-contained components each implementing a specific functionality. These components can then be loaded into a Python interpreter, effectively extending the interpreter with new functionality. Moreover, this component naturally becomes available to any application running a Python interpreter.

As modern computational biology moves toward the study of larger systems, it will require the integration of computational methods and experimental data from a variety of scientific fields. Furthermore, the complexity of the models will require an unprecedented level of flexibility in the software tools to allow investigators to formulate and validate new hypotheses. The interactive and dynamic nature of interpretive languages, their relative simplicity, and their ease-of-use, make these languages excellent choices for creating software tools with these advanced and complex requirements. Therefore, these languages will play an increasingly significant role in the era of digital biology.

Acknowledgments

The development of the Python-based software components described in this article was supported through the NSF grant CA ACI9619020 and the NIH grant

RR08605. We would also like to thank Iddo Friedberg for helpful discussions. This is manuscript #17185-MB from the Scripps Research Institute.

References

Adams PD, Grosse-Kunstleve RW, Hung LW, Loerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK and Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 11), 1948–1954.

Aho A, Kernighan B and Weinberger P (1988). *The AWK Programming Language*, Addison Wesley.

BioPerl project home page (2005) **http://bio.perl.org/**.

BioPython project home page (2005) **http://www.biopython.org/**.

Coon S, Sanner MF and Olson AJ (2001) *Re-Usable components for Structural Bioinformatics*. 9th International Python conference (IPC9), Long Beach.

Cowell-Shah CW (2004) Nine Language Performance Round-up: Benchmarking Math & File I/O **http://www.osnews.com/story.php?news_id=5602**, Accessed year 2005.

van Daelen T (2005) The Insight Scripting Language **http://www.chem.ac.ru/Chemistry/Soft/BIOLANGU.en.html**.

Delano WL (2002) The PyMOL Molecular Graphics System. **http://www.pymol.org**, Accessed year 2005.

Dybvig RK (1996) *The Scheme Programming Language*, Second Edition, Prentice Hall PTR.

Gillet A, Goodsell D, Sanner MF, Stoffler D and Oslon AJ (2004a) *A Tangible Model Augmented Reality Application for Molecular Biology*. IEEE Visualization Vis04, Austin.

Gillet A, Goodsell D, Sanner MF, Stoffler D, Weghorst S, Winn W and Olson AJ (2004b) *Computer-Linked Autofabricated 3D Models For Teaching Structural Biology*. *SigGraph 2004*, Los Angeles, Convention Center.

Hanselman D and Littlefield BR (2001) *Mastering MATLAB 6, 1/e*, Prentice Hall.

Hinsen K (1997) *The molecular modeling toolkit: A case study of a large scientific application in Python*. *6th International Python Conference*, San Jose.

Huang CC, Couch GS, Pettersen EF and Ferrin TE (1996) *Chimera: An extensible molecular modeling application constructed using standard components*. *Pacific Symposium on Biocomputing*, The Big Island of Hawaii.

Humphrey W, Dalke A and Schulten F (1996) VMD: visual molecular dynamics. *Journal of Molecular Graphics*, **14**(1), 27–28.

Lien DA (1986) *The Basic Handbook: Encyclopedia of the BASIC Computer Language*, Compusoft Publishing.

Lutz M and Asher D (1999) *Learning Python*. O'reilly & Associates, Sebastopol, CA.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK and Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, **19**(14), 1639–1662.

NCBR Download statistics for MGLTools (PMV, ADT, Vision) (2005) **http://nbcr.sdsc.edu/nbcrstats/adttest.cgi**.

Ousterhout J (1994) *Tcl and the Tk Toolkit*, Addison Wesley.

Raymond ES (2001) *Why Python? Linux Journal*.

Sanner MF (1999) Python: a programming language for software integration and development. *Journal of Molecular Graphics & Modelling*, **17**(1), 57–61.

Sanner MF, Stoffler D and Olson AJ (2002) ViPEr, a visual programming environment for Python. *Proceedings 10th International Python Conference, Alexandria*.

Santavy M and Labute P (2005) SVL: The Scientific Vector Language **http://www.chemcomp.com/Journal_of_CCG/Features/svl.htm**.

Schrodinger MacroModel (2005) **http://www.schrodinger.com/**.

Schuerer K and Letondal C (2004) Python course in Bioinformatics **http://www.pasteur.fr/recherche/unites/sis/formation/python/**, Accessed year 2005.

Steele GL (1990) *Common Lisp the Language*, Second Edition, Digital Press.

SWIG Simple Wrapper Interface Generator (SWIG) (2005) **http://www.swig.org/**.

Thomas D and Hunt A (2001) *Programming Ruby - The Pragmatic Programmer's Guide*, Addison Wesley Longman, Inc.

Tisdall J (2003) *Mastering Perl for Bioinformatics*, O'reilly.

Tripos SYBYL (2005) **http://www.tripos.com**.

Vriend G (1990) WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics*, **8**, 52–56.

Short Specialist Review

Perl in bioinformatics

R. Hannes Niedner and T. Murlidharan Nair
University of California, San Diego, CA, USA

Michael Gribskov
Purdue University, West Lafayette, IN, USA

1. Introduction

Advances in biology have generated an ocean of data that needs careful computational analysis to extract useful information. Bioinformatics uses computation and mathematics to interpret biological data, such as DNA sequence, protein sequence, or three-dimensional structures, and depends on the ability to integrate data of various types. The Perl language is well suited to this purpose.

An important practical skill in bioinformatics is the ability to quickly develop scripts (short programs) for scanning or transforming large amounts of data. Perl is an excellent language for scripting, because of its compact syntax, broad array of functions, and data orientation. The following are a few of the attributes of Perl that make it an attractive choice.

- Perl provides powerful ways to match and manipulate strings through the use of regular expressions. Changing file formats from one to another is a matter of contorting strings as required.
- Modularity that makes it easy to write programs as libraries (called modules).
- Perl system calls and pipes can be used to incorporate external programs.
- Dynamic loaders of Perl that help extend Perl with programs written in C as well as create compiled libraries that can be interpreted by the Perl interpreter.
- Perl is a good prototyping language and is easy to code. New algorithms can be easily tested in Perl before using a rigorous language.
- Perl is excellent for writing CGI scripts to interface with the Web.
- Perl provides support for object-oriented program development.

Projects such as the sequencing of the human genome produce vast amounts of textual data. In its early stages, the human genome project faced issues with data interchange between groups that were developing software, and Lincoln Stein has noted how Perl came to the rescue (Stein, 1996). Perl is certainly not the only

language that possesses these positive features – many of them are found, as well, in other scripting languages such as Python (*see* Article 103, **Using the Python programming language for bioinformatics**, Volume 8).

In the following section, we discuss some of the important Perl features that make it attractive to bioinformaticians. We refer in this article to Perl version 5.6–5.8, but encourage you to look into the exciting new features of Perl 6, which has been reengineered from the ground up (**http://dev.perl.org/perl6/**, 2005).

2. Why use Perl?

Looking at today's landscape of programming languages, one is offered a wide variety from which to choose. The difference between compiled and interpreted languages is explained sufficiently in other articles in this series (Sanner, 2004), thus it suffices here to say that compiled languages are usually selected for their superior performance, as reflected by the C/C++ implementation of computationally demanding algorithms such as BLAST, ClustalW, or HMMer, or for their specific capabilities such as comprehensive user interfaces or graphical libraries. Scripting languages (often used synonymously for interpreted languages) such as Perl, Python, PHP, Ruby, Tcl, and so on, on the other hand, are often favored for fast prototyping since the development cycle does not require code recompilation. With the ever-increasing computational power (Moore's Law) (Moore, 1965) available for bioinformatics tasks, performance issues lose their importance in contrast to ease of development and code maintainability. In the search for the right programming language, one might also encounter the distinction of object-oriented and procedural programming accompanied with evaluations of which languages to choose for each of these programming approaches. Again, detailed publications exist on this topic (Valdes, 1994; Kindler, 1988; **http://www.archwing.com/technet/technet_OO.html**, 2001; **http://search.cpan.org/~nwclark/perl-5.8.6/pod/perltoot.pod**, 2005); thus, we restrict ourselves to some brief explanations here.

Until recently, the most popular programming languages (e.g., FORTRAN, BASIC, and C) were procedural – that is, they focused on the ordered series of steps needed to produce a result. Object-oriented programming (OOP) differs by combining data and code into indivisible components, called objects, which are abstract representations of "real-life" items. These objects model all information and functionality as required by the application. OOP languages include features such as "class", "instance", "inheritance", and "polymorphism" that increase the power and flexibility of an object.

In bioinformatics, a DNA sequence might be represented as an object inheriting from a more general implementation that covers the properties of all biological sequences. OOP would code this object by describing its properties such as length, checksum, and certainly the string of letters that comprises the sequence itself. Then one would implement accessor methods to retrieve or set these properties, and also more complex functions such as transcribe() that would take as an argument an organism-specific codon matrix and transform the DNA object to an RNA object. In procedural (or sequential) programming, the code flow is driven by the "natural" sequence of events. OOP is generally regarded as

producing more reusable software libraries and better APIs (application programming interfaces) (**http://www.archwing.com/technet/technet_OO.html**; **http://search.cpan.org/~nwclark/perl-5.8.6/pod/perltoot.pod**, 2005).

While all these aspects certainly matter in one's choice of a language, our experience is that bioinformatics programming frequently is not done by computer scientists or trained programmers, but rather by researchers in bioscience who are trying to make sense out of their experiments. The priority is not the production of elegant and maintainable code but rather speed and ease of programming. As Larry Wall puts it, " ... you can get your job done with any 'complete' computer language, theoretically speaking. But we know from experience that computer languages differ not so much in what they make possible, but in what they make easy." In that regard, "Perl is designed to make the easy jobs easy, without making the hard jobs impossible" (Wall *et al*., 2000).

The high-level data structures built into Perl scale to any size and are merely restricted by the boundaries of the operating system and the amount of memory available on the host machine. Perl supports both procedural and object-oriented programming approaches and runs on virtually on all flavors of Unix and Linux. In addition, ActivePerl from ActiveState (**http://www.activestate.com/Products/ActivePerl/**, 2005) allows Perl to run even on Windows computers, and MacPerl (**http://dev.macperl.org/**, 2005) on Apple computers running System 9 and lower.

3. But Perl code is cryptic?!

At its simplest, Perl requires no declaration of variables. Variables belong to one of three classes: scalars (the scalar class supports strings, integers and floating point numbers), arrays, and associative arrays (arrays indexed by strings). Each type of variable is implicitly identified by beginning with a specific character: "$" for scalars, "@" for arrays, and "%" for associative arrays. Storage for variables is created when they are first referenced and released automatically when they pass out of scope. This automatic garbage collection is one of the high-level language features that makes Perl very convenient. Strings, integers, and floating point numbers are automatically identified and converted from one type to another as required. This is another powerful simplifying feature of Perl.

As with any high-level language, one can write very cryptic code in Perl – the distinction is that Perl has frequently been criticized for this. While the lack of a requirement for declaring variables exacerbates this problem and the automatic creation and initiation of variables can be a headache when typographical errors occur, these Perl-specific criticisms are, in our view, largely a red herring. Many style guides (e.g., perlstyle, (**http://search.cpan.org/~krishpl/pod2texi-0.1/perlstyle.pod**, 2005)) point out some common practices, not necessarily specific to Perl, that lead to "cryptic" code:

- usage of meaningless variable and function names (e.g., x1);
- removing extra (but structuring) white spaces and parentheses;
- writing more than one instruction (ended by a semicolon) on one line; or

- active attempts to obfuscate your code such as substitution of numeric values with the arithmetic expressions or usage of hexadecimal character codes for all characters in strings to name just a few.

Perl provides support for better programming style (for instance, the "strict" pragma that requires explicit declaration of variables) and for internal documentation (via the pod mechanism). Perl also supports high-level constructs such as subroutines (with support for both call-by-reference and call-by-value), libraries (called modules in Perl), and classes.

Perl is not strict about the completeness of the data. Missing or odd characters are not penalized. This can be thought as a downside of Perl, since Perl will perform operations on mixed types without complaining. This can come back to haunt you when spelling mistakes lead to logical errors, but can be easily circumvented by using the strict pragma and taking advantage of the many debugging features built into Perl, such as the using the -w switch to increase the verbosity of warnings and messages (**http://search.cpan.org/~nwclark/perl-5.8.6/pod/perlmodlib.pod#Pragmatic_Modules**, 2005; Gutschmidt, 2005).

4. Numerical calculations

With all the characteristics of a loosely typed language discussed so far, one might conclude that Perl cannot be used for numerical computations, but this is certainly not the case. Perl performs numerical calculations as double-precision floating point operations, and thus can be used for most computations that do not involve very high numerical precision. The Math::BigInt (**http://search.cpan.org/~tels/Math-BigInt-1.77/lib/Math/BigInt.pm**, 2005) and Math::BigFloat (**http://search.cpan.org/~tels/Math-BigInt-1.77/lib/Math/BigFloat.pm**, 2005) provide high-precision arithmetic for integers and floating-point numbers at the cost of somewhat slower execution in case you require more precision.

One can thus comfortably implement the classical algorithms such as learning by back-propagation of errors or Cleveland's locally weighted regression (LOWESS) in Perl. These are very frequently used algorithms in Bioinformatics for pattern recognition and microarray data normalization respectively. Having said this, however, Perl is nearly always a poor choice for large "number crunching". This is because Perl represents floating point numbers internally as strings and every numerical operation requires a conversion of string to floating point and back again. This can make Perl programs very slow if extensive calculations are required.

5. Regular expressions – Perl's most famous strong point

Pattern matching in text is simple and straightforward in Perl. Our experience is that code that can extend over a half a page in C or C++ can be easily written in one or two lines of Perl. Frequently, bioinformatics programs screen text documents for tokens that are identified by complex patterns. Regular expressions ("regex's" for short) are sets of symbols and syntactic elements used to match patterns of text.

They make it easy to represent complex patterns. For instance, searching for the pattern "ATT(C-or-G)(C-or-G)AAT" within 500 bases upstream of a promoter can be easily accomplished using this regular expression:

```
ATT[CG][CG]AAT.{0,500}TATAA
```

(the promoter is characterized by the Goldstein-Hogness or TATA box (Sudhof *et al*., 1987)).

Perl's built-in support for regular expressions is second to no other language, and there are three typical operations done with regular expressions:

```
Match:   m/PATTERN/;
Replace: s/PATTERN/REPLACEMENT/;
Split:   @tokenlist=split /PATTERN/, $string;
```

Additional facilities allow the pattern matching to be very flexible, for instance to be case-insensitive, to apply to every occurrence (global matching), or to translate sets of characters (useful for changing case or converting DNA Ts to RNA Us). It is simple to transform genetic sequences formatted with spaces and numbers into plain strings of letters as often required for further analyses:

$$\$\text{sequence} = \text{s/[\^{}ACGT]//g} \tag{1}$$

The above expression replaces every character that is NOT a capital A, C, G, or T with an empty string (well nothing), thus stripping all nonsequence characters (including carriage returns), lower case letters, and numbers from the sequence string. It is outside the scope of this article to provide a tutorial on regular expression, but several on-line and printed publications on this topic are available (Friedl, 2002; **http://www.regular-expressions.info/**, 2005).

Regular expressions also support bioinformatics resources such as, for example, Prosite (Sigrist *et al*., 2002) that provide access to organized collections of sequence motifs expressed as patterns.

6. File handling and databases

"Bioinformation", that is, data produced within the biosciences, is often produced as flat files (tag-field formatted text files). File formats developed by the Protein Data Bank (PDB) (Berman *et al*., 2002), Swissprot (Bairoch and Boeckmann, 1991), and Genbank (Benson *et al*., 2003) have become quasi-standards within the bioinformatics community. Perl makes it very easy to traverse large file systems, reading from and writing to files, and using its rich regular expression syntax to parse and transform textual file content. An example would be extracting the sequence string out of hundreds of individual files in Genbank format, while leaving out the information describing taxonomy, sequence features, bibliographic references, and other annotations. With these features, it is easy to understand why Perl naturally became the number one choice for the bioinformatics pioneers.

With the advent of relational databases, many bioinformatics repositories have been undergoing major reengineering efforts to reorganize their data storage technology from file repositories to relational databases. In addition, many labs and companies use relational database management systems (RDBMS) to store and organize annotation and experimental data. In Perl the Database Interface module (DBI) (**http://dbi.perl.org/**, 2005) provides a standardized and common interface for writing scripts that reference RDBMSs. Specific database drivers (DBDs) are called by DBI to communicate with particular RDBMS. The following code sample establishes a database connection ($dbh), which can be used to execute SQL commands:

```
use DBI;
my $dsn = 'DBI:mysql:my_database:localhost';
my $db_user_name = 'admin';
my $db_password = 'secret';
my $dbh = DBI->connect($dsn, $db_user_name,$db_password);
```

This modular architecture allows programmers to clearly separate common and RDBMS-specific code. While facilitating the production of database-independent code, DBI allows one to take full advantage of the features of a particular RDBMS via the specific functionality implemented in the DBD, or by handing off vendor-specific SQL statements directly to the RDBMS.

7. CPAN – the on-line Perl code library

Perl has been widely used not only by bioinformaticians but also by programmers in many other areas. Consequentially, there is an enormous wealth of Perl code available in the form of reusable software libraries, also called Perl modules (e.g., the DBI module mentioned above). The open repository for these Perl modules is the Comprehensive Perl Archive Network or CPAN for short. CPAN can be accessed not only at **http://www.cpan.org** (or many mirror sites) but also from within the standard Perl installation using the CPAN.pm module (**http://search.cpan.org/~andk/CPAN-1.76/lib/CPAN.pm**, 2005). When using Perl, you are part of a huge community – take advantage of it. Check CPAN first before you start implementing any major programming task since there is a very good chance that at least the core functionality is already available there.

8. Perl and the World Wide Web

Perl is widely used to develop web applications, and was adopted as the de facto language for creating content on the World Wide Web. Perl's powerful text manipulation facilities have made it an obvious choice for writing Common Gateway Interface (CGI) scripts. CGI is a standard for external gateway programs to interface with information servers such as http (or web) servers. An HTML document

that the web server retrieves is static, a CGI script, on the other hand, is executed in real time upon user request (via the web client) and helps to dynamically create a webpage. The standard Perl installation comes with the CGI module (CGI.pm), which can be used to streamline the creation of web pages. CGI.pm provides functionality to handle the GET and POST protocols, to receive and parse CGI-queries, to create Web fill-out forms on the fly, and to parse their contents and to generate HTML elements (e.g., lists and buttons) as a series of Perl functions, thus sparing you from the need to incorporate HTML syntax into your code (**http://stein.cshl.org/WWW/software/CGI/**, 2005).

The results of data analyses often require display in the form of charts and graphics. The GD module provides basic tools necessary for this. GD.pm is a Perl interface to Thomas Boutell's C-based gd graphics library that enables Perl programs to create color drawings using a large number of graphics primitives, and to format the images in PNG (**http://stein.cshl.org/WWW/software/GD/**, 2005) and other formats.

Perl also has a set of ready-to-use libraries to implement conventional http-based as well as the more recently developed RSS-, RPC-, and SOAP-based web services. This is important when one implements fully automated data retrieval from online services. For example, SOAP::Lite for Perl is a collection of Perl modules that provides a simple and lightweight interface to the Simple Object Access Protocol (SOAP) both on client and server side (**http://www.soaplite.com/**, 2005).

The LWP (Library of WWW access in Perl) modules provide the core of functionality for web programming in Perl. It contains the foundations for networking applications, protocol implementations, media type definitions, and debugging ability. Most notably, with LWP::UserAgent you can build "robots" to access remote websites as part of a program or even build your own robust web client (**http://search.cpan.org/~gaas/libwww-perl-5.803/lib/LWP.pm**, 2005).

9. XML processing

Perl's strong text parsing abilities make it no wonder that a host of modules have been developed to apply the power of Perl (and especially its regular expression syntax) to XML. The available modules cover the full range of XML standards, such as SAX and DOM, and are implemented either in Perl or provide a Perl interface to C-based libraries such as xerces, expat, or libxml/libxslt (**http://perl-xml.sourceforge.net/**, 2005).

10. BioPerl

BioPerl, the first in the series of Bio* projects, is an international association of developers of open source Perl tools for bioinformatics, genomics, and life science research formed in the early nineties and officially organized in 1995. BioPerl is a coordinated effort to convert and collect computational methods routinely used in bioinformatics and life science research into a set of standard CPAN-style, well-documented, and freely available Perl modules (**http://bioperl.org/**, 2005). BioPerl

provides you with a rather complete set of sequence analysis functions and more, thus the following is by no means a comprehensive listing:

- accessing the major biological databases for data retrieval;
- reading and converting all major sequence file formats;
- extracting sequence, parameters, and annotation from flat files (source data as well as program outputs);
- providing an interface to ClustalW, HMMer, BLAST, FastA, and other standard bioinformatics applications;
- parsing/representing protein structure (PDB) data;
- traversal of phylogenetic trees.

Usually, there is a steep learning curve before a novice programmer gets around to actually using BioPerl. This is because it is a complex collection of modules, it is not trivial to install, and requires an understanding of object-oriented programming. It is worth the effort required to work through the comprehensive documentation and examples provided on the World Wide Web since BioPerl can tremendously accelerate the development of complex and feature-rich applications after one masters the initial hurdles. In the event that you decide to go with your own implementation, keep in mind that you do not have to install BioPerl to use it – the BioPerl developers recommend "you just steal the routines in there if you find any of them useful" (**http://bio.cc/Bioperl/index_bioperl_original.html**, 2005).

11. Where to learn more

A tutorial on using Perl in Bioinformatics is presented in Article 112, **A brief Perl tutorial for bioinformatics**, Volume 8. Additional sources include:

Books:

- "Beginning Perl for Bioinformatics" by James Tisdall and published by O'Reilly and Associates Inc.
- "Sequence Analysis in a Nutshell" by Darryl León, Scott Markel and published by O'Reilly and Associates Inc.
- "Learning Perl" by Randal L. Schwartz, Tom Christiansen and published by O'Reilly and Associates Inc.
- "Programming Perl" by Larry Wall, Tom Christiansen and Randal L. Schwartz and published by O'Reilly and Associates Inc. (the Camel Book)
- "Perl Cookbook" by Tom Christiansen and Nathan Torkinton and published by O'Reilly and Associates Inc.
- "Mastering Algorithms with Perl" Jon Orwant, Jarkko Hietaniemi & John Macdonald and published by O'Reilly and Associates Inc.
- "Data Munging with Perl" by David Cross and published by Manning.
- "Object Oriented Perl" by Damian Conway and published by Manning.
- "Mastering Regular Expressions" by Jeffrey E. F. Friedl and published by O'Reilly and Associates Inc.

Websites:

- **http://www.cpan.org/** – the main CPAN site
- **http://www.bioperl.org/** – the BioPerl home
- **http://www.tpj.com/** – the Perl Journal
- **http://www.perl.com/** – the O'Reilly Perl site
- **http://bio.oreilly.com/** – the O'Reilly Bioinformatics site

12. Conclusion

Bioinformatics is surging ahead at an increasing pace. The demand for the development of tools needed to analyze biological data is ever increasing. Developing scripts in Perl for prototyping, or suites of reusable Perl modules, allows one to contribute to a large research community developing and sharing Perl programs. As with any other programming language, fluency in the language comes with experience; clarity of the code largely depends on the programmer and his or her use of good programming practices. With the wealth of code already freely available, Perl is and will remain a language of significant impact in computational biology for many years to come.

Perl is a "feel-good" language that does not impose a particular style on you but rather serves your way of doing things. Perl therefore promotes the three virtues of a programmer as expressed in the editorial of the Camel book (Wall *et al*., 2000): Laziness, Impatience, and Hubris (as explained at **http://c2.com/cgi/wiki?LazinessImpatienceHubris**).

References

ActiveState (2005) *ActivePerl – The industry-standard Perl distribution for Linux, Solaris and Windows* @ **http://www.activestate.com/Products/ActivePerl/**.

archwing.com (2001) *Object-Oriented Programming Overview* @ **http://www.archwing.com/technet/technet_OO.html**.

Bairoch A and Boeckmann B (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*, **19**(Suppl), 2247–2249.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (2003) GenBank. *Nucleic Acids Research*, **31**(1), 23–27.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, *et al*. (2002) The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 6 1), 899–907.

bio.cc (2005) *Bio Perl* @ **http://bio.cc/Bioperl/index_bioperl_original.html**.

bioperl.org (2005) *bioperl.org* @ **http://bioperl.org/**.

cpan.org (2005) *CPAN – Comprehensive Perl Archive Network* @ **http://cpan.org/**.

dbi.perl.org (2005) *DBI – a database interface module for Perl* @ **http://dbi.perl.org/**.

Friedl JEF (2002) *Mastering Regular Expressions*, *Second Edition*, O'Reilly.

Gutschmidt T (2005) *Perl: Strict, Warnings, and Taint* @ **http://www.developer.com/lang/perl/article.php/1478301**.

Kindler E (1988) Object oriented programming and general principles of modelling complex biological systems. *Acta Universitatis Carolinae. Medica*, **34**(3–4), 123–147.

macperl.org (2005) *MacPerl Development* @ **http://dev.macperl.org/**.

Moore GE (1965) Cramming more components onto integrated circuits. *Electronics*, **38**(8), 114–117.
perl.org (2005) *Perl6* @ **http://dev.perl.org/perl6/**.
perldoc.com (2005) *CPAN.pm* @ **http://search.cpan.org/~andk/CPAN-1.76/lib/CPAN.pm**.
perldoc.com (2005) *LWP – Library for WWW access in Perl* @ **http://search.cpan.org/~gaas/libwww-perl-5.803/lib/LWP.pm**.
perldoc.com (2005) *Math::BigFloat – Arbitrary length float math package* @ **http://search.cpan.org/~tels/Math-BigInt-1.77/lib/Math/BigFloat.pm**.
perldoc.com (2005) *Math::BigInt – Arbitrary size integer math package* @ **http://search.cpan.org/~tels/Math-BigInt-1.77/lib/Math/BigInt.pm**.
perldoc.com (2005) *perlstyle – Perl style guide* @ **http://search.cpan.org/~krishpl/pod2texi-0.1/perlstyle.pod**.
perldoc.com (2005) *perltoot – Tom's object-oriented tutorial for perl* @ **http://search.cpan.org/~nwclark/perl-5.8.6/pod/perltoot.pod**.
perldoc.com (2005) *strict – Perl pragma to restrict unsafe constructs* @ **http://search.cpan.org/~nwclark/perl-5.8.6/pod/perlmodlib.pod#Pragmatic_Modules**.
regular-expressions.info (2005) *Regular Expressions* @ **http://www.regular-expressions.info/**.
Sanner MF (2004) *Using the Python Programming Language for Bioinformatics*, pp. 5–9.
Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A and Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatic*, **3**(3), 265–274.
soaplite.com (2005) *SOAP::Lite for Perl* @ **http://www.soaplite.com/**.
sourceforge.net (2005) *Perl XML Project Home Page* @ **http://perl-xml.sourceforge.net/**.
Stein L (1996) How Perl saved the human genome project. *The Perl Journal*, **1**(2).
stein.cshl.org (2005) *CGI.pm – a Perl5 CGI Library* @ **http://stein.cshl.org/WWW/software/CGI/**.
stein.cshl.org (2005) *GD.pm – Interface to Gd Graphics Library* @ **http://stein.cshl.org/WWW/software/GD/**.
Sudhof TC, Van der Westhuyzen DR, Goldstein JL, Brown MS and Russell DW (1987) Three direct repeats and a TATA-like sequence are required for regulated expression of the human low density lipoprotein receptor gene. *The Journal of Biological Chemistry*, **262**(22), 10773–10779.
Valdes IH (1994) Advantages of object-oriented programming. *M.D. Computing*, **11**(5), 282–283.
Wall L, Christiansen T and Orwan J (2000) *Programming Perl*, *Third Edition*, O'Reilly.

Short Specialist Review

The MATLAB bioinformatics toolbox

Robert Henson and Lucio Cetto
The MathWorks, Inc., Natick, MA, USA

1. MATLAB overview

MATLAB® (The MathWorks, Inc.) is a general-purpose technical computing language (*see* Article 103, **Using the Python programming language for bioinformatics**, Volume 8) and development environment that is widely used in scientific and engineering applications. MATLAB is used in many aspects of industrial and academic bioinformatics, including base calling algorithms for DNA sequencing, image analysis of microarrays, signal processing and classification of protein mass spectra, and pathway inference from gene expression results.

MATLAB includes many mathematical, statistical, and engineering functions as well as graphics and visualization tools. Toolboxes are collections of algorithms and functions that provide application-specific numerical, analysis, and graphical capabilities.

The Bioinformatics Toolbox provides access from within the MATLAB environment to genomic and proteomic data formats, analysis techniques, and specialized visualizations. It is designed for implementing genomic and proteomic sequence and microarray analysis techniques. Most functions in the toolbox are implemented in the open MATLAB language, enabling you to explore and customize the algorithms.

2. Bioinformatics toolbox features

Functions in the Bioinformatics Toolbox enable you to access many standard file formats for biological data, Web-based databases, and other on-line data sources. Supported file formats include FASTA, PDB, and SCF, and commonly used microarray data formats, such as Affymetrix®, GenePix®, and Imagene®. You can also directly interface with major Web-based databases, such as GenBank, EMBL, PIR, and PDB.

Once you have imported the sequences into the MATLAB environment, you can easily manipulate and analyze your sequences to gain a deeper understanding of your data. The toolbox provides routines for standard operations, such as converting DNA or RNA sequences to amino acid sequences using the genetic code.

You can report statistics about the sequences and search for specific patterns within a sequence. You can further manipulate your results by applying restriction enzymes and proteases to perform in silico digestion of sequences or create random sequences for test cases.

The toolbox also provides implementations in the MATLAB language of standard pairwise and multiple sequence alignment algorithms, including the Needleman–Wunsch (Needleman and Wunsch, 1970), Smith–Waterman (Smith and Waterman, 1981), and profile hidden Markov models (*see* Article 98, **Hidden Markov models and neural networks**, Volume 8) (Durbin *et al*., 1998). Once you have aligned your data, you can visualize your sequence alignments. You can also perform phylogenetic analysis and use the graphical user interface to explore phylogenetic trees. An example of using the toolbox for phylogenetic analysis is given below.

For work with amino acid sequences, the toolbox provides several protein-analysis methods, as well as routines to calculate properties of peptide sequences, such as isoelectric point and molecular weight. A graphical user interface lets you visually study protein properties such as hydrophobicity and create standard plots such as the Ramachandran plot of PDB data.

The Bioinformatics Toolbox also provides functions for filtering and normalizing microarray data, including lowess, global mean, and median absolute deviation (MAD) normalization. Specialized routines for visualizing microarray data include box plots, log–log, I-R plots, and spatial heat maps of the microarray. Using routines from the Statistics Toolbox, you can perform hierarchical and K-means clustering or use other statistical methods to classify your results (*see* Article 50, **Integrating statistical approaches in experimental design and data analysis**, Volume 7).

3. Phylogenetic analysis example

This example shows how the toolbox functions can be used to construct phylogenetic trees from Human Immunodeficiency Virus (HIV) and Simian Immunodeficiency Virus (SIV) sequence data.

Mutations accumulate in the genomes of pathogens, in this case the human/simian immunodeficiency virus, during the spread of an infection. This information can be used to study the history of transmission events and also as evidence for the origins of the different viral strains. There are two characterized strains of human AIDS viruses: type 1 (HIV-1) and type 2 (HIV-2). Both strains represent cross-species infections. The primate reservoir of HIV-2 has been clearly identified as the sooty mangabey (*Cercocebus atys*). The origin of HIV-1 is believed to be the common chimpanzee (*Pan troglodytes*) (Gao *et al*., 1999).

In this example, the variations in three coding regions from the complete genomes of 16 different isolated strains of the human and simian immunodeficiency viruses are used to construct a phylogenetic tree. These regions were chosen because they are relatively long and contain well-conserved domains (Alizon *et al*., 1986) (Rambaut *et al*., 2004). The sequences for these virus strains can be retrieved from GenBank, using their accession numbers. The three coding regions of interest, the

gag protein, the pol polyprotein, the least stable of the chosen regions, and the envelope polyprotein precursor, can then be extracted from the sequences.

In order to access the data from GenBank, we first create an array containing the information about the sequences. This includes a brief description, the accession number, and the indices of the coding sequences (CDS) in the genomes corresponding to the regions of interest. Lines starting with a % sign are comments in the MATLAB language.

```
%          Description                        Accession   CDS:gag/pol/env
data =  { 'HIV-1 (Zaire)'                    'K03454'     [1  2  8]    ;
          'HIV1-NDK (Zaire)'                 'M27323'     [1  2  8]    ;
          'HIV-2 (Senegal)'                  'M15390'     [1  2  8]    ;
          'HIV2-MCR35 (Portugal)'            'M31113'     [1  2  8]    ;
          'HIV-2UC1 (IvoryCoast)'            'L07625'     [1  2  8]    ;
          'SIVMM251 Macaque'                 'M19499'     [1  2  8]    ;
          'SIVAGM677A Green monkey'          'M58410'     [1  2  7]    ;
          'SIVlhoest L'Hoest monkeys'        'AF075269'   [1  2  7]    ;
          'SIVcpz Chimpanzees Cameroon'      'AF115393'   [1  2  8]    ;
          'SIVmnd5440 Mandrillus sphinx'     'AY159322'   [1  2  8]    ;
          'SIVAGM3 Green monkeys'            'M30931'     [1  2  7]    ;
          'SIVMM239 Simian macaque'          'M33262'     [1  2  8]    ;
          'CIVcpzUS Chimpanzee'              'AF103818'   [1  2  8]    ;
          'SIVmon Cercopithecus Monkeys'     'AY340701'   [1  2  8]    ;
          'SIVcpzTAN1 Chimpanzee'            'AF447763'   [1  2  8]    ;
          'SIVsmSL92b Sooty Mangabey'        'AF334679'   [1  2  8]    ;
          };
names = data(: , 1); acc = data(: , 2); cds = data(: , 3);
```

We can now retrieve the sequence information from the NCBI GenBank database, using the *getgenbank* function.

```
% Store the number of viruses
numViruses = size(data,1);
% Loop over each entry and download data from GenBank
for i = 1:numViruses
   seqs_hiv(i) = getgenbank(acc{i});
end
```

The next step is to extract the CDS for the GAG, POL, and ENV coding regions from the sequences. Then, extract the nucleotide sequences using the CDS pointers. The code below shows how to do this for the first sequence.

```
% Extract the sequence and CDS for the first sequence
theSequence = seqs_hiv(i).Sequence;
CDSs = seqs_hiv(i).CDS(cds{1},:);

% Extract the coding regions
gag{i} = theSequence(CDSs(1,1):CDSs(1,2));
pol{i} = theSequence(CDSs(2,1):CDSs(2,2));
env{i} = theSequence(CDSs(3,1):CDSs(3,2));
```

The Bioinformatics Toolbox can generate phylogenetic trees from either nucleotide or amino acid sequences. In this example, we will use the amino acid sequence of the coding regions to construct the trees. The translated sequence information is stored in the GenBank data structure, so we can extract the amino acid sequences from there or we can use the toolbox to calculate the sequences.

```
% Use nt2aa to convert nucleotide sequences to amino acid sequences
aagag{i} = nt2aa(gag{i});
aapol{i} = nt2aa(pol{i});
aaenv{i} = nt2aa(env{i});
```

To get some idea of how closely related the sequences are, we can use the global alignment functions from the toolbox to align some of the sequences.

```
% Align the two HIV-1 sequences using the default options
% score is in bits
hiv1_hiv1NDK_score = nwalign(aagag{1},aagag{2})
hiv1_hiv1NDK_score =
  1060.67
```

Figure 1 shows the global alignment of the two HIV-1 GAG proteins.

```
% Align the HIV-1 and HIV-2 sequences - score is in bits
hiv1_hiv2_score = nwalign(aagag{1},aagag{3})
hiv1_hiv2_score =
  570.00
```

Figure 2 shows the global alignment of the HIV-1 (Zaire) and HIV-2 Senegal GAG proteins.

The first two sequences are clearly very closely related. The HIV-1 and HIV-2 sequences have regions of similarity but also large regions that are significantly different.

In this example, we will use a distance-based method to generate the phylogenetic tree. There are two steps to creating distance-based phylogenetic trees (Page and Holmes, 1999). We first calculate all pairwise distances between the sequences and then construct the hierarchy from these distances. We can use many different metrics to measure the distance between two sequences, based on statistical and evolutionary models of how mutations occur. The Bioinformatics Toolbox provides many of the standard metrics including the Jukes–Cantor (Jukes and Cantor, 1969), Kimura (Kimura, 1980), and Tajima–Nei (Tajima and Nei, 1984) methods. You can also create a custom function to add your own metric to the toolbox. In this example, we will use the Jukes–Cantor distance, which assumes an equal rate of mutation for all nucleotides or amino acids. This approach clearly has some limitations and does not take into account the possibility of commonly occurring events in retroviruses such as crossover or recombination (Salemi *et al*., 2003).

```
gagd = seqpdist(aagag,'method','Jukes-Cantor');
```



Figure 1 Global pairwise alignment of GAG proteins from HIV-1 (Zaire) and HIV-1 NDK (Zaire)

This function creates a matrix of the pairwise distances between the 16 sequences. We can now use this information to build a hierarchical tree. There are several methods used for building trees from pairwise distances. The most commonly used is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method. You can also choose from other linkage methods including single, complete, and weighted PGMA (Durbin *et al*., 1998).

```
gagtree = seqlinkage(gagd,'UPGMA',names)
```



Figure 2 Global pairwise alignment of GAG proteins from HIV-1 (Zaire) and HIV-2 (Senegal)

We can now plot the tree (Figure 3) and add a title.

```
plot(gagtree,'type','cladogram');
title('Immunodeficiency virus (GAG protein)')
```

If we repeat this process with the ENV (Figure 4) and POL (Figure 5) proteins, we see similar but not identical trees.

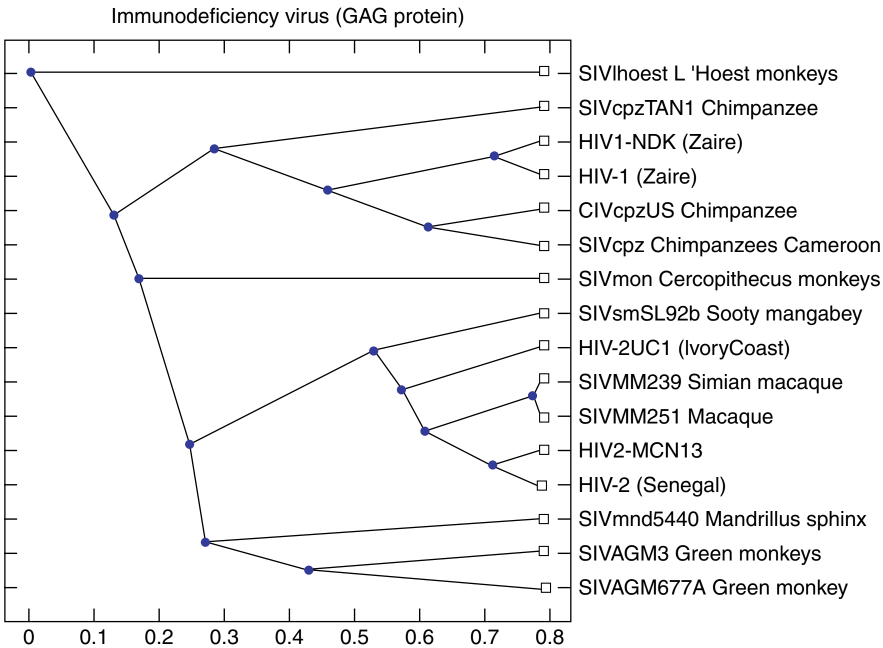


Figure 3 Phylogenetic Tree for GAG protein

The trees are slightly different, although they do show some clear trends. For example, the HIV-1 sequences always cluster together with the chimpanzee sequences and the HIV-2 sequences cluster with the SIV sequences from the sooty mangabey, as expected. One way to average out the three trees is to create a weighted consensus tree (Figure 6).

```
% calculate weights
weights = [sum(gagd) sum(pold) sum(envd)];
weights = weights / sum(weights);
% weighted average
dist = gagd .* weights(1) + pold .* weights(2) + envd .* weights(3);
% construct tree
tree_hiv = seqlinkage(dist,‘UPGMA’,names);
plot(tree_hiv,‘type’,‘cladogram’);
title(‘Immunodeficiency virus (Weighted Consensus Tree)’)
```

Using the same sequences to generate a maximum parsimony tree in PHYLIP (Felsenstein, 1989) produces a similar tree with the same distinct Chimpanzee/HIV-1 and Sooty mangabey/HIV-2 clusters.

4. Conclusion

Bioinformaticists have traditionally had to invest a great deal of time programming math and statistics algorithms in a short time frame. MATLAB and the Bio-

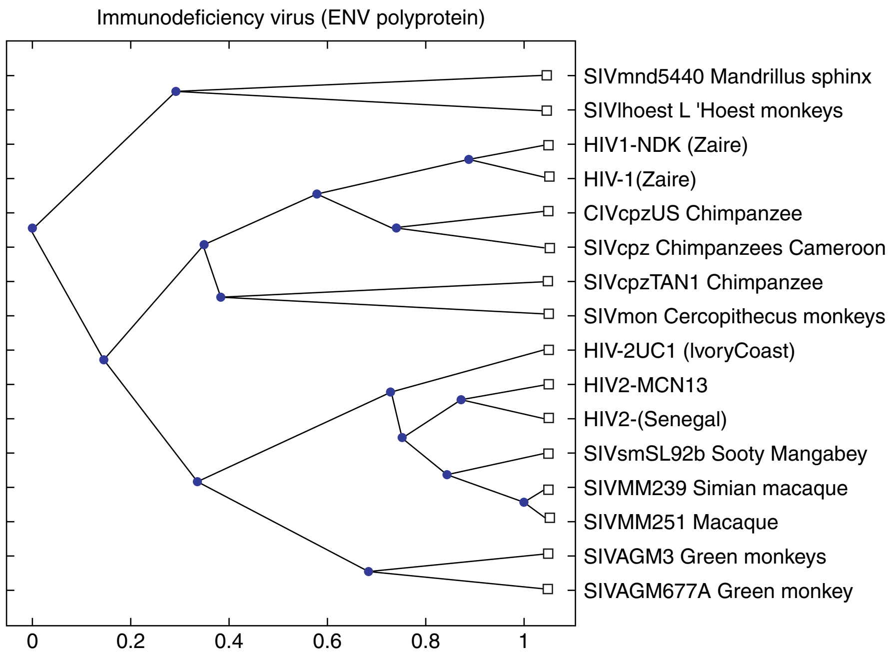


Figure 4 Phylogenetic Tree for ENV polyprotein

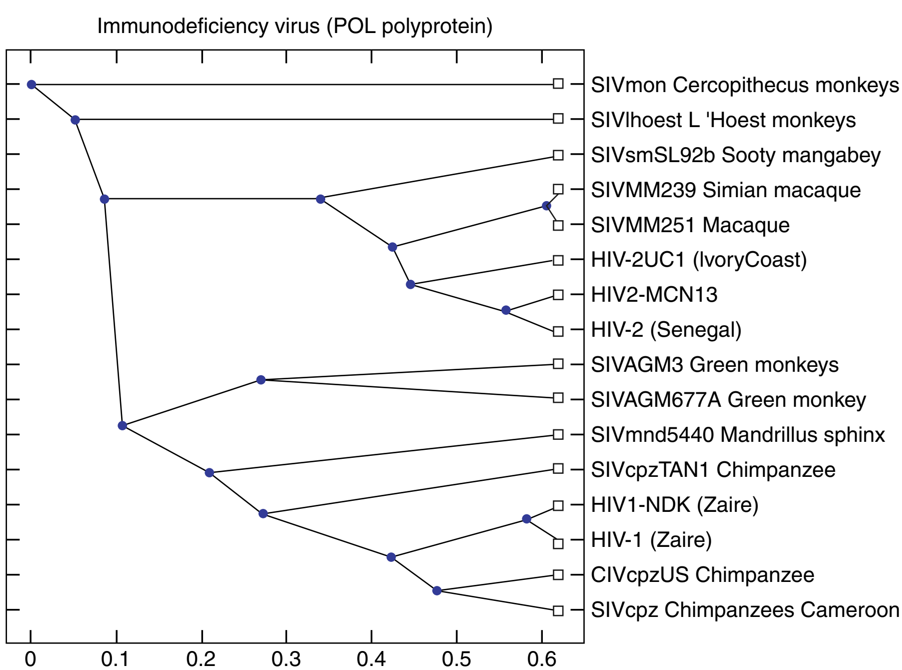


Figure 5 Phylogenetic Tree for POL polyprotein

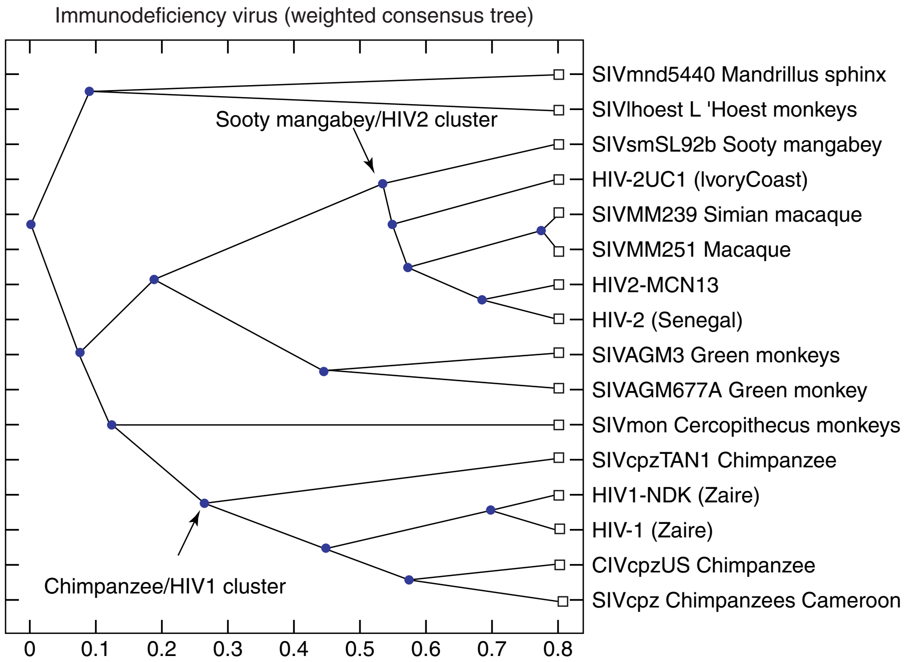


Figure 6 Weighted Consensus Phylogenetic Tree for HIV Polyproteins

informatics Toolbox provide bioinformaticists with a powerful development environment tool set for their mathematical and statistical work.

Further information

More information about MATLAB and the Bioinformatics Toolbox is available from **http://www.mathworks.com/products/bioinfo**.

Full documentation and many examples from the Bioinformatics Toolbox are available at **http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo**.

Example code and many useful tools are available from the MATLAB Central File Repository **http://www.mathworks.com/matlabcentral**.

References

Alizon M, Wain-Hobson S, Montagnier L and Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. *Cell*, **46**(1), 63–74.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis*, Cambridge University Press: Cambridge.

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, *et al*. (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, **397**(6718), 436–441.

Jukes T and Cantor C (1969) *Evolution of Protein Molecules*, Academic Press: New York.
Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.
Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, **48**, 443–453.
Page RDM and Holmes EC (1999) *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science Ltd.: Oxford.
Rambaut A, Posada D, Crandall KA and Holmes EC (2004) The causes and consequences of HIV evolution. *Nature Reviews. Genetics*, **5**(1), 52–61.
Salemi M, De Oliveira T, Courgnaud V, Moulton V, Holland B, Cassol S, Switzer WM and Vandamme AM (2003) Mosaic genomes of the six major primate lentivirus lineages revealed by phylogenetic analyses. *Journal of Virology*, **77**(13), 7202–7213.
Smith TF and Waterman MS (1981) The identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.
Tajima F and Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, **1**, 269–285.

Short Specialist Review

Gibbs sampling and bioinformatics

Xiaole Shirley Liu

Dana-Farber Cancer Institute, Boston, MA, USA

1. Gibbs sampling introduction

Gibbs sampling is a variation of the Metropolis–Hastings algorithm, and one of the best-known Markov chain Monte Carlo methods. The algorithm was developed by Geman and Geman (1984), and named after the American physicist Josiah Willard Gibbs (1839–1903) in reference to the similarity between the sampling algorithm and statistical physics. It is useful when the joint or marginal distribution of two or more random variables is too complex to directly sample, but the conditional densities of each variable are available. Starting from initial values, a Gibbs sampler draws samples from the distribution of each variable in turn, conditional on the current values of the other variables. The sequence of the samples is a Markov chain. As the chain approaches infinity, the distribution of all variables approximates the joint distribution and the distribution of each variable converges to its marginal distribution (see Gelfand and Smith, 1990 and Casella and George, 1992 for details).

2. Gibbs sampling application to sequence motif finding

Gibbs sampling was first introduced to bioinformatics to detect local sequence patterns or motifs shared by multiple sequences (Lawrence *et al.*, 1993). These patterns often reflect similar biological functions, for example, motifs shared by promoters of multiple genes may indicate similar gene expression regulation. In the motif finding problem, there are two random variables involved: a position-specific weight matrix θ of width w to represent residue frequencies at each position of the motif; an alignment $A = (a_1, a_2, \ldots, a_n)$ to describe the location of the motif occurrence within each of the n sequences. Starting from random initial alignment, the Gibbs sampler iteratively updates the motif θ and the alignment A in turn conditional on the other variable. Liu (1994) observed that θ can be integrated out explicitly or "collapsed", and the alignment location of one sequence can be updated on the basis of the alignment of all other sequences. The collapsed Gibbs sampler picks a sequence i at random (or in a certain order), and estimates θ by counting the residues at each aligned position from all other sequences (plus some

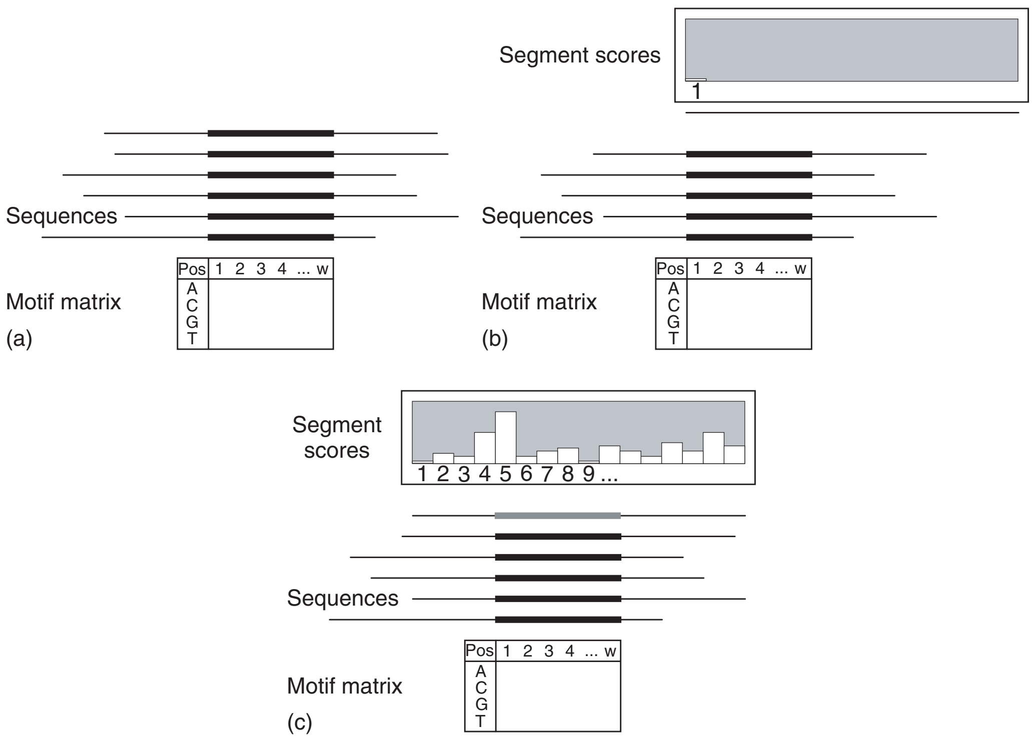


Figure 1 Gibbs sampling procedure for sequence motif finding. (a) One w-mer from each sequence is randomly picked to establish the initial motif probability matrix and initial alignment. (b) During Gibbs sampling iterations, a sequence i is picked at random and its contributing w-mer removed from the motif. The current motif is used to score every w-mer in sequence i. (c) A new w-mer in sequence i is sampled to add to the motif and put sequence i back to the alignment. The probability of picking a w-mer is proportional to its score. In Gibbs sampling, processes (b) and (c) are iterated until the motif converges

pseudocounts). Given this θ, each segment of width w in sequence i can be scored, and a new is sampled from this score distribution (Figure 1). The score is the probability of generating the segment by the motif over the probability of generating it by the general background, and thus preference is given to a segment that is similar to the motif yet different from the background.

Since the original Gibbs sampler was applied to the motif finding problem, there have been many improvements to make it more flexible in analyzing biological sequences. The Gibbs Motif Sampler (Liu *et al*., 1995) allows variable motif occurrences in each sequence by concatenating the sequences into one and checking each segment's sampled rate at stationary distribution. It also allows gaps in motifs by sampling only information-rich positions. AlignACE (Roth *et al*., 1998) finds multiple distinct motifs in the sequences by iteratively masking out reported motifs, and measures the goodness of a motif on the basis of whether the pattern is enriched compared to the whole-genome sequence. BioProspector (Liu *et al*., 2001) improves the motif specificity by introducing Markov dependency in the nonmotif background and can find two-blocked motifs with variable-sized gaps

in each sequence. Zhou and Liu (2004) improved the motif models by considering pairs of correlated motif positions. In higher eukaryotes, the sequences to be analyzed are much longer, and motifs (either the same or different motifs) often occur in close proximity to form modules. CompareProspector (Liu *et al*., 2004) reduces the search space by starting the search from subsequences conserved in evolution and sampling alignments weighted by evolutionary conservations. CisModule (Zhou and Wong, 2004) looks for modules where motif clusters reside. Given the module and motif locations, CisModule updates the motifs; given the motifs, CisModule first updates the module locations, and then samples the motif locations within each module.

3. Other Gibbs sampling applications

In the postgenomics era, high-throughput experiments are routine procedures in biomedical research. Gibbs sampling is well suited for conducting inference on these high-throughput data. We will discuss two of the most successful Gibbs sampling applications: haplotype inference and microarray bicluster analysis.

Single-nucleotide polymorphisms (SNPs) in the genome may help map complex disease genes and influence drug response. However, high-throughput SNP genotyping and analysis is inefficient and error prone. Usually, a set of closely linked loci has a limited collection of genotypes called haplotypes. Given a population of observed multilocus phenotypes, inferring each individual's haplotype is more informative and robust. The two variables involved here are the population haplotype frequencies θ and the assigned haplotype pairs $Z = ((z_{11}, z_{12}), \ldots, (z_{n1}, z_{n2}))$ that are compatible with the observed phenotypes. Gibbs sampler can converge on the stationary distribution of the two variables by iteratively updating the two variables in turn conditional on the other. Stephens *et al*. (2001) use the collapsed Gibbs sampler by iteratively picking an individual at random and updating his/her haplotype on the basis of the haplotype assignments of all other people. This method could be too computationally intensive when the number of SNPs in linked loci is too large. Niu *et al*. (2002) adopt the partition ligation strategy by dividing the long loci into small blocks, performing haplotype inference on each, and finally ligating the short haplotypes together.

Over the last 8 years, microarray experiments have become the standard procedure to profile the genome-level gene expressions. After microarray data are collected over a large number of biological samples or experimental conditions, clustering analysis can be conducted to find group of genes or conditions with similar expression profiles. When the number of samples or conditions is very large, often, only a subset of genes over a subset of conditions forms a tight enough cluster. Deciding the genes and conditions assigned to such a bicluster can be achieved by Gibbs sampling. Sheng *et al*. (2003) first discretize the expression values, randomly assign initial genes and conditions to a bicluster, then update genes and clusters in turn. With fixed conditions in the bicluster, every gene i is reassigned on the basis of whether its expression profile under the clustered conditions is similar to the expression profiles of other genes in the bicluster. With fixed genes in the bicluster, every condition j is reassigned on the basis of whether the expression profile of

clustered genes under condition j is similar to those under all other conditions in the bicluster. Wu *et al*. (2004) use a Gibbs sampler that aims to find a bicluster with the largest number of genes whose expression profiles pass a similarity threshold under at least a certain number of conditions. Each sampling iteration removes a condition i in the bicluster, scores each nonclustered condition by counting the number of genes with similar expression profiles if the condition is added to the bicluster, and samples a new condition to the bicluster on the basis of this score. Both methods identify one bicluster at a time, and multiple biclusters can be detected by iteratively masking out the expression values assigned to a converged bicluster.

References

Casella G and George EI (1992) Explaining the Gibbs sampler. *American Statistician*, **46**, 167–174.

Gelfand AE and Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Liu JS (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966.

Liu X, Brutlag DL and Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Proceedings of Pacific Symposium on Biocomputing*, **2001**, 127–138.

Liu Y, Liu XS, Wei L, Altman RB and Batzoglou S (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Research*, **14**, 451–458.

Liu JS, Neuwald AF and Lawrence CE (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, **90**, 1156–1170.

Niu T, Qin ZS, Xu X and Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, **70**, 157–169.

Roth FP, Hughes JD, Estep PW and Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, **16**, 939–945.

Sheng Q, Moreau Y and De Moor B (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics*, **19**(Suppl 2), II196–II205.

Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Wu CJ, Fu Y, Murali TM and Kasif S (2004) Gene expression module discovery using Gibbs sampling. *Genome Informatics*, **15**, 239–248.

Zhou Q and Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.

Zhou Q and Wong WH (2004) CisModule: De Novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12114–12119.

Short Specialist Review

Applications of RNA minimum free energy computations

Peter Clote
Boston College, Chestnut Hill, MA, USA

1. Introduction

The article "RNA secondary structure prediction" (*see* Article 96, **RNA secondary structure prediction**, Volume 8) discussed dynamic programming methods to predict the minimum free energy (mfe) E_0 and minimum free energy secondary structure S_0 of a given RNA sequence, using the Turner energy model (Xia *et al*., 1999), with experimentally measured negative, stabilizing base stacking energies and positive, destabilizing loop energies (hairpin loop, interior loop, etc.). Here, we survey a few applications of this method to determine regulatory regions of RNA and more generally to determine noncoding RNA genes.

2. Methods

A general, often-used approach in genomic motif finding is to fix a window size n, and scan through a chromosome or genome, repeatedly moving the window forward one position. The window contents may then be scored using machine-learning algorithms, such as weight matrices (Gribskov *et al*., 1987; Bucher, 1990), hidden Markov models (Baldi *et al*., 1994; Eddy *et al*., 1995; *see also* Article 98, **Hidden Markov models and neural networks**, Volume 8), neural networks (Nielsen *et al*., 1997; *see also* Article 98, **Hidden Markov models and neural networks**, Volume 8), and support vector machines (Vert, 2002; *see also* Article 110, **Support vector machine software**, Volume 8). While accurate detection of protein coding genes can be achieved using hidden Markov models (Borodovsky and McIninch, 1993; Burge and Karlin, 1997), by exploiting the nucleotide bias present in a succession of codons, such signals are less apparent in noncoding RNA genes.

Noncoding RNA (ncRNA) (Eddy, 2001; Eddy, 2002) is transcribed from genomic DNA and plays a biologically important role, although it is not translated into protein. Examples include tRNA, rRNA, XIST (which in mammalian males suppresses expression of genes on the X chromosome) (Brown *et al*., 1992), metabolite-sensing mRNAs, called *riboswitches*, discovered to interact with small ligands and up- or downregulate certain genes (Barrick *et al*., 2004), tiny noncoding RNA

(tncRNA) (Ambros *et al*., 2003), and miRNA (microRNA). MicroRNAs are ~21 nucleotide (nt) sequences, which are processed from a stem-loop precursor by Dicer (Tuschl, 2003; Lim *et al*., 2003) – see Figure 1, which depicts the predicted secondary structure for *C. elegans* let-7 precursor RNA. MicroRNA is (approximately) the reverse complement of a portion of transcribed mRNA, and has been shown to prevent the translation of protein from mRNA – this is an example of posttranscriptional regulation.

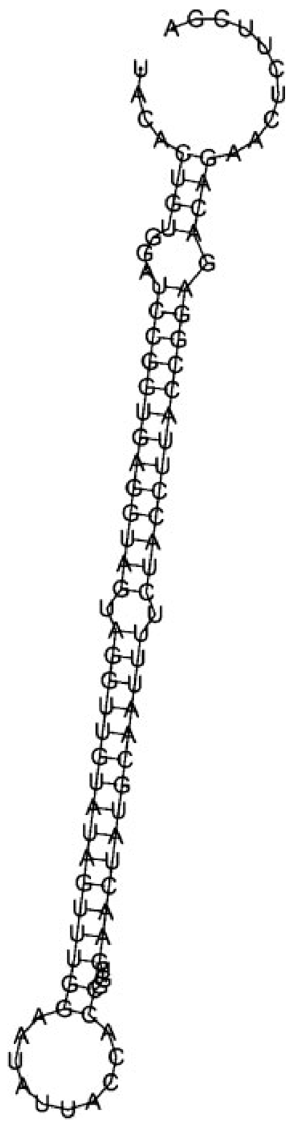


Figure 1 Predicted minimum free energy secondary structure of *C. elegans* let-7 precursor RNA; sequence taken from Rfam. Predicted minimum free energy for this 99-nt sequence is −42.90 kcal mol^{-1} (prediction made using `Vienna RNA package`)

For certain classes of ncRNA, there is a sufficiently well-defined sequence consensus or common secondary structure shared by experimentally determined examples, so that machine-learning methods such as *stochastic context-free grammars* (SCFG) have proven successful. RNA secondary structures can be depicted as a balanced parenthesis expression with dots, where balanced left and right parentheses correspond to base pairs and dots to unpaired bases.

In particular, by training an SCFG on many examples of tRNA, additionally using promoter detection with heuristics, T. Lowe and S. Eddy's program `tRNAscan-SE` identifies "99–100% of transfer RNA genes in DNA sequence while giving less than one false-positive per 15 gigabases" (Lowe and Eddy, 1997).

Exploiting the fact that ncRNA genes of the AT-rich thermophiles *Methanococcus jannaschii* and *Pyrococcus furiosus* have high $G + C$ content, Klein *et al*. (2002) describe a surprisingly simple yet accurate noncoding RNA gene finder for these and related bacteria. Lim *et al*. (2003) describe a novel computational procedure, `MiRscan`, to identify vertebrate microRNA genes. In a moving-window scan of the noncoding portion of the human genome, `MiRscan` uses `RNAfold` from the Vienna RNA Package (Hofacker *et al*., 1994) to search for stem-loop structures having at least 25 bp and predicted mfe of -25 kcal mol^{-1} or less. Subsequently, `MiRscan` passes a 21-nt window over each conserved stem-loop, then assigns a log-likelihood score to each window to determine how well its attributes resemble those of certain experimentally verified miRNAs of *Caenorhabditis elegans* and *Caenorhabditis briggsae* homologs.

Using the power of comparative genomics (alignments of homologous ncRNA genes from different organisms), Rivas and Eddy (2001) developed the program `QRNA` that trains a *pair stochastic context-free grammar*, given pairs of homologous ncRNA genes. Coventry *et al*. (2004) developed the algorithm `MSARI`, which assigns appropriate weights for local shifts of a ClustalW multiple sequence alignment of many (e.g., 11) homologous ncRNAs, in order to detect a conserved pattern of secondary structure. The authors suggest that a gene finder might then be trained on automatically generated multiple sequence alignments of RNAs, suitably corrected by their algorithm to identify the underlying sequence/structure alignment.

A related and equally important algorithmic task is the detection of regulatory and retranslation signals in the untranslated region (UTR), both upstream $5'$ and downstream $3'$ of the coding sequence (cds) of messenger RNA. For instance, Lescure *et al*. (1999) used Vienna RNA Package `RNAfold` in a simple screen to determine putative selenocysteine insertion sequence (SECIS) elements (see Hüttenhofer and Böck, 1998 for a review of selenocysteine incorporation); the authors subsequently performed (wet-bench) experiments to validate certain SECIS elements. Grate (1998) applied Eddy's RNA structure pattern searching algorithm program `RNABOB` in the search for SECIS elements in HIV. Bekaert *et al*. (2003) developed a model for -1 eukaryotic ribosomal frame-shifting sites, on the basis of a *slippery sequence* and a predicted *pseudo-knot* structure.

Recently, Washietl *et al*. (2005) described a noncoding RNA gene finder, based on a combination of mfe Z-score computations and comparative genomics. Here, the Z-score of the content of a current window of size n is defined by $\frac{x-\mu}{\sigma}$, where x is the mfe of the window contents, while μ, σ are respectively the mean and standard deviation of the mfe of random length n sequences having the same mono-

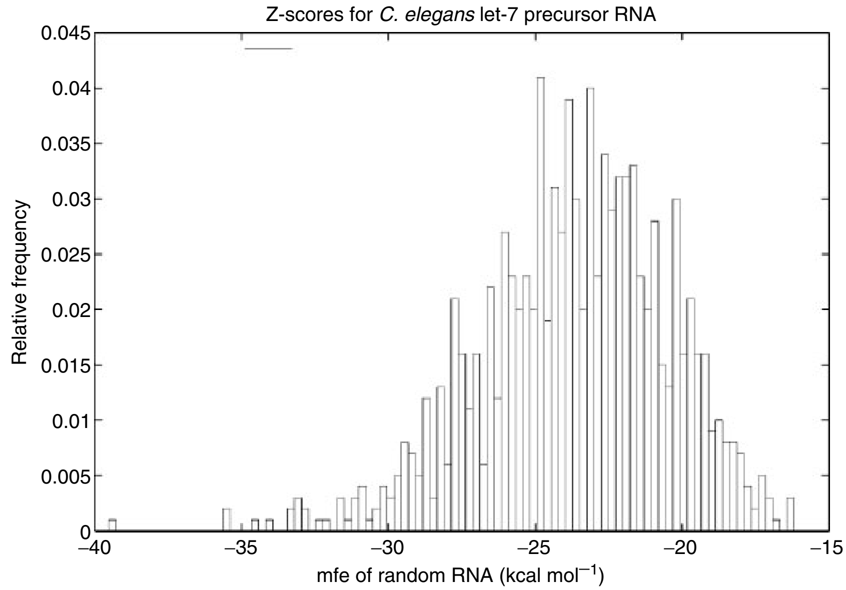


Figure 2 Histogram of the mfe for 1000 random RNAs, each having the same (exact) dinucleotide frequency as that in *C. elegans* let-7 precursor RNA. Mean mfe is −23.54 kcal mol^{-1} with standard deviation 3.23, hence the Z-score for let-7 precursor RNA is $\frac{-42.90-(-23.54)}{3.23}$ or roughly −6. Random RNA produced by the method of Workman and Krogh (1999) as implemented in Clote *et al*. (2005) (minimum free energy computed using `RNAfold`)

or possibly dinucleotide frequencies as that of the window contents (see Workman and Krogh, 1999; Clote *et al*., 2005 for discussion, and Figure 2 for an example). A Z-score of x that is approximately zero means that the mfe of sequence x is indistinguishable from that of its randomizations (i.e., the mfe of a randomization of x is just as often lower as higher than that of x). Similarly, a negative Z-score of x means that the mfe of x is lower than that of most of its randomizations.

Results from Rivas and Eddy (2000) indicate that using Z-score alone is not sufficiently statistically significant to be used to find ncRNA genes. Nevertheless, Washietl *et al*. (2005) combine the use of Z-scores with comparative genomics to develop a remarkably accurate and computationally efficient noncoding RNA gene finder. The authors make novel use of a support vector machine to compute the mean μ and standard deviation σ, rather than relying on slow repeated randomizations of window contents.

References

Ambros V, Lee R, Lavanway A, Williams P and Jewell D (2003) MicroRNAs and other tiny endogenous RNAs in *c. elegans*. *Current Biology*, **13**, 807–818.

Baldi P, Chauvin Y, Hunkapiller T and McClure MA (1994) Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 1059–1063.

Barrick J, Corbino K, Winkler W, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, *et al*. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial

genetic control. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(17), 6421–6426.

Bekaert M, Bidou L, Denise A, Duchateau-Nguyen G, Forest J, Froidevaux C, Hatin I, Rousset J and Termier M (2003) Towards a computational model for −1 eukaryotic frameshifting sites. *Bioinformatics*, **19**, 327–335.

Borodovsky M and McIninch J (1993) Genmark: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, **17**(2), 123–133.

Brown C, Hendrich B, Rupert J, Lafreniere R, Xing Y, Lawrence J and Willard H (1992) The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, **212**, 563–578.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Clote P, Ferrè F, Kranakis E and Krizanc D (2005) Structural rna has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**(5), 578–591.

Coventry A, Kleitman D and Berger B (2004) MSARi: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(33), 12102–12107.

Eddy SR (2001) Non-codingRNA genes and the modern RNA world. *Nature Reviews*, **2**, 919–929.

Eddy SR (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.

Eddy SR, Mitchison G and Durbin R (1995) Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology*, **2**(1), 9–24.

Grate L (1998) Potential SECIS elements in HIV-1 strain HXB2. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **17**(5), 398–403.

Gribskov M, McLachlan A and Eisenberg D (1987) Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 4355–4358.

Hüttenhofer A and Böck A (1998) RNA structures involved in selenoprotein synthesis. In *RNA Structure and Function*, Cold Spring Harbor Laboratory Press, 603–639.

Hofacker IL, Fontana W, Stadler P, Bonhoeffer L, Tacker M and Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie*, **125**, 167–188.

Klein R, Misulovin Z and Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7542–7547.

Lescure A, Gautheret D, Carbon P and Krol A (1999) Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *The Journal of Biological Chemistry*, **274**(53), 38147–38154.

Lim L, Glasner M, Yekta S, Burge C and Bartel D (2003) Vertebrate microRNA genes. *Science*, **299**(5612), 1540.

Lowe T and Eddy S (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleid Acids Research*, **25**(5), 955–964.

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, **10**(1), 1–6.

Rivas E and Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNA. *Bioinformatics*, **16**, 573–585.

Rivas E and Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**(8), **http://www.biomedcentral.com/1471-2105/2/8**.

Tuschl T (2003) Functional genomics: RNA sets the standard. *Nature*, **421**, 220–221.

Vert J-P (2002) Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. In *Pacific Symposium on Biocomputing 2002*, Altman R, Dunker A, Hunter L, Lauderdale K and Klein T (Eds.), World Scientific, 649–660.

Washietl S, Hofacker IL and Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, **19**, 327–335.

Workman C and Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research*, **27**, 4816–4822.

Xia T, SantaLucia J, Burkard M, Kierzek R, Schroeder S, Jiao X, Cox C and Turner D (1999) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Cluster architecture

Chris Dagdigian
BioTeam Inc., Cambridge, MA, USA

1. Clusters

Cluster computing allows mass-market PC and server systems to be networked together to form an extremely cost-effective system capable of handling supercomputer-scale workloads. Cluster sizes can range from two machines up through many thousands of interconnected systems.

Biologists have found that clusters can be used both to expand the scale of existing informatics research efforts as well as to investigate research areas previously written off as prohibitively expensive or computationally infeasible.

The use of the term "cluster" in this article refers to systems operating on the same network or within the same cabinet, datacenter, building, or campus. This differs from "grid computing", which is typically a term associated with the use of many clusters or diverse distributed systems linked together via the Internet or other wide area networking (WAN) technologies. Clusters typically have a single administrative domain, whereas grids can be composed of geographically separated systems and services each of which with its own administrative domain and access policies. The term "Beowulf cluster" typically refers systems purpose-built for parallel computation.

2. Life science cluster characteristics

For maximum utility, flexibility, and capability, scientific and research goals are the primary drivers for cluster architecture decisions. To do otherwise is to risk unintended consequences that limit how the cluster may be used as a research or data processing tool.

The performance characteristics and runtime requirements of the intended scientific application mix play a major role in hardware selection and overall system design. Researchers with a significant need for bioinformatics sequence analysis often find that many of their applications are performance-bound by the amount of physical memory (RAM) in a machine and the speed of underlying storage and I/O subsystems. Users running large parallel applications will find that the speed and latency characteristics of the cluster network will often be the most important factor in optimizing performance and throughput. Some applications including chemistry and molecular modeling codes can be CPU-bound, and run best on systems with

very fast CPUs and high data transfer rates between processor, onboard cache, and external memory.

Understanding the performance-affecting requirements of the scientific application mix is essential when planning new clusters or even upgrading existing systems. Wherever possible, benchmarks reflecting real-world usage and workflows should be performed.

3. Serial or "batch computing" versus parallel computing

Unfortunately, many cluster references, resources, and cluster "kits" are biased toward a process of parallel cluster computing that is not commonly used on many life science settings. A single parallel application is designed to run across many systems simultaneously. The most commonly seen parallel applications are based on PVM (Parallel Virtual Machine) or MPI (Message Passing Interface) standards.

The use of PVM- or MPI-aware parallel applications tends to be rare in the life sciences. The exception tends to be in the areas of molecular modeling and computational chemistry, where there exists a significant body of parallel software available and in use.

A far more common requirement is the need to repeatedly run large numbers of traditional nonparallel scientific applications or algorithms. Each application instance becomes a stand-alone job that can be efficiently scheduled and independently distributed across a cluster.

Large computational biology problems such as bioinformatics sequence analysis fit nicely into this paradigm – every large analysis task is capable of being broken down into individual pieces that can be executed in any order, independent of any other segment. This approach is known as "serial" or "batch" computing. Problems that can be broken up for serial or batch distribution across a cluster are also referred to as "embarrassingly parallel" problems.

The workflow bias toward serial computing rather than parallel computing is one of the main distinguishing characteristics of life science clusters.

4. Cluster topology

Clusters tend to use variations on a "portal" architecture in which all of the cluster compute nodes are kept isolated on a private network (Figure 1). Management and usage of the cluster is achieved via use of a machine that is attached to both the public organizational network and the private cluster network. Additional servers, storage devices, database servers, and management servers are also "multi-homed" to both networks as needed.

A schematic representation of the portal style cluster architecture is seen in Figure 2.

Advantages of the portal architecture approach include:

- Easier management and administration. Cluster operators are free to control, customize, and modify essential network services such as DHCP, TFTP, PXE, LDAP, NIS, and so on, without affecting the organizational network.



Figure 1 A small bioinformatics research cluster using Apple G4 and Intel Xeon-based server systems

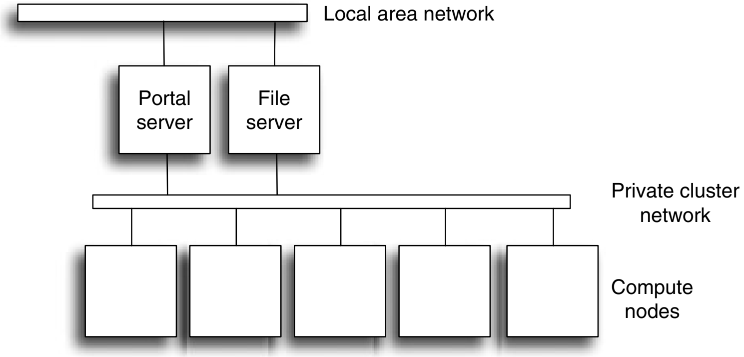


Figure 2 Logical view – portal architecture

- Security and abstraction of computing resources. The architecture prevents large numbers of compute nodes from being directly accessible to the public network. Cluster users are encouraged to think of the cluster nodes as anonymous and interchangeable.

In instances in which jobs running on the cluster may need to communicate with systems or services outside of the cluster (database or LIMS systems, etc.), it is a simple matter to set up NAT (network address translation) or proxy services.

5. Network and interconnects

General-purpose clusters use standard switched Ethernet networking components as the primary method of interconnecting cluster nodes. Some cluster operators find that running a second private "management" network alongside the primary network has administrative advantages. The cost of copper-based Gigabit Ethernet networking products has plummeted to the point where it has become the default choice for clusters of all sizes.

Multiple Gigabit Ethernet links can be "trunked" or bonded together to achieve higher performance. In many cases, the cluster services that best benefit from added bandwidth and network performance capability are cluster file-servers and other data staging or storage systems.

Ethernet networking may not be suitable for all network-dependent tasks and use-cases. In particular, some parallel PVM or MPI applications may be performance limited when run over an Ethernet network due to the relatively high latency between transported packets. Some parallel or global distributed file-system technologies also prefer or may even require the use of a special high-speed, low-latency interconnect.

There are a number of available products and technologies aimed at providing clusters with a higher-performance interconnect. Examples include Myrinet and Infiniband. These can be deployed cluster-wide to complement an existing Ethernet network or deployed to a limited subset of cluster nodes to support parallel applications.

Given the relative lack of latency-sensitive, massively parallel scientific software in the life sciences, the use of interconnect technologies other than Ethernet is quite rare.

6. Distributed resource management

An essential component, especially systems supporting multiple groups or research efforts, is the software layer that handles resource allocation and all aspects of job scheduling and execution across many machines. Generally known as "distributed resource management" (DRM), these software products are critical to successful cluster operation. The most commonly seen DRM products in life science settings are Platform LSF and Sun Grid Engine. Other DRM suites include Portable Batch System (PBS) and Condor. Proper selection and configuration of the DRM software layer is extremely important, as the DRM layer is the "glue" that ties the cluster together.

7. Storage

The most commonly encountered performance-limiting bottleneck in life science clusters is the speed of both local and network-resident storage systems. Network attached storage (NAS) devices providing NFS-based file systems are often used

as a way of making vast amounts of raw research data available for analysis within the cluster. These devices (or the network itself) can quickly be saturated on even moderately busy clusters.

The cost of acquiring a few terabytes of raw NAS can vary by as much as $50 000 USD or more between competitive storage products. Fortunately, the wide price range allows for a healthy ecosystem of differentiated storage products offering different levels of performance, resiliency, cost, and capability.

A popular method of increasing local disk performance in life science clusters involves populating compute nodes with multiple large but inexpensive ATA drives that are mirrored or striped together via software RAID. In addition to vastly increased local I/O performance, these disks can also be used to cache popular databases or files from the central file-server. Very significant amounts of cluster network traffic and file-server load can be eliminated simply by staging data to the local compute nodes prior to launching a large analytical job.

8. Reducing administrative burden

Clusters of loosely interconnected server systems can represent a significant operational challenge. Several inexpensive hardware or software-based methodologies can greatly reduce the amount of effort needed to maintain cluster systems. Approaches include various techniques for performing unattended operating system installations (or reinstallation), remote power control products, and serial console access concentrators.

9. Additional resources

The **bioclusters@bioinformatics.org** mailing list is a 600+ member on-line community of life science cluster users and practitioners. To subscribe or view the list discussion archives, visit **http://bioinformatics.org/lists/bioclusters**.

Further reading

Dagdigian C (2005) Biocluster whitepapers, HOWTO's and conference presentations, available online at **http://bioteam.net/dag/**

Sterling T (Ed.) (1999) *How to Build a Beowulf*, MIT Press: Cambridge.

Sterling T (Ed.) (2002) *Beowulf Cluster Computing With Linux*, MIT Press: Cambridge.

Relational databases in bioinformatics

Hans-Peter Kriegel, Peer Kröger and Stefan Schönauer
University of Munich, Munich, Germany

1. Introduction

The rapidly growing amount of data in bioinformatics makes the use of a database management system (DBMS) indispensable. Additionally, modern DBMS provide advantages like concurrent multiuser access, security features, and easy web integration.

The relational data model, introduced by Codd in 1970, is the most successful model for DBMS, and is used in numerous commercial products.

2. Relational data management

The relational data model is based on the single concept of relations for data representation. Relations are perceived by the user as tables whose rows are called *tuples* representing the objects. All entries in a column of a table are atomic values from the same domain, and a single column is called an *attribute*. The name of a relation together with its attribute names and domains form the schema of the relation. A minimal subset of the attributes from a relation whose value combination uniquely distinguishes each possible tuple of the relation is called a *key*. Relationships between object sets represented by relations are also stored in tables built from the keys of the relations participating in the relationship. All operations defined on relations take relations as input and generate relations as output. Queries are usually expressed in the Standard Query Language (SQL), which allows data definition as well as data manipulation statements but no procedural elements like loops. Consequently, SQL statements have to be embedded in application programs written in a different programming language.

A sample relational data modeling is illustrated in Figure 1. The example models complexes of proteins. A complex has properties such as name and function and can consist of several proteins. A protein also has properties such as name and function and can participate in several complexes. Thus, two tables "Protein" and "Complex" are generated with particular attributes. Each table specifies attributes

Protein

pID	Name	Function	–
–	–	–	–
101	xyz	Protease	–
–	–	–	–

Complex

cID	Name	Function	–
–	–	–	–
075	abc	null	–
–	–	–	–

Contains

cID	pID
–	–
075	101
–	–

Figure 1 Sample relational model

as primary keys, in this case artificial keys such as "pID" (protein ID) and "cID" (complex ID) for unique identification of its tuples. A further table "Contains" consisting of the primary keys of Complex and Protein links both tables, modeling the relationship between a complex and its contained proteins. A protein may be part of several complexes. Let us note that for some tuples in the tables the value might be unknown, for example, for complex "abc" with cID = 075, the function is not known. In this case, the value is *null*.

A sample query "Find the names of all complexes that contain protein 'xyz'" could be formulated in SQL as follows:

select Complex.name
from Protein, Complex, Contains **where** Protein.name = "xyz"
and Protein.pID = Contains.pID **and** Contains.cID = Complex.cID

3. Advantages

The use of relational DBMSs has several advantages. Owing to its widespread applicability, several commercial and even open-source relational DBMSs are available. With SQL, there exists a standardized and easy-to-learn query language. Several standard interfaces available for SQL allow the integration of databases into web applications or other information systems. Along with the widespread use of relational DBMSs, many powerful tools for such systems were developed, providing support for application development or database administration.

4. Limitations

Although relational databases are frequently used for bioinformatics applications, they cause several major drawbacks, which are described in the following.

4.1. Complex schemas

Bioinformatics research deals with complex objects. These complex objects such as proteins, genes, metabolic pathways, and so on, can often not be modeled adequately with the constructs of relational data models. The resulting data schemas are often rather complex and not intuitive anymore, and thus they are hard to understand and administrate. The information about a single complex biological object is spread over several relations, each describing a single aspect of the object. In addition, data models for bioinformatics databases tend to evolve frequently, increasing the problems of administration.

4.2. Managing biospecific objects

Bioinformatics objects are not only complex to model. Biological entities such as genes and proteins also provide bulky data types that are difficult to model and manage with traditional relational methods. Typically, bulky data types include sequences, for example, nucleotide or amino acid sequences, images, for example, MNR images, set-valued attributes such as molecules consisting of several atoms, and graph- or tree-structured data, for example, pathways or phylogenetic trees. In addition, bioinformatics databases also have to cope with missing attribute values that cannot be adequately and efficiently supported by traditional relational data management concepts.

4.3. Querying

Using SQL is the traditional, most powerful, and most comfortable way to query and extract information from relational databases. To formulate an SQL query, the user must have an overview of the database schema and must know which relations and attributes are relevant for the intended query. Obviously, complex and unintuitive relational schemas are hard to query. Furthermore, in order to manipulate an entire object distributed over several relations, these relations have to be joined during query processing, which is often very time consuming. As a consequence, many bioinformatics databases prevent their users from querying the databases directly using SQL, but instead provide the so-called fixed-form query interfaces. Such fixed-form query interfaces provide a view on the database and allow using a predetermined set of relations and attributes. The queries are allowed only against these interfaces.

5. Discussion

Despite several limitations, relational database management is rather widespread in bioinformatics. Nevertheless, several interesting approaches have been developed recently to cope with the problems and limitations of relational data modeling mentioned above. Some of them are bioinformatics-specific solutions, that is, the

approaches are directly motivated by bioinformatics data management. One of these solutions is ACEDB, a DBMS providing common DBMS features such as concurrency and security features but relying on a data model that is similar to the semistructured data model and thus is more flexible and powerful in the modeling aspect. A further bioinformatics-specific solution is OPM, a commercial suite of tools providing a more powerful, extended object-oriented data model including an appropriate query language and a mapping of such OPM models to standard relational DBMSs. In addition, a more general approach is the concept of object-relational DBMSs, developed for all kinds of applications dealing with complex and spatial objects. An object-relational DBMS provides most features of a traditional relational DBMS but supports object-oriented modeling constructs such as nonatomic attribute types. These systems provide the possibility to specify objects with user-defined attribute types and user-defined functions to determine the behavior of such objects. Via a standard interface, these objects are usually accessible by standard SQL. The advantage of these concepts is that the schemas are less complex since the model is more powerful and expressive than the traditional relational model. Furthermore, the object-relational approach also provides all features of relational DBMSs such as SQL querying, concurrency, query optimization, and so on. The newest versions of DBMSs such as ORACLE, DB2, or MS SQL Server are in fact object-relational DBMSs. However, several bioinformatics databases do not use their features yet.

References and links

Articles on bioinformatics databases and books on relational data management

Date CJ (2003) *An Introduction to Database Systems*, Vol. I, Eighth Edition, Addison-Wesley: Bostan, MA.

Elmasri R and Navathe S (2000) *Fundamentals of Database Systems*, Third Edition, Benjamin/Cummings, Redwood City, CA.

François Bry and Peer Kroeger (2003) A computational biology database digest: data, data Analysis, and data management. In *Distributed and Parallel Databases*, Vol. 13, Kluwer Academic Press: Bostan, MA. pp. 7–42.

Watson RT (2003) *Data Management: Databases and Organizations*, Fourth Edition, Wiley: Hoboken, NJ.

Useful links

ACEDB: **http://www.acedb.org/**
European Bioinformatics Institute (EBI): **http://www.ebi.ac.uk/**
Kyoto Encyclopedia of Genes and Genomes (KEGG): **http://www.genome.ad.jp/kegg/**
MySQL: **www.mysql.com**
National Centre for Biotechnology Information (NCBI): **http://www.ncbi.nlm.nih.gov/**

Basic Techniques and Approaches

Support vector machine software

William Stafford Noble
University of Washington, Seattle, WA, USA

The support vector machine (SVM) (Boser *et al*., 1992; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) is a supervised learning algorithm, useful for recognizing subtle patterns in complex data sets. The algorithm performs discriminative classification, learning by example to predict the classifications of previously unseen data. The algorithm has been applied in domains as diverse as text categorization, image recognition, and hand-written digit recognition (Cristianini and Shawe-Tyalor, 2000). Recently, SVMs have been applied in numerous bioinformatics domains, including recognition of translation start sites, protein remote homology detection, protein fold recognition, microarray gene expression analysis, functional classification of promoter regions, prediction of protein–protein interactions, and peptide identification from mass spectrometry data (reviewed in Noble (2004)).

The popularity of the SVM algorithm stems from four primary factors. First, the algorithm boasts a strong theoretical foundation, based upon the dual ideas of VC dimension and structural risk minimization (Vapnik, 1998). Second, the SVM algorithm is well-behaved, in the sense that it is guaranteed to find a global minimum and that it scales well to large data sets (Platt, 1999a). Third, the SVM algorithm is flexible, as evidenced by the list of applications above. This flexibility is due in part to the robustness of the algorithm itself, and in part to the parameterization of the SVM via a broad class of functions, called *kernel functions*. The behavior of the SVM can be modified to incorporate prior knowledge of a classification task simply by modifying the underlying kernel function. The fourth and most important explanation for the popularity of the SVM algorithm is its accuracy. Although the underlying theory suggests explanations for the SVM's excellent learning performance, its widespread application is due in large part to the empirical success the algorithm has achieved.

An early successful application of SVMs to biological data involved the classification of microarray gene expression data. Brown *et al*. (2000) used SVMs to classify yeast genes into functional categories on the basis of their expression profiles across a collection of 79 experimental conditions (Eisen *et al*., 1998). Figure 1 shows a subset of the data set, divided into genes whose protein products participate and do not participate in the cytoplasmic ribosome. The ribosomal genes show a clear pattern, which the SVM is able to learn relatively easily. The SVM solution is a hyperplane in the 79-dimensional expression space. Subsequently, a

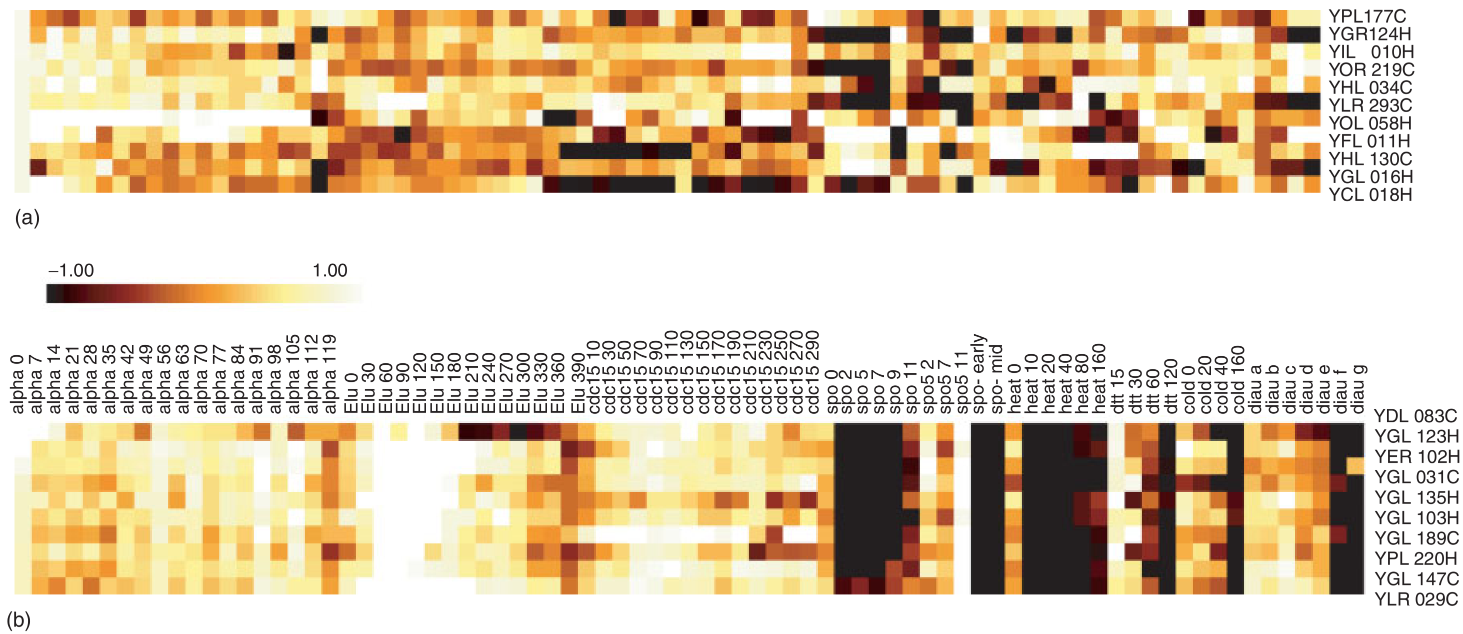


Figure 1 Gene expression profiles of ribosomal and nonribosomal genes. The figure is a heat map representation of the expression profiles of 20 randomly selected genes. Values in the matrix are log ratios of the two channels on the array. The upper 10 genes are from the "cytoplasmic ribosomal" class in the MIPS Yeast Genome Database, as listed at **www.cse.ucsc.edu/research/compbio/genex/genex.html**. An SVM can learn to differentiate between the characteristic gene expression pattern of ribosomal proteins and non-ribosomal proteins. The figure was produced using matrix2png (Pavlidis P and Noble WS (2003) Matrix2png: A utility for visualizing matrix data. *Bioinformatics*, **19**(2), 295–296)

given gene's protein product can be predicted to localize within or outside the ribosome, depending upon the location of that gene's expression profile with respect to the SVM hyperplane. SVMs have also been used successfully to classify along the other dimension of gene expression data: placing entire gene expression experiments into categories on the basis of, for example, the disease state of the individual from which the expression profile was derived (Furey *et al*., 2001; Ramaswamy *et al*., 2001; Segal *et al*., 2003).

A scientist who wishes to apply support vector machine learning to a particular biological problem faces first the question of which kernel function to apply to the data. Essentially, the kernel function defines a notion of similarity between pairs of objects. In the case of gene expression data, for example, the kernel value for two genes with similar expression profiles will be large, and vice versa. Mathematically, the kernel function must follow certain rules in order for the SVM optimization to work properly. Specifically, the kernel must be positive semidefinite, meaning that, for any given data set, the square matrix of pairwise kernel values defined from that data set will have nonnegative eigenvalues. In practice, however, most users of SVM software do not have to worry about these mathematical details, because the software typically provides a relatively small collection of valid kernel functions to choose from.

In general, a good rule of thumb for kernel selection is to start simple. The simplest kernel function is a linear scalar product, in which the products of corresponding elements in each of the two items being compared are summed. Normalizing the kernel, which amounts to projecting the data onto a unit sphere, is almost always a good idea. Centering the kernel, so that the points lie around the origin, is also helpful, though this operation is difficult to perform properly if the (unlabeled) test data is not available during training.

A slightly more complex class of kernel functions is the set of polynomials, in which the scalar product is raised to a positive power. The degree of the polynomial specifies the n-way correlations that the kernel takes into account. Thus, for the 79-element gene expression data mentioned above, a quadratic kernel implicitly accounts for all $79 \times 79 = 6241$ possible pairwise correlations among expression measurements. Obviously, as the polynomial degree gets larger, the number of features increases exponentially, eventually overwhelming the SVM's learning ability. It is necessary, therefore, to hold out a portion of the labeled data from the training phase and to use that data to evaluate the quality of the trained SVM with different kernels. The use of a hold-out set is the basis of a more general technique known as *cross-validation* (Duda and Hart, 1973). Figure 2 shows the effect of the polynomial degree on the SVM's ability to recognize members of the cytoplasmic ribosomal proteins. As the degree increases, accuracy improves slightly, but then drops again when the number of features becomes too large.

Besides polynomial kernels, the other common kernel is the radial basis function. This kernel warps the space where the data resides by putting a Gaussian over every data point. The kernel then measures similarities in that warped space. In this case, the user-controlled parameter is the width of this Gaussian. Like the polynomial degree parameter, the Gaussian width is typically selected via cross-validation. Some software will automatically select a reasonable (though not

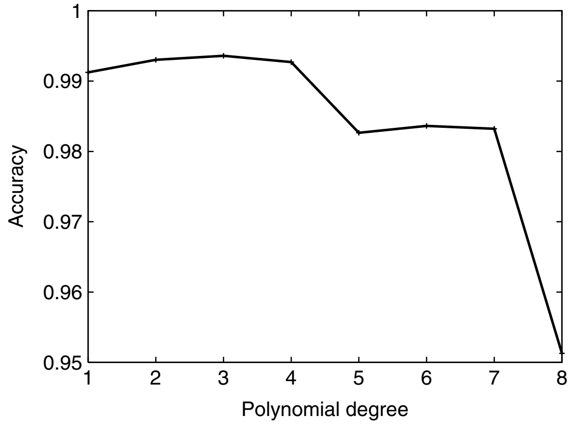


Figure 2 The effect of polynomial degree on SVM performance. The figure plots the classification accuracy of an SVM as a function of the polynomial degree. The SVM was trained to recognize ribosomal proteins, using data from Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Furey TS, Ares Jr M and Haussler D (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(1), 262–267 and using the Gist software with default parameters. Accuracy was measured using threefold cross-validation, repeated five times

necessarily optimal) width by examining the average distance between positive and negative examples in the training set.

The kernels discussed so far are applicable only to vector data, in which each object being classified can be represented as a fixed-length vector of real numbers. Gene expression data fits this paradigm, but many other types of biological data – protein sequences, promoter regions, protein–protein interaction data, and so on – do not. Often, it is possible to define a relatively simple kernel function from nonvector data by explicitly constructing a vector representation from the nonvector data and applying a linear kernel. For example, a protein can be represented as a vector of pairwise sequence comparison (BLAST or Smith–Waterman) scores with respect to a fixed set of proteins (Liao and Noble, 2002) or simply as a vector of 1's and 0's, indicating the presence or absence of all possible length-k substrings within that protein (Leslie *et al*., 2002). Similarly, protein–protein interaction data for a given protein can be summarized simply as a vector of 1's and 0's, where each bit indicates whether the protein interacts with one other protein. Other, more complex kernels, such as the diffusion kernel (Kondor and Lafferty, 2002), can also be constructed without relying upon an explicit vector representation. In general, selecting the kernel function is analogous to selecting a prior when building a Bayesian model. Thus, while simple kernels often work reasonably well, much research focuses on the development of complex kernel functions that incorporate domain knowledge about particular types of biological data.

In addition to selecting the kernel function, the user must also set a parameter that controls the penalty for misclassifications made during the SVM training phase. For a noise-free data set, in which no overlap is expected between the two classes being discriminated, a hard margin SVM can be employed. In this case,

the hyperplane that the SVM finds must perfectly separate the two classes, with no example falling on the wrong side. In practice, however, it is often the case that some small percentage of the training set samples are measured poorly or are improperly labeled. In such situations, the hard margin SVM will fail to find any separating hyperplane, and a soft margin must be employed instead. The soft margin incorporates a penalty term, and penalties are assigned to each misclassified point proportional to the point's distance from the hyperplane. Thus, a misclassified point that is close to the hyperplane will receive a small penalty, and vice versa. SVM practitioners differ over whether a linear (1-norm) or quadratic (2-norm) penalty yields better performance in general; however, regardless of the type of penalty imposed, the proper magnitude of the penalty is definitely problem-specific. Hence, this penalty parameter is typically selected either using cross-validation or by minimizing the number of support vectors (i.e., nonzero weights) in the SVM training output. The effect of this soft margin parameter on SVM accuracy is typically large, outweighing the effect, for example, of polynomial degree shown in Figure 2.

Setting the soft margin parameter is further complicated when the relative sizes of the two groups of data being discriminated are skewed. In many pattern recognition problems, this type of skew is typical: the interesting (positive) class of examples is small, and it is being discriminated from a relatively large (negative) class of uninteresting examples. In such a situation, the soft margin must be modified to asymmetrically penalize errors in the two classes: making a mistake on one of the relatively few positive examples is much worse than making a mistake on one of the many negative examples. A reasonable heuristic is to scale the penalty according to the relative class sizes. Hence, if there are 10 times as many negative examples as positive examples, then each positive misclassification receives 10 times as much penalty. In general, the degree of asymmetry should depend upon the relative cost associated with false-positive and false-negative predictions.

So far, we have discussed the SVM only in the context of discriminating between two classes of examples (e.g., ribosomal and nonribosomal genes). Many prospective SVM users worry that the binary nature of the algorithm is a fundamental limitation. This is not the case. Generalizing the SVM to perform multiclass pattern recognition is trivial. The most straightforward, and often the most successful, means of carrying out this generalization is via a so-called one-versus-all training paradigm. In order to discriminate among n different classes of examples, n SVMs are trained independently. Each SVM learns to differentiate between one class and all of the other classes. A new object is classified by running it through each of the SVMs and finding the one that produces the largest output value. This approach has been used successfully, for example, to classify 14 types of cancer from gene expression profiles (Ramaswamy *et al*., 2001).

Finally, a desirable feature of any classification algorithm is the ability to produce probabilistic outputs. In general, an SVM produces a prediction that has no units. This discriminant score is proportional to the example's distance from the hyperplane, so a large positive value implies high confidence that the example lies in the positive class. A relatively simple postprocessing step involving fitting a sigmoid curve can convert these discriminants to probabilities (Platt, 1999b).

Perhaps the easiest way for a novice to apply an SVM to a particular data set is via the web interface at **svm.sdsc.edu**. This server allows for the training of an

SVM on a labeled data set. The SVM can then be used to make predictions on a second, unlabeled data set. Additional flexibility, including leave-one-out cross-validation and empirical curve fitting to get probabilistic outputs, is available by downloading the underlying Gist software (**microarray.cpmc.columbia.edu/gist**) (Pavlidis *et al*., 2004). Many other SVM implementations are available in the "Software" section of **www.kernel-machines.org**. Among these, perhaps the most popular is SVMlight (**svmlight.joachims.org**), which implements a generalization of the fast sequential minimal optimization algorithm (Platt, 1999a) and also offers a range of features. Other popular implementations include mySVM (**www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM**), SVMTorch (**www.torch.ch**), and libSVM (**www.csie.ntu.edu.tw/~cjlin/libsvm**).

References

Boser BE, Guyon IM and Vapnik VN (1992) A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on COLT*, Haussler D (Ed.), ACM Press: Pittsburgh, PA, pp. 144–152.

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Furey TS, Ares Jr M and Haussler D (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(1), 262–267.

Cristianini N and Shawe-Taylor J (2000) *An Introduction to Support Vector Machines*, Cambridge University Press: Cambridge.

Duda RO and Hart PE (1973) *Pattern Classification and Scene Analysis*, Wiley: New York.

Eisen M, Spellman P, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D (2001) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**(10), 906–914.

Kondor RI and Lafferty J (2002) Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the International Conference on Machine Learning*, Sammut C and Hoffmann A (Eds.), Morgan Kaufmann: San Francisco, CA.

Leslie C, Eskin E and Noble WS (2002) The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, Altman RB, Dunker AK, Hunter L, Lauderdale K and Klein TE (Eds.), World Scientific: New Jersey, pp. 564–575.

Liao L and Noble WS (2002) Combining Pairwise Sequence Similarity and Support Vector Machines for Remote Protein Homology Detection. *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology, Washington, 18-21 April 2002*, pp. 225–232.

Noble WS (2004) Support vector machine applications in computational biology. *Kernel Methods in Computational Biology*, MIT Press: Cambridge, MA.

Pavlidis P and Noble WS (2003) Matrix2png: A utility for visualizing matrix data. *Bioinformatics*, **19**(2), 295–296.

Pavlidis P Wapinski I and Noble WS (2004) Support vector machine classification on the web. *Bioinformatics*, **20**(4), 586–587.

Platt JC (1999a) Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods*, Schoelkopf B, Burges CJC and Smola AJ (Eds.), MIT Press: Cambridge, MA.

Platt JC (1999b) Probabilities for support vector machines. In *Advances in Large Margin Classifiers*, Smola A, Bartlett P, Schoelkopf B and Schuurmans D (Eds.), MIT Press: Cambridge, MA. pp. 61–74.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(26), 15149–15154.

Segal NH, Pavlidis P, Noble WS, Antonescu CR, Viale A, Wesley UV, Busam K, Gallardo H, DeSantis D, Brennan MF, *et al*. (2003) Classification of clear cell sarcoma as melanoma of soft parts by genomic profiling. *Journal of Clinical Oncology*, **21**, 1775–1781.

Vapnik VN (1998) *Statistical Learning Theory*, Wiley: New York.

Brief Python tutorial for bioinformatics

Michael Poidinger
Johnson and Johnson Research Pty Ltd, Sydney, Australia

1. Introduction

Python is an interpreted, interactive, object-orientated programming language first created and released by Guido van Rossen in 1989–1991. It is called Python after the BBC comedy series, "Monty Pythons's Flying Circus". Python is currently an open source software project, which freely welcomes anyone with an interest in the language to contribute to it. The primary source of information concerning Python can be found at **www.python.org**, which includes downloads of the most recent release of the language (2.3.4) for all major operating systems. Other references to bioinformatic scripting in general and Python in particular can be found in this encyclopedia in other article references (*see* Article 103, **Using the Python programming language for bioinformatics**, Volume 8, Article 112, **A brief Perl tutorial for bioinformatics**, Volume 8, and Article 104, **Perl in bioinformatics**, Volume 8).

2. Language basics

Python uses a natural language syntax and indentation to delineate code blocks (as opposed to {} brackets used by most other programming languages). Control statements are terminated by a colon. Python is case sensitive, x and X could be the names of different variables. Python is a dynamically/implicitly typed language, which means that variables do not need to be formally declared at compile time. Variables can also be reused to hold any type of data during the runtime of the program. The Python interpreter can be run in interactive mode, either from a Unix style command line or by running pythonwin in Win32. This shell is exited with control-D. Inside the interpreter, lines are prefixed by `">>>"`. A Python script can be run by passing it as a parameter to the Python command. The –i parameter will invoke the interactive interpreter after script execution.

For example, assume that the script myscript.py contains the line x = 1:

```
python -i myscript.py
>>>print x
1
```

Basic data types: Python has four basic data types: numbers, strings, 0-based arrays, and dictionaries (hash tables). Literal strings can be enclosed by either single or double quotes. There are two types of arrays: mutable lists (contents can be changed after instantiation) and immutable tuples.

Assignment uses the = operator:

```
x = "hello world" #string
x = 1 #integer
x = anObject #user defined/built-in object
x = aMethod #user defined/built-in method
x = [1,"hello world",anObject,aMethod] #a list
x = (1,"hello world",anObject,aMethod) #a tuple
x = {1: 1, 2: "hello world", 3: anObject, 4: aMethod} #a dictionary
```

```
Note: Because Python is dynamically typed, the above 7 lines
would be valid as successive lines of a Python program

alternative dictionary implementation
x = {}
x[1] = 1
x[2] = "hello world"
x[3] = anObject
x[4] = aMethod
dictionary keys and values can be any valid Python data type,
including user defined.
```

3. Basic data manipulation

Strings and arrays can be sliced, which means you can extract the *n*th element or the *m*th to *n*th element.

```
python
>>>x = "hello world"
>>>y = ["h","e","l","l","o"," ","w","o","r","l","d"]
>>>print x[0]
h
>>>print y[0]
h
>>>print x[6:8]
wo
print y[6:8]
["w","o"]
```

Strings can be instantiated with % substitution

```
assume v1 is a string and v2 is an integer

In other languages:
myString = "hello" + v1 + "world" + str(v2)

same example using % substitution
myString = "hello %s world %d" % (v1, v2)
```

File handlers can be assigned using the built-in open function. Files can be opened in read (r), write (w), append (a), and binary (b) modes. r is the default mode.

```
fh = open("myfile.txt", "w") #open file for writing
fh = open("myfile.jpg", "rb") #open file for binary reading
fh = open("myfile.txt") #open file for reading
x = fh.read() #assign the entire contents of the file as a
    string to x
x = fh.readline() #assign the first line of the file as a
    string to x
x = fh.readlines() #assign the entire contents as a list of
    lines to x
fh.write("hello world") #write hello world to a file opened for
writing
```

Variables can be assigned a null/nil value with the keyword None.
Boolean operators: Python uses natural language : and, or, not, in.
Equality operators: == (equals), != (not equals), < (less than), > (greater than).

```
if (x != 2 or y <  = 3) and not (z == 4 or w in [1,2,3]):
```

Any variable can be used in a boolean statement. None, False, “”, 0, [], and () equate to false, any other value equates to true. The boolean types True and False can also be used.

```
if x:
    #do this code if x is true
else:
    #do this code if x is 0, "", [], (), None or False
```

4. Importing modules

External Python modules are imported using the `import` key word. Modules can be any other Python scripts written by anyone, defined modules that come with the language, or third-party modules downloaded and compiled into the Python program. Paths to Python modules in the first case can be defined either with the environment variable PYTHONPATH or at runtime using the path function of

the sys module (see below). Import statements can occur anywhere in Python code (including inside if ... else statements). There are two syntaxes for the import statement, which affect how members of the module are referenced. The following example assumes a file called mymodule.py, which is located in mydirectory/mypythonscripts.

```
import sys
sys.path.append("mydirectory/mypythonscripts")
import mymodule
x = mymodule.myfunction()

from mymodule import myfunction
x = myfunction()
```

You can also use * to reference everything in a module.

```
from mymodule import *
x = myfunction()
```

5. Control statements

Python uses if ... elif ... else, for and while for execution control.

```
for i in range(0,100,2): #for each number from 0 to 99,
increment by 2
for item in [item1, item2, item3]:

while x > 3:

if x > 3:
     #code block 1
elif x < 1:
     #code block 2
else:
     #code block 3
```

While and for loops can be exited with the `break` statement.

```
fh = open("myfile.txt")
while 1: #always true
     line = fh.readline()
     if not line: #line is an empty string if at end of file
          break #exit the while loop
```

Python uses `class` to define a new class, and `def` to define a function/subroutine or class method. Function parameters can be assigned a default value,

and those parameters with default values are optionally passed to the function when called.

```
python
>>>def increment(value, inc=1):
... return value + inc
...
>>> print increment(2)
3
>>>print increment(2,3)
5
>>>print increment(2, inc=3)
5
```

Class syntax is shown in the examples below. Python has no concept of class scope, such as private or protected, as found in languages such as Java. It is common to use single or double underscore prefixes as a "reminder" that a method has restricted scope, but this rule is not enforced by Python. You can overload built-in functions using methods both prefixed and suffixed by double underscores; examples are given below. Class methods always receive the keyword self as the first parameter; class attributes and methods are also referenced within the class by the self keyword.

Python is inherently reflective. Variables and functions can be accessed through the built-in locals() and globals() functions, which return dictionaries of all elements of the script. Classes can be manipulated with the hasattr, getattr, and setattr built-in methods. Examples of these are below.

6. Python bioinformatic resources

There is an open source project located at **www.biopython.org**, which contains a large number of script and class modules to handle biological data and which acts as a framework for building your own programs. A short example is shown in the examples section.

7. Examples

The following example scripts display the flexibility of the Python language for building useful code for biological data manipulation. The code can be cut and pasted into files, as indicated.

```
#--------------------cut & paste into a file called bioclasses.
py------
import string

class CodonTable:
```

```
    def __init__(self):
         self.codons = {}
         self.codons["AAA"] = "L"
         self.codons["AAC"] = "N"
         self.codons["GGG"] = "G"
         self.codons["TAA"] = "*"
    #table not finished due to space limitations


    def getResidue(self, codon):
         if self.codons.has_key(codon):
              return self.codons[codon]
         else:
       return "?"



class Writeable:
#a basic class which allows an object to be written to a file

     def write(self, fh, format, lineLength = -1):
     #method parameters:
     #fh : a file handle
     #format : a string representing the format
     #lineLength : optional integer indicating number of characters
     #on a line
         method = "_as%s%s" % (string.upper(format[0]),format[1:])
         #The method variable will be assigned the string:
         #_as + uppercase first letter of format + rest of format

         if hasattr(self, method):
         #if format == "fasta" then method == "asFasta"
         #hasattr will test for the existence of the asFasta method
              data = getattr(self, method)(lineLength)
              fh.write(data)
         else:
              print "Don't know how to generate %s format" % format



class Gene(Writeable):
#class Gene inherits from Writeable

    def __init__(self, name, seq):
    #constructor
         self.name = name
         self.seq = seq

         self.utr3 = None
         self.utr5 = None
         self.exons = []
         self.transcripts = []

    def set3utr(self, start, end):
    #set the start and stop positions of the 3' UTR
         self.utr3 = (start, end)
```

```
    def set5utr(self, start, end):
    #set the start and stop positions of the 5' UTR
         self.utr5 = (start, end)

    def setExon(self, start, end):
    #create a list of co-ordinates for exon boundaries
         self.exons.append((start, end))
         self.exons.sort()

    def setTranscript(self, exonList):
    #define a transcript as a list of exons
    #as set up in the method above
         self.transcripts.append(exonList)
         self.transcripts.sort()


    def getTranscript(self, index):
    #concatenate the exons into a single sequence
         exonList = self.transcripts[index]
         subSeq = ""
         for e in exonList:
              (start, end) = self.exons[e]
              subSeq = subSeq + self.seq[start:end]

         return subSeq

    def getPeptide(self, index, codonTable):
    #translate a transcript
    #index refers to the transcript
         tscript = self.getTranscript(index)
         peptide = ""
         for i in range(0, len(tscript), 3):
              codon = tscript[i:i+3]
              residue = codonTable.getResidue(codon)
              peptide = peptide + residue
         return peptide

    def _asFasta(self, lineLength):
    #return fasta format sequence with any line length
    #prefixed with a _ to indicate method should not be called
    #outside the class
        if lineLength == -1:
             result = ">%s\n%s\n" % (self.name, self.seq)
             #\n is a special character which denotes a line return
        else:
             result = ">%s\n" % self.name
             for i in range(0,len(self.seq), lineLength):
                  result = "%s%s\n" % (result,
self.seq[i:i+lineLength])
          return result

    def __getitem__(self,cmd):
    #this method overloads the slice operator of strings and arrays
    #It will return a subsequence as a Sequence object
    #if asked for a single element (e.g. Sequence[10])
```

```
    #cmd will equal the integer position of the base to be returned
    #else (e.g. Sequence[10:20]) cmd will be an object with start
    #and stop attributes

        if type(cmd) == type(1):
        #check if the cmd variable is an integer
             subName = '%s_%d' % (self.name, cmd)
             subSeq = self.seq[cmd]
        else:
             start = cmd.start
             stop = cmd.stop
             subName = '%s_%d_%d' % (self.name, start, stop)
             subSeq = self.seq[start:stop]
        result = Gene(subName, subSeq)
        return result
#----------------end file bioclasses.py---------------------



#------cut and paste into a file called biorun.py--------------------

#ideally, the information for a gene would be derived from
#a GenBank or similar record that has been parsed. For this
#example, the information will be artificially created


import bioclasses

c1 = "AAAAAAAAC"
c2 = "GGGAAAGGGAACAAAGGG"
c3 = "GGGAACAAAGGGAAAAACTAA"
#3 codons

i1 = "CTGCGCGCTAAGATCGCT"
i2 = "CGCTAGAGCTCGGGAATAGCGCTA"
#introns

seq = c1 + i1 + c2 + i2 + c3

gene = bioclasses.Gene("test", seq)

marker = 0
gene.setExon(marker, len(c1))
#This will be the 0th exon in the object exon list

marker = len(c1) + len(i1)
gene.setExon(marker, marker + len(c2))
#This will be the 1st exon in the object exon list

marker = marker + len(c2) + len(i2)
gene.setExon(marker, marker + len(c3))
#This will be the 2nd exon in the object exon list

gene.setTranscript((0,1))
#1st transcript comprises exons 0 and 1
gene.setTranscript((0,2))
```

```
gene.setTranscript((0,1,2))


cTable = bioclasses.CodonTable()

print gene.getPeptide(0, cTable)
print gene.getPeptide(1, cTable)
print gene.getPeptide(2, cTable)

#some examples of overloading, and reflection

sub1 = gene[10]          #use the __getitem__ method
sub2 = gene[10:]         #which overwrites the slice operator
sub3 = gene[10:20]
subList = [sub1, sub2, sub3]   #make a list of subsequences

fh = open("sequences.txt","w")  #open a file for writing
for s in subList:
     s.write(fh, "fasta",20)          #use the write method
                                #of the parent class

s.write(fh, "staden",20)
fh.close()

#-------------------end file biorun.py-------------------------
```

Running the biorun script (Python biorun.py from the command line) will produce the following output (from the print statements):

```
LLNGNLGLN*
LLNGLGNLG
LLNGLGNLGGNLGLN*
Don't know how to generate staden format
```

It will also produce a file called “sequences.txt,” which contains the following:

```
>test_10
T
>test_10_2147483647
TGCGCGCTAAGATCGCTGGG
AAAGGGAACAAAGGGCGCTA
GAGCTCGGGAATAGCGCTAG
GGAACAAAGGGAAAAACTAA
>test_10_20
TGCGCGCTAA
```

8. Biopython example

The following is paraphrased from **www.biopython.org/docs/tutorial/Tutorial004.html**

Assuming that you have downloaded and installed the Biopython modules from **www.biopython.org** and that you have a local copy of a blast search,

```
from Bio.Blast import NCBIStandalone
#import the relevant part of Biopython

blastFH = open('my_file_of_blast_output', 'r')
#create a file handle to your blast result

bParser = NCBIStandalone.BlastParser()
#instantiate a parser

bRecord = bParser.parse(blastFH)
#Create a blast record object from your file

#That's pretty much it. Now you can manipulate the
#record object to display the different sections
#of the blast file.

#for instance, to see all the alignments with Evalue < = 0.04...

E_VALUE_THRESH = 0.04

for alignment in bRecord.alignments:
    for hsp in alignment.hsps:
        if hsp.expect <  E_VALUE_THRESH:
           print '****Alignment****'
           print 'sequence:', alignment.title
           print 'length:', alignment.length
           print 'e value:', hsp.expect
           print hsp.query[0:75] + '...'
           print hsp.match[0:75] + '...'
           print hsp.sbjct[0:75] + '...'
```

Further reading

There is a wide range of books for Python, aimed at beginner, intermediate, and advanced users.

An extensive list can be found at **http://www.python.org/cgi-bin/moinmoin/PythonBooks**

A brief Perl tutorial for bioinformatics

Michael J. Moorhouse
Erasmus MC, Rotterdam, The Netherlands

1. Introduction

This tutorial gives an introduction to programming in the Perl programming language by covering the basic syntax and key concepts important in two examples of Perl. For a review of Perl and its use in the Bioinformatics and Bioscience, Article 104, **Perl in bioinformatics**, Volume 8 (*also see* Article 103, **Using the Python programming language for bioinformatics**, Volume 8). Presented first is an overview of Perl syntax and key concepts. More extensive documentation can be accessed by running "perldoc perl" with a default installation of the Perl or from the excellent on-line resource: **http://www.perldoc.com**.

2. Perl syntax and concepts

The conversion of the human readable source code, listed in (say) file foo.pl, to machine executable format is made by the Perl runtime system. This can be done in two ways:

1. Calling the Perl interpreter (e.g., /usr/bin/perl on some Unix systems) on the command line with the source code file name.
2. On Unix systems, setting the executable flag on the source code file and including a line starting with "#!" followed by the path to the Perl interpreter (e.g., #! /usr/bin/perl). Technically, Perl is interpreted language that is then executed in a runtime system rather than being compiled.

In Perl, there are three basic datatypes (see the "perlvar" manual page for more information on these):

1. "Scalars" that store numbers (integers or floating point), characters, strings or a combination of these. These are referred using a "$" symbol.
2. "Arrays of scalars", which are ordered sets of scalars variables. The array as a whole is referred to using the "@" symbol and the individual elements using the scalar "$" symbol (as these are scalars).

3. “Hash (or associative) arrays”, which store scalar values that are referred to using keys with no assumed order between the key-value pairings. The entire array is referred to using the “%” symbol and the values using the scalar “$”.

Owing to its history as a scripting/text processing language, Perl has very good file-handling capabilities. These are complemented by inbuilt POSIX compliant regular expressions that allow the flexible matching of text patterns.

It is very easy to run through all the files passed on the command line or the text stream passed via the standard input (STDIN) and search for occurrences in the string “kinase” is very easy:

```
while (<>) #Iterates through each line of the file / STDIN
      {
#'kinase' found in $_ (special variable meaning "current line")?
      if (/kinase/)
            {print "$_";}    #If so, print the line
      }                      #Otherwise, do nothing and load next line
```

Certain “metacharacters” can be used to match specific features of text, for example:

- “.” matches “any character” (apart from new line characters: “\n” for Unix, “\r\n” for Microsoft Windows or “\r” for Apple Macintosh).
- “ ^ ” means “the start of the string”.
- “*” means “zero or more characters”. A useful modifier here is “*?”, which means “match the shortest string possible rather than the longest”.
- “$” means “end of line” (or end of the string depending on the context).

A more complete list is giving the “perlre” manpage.

Regular expressions are also useful for “capturing” data using brackets “()” to store a particular section of text matched.

For example,

```
(my $GO_ID, my $Description, my $Type) = m/^(GO:.*?)\t(.*?)\t(.)/;
```

which when scanning the text

```
"GO:0000001     mitochondrion inheritance     P"
```

results in the text being split into its component parts and the substrings being placed into the correct scalar variables.

Substituting strings is also easy using the “s/ / /” construct, for example

```
$Test = "Bioinformatics";
$Test =~ s/informatics/science/;
$Test is now "Bioscience"
```

Also useful are standard solutions to common regex tasks that have been developed, as in the following example, which converts the string into lower case text with the start of each word in upper case (see "perlfaq4" for a fuller discussion or the "Text::Autoformat", which does better "title case" capitalization):

```
$Description =~ s/(\w+)/\u\L$1/g;
```

3. Program 1 – basic Perl

This example demonstrates the basic features and syntax of Perl: variable declaration, conditional tests, iterative loops, and the two types of array access (Hash and Index) (see Figure 1a).

When run, the program checks IDs supplied on the command line against the list downloaded from the Gene Ontology consortium (Ashburner *et al*., 2000; Harris *et al*., 2004), and prints out the human readable text for those recognized as valid. (*See also* Article 82, **The Gene Ontology project**, Volume 8, Article 83, **Ontologies for information retrieval**, Volume 8.) The output is shown in Figure 1(b).

4. Program 2 – graphical BioPerl

This example demonstrates the use of the Bio::Graphics module, which plots the positions of features in a section of DNA in the form of an image (Stajich *et al*., 2002). See Figure 2(b) for the image and Figure 2(a) for the source code.

As Perl has no native graphic drawing capabilities, Bio::Graphics uses the external gd-lib drawing library (written in "C" and coded by Tom Boutell, see **http://www.boutell.com/gd**) via the GD.pm Perl interface coded by Lincoln Stein (see **http://stein.cshl.org/WWW/software/GD/**). GD.pm is a generic image creation library that supplies a set of functions "graphics primitives" that can draw lines, arcs, rectangles, polygons, and other simple image manipulation tasks, harnessed by Bio::Graphics to produce a more complex display. This is as used in the "Ensembl" genome browser to visualize the location of "features" (genes, exons, mutation sites, etc.) (Birney *et al*., 2004).

For this example to run, you need the GD.pm module and the GD Library installed on your system along with the libpng and zlib "C" libraries needed by GD.

The biological data displayed in this example are the protein coding regions of entry "ISTN501" in the EMBL database (Brown *et al*., 1983; Nascimento and Chartone-Souza, 2003), see Table 1.

This contains the Tn501 transposon, which codes for the biological Mercury detoxification system (aka. the "Mer Operon"). This is studied in depth in "Bioinformatics, Biocomputing and Perl" written by Moorhouse and Barry.

The example demonstrates two important points:

1. The power of modules and how to call them.

```
#!/usr/bin/perl
#
use strict;     #Force declarations of variables (using 'my') before use

my $GO_List = shift;   #Recover list from 1st Command Line argument
unless (-e $ARGV[0]) {die "Cannot continue without file: '$ARGV[0]'\n"; }
$GO_List =~ s/ //g;    #Remove any spaces user might have added
my @GO_List = split (",",$GO_List);   #Split list in an index array '@GO_List' on commas

my %GO_Lookup_Table;   #Create hash array to store GO ID -> Text mappings
while (<>)                      #Cycle through remaining file:
        {                       #each line is loaded in '$_'
        if (/^!/) {next;}       #Skip all header lines (tests '$_' implicitly using regex)
#Use regex captured parts of line between '()' to variables $GO_ID, $Description and
$Type
        (my $GO_ID, my $Description, my $Type) = m/^(GO:.*?)\t(.*?)\t(.)/;
        unless ($Type eq "P") {next;}  #Check if GO Term is of type "biological process"
        $Description =~ s/(\w+)/\u\L$1/g;#Complex regex to capitalize text (see perlfaq4)
        $GO_Lookup_Table {$GO_ID} = $Description;    #Populate the hash array
        }
#Run through '@GO_List' one ID, '$Current_GO_ID', at a time:
foreach my $Current_GO_ID (@GO_List)
        {
        if (exists $GO_Lookup_Table{$Current_GO_ID}) #Was the ID one found in the file?
                {                               #Yes it was, so print out the term:
                print "$Current_GO_ID\t:\t'",$GO_Lookup_Table{$Current_GO_ID},"'\n"
                }
                else                            #No it was not, so note this:
                {print "'$Current_GO_ID'\t- NOT FOUND!\n";}
        }       #Move on to next @GO_ID / '$Current_GO_ID'
```

(a)

```
GO:0000025     :       'Maltose Catabolism'
GO:0000002     :       'Mitochondrial Genome Maintenance'
'GO:9999999'   - NOT FOUND!
```

(b)

Figure 1 (a) Source code of Program #1. This “looks-up” the textual description of GO ID. Syntax highlighting, based on the styles used in ‘nedit’ editor program has been used to make the code more readable. (b) Output for Program #1 as run in October 2004 (results may vary slightly due to continual updating of the GO)

```
#!/usr/bin/perl

# This is code example 1 in the Graphics-HOWTO
use strict;
use Bio::Graphics;
use Bio::SeqFeature::Generic;

my $panel = Bio::Graphics::Panel->new(-length => 8000,-width  => 800);
my $track = $panel->add_track(-glyph => 'generic',-label  => 1);

while (<>) { # read blast file
  chomp;
  next if /^\#/;  # ignore comments
  my($name,$score,$start,$end) = split /\t+/;
  my $feature = Bio::SeqFeature::Generic->new(-display_name=>$name,-score=>$score,-
start=>$start,-end=>$end);
  $track->add_feature($feature);
}
open OUT, ">panel.png";
print OUT $panel->png;
close OUT;
```

(a)



(b)

Figure 2 (a) Source code of Program #2. This is a fully functional program fragment: See text for description. Syntax highlighting as in the first program, in Figure 1(a). (b) Output of Program #2. This uses the Bio::Graphics 'BioPerl' module

Table 1 Input file/table for program #2. These data are the extracted protein coding regions for the 'Tn501' Operon

#EMBL Original File: 'ISTN501'			
#Gene	Score	Start	End
MerR	0	114	548
MerT	0	620	970
MerP	0	983	1258
MerA	0	1330	3015
MerD	0	3033	3398
MerE	0	3395	3631
ORF2	0	3628	4617
TnpR	0	4792	5352
TnpA	0	5356	8322

Drawing the graphic shown in Figure 2a is very easy because the complexity is contained in the Perl modules and the underlying libraries they call. The two lines:

```
use Bio::Graphics;
use Bio::SeqFeature::Generic;
```

instruct Perl to expect calls to subroutines and data objects contained in these modules. The "::" syntax denotes the organizational hierarchy of the module and allows either all the features of the module to be accessible (in the first case above) or just a subset (as in the second case).
The ultimate result is a simple interface that leaves the programmer free to concentrate on data manipulation rather than on rudimentary layout of the graphics primitives.

2. The object orientated syntax of Perl, for example:

```
my $feature = Bio::SeqFeature::Generic->new(-
display_name=>$name,-score=>$score, -start=>$start, -
end=>$end);
```

is of the generic type:

```
my Handle = Module::SubModule::Function -> new (hash array
containing parameter list);
```

often used by modules. If you wish to code extensively using other modules, such as GD.pm, you will need to understand the basics of his syntax, but with this, it is easy to get started.

The Bio::Graphics module is far more capable than presented here; see the online tutorial at **http://bioperl.org/HOWTOs/html/Graphics-HOWTO.html**.

5. Summary

Perl is widely used in Bioinformatics today because it is easy to use and it has a wide variety of extra code modules – many of which have been developed to solve Bioinformatics tasks. The appeal of Perl seems to be a combination of the ease of programming, excellent native support for text handling and the extensive stock of modules that exists. Perl really is a biological "gem".

Further reading

Moorhouse MJ and Barry P (2004) *Bioinformatics Biocomputing and Perl: An Introduction to Bioinformatics Computing Skills and Practice*, John Wiley & Sons: Chichester.

Stein LD (1998) *Official Guide to Programming with CGI.pm*, John Wiley & Sons: New York.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, *et al*. (2004) An overview of Ensembl. *Genome Research*, **14**, 925–928.

Brown NL, Ford SJ, Pridmore RD and Fritzinger DC (1983). Nucleotide sequence of a gene from the Pseudomonas transposon Tn501 encoding mercuric reductase. *Biochemistry*, **22**(17), 4089–4095.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*. (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32** Database issue, D258–D261.

Nascimento AM and Chartone-Souza E (2003) Operon mer: bacterial resistance to mercury and potential for bioremediation of contaminated environments. *Genetics and Molecular Research*, **2**, 92–101.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**, 1611–1618.

Tutorial

Grid technologies

Douglas W. O'Neal
Delaware Biotechnology Institute, Newark, DE, US

In their 1998 book *The Grid*: *Blueprint for a New Computing Infrastructure*, Foster and Kesselman gave an initial definition of a grid: "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities" (Foster and Kesselman, 1998). This definition is built on earlier applications in metacomputing and network-aware applications and is general enough to encompass several types of grids in existence today. The analogy commonly used to describe grid computing is to a power grid. When an appliance is plugged into a matching receptacle, it is understood that the power of the correct voltage will be immediately available. How that power is generated and delivered may not be known to the end user and the complexity of the system is irrelevant to the functioning of the appliance. This vision translates to a view of a virtual computer that provides unlimited capacity on demand without regard to the complexity of the underlying hardware.

Capacity can be measured in multiple ways and grids can be developed to match the requirements. The most common grids in existence today fall into the categories summarized below. These types are not strict definitions and any given grid may combine characteristics of several of these.

1. Computational grid

This is the most commonly considered grid, primarily using high-performance servers and/or clusters of systems to deliver raw computational power. A job may make use of this grid's resources in different ways depending on the job's algorithms. The easiest is to just have the grid find an available high-performance system on which to run the job instead of using the local resources. Second, if the job can be split into several independent pieces, then each piece can be sent to a different processor and the results merged after the last piece is finished. The third way is the most involved and is where the application is redesigned to work in parallel on multiple processors in the grid.

2. Data grid

This type of grid provides for the storage and retrieval of data in a secure and reliable manner. Data may be replicated across multiple sites, with access granted to other organizations. The grid manages the data storage, grants access to authorized users, and provides update procedures for multiple writers.

3. Communications grid

High-bandwidth communication within a grid is a necessary component for the utilization of other resources. For example, a job may be processing a large data set through multiple passes and the data set may not be local to the machine running the job. External communications to the Internet may be the primary purpose of a grid. Services such as search engines have large bandwidth and redundancy requirements that can be met through a grid with multiple external Internet connections.

4. Scavenging grid

This is a special case of a computational grid in which unused computing resources from desktop systems are harvested. The use of these machines may be limited to after business hours to minimize the impact on the users of the systems. SETI@Home (2006) and similar projects also scavenge central processing unit (CPU) cycles while a desktop system is otherwise idle.

For any grid, there are several basic requirements that must be met. All have a technological base but some also have political overtones. These core technologies include the following:

5. User authentication and authorization

The first steps in accessing resources on the grid are identifying yourself to the system and determining to which resources you are allowed access. The first of these is authentication, a process that needs to be made a single time per session and then propagated to any resource used in that session. The second is authorization – a more difficult issue. Given that organizations tend to be protective of the resources they contribute to the grid, the granting of these resources to others is a social, business, and, possibly, a legal issue.

6. Resource discovery

Services on the grid must not only be made available but their availability must be made known to the potential users of those services. Since systems are frequently entering or leaving the grid, this list of available resources must be managed dynamically. Usage of the resources needs to be tracked for both accounting purposes and for scheduling.

7. Resource management/scheduling

At the most basic level, this involves the actual allocation of resources to specific jobs. Allocation rules may take into account the user's authorization level, priorities given to the user, previous usage patterns, and resource requirements such as memory needed, expected run time, or a hardware platform. Scheduling may also be included in order to have the necessary resources available for a process that may need to run at a certain time or to be able to block off large amounts of resources.

8. Data management

In a grid, it is unlikely that data will be fully shared across all resources. Thus, a job must be able to locate the necessary data and possibly transfer a copy to local storage. In large grids, data redundancy may be necessary to ensure timely access.

9. Security

The requirements for security cut across all aspects of a grid. User authentication must be reliable, data confidentiality must be respected, local resources protected, etc. With such broad requirements, security must be designed into grid software from the beginning and not put in place after the other aspects are designed.

The translation of these requirements into a working computational environment is an ongoing project that has not been fully developed. The Global Grid Forum (GGF) (Global Grid Forum, 2006) is an international body of vendors, developers, and users working to define a set of standards for interoperable grid software. It is an open forum with participants from all aspects of grid development, and most vendors with grid products are members of the GGF. The reference implementation of the GGF standards is the Globus Toolkit (Globus, 2006). Globus is designed as a layered architecture where core low-level services are used to construct high-level global services. This modular design allows a wide variety of applications to use the services necessary to meet their particular needs.

The Globus implementation has evolved with changing standards. Now, at version 4, its programming model is based on the Open Grid Services Architecture (OGSA). OGSA uses Web services to make resources available to grid users. The benefits of this approach include the following:

- Well-known open standards such as Simple Object Access Protocol (SOAP) and Extensible Markup Language (XML) can be used to define and access grid services.
- Additional services can be integrated into the existing infrastructure in a consistent fashion.
- New resources in the grid can be identified and utilized in a standard fashion.
- Interoperability between grids is possible based on standard toolkits and mutual trust.

The Open Grid Services Interface (OGSI) specification is the implementation of the OGSA standard. The OGSI defines the interfaces and protocols used between services in a grid environment. The transparent interoperability provided by OGSI does come at a cost in performance in the current implementation. Certain vital operations such as time-critical data transfer may require the use of other protocols to meet their performance goals until faster implementations emerge.

As the *de facto* standard, the Globus Toolkit has been used by most large public grids. The Biomedical Informatics Research Network (BIRN) (Biomedical Informatics Research Network, 2006) is a consortium of 30 research sites at 21 universities and hospitals. Set up to encourage collaboration among the research institutions, it has used Globus to share data collections and analysis tools. Similarly, the Open Science Grid (2006), the TeraGrid (2006), and the Enabling Grids for E-sciencE (EGEE) Project (Enabling Grids for E-sciencE, 2006) each bring together researchers across multiple disciplines to share resources. The EGEE Grid is possibly the largest, with over 30 000 CPUs and 5 Petabytes of storage available to the average of 10 000 jobs running at any given time.

It is important to note that the Globus Toolkit is in a turnkey package, but is a collection of components used to develop a grid in a specific environment. Smaller organizations may find the task of building a custom grid daunting and there are several packages available to create a departmental grid easily. The open source package Grid Engine, (2006) provides a complete package for building and running a compute or scavenging grid and can be installed as a turnkey system. When coupled with Grid Engine Portal, the end users are presented with a web portal allowing them to choose resources, submit and monitor jobs, and collect results. The web portal uses XML-based configuration files to create customized job submission pages for each grid application. Use of the web portal is not necessary in a Grid Engine environment and advanced users may find it easier to use command-line utilities, but the portal hides the command-line interface from the users who do not know or wish to use the underlying interface.

Despite the successes of the grids mentioned above, grid technologies are still in their infancy. Applications that migrate well to current grids are mostly limited to life sciences and physical sciences, and computation platforms generally use the

Unix or Linux operating system. To expand past the present base, more resources will be linked by more and better networks. The inclusion of lower-end devices (Personal Digital Assistants (PDAs), cell phones, sensor networks) will create grids of greater heterogeneity with large variations in the performance of individual components. The end result will be moving closer to the goals of autonomic computing and self-healing networks. Programming models will also change to take advantage of these resources, with interfaces developed to cope with resource discovery and selection. The end result can merge the incoming data flow from sensor networks or experimental instrumentation with the compute power of a large grid to produce real-time analysis of data on a scale not seen before.

Further reading

Berlich R (2004) Linking data networks. *Linux Magazine*, **44**: 48–51.

Berman F, Fox G and Hey T (eds.) (2003) *Grid Computing: Making the Global Infrastructure a Reality*, John Wiley & Sons Ltd: Chichester, West Sussex.

Joseph J and Fellenstein C (2004) *Grid Computing*, Prentice Hall: Upper Saddle River, New Jersey.

References

Biomedical Informatics Research Network (2006) **http://www.nbirn.net**.

Enabling Grids for E-sciencE (2006) **http://www.eu-egee.org**.

Foster I, Kesselman C (eds.) (1998) *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers: San Francisco.

Global Grid Forum (2006) **http://www.ggf.org**.

Globus (2006) Project, **http://www.globus.org**.

Grid Engine (2006) Project Home, **http://gridengine.sunsource.net**.

Open Science Grid (2006) **http://www.opensciencegrid.org**.

SETI@Home (2006) **http://setiathome.ssl.berkeley.edu**.

TeraGrid (2006) **http://www.teragrid.org**.

Tutorial

Attacking performance bottlenecks

Ruud van der Pas
Sun Microsystems, Inc., Amersfoort, The Netherlands

1. Introduction

More than ever, application tuning is of paramount importance to reduce the time to process the data. This is increasingly the case in bioinformatics, as witnessed in the parallelization of NCBI BLAST (**http://mpiblast.lanl.gov/**).

Typically, the data sets have grown in size over time. In many cases, the algorithms used to process the data increase nonlinearly with respect to the execution time.

Meanwhile, a quiet revolution is taking place in microprocessor technology. Instead of focusing on improving the performance of a single, monolithic processor, attention has shifted to multicore technology.

With this technology, a single processor turns into a relatively small parallel computer. This is done at the expense of squeezing everything out of the single core performance. Typically, the clock speed increase is no longer as aggressive as it used to be. Caches are smaller, and there is a certain level of cache sharing too.

This paper is organized as follows. In Section 2, the recent hardware trends are presented and briefly discussed. In Sections 3, 4 and 5, the impact of this on the developer is discussed. Section 6 covers the importance of tools in this area. A summary of a selection of the tools Sun Microsystems has to offer is given in Section 7.

2. Technology – recent history

The performance of today's computers relies heavily on the use of fast buffer memory, called a *cache*. This has been the case for quite a number of years now.

Recently, the so-called *multicore* processor designs have emerged. This not only provides parallelism on the chip but it also adds another dimension to the structure of the cache subsystem.

In this section, these new concepts, to the extent relevant for application performance and developers, are briefly introduced and discussed.

2.1. Cache memories

The speed of the main memory system has not kept up with the increase in processor clock speed. Microprocessor technology follows Moore's law

(**http://www.intel.com/technology/silicon/mooreslaw/** The original paper can be downloaded from **ftp://download.intel.com/research/silicon/moorespaper.pdf**), which states that the number of transistors on a chip roughly doubles every 18 months. Although not automatically implied, for a long time this has translated to a proportional increase in the clock speed of the processor.

This law is however not applicable to the speed of main memory. The rate of increase is less, giving rise to an ever bigger performance bottleneck.

This discrepancy has existed for many years now and has been addressed by computer architects primarily with the use of cache memory. A cache is nothing more, or less, than a relatively small buffer memory, substantially faster than main memory.

Modern systems have a hierarchy of caches, logically placed between the CPU and main memory. The size and cache architecture details are system specific, but the rule is easy. The closer the cache is to the CPU (measured in terms of access time, not physical distance per se), the smaller it tends to be.

Caches are typically used to buffer three types of information needed to execute an application: data, instructions, and address mapping structures. The latter is generally referred to as the translation lookaside buffer (TLB) cache.

The number of caches, functionality, respective sizes, and internal architecture details are decided by the microprocessor design team. These choices are driven by technology, power requirements, target market(s), and time to market.

2.2. Multicore processor designs

Not long ago, microprocessors were only able to execute one single thread of execution at any point in time. Through context switching, it appeared as if the system was executing more than one task simultaneously, but this was not the case in reality.

With the advent of multicore designs, this scenario has changed.

Basically, this technological trend is driven by the increased complexity, power consumption, and heat dissipation of the conventional microprocessor designs. Caches only help to a limited extent, keeping the processor busy doing useful work. In particular, because of the widening memory gap, processors are increasingly burning more power doing nothing, waiting for the data to arrive.

The increased complexity of these designs magnifies this problem. For a number of years, highly specialized features have been added. Only a few applications benefit from these features, but meanwhile both the processor design time as well as the demand on power have gone up.

To counter this problem, multiple *cores* were placed on a single processor. Although the word “core” is not well defined, and vast differences exist between various designs, a fairly accurate way to view it is as a relatively simpler compute engine, capable of executing instructions, producing results, and having its own state information, like a program counter and registers.

The advantages of this multicore design are an impressive aggregate performance from a single processor, reduced or stationary power consumption, and a reduced time to market.

Now, all microprocessor vendors have realized these advantages. Either they are already shipping multicore designs, or will do so in the near future.

There is also a price to pay though. When it comes to performance, each core is relatively weak, compared to a traditional, but fully fledged, single monolithic design. Not only is the clock speed relatively lower, the caches are also smaller.

Some of these designs also use a combination of caches *shared* between the cores, together with caches that are private to the core.

For the purpose of this introductory paper, this aspect will be ignored. All that is relevant in as far as the memory system goes is the existence of one or more caches, each with potentially different sizes and access times, but always significantly faster than main memory.

3. Why can't I simply wait for faster processors?

Many do not consider tuning the performance of an application as an efficient investment of time, and therefore money. Simply waiting for faster processors to appear on the market is an easier and cheaper way to realize the performance level desired. (Of course, this does not hold for a relatively small category of power users who need to squeeze every single drop of performance out of the system.)

Even though the clock speed of processors has steadily increased and is well into the gigahertz range these days, fewer and fewer users have enjoyed a corresponding increase in performance when upgrading their system. The widening memory gap is a big contributor to this anomaly.

Given the multicore trend, one should no longer expect to see dramatic increases in the clock speed of the individual cores. These designs favor the overall workload by getting more work done in the same amount of time, but the performance of an individual application does not increase as dramatically as before.

The method to improve the performance of a single application is to exploit the parallelism inherent in these multicore designs (*"A Fundamental Turn Toward Concurrency in Software"*, Herb Sutter, **http://www.ddj.com**).

4. Five different ways to optimize an application

How does the application developer take into account all of the above-mentioned hardware performance features *and* become prepared for future developments in technology?

The answer is to leverage and use the possibilities listed in the following text, preferably in the order of appearance.

1. *Operating system features* Each operating system has a specific set of performance-oriented features that are worth exploring and, where applicable, using.

An operating system has many parameters and comes with several features that affect performance. These have to be set and chosen such that a wide range of applications and tools perform well enough, even under varying operational conditions.

These choices need not be optimal for a single program or a limited set of programs. Knowing the characteristics of a specific workload helps in identifying different settings and selecting special features that improve performance.

2. *Optimized libraries* Through the operating system and externally provided optimized libraries, the developer is not only able to leverage the tuning investment put in by others but is also automatically able to derive benefits from specific performance features in case the software executes on different hardware that runs the same underlying software layer.
3. *Compiler features* Today's compilers are powerful, but also complex, tools. A rich set of options is available to the user, but, just as with an operating system, the default settings have been chosen to get good performance across a wide variety of applications.

   If one knows the runtime behavior of the application, the use of specific additional options, or change in the settings of those already used, can make all the difference when it comes to performance. It is not uncommon to obtain a speedup of a factor of 5–10 when switching on the (right) optimization options on the compiler.

   In general, it is important to have a good understanding of what modern compilers are capable of. In some cases, developers are overly pessimistic, or optimistic, as to what a compiler can do when it comes to automatically restructuring the source code to improve performance.

   It is difficult to give general advice on this aspect. Some compilers are smarter and offer more features than others. It is a good investment in time to study the documentation that comes with the compiler.
4. *Source code changes* The effort spent on this can be unlimited, but we have found that quite often there is low hanging fruit in many applications.

   By addressing these, an impressive performance gain with a relatively modest investment can be achieved.

   To be honest, this need not always be the case, though. Some modifications can have a profound impact. If, for example, the way the data is organized in memory has to be changed, the definition and usage of the data structure(s) involved are affected throughout the application. The reward is usually not only high, but tends to also have a positive impact on other systems besides the target hardware.

   Finding the balance by writing efficient and readable code that can still be maintained is key. This is also related to a good understanding of the compiler one is using.

   What complicates the matter is that this is somewhat of a moving target as well. As time goes by, compiler technology improves. Some compilers implement an impressive set of transformations and optimizations. In particular, one should be careful while manually implementing low-level optimizations. These may not only jeopardize performance in a future release of the compiler, but could also backfire if used on a different platform.

5. *Parallelization* With parallelization, sometimes also referred to as *multithreading*, multiple processors (Nowadays, these could be cores too.) are used to execute a single application. By assigning different tasks to the processors, the turnaround time is reduced. Ideally, this decrease in execution time is proportional to the number of processors used, or phrased differently, if P processors are used, the program finishes in $1/P$ of the original time.
   This is the ultimate method to increase performance. The downside is that it could be a relatively time-consuming process to identify and implement the parallelism.
   To a certain extent, the effort needed to parallelize an application also depends on the programming model selected. The intrinsic parallelization potential within the application plays a role too. Some programs are easier to parallelize than others.

To wrap this up, it is important to realize that the preceding five steps are best addressed in the order they are presented.

The first two phases (operating system features and optimized libraries) are the lowest hanging fruit. Merely by taking advantage of certain features and optimization efforts put in by others, a significant performance increase is already realized at virtually no cost.

The third step (compiler features) requires some more work, but it does not go beyond reading the compiler documentation and possibly conducting some additional experiments with various options that affect performance. This could still be considered to be low hanging fruit.

Step four (source code changes) goes beyond that and should only be considered once the other three options have been explored in detail. How much effort is to be put into implementing source code changes to improve the performance depends on various factors, including the amount of time available for the task at hand.

The use of high-quality tools to guide this effort is highly recommended. Further information on this can be found in Section 6.

In view of the last step (parallelization), the main focus of these sequential, single thread, tuning efforts should be on the memory behavior of the application.

In particular, the caches should be used in the best possible way. In addition to the reduced execution time of the application, the parallel performance, as a function of the number of threads, also benefits from this kind of optimization. This is in particular true on a multicore processor, or shared memory system with a shared interconnect.

Because of increased use of the cache(s), the number of memory transactions is reduced. This in turn implies that there is less traffic per processor on the interconnect. Therefore, the available total bandwidth can be shared by more processors, or cores in case of a multicore design.

5. Five parallel programming models

In a way, there is no escape for those who need to squeeze the most out of a multicore design in order to reduce the turnaround time of a single application.

Parallelization is the way to accomplish this and those who start today have a head start. We will not make things any prettier than they are: parallelizing an application may be hard work. The reward is high however, and a careful choice of programming model and software development environment can greatly help to ease the task.

Many parallel programming paradigms are available, each with its own set of pros and cons. Some of these are very briefly summarized in this section. In Section 6 an overview of the recommended software environment is given.

1. *Automatic parallelization* To start with, certain compilers support automatic parallelization, or "autopar" for short. Although not a programming model in the strict sense, it is mentioned here, because it is worth exploring.
   Through a compiler option, the user requests the compiler to identify portions of the application that can be executed in parallel. If such occurrences are found, the compiler also generates the corresponding runtime infrastructure. Typically, this is in the form of calls to a multithreading library.
   The resulting binary that is generated this way can be used to execute on a shared memory and/or multicore system. It cannot be run on a cluster of systems that do not share memory.
   Success or failure of automatic parallelization depends on the programming language used, the application (area), the coding style, and the quality of the compiler, in particular, the dependence analysis component. The mileage will vary, but it is certainly recommended that this feature be tried, in case it is supported by the compiler at hand.
   The alternative approach is called *explicit parallelization*. In this case, the developer is fully in charge and decides what parts of the application are to be parallelized and how the parallelism is to be implemented. (On some systems both autopar and certain explicit parallel programming models can be combined.)
2. *POSIX threads* This powerful model for C and C++ supports parallelization within one address space. In other words, it is not possible to run an application parallelized with this model over different physical systems that do not share the memory.
   Through explicit insertion of function calls, the programmer implements and controls the parallelism.
   The advantage is that this threading model has been standardized. A disadvantage is that it is fairly low level: there are more details to worry about and the increase in the number of source lines is significant.
3. *Java threads* This model is similar to POSIX threads, but specific to the Java programming language. It essentially has the same pros and cons as POSIX threads.
4. *MPI, the message passing interface* There are specifications for C/C++ and Fortran in this programming model (**http://www-unix.mcs.anl.gov/mpi/standard.html**). The distributed memory model is supported: each process has exclusive access to private memory only. The other message passing interface (MPI) threads are not able to access it. (This is true even if multiple MPI processes are executed on a single shared memory system.) Information is exchanged and data shared by sending and receiving messages.

This model provides for great flexibility with respect to the execution environment. An application parallelized with MPI can be run on any cluster of computers that supports an MPI runtime environment.

The downside is that almost the entire burden is on the developer. All of the data exchange has to be put in by the programmer through the insertion of explicit, MPI- specific function calls to control the parallelism and exchange information by sending and receiving messages.

Moreover, care needs to be taken that the time spent in the communication part of the application is (significantly) less than the computational time. For example, it is often more favorable to combine several smaller messages into one large packet.

MPI is not an official standard, but the various implementations available adhere very well to the specifications. Therefore, portability is generally not an issue.

5. *OpenMP* The use of this shared memory programming model is on the rise. The specifications for C/C++/Fortran can be downloaded from (**http://www.openmp.org**).

   In contrast with the models discussed earlier, OpenMP provides a higher level interface. Through the so-called directives, the developer implements and controls the parallelism. In addition to this, runtime functions to query and change the execution environment are available. An example is the `omp_get_num_threads()` function call to change the number of threads used.

   In C/C++, a directive is a pragma with an OpenMP-specific syntax, for example, `#pragma omp parallel for` to parallelize a for-loop.

   In Fortran, directives use a specific language comment string, such as `!$omp parallel do` to parallelize a do-loop.

   This approach ensures portability, as a non-OpenMP compiler simply ignores the pragmas or comments, whereas an OpenMP compiler triggers on the keywords and translates these into the appropriate parallel infrastructure, typically calls to a lower-level parallelization library.

   OpenMP has several advantages over other programming models. The application can be parallelized incrementally. This also allows for incremental testing.

   Another useful feature is that the OpenMP application can be written such that the sequential version is preserved. This is convenient in case a bug is revealed. By not compiling specific parts of the source for OpenMP, the directives are ignored and deactivated/activated. This not only speeds up finding the root cause of the problem, but it can also be used as a workaround if time to market is important and troubleshooting is not feasible before releasing the application.

   Last, but not least, OpenMP compilers can assist the user too. By issuing warning messages, for example, the compiler may point the developer to a possible error. This is much harder, if not impossible, to do in other explicit parallel programming models.

The choice of programming model depends on several important factors. It is therefore hard to give recommendations, but we think that the OpenMP

programming model offers several attractive features justifying a serious look into it as a model of choice.

Thanks to the multicore trend, hardware availability is increasingly less of an issue. In the not too distant future, all new computers will be equipped with these kinds of processors, turning the system into a shared memory parallel computer. This is even true for laptops.

The programming benefits that OpenMP offers make it suitable for the beginner as well as for the more advanced developer.

6. The importance of application development tools

Without the right tools, it is very hard to develop an efficiently performing application. (Of course, correctness is always an issue; in case parallelism is considered, this is extra hard.)

At the tools level, application tuning starts with the compiler. A good compiler takes advantage of the features the hardware offers. Exploiting the cache hierarchy is a case in point, but there are many other optimizations an advanced compiler can perform. Ideally, this can be achieved with a minimal effort by the user.

This is the starting point only though. Application tuning has to be a guided effort, focusing on those parts where the application spends most of the time.

A performance analysis tool is extremely useful to identify such hot spots.

An important aspect to consider is the level of detail the tool provides. Knowing only the most expensive function(s) is usually not sufficient. Especially if the function is complicated, source line level performance information is needed to identify the time-consuming parts within the function. In some cases, there is even a need to drill deeper and find the most expensive instruction(s).

Modern microprocessors offer hardware performance counters that give low-level information on the various activities within the processor.

With this feature, one can, for example, measure how often the application accesses the cache(s) and the number of times the data requested is in the cache(s). The ratio of these two numbers is called the *cache hit rate* and is an indicator of how well behaved the application is from the point of view of a cache memory. In addition to cache-related activities, other events too can be measured.

Although hardware counters are useful, it is not always easy to access them and gather the data. A tool to assist the user with this is a must have for those interested in performance tuning.

Parallel processing adds another dimension to performance analysis. In this case, it is important to know how the various processors perform relative to each other. For example, are they all marching along in an efficient and synchronized way, or does one processor need more time than the others? In case of the latter, the parallel performance is most likely not optimal.

A tool to identify these kinds of bottlenecks is invaluable to improve the efficiency of a parallel application.

Last, but not least, ensuring correct execution of a parallel program is nontrivial.

In case of an MPI application, for example, the exchange of messages might be incorrect, giving rise to wrong results. Another source of errors is a missing, or

incorrect, synchronization operation. These and other errors are bugs causing the parallel application to produce wrong results or hang.

What makes it extra hard is that some of these problems only manifest themselves on a specific processor configuration. They could even be dependent on the actual load on the system and interconnect.

Troubleshooting these types of errors requires a debugger specifically developed for MPI applications. Examples are TotalView (**http://www.etnus.com**) and DDT (**http://www.allinea.com**), but these are not the only MPI debuggers on the market.

The shared memory programming model comes with its own set of possible bugs. A so-called data race is one of the most difficult errors to find.

Simply stated, with a data race, the update of a shared variable is not well protected, for example, when different processors simultaneously update the same shared variable with a different value.

This gives rise to unpredictable results. The variable could basically take any value, depending on the number of processors used and the order in which the write operations to memory are issued.

The behavior of a parallel application that has a data race bug is unpredictable. Even if the same number of processors are used, the error may or may not appear. Therefore, extensive testing may not reveal a data race and a tool to detect these is of great importance when developing a shared memory parallel application.

7. Tools by Sun Microsystems

In this section, we briefly discuss some of the Sun Microsystems software products that are available to assist the user in developing an application for multicore processors.

The Solaris Operating System (**http://www.sun.com/software/solaris**) offers many performance-related features that are worth exploring. For example, the DTrace tool that is part of Solaris 10 is very helpful to diagnose and analyze performance bottlenecks at the system level in particular.

The Sun HPC Clustertools product provides an efficient implementation of the MPI-2 specification (**http://www.sun.com/products-n-solutions/hardware/docs/Software/Sun_HPC_ClusterTools_Software**).

The Sun Studio Compilers and Tools (**http://developers.sun.com/prodtech/cc**) suite offers a comprehensive set of tools to assist the developer with the development and optimization of sequential as well as shared memory parallel applications.

The Sun compilers implement state-of-the-art optimizations, support automatic parallelization, and implement the most recent, 2.5, OpenMP specification.

The compilers have extensive support for OpenMP. Runtime performance optimizations and debugging features, as well as options to assist the user during the development cycle are provided.

An Integrated Development Environment (IDE) is part of Sun Studio as well. With that a debugger, editors, and so on, are available. It also includes the Sun Studio Performance Analyzer (**http://docs.sun.com/app/docs/doc/819-3687**), a performance analysis tool for applications written in C, C++, Fortran or Java. At

the parallel programming level, POSIX threads, Solaris threads, Java threads, MPI, automatic parallelization, and OpenMP are supported.

This analyzer presents performance information at the function, source line, and instruction level. Easy access to hardware event counters is also supported.

The Sun Studio Thread Analyzer is a tool that helps in finding data race conditions in a shared memory parallel application. It is available for download as part of the Sun Studio Express program (**http://developers.sun.com/prodtech/cc/downloads/express.jsp**).

8. Conclusions

With the advent of multicore processors, the developer is faced with two challenges. The single core performance will no longer improve as dramatically as it has done in the past. Secondly, parallelization is the way to exploit the computational power these architectures offer.

The OpenMP programming model in particular is very suitable for taking advantage of multicore designs.

Good tools are an absolute must when it comes to developing or adapting applications to this new paradigm.

Acknowledgments

The author is indebted to Mark Woodyard and Partha Tirumalai at Sun Microsystems for their feedback on an earlier version of this paper.

Glossary Terms

Glossary compiled by Clare E. Sansom

School of Crystallography, Birkbeck College, London, UK and freelance bioinformatics consultant and science writer

Accessibility The amount by which an amino acid within a protein structure is exposed on the surface of the protein and thus exposed to the polar solvent. Numerical values for accessibility can be used in protein structure prediction; polar and charged amino acids most often have high accessibility values.

Accession Number A number (most often an alphanumeric) that is assigned to an data entity – for example, a gene or protein sequence – when it is added to a database. The accession number is one of the primary keys for database searching. Each database uses a different series of accession numbers. Cross references to accession numbers are used to link between databases.

Adenovirus A DNA virus with a core made of DNA/protein and a capsid composed of 252 capsomers. The double-stranded DNA genome is about 36-kb long and contains inverted terminal repeats at its ends. Adenoviruses are often used as vectors in gene therapy. They have many advantages for this, among them that it is possible to delete large parts of the viral genome without destroying viral function.

Aequorin A bioluminescent protein that is used in functional proteomics as a sensitive assay for calcium and that does not disrupt cell function. A blue light is produced when calcium ions bind to a complex of aequorin protein with molecular oxygen and coelenterazine. Aequorin was originally obtained from the jellyfish Aequorea victoria, but a recombinant form is now available.

AI *See* Artificial Intelligence

Algorithm A mathematical method of data analysis that is (usually) progammed into a computer and that can be proven to produce the solution desired. Examples in bioinformatics include the Smith–Waterman and Needleman–Wunsch algorithms for sequence alignment. There is often more than one possible algorithm to produce a given solution; fast and elegant solutions are preferred.

Algorithm, greedy *See* Greedy Algorithm

Alignment An arrangement of two or more protein or nucleic acid sequences to maximize the number of matches and, with protein sequences, the number of near matches. Alignment algorithms are some of the most important and commonly

used in sequence analysis. A number of more or less rigorous algorithms for local and global alignment are widely available.

Alignment, Gapped *See* Gapped Alignment

Alignment, Global *See* Global Alignment

Alignment, Local *See* Local Alignment

Alignment, multiple *See* Multiple Alignment

Allele One of two or more differing forms of a gene occupying the same locus on the same chromosome. An allele may differ from all other alleles of that particular gene at one or more mutation sites; each gene may have up to 1000 different potential sites of mutation. If the only mutations are silent (leading to no change in the amino acid), the phenotypes of two or more alleles may be the same.
Source: Kahl, G, The Dictionary of Gene Technology (Wiley-VCH, 2001).

Allelomorph *See* Allele

Alpha Bundle A group of alpha helices clustered together to form a stable unit within a protein structure. An alpha bundle may form all or part of a protein domain.

Alpha Helix One of the two main types of secondary structure found in proteins. An alpha helix is a tightly coiled protein conformation with 3.6 residues per helical turn and a rise of 1.5 Angstroms per residue. The carbonyl group of residue "n" in the helix is hydrogen bonded to the amino group of residue "$n+4$". Alpha helices are extremely stable and are found in almost all proteins.

Alternative Splicing The ligation of exons from a gene to form a different mRNA, and thence a protein with a different sequence (and potentially a different structure and/or function) from the conventional one. Some exons may be left out of the mRNA, or the exons may be spliced in a nonconventional order. This is one of the mechanisms for increasing the number of protein products from a genome of a given size, and it is more common in higher organisms.

Amphipathic A molecule is defined as amphipathic or amphiphilic if it has one face that is much more hydrophobic than the other. Amphipathic helices are present both on the surfaces of soluble proteins, with the hydrophobic face pointing into the protein centre, and in transmembrane helix bundles, with the hydrophobic face pointing "out" toward the membrane phosopholipids. This feature can be used to predict helix locations and orientations.

Amphiphilic *See* Amphipathic

a.m.u., Dalton *See* Atomic Mass Unit

Analog Proteins that have evolved separately to perform similar functions, through the process of convergent evolution, are known as analogs. The folds of protein analogs may, but will not necessarily, be similar. The alpha/beta barrel (TIM barrel) proteins, which are all enzymes, are examples of protein analogs with similar folds; the serine proteases trypsin and subtilisin have closer functions but completely different folds. It is much harder to model a protein structure from an analog than from a known homolog.

Anaphase In the cell cycle, the phase between metaphase and telophase during which the daughter chromosomes are drawn toward either end of the dividing cell by the microtubules that are attached to the chromosome centromeres. Chromosome separation errors during both meiosis and mitosis are often picked up at this stage, blocking further cell division. However, these errors may not be detected during meiosis in the female, so most human trisomies result from nondisjunction errors arising in the egg cell.

Aneuploid A polyploid cell is defined as aneuploid if its chromosome number is not an exact multiple of the haploid number, caused by an error in mitosis: an aneuploid individual is one with a countable number of aneuploid cells. Individuals suffering from trisomies, most commonly trisomy 21, are aneuploid. The normal state is known as euploidy.

Annealing In the polymerase chain reaction (PCR) and similar technologies, the word annealing is used to mean the initial attachment of a complementary oliogonucleotide primer to a DNA or RNA sequence prior to the start of the reaction. It is not to be confused with the molecular dynamics technique of simulated annealing, used, for example, in ab initio protein structure prediction.

Annotation The process of determining the features encoded by a genome sequence, and marking the genome sequence with those features in order. Genome annotations include the function of encoded genes, the position of introns in eukaryotic genes, recognized regulatory sequences, and statistical measures such as CG content. One example of a widely used annotation tool is Artemis, which is freely available and widely used for prokaryotic and simple eukaryotic genomes.

Antibody An important part of the mammalian immune system; a serum protein that is secreted by plasma cells after contact with a foreign molecule (antigen). Antibody-antigen binding is a signal to other components of the immune system that the antigen is a foreign substance. Potentially toxic proteins are precipitated and removed from solution, whereas bacteria are agglutinated. There are several varieties of antibody, composed of differing numbers and arrangements of immunoglobulin domains; the most common is the IgG, with four such domains.

Antigen Any molecule, or part of a molecule, that is recognized and bound by an antibody (immunoglobulin) as part of the mammalian immune response. Antigens may be small organic molecules or loops on the surface of proteins. Identification of

potential antigenic regions in proteins is a useful bioinformatics exercise in vaccine design.

Apoptosis In short, apoptosis is programmed cell death, and it is common in multicellular organisms. In apoptosis, cells die in a controlled, regulated fashion, in response to a complex series of stimuli. Thus, the cells play an active part in their own death (which is why apoptosis is sometimes referred to as cell suicide). Unlike necrosis, it is an essential part of the development of all organisms.

Archaea A third class of organisms, distinct from both the bacteria and the eukaryotes, originally distinguished from the bacteria by an analysis of the evolution of rRNA structure. Archaea are now thought to be the first ancestors of eukaryotes. They are, however, prokaryotes, as they are single-celled organisms with no nucleus. Extremophiles, which thrive in "extreme" conditions such as high temperature or high salinity, are archaea.

Artificial Intelligence A type of algorithm that depends on modeling tasks and processes that are normally associated with the need for human intelligence. It usually involves encoding facts from an area of human knowledge into a complex set of rules and applying these to a problem. Artificial intelligence algorithms were particularly popular in some areas of bioinformatics during the 1970s and 1980s. They are less popular today, but they are still occasionally used.

Association Mapping An approach to genetic mapping involving the testing for functional polymorphisms or mutations using genetic markers that are in linkage disequilibrium (LD) with the mutation under test. Investigations typically compare sets of cases and controls, although other methodologies are available. Association mapping is more sensitive than straightforward linkage mapping in detecting small and moderate genetic effects involved in common, multigenic disorders. It is not yet possible at the whole-genome level.

Association Study A type of study in medical genetics in which the association of given genetic patterns (i.e., SNP markers for particular genes) with given phenotypes, most often diseases, is studied. If the marker is found significantly more frequently in the cases than in the controls, it indicates an association between the disease or trait and the marker (and therefore the gene) under study.

Atomic Mass Unit A unit of mass that is used to measure atomic and molecular masses, defined as 1/12 of the mass of one atom of the isotope Carbon-12. It is (approximately) equal to $1.66053886 \times 10-27$ kg. In biochemistry and molecular biology, the term "dalton" is most often used as a synonym. However, atomic mass unit (or a.m.u.) is used to refer to masses, and errors in masses, of peptide ions measured by mass spectrometry in proteomics experiments.

Autosome The autosomes are those chromosomes that do not determine the sex of an individual: for example, chromosomes 1-22 of the human genome are

autosomes. Traits that are determined by genes on the autosomes are inherited in an autosomal fashion; they are not sex specific.

Autozygosity Mapping A mapping strategy for identifying the loci and identity of genes involved in autosomal recessive (simple Mendelian) diseases. The allele studied is not only homozygous but identical by descent; it is often used with consanguineous families. The technique was discovered in theory in the 1950s but not exploited until the late 1980s.

BAC A 6.5 kb bacterial cloning vector, based on a single-copy F-factor of *E. coli*, that allows the cloning of DNA fragments of greater than 300 kb (although by no means all cloned fragments are of this size). The BAC is composed of the *E. coli* plasmid pMBO 131, carrying a chloramphenicol resistance gene, HindIII an BamH1 cloning sites, sites for rare cutters and a bacteriophage lambda cos N and lox P site.
Source: Kahl, G, The Dictionary of Gene Technology.

Backcross A type of genetics experiment in which a hybrid individual animal is crossed with one of its (typically inbred) parents, or an animal bred through such an experiment. The mouse is by far the animal that has been most widely studied in this way; laboratories such as the Jackson Lab hold panels of DNA samples from mouse backcrosses.

Bacterial Artificial Chromosome *See* BAC

Bacteriophage *See* Phage

Balancing Selection *See* Heterozygote Advantage

Basic Local Alignment Search Tool *See* BLAST

Beta Propeller A fairly common type of all-beta protein fold, found in, for example, the influenza virus surface protein, neuraminidase. A large number of beta strands are arranged into a number (generally 4–6) of consecutive motifs, each made up of four short antiparallel beta strands. These form the "blades" of the propeller and are arranged in a circular structure, with the active or binding site in the center.

Beta Sheet One of the two main types of secondary structure found in proteins. A beta sheet consists of two or, much more often, more extended beta strands, held together through a network of main chain–main chain hydrogen bonds between adjacent strands. Adjacent strands may be parallel or antiparallel, and the hydrogen bonding patterns between the different types are different. Whole sheets may be parallel, antiparallel, or mixed. At least one sheet is found in the majority of proteins.

Beta Trefoil A type of all-beta protein fold, found in, for example, cytokines, lectins and agglutinin. It contains a closed beta barrel and a hairpin triplet, and has

internal threefold symmetry. The beta trefoil fold is defined as Architecture 2.80 in the CATH database, and it has only one Topology sublevel (2.80.10 Trefoil).

Biallelic A mutation, or a marker, is described as biallelic if the gene involved has only two different forms. Single nucleotide polymorphisms are generally biallelic, in that only two amino acids are commonly formed at the SNP position. In contrast, microsatellites and other tandem repeats are not biallelic because they may have many different lengths.

Bilayer A double layer. Most often used to describe the double layer of phospholipids that makes up the membrane surrounding all cells and organelles. The hydrophilic head groups and parallel hydrophobic tails of the phospholipids in a membrane bilayer strongly affect the amino acid composition of proteins associated with membranes.

Binominal A binominal distribution is a statistical distribution derived from many different trials, each of which may have only two possible outcomes – for example, the toss of a coin, which may only result in a "head" or a "tail". The distribution of the numbers of "heads" that may possibly be obtained from x tosses of a coin is a binominal one.

Bioconductor An open source and open development software project to produce code for bioinformatics, and particularly for the analysis of genomic data. There is a strong focus on microarray analysis, although other topics are also addressed. The public domain statistical package R, which is also often used for microarray analysis, is a prerequisite for the development of Bioconductor software.

Bioluminescence The emission of photons of visible light by a living organism (generally a lower one, although some species of fish produce bioluminescence). It is usually generated by the oxidation of a substrate. The luciferase bioluminescence system, found in nature in the firefly *Photinus pyralis*, is often used as a marker for transient gene expression or genetic transformation.
Source: Kahl, G, The Dictionary of Gene Technology.

BioPerl An international association of software developers of open source bioinformatics programs and tools using the Perl language. It provides a set of free on-line resources, tutorials, and libraries of Perl scripts for developers to use. It is facilitated by the Open Bioinformatics Foundation, and has links with other open source software developers including BioPython, BioJava, and EMBOSS.

BioPython An international association of software developers of open source bioinformatics programs and tools using the Python scripting language. It provides a set of free on-line resources, tutorials, and libraries of Python scripts for developers to use. It is facilitated by the Open Bioinformatics Foundation and has links with other open source software developers including BioPerl, BioJava, and EMBOSS.

Biosynthetic Pathway *See* Metabolic Pathway

BLAST A very popular and very fast bioinformatics tool for searching a gene or protein sequence database with a single sequence. The test sequence is aligned with each database sequence in turn, using a crude, rapid algorithm that is based on initial identification of short exact matches, and the best matches are reported. There are many variants of BLAST, including the iterative PSI-BLAST which is only available for protein sequences.

Block In bioinformatics theory, a block is an alignment of a section of (usually protein) sequences where there is a high degree of identity between the amino acids in the block. Each protein family is represented by a number of blocks, and this family information is held in the BLOCKS database. The empirical data in this database was used to derive the popular BLOSUM family of amino acid substitution matrices.

BLOSUM Matrices A series of amino acid substitution matrices generated empirically using the protein sequence alignments in the Blocks database. The numbers reflect the number of substitutions observed in the Blocks alignments. There is a series of matrices defined using different Blocks; for example, the popular BLOSUM62 matrix is derived from only those blocks that have 62% or higher sequence identity. The BLOSUM matrices are now more popular than the classic PAM matrices, although the latter are still used.

Boolean Logic The simple and ubiquitous logic in which states are combined using NOT, AND, and OR relations (sometimes combined into NAND and NOR). Boolean logic underlies the whole computer revolution, but in bioinformatics it is most obviously used in the more sophisticated type of database search. SRS is an example of a well-known bioinformatics software product that allows users to perform searches using complex Boolean logic.

Candidate Gene Approach A strategy used in the identification of genes that are linked with disease (particularly the many genetic risk factors for complex diseases). In it, one or more likely genes (the candidate genes) are first identified and the effects of polymorphisms of these genes are tested in an association study. It is possible to use this approach to identify genes that are weakly linked with disease. However, it has one key disadvantage: the disease must be understood well enough for suitable candidate genes to be identified.

Capillary Electrophoresis A technique to separate ionized molecules in silica capillaries with a diameter of 25–100 μm and a length of 20–100 cm by electro-osmotic flow. It combines the advantages of effective heat dissipation with reduced sample volume. The detection limit may be extended to sub-attomolar concentrations of the ions using high-sensitivity detectors such as laser-induced fluorescence detectors. Capillary array electrophoresis uses an array of 96 or 384 capillaries filled with polyacrylamide gel.
Source: Kahl, G, The Dictionary of Gene Technology.

Capsid The DNA or RNA genomes of all viruses are surrounded by protein coats known as capsids. Some also have a lipid bilayer membrane enclosing the

capsid. The capsids of simple spherical viruses are roughly spherical shells made up of many copies of each of a small number of subunits. These subunits are generally assembled into the capsid with icosahedral (=20-fold) symmetry; the smallest capsids are composed of 60 identical subunits.

CCD *See* Charge Coupled Device

cDNA A single- or double-stranded DNA molecule that is complementary to an RNA (usually mRNA) template from which it has been copied by RNA-dependent DNA polymerase (reverse transcriptase). Complementary DNAs immobilized onto "chips" and able to hybridize to mRNAs present in a cell sample are the basis of microarray technology.
Source: Kahl, G, A Dictionary of Gene Technology (Wiley-VCH, 2001).

cDNA array, DNA chip, gene chip *See* Microarray

cDNA Clone Set A set of DNA duplexes complementary to mRNA molecules, generated by the reverse transcription of the mRNA messages using reverse transcriptase and cloned into a plasmid or other cloning vector. Clone sets from eukaryotic genomes may be very large, and many have now been made freely available for recognized not-for-profit researchers.
Source: Kahl, G, The Dictionary of Gene Technology.

Centromere That part of a eukaryotic chromosome (as it is visible under the light microscope during metaphase and with the morphology generally associated with the word "chromosome") at which the two chromosome copies are held together. The centromere separates the p and q arms of the chromosome. Very few genes are found in the centromeric regions of chromosomes.

Chaperone Proteins that help other proteins to fold are known as chaperones (or molecular chaperones). They are found in the proteomes of all organisms except viruses, but have been studied in most detail in *E. coli*. The *E. coli* proteome contains two distinct families of chaperones, the Hsp70 chaperones and the chaperonins (exemplified by the GroEL/GroES complex). Chaperonins do not define other proteins' folds, but enable them to find their correct structures. Different chaperonin families have different mechanisms.

Charge Coupled Device A light-sensitive integrated circuit that stores data from an image, pixel by pixel, in such a way that the charge intensity stored is related to the color of the image. CCDs have many uses, but, in proteomics, they can be linked to signals from fluorescent proteins, converting the light signals to electric charge.

Chemical Shift The key measurement in NMR spectroscopy: that is, the difference between the observed resonance energy of a nucleus (of non-integral spin, e.g., 1H, 13C or 15N) and the standard value expected for that nucleus.

This difference can be correlated with the chemical environment of the nucleus concerned, and thus used to deduce structural details of that part of the molecule.

Chemiluminescence The emission of visible light as a consequence of the excitation of atoms or molecules by absorption of free energy from a chemical reaction. Metastable energy-rich intermediates are created, and these emit visible light as they decompose into their ground states. Chemoluminescence detection is a sensitive method for detecting specific proteins or DNA molecules using an enzyme-linked probe.
Source: Kahn, G, The Dictionary of Gene Technology.

Chemotaxis The directed movement of a microbe, or a motile cell, following a chemical concentration gradient. Positive chemotaxis is movement toward a higher chemical concentration, negative chemotaxis movement toward a lower one. Many proteins have been identified as being involved in mediating or responding to chemotaxis through activation of intracellular signaling pathways or remodeling of the cytoskeleton through the activation or inhibition of actin-binding proteins.

Chimerism The word chimera (chimaera) is used to mean both an organism composed of two or more genetically different cell types, and a DNA construct composed of sequences with different origins. Similarly, a chimeric gene (fused gene) is a genetic construct comprising coding seqences from one gene expressed under the transcriptional control of another.

Chromatin Chromatin is the name given to the complex of DNA and proteins that makes up eukaryotic chromosomes. In these, DNA is associated with DNA-binding proteins called histones. During the 1970s, both nuclease protection experiments and electron microscopy proved that histones are spaced regularly along the DNA, like beads on a string. Each bead is termed a nucleosome. The compact DNA-protein complex around the periphery of nondividing nuclei is termed heterochromatin.

Chromosomal Rearrangement Simply, the rearrangement of chromosomal parts, leading to chromosomes that contain parts of others. Chromosomal rearrangement is common in many genetic diseases and in cancers. Some cancers are associated with characteristic chromosomal rearrangements; one example is chronic myeloid leukemia, which is almost always associated with a translocation of part of chromosome 22 onto chromosome 9.

Chromosome Banding Chromosomes may be stained to produce the characteristic binding patterns that are illustrated in a karyogram and used to identify them by number. For example, the common G-banding procedure, involving mild proteolysis followed by staining with the Giesma stain, will produce dark bands where the DNA is AT-rich and light bands where it is GC-rich.

Chromosome substitution strain, CSS *See* Consomic Strain

CID *See* Collision-induced Dissociation

Cis-Element Cluster A cluster of sequence elements cis of (that is, in the 5′ untranslated region of) a gene. Many of these DNA sequences bind regulatory factors and are therefore involved in the regulation of gene transcription; they include, for example, TATA boxes and Ets sequences. Cis-element clusters can be detected from sequence patterns using programs such as Cister, and their presence is used in gene-finding algorithms as signals of functional genes.

Cis-splicing The joining (splicing) of two exons from the same pre-mRNA, with the removing of the intervening intron sequence. This self-evidently occurs only in eukaryotes, since prokaryotic genes contain no introns. Exons and introns are partly defined by splice site sequences, but there is not a strong enough consensus in these for them to be sufficient. Proteins are involved in splice site identification in the context of cis-splicing, and these also influence alternative splicing events.

Clade In phylogenetics, any subtree growing from a single node, whether it is a single sequence or taxon or a large group, is termed a clade. Thus, the vertebrates, the mammals, and the genus Homo may all be described as clades. There is always one common ancestor that organisms in a clade share with each other but not with others in the phylogenetic tree.

Clonality A clonal population – that is, one displaying clonality – is a population (of cells or of bacteria) that is derived from a single cell. Thus, a population of cancer cells derived from a single mutant cell will be clonal. In microbial genetics, a population of bacteria displays clonality if it has evolved without recombination; there may, however, be significant sequence diversity derived from mutation.

Cluster of Orthologous Groups *See* COG

Clustering A mathematical technique in which points in large data are "clustered" into groups depending on their properties. In bioinformatics, clustering is often used to analyse microarray datasets and group genes with similar expression profiles. Clustering algorithms used in microarray analysis include hierarchical clustering, self-organizing maps (SOMs), k-means clustering, and principal component analysis.

Codon A triplet of bases that codes for a particular amino acid or for a START or STOP signal. Four bases give a total of 4ˆ3 or 64 different possible codons, so with 20 coded amino acids there is significant redundancy. In the standard genetic code, some amino acids (e.g., Trp) are coded by only one codon, others (e.g., Leu) by as many as six.

Coevolution A reciprocal evolutionary change in two or more species that interact together, or, a change in the genetic composition of one species as a result of a genetic change in another that interacts with it. The term is usually attributed

to Ehrlich and Raven (1964) who studied genetic diversity in butterflies and their host plants in relation to the interactions between them.

Co-evolution *See* Coevolution

Coexpression The simultaneous transcription of two or more different genes, and the translation of the resulting mRNAs in a cell. Coexpression of different transgenes may be achieved by fusing them to a specific type of promoter, or by placing two different promoters in opposite orientation between the genes.
Source: Kahl, G, The Dictionary of Gene Technology.

Cofactor If an enzyme requires another molecule to take part in the reaction that it catalyzes, and if that molecule is not changed by the reaction, the molecule is termed a co-factor for that reaction. Co-factors are usually fairly small organic molecules, and nucleotides and nucleotide phosphates are common ones. The co-factor generally binds to a different site on the enzyme from the substrate.

COG A third class of organisms, distinct from both the bacteria and the eukaryotes, originally distinguished from the bacteria by an analysis of the evolution of rRNA structure. Archaea are now thought to be the first ancestors of eukaryotes. They are, however, prokaryotes, as they are single-celled organisms with no nucleus. Extremophiles, which thrive in "extreme" conditions such as high temperature or high salinity, are archaea.

Coiled Coil A protein structure composed of two or three parallel alpha helices packed or wound together (coiled round each other); such a structure is more stable than straight helices lying side by side. They are often found in fibrous proteins. The helices display a pattern of hydrophobic and hydrophilic residues that repeats every seven residues, and several bioinformatics programs have been written to predict coiled coils from sequence using these rules.

Coimmunoprecipitation A purification procedure that is used to determine whether two proteins interact. An antibody to one protein is added to a cell lysis. The antibody-protein complex then is pelleted, usually using protein-G sepharose (which most antibodies will bind to). Proteins that bind to the first one will also be pelleted, and can therefore be identified either by Western blotting or by protein sequencing.

Co-IP *See* Coimmunoprecipitation

Collision-induced Dissociation A technique used for determining the sequences of proteins or peptides using mass spectrometry. In it, the mass-charge ratio of selected ions is first determined and then those ions are further fragmented through collisions with (usually) noble gas atoms. The sequence of the original fragment can be determined from the mass-charge ratios of the ions generated from the second set of collisions.

Compartment, Subcellular *See* Subcellular Compartment

Complementary DNA, copy DNA *See* cDNA

Complex Disease *See* Multigenic Disease

Concept Disambiguation The process by which, in order to provide the interpretation or definition of a complex term (e.g., when constructing an ontology) the different concepts within the term are disambiguated, that is, the particular sense of the term in the appropriate context is determined. (One simple example of a term that has different meanings in different contexts is the noun "virus", which has different meanings in microbiology and in computer science.)

Conformational Sampling Any method of modeling the likely conformations of a molecular structure (e.g., a protein or protein–ligand complex) that involves taking a random sample of conformations and calculating their energy. Systematic sampling is only possible if the molecules are fairly small. Monte Carlo analysis is a good example of a nonsystematic conformational sampling technique.

Congenic Strain A congenic strain of an experimental animal (most typically a mouse) is one that is genetically identical to the original, or host, strain at all positions except those linked to the gene of interest. It is obtained by a series of backcrosses between the host strain and a strain carrying the mutation of interest, usually with a mixed genetic background.

Consensus The dictionary definition of "consensus" is just "a general or widespread opinion" (Collins English Dictionary, 1986). The term has several specific meanings in bioinformatics. A consensus method of, for example, secondary structure prediction is a method in which the same prediction is made using several different algorithms and the result recorded only where there is general agreement. A consensus sequence is a sequence giving only those residues conserved in a multiple sequence alignment, with no residue recorded in the nonconserved positions.

Conservative Substitution A substitution of one amino acid for another that does not cause a significant change in the protein's function (hence conservative) and that is commonly seen in alignments of homologous sequences. Usually, the chemical nature of the amino acid pairs involved in conservative substitutions is similar. Phenylalanine for tyrosine (or vice versa) and leucine for isoleucine or valine are examples of conservative substitutions.

Consomic Strain A consomic strain of an experimental animal (most typically a mouse) is one in which a single chromosome from one inbred strain (the donor strain) is transferred onto the background of another strain (the host strain) by repeated backcrossing. It usually takes about 10 backcrosses to create such a strain. It is possible to create a panel of consomic strains using the same donor and host strains but with a different chromosome replaced in each one.

Constitutive Gene *See* Housekeeping Gene

Construct Simply, an informal (possibly “laboratory slang”) term for any recombinant DNA molecule: that is, any DNA molecule formed in vitro through the ligation of two or more nonhomologous DNA molecules. For example, a recombinant plasmid containing one or more inserts of cloned foreign DNA. The term may be derived from “in vitro constructed DNA”.
Source: Kaul, G, The Dictionary of Gene Technology.

Contig An abbreviation for contiguous segment; one of many genomic clones produced during a genome sequencing project that contains mutually overlapping DNA sequences. Obtaining contig sequences is now relatively straightforward: the latter stages of the genome project, called finishing, when the contigs are assembled in the correct order and joined to form the completed sequences, are the most time consuming.

Control Element A DNA sequence, such as a promoter or operator, that responds to an external signal (e.g., light, temperature) or an internal one (e.g., the presence of a hormone or other chemical signal) and so controls whether or not the gene associated with the control element is expressed. Transposons (transposable elements or mobile elements) in plants such as corn may also be termed control elements.
Source: Kahl, G, The Dictionary of Gene Technology.

Controlling Element *See* Control Element

Convergence In mathematics and computer science, an iterative algorithm is said to have converged (reached convergence) when the result is no longer changed from one iteration to the next. One example in bioinformatics is the iterative protein search program PSI-BLAST; this is said to have converged when a database search cycle adds no more protein sequences to the proposed list of matches.

Convergent Evolution If the same solution (in terms of protein structure and/or function) is arrived at two or more times during evolution, giving rise to two or more similar solutions with no direct evolutionary link between them, that is termed convergent evolution. Examples include the serine proteases trypsin and subtilisin, which have the same mechanisms but completely different structures, and the various families of unrelated enzymes with the TIM barrel fold.

Copy Number The number of a particular plasmid in a cell, or the number of a particular gene (or chromosome) in a genome. A low copy or “low cop” mutation is a mutation that leads to a decrease in the copy number of plasmids in a cell; it is not favored in recombinant DNA experiments. If there are only a few copies of a plasmid in a cell, that plasmid is termed a low copy number plasmid or stringent plasmid.
Source: Kahl, G, The Dictionary of Gene Technology.

Cosegregation The transmission of two or more genes on the same chromosome to the same individual in the next generation as a result of them having been included in the same gamete. The closer two genes are on a chromosome, the more likely they are to cosegregate. The opposite process, in which genes on the same chromosome are separated in the gametes, is recombination.

Cosmid A small extra chromosome found in bacteria, into which fragments of about 40 kb of DNA from various external sources may be introduced using recombinant DNA methods. This allows the foreign DNA to be replicated every time the bacteria divide; it is a common and useful technique in genome sequencing. This was one of the first methods to be developed for amplifying DNA in sequencing projects.

Co-transcription Two (or more) genes are cotranscribed if they are transcribed together into the same mRNA, and, usually, then translated together into protein. Genes that are cotranscribed must be linked and are almost always contiguous on the chromosome. Cotranscription and cotranslation gives rise to the presence of equimolar concentrations of proteins in the cell.

Covalent The atoms in organic molecules are connected by strong covalent bonds (single, double, partial double, or triple). Unlike noncovalent bonds, such as hydrogen bonds, they are extremely strong and can only be broken in chemical reactions. In molecular mechanics simulations, bond lengths are modeled using simple Newtonian mechanics, using parameters for bond length and strength that vary according to atom type and bond order.

Coverage The number of times, on average, that any piece of DNA has been sequenced in a genome sequencing project. As it is still only possible to sequence DNA in fairly small segments, each segment must be sequenced many times to make sure that the fragments will be assembled in the correct order. In general, the higher the copy number, the more accurate the assembled sequence will be.

Cryptic Relatedness A problem (or complication) in medical genetics, particularly in case–control studies, arising from correlation between alleles in a subpopulation that must be taken as independent. In case–control studies, where it is assumed that there is no genetic link between individuals, the "cases" must be assumed to be more likely to be related than the "controls" because they share a deleterious gene or genes.

2D-PAGE A technique for separating complex mixtures of proteins according to isoelectric point and mass, which has revolutionized the discipline of proteomics. The protein mixture is usually subjected to electrofocusing to separate by pI; the resulting 1D gel is separated by mass using SDS-PAGE.
Source: Kahl, G, Dictionary of Gene Technology, Wiley 2001.

Data Structure Trivially, a complete and self-consistent structure that can be imposed on a set of data, such as the elements of a database. Data structures have

been described as one of two core components of computer science, the other being algorithms. Ontologies and markup languages schemas are two types of data structure. Sequence and structure file formats (such as, respectively, the FASTA and PDB formats) can be thought of as specialized data structures.

Degeneracy Generally, any code that uses a "many to one" mapping, so that more than one term in the original map onto a single term in the code, is termed degenerate. Therefore, the genetic code is an example of a degenerate code: 64 different three-base codons map onto 20 amino acids plus the START and STOP codons, so more than one codon usually code for the same amino acid. The maximum number of codons that code for any one amino acid is six; the minimum number is one.

Deletant *See* Deletion Mutant

Deletion Mutant A mutant that has been generated by the loss of one or more base pairs from the DNA of its genome. Similarly, a deletion mutation is any mutation that results in base pair loss. If the deletion mutation occurs in a coding region, and unless the number of base pairs deleted is a multiple of three, a deletion mutation will cause a frame shift leading to a truncated gene and probably to a loss of gene function.

Dendrimer A highly branched multilayered DNA scaffold structure for the annealing of multiple oligonucleotide probes. A dendrimeric unit consists of a central double-stranded region with four single-stranded arms, which are complementary to the single-stranded arms of another unit. DNA dendrimer technology is useful for the detection of rare DNA or RNA sequences.
Source: Kahl, G, A Dictionary of Gene Technology.

Desolvation The process of breaking interactions between the atoms of (usually) a protein molecule and the molecules of the solvent in which the protein is dissolved, in order for the protein to fold. Generally, the interactions between protein and solvent atoms are stronger than those between atoms within the protein, so the process of desolvation requires energy. Desolvating hydrophobic atoms and groups is more energetically favorable than desolvating hydrophilic ones; desolvating polar atoms removes ordered hydrogen-bonding interactions from the solvent surrounding the protein and so increases the entropy of the system.

Difference Gel Electrophoresis *See* DIGE

Differential Expression Genes are not expressed equally in all cells and tissues; which of the potentially active genes are expressed in a given cell at a given time depends on the cell type, developmental stage, and even on whether or not the cell is diseased. Microarrays are used to determine differential expression – that is, differences in expression patterns – between genes in different cell types or under different conditions. This has important implications for differential diagnosis and for the selection of potential targets for drug design.

Differentially Methylated Region Regions of DNA in eukaryotic chromosomes that are methylated differentially on the maternal and paternal alleles are termed differentially methylated regions. Many of them are CpG islands. They are divided into two classes, one methylated during gametogenesis and the other methylated after fertilization. It has been suggested that methylation is the epigenetic mark (imprint) that differentiates the paternal and maternal alleles.

DIGE An emerging technique in proteomics where the protein samples are labeled with fluorescent dyes prior to separation using a 2D-PAGE gel. This enables up to three different images of the same gel to be captured (using the wavelengths of the different dyes) and differences between the images determined using image analysis software. The main advantage of this technique is the avoidance of gel-to-gel fluctuations: the main disadvantage is cost.

Dihedral Angle The torsion or twist angle that defines the geometric position of atoms separated by two other atoms in a covalently bonded chain: thus, if four atoms A–B–C–D are connected together, the dihedral angle is the twist angle about the B–C bond that defines the position of D with respect to A. Free rotation is only allowable about single covalent bonds. The dihedral angles that describe the structure of amino acids within proteins are named using Greek letters: the backbone torsion angles as phi, psi, and omega and the side chain angles as chi1, chi2 etc.

Diploid A eukaryotic cell is described as diploid if it contains a full copy of chromosome pairs. Thus, a human diploid cell contains two copies of each of the 22 autosomes and two sex chromosomes (XX for females and XY for males) making 46 chromosomes in all. All normal human cells other than gametes are diploid: these cells are also termed somatic cells.

Directed Graph In graph theory, a graph is defined as directed if its edges have direction (so that they join one node "to" another rather than joining them together). A graph describing relationships between genes will be directed if the relationship defining the edges is, for example, "regulates". Many applications of graph theory to bioinformatics use directed graphs.

Directon A contiguous series of genes on a (normally bacterial) chromosome, such that all the genes in the directon are transcribed in the same direction. Directons are easily identified and counted once the complete genome sequence is known. One directon may contain one, or more than one, operons (defined as transcription units containing two or more genes).

Discriminant Function A method of distinguishing between data based on the measurement of a set of properties. For example, in gene prediction, a method of distinguishing between genic and intergenic DNA based on patterns and frequencies of the different bases (e.g., GC content) would be described as a discriminant function. Although this is the most obvious one, the phrase can be used in other contexts in bioinformatics.

Distance, Evolutionary *See* Evolutionary Distance

Distance Matrix A matrix holding data on the evolutionary distance between amino acids (i.e., the probability that a substitution of one amino acid by another will be accepted). The most widely used examples are the BLOSUM series and the older, but classic PAM series. They are used in all protein–protein sequence alignments and database searches. The equivalent matrix in DNA sequence analysis is usally the identity matrix.

Disulphide Bond A contiguous series of genes on a (normally bacterial) chromosome, such that all the genes in the directon are transcribed in the same direction. Directons are easily identified and counted once the complete genome sequence is known. One directon may contain one or more than one operons (defined as transcription units containing two or more genes).

Disulphide Bridge *See* Disulphide Bond

Divergent Evolution The usual process of evolution in which DNA molecules diverge in sequence (and, thence, the protein products of the genes they code for may (or may not) diverge in structure and/or function can be described as divergent as opposed to convergent evolution. Divergent evolution may take place within or between genomes.

DMR *See* Differentially Methylated Region

DNA dendrimer *See* Dendrimer

DNA Methylation The enzymatic transfer of methyl groups onto nucleotides in DNA molecules, more precisely from *S*-adenosyl methionine onto C5 of cytosine residues (generally in eukaryotes) or N6 of adenine residues (generally in prokaryotes). Methyltransferases in prokaryotes are part of a restriction modification system: in eukaryotes, methylation of bases in the promoter region of genes can modulate gene transcription.
Source: Kahl, G, The Dictionary of Gene Technology.

DNA Polymerase Enzymes that catalyze the polymerization of deoxyribosenucleotide triphosphates into the DNA polymer based on the sequence template of a single-stranded DNA molecule. Polymerization proceeds in the 5′-3′ direction. DNA polymerases are used for DNA repair as well as replication. All organisms, apart from some viruses, contain one or more DNA polymerases. Some DNA polymerases also possess exonuclease activity, and some have important uses in molecular biology.

DNA Transfection; Transformation-Infection *See* Transfection

Docking A technique used in protein bioinformatics and especially in drug design, where a small organic molecule (e.g., a putative substrate or inhibitor)

is placed within the binding site of a protein and its geometry and position varied to minimize the energy of the system. (The term is derived by analogy with a ship docking in a port.) Many commercial and public domain automatic docking programs are now available.

Domain Most often used to describe part of a protein chain that will fold (and may be crystallized) independently. However, domains are also distinguishable by both function (e.g., a serine protease domain, a tyrosine kinase domain) and by sequence. The domain is the basic unit in both databases of protein structure (e.g., SCOP, CATH) and of protein sequence/function (e.g., Interpro).

Domain Chaining A phenomenon in protein function prediction that may lead to errors in domain identification. If protein X contains domains A and B, a sequence search will pick up similarity to protein Y containing domains B and C. If protein Z, containing domains C and D, is also identified – wrongly – as having similarities to X, domain chaining has occurred. Careless use of the iterative search program PSI-BLAST may lead to domain chaining.

Domain Recombination The process through which multidomain proteins with different structures and functions can be generated by combining domains in different numbers and orders. A large proportion of proteins contain more than one domain, and a relatively small number of domains (compact units with similar 3D structures and, generally, similar functions) are found in a wide variety of multidomain proteins. Domain recombination is much less common in integral membrane proteins.

Domain Swapping A mechanism for forming oligomeric proteins from their monomers, in which one domain of a multidomain, monomeric protein is replaced in the protein 3D structure by the same domain from an identical protein chain. The resulting intertwined dimeric or oligomeric structure is extremely stable. Large domains, supersecondary structures and single helices or strands may be involved in domain swapping. In one example, some cysteine proteases crystallize as domain swapped dimers.

Dominant In Mendelian genetics, a phenotype is defined as dominant if it is displayed in an individual that is heterozygous for the characteristic (containing one allele with the gene for the trait and the other without). Thus the normal phenotype is dominant in cystic fibrosis (recessive inheritance), but the disease phenotype is dominant in Huntington's disease (so all individuals who inherit the HD allele will develop the condition if they live long enough; dominant inheritance). The term incomplete dominance is used when the heterozygous phenotype is intermediate between the two homozygous ones, and codominance where both homozygous phenotypes are observed to some extent in heterozygote.

Dominant Negative Mutation A mutation that, when expressed in a heterozygote, produces a gene product that interferes with the functioning of the normal

gene product is termed a dominant negative mutation. The mutant protein usually combines physically with the normal one to form a dimer. In these cases, heterozygotes invariably exhibit a disease phenotype (which may or may not be the same as that of homozygotes for the mutant gene) and the inheritance will be described as dominant, codominant, or incomplete dominant. The allele involved is an antimorphic allele.

Dot Matrix A method of comparing two sequences (or whole genomes) as a simple plot of one sequence against the other, by sliding a window along each sequence in turn. If the two segments under consideration are similar enough, a dot is drawn at the specified point on the plot; lengths of sequence similarity are therefore represented by diagonal lines. The degree of similarity between the sequence segments required for a dot to be drawn is termed the stringency of the plot.

Dot Plot *See* Dot Matrix

Downstream Toward the 3′ end of a DNA sequence. The term is most often used for sequence that is located 3′ of the coding sequence of a gene. The 3′-most region of that sequence that is transcribed into the pre-mRNA (but not translated into protein) is known as the 3′-untranslated region (3′-UTR). The region downstream of the 3′UTR, which is never transcribed into mRNA, may also be of interest.

Drug Target A protein that is involved in a disease state and where it is thought that chemical intervention (most often, but not always, using a small organic drug molecule) may alter that activity and affect the progress of the disease. Out of tens of thousands of proteins in the human and pathogen proteomes, only about 400 had been used as drug targets by the turn of the twenty-first century. Proteins that are thought to be good drug targets are (self-evidently) frequently selected by (e.g.) structural proteomics programs.

Dyad The word dyad is used in general terms to describe a pair, so some people may refer to, for example, a hydrogen bond donor – acceptor dyad or a protein – ligand dyad. However, in DNA sequence analysis, the word has a specific meaning: it is used to describe a sequence fragment that is immediately followed by its reverse complement. Statistical analysis has been used to determine whether dyads are distributed evenly within chromosomes or genomes.

Dynamic Programming A programmming technique that involves breaking the problem down into constituent subproblems, which may share some components (sub-subproblems) and storing ("memorizing") the results of the shared components. In bioinformatics, dynamic programming is used in, for example, the Needleman–Wunsch and Smith–Waterman alignment algorithms. In these, the use of dynamic programming allows for the computationally efficient incorporation of gaps.

Dynamic Range The range of, for example, protein concentrations that may be detected by a given technique, expressed in orders of magnitude. Commonly used

silver staining techniques, for example, may detect protein concentrations over a range of only two orders of magnitude, whereas the newer fluorescent techniques may have a range of up to five orders of magnitude. A technique is said to have a linear dynamic range if the relationship between a signal and the respective concentration is one of simple proportion (so plotting signal against concentration gives a straight line).

Dysmorphology Dysmorphology is the study of abnormal form in humans, and, therefore, of genetic disorders that cause external malformations, whether simple or complex, mild (often hardly visible) or severe enough to cause disability. Many such abnormalities are caused by the malfunction of developmental genes. Large numbers of familial dysmorphology syndromes, inherited in the Mendelian pattern, are known, and the mapping of these has led to the development of molecular dysmorphology and to the mapping of developmental genes.

Edge In graph theory, objects, known as nodes, are connected by lines indicating relationship; these are known as edges. The edges may or may not have directionality, depending on the relationship modeled, and they be labelled to represent different degrees of relationship. Graph theory has many applications in bioinformatics, where is used to cluster the most closely related objects (typically genes or proteins) together. Edges may represent, for example, relationships such as "is coexpressed by" (for genes) or "interacts with" (for proteins).

Electrophoresis In theory, the movement of charged molecules in solution within an electric field. In practice, any method for separating charged molecules that uses an electric field, exploiting differences in the charge, shape, and size of the particles. The core proteomics technology of two-dimensional gel electrophoresis involves the separation of particles in two dimensions, according to both charge and size.
Source: Kahl, G, The Dictionary of Gene Technology.

Electroporation A method for the direct transfer of macromolecules, most often DNA, into cells by perforating the cell membrane with a short electric pulse and a potential gradient, leading to a transient permeabilization of the cell membrane. The process will allow the entry of large DNA molecules that may be integrated into a nuclear or organelle genome.
Source: Kahl, G, The Dictionary of Gene Technology.

Electrospray Ionization A technique used in protein mass spectroscopy, as applied to the analysis and identification of separated proteins obtained in proteomics experiments (e.g., 2DPAGE). A solution of the digested peptides is passed through a thin needle with a nebulizing gas, and a high voltage applied to the tip. This generates a spray of droplets containing ions. The droplets evaporate leaving peptide ions which pass through a series of electrodes and samplers to the mass analyzer, which usually works on the time-of-flight principle.

Electrotransformation *See* Electroporation

Embryonic Stem Cells A type of nondifferentiated cell that is derived from a very early embryo and has the potential to differentiate into any one of many different types of cell. Embryonic stem cells have enormous potential in medicine, but their use has important ethical implications, is extremely controversial, and has been banned in many countries. Rather less controversial technologies involving ES cells include the production of transgenic mice by incorporating genes containing mutations into mouse ES cells to produce lines of transgenic mice.

Endogenous Retrovirus A complete retrovirus sequence that is integrated into a recipient genome. ERV sequences may be complete, containing the full gene complement of exogenous retroviruses, or incomplete. The human genome contains several ERV families (known as HERVs). These are transcribed but not normally translated, and all human endogenous retroviruses are noninfectious.
Source: Kahl, G, The Dictionary of Gene Technology.

Endosome A vesicle, derived from the cell membrane, through which cells absorb material (such as nutrients) from the environment in a process known as endocytosis. Typically, the cell membrane forms a pocket containing the nutrient, which pinches off to form a vesicle that ruptures to release its contents into the cytoplasm. It is the opposite of exocytosis, the process through which waste products and toxins are excreted from the cell.

Endosymbiosis The evolutionary theory that states that organelles within eukaryotic cells must have evolved from prokaryotes that were engulfed by the primitive eukaryotes. Specifically, mitochondria are believed to have evolved from aerobic bacteria (probably related to the rickettsias) living within the host cells, and chlorplasts are believed to have evolved from endosymbiotic cyanobacteria (autotrophic prokaryotes). This theory explains the fact that organelles have their own genomes.

Energy Minimization The technique in which the geometry of a molecular system is altered in order to minimize its energy. Most energy minimization programs work with a molecular mechanics model of the molecule, which defines it in terms of its geometry (bond lengths, angles, and torsion angles), atom types, and partial atomic charges. There are many problems with this methodology, not least the likelihood that the molecule will simply move into the nearest local minimum conformation rather than the global energy minimum. However, it has its uses in the final stages of homology modeling and in estimating interaction energies of molecular complexes.

Epigenetic Factor Any factor that affects gene expression without altering the sequence of the genes involved. Some epigenetic factors, such as DNA methylation, involve permanent, covalent bonds to the DNA molecules; others involve the structure of molecules such as chromatin that interact with the DNA via nonbonded interactions. If chromatin is particularly dense, transcription factors may be physically prevented from reaching the DNA, preventing transcription.

Epigenetics The altering of gene expression by any means that is not related to the sequence of bases along the DNA molecule. Epigenetic modifications of DNA, such as DNA methylation, changes in chromatin structure and histone deacetylation affect the expression of many genes and are involved in diseases such as cancer. Epigenetics also goes part way to explain differences, for example in complex disease incidence, between identical twins (who, of course, have identical DNA sequences).

Ergodic Theory A mathematical theory related to the theory behind Markov modeling, which has many applications in bioinformatics (e.g., gene finding and protein function prediction) and that also has applications in thermodynamics. It is a statistical theory saying, in simple terms, that if all states of a system are equally accessible, they will be occupied with equal probability if sampled over a long period of time.

ERV *See* Endogenous Retrovirus

ES Cells *See* Embryonic Stem Cells

ESI *See* Electrospray Ionization

EST *See* Expressed Sequence Tag

Eutherian A Eutherian mammal is a member of the subclass Eutheria of the class Mammalia: that is, any mammal with a placenta (the structure in the uterus of a pregnant female that allows the transport of nutrients to, and waste products from, the developing fetus). The only mammals that are not Eutherian are the marsupials and the monotremes.

E-value A statistical measure of the similarity between two sequences, related to the probability of the sequences being unrelated (thus, the lower the E-value, the more significant the match). The E-value, or expect value, is a measure of the number of unrelated sequences in a database the size of the one used that would be expected to be at least as similar to the test sequence as the one chosen.

Evolutionary Distance A measure of the closeness of the relationship between the genomes of two organisms (generally two species) measured from the amount of divergence between aligned homologous regions of DNA. Evolutionary distance is most commonly measured using the number of nucleotide substitutions per site, although this method is not always the most useful for reconstructing phylogenetic trees. One well-known example is the evolutionary distance between mouse and rat, which can be given as 0.014 indels per site.

Evolutionary Dynamics The study of the dynamics (forces on and changes in over time) of populations undergoing evolution – that is, of populations capable of mutation and subject to selective pressure. Evolutionary dynamics may be applied at the molecular or organism/ecosystem level.

Evolution, Convergent *See* Convergent Evolution

Evolution, divergent *See* Divergent Evolution

Exon Those sequences within a eukaryotic gene that are conserved during pre-mRNA processing and so make up the mature message (c.f. Intron). The introns on the 5′ and 3′ ends of the mRNA may contain sequences that signal initiation or termination of processing, respectively, and so are not translated into protein. Prokaryotic genes do not contain introns, so the concept of exons is meaningless there.

Exon–intron Structure The structure of a eukaryotic gene, in terms of the number, order and size of the coding exons and the noncoding exons that separate them. Eukaryotic genes vary widely in exon–intron structure, from single exon genes, through genes with a simple structure such as insulin with its single intron, to genes where over 90% of the genetic material consists of introns. It is even possible for one gene to be embedded in an intron of another gene.

Exon Skipping The elimination of one or more exons from a transcript during splicing, such that the combination of exons remaining results in a different mRNA and hence in a translated protein with a different arrangement of domains. For example, a gene with three exons, A, B, and C may give rise to a protein containing domains coded from only exons A and C as a result of the skipping of exon B.
Source: Kahl, G, The Dictionary of Gene Technology.

Expect Value, Expectation Value *See* E-value

Expert System A form of artificial intelligence in which the knowledge associated with an area of human expertise is codified into a set of rules that are then applied to the analysis and classification of data. Like artificial intelligence, more generally, expert systems are less common in bioinformatics than they were 20 or 30 years ago. They do, however, have some applications there, and they are more common in clinical applications.

Expressed Sequence Tag A short, synthetic oligonucleotide of 300–500 bp, complementary to the 5′ or 3′ end of a specific mRNA and usually derived from a cDNA library by random sequencing. ESTs represent tags for the state of gene expression in a given cell type at a given time (disease status and/or developmental stage). Millions of sequenced ESTs have been deposited in public databases: ESTs represent the largest subdivision of the EMBL and GenBank databases.
Source: Kahn, G, The Dictionary of Gene Technology (Wiley-VCH, 2001).

Expression Cartridge *See* Expression Cassette

Expression Cassette A DNA fragment (usually synthetic) into which foreign DNA can be cloned and expressed. The expression cassette is usually part of an expression vector and encodes a control region (e.g., promoter) with an

adjacent Shine–Dalgarno sequence (for expression in prokaryotes), a signal peptide sequence if necessary, a polylinker, and a termination sequence.
Source: Kahl, G, The Dictionary of Gene Technology.

Expression Profile The simultaneously measured levels of many thousands of mRNAs (expressed genes) in a cell or tissue type, detected using a microarray. The classic microarray experiment measures differences between the expression profiles of different cell types, or the same type under different conditions.

Expression vector; expression cloning vector; transcription vector *See* Vector

External Spike-in Control A standard pool of cDNA sequences showing no sequence identity to the human genome (or the other genome under analysis) used as controls in microarray experiments. The samples, which span a large concentration range, are added to the RNA samples before labeling and can be used to test for efficiency of cDNA synthesis/labeling, uniformity of hybridization, and sensitivity of detection.

FASTA A bioinformatics tool for searching a gene or protein sequence database with a single sequence. FASTA is complementary to BLAST; it is slightly more precise, but runs much more slowly. The database is scanned for those sequences that contain the largest number of perfect matches to short subsequences of the test sequence, and each of these best matches is then aligned more precisely. FASTA has also given its name to the most commonly used sequence file format, where the sequence title is given on the first line preceded by a greater-than sign.

FID *See* Free Induction Decay

Fingerprint A pattern of peptides obtained from a protein by proteolytic degradation (very often using the protease trypsin) and then separated and identified using mass spectrometry and optionally peptide sequencing. The peptide fingerprint produced by a particular protease is characteristic for each protein and can be used for protein identification.

FISH A technique used to study the details of an individual's chromosomes, and particularly to determine chromosomal abnormalities either before or after birth. It involves labeling segments of single-stranded DNA (known as probes) with fluorescent dye. If the cells of the individual under study contain chromosomal DNA that is complementary to the probes, they will bind and the fluorescent signal will be detected. Unlike most techniques for chromosomal analysis, it does not need to be performed on dividing cells.

Fluorescence In Situ Hybridization *See* FISH

Fluorescence Resonance Energy Transfer A technique used in many proteomics applications, for example, in probing the structure of proteins, and protein–protein and protein–ligand interactions. It uses the fact that energy can,

under certain circumstances, be transferred between two dye molecules (donor and acceptor) in close proximity without the emission of a photon. The absorption spectrum of the acceptor must overlap the fluorescence emission spectrum of the donor and the dipoles of the two molecules must be approximately parallel. It can be used to determine intermolecular distances of the order of tens of Angstroms.

Fluorophore Any molecule that produces fluorescence and that can therefore be used as a probe, typically in proteomics experiments. In proteomics, fluorescence may be described as intrinsic or extrinsic. Intrinsic fluorophores may be fluorescent molecules that are naturally bound to the protein, perhaps as cofactors, or amino acids with intrinsic fluorescence. Green fluorescent protein contains the unusual intrinsic fluorophore of a STG tripeptide, which is posttranslationally modified to a 4-(p-hydroxybenzylidene)-imidazolidin-5-one. Extrinsic fluorphores are fluorescent molecules that may be artificially bonded to proteins.

Fold The level in protein structure classification that groups together proteins with the same secondary structure elements connected in the same order. The FOLD level in the SCOP database equates to the Topology level in CATH. Proteins with the same fold may have a common evolutionary origin, but homology cannot be assumed, particularly with the highly populated so-called superfolds (examples being the alpha-beta barrel and the immunoglobulin fold).

Fold Prediction A method for predicting the tertiary (three-dimensional) structure of a protein that does not necessarily require the structure of a homologous protein to be available. It involves aligning the sequence of the protein to be modeled with known protein structures and using “threading” or a similar algorithm to select structures that are most compatible to the test sequence. It is a low precision method and cannot be used to predict novel folds, but it has had some notable successes, particularly in selecting remotely homologous structures as templates for homologous modeling.

Force Field A self-consistent representation of a molecule or molecular system (e.g., protein, ligand, and solvent molecules) using Newtonian mechanics, which can be used for energy minimization or molecular dynamics simulations. Atoms are represented by points with size, mass, and partial atomic charge, and bonds as “springs” separating them. Bond lengths, bond angles, and torsion (twist) angles are maintained close to optimum positions using energy terms, and other energy terms define the nonbonded forces between atoms. Given the crudity of this model, it can produce some surprisingly good results, but these techniques must be used with caution.

Founder Population The small population that first invades an isolated area such as an island. Descendents of a founder population will exhibit reduced genetic variation as a result of this population bottleneck. Human communities that descended from founder populations may exhibit unusually high prevalence of Mendelian diseases.

Fourier Transform A mathematical transform that expresses a function as a sum or integral of sinusoidal functions multiplied by constants known as amplitudes. There are many distinct forms of Fourier transforms, which were named for French mathematician Jean Baptiste Fourier. The most common applications of Fourier transforms in areas related to molecular biology are in structural biology, in deconvoluting X-ray and electron diffraction patterns to obtain the structures of macromolecules.

Fourier Transform Ion Coupled Resonance A mass spectrometry method that is able to measure masses of, for example, peptide ions, with very high precision and accuracy. It is a versatile analysis method that may be used with both MALDI and ESI ionization technologies. It is an ion-trapping method, based on the principles of ion cyclotron resonance (ICR) spectrometry; one of its few disadvantages is that it is very sensitive to pressure and requires near vacuum conditions.

Frameshift An alteration of the reading frame of a gene in which the sequence is read (i.e., translated into protein), caused by a change in sequence length (i.e., an insertion or deletion) of a number of nucleotides that is not divisible by 3. It usually results in the production of a truncated, nonfunctional amino acid because all the sequence downstream of the change will be translated incorrectly.

Free Induction Decay In structure determination by NMR, free induction decay (FID) is a transient signal that decays (or relaxes) exponentially with time and that is caused by dephasing in an inhomogeneous magnetic field. The signal is sinusoidal and it is generated by spins in the $x-y$ plane; it decays over time as magnetization returns to its equilibrium level.

FRET *See* Fluorescence Resonance Energy Transfer

FTICR *See* Fourier Transform Ion Coupled Resonance

Gain of Function Mutation Trivially, any mutation where the protein produced by the mutated gene displays extra functionality (either a different function altogether or an enhancement in normal function) that is not present in the wild type. Gain of function mutations may be beneficial or deleterious; inheritance of these mutations is usually, if not always, dominant.

Gapped Alignment Any alignment of two or more sequences that includes gaps, that is, that allows for insertions and deletions in the sequences. In practice, almost all alignment methodologies in modern bioinformatics produce gapped alignments; the only nongapped alignments are local alignments used where speed is at a premium. Early versions of BLAST produced nongapped alignments.

Gap Penalty A penalty that is deducted from an alignment score for the addition of a gap in a sequence alignment. Alignment programs generally use two different gap scores: a large penalty for starting a gap (the gap insertion penalty) and a smaller penalty for extending one (the gap extension penalty). This reflects the fact

that there is a greater difference between an ungapped alignment and one with an insertion or deletion of a single character (base or amino acid) than there is between an alignment with a single insertion or deletion and one with two.

Gene Duplication An event during evolution in which a single gene is duplicated, giving rise to two different genes in the same genome. These genes will gradually diverge on an evolutionary timescale, giving rise to gene products with different sequences and usually different, although related, functions. Genes in the same genome that are related in this way are known as paralogs.

Gene Fusion The use of recombinant DNA techniques to join (fuse) together two or more genes coding for different products so that they are expressed under the control of the same regulatory system.
Source: Kahl, G, The Dictionary of Gene Technology (Second Edition).

Gene Gun A piece of apparatus used for inserting transgenes into plant cells. Genes are loaded on to very small gold or tungsten pellets. These are then fired at the leaves or other tissues of the target plant, using the gene gun. The pellets pass through the plant tissue, but the genes are physically wiped off the pellets and may be incorporated into the plant chromosomes.

Gene Index A list of genes; specifically, an annotated, nonredundant list of the genes in a genome, generally including other related genetic information and links. The TIGR Human Gene Index, which is freely available, contains data on the expression patterns, functions, and evolutionary relationships of the genes in the index. Gene indexes for other, less well studied genomes will tend to be less complete.

Gene Knockout An informal term used in the lab and less frequently in non-specialist publications for the disruption of a gene by the addition or deletion of base sequences so that the function of the gene is abolished. Knockout mice, or mice in which the function of one gene has been removed, have very important uses in the study of genetic diseases.

Gene Silencing The inactivation of a previously active (i.e., previously transcribed) gene. Its converse, gene activation, is used to mean the activation of a previously silent gene. Silencing (or activation) may take place by altering the transcription mechanism of the gene rather than the sequence of its coding regions.

Genetic Anticipation A phenomenon in which some genetic diseases are observed to appear in more severe forms, either in terms of symptom severity, age of onset, or both, in subsequent generations. Genetic anticipation is common, and has been widely studied, in the trinucleotide repeat disorders (e.g., Huntington's disease) but it has also been observed in some more common complex diseases with a strong genetic component, including Crohn's Disease and biolar disorder.

Genetic Drift A random change in the frequencies of different alleles in a population that is neither deleterious nor beneficial. The term is also used to mean

random change as a mechanism of evolution; it is believed to be one of the two most important such mechanisms, with natural selection. Moran, in the reference below, states that random genetic drift is by definition a stochastic mechanism.

Genetic Imprinting An epigenetic process by which the male and female germline of viviparous species confer specific marks on certain chromosomal regions, leading to the activation of either the paternal or the maternal allele only in somatic cells. Imprinted regions are characterized by increased and specific DNA methylation at particular CpG nucleotides. About 100–200 genes are believed to be imprinted in mammals, including man.
Source: Kahl, G, The Dictionary of Gene Technology.

Genetic Profile Simply, a profile of the variation in one or many genes in an individual or population. In medical genetics, the genetic profile of an individual may be used to predict their likely susceptibilty to disease; in evolution, and particularly in microbial evolution, the genetic profile of a population may be used to track its changes over time or space.

Gene Transfer The main mechanism through which genetic material is transferred between species of bacteria. Bacteria are unable to reproduce sexually, so horizontal gene transfer is the only mechanism other than mutation through which variation is introduced into bacteria. The main methods are the uptake of "naked" DNA from one DNA species into the chromosome of the other, the transfer of plasmids or transposons, and the transfer of DNA using phages.

Gene Trap A method of creating large numbers of insertional mutants in the mouse genome, which is both high throughput and cost-effective. A gene trap vector is inserted at random into the genome of mouse embryonic stem cells, simultaneously disrupting the gene at the site of the insertion. A reporter gene is used to monitor the expression of the inserted gene. The resulting databases of mutant stem cell lines may be used to establish mutant strains of mice via the creation of chimaeras.

Genomic Control A method of reducing the chance of finding spurious associations between genes and disease in the large populations that are necessary for the study of the genetics of complex diseases, caused by population heterogeneity. By studying multiple polymorphisms scattered through the genome, many of which are known not to be associated with the disease in question, it is possible to estimate population heterogeneity and take it into account.

Genomic Imprinting *See* Genetic Imprinting, Imprinting

Genomic Orphan, ORFan *See* Orphan Gene

Genotype The genetic composition of an individual (of any species), as opposed to the physical features imposed by that genotype (the phenotype). The term may be used either to describe the alleles that are present at a particular locus or

to describe the organism's overall genetic composition. Thus, to take a simple case, the genotype of a cystic fibrosis carrier at the CF locus is different from that of an unaffected individual, although their phenotypes are (in that aspect) indistinguishable.

Germ Cell A eukaryotic cell that has been produced by meiosis and that is therefore haploid (containing only one copy of each chromosome). Germ cells are egg cells in the female and sperm cells in the male. Most other types of eukaryotic cell cannot pass information to the next generation and are known as somatic cells.

Germline All cells in an individual that contain genetic material that can be passed on to that person's children are part of the germline. Self-evidently, this includes the egg and sperm cells, but it also includes the cells from which those cells are derived – the gametocytes. If a mutation occurs in a germline cell (a germline mutation), that change may be passed on to future generations.

GFP *See* Green Fluorescent Protein

Gibbs Sampling An algorithm for finding patterns (corresponding to, e.g., structural properties or functional motifs) within a set of DNA or protein sequences. One sequence is left out of the set; the other sequences are aligned and the alignment used to produce a scoring matrix. This is matched to the extra sequence and used to predict the pattern; a second sequence is then left out of the resulting complete alignment and the procedure repeated until the matrix can no longer be improved.

Global Alignment Any sequence alignment technique in which the assumption is made that the (gene or protein) sequences are homologous along their entire lengths. Gaps are inserted into one or both sequences in an attempt to stretch the alignment to cover all the sequences. This method is appropriate for, for example, aligning orthologs from different genomes: it is not appropriate for aligning whole genes with partial ones or cDNA with genomic DNA.

Global Free Energy Minimum The conformation of a molecule (or molecular complex) that has the lowest free energy, as measured (generally) using a molecular mechanics force field. The global minimum is distinct from a large number of local energy minima in different parts of conformational space. Molecular mechanics calculations make the assumption that a molecule or system is most likely to be found in its global minimum, and the difficulty of distinguishing this from the local minima is one of the drawbacks of this methodology.

Global Minimum *See* Global Free Energy Minimum

Glycoconjugate A generic term used to describe any macromolecule that consists of an oligo- or polysaccharide (i.e., a glycan) covalently bound, or conjugated, to another type of molecule. Glycoproteins and glycolipids are examples of glycoconjugates.

Glycomics The study of the glycome of a cell or organism. By analogy with genomics and proteomics, the glycome is defined as the complete set of simple and complex carbohydrates that it makes. The glycome, like the proteome, is many times more complex than the genome.

Glycoprotein A glycoprotein is a protein that is glycosylated – that is, one in which one or (more often) more asparagine, serine, or threonine side chains have been covalently linked to sugar moeities. Glycosylation is a posttranslational modification.

Glycosidic Linkage The covalent link between the protein and carbohydrate parts of a glycoprotein or proteoglycan. There are two main types of such linkages: N-glycosidic linkages, where the oligosaccharide is attached to the amide nitrogen of an asparagine residue, and O-glycosylation, where the oligosaccharide is linked via the side chain hydroxyl group of a serine or threonine residue.

Glycosylation Island A locus on a eukaryotic chromosome that contains genes that code for proteins involved in glycosylation. The genes are close enough together for the locus to be defined as an operon.

Glycosylphosphatidylinositol Anchor One of the three main groups of post-translational glycosylation modifications of protein sequences, the others being O-glycosylation and N-glycosylation. Proteins with GPI anchors are attached to the cell membrane by means of the anchors; they are found in all eukaryotic genomes. It is possible to predict GPI anchor attachment sites from sequence using bioinformatics tools.

Goldberg–Hogness Box, Hogness Box *See* TATA Box

GPI Anchor *See* Glycosylphosphatidylinositol Anchor

Graph, Directed *See* Directed Graph

Graph Theory The branch of mathematics that is concerned with the study of graphs. A graph is defined as an array of points (vertices or nodes) that are connected by lines (edges or arcs). In bioinformatics, graph theory may be used, for example, to analyze the expression patterns of a group of genes. If the edges have a direction (e.g., representing the fact that one gene controls the expression of another), the graph is termed a directed graph.

Greedy Algorithm An algorithm that always takes the best immediate, or local, solution while finding an answer. Greedy algorithms find the overall, or globally, optimal solution for some optimization problems, but may find less-than-optimal solutions for some instances of other problems.
Source: Black, Paul E., NIST Dictionary of Algorithms and Data Structures.

Grid Computing The Grid has been termed “the second-generation Internet”. It is a vision, which is slowly becoming realized, of networked computers set up so

that processing power is as accessible as data is (via the World Wide Web) today. Each computer linked to the grid will be able to "plug in" to a range of services including processing power, communications and storage facilities. The so-called "at home" services in which "spare" PC power is used to solve complex problems, such as Folding@Home in protein folding, are early examples of grid computing. Bioinformatics protocols haved already been set up using the grid.

Hairpin A structural element in either protein or RNA in which a linear chain folds back on itself forming a relatively straight piece of structure with a short loop at one end. In proteins, the two linear regions of chain are beta strands held together with main chain–main chain hydrogen bonds, and the structure is also known as a beta hairpin. In RNA, the linear regions are held together by base pairing, and the structure may also be known as a stem-loop.

Hamming Distance In information theory, the Hamming distance is the number of positions in two character strings where the characters are not identical. The strings are of equal length. This has obvious implications for sequence comparison and pattern matching of genes and proteins, for example, the Hamming distance between the two fragments of protein sequence AFDTGH and VGDTGN is three.

Haploblock A block of DNA sequence that is usually inherited as a whole, at least in a specific population: that is, a block of sequence where linkage disequilibrium is low. The identification of haploblocks is of great value in identifying and mapping genetic associations for complex diseases.

Haploid A eukaryotic cell is described as haploid if it contains only one copy of each chromosome. Thus, a human haploid cell contains one copy of each of the 22 autosomes and one sex chromosomes (either X or Y), making 23 chromosomes in all. Normal gamete (egg and sperm) cells are haploid.

Haploinsufficiency The reduction in gene dosage caused by the mutation of one allele of a gene such that the mutated allele cannot be expressed (i.e., the mutant protein is nonfunctional, truncated, or rapidly degraded). The nonmutant allele, however, is synthesized normally, resulting in the concentration of that protein in a cell being approximately half the normal concentration.
Source: Kahl, G, The Dictionary of Gene Technology.

Haplotype The specific pattern and order of alleles on a chromosome (a specific strand of DNA). Haplotypes tend to be conserved from generation to generation; in particular, alleles that are located close together on a chromosome are likely to be inherited together.

Haplotype Map A map of a chromosome showing the location of specific haplotype blocks. A haplotype block is a block of alleles that are normally inherited together: that is, a stretch of DNA between two areas of high linkage disequilibrium. Haplotype mapping may be used for the detection of genes associated with common, multigenic disorders.

Haplotype tag SNP, htSNP *See* Tag SNP

Helix Bundle *See* Alpha Bundle

Helix Packing The way in which alpha helices pack together in protein structures, to maximize the attractive interactions between the helices. The helix side chains pack together in a way described as the "knobs in holes" model; interhelical angles of 20 degrees (as in four-helix bundles) and 50 degrees (as in the globin family) are preferred.

Heterochromatin This term may be used to mean, either, the part of chromatin that is maximally condensed in interphase nuclei, replicates late in the S phase and is mostly transcriptionally inactive (such as satellite DNA); or, in a different context, the DNA content of the sex-linked chromosomes (such as human X and Y), which are sometimes termed heterosomes or heterochromosomes.
Source: Kahl, G, The Dictionary of Gene Technology.

Heteroduplex Any double-stranded nucleic acid molecule (or duplex) in which the two strands have different origins, whatever those origins are; they may be DNA sequences arising from different genomes or from paralogous genes in the same genome, or they may be an mRNA with its parent DNA. Heteroduplexes may contain loops of single-stranded material lacking a complementary sequence on the opposite strand.

Heterologous Gene Any gene that has been isolated from one organism and transferred into another (i.e., a transgene). Heterologous genes may be contrasted with homologous genes, which are genes that have been taken out of one organism, manipulated (e.g., by introducing site directed mutations) and then transferred back into the same organism.
Source: Kahl, G, The Dictionary of Gene Technology.

Heterozygote Advantage A case where the disadvantage conferred on homozygotes for a particular allele is balanced by an advantage conferred on heterozygotes. If heterozygotes have sufficient survival advantage over individuals without the allele, the allele will increase in frequency despite poorer survival of those with two copies. The allele for sickle cell hemoglobin is a well known example: heterozygotes (with so-called sickle cell trait) are less susceptible to malaria than those without the trait, which balances the disadvantage of homozygotes suffering from sickle cell anemia.

HGP *See* Human Genome Project

Hidden Markov Model A complex, powerful probabilistic prediction technique that has many applications in bioinformatics: for example, predicting gene structure from DNA sequences, protein secondary structure from protein sequences, and classifying genes and proteins into families. The algorithm involves the prediction

of hidden states (e.g., whether a particular base is or is not coding) based on observable ones (e.g., the nucleic acid sequence).

High Pressure Liquid Chromatography HPLC is a very commonly used separation technique with many applications in biotechnology, particularly in proteomics. A complex mixture is passed through a matrix material under high pressure, which separates the components of the complex by mass. HPLC is used for protein separation, protein and nucleic acid purification, and peptide sequencing.

Histone Histones are basic proteins that bind DNA and that are used to package long DNA molecules into the nuclei of eukaryotic cells. This must be a complex process as the average length of a human chromosome when extended is 4–5 cm. DNA-histone complexes are termed chromatin. Posttranslational modification of histone sequences has been implicated in imprinting.

HMM *See* Hidden Markov Model

Holocentric During mitosis, the chromosomes of some eukaryotic species bind to the microtubules along their entire length, and move from there to the poles broadside. These chromosomes are termed holocentric, in contrast to monocentric chromosomes, which bind to the microtubules at the centromere and move toward the poles with that leading. The majority of eukaryotic species, including most model organisms, have monocentric chromosomes; however, the chromosomes of the nematode *C. elegans* are holocentric.

Homeobox A family of genes involved in the control of development in eukaryotes. They code for transcription factors that have been implicated in the formation and differentiation of many tissue and organ types. Homeobox gene sequences are well conserved throughout the evolutionary history of eukaryotes and they have been used to study mechanisms of evolution.

Homeostasis Briefly, homeostasis is the maintenance of equilibrium, or resistance to change. It is a feature of living organisms at all levels, from the molecular, through the cellular to the level of the whole organism. In higher eukaryotes, the maintenance of equilibrium is complex and requires the interaction of many different feedback mechanisms. The mechanisms by which the presence of a metabolite can inhibit the enzyme reactions necessary for its production are very simple examples of these.

Homolog Gene or protein sequences are defined as homologs (or homologous sequences) if and only if they are related by divergent evolution from a common ancestor. Sequence analysis programs determine the degree of identity between sequences; homology can only be inferred from probability, often using functional information.

Homologue *See* Homolog

Homology Modeling A technique for predicting the structure of a protein from its sequence using one or more structures of homologous proteins. This is the most accurate method of predicting protein structure, and can be as accurate as a medium resolution X-ray crystal structure. It is based on a multiple alignment of the test sequence with the sequences of known structure. Generally, conserved regions of structural or functional importance are copied from one of the known proteins and loops are then modeled separately.

Homoplasy Any similarity between two or more sequences (gene or protein), or two or more phenotypic traits, that is not an indication of a common evolutionary origin. Convergent evolution, where the same or a similar solution to a particular problem arises independently more than once, is an example of a process that may lead to homoplasy.
Source: Kahl, G, The Dictionary of Gene Technology.

Horizontal Gene Transfer *See* Gene Transfer

Hot Spot *See* Recombination Hot Spot

Housekeeping Gene A gene that is constitutively active in all cells of an organism and at most developmental stages, because the protein that it encodes is essential for the maintenance of life (e.g., an enzyme that forms part of a general anabolic or catabolic pathway). The concentration of the proteins encoded by these genes is kept at a fairly constant level within the cell. Genes that are only active under some conditions are termed inducible genes. One classic example is the COX family of enzymes; COX-1 is a constitutive gene, whereas COX-2 is induced as part of the inflammatory response.

HPLC *See* High Pressure Liquid Chromatography

Human Genome Project Trivially, the project to sequence the human genome. It was set up in 1990 and expected to take 15 years; however, thanks, largely, to the rivalry between the original public collaboration led by Drs Francis Collins at the NIH and John Sulston at the UK's Sanger Institute, and the private company Celera Genomics founded by Craig Venter it finished 2 years ahead of schedule. The working draft was published in February 2001 and the complete sequence in April 2003. All human genome data is now freely available.

Hybridization The formation of a nucleic acid duplex from two complementary (or near complementary) single strands, either naturally or induced. Hybridization experiments are used to detect sequence similarities and form the basis of microarray technology. In this, which is one of the mainstays of modern bioinformatics, mRNA molecules are detected by hybridization with fragments of complementary cDNA, immobilized on the microarray (or so-called "DNA chip").

Hydropathy plot *See* Hydropathy Profile

Hydropathy Profile A graph that plots the average hydrophobicity of a segment of a protein chain against the amino acid at the centre of that segment. The average hydrophobicity is calculated from the amino acid content of the segment using a hydrophobicity scale. In most widely used scales, very hydrophobic amino acids are given high positive scores, so hydrophobic regions of the sequence – which may, for example, represent transmembrane regions – are represented as "peaks" on the hydropathy plot.

Hydrophobicity Hydrophobic literally means "water-hating". Molecules that are hydrophobic (such as hydrocarbons) are more soluble in oily solvents, such as octanol, than they are in water. Over a third of the amino acids that occur naturally in proteins are hydrophobic; phenylalanine, leucine, and valine are good examples. The fact that these amino acids will be driven into the interior of the protein, away from the solvent, is one of the principal factors driving protein folding.

Hydrophobic Effect The force that drives hydrophobic molecules or parts of molecules (such as hydrophobic side chains in amino acids) away from solvent molecules and into contact with other hydrophobic molecules. The hydrophobic effect drives the formation of the hydrophobic core of globular proteins and is the principal force driving their folding. The solvent accessible surface of proteins is principally formed by hydrophilic amino acids.

Hydrophobic Moment Many alpha helices are significantly amphipathic, with hydrophobic amino acids clustered on one side of the helix and polar and charged ones on the other. In protein structures, amphipathic helices will often be found with the hydrophobic face pointing toward the more hydrophobic environment (the interior of a soluble protein or the lipids of a cell membrane). The hydrophobic moment of a helix is a mathematical concept that measures amphiphilicity, and that is used in protein structure prediction. It is determined by summing the set of vectors in the direction of each amino acid with lengths proportional to their hydrophobicity.

Hypomorphic Allele An allele that produces a protein that has the same function as the wild type protein but with a reduced level of activity, or, alternatively, an allele that produces the wild type protein at lower levels of expression. There will be serious consequences if the function of that gene product is concentration dependent. Hypomorphic alleles are produced by hypomorphic mutations.

Immobilized pH Gradient A polyacrylamide support matrix, which contains chemically immobilized carrier ampholytes such that a stabilized pH gradient is generated along the strip. IPGs allow the separation of larger amounts of protein than is possible using conventional isoelectric focusing techniques.
Source: Kahl, G, The Dictionary of Gene Technology (Second Edition).

Immunocytochemistry A technique that uses antibody-antigen binding to prove protein expression and locate a protein within a cell or tissue. Proteins are located using specific antibodies that are conjugated to dye molecules, and the dye located

under a microscope. All techniques that use dye stains for molecular localization are collectively termed cytochemistry.

Immunodetection *See* Immunoprecipitation

Immunoglobulin *See* Antibody

Immunoprecipitation Any method for locating specific protein antigens in cells or tissues using an antibody that is specific for that antigen that is conjugated with a peroxidase. The antibody-antigen complex is detected by, for example, the peroxide-dependent conversion of luminol, which is accompanied by the emission of light.
Source: Kahl, G, The Dictionary of Gene Technology (Second Edition).

Imprinting Loosely, a phenomenon in which the phenotype expressed by an allele differs according to the sex of the parent who passed on that chromosome. In mammals, the term is usually restricted to those cases where the gene from either the material or the paternal chromosome is inactivated. The gene in question can be referred to as an imprinted gene. In some cases, the phenotype of a genetic disease will depend on whether the defective gene was inherited from the mother or the father. Imprinting is thought to derive from epigenetic differences between the maternal and paternal alleles.

Imprinting Centre Imprinting is a phenomenon in which the phenotype expressed by an allele differs according to the sex of the parent who passed on that chromosome. It arises because some genes from either the maternal or the chromosome are normally inactivated during germ cell development. The chromosomal regions that determine this are known as imprinting centres. Deletions of and errors in imprinting centres give rise to inappropriate imprinting and therefore to genetic disorders.

Indegree In graph theory, the indegree of a node in a directed graph is the number of edges that terminate at that node. This is often applied to the analysis of gene networks derived from microarray experiments, where the relationship denoted by an edge is that one gene affects the transcription of another. A gene with a high indegree is one that is affected by many others, that is, which is highly regulated. Experiments with yeast microarrays have found that most of the genes with high indegree are involved in metabolism.

Indel A shorthand way of expressing “insertion or deletion” in a sequence alignment, expressing the fact that it is impossible to tell (at least without very detailed phylogenetic analysis) whether a gap in an alignment arose from an insertion in one sequence or a deletion in another. In some contexts, the word “indel” may be used synonymously with “gap”.

Index Case In studies of infectious disease, the index case is the first person to become infected with a disease, and so the source of the outbreak. In studies

of genetic disease, the term has been generalized to mean the affected individual through whom an inherited disease-causing mutation is identified in a family.

Inducer A chemical substance, generally of low molecular weight, that binds to a regulator protein and alters its activity in such a way that the transcription of a specific gene or operon, which has previously been repressed, is reactivated. The generic term "effector" is used to indicate a chemical that binds to a regulator and so controls its activity.
Source: Kahl, G, The Dictionary of Gene Technology.

Integrative Biology Integrative biology is often used as a synonym for systems biology. As such, it can be defined trivially as the computer-based analysis or simulation of molecular data within the context of a system. A system may be as (relatively) simple as a metabolic or regulatory network within a single cell, or it may be a cell, tissue, organ, or organism. Integrative or systems biology may therefore include models of different types and of different levels of precision.

Interphase In the cell cycle, the period between cell divisions in which the chromosomes are in an extended form within the cell nucleus and cannot be distinguished separately. Interphase is the phase of the cell cycle during which cells grow and carry out their functions. Cytogenetic tests such as FISH are easier if they can be carried out during interphase, as cell culture is not necessary, but chromosomal abnormalities can usually not be identified.

Intron Those sequences within a eukaryotic gene that are not conserved during pre-mRNA processing and so do not make up the mature message. The introns on the 5′ and 3′ ends of the mRNA may contain sequences that signal initiation or termination of processing, respectively. Prokaryotic genes do not contain introns.

Inverse Problem *See* Inverse Protein Folding Problem

Inverse Protein Folding Problem The problem of finding sequences that conform to (i.e., that are likely to fold into) a given protein topology. It is so-called because it is the inverse of the more common problem of finding the structure that a particular sequence is likely to fold into.

Inverted Terminal Repeat Sequence motifs that flank transposons and that are identical or partly identical and present in inverse orientations. Their function is as recognition sites for the excision of transposons.
Source: Kahl, G, The Dictionary of Gene Technology (2nd Edition).

Ion Mirror *See* Reflectron

Ion Trapping A term used for a group of mass spectrometry methods that are able to measure masses of, for example, peptide ions, with very high precision and accuracy. The peptide ions may be created by any standard MS method (e.g., MALDI or ESI); they are focused into the helium-filled ion trap using an

electrostatic lens. The positions at which the ions are stably trapped depends on the equipment parameters and their mass/charge ratios, and this enables the *m/z* ratios to be calculated.

IPG *See* Immobilized pH Gradient

Isobaric Residues Amino acid residues that have the same molecular mass, and that therefore cannot be distinguished in peptide sequencing using mass spectroscopy (e.g., leucine and isoleucine) are termed isobaric residues.

Isoelectric Point The isoelectric point of a protein is defined as that point on the pH scale where its net positive and negative charge(s) equal zero. During electrophoresis, a protein migrates to a position on a stabilized pH gradient where the pH is equivalent to its isoelectric point.

Isoenzyme *See* Isozyme

Isozyme Multiple forms of the same enzyme, which catalyze the same reaction, but may differ in amino acid sequence, physical properties, and regulation. Isozymes may consist of complexes of different, possibly randomly selected, polypeptide chains. They may be separated by conventional biochemical methods.

Iterative Improvement An algorithmic technique that solves a problem by repeatedly estimating a "slightly wrong" solution, estimating the slight error and subtracting it from the wrong solution to give an improved solution. The process is repeated until the error is smaller than a set value.

ITR, Terminal Inverted Repeat, TIR *See* Inverted Terminal Repeat

Karyotype The complete set of chromosomes in a cell, an individual or a species. The karyotype of a cell or an individual will include gross chromosomal abnormalilties (e.g., in chromosomal number). The word karyotyping is used to describe, generically, a number of techniques for determining the karyotype of an individual; FISH is one example. These may be used to detect aneuploidies such as trisomy 21 (Down's syndrome).

Knockout An informal term for an animal model (very often, but not invariably, a mouse) in which a single gene has been inactivated (silenced or "knocked out") by either random or site directed mutagenesis. Phenotypically, gene knockout animals range from normal to nonviable (i.e., embryonic lethal mutations). The term "knock-in" may be used to describe a model in which the function of an inactive gene is restored by mutation.

Knockout Model, Animal Knockout *See* Knockout

Laboratory Information Management System *See* LIMS

Lagging Strand The DNA strand that is discontinuously synthesized in a 5′ to 3′ direction away from the replication fork during DNA replication. It contains the ligated Okazaki fragments that are linked by ligases to form a continuous strand: each of these is several thousands of nucleotides long in prokaryotes or several hundred nucleotides long in eukaryotes.
Source: Kahl, G, The Dictionary of Gene Technology.

LC *See* Liquid Chromatography

LC-MS/MS *See* Liquid Chromatography/Mass Spectroscopy

LCR *See* Locus Control Region

Leading Strand The DNA strand that is continuously synthesized in a 5′ to 3′ direction toward the replication fork during DNA replication. The opposite strand is the lagging strand, which is synthesized discontinuously.
Source: Kahl, G, The Dictionary of Gene Technology.

Leucine-rich Repeat Short amino acid sequence repeats with a high proportion of leucine residues that are found in tandem arrays in many proteins from different functional families. They are believed to provide a versatile structural framework for the formation of protein–protein interactions, and to be necessary for cytoskeleton morphology and dynamics.

Ligand A generic term for a nonprotein molecule that must be bound to a protein in order for that protein to function. Ligands are usually, but not always, of low molecular weight. In receptor theory, the term ligand is used to indicate the naturally occurring compound that binds to the receptor in order to elicit a response, as opposed to an agonist or antagonist that is added artifically. However, the term may be used to indicate, for instance, an enzyme inhibitor.

Ligase Chain Reaction An in vitro DNA amplification procedure that uses the enzyme DNA ligase to amplify a template. A pair of synthetic oligonucleotides is allowed to anneal to adjacent complementary regions of one strand of the target double stranded DNA, and two other oligos anneal to adjacent complementary regions of the other strand. Each pair of oligos is ligated by DNA ligase, and the ligation product used as a template for subsequent ligation cycles.
Source: Kahl, G, The Dictionary of Gene Technology.

Ligation Amplification Reaction *See* Ligase Chain Reaction

LIMS Computer software used for the automatic management of laboratory functions, which could involve anything from the management of samples and standards to invoicing. LIMS as used to control workflow in complex biotechnology laboratories can be considered a branch of bioinformatics, but it is currently only used to any extent in an industrial context, such as managing high-throughput screening in the pharmaceutical industry.

Lineage Any group of individuals that are derived from a common ancestor may be termed a lineage. Thus, in phylogenetics, the term lineage is synonymous with clade. However, lineage is also used to refer to a family of individual (human or nonhuman) organisms, or, alternatively, a population of differentiated cells derived from an individual precursor (as in "tumour cell lineage").

Linear Ion Trap Ion trapping is a mass spectrometry method that are able to measure masses of, for example, peptide ions, with very high precision and accuracy and in which the peptide ions are focused into the ion trap using an electrostatic lens. A linear ion trap is an enhancement that reduces the number of dimensions of the ion trap from three to two; the ions are trapped radially by a radio frequency containment field, but axially by a static electric field. Linear traps have increased efficiency, sensitivity, and dynamic rang.

Linear Trap *See* Linear Ion Trap

Linkage Analysis *See* Linkage Mapping

Linkage Disequilibrium The occurrence of two or more linked alleles together at a higher frequency than would be expected from their individual frequency in a particular population. The tighter the genetic linkage between a pair of loci is, the higher degree of linkage disequilibrium is observed.
Source: Kahl, G, The Dictionary of Gene Technology.

Linkage Disequilibrium Mapping *See* Association Mapping

Linkage, Glycosidic *See* Glycosidic Linkage

Linkage Mapping The process of deriving a linkage map (or genetic map) of a chromosome location from DNA samples from related and nonrelated individuals, plotting the relative positions of markers based on the frequency of crossovers or recombinations. The genetic distance between two markers – that is, the average number of crossovers during meiosis at the two loci – is given in centiMorgans (cM).

Lipid Raft A small area within a cell membrane that is particularly rich in different kinds of lipids: glycolipids, sphingolipids, and cholesterol. Lipid rafts also contain proteins embedded in the membrane using GPI anchors. Many of these proteins are involved in cell signaling, and lipid rafts are also thought to play a role in signaling processes. They are found in both prokaryotic and eukaryotic cells.

Lipophilicity *See* Hydrophobicity

Lipoplex A complex formed between cationic lipids and DNA, used in nonviral vectors for gene therapy. Complexes formed by cationic polymers for the same reason are termed polyplexes. The DNA in both these types of complexes

is protected from degradation by nucleases. Cationic lipids are very useful as components of gene therapy vectors as they are easy to prepare and characterize.

Liquid Chromatography Any separation technique in which a liquid sample of a complex mixture is passed through a column containing a matrix in such a way that the components of the mixture are separated (e.g., according to their mass). High pressure liquid chromatography (HPLC) has many applications in proteomics and in biotechnology in general.

Liquid Chromatography/Mass Spectroscopy A reliable method for the separation and identification of proteins, involving linking the output from a liquid chromatographic system to a mass spectrometer. Separation of proteins using liquid chromatography is considered to be competitive with the more widely used 2D-PAGE method. Often, the mass spectrometry step is repeated, leading to protein identification: hence LC-MS/MS.

Local Alignment Any pairwise sequence alignment technique in which the assumption is made that the (gene or protein) sequences are not homologous along their entire lengths. Local alignment programs report one or more regions of sequence similarity; where multiple regions are reported, these do not necessarily need to be in the same order in both sequences. This method is appropriate for, for example, aligning whole genes with partial ones, cDNA sequences with genomic DNA, or single domains within multidomain proteins.

Locus The locus of a gene is its location on a chromosome or on a gene map. A single locus may contain several contiguous genes, which are likely to be functionally and/or evolutionarily related: for example, the human cytochrome P450 3A locus on chromosome 7 contains the genes for three different CYP450 isoforms and related pseudogenes.

Locus Control Region Any DNA sequence that exerts a dominant, activating effect on the transcription of genes in a large chromatin domain (10–100 kb). LCRs prevent the influence of heterochromatic silencing on neighboring sequences. They are therefore used in transgenic experiments as insulator elements that protect themselves and linked genes against the repressive action of heterochromatin.
Source: Kahl, G, The Dictionary of Gene Technology.

LOD Score A mathematical description of genetic linkage. The LOD score is defined as the logarithm (to base 10) of the ratio of probabilities that the observed results are produced by linked or unlinked loci. A LOD score of 3 or more indicates that the loci are linked.
Source: Kahl, G, A Dictionary of Gene Technology.

Log of Odds Score *See* LOD Score

Long-branch Attraction In phylogenetic analysis, a phenomenon that is thought to bias any attempt to root the universal tree of life toward a eubacterial root. Since,

in a universal tree, the eubacterial brance is always the longest one, its selection as the universal root may be explained by an attraction between this branch and the long branch of the outgroup.

Long Terminal Repeat The repeat sequences at the ends of a retroviral nucleic acid. In proviruses, the upstream LTR functions as a promoter/enhancer and the downstream LTR as a poly-A addition signal. Long terminal repeats are several hundreds of base pairs in length and the repeated sequence is of 4–6 bp. These sequences can be used as elements of integration vectors.
Source: Kahl, G, The Dictionary of Gene Technology.

Low Complexity Regions Regions of DNA or protein sequence that either repeat a single residue or short residue pattern, or else contain a much higher than average percentage of a particular residue type. In gene sequences, low complexity regions are often microsatellites; in protein sequences they may represent, for example, glycine or cysteine rich regions. Low complexity regions are often masked (ignored) in sequence alignments and searches because they may generate spurious (unrelated) matches.

LTR *See* Long Terminal Repeat

Luciferase An enzyme, isolated from fireflies and some bacteria, which catalyzes the decarboxylation of D-luciferin to oxyluciferin. This reaction generates a flash of light, which can be easily monitored. Luciferin is often used to monitor protein expression, particularly in transgenic cells.

Luciferin Any compound that is a natural substrate for the luminescent enzyme luciferase, which is used in proteomics as a reporter for gene expression. Structurally unrelated substrates of luciferase have been isolated from various species including the firefly *Photinus pyralis* and the ostracode Cypridina. These are generally medium-sized organic molecules containing aromatic heterocycles.
Source: Kahl, G, The Dictionary of Gene Technology.

M, 100 centiMorgans, 100 cM *See* Morgan

Machine Learning An area of artificial intelligence in which a computer is allowed to "learn" the pattern and structure of a dataset by analyzing it, and use that to classify data not in the original dataset. Machine learning overlaps significantly with statistical analysis. It is more popular than other areas of artificial intelligence in modern bioinformatics, finding applications in sequence analysis and in the analysis of microarray data.

Main Chain The backbone of a polypeptide chain, consisting of linked peptide groups and alpha-carbon atoms: thus, the main chain atoms of a single amino acid can be written using standard terminology as –C(O)–C–N(H)–. The peptide group is planar, which restricts the geometric conformation of the main chain. It

is the side chains bonded to the alpha-carbon atoms that give the amino acids, and therefore the proteins, their chemical diversity.

MALDI *See* Matrix Assisted Laser Desorption/ionization

Malecot Model An algorithm for prediction of the decay of linkage disequilibrium with distance, using three parameters. Distance can be measured either in centimorgans or, if it can be assumed that recombination is uniform over the region, in kilobases.

Map-based Cloning; Map-assisted Cloning; MAC *See* Positional Cloning

Markov Chain Analysis *See* Markov Model

Markov Model A probabilistic statistical model used in many bioinformatics applications. One example is its use in sequence analysis, where the probability of each nucleotide or amino acid occurring is dependent on those preceding it. A hidden Markov model (HMM) is a Markov model in which one or more variables are hidden.
Source: Westhead et al, DR, Instant Notes in Bioinformatics.

Mass/charge Ratio *See* m/z Ratio

Mass Fingerprint *See* Peptide Mass Fingerprint

Mass Resolution The extent to which a mass spectrometer can distinguish between samples of similar mass. Modern Fourier Transform mass spectrometers have very good mass resolution, being capable of identifying peptides and even distinguishing between isotopes.

Mass Spectrometry In proteomics, mass spectrometry is used to identify proteins from small samples that have been separated by, for example, 2DPAGE. The proteins are first fragmented into peptides using proteases (typically trypsin). Mass spectroscopy involves the ionization of the peptide sample, the separation of the ions by mass-charge (m/z) ratio, and the analysis of the separated ions and identification of the protein from its constituent "peptide fingerprint". Different technologies exist for peptide ionization (e.g., MALDI and ESI), ion separation (e.g., time-of-flight) and mass fingerprint analysis.

Mass Spectrometry, Tandem; Tandem Mass Spectroscopy; MS/MS *See* Tandem Mass Spectrometry

Mass Tolerance A measure of the precision expected from a mass spectrometry experiment, which is used in determining whether, for example, an experimentally measured ion corresponds to a certain peptide. The Mowse scoring algorithm for MS matches states that "each calculated value which falls within a given mass

tolerance of an experimental value counts as a match". Typical mass tolerance values are 2.0 a.m.u. for peptides and 0.8 a.m.u. for fragment ions.

Matrix Assisted Laser Desorption/ionization A technique used in protein mass spectroscopy, as applied to the analysis and identification of separated proteins obtained in proteomics experiments (e.g., 2DPAGE). A solution of the digested peptides is passed through a thin needle with a nebulizing gas, and a high voltage applied to the tip. This generates a spray of droplets containing ions. The droplets evaporate leaving peptide ions that pass through a series of electrodes and samplers to the mass analyzer, which usually works on the time-of-flight principle.

Maximum Likelihood A statistical method pioneered by a geneticist, Sir Ronald A. Fisher. It is a method of point estimation that estimates the value of an unobservable parameter as that value that maximizes the likelihood function. The log of the likelihood (the log-likelihood) is an often quoted value.

Maximum Parsimony One of three methods commonly used in phylogeny to select the most probably phylogenetic tree relating a set of sequences. In maximum parsimony, the "correct" tree is assumed to be the one that minimizes the number of step changes (i.e., single base or amino acid changes) from the presumed common ancestor that are needed to complete the tree. It generates unrooted trees; it is a very reliable method, but is time consuming and CPU intensive, so best used with small numbers of similar sequences.

MCS *See* Multispecies Conserved Sequence

MD *See* Molecular Dynamics

Membrane Anchor A single segment of protein chain that either embeds in, or passes through, a cell or organelle membrane, anchoring the protein to that membrane. Proteins containing membrane anchors must contain either a cytoplasmic or an extracellular functional domain, and may contain both these: the function of the anchor is to attach that protein to a particular point at the surface of the cell or organelle. If such a protein contains more than one domain, the anchor is bound to act as a domain boundary.

Membrane Protein A protein that passes through a cell membrane, either once or more than once. Apart from proteins that are embedded in the outer membranes of Gram negative bacteria, which are beta-barrels, all membrane proteins contain one or more transmembrane helices. Type I and type II membrane proteins have a single transmembrane helix separating extracellular and cytoplasmic domains. Integral membrane proteins contain helix bundles that are embedded in the cell membrane. Some types of membrane protein contain signal anchor sequences towards their N-terminal ends.

MEMS *See* Microelectromechanical Systems

Mendelian Disease A genetic disease or disorder that is carried by a single gene. Mendelian diseases have a penetrance approaching 100%, that is, all people who carry an abnormal variant of the gene on one or more alleles (depending on the inheritance pattern) will suffer the disease to a greater or lesser extent. Examples include cystic fibrosis (recessive inheritance), Huntington's disease (dominant inheritance), and hemophilia (sex-linked inheritance).

Metabolic Pathway The linking of small, biosynthetic molecules via the enzymes that synthesize them in the normal metabolism of any species, to form a network; one widely studied example is the glycosidic pathway through which glucose is hydrolyzed to pyruvate with the release of ATP (energy). Information about metabolic pathways is held in databases including KEGG and WIT.

Metabolomics The study of the metabolome. This is defined, by analogy with "genome" and "proteome" , as the sum total of all metabolites (the "small molecules" that are substrates, intermediates and products in metabolic reactions within a cell. Like the proteome, the metabolome varies between cell types and, within a cell type, according to developmental stage and environmental conditions.

Meta-Data Data held within a database that is accessory to and associated with the primary data in the database. For example, the metadata held in a protein sequence database might include gene name and chromosomal location, Gene Ontology annotations, enzyme activity and metabolic pathway involvement. The term metadata may also be used to describe information about an HTML document that is held within the file but not displayed by a browser.

Metadata *See* Meta-Data

Metaphase The phase during eukaryotic cell division (mitosis or meiosis) between prophase and anaphase, in which the nuclear membrane has broken down and the daughter chromosomes align in the center of the cell before being drawn toward its ends by the microtubules. This is the stage in mitosis when the chromosome pairs are most clearly visible, so it is useful for cytogenetic analysis.

Meta Server A server on the Internet that provides access to a number of other servers that provide programs with very similar functions (but probably different methodologies), for example, protein structure prediction, allowing users to compare the results of the different programs. Groups that provide meta servers often do not provide their own methods, but do give anaylsis of the different methodologies.

Metric Map Any map of a genome or chromosome, whether defined by linkage, marker or polymorphism, in which the distance between the elements is recorded as well as their order. Linkage disequilibrium maps may be made into metric maps when the linkage disequilibrium is plotted against the physical distance between markers on the chromosome.

MIAME A series of guidelines set up by the MGED (Microarray Gene Expression Data) Society to enable sharing of microarray data within the gene expression profiling community. The guidelines are designed so one group will be able to reproduce exactly a microarray experiment produced by another. This has been facilitiated by the invention of an XMA format markup language, MAGE-ML, for the storage of microarray data.

MIAPE A series of guidelines set up by the Human Proteomics Society, by analogy with the MIAME guidelines for microarray experiments, to enable sharing of proteomics data within the community. The guidelines are designed so one group will be able to reproduce exactly a proteomics experiment produced by another.

Michaelis–Menten Equation

Microarray An ordered array of (usually) cDNA fragments, arranged at extremely high density on a solid support, and used for analysis of the mRNA content (transcriptome) of a cell. The experiment is set up so that a signal is generated if the sample contains mRNA molecules that can hybridize to a given cDNA.

Microchimerism A relatively common phenomenon in which cell lines with different chromosomal compositions are found in one individual. Unlike mosaicism, however, in microchimerism, the cells are derived from two separate individuals. Sometimes cells are exchanged between twin fetuses in the uterus (so-called twin-to-twin transfusion); more often, there is an exchange of cells between mother and fetus during pregnancy, and the mother's cells may persist throughout the lifespan of the offspring. Microchimerism has been implicated in a number of autoimmune diseases.

Microelectromechanical Systems Extremely small, microfabricated electrophoresis systems that have been proposed as a potential solution to the remaining cost limitations of genome sequencing. The technology requires multichannel devices and the ability to process samples on the nanoliter scale. Many such devices have short read lengths and these may be most suited to resequencing or genotyping.

Micro-RNA *See* miRNA

Microsatellite Any short (typically 1–6 bp) tandem repeat in a genome sequence, that is, any short base pattern repeated a number of times. Microsatellites are common throughout eukaryotic DNA. They are often "masked" in sequence searches because a microsatellite match may swamp a match to a distant homolog. Most microsatellites occur in intergenic DNA (so-called "junk DNA") but occasionally one occurs in a coding region, for example, the (CAG)n motif in the huntingtin gene which is expanded in Huntington's disease patients.

Middleware A type of software that is used as an intermediary between different components; for example, the different components of software that sit between a

database user on a client system and the database server. There is a sense in which an ontology can be described as a piece of middleware.

Minimum Information about Microarray Experiment *See* MIAME

Minimum Information about Proteomics Experiment *See* MIAPE

Minisatellite A short, repetitive, usually GC-rich tandemly arranged DNA sequence. Minisatellites (9–64 bp) are longer than microsatellites (1–8 bp). They occur in all eukaryotic genomes, but are more common in large genomes of complex organisms. Minisatellites tend to show significant length polymorphism.

miRNA Very small mRNA molecules, only 20–25 nucleotides long, that are involved in the regulation of gene expression. They are transcribed from DNA sequences, initially as longer sequences that contain the miRNA and an almost self-complementary sequence that forms a hairpin. The mature miRNA is cleaved out of the precursor sequence by enzymes. It is complementary to part of a coding gene and may anneal to the mRNA, preventing protein translation.

Missegregation Any process by which chromosomes fail to segregate correctly during cell division, leading to the formation of daughter cells with abnormal and/or missing chromosomes. Chromosomal missegregation often occurs during the division of cancer cells, leading to further errors. The production of abnormal and even extra spindle poles has recently been implicated in this process.

Missense Mutation An ordered array of (usually) cDNA fragments, arranged at extremely high density on a solid support, and used for analysis of the mRNA content (transcriptome) of a cell. The experiment is set up so that a signal is generated if the sample contains mRNA molecules that can hybridize to a given cDNA.

Mitosis The process of cell division that takes place in eukaryotic cells at all times except gametogenesis, and in which the chromosomes are replicated, maintaining chromosome number. Thus, one diploid cell will – in the absence of replication errors – produce a pair of identical diploid cells.

Model-based Analysis A type of test used in statistical genetics in which the frequency and penetrance of an allele that has been implicated in disease can be estimated with sufficient accuracy to be used in a mathematical model. It is most commonly used for simple genetic diseases; model-free analysis is usually used to model complex diseases. The alternative terms of parametric and nonparametric analyses are regarded as less accurate because some mathematical parameters are generally used in model-free analysis.

Model Organism An organism that is widely studied by geneticists not because of its pathogenicity or utility but as a genetic "model" for higher organisms. Model organisms are generally common, small, and tractable, and have short

life cycles: thus, the nematode word, Drosophila, Arabidopsis and the common laboratory mouse are all model organisms. By 2004, the genomes of most organisms commonly used as models had been made publicly available.

Modular Protein *See* Mosaic Protein

Module Domains within proteins may also be referred to as modules. This terminology is most often used of domains that are relatively small, that are present in many protein families with different functions, and that can occur multiple times in the same protein. The immunoglobulin, SH2, and SH3 domains are examples of domains with these properties.

Molecular Clock The molecular clock hypothesis is the assumption that evolution occurs at the same rate along branches of a phylogenetic tree that emerge from the same node – that is, that branches of a tree that share a common node will be of the same length. It is often a reasonable assumption, particularly if the sequences are closely related, but there are many instances where it cannot be applied because one taxon has undergone more mutations since divergence than another. This hypothesis is built into some phylogeny methods.

Molecular Dynamics A molecular modeling technique in which the motion of a single molecule or, more often, a molecular system (such as a protein and its ligands in a "bath" of solvent molecules) is simulated. This allows a fuller exploration of conformational space than the related technique of energy minimization. Most often, the molecules are described using a simple molecular mechanics force field: nevertheless, it is very expensive in CPU time. Simulations cover times that are typically of the order of nanoseconds.

Monocistronic A messenger RNA is defined as monocistronic if it codes for a single polypeptide chain (i.e., a single protein). An mRNA that codes for more than one protein, such as that produced from a single prokaryotic operon, is said to be polycistronic. The majority of eukaryotic mRNAs are monocistronic.

Monogenic Disease *See* Mendelian Disease

Monophyletic In phylogeny, a taxonomic group is defined to be monophyletic if all organisms in that group are known to be descended from a common ancestor, and if all the descendants of that ancestor are included in that group. Thus, the genus Homo is classified as monophyletic because all organisms in that genus are believed to derive from a common ancestor, and no other descendants of that ancestor occur outside Homo. Taxonomists prefer to define monophyletic groups if at all possible.

Monte Carlo Algorithm A type of numerical method that involves statistical simulation using sequences of random numbers. In bioinformatics, Monte Carlo methods are regularly used, for example, in simulating the motion of a macromolecule or complex.

Morgan A measure for the relative distance between two genes on a chromosome, or for the frequency of recombination between two genetic markers. One Morgan corresponds to that length of chromosome in which, on average, one recombination event occurs each time a gamete is formed. Genetic distances are more usually recorded in centiMorgans (0.01 M).
Source: Kahn, G, The Dictionary of Gene Technology.

Mosaicism A type of genotype in which two cell lines with different chromosomal compositions, derived from a single fertilization, are found in a single individual. Generally, one cell line will be normal and the other contain a chromosomal aberration such as aneuploidy. The resulting phenotype depends on the proportion of abnormal cells as well as the type of aberration, and ranges from normal through minor abnormalities to malformations incompatible with life, posing serious problems in genetic counseling.

Mosaic Protein A protein that is composed of a number of different domains (or modules). Some mosaic proteins contain very large numbers of domains. The domains that are present in many mosaic proteins are often relatively small, and some domains are found in an enormous range of different proteins with a wide variety of functions. A protein containing only 2–4 domains would not be termed mosaic: it is not a synonym for multidomain.

Motif A (generally small) sequence of amino acids within a protein sequence, or bases between a nucleic acid sequence, that are characteristic of a particular family, a generic function or a structural pattern. Examples of protein motifs include the helix-turn-helix and the zinc finger, which both bind DNA. The smallest motifs, which can involve only 3 or 4 amino acids, represent potential locations of posttranslational modifications, The main database of protein motifs is PROSITE.

MS, Mass Spectroscopy *See* Mass Spectrometry

Multiallelic A gene, or a genetic marker that has more than two forms; in contrast, almost all single nucleotide polymorphisms (SNPs) have only two base variants (e.g., a position may be A or T but not G or C) and are therefore termed biallelic. Genotyping individuals at the sites of multiallelic markers can be very useful in the mapping of genes involved with complex diseases.

Multifocal A disease that is present at more than one site in the body (i.e., which has more than one focus) is termed a multifocal disease. Bilateral breast cancer, in which the cancer is found in both breasts, is an example. Where a disease, such as breast cancer, is heterogenous and is only sometimes multifocal, the presence of multifocal disease is one characteristic that can suggest a high genetic component and thence increase risk in blood relatives.

Multigenic Disease A disease or deleterious trait that is caused by mutations in many genes, rather than, as is the case in monogenic disorders, by a single mutation in one gene. Many common diseases, such as asthma, some types of cancer and

some forms of heart disease, are multigenic. The same disease phenotype may have many possible complex genetic causes.

Multiple Alignment An alignment of more than two gene or protein sequences. Each row in a multiple alignment consists of a single sequence padded by gaps, with the columns highlighting similarity/conservation between positions. An optimal multiple alignment is one with the highest degree of similarity between the sequences. CLUSTAL is a commonly used public domain multiple alignment program.

Multiple Marker Screening Any test that involves obtaining values from several different markers and combining their results to predict the most likely outcome. The term is often applied to one particular test: the measurement of alphafetalprotein and hormone levels in an attempt to detect pregnancies with a high probability of Down's syndrome or another genetic abnormality. These tests generally give a large number of false-positive results.

Multispecies Conserved Sequence A sequence – generally a DNA sequence – that is conserved throughout a large number of species, often highly divergent species. Highly conserved regions have been subjected to extremely strong evolutionary pressures, and therefore code for elements that are necessary for the survival of complete clades (e.g., all vertebrates).

Mutagen Any physical or chemical agent that increases the frequency of mutations in DNA above the spontaneous background level. Mutagenic agents include ionizing radiation, UV irradiation, chemicals (e.g., alkylating agents) and nucleotide base analogs. Mutation may take place in the test tube or in vivo.
Source: Kahl, G, The Dictionary of Gene Technology.

Mutagenesis The process of introducing a change – that is, a mutation – into a DNA sequence. Mutagenesis does, of course, occur naturally, and it may be silent (produce no change in the resulting protein). However, the term is most often used to indicate an artificially induced change. Point mutations are introduced into protein sequences via site directed mutagenesis. The term is also used for methodologies used to create strains of transgenic mice (e.g., gene-trap mutagenesis).

Mutagenesis, Site-directed; Site-specific Mutagenesis *See* Site directed mutagenesis

Mutagenic Agent *See* Mutagen

Mutation Any compositional change in a DNA sequence that is not caused by normal segregation or genetic recombination. Mutations may involve base changes (giving rise to single nucleotide polymorphisms), insertions or deletions; they may occur in coding or noncoding sequence. Mutations in coding sequence will lead to a change in the protein sequence unless the change is to a synonymous codon;

mutations in noncoding sequence may have phenotypic consequences if they change the expression patterns of genes.

Mutation, Missense *See* Missense Mutation

Mutation, Nonsense *See* Nonsense Mutation

***m/z* Ratio** The ratio of the mass of a molecular ion to its charge. This is the quantity by which particles are sorted in a mass spectrometer in the separation experiments that are key to peptide identification in proteomics (most usually by time of flight). It is, therefore, important to work out the charge of each species if its mass – by which it is identified – is to be calculated correctly.

N-glycosylation One of two types of linkage between the side chain of an amino acid within a protein and a simple sugar or oligosaccharide. The sugar moiety is attached to the protein chain via a covalent bond between *N*-acetyl-D-glucosamine and a nitrogen atom in the side chain of an asparagine (N) residue which must lie in the context of the simple motif N-X-S/T).

Needleman–Wunsch Algorithm A dynamic programming method of aligning pairs of sequences that produces a global alignment between the whole sequences. The alignment score is defined as the sum of the scores at each individual position; the sequences are moved and gaps introduced to maximize the total score along the sequence lengths. Gap penalties are often, but not always, applied to gaps at the end of sequences. This method is used in, for example, the EMBOSS global alignment program, NEEDLE.

Network Complex relationships between entities may be represented using networks of connections. Each entity (a gene or protein) is represented as a point, or node, in the network, and the relationships between them are represented by lines joining nodes (known as edges). Graph theory is used to classify and cluster the nodes in a network, discovering relationships that may not be visible from a simple examination of the raw data.

Neural Network A programming methodology often used in bioinformatics for predicting features from sequence data (e.g., predicting genes from DNA sequences or protein structure from amino acid sequences). A neural network program consists, simply, of a series of “neurons” that read data (e.g., sequences) and pass information about that data as signals to other neurons; the final neuron makes the prediction. Sequences containing the known features must first be used to “train” the network.

NMR *See* Nuclear Magnetic Resonance

Node In graph theory, the objects are known as nodes, and they are connected by lines indicating relationship; these are edges. The edges may or may not have directionality, depending on the relationship modeled. Graph theory has many

applications in bioinformatics, where is used to cluster the most closely related objects (typically genes or proteins) together. Nodes are most often used to represent individual genes or proteins, connected by relationships such as “is coexpressed by” (for genes) or “interacts with” (for proteins).

Nonsense Mutation Any mutation that converts a sense codon (coding for an amino acid) into a stop codon (TTA, TAG, or TGA in the standard code), or, conversely, a stop codon into a sense codon. This leads to the production of a polypeptide chain that is either truncated or extended, and, consequently, the function of the protein will be either severely limited or completely abolished.

Nonsynonymous A base change (mutation) is described as nonsymonymous if it occurs in coding DNA and gives rise to a change in the amino acid that is coded for: thus, a change from T to A that changes the codon CAT to CAA, and thus changes the amino acid histidine to glutamine in the resulting protein is synonymous, whilst a T–A change that changes CCT to CCA is synonymous as both codons code for proline. Nonsynonymous changes are self-evidently more important in evolution than synonymous ones; they are also less common in coding DNA.

Normalization The equalization of the concentrations of transcripts present in a cell at extremely different levels, balancing the unequal representation of the messages in a cDNA library (which often vary by more than 5 orders of magnitude) by reducing the number of highly expressed mRNAs and enriching rarely expressed message.

Northern Blotting A gel blotting technique in which RNA molecules, separated according to size by agarose or polyacrylamide gel electrophoresis, are transferred directly to a filter by electric or capillary forces. Single-stranded nucleic acids may be fixed to the filter by baking and are thus immobilized. Hybridization of single-stranded probes to the immobilized RNAs allows the detection of individual RNAs out of complex mixtures.
Source: Kahl, G, The Dictionary of Gene Technology.

Nuclear Magnetic Resonance An analysis technique in which molecules are identified, and molecular structures detected, by monitoring signals generated by certain atomic nuclei (those of nonintegral spin, most often protons) in oscillating high magnetic fields. Two-dimensional nuclear magnetic resonance (2D NMR) is often used for determining protein structures. This technique has the advantage of generating the structures of proteins in solution, but the disadvantage that it can only be used with relatively small proteins.

Null Hypothesis In statistical analysis, a hypothesis is chosen at the beginning of an experiment; the objective is to collect enough data to prove or disprove that hypothesis. The null hypothesis states that a condition, for example, that a given proportion of the data has a particular value or range of values, (or will not) be met; the objective of the test is to accept or reject that hypothesis.

O-glycosylation One of two types of linkage between the side chain of an amino acid within a protein and a simple sugar or oligosaccharide. The sugar moiety is attached to the protein chain via a covalent bond between *N*-acetyl-D-galactosamine and the hydroxyl group of a serine or threonine residue in most protein, or of a (nonstandard) hydroxylysine residue in the protein collagen.

O-mannosylation The transfer of a mannose residue to dolichyl activated mannose to serine or threonine residues of secretory proteins, catalyzed by protein *O*-mannosyltransferases. Mannosylation was first observed in fungi, but mannosyltransferase orthologs have now been identified in the genomes of higher eukaryotes.

Object Oriented Programming A programming paradigm, adopted in the programming language C and more modern languages influenced by it including C++ and Java, in which data types are defined as objects. An object includes both data and the operations (functions) that can be applied to it. Most programming languages frequently used in bioinformatics are fully object oriented; one exception is the popular and easy to learn scripting language, Perl.

Obligate Able to live only in a particular set of conditions; that is, an obligate parasite is unable to survive and reproduce outside its host. The bacterium *Chlamydia trachomatis* is an obligate intracellular human pathogen that is unable to reproduce outside human cells.
Source: http://www.biology-online.org/dictionary.asp?Term=Obligate.

Oligo *See* Oligonucleotide

Oligomer A relatively small number of molecular units joined or associated together. These may be covalently bonded, as in nucleotides (to form an oligonucleotide) or amino acids (to form an oligopeptide or, simply, a peptide) or noncovalently associated, as in several protein chains forming a functional protein complex. Associations of two, three and four units are termed dimers, trimers and tetramers respectively. The DNA double helix is, therefore, a noncovalently bonded dimer.

Oligonucleotide A short segment of nucleic acid, which may be single- or double-stranded. The term is generally used for segments containing up to 100 nucleotides or base pairs. The short form "oligo" is almost always used informally by experimental molecular biologists. Oligos may consist of deoxy- or ribonucleotides, or of a mixture of the two.

Oligosaccharide A molecule made up of a relatively small (say 10–100) number of sugar units (=monosaccharides), joined together by condensation reactions to form linear or branched chains. Oligosaccharides are frequently attached to protein molecules to form glycoproteins. Larger numbers of monosaccharide linked together form polysaccharides (also termed complex carbohydrates).

Oncogene Genes that control normal cellular growth and development are known as proto-oncogenes. In normal cells, these are kept under tight control, so growth and development signals are only sent when required. When an oncogene is mutated (by point mutation or simply gene amplification), it can become altered so its protein product is always activated, so growth/division signals are always sent. Uncontrolled cellular growth and development is the hallmark of cancer; the altered proto-oncogene is known as an oncogene.

Ontology In computer science and allied fields, the word ontology – defined philosophically by Aristotle as ”the science of being qua being – is used to describe a strict conceptual schema of data or concepts within a given domain. This has been applied to the derivation of structured, consistent vocabularies in different areas of knowledge, including the life sciences. The most well known ontology in the molecular life sciences is undoubtedly the Gene Ontology (http://www.geneontology.org).

Open Reading Frame Trivially, a region of genome sequence that starts with an initiation (START) codon and ends with a termination (STOP) codon, and so is translated into protein. A scan of a genome sequence for long ORFs is the first and easiest stage of gene prediction. In practice, the situation is much easier in prokaryotic genomes than in eukaryotic genomes, which are complicated by the extreme length of some genes, the presence of introns, and the necessity of identifying splice sites.

Open Reading Frame EST Sequencing *See* ORESTES

Open Source A software product that is not only deliberately given away free, but where the code is made freely available and where modification is not only allowed but encouraged, is termed open source software. It may be protected by agreements that are analogous to the way copyright laws work in the commercial sector; one of these is known as “copyleft”. Examples of open source software products include the Linux operating system for PCs and the general bioinformatics package EMBOSS (European Molecular Biology Open Software Suite).

Operator The stretches of prokaryotic genome sequence, adjacent to the promoter regions of genes, that regulate gene expression by binding proteins. The first regulatory mechanism to be understood was that of the lactose operon: here, it is the binding of the lac repressor to the operator region that prevents the attachment of RNA polymerase and therefore gene expression.

Operon Operons are only found in prokaryotes. They are series of genes, normally functionally related, that are adjacent on the bacterial chromosome, are under the control of a single promoter, and are synthesized into a single, polycistronic mRNA that is translated into the constituent proteins.

ORESTES Normally, ESTs are derived from the $3'$ and the $5'$ ends of cDNAs, and fragments from the centre of transcripts are underrepresented in EST libraries.

ORESTES is a novel technique for generating ESTs that preferentially amplifies the central portion of transcripts, and which can therefore be used to add many novel sequences to EST databases. It involves the amplification of the expressed gene transcripts by reverse transcription-PCR using arbitrarily chosen primers.

ORF *See* Open Reading Frame

Origin *See* Origin of Replication

Origin of Replication The sequence or region on a DNA strand or chromosome where replication begins – that is, the replication-initiation focus. In eukaryotes, the segment of DNA that is under the control of one replication-initiation focus, and which therefore acts as an autonomous unit during replication, is termed a replicon.
Source: Kahl, G, The Dictionary of Gene Technology, 2nd edition (Wiley-VCH, 2001).

Orphan Gene A gene that does not have any known orthologs in any other species – that is, a gene that is, as far as is known, found in one species only. Generally, the function of an orphan gene is unknown. The term is also applied to open reading frames (ORFs) that are not (yet) validated genes, hence the alternative term ORFan. Of course, it is possible that a gene that is thought to be an orphan gene may not be because its homologs are distant enough to be undetectable at the sequence level or because all its orthologs are in genomes that have not yet been sequenced.

Ortholog Two homologous (evolutionarily related) genes are defined to be orthologous (i.e., they are orthologs of each other) if they are essentially the same gene, with the same function, in different organisms. Thus, human hemoglobin, mouse hemoglobin, and sperm whale hemoglobin are orthologs.

Outdegree In graph theory, the outdegree of a node in a directed graph is the number of edges that start at that node. This is often applied to the analysis of gene networks derived from microarray experiments, where the relationship denoted by an edge is that one gene affects the transcription of another. A gene with a high outdegree is one that affects many others, that is, which is a central regulator of the network. Experiments with yeast microarrays have found that most of the genes with high outdegree are transcriptional regulators.

Outgroup A sequence (or group of sequences) included in a phylogenetic analysis precisely because it is known to be more distantly related to the other sequences than any of them are to each other. The outgroup will diverge from the other sequences near the root of a rooted tree. Outgroups are useful as external references, and including one may lead to more accurate ordering of the other sequences.

Pair Potential In molecular mechanics calculations or (much more often) molecular dynamics simulations, parameters to be used in equations defining

the non-bonded interactions between different types of atom. Each pair of atom types will have a different pair potential. The parameters are inserted in a standard equation defining nonbonded interactions (e.g., the Lennard–Jones or the Buckingham potential equation).

Palindrome In the study of language, a palindrome is a word or sentence that reads the same forward as backward, but that nevertheless makes sense: one example is the word MADAM. In genetics and genomics, the word is used analogously to mean a sequence of DNA where identical sequences run in opposite directions, so each strand reads the same in the 5′ to 3′ direction. Palindromic DNA sequences can be the target for DNA binding proteins and they often occur in regulatory regions of DNA.

PAM Matrices One of the two most widely used sets of matrices that hold data on the evolutionary distance between amino acids (i.e., the probability that a substitution of one amino acid by another will be accepted), the other being the BLOSUM matrices. PAM stands for "point accepted mutation" although "accepted point mutation" would be clearer. The PAM 1 matrix is the substitution matrix for a situation where exactly one mutation has occurred per 100 amino acids. The most widely used matrix is PAM 250, which corresponds to approximately 20% identity between the sequences.

Panmixis Simply, random mating – that is, sexual reproduction where the choice of mates is not influenced by their genotypes. The word is derived from the Greek word mixis (mixture).

Paralog Two homologous (evolutionarily related) genes are defined to be paralogous (i.e., they are paralologs of each other) if they have different (although almost always related) functions. They may or may not occur in the same genome: paralogs that occur in the same genome will have evolved through gene duplication. Thus, human hemoglobin and human myoglobin are paralogs, but so are human myoglobin and sperm whale hemoglobin.

Parametric Analysis *See* Model-based Analysis

Parsimony *See* Maximum Parsimony

Pattern Recognition Trivially, any tool or technique for recognizing patterns in sequences. The technique of pattern recognition is most often applied to protein function detection, as short groupings and/or more complex patterns of amino acids often have implications for the function of the protein. The database PROSITE contains data on many hundreds of amino acid patterns that have been associated either with protein families, functions, or (for the shortest patterns) posttranslational modifications.

PCR *See* Polymerase Chain Reaction

Pedigree Simply, a chart or diagram showing the relationships within a human (or model organism) family that can be used to study the inheritance pattern of an allele, marker, or disease. Large pedigrees over many generations and within relatively isolated populations, such as those studied in Decode's Icelandic genome project, have been used to map the loci and alleles involved in complex diseases.

Penetrance A gene is said to have high penetrance if the properties that it codes for will always or almost always be present in the phenotype, and low penetrance if the amount to which it is observed in the phenotype is more dependent on environmental variables. Thus, the CFTR gene has higher penetrance than the BRCA1 gene because a mutation in CFTR will almost always cause cystic fibrosis, whereas one in BRCA1 only increases the lifetime chances of contracting certain cancers.

Peptide Fingerprint *See* Fingerprint

Peptide Mass Fingerprint Analysis of a protein by mass spectroscopy produces a series of masses of the peptides that were generated from the original protein by protease cleavage. Knowing the mass series and the protease used, it is often possible to identify the protein. The mass series is referred to as a peptide mass fingerprint (or mass fingerprint) for the protein concerned. Mass fingerprinting cannot be perfectly reliable because of, for example, the existence of isobaric residues: sequencing at least parts of the fragments is often needed to fully identify the protein.

Peptide Sequence Tag A short string of peptide mass differences corresponding to a peptide sequence that can be used to identify a longer protein. In a technique developed by Matthias Mann and Matthias Wilm at EMBL in the 1990s, mass spectra of protein fragments derived from MS/MS analysis of proteomics experiments are searched for the presence of sequence tags, and these are used to identify the original protein.

Pericentromere The regions of eukaryotic chromosomes that immediately flank the centromere. Like centromeres, pericentromeres contain a large proportion of repetitive sequences and few genes. The pericentromere is a structural domain of the chromosome that is essential for chromosomal segregation; it has been implicated in the cohesion of the chromosome pairs.

Perl A programming or scripting language that is particularly useful for interpreting and reformatting large quantities of textual data. It is available for all common computer platforms; it is regarded as being easy to learn and use and is the programming language of choice of most bioinformatics professionals who were not trained as programmers. Libraries of Perl scripts for bioinformatics tasks (e.g., BioPerl) have been made available.

Phage Trivially, any virus that infects bacteria. Phages are simple viruses, consisting of a core of either DNA or RNA surrounded by a protein coat. Some phages

are virulent; infection with a virulent phage inevitably leads to viral replication, and the death and lysis of the host cell. Other phages, known as temperate phages, may insert their DNA into the host chromosome where it remains transcriptionally silent. Phages have many uses in modern molecular biology.
Source: Kahl, G, The Dictionary of Gene Technology.

Phage Display A technique for the presentation of distinct proteins or peptides on bacterial surfaces, using bacteriophages as carriers. Genes for the proteins to be displayed are integrated into the phage genome, and the proteins expressed as fusions with a viral coat protein. This exposes the display proteins on the bacterial surface. The technique enables the identification of proteins with particular binding properties.
Source: Kahl, G, The Dictionary of Gene Technology (2nd Edition).

Pharmacogenetics The influence of genetics and, especially, of genetic variation, on pharmacology: that is, how differences in people's genetic makeup influence their response to drugs. One important aspect of this is the effect of common polymorphisms in the P450 protein family on drug metabolism on the optimum dose of each drug for different individuals.

Phenotype The observable characteristics of an organism. These may be structural, functional, or (with higher animals, particularly man) behavioral, and they derive from both the organism's genetics and its environment.

Phosphorylation Generally, the addition of a phosphate (PO3-) group to any (usually organic) molecule. In proteomics, the addition of a phosphate group to the hydroxyl group of a serine, threonine or tyrosine residue of a protein. Protein phosphorylation is controlled by the large kinase family of enzymes and is extremely important in cellular signaling pathways.

Phylogenetic Footprinting A bioinformatics technique for identifying regulatory elements in DNA by locating regions of orthologous noncoding DNA that show unexpectedly high conservation between species.

Phylogenetic Marker A gene, coding for either RNA or protein, that can be used for phylogenetic analysis because changes in its sequence can be consistently followed throughout a relevant period of evolutionary history. Genes that are highly conserved throughout long evolutionary distances, such as RNA genes and certain essential proteins such as vacuolar ATPases (ubiquitous in eukaryotes) and cytochromes, are commonly used as phylogenetic markers.

Phylogenetic Profile A binomial string that describes the presence or absence of a particular gene in all fully sequenced genomes – thus, if a gene is present in a species' genome a "1" will be entered in that position in the string, whereas if it is not, a "0" will be entered. As proteins that take part in, for instance, the same metabolic pathway or process are likely to evolve in a correlated fashion, proteins with similar phylogenetic profiles are thought likely to be functionally related.

Phylogenetic Tree A tree diagram showing the evolutionary relationships between species, or between genes or proteins in a family, that are believed to have a common ancestor. The edge lengths of the "branches" correspond to estimates of the distance between the entities in evolutionary time. In a rooted tree, there is a unique node at the bottom of the tree that represents the (putative) most recent common ancestor of the entities at the "leaves".

Phylogenomics The use of molecular evolution (phylogeny) to help deduce the function of proteins. These techniques rely on the fact that genes that have diverged in speciation events (orthologs, e.g., human hemoglobin and mouse hemoglobin) are generally closer in function than genes that have diverged in duplication events (paralogs, e.g., human hemoglobin and human myoglobin). Phylogenetic analysis is used to deduce orthologs of genes of unknown function and information from those used in the annotation of the new genome.

pI *See* Isoelectric Point

Plasmid A piece of closed, circular, autonomously replicated, double-stranded DNA. Plasmids range in size between 1 and >200 kb. They are found mainly in bacterial cells, with copy numbers from one to several hundred per cell. They are one of the main means of horizontal gene transfer (and so the transfer of traits such as antibiotic resistance) between prokaryotes; modified plasmids are used in the construction of cloning vectors. In eukaryotes, plasmids may be found in mitochondria and plastids.
Source: Kahl, G, The Dictionary of Gene Technology (2nd Edition).

Pleiotropic A gene is defined as pleiotropic if mutations in that gene have different clinical effects. For example, mutations in the gene for fibrillin-1, located on human chromosome 15, cause Marfan syndrome, but that syndrome may bave strikingly different clinical effects, involving one or more of the skeletal, ocular, and cardiovascular systems. The fibrillin-1 gene is therefore described as strongly pleiotropic. The disease or condition concerned (in this case Marfan syndrome) may also be described as pleiotropic.

Poisson Distribution A probabilistic distribution used in statistical analysis to predict the likelihood of success of a trial in situations in which a large number of trials have been conducted but the probability of success in each individual trial is small. In bioinformatics, it can be applied, for example, to the probability of two sequences chosen at random having a similarity score similar to one that could be expected with sequences that have a common ancestor.

Polarity The phenomenon in which a nonsense mutation introduced into a gene transcribed early in an operon has the secondary effect of repressing expression of nonmutated genes downstream of the mutated gene. The mutation involved is termed a polar mutation or dual effect mutation.
Source: Kahl, G, The Dictionary of Gene Technology, 2nd edition (Wiley-VCH, 2001).

Poly-A Tail A sequence of 60–200 adenine nucleotides added to the 3′ end of most eukaryotic mRNAs after transcription by a template-independent poly(A) polymerase. Its role is to add stability to the mRNA.
Source: Kahl, G, A Dictionary of Gene Technology (Wiley-VCH, 2001).

Polycistronic An mRNA is said to be polycistronic if it contains the transcript of more than one gene, expressed under the control of a single set of transcription factors. Generally, a polycistronic mRNA will contain the transcripts expressed from a single operon (an example being the lac operon in *E. coli*). Most often, the proteins are synthesized separately, but sometimes the entire message will be transcribed into a polyprotein.

Polymerase Chain Reaction A technique used for the selective amplification of a region of target DNA between two annealed primers, by the DNA polymerase-driven extension of those primers in the 5′ to 3′ direction. It initially uses target DNA as the sequence template. The target DNA is first heated with an excess of primers, nucleotides, and DNA polymerase to over 93°C to separate the strands. It is then cooled, and the primers anneal to the original DNA. When the temperature is raised again, the polymerase catalyzes the extension of the primer strands. This produces two new duplexes and the cycle then repeats with the system being heated again to break the hydrogen bonds in the new duplexes.
Source: Savva, R, Techniques in Structural Molecular Biology (Birkbeck College Advanced Certificate) section 4: DNA Technology.

Polymorphism Any specific change in a DNA sequence that is found in some individuals, leading to heterogenicity in a population. This change in genotype may or may not lead to a change in phenotype. Polymporphisms may be in coding or noncoding DNA and may consist of deletions, insertions, inversions, genetic rearrangements, or single base changes. The last named are known as Single Nucleotide Polymorphisms or SNPs.

Polyphyletic In phylogeny, a taxonomic group is defined to be polyphyletic if it is not monophyletic or paraphyletic. A monophyletic group (or clade) consists of a single organism plus all its descendents; a paraphyletic group is a monophyletic group minus one or more distinct subclades. All other groupings (polyphyletic groups) are considered to be unnatural assemblages and are not used in phylogeny, even if there is a phenotype common to the organisms. An example of a polyphyletic group is the group of warm-blooded animals (mammals + bird).

Polyplex A complex formed between cationic polymers and DNA, used in nonviral vectors for gene therapy. Complexes formed by cationic lipids for the same reason are termed lipoplexes. The DNA in both these types of complexes is protected from degradation by nucleases. The linear polymer poly-L-lysine was the first cationic polymer to be used in this type of gene delivery, in 1998.

Polytopic Membrane proteins that are embedded in a cell or organelle membrane, that is, that cross the membrane more than once, and in contrast to single-term

membrane proteins. The term is usually reserved for the alpha-helical type of membrane proteins, which are not found in the outer membranes of Gram negative bacteria.

Population Isolate A population, generally of humans but potentially of other species that has been genetically isolated by geography and lack of outbreeding and that will therefore exhibit less genetic heterogeneity and a higher degree of linkage disequilibrium. Population isolates are very useful for the study of genetic diseases, as mutations may accumulate leading to unusually high prevalence of certain genetic diseases.

Population Structure In population genetics, a population has a structure if its distribution of genetic material is nonrandom. If population structure is undetected, genetic association studies can give both false-positive and false-negative results. Studies of common, multigenic disorders are particularly prone to this problem.

Position-specific Iterative BLAST *See* PSI-BLAST

Position-specific Scoring Matrix A matrix of numbers representing the likelihood of finding a particular base or amino acid at each position of a domain or motif. Each row of the matrix represents a base or amino acid type and each column represents a position in the motif or domain sequence, from the first to the last. The values in the matrix give the log odds of finding each residue at each position. These matrices are used to select regions that are similar to the sequence family modelled. In this method, gaps are not allowed in the motifs modeled.

Positional Cloning The cloning of a specific gene in the absence of a transcript or a protein product, using genetic markers tightly linked to the target gene and a direct or random chromosome walk by linking overlapping clones from a genomic library.
Source: Kahl, G, The Dictionary of Gene Technology, 2nd Edition (Wiley-VCH, 2001).

Positive Inside Rule A rule that states that the segments or loops of a polytopic membrane protein that lie inside the cell (i.e., in the cytoplasm) contain more positively charged residues than those that lie outside the cell, in the periplasm or the extracellular medium. It is often used with hydropathy analysis to predict the number, location and topology of helices in these proteins. TopPred and TMHMM are examples of publicly available algorithms that use this rule in their prediction of transmembrane helix topology.

Posterior Probability In Bayesian probability theory, the conditional probability of an event when empirical data has been taken into account. It may be calculated from the prior probability and the likelihood using Bayes' Theorem.

Posttranslational Modification Any chemical modification to a protein that is made once the protein has been transcribed from its mRNA. There are thought to

be several hundred different post-translational modifications, ranging from cross-linking with disulphide bonds and simple glycosylation and phosphorylation to the covalent binding of complex cofactors. The large number of posttranslational modifications in higher eukaryotes is one reason why their proteomes are much larger than their genomes.

Practical Extraction and Report Language *See* Perl

Preinitiation Complex Trivially, the protein–DNA complex that is assembled prior to the transcription of a gene. In practice, the assembly of all components of the basal transcriptional machinery – that is, the complex of universal nuclear proteins, comprising RNA polymerase II(B) and transcription factors – on the core promoter. The assembly of the preinitiation complex initiates transcription.
Source: Kahl, G, The Dictionary of Gene Technology.

Pre-mRNA Any complete primary transcript of a structural (protein-coding gene) before it is modified to form the mature transcript, which is, in turn, translated into protein. In eukaryotes, pre-mRNA includes transcripts of the exons as well as the introns. Spliceosomes – small organelles made up of protein and RNA – excise the introns and add new noncoding sequences to the 5′and 3′ ends of the DNA.

Premutation Allele An allele of a gene for one of the so-called triplet expansion diseases (e.g., Huntington's disease, Fragile X syndrome) that is toward the high end of the phenotypically normal range. Individuals carrying premutation alleles are at greatly increased risk of passing on a defective allele to their offspring as a result of further expansion. For example, individuals with normal alleles for Fragile X sydrome carry between 6 and 50 CGG repeats in that gene, individuals with premutation alleles between 50 and 200 repeats, and affected individuals often well over 200.

Prevalence The number, or percentage, of cases (generally but not necessarily of disease) present in a population at a given time. This is to be compared with the incidence of the disease, which is the rate of occurrence of new cases of the disease during a given period. A chronic and relatively benign disease such as asthma or arthritis will have a much greater prevalence than incidence.

Primary Structure The amino acid sequence of a protein. In practice, the term "primary structure" is only used as a synonym for sequence in structural proteomics, where it is viewed as the first grouping in the hierarchy of protein structure classification, coming before secondary structure (alpha helices and beta strands), tertiary structure (the fold of a single polypeptide chain) and quaternary structure (the arrangement of chains to form a functional protein).

Primer A short, generally synthetic oligonucleotide that is complementary to part of a larger DNA molecule. Primers form the 3′ end of substrates onto which DNA polymerases can add nucleotides to grow a new DNA chain. Primers are used

as templates in the polymerase chain reaction, and so must be chosen carefully if only the correct sequence of DNA is to be amplified.

Proband *See* Index Case

Profile A way of representing a multiple sequence alignment numerically as a matrix of scores, where each score represents the probability of finding a particular amino acid (or base) at a particular position in the profile. Profiles are often used for classifying protein domains into functional families; they can be used to model DNA sequence alignments, but this is much less common. Some databases, such as Pfam and Smart, use profiles generated using hidden Markov models.

Promoter A region of DNA located upstream of the initiation site, to which RNA polymerase binds to initiate transcription. The sequences of prokaryotic promoters, and of eukaryotic promoters that bind different types of RNA polymerase, have very different sequences. Promoter sequences of a particular type have quite divergent sequences but are characterized by specific short sequence patterns: for example, prokaryotic promoters have the so-called "Pribnow box" sequence at approximately position −10 and eukaryotic promoters that bind RNA polymerase II have the "TATA box" sequence at approximately position −25.

Prophase In the cell cycle, the first phase of cell division during which DNA replication occurs and the chromosomes condense. By the end of prophase, the chromosome pairs are visible under the light microscope, with each pair of daughter chromosomes held together by the centromere. Details of the chromosomes, including abnormalities, can be viewed easily during prophase.

Protease An enzyme that breaks peptide bonds by hydrolysis, thus breaking a protein into peptides. Most proteases are specific, that is, they only break bonds before and/or after particular patterns of amino acids. They have been divided into four main families based on the functional groups in their active sites: the aspartic (or acid) proteases, cysteine proteases, serine proteases, and zinc (or metallo-) proteases. In proteomics, proteases are used to break separated proteins into peptides prior to identification.

Protein Blotting *See* Western Blotting

Protein Fold, Fold Family *See* Fold

Protein Interaction Map A map showing the complex network of interactions between (preferably) a large subset of the proteins expressed in a given cell type at a given time. Protein interaction maps may be generated using two-hybrid technology.
Source: Kahl, G, The Dictionary of Gene Technology.

Protein Microarray An array of probes that is used, by analogy with cDNA microarrays, to determine which proteins are present in a sample. A signal is

detected whenever a protein binds to a probe (which may be, for example, an antibody). Protein microarrays much less advanced, technologically, than cDNA microarrays, but they are now becoming more available.

Protein Profiling Any technology that is used to quantify the expression level of every protein in a tissue sample may be described as a protein profiling technology. It is, essentially, the equivalent in proteomics of the DNA microarray in transcriptomics. The technologies involved are still very much in development but some of the most promising developments involve arrays of spotted antibodies or spotted protein antigens.

Protein–protein Interaction Map *See* Protein Interaction Map

Protein Sequence-structure Space *See* Sequence-Structure Space

Protein Trafficking The processes by which proteins synthesized in a cell nucleus move through a cell to their eventual destinations – within the cytoplasm or an organelle, embedded in a cell or organelle membrane, or secreted from the cell – are generically known as protein trafficking. The endoplasmic reticulum and Golgi bodies are involved in trafficking. Some parts of protein sequences, such as signal peptides, may determine protein location.

Proteinase, Peptidase *See* Protease

Proteoglycan A particular type of glycoprotein (or protein-saccharide conjugate) that is heavily glycosylated, that is, has a high proportion by mass of saccharide. Proteoglycans always consist of a core peptide chain with one or more linear chains of glycosaminoglycans that have sulphate and/or urate groups attached and so are negatively charged. Proteoglycans can have a variety of forms and functions.

Proteolysis The breakdown of a protein into peptides by a protease. There are hundreds of different proteases known, each with a different specificity. In complete proteolysis the protein is broken down into its constituent amino acids. Proteolysis is a natural function that occurs in all organisms, even viruses, but it is also an important part of many scientific analyses. In core proteomics methodologies, separated proteins are broken down into short peptides by proteolysis before mass analysis. Trypsin is one enzyme that is very commonly used for this.

Proteolytic Peptide Peptides that are produced from the digestion of a protein by a protease trypsin are termed proteolytic peptides. The digestion of proteins into fragments, using proteases – most often trypsin (hence tryptic peptide) – is the first step in the procedure of protein identification using mass spectrometry. Trypsin cleaves preferentially after the positively charged amino acids lysine and arginine.

Proteolytic Processing *See* Proteolysis

Provirus Any viral DNA that becomes an integral part of the host cell chromosome and is therefore transmitted from one cell generation to another without lysis of the host cell. A retrovirus that has been integrated into a host chromosome is an example of a provirus. Similarly, a prophage is bacteriophage DNA that becomes integrated into the chromosomal DNA of a bacterial host.
Source: Kahl, G, The Dictionary of Gene Technology.

Pseudogene A nonfunctional derivative of a functional gene that has been inherited by an organism but is no longer needed. During evolutionary history, the sequences of pseudogenes mutate to prevent normal gene expression, for example, by changing promoter region sequences or inserting stop codons. Pseudogenes often retain significant similarity to functional genes in other species and so may be found by homology-based gene finding programs.

Pseudoknot A complex structural interaction between local regions of an mRNA molecule in which one strand of an RNA hairpin is folded back on itself to form, first, a second loop and then a series of base pairs with bases in the loop region of the first hairpin.

PSI-BLAST A form of BLAST in which the database is searched with a profile of sequences rather than a single sequence. The first cycle of PSI-BLAST is a traditional BLAST run; then a profile is constructed from the sequences that match the first sequence and a second cycle of BLAST run using that profile. The process is then repeated until no further sequences are added in a run, at which point the run is said to have conserved. PSI-BLAST can only be run with protein sequences. It is a sensitive method of finding distant homologs.

PSSM, Weight Matrix *See* Position-specific Scoring Matrix

QTL *See* Quantitative Trait Locus

Quadrupole Ion Trap A type of mass spectrometer that has been developed relatively recently and which is both sensitive and versatile. Ions are focused into the ion trap machine using electrodes, and the ions are injected into the trap using an electrostatic ion gate, which pulses open and closed. The ion trap is filled with helium. The kinetic energy of the ions is reduced by collisions with helium atoms, and the ions are trapped with their movement depending on their mass and charge. This allows the precise determination of mass/charge ratios.

Quantitative Trait Locus A genomic region that contains two or more genes or two or more separate genetic loci (map positions) that are known to contribute cooperatively to the establishment of a specific phenotype or trait.
Source: Kahl, G, The Dictionary of Gene Technology.

Quaternary Structure The highest level in the hierarchy of protein structure. Quaternary structure describes the association of more than one separate protein chain into an active protein complex, held together by disulfide bonds or

noncovalent interactions. The association of the four globin chains that make up hemoglobin is a simple example.

Radiation Hybrid Map A dense map of a mammalian chromosome that is created with a somatic cell hybrid technique. Pairs of genes are localized by using two gene-specific primers to amplify a single PCR product, which is then given a radioactive label and hybridized to a panel of radiation hybrid clones. Hybridization indicates that sequences complementary to the PCR product are present in a radiation hybrid clone.

Ramachandran Diagram *See* Ramachandran Plot

Ramachandran Plot A plot of the two torsion angles that, between them, describe the backbone conformation of an amino acid within a protein against each other. Generally phi, the torsion angle about the N–CA bond, is plotted on the x axis and psi, the torsion angle about the CA–C bond, on the y axis. Each amino acid will therefore be represented by a single point. The positions that may be adopted in real proteins are limited by steric hindrance to regions corresponding to alpha helices and beta strands, and a smaller, less populated region corresponding to the left-handed alpha helix.

Read Length The length (in bases) of the small pieces of DNA that are sequenced using Sanger's sequencing techniques and then assembled into longer pieces (and eventually into chromosome and genome sequences) is termed the read length. Computer programs are used to assemble the resulting short sequences. A typical read length in many sequencing projects is 500 bases.

Reading Frame The position from which the codons defining amino acids are read when a DNA sequence is translated into protein. As codons contain three bases, each strand may be read in three ways, so, for instance, a sequence beginning ACGT . . . may be read starting from the A, the C and the G; if the sequence starts from the T, it is read in the first reading frame with the first codon (amino acid) missing. Any gene sequence may be read in six reading frames, three on the forward strand and three on the reverse strand.

Real Time PCR A method of monitoring the amount of a gene produced during the polymerase chain reaction (PCR) using a fluorescent reporter. The amount of fluorescence, which can be measured directly in real time, is directly dependent on the amount of reporter and thus on the amount of amplicon present. It can thus be used throughout the PCR reaction, including the exponential phase, and not just at the end. Real time PCR is sensitive, specific, and reproducible over a wide range of concentration ranges.

Receptor A protein that recognizes another molecule (known as its ligand) and becomes activated when that ligand binds. This activity may take one of a number of forms, including, for instance, conformational changes and binding further molecules. The genomes of free-living organisms contain genes for many hundreds,

if not thousands, of different receptors. They are probably the most important large family of protein drug targets. Drugs targeted at receptors may duplicate the receptor activity (agonists) or block the receptor site preventing activity (antagonists).

Recessive A recessive trait or phenotype is one that is not expressed unless an individual carries two alleles for that trait. Many genetic conditions, including the relatively common cystic fibrosis, are inherited as recessive traits. The existence of dominant and recessive traits was one of the key discoveries of Gregor Mendel, which led to the invention of genetic mapping. It is now known to be an oversimplification.

Recombination Hot Spot The rate of recombination is the rate at which genes are combined in a "child" cell or organism in a different pattern from that in which they are found in either of the parents, for example, due to exchange of DNA between chromosomes. The rate of recombination is not constant throughout a genome, or even an individual chromosome: areas of the genome where recombination is particularly common are known as recombination hot spots. (Similarly, regions where recombination rates are low are known as cold spots.)

Redundancy Generally, a code is said to be redundant if part of it is unnecessary because more than one entity in the code maps on to the same entity in the translation. In molecular biology and bioinformatics, the genetic code is said to be redundant because 64 different codons code for 20 amino acids plus the stop signal. The number of codons that code for individual amino acids ranges between one and six.

Reflectron A mass analyzer, used in mass spectrometry and thence in proteomics, that focuses a beam of ions by reversing the direction of the ions using a retarding electric field. The result is a reduction in the spread of kinetic energies in the ion beam. Types of reflectrons available include single-stage (the simplest), dual-stage, quadratic, and curved-field.

Regular Expression An expression of characters, normally alphanumerics and symbols, that can be matched automatically using pattern matching software. Regular expression matching is commonly used in bioinformatics algorithms, in, for example, searching for amino acid or base patterns. The Unix tools sed, awk, and grep use regular expression matching and the programming language Perl has been optimized for this programming task.

Regulator A gene coding for a protein, known as a repressor, that blocks the activity (DNA binding) of an operator and so prevents transcription of the adjacent operon. In these cases, transcription is often induced by effectors that bind to the repressor proteins, causing changes to the repressor structure and preventing its binding.
Source: Kahl, G, The Dictionary of Gene Technology (2nd edition).

Regulator Gene *See* Regulator

Regulatory Network A network of interactions between genes in which the condition represented by the edges is regulation; that is, two genes (nodes) are joined by an edge if the expression of one regulates the expression of the other. It is self-evident that all regulatory networks are directional.

Relational Database Any database that is built using the relational model is termed a relational database; the best-known commercial example is probably Oracle. A relational model is a logical data structure defined using set theory, so each data item is a member of one or (usually) many more than one set. It can be stated more simply by saying that the data is collected in tables that are linked using keys, so relationships may be modeled across tables.

Replication Competent Simply, a piece of DNA that is able to replicate is replication competent (the opposite being replication deficient). The term is often used in virology and gene therapy; a virus vector for gene therapy that is replication competent will be able to multiply and distribute the introduced gene round the body.

Replication Fork Trivially, a region of genome sequence that starts with an initiation (START) codon and ends with a termination (STOP) codon, and so is translated into protein. A scan of a genome sequence for long ORFs is the first and easiest stage of gene prediction. In practice, the situation is much easier in prokaryotic genomes than in eukaryotic genomes, which are complicated by the extreme length of some genes, the presence of introns, and the necessity of identifying splice sites.

Replicon A segment of DNA under the control of a single replication-initiation locus and behaving as an autonomous unit during DNA replication. Whole plasmids and bacterial chromosomes are replicons. In eukaryotes, the number of replicons tends to increase with increasing genome size and organism complexity (e.g., yeast: 500 replicons, average size 40 kb; mouse: 25 000 replicons, average size 150 kb).
Source: Kahl, G, The Dictionary of Gene Technology (2nd edition).

Reporter Gene A network of interactions between genes in which the condition represented by the edges is regulation; that is, two genes (nodes) are joined by an edge if the expression of one regulates the expression of the other. It is self-evident that all regulatory networks are directional.

Repressor A protein that binds specifically to the regulatory sequence of an operator gene, blocking the movement of the RNA polymerase along the operator DNA and therefore blocking the initation of transcription. The affinity of repressor proteins can be modulated by small molecules that are known as effectors. Many repressors use the helix-turn-helix motif for binding the operator DNA.

Restriction Endonuclease *See* Restriction Enzyme

Restriction Enzyme An enzyme that recognizes specific short target sequences in double-stranded DNA and catalyzes the formation of double-strand breaks.

Restriction enzymes are natural enzymes that protect cells from foreign DNA, and are frequently used in molecular biology. Many different restriction enzymes that are known to recognize different oligonucleotides are used to detect small differences between DNA sequences.

Restriction Fragment Length Polymorphism A polymorphism in which different alleles have different sequences at one or more restriction enzyme cut sits, so that in at least one case, a cut site is added or removed. Cutting the gene sequence using the restriction enzyme concerned will therefore produce fragments of different lengths that can be easily identified. This is a simple and cost-effective way of detecting polymorphisms.

Retrovirus A member of a class of viruses that infect eukaryotic cells and that has single-stranded RNA as its genetic material. After the virus infects a cell, its RNA is reverse-transcribed into a copy of the eukaryotic DNA by the enzyme reverse transcriptase; the integrated (endogenous) retrovirus is termed a provirus. When the endogenous retrovirus is transcribed viral proteins that can associate into new virus particles are formed. Human immunodeficiency virus (HIV), which causes AIDS, is the best known retrovirus.

Reverse Transcription The process, catalyzed by the enzyme reverse transcriptase, by which a double-stranded DNA molecule is transcribed using a single-stranded RNA molecule as a template and a primer. Reverse transposition is used in recombinant DNA technology for the synthesis of cDNA from messenger RNA. It is also the process by which retrovirus DNA is integrated into the eukaryotic genome.

RFLP *See* Restriction Fragment Length Polymorphism

RH Map *See* Radiation Hybrid Map

Risk Factor Any feature that is known to increase a person's chance of developing a disease is termed a risk factor. Risk factors may be lifestyle related (e.g., smoking, obesity, sun exposure) or genetic; well-known examples of the latter are the defective alleles of the BRCA1 and BRCA2 genes, which convey a greatly enhanced risk of developing breast or ovarian cancer.

RMSD *See* Root Mean Square Deviation

RNA Interference The silencing of a specific gene (i.e., the blocking of gene expression) by micro-injection single- or double-stranded RNA that is complementary to the gene to be silenced into cells. It is used in comparative genomics for determining the function of a gene. The injected RNA may in some circumstances be transmitted to germline cells and observed in the experimental organism's progeny. RNA interference is also a natural mechanism for silencing gene expression.

RNAi *See* RNA Interference

Robustness A bioinformatics method or algorithm is described as robust if is reliable and both sensitive and specific, that is, if it predicts features (for example) with few false positives and false negatives. Signals within sequences, defined as patterns or profiles, can also be defined as robust if they predict family members with great accuracy. Generally, increasing the number of sequences that contribute to a pattern or profile will increase its robustness.

Root Mean Square Deviation A measure of the similarity of two structures as the square root of the mean of the squares of the (scalar) distances between selected points. In an analysis of protein structures, the points chosen are self-evidently the atoms; generally, only main chain or alpha-carbon atoms are used. The root mean squared deviation between a model protein structure and the experimentally determined structure of the same protein is a very useful measure of the quality of the prediction; a good model based on a close homolog may have a RMSD of less than 1 Angstrom from the experimental structure.

Rotamer The side chains of amino acids in protein structures are restricted by steric hindrance and therefore preferentially take up certain conformations, which are known as rotamers. Libraries of rotamer conformations are included in programs for three-dimensional protein structure determination and homology modeling, where they are used to suggest likely side chain positions.

RT-PCR *See* Real Time PCR

S-nitrosylation A posttranslational modification of cysteine residues in proteins, involving the addition of a nitro group to the free thiol. Like phosphorylation, it is thought to represent a mechanism for reversible posttranslational regulation of protein activity and consequently of cellular function.

SA *See* Structured Association

SAGE *See* Serial Analysis of Gene Expression

Scale free A network can be described as scale free if the number of connections at each node is distributed very unevenly, that is, if there are a small number of very highly connected nodes. In these networks, the probability that a given node is connected by a given number of connections is determined by a power law. The highly connected nodes are termed the hubs of the network. Many examples of scale free networks can be taken from bioinformatics and other disciplines. Gene networks can be scale free if they contain genes with particularly high numbers of connections.

Schema A term used in several disciplines within computer science to mean a model. For example, in database theory, the schema is the structure of the database,

and in XML, the XML schema defines the structure of the XML documents. The term may also be used to mean a specialized type of ontology.

ScRNA The RNA component of small cytoplasmic nucleoproteins (scRNPs), which are found in the cytoplasm of eukaryotic cells. These scRNPs are RNA-protein complexes that are involved in the splicing of nuclear percursor RNA after its transport into the cytoplasm. They are released from the mRNA before it is translated into protein.
Source: Kahl, G, The Dictionary of Gene Technology.

SDS-PAGE The most commonly used method for separating proteins by molecular mass; one of the two methods routinely used in two-dimensional protein separation. The protein mixture is loaded with a detergent, most often sodium dodecyl sulphate (SDS); this denatures them and confers a negative charge that is proportional to their molecular mass. Migration of SDS-protein complexes on polyacrylamide gels will therefore depend largely on molecular mass.

Second Messenger A signal transduction protein is termed a second messenger if it passes a signal on; that is, second messengers relay signals received by cells following the activation of cell surface receptors to their target molecules. Second messengers also amplify the strength of the signal received. There are three main classes of second messenger: cyclic nucleotides, inositol triphosphate and diacyl glycerol, and calcium ions.

Secondary Structure The second level in the hierarchy of protein structure, describing short stretches of the polypeptide chain that have regular backbone geometry and patterns of main chain–main chain hydrogen bonding. There are two common types of secondary structure, the alpha helix and the beta strand: beta strands associate into sheets, and these sheets – and certain types of tight turn between beta strands – are sometimes included in this structural category.

Segmental Duplication A duplication of a relatively large segment of DNA within a genome sequence is termed a segmental duplication. A high degree of sequence identity within the duplicate regions indicates that the duplication event occurred relatively recently in evolutionary history. The human segmental duplication database contains all duplicates that are greater than 1kb in length and share greater than 90% sequence identity.

Segregation The process by which chromosomes separate during meiosis and mitosis and migrate toward opposite ends of the cell is termed segregation. Segregation occurs during late metaphase and anaphase; the daughter chromosomes are drawn toward the ends of the cell by the microtubules.

Selective Pressure Any change in the environment of a species that leads to some variants being more favored than others, and which therefore leads to those variants surviving and reproducing in greater numbers, is a selective pressure on that species. A classic example, often described in elementary texts, is the rise in

pollution during the Industrial Revolution, which increased the chances of survival of dark pigmented moths.

Selective Sweep An evolutionary event in which a favourable mutation is incorporated into the genome of a species so that it becomes the dominant variant so quickly that alleles that are linked to that mutation also become incorporated into the genome. It can therefore be difficult to identify the allele that is the original target of the selective sweep. The linked genes can be said to have been "hitch-hiked" into the genome.

Selectivity A value that indicates how successful a test is in selecting mismatches (=negatives) from a sample set. It is calculated as the ratio of true negatives (TN; samples that do not have the feature tested for and also fail the test) to all those that do not have the feature (false positives and true negatives); thus specificity = TN/(TN+FP). Specificity ranges between 0 and 1; a perfect test will have a specificity of 1.

Sensitivity A value that indicates how successful a test is in selecting matches (=positives) from a sample set. It is calculated as the ratio of true positives (TP; samples that have the feature tested for and also pass the test) to all those that have the feature (true positives and false negatives); thus sensitivity = TP/(TP+FN). Sensitivity ranges between 0 and 1; a perfect test will have a sensitivity of 1.

Separation Matrix In proteomics experiments, the first step is very often the separation of a protein mixture by mass and/or charge. The protein sample must be dissolved in a compound, such as a gel, before separation can take place; this is the separation matrix. Polyacrylamide is most commonly used (hence the term 2D-PAGE, or 2-G polyacrylamide gel electrophoresis) but others, including polyethylene oxide and hydroxycellulose, have also been used.

Sequence Gap *See* Sequence-Structure Gap

Sequence Profile *See* Profile

Sequence Signature A pattern of residues within a protein that is associated with, for example, a particular functionality or a type of posttranslational modification. Sequence signatures may be long or short, simple or complex, and based on regular expressions, weight matrices, profiles or hidden Markov models. There are many databases containing information on sequence signatures, some of the best known being PROSITE, Pfam, Prints and Smart, and there is one meta-database, Interpro, that collects together information from these and other source databases.

Sequence-structure Alignment The alignment of the sequence of one protein with the structure of another, which may be either homologous or analogous. Sequence-structure alignment is computationally a much harder technique than sequence–sequence alignment. This method is used prior to protein structure

prediction only when no structures of proteins that are clearly recognizable as homologs at the sequence level are available.

Sequence-structure Gap The gap between the number of proteins of known sequence and the number of proteins of known structure, that is, the number of known proteins with no experimentally determined tertiary structure. This gap is still growing, as the structure determination is not keeping place with the number of translated gene sequences coming out of genome projects. It can be narrowed by using homology modeling to predict the structures of proteins that are homologous to proteins of known structure.

Sequence-structure Space A conceptual term used to describe the complete range of protein sequences and structures that have been generated by evolution. Structural proteomics (or structural genomics) program that aim to find unknown folds or solve the structures of unknown sequences without taking much account of the known or predicted function of the proteins are described as searching or filling protein sequence-structure space.

Sequence Tag *See* Peptide Sequence Tag

Serial Analysis of Gene Expression A high-throughput technique for the simultaneous detection and analysis of almost all the genes that are expressed in a given cell at a given time. It is based on the isolation of a short sequence tag, a so-called “SAGE tag” or “diagnostic tag” , from a defined location within the transcript. This tag contains sufficient sequence information to uniquely identify the transcript. The tags are concatenated into a single DNA molecule for sequencing, which aids rapid identification of the tags and therefore the genes from which they are derived. The software used for gene identification with SAGE is also able to determine expression levels.
Source: Kahl, G, The Dictionary of Gene Technology, 2nd edition (Wiley-VCH, 2001).

Short Tandem Repeat, STR *See* Microsatellite

Shotgun Sequencing The determination of the sequence of bases in a complete genome by a method that involves the fragmentation of the target genome or its chromosomes by physical or enzymatic means, the cloning and sequencing of the resulting fragments and the reconstruction of the complete sequence by ordering the fragments. It was used in the publicly funded Human Genome Project.
Source: Kahl, G, A Dictionary of Gene Technology.

Side Chain The part of an amino acid within a protein that is covalently bonded to the alpha carbon atom and, therefore, not part of the continuous main chain of the protein. The genetic code can code for 20 amino acids with different side chains; some other variants can be created by posttranslational modification. The chemical nature of amino acid side chains determines the biochemical properties of the amino acids and, thence, the function of the proteins that they are built from.

Signal Peptide A sequence of generally 15–30 mainly hydrophobic residues at the N-terminal end of a protein. Its function is to target the protein to, and then across, the cell membrane. The protein will then be cleaved at the end of the signal peptide, releasing the mature protein from the cell. There are often positive charges at the far N-terminus of the sequence and negative ones just C-terminal of the cleavage site. Many reliable programs for the recognition of signal peptides and consequent prediction of cellular location are available.

Signal Sequence, Leader Peptide *See* Signal Peptide

Signal Transduction *See* Transduction

Silent gene *See* Pseudogene

Similarity Search A general term for a program for searching a DNA or protein sequence database for sequences that are similar to a test sequence, such as BLAST or FastA. Similarity searches are often termed homology searches, but this is a misnomer: the programs do not explicitly determine homology and this must be inferred by the expert user. There is always a "grey area" where sequence similarity may or may not be statistically significant.

Simulated Annealing A technique used in simulation of macromolecular (or any molecular) structure in which the molecule is "heated up" (i.e., its kinetic energy is increased) during a molecular dynamics simulation. The increase in kinetic energy allows the structure to cross energy barriers. The temperature is then reduced to more physiologically appropriate ones. Repeated simulated annealing experiments allow the molecule to sample more of the available conformational space, increasing the chances of approaching the global energy minimum for the structure.

Single Nucleotide Polymorphism Strictly, a type of polymorphism in which there is a base change at a single position only (e.g., a single A is changed into T, C, or G with the surrounding sequence left unchanged). SNPs occur in coding and noncoding regions, and coding SNPs may be silent (i.e., the codon change does not affect the coded amino acid). Sometimes, small insertions and deletions are included in the same category as SNPs. There are estimated to be 3–30 million SNPs in the human genome.

Single Transmembrane Segment *See* Membrane Anchor

Site-directed Mutagenesis A technique for introducing single amino acid changes into a protein by making specific changes to single base pairs at specific sites in a target DNA. Often used to probe the effects of small changes on the structure or function of a protein.

Small Cytoplasmic RNA *See* ScRNA

Small Nuclear RNA *See* snRNA

Small Nucleolar RNA *See* snoRNA

Smith–Waterman Algorithm A dynamic programming method of aligning pairs of sequences that was adapted from the Needleman–Wunsch method to produce local alignments between the whole sequences. The output will be one or more high scoring sequence segments, with or without internal gaps; gaps at the end of sequences will be removed. The order of matching regions may differ between the sequences. Local alignments and methods that depend on them are often used to identify conserved domains. This method is used in, for example, the EMBOSS local alignment program, WATER.

snoRNA A class of small RNA molecules that are involved in chemical modifications of other RNA genes. They are a component of the small nucleolar ribonucleoprotein complex (snoRNP), which also contains protein. The snoRNA guides the snoRNP complex to the modification site of the target RNA gene via snoRNA sequences that hybridize to the target site.

SNP *See* Single Nucleotide Polymorphism

snRNA An abundant class of relatively small, uridine rich RNA molecules, 100–300 nucleotides in length, which are associated with small nuclear ribonucleoprotein particles. These are found in the nucleus and needed for RNA splicing. The U-RNA family is a strongly conserved family of snRNAs, and its members are designated U1-U10.
Source: Kahl, G, The Dictionary of Gene Technology.

Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis *See* SDS-PAGE

Solvent Accessible Surface In computer-based molecular modeling, a surface around a molecule that is created by virtually rolling a “probe” , usually the size of a water molecule, around the molecule in direct contact with it and plotting the trajectory of the center of the probe. It is larger and “smoother” surface than that built using atomic van der Waals radii and includes those parts of the molecule that are accessible to solvent.

Somatic Cell Any eukaryotic cell other than a germ cell – that is, any cell that is normally diploid. The term somatic gene therapy, or somatic gene therapy, is used for any process by which the genomes of somatic cells are artificially altered. This is safer than altering germ cells, and raises fewer ethical questions, as the germline of the individual is not altered.

Southern Blotting A well-known method for the detection of specific DNA fragments. The DNA fragments are first separated using agarose gel electrophoresis; then the separated fragments are blotted onto nitrocellulose paper. Labeled cDNA probes are hybridised onto the separated bands, which can then be viewed using autoradiography. The similar technique of Northern blotting is used to detect sequences in RNA.

Space Charge Effect An effect that limits the current in a beam of ions of like charge, such as that in a mass spectrometer, arising from mutual repulsion between the ions. Space charge is a consequence of Coulomb's Law, which states the repulsion between particles of like electrostatic charge. In practice, the space charge effect leads to an expansion in radius of the charged ion beam.

Speciation The evolutionary process by which a branch of the Tree of Life bifurcates to form two distinct species. A species is defined as a population of individual organisms that share extremely similar phenotypes and genetic makeup. Where sexual reproduction occurs, male and female individuals of the same species must be able to produce fertile offspring. In prokaryotes, the basic definition of speciation is confused by horizontal gene transfer.

Specificity Window Many procedures in sequence analysis (e.g., hydropathy profiles, dotplots) involve defining a stretch of contiguous residues, calculating a given property and then recalculating the same property for each stretch of residues along the entire sequence. This is known as defining a "window" that is "slid" along the sequence.

Spike-in Control, External Standard, Exogenous Control *See* External Spike-in Control

Splice Acceptor Element A segment of DNA that, if included in a vector upstream of a gene sequence, will be read as a splice acceptor signal (intron–exon boundary) and thus enable that gene to be transcribed if it is inserted within an intron of a transcribed gene. Splice acceptor and donor elements are used, for example, in the technique of gene trap mutagenesis, in which mutated genes are inserted into the mouse germline.

Splice Acceptor Site/Splice Donor Site; Acceptor Splice Junction/Donor Splice Junction; other variants *See* Splice Site

Splice Donor Element A segment of DNA that, if included in a vector downstream of a gene sequence, will be read as a splice donor signal (exon–intron boundary). If a splice donor site is incorporated without a poly-A sequence the gene will only be transcribed if a poly-A signal can be obtained from the endogenous gene. Splice acceptor and donor elements are used, for example, in the technique of gene trap mutagenesis, in which mutated genes are inserted into the mouse germline.

Splice Site In eukaryotic genes containing introns, the junction between an exon and intron at the 3′ end of the intron (splice acceptor site) and the junction between the intron and the exon at the 5′ end of the intron (splice donor site). Both types of splice sites may be identified from consensus DNA sequences.

Spliced Alignment A method of gene finding from a DNA sequence and a set of candidate predicted exons, by searching the set of possible exon chains for the one with the best fit to a related protein sequence. The original exon set is constructed

by considering all possible donor and acceptor splice sites in the genomic sequence. Although this gives an enormous number of candidate exons, most of which will be false positives, the method can be very fast.

Statistical Overrepresentation If certain gene or protein sequence patterns are more frequently found in a longer sequence than they would statistically be expected to be, they are said to be statistically overrepresented in the longer sequence. The opposite phenomenon is termed statistical under-representation. Sequences that are overrepresented may be functionally important whereas underrepresentation indicates that that particular (possibly functional) motif may have deleterious consequences to the organism.

Stem Cells A type of cell that has the potential to differentiate into any one of many different types of cell. Stem cells potentially have very important applications in therapy, particularly for degenerative diseases. The most effective stem cells are embryonic stem cells, derived from early embryos, but the use of these is extremely controversial. Stem cells may also be derived from the umbilical cord or produced from certain types of ordinary cell using chemicals.

Stratification *See* Substructure

Stringency In general terms, the extent to which errors or mismatches are tolerated in a detection experiment or a bioinformatics calculation. Thus, in sequence analysis, stringency is used to set the number of mismatches allowable in a sequence segment defined as a hit (e.g., one that generates a dot in a dotplot). Similarly, a PCR reaction that is relatively tolerant of errors in the resulting sequences is defined as a low stringency reaction.

Structural Genomics *See* Structural Proteomics

Structural Proteomics Any experimental program for solving protein structures by X-ray crystallography or nuclear magnetic resonance that may be described as high throughput, aiming to solve a large number of different structures in a short time, may be termed a structural proteomics (or, confusingly, structural genomics) program. There are two main approaches: either experimentalists concentrate on solving proteins from a particular bacterium or involved in a particular disease, or they attempt to increase the coverage of sequence-structure space by picking proteins predicted to have unknown folds.

Structured Association A method of reducing the chance of finding spurious associations between genes and disease in the large populations that are necessary for the study of the genetics of complex diseases, caused by population heterogeneity. In structural association methods, the details of the population substructure are inferred during the early stages of association testing, so they can be taken into account during the rest of the analysis.

Subcellular Compartment In simple terms, a part of a cell, such as the nucleus, the cytoplasm, or the cell membrane. Bioinformatics tools may be used to predict,

with reasonable accuracy, the subcellular compartment or compartments in which a protein is most likely to be found: to take a simple example, proteins containing hydrophobic segments of a certain length are likely to be found embedded in the cell membrane.

Subgraph Graph theory is used in bioinformatics, for example, to analyze the expression patterns of a group of genes. In graph theory, “A subgraph of the graph G is defined as a graph whose vertex set is a subset of the vertex set of G, whose edge set is a subset of the edge set of G, and such that the map w is the restriction of the map from G.”

Substitution Matrix *See* Weight Matrix

Substructure The genetic heterogeneity in large populations that causes problems such as spurious association (false positives) in population-based studies of genetics of complex diseases is generally known as population substructure or stratification.

Supertree Phylogenetic trees of prokaryotic species based on single genes are limited by, among other factors, the amount of horizontal gene transfer between species. It is considered that phylogenetic trees derived from many genes or even whole genomes may reconstruct the prokaryotic “tree of life” more accurately. The supertree approach to combining phylogenetic information combines single-gene trees rather than sequence alignments. It can be used to combine trees that share only a few species, and it can be used where whole-genome sequences are not available for all species.

Synteny Regions of genomes (generally from quite closely related species) that share at least gene content and often gene order are said to exhibit synteny. Often gene order will be partly disrupted by gene loss, inversion and duplication events. In eukaryotic genomes, large chromosomal regions are conserved throughout chromosomal rearrangements, so regions of one chromosome will be syntenous with parts of different chromosome from related species.

Systems Biology Any one of a number of disparate techniques to study, and specifically to model using mathematics and engineering techniques, the various components of a biological system as the integrated system, for example, modeling a cell using its component molecules, or an organ or tissue using its component cells. It may also be thought of as a mathematical way of thinking of physiology.

Tag SNP Single nucleotide polymorphisms (SNPs) that are known to be associated with haplotype blocks (DNA segments located between recombination hot spots that are usually inherited as blocks). It is possible to genotype individuals for susceptibility to complex diseases using fewer SNPs in total if only the known tag SNPs are used.

Tandem Mass Spectrometry A method for separating and identifying proteins using mass spectrometry (MS) alone, without an initial electrophoresis step. It

uses two mass spectrometry steps, hence the term "tandem MS". The first MS step separates a single protein ion from a mixture. The second step fragments the protein into a series of peptides and analysis of the fragmentation pattern gives ride to short sequence fragments that can be used to identify the protein. Often, electrospray is used for the first ionisation: this technique is known as ESI-MS-MS.

Tandem Repeat In gene sequence analysis, the arrangement of two or more identical sequences within a DNA molecule so that they are close neighbors. These can either be direct (head-to-head) or indirect (head-to-tail), in which case one of the sequences is reversed. The term may also be used to refer to two or more chromosomal segments that are arranged as close neighbors within the chromosome.
Source: Kahl, G, The Dictionary of Gene Technology.

Taq Polymerase An enzyme (EC 2.7.7.7) from the thermophilic eubacterium *Thermus aquaticus* (strain YT 1 or BM), which polymerizes deoxynucleotides with little or no 3′-5′ or 5′-3′ exonuclease activity. It is thermostable (optimum temperature 70–75°C) and allows the selective amplification of any cloned DNA about 10 million-fold with high specificity and fidelity. It is also used to label DNA fragments with radioactive nucleotides, biotin or digoxygenin, and it can be used in Sanger sequencing.
Source: Kahl, G, The Dictionary of Gene Technology.

TATA Box An AT-rich DNA region with the consensus sequence TATAT/AAT/A (in plants, the consensus is TATAATA) most frequently located a few tens of base pairs upstream of the transcription initiation site of eukaryotic genes. It represents the transcription factor binding site; it is essential for accurate initiation of transcription, but not necessary for quantitative expression. It is not found in the promoters of most constitutively expressed ("housekeeping") genes.
Source: Kahl, G, The Dictionary of Gene Technology.

Taxon In phylogenetic analysis, each individual sequence (or species) represented in a phylogenetic tree is described as a taxon (plural: taxa). Each taxon is a phylogenetically distinct unit and appears on the tree as a point at the top level of the tree (so following the analogy further, the taxa are the leaves).

TCS *See* Tentative Consensus Sequence

TDT *See* Transmission-disequilibrium Test

Telomere The ends of eukaryotic chromosomes are known as telomeres. They are usually gene poor, and the telomere tips contain highly repetitive DNA sequences. Telomeres preserve the integrity of chromosomes during replication, and their length tends to decrease as the age of the organism increases. Telomere elongation is catalyzed by the enzyme telomerase.

Telophase The final phase of the cell cycle, during which nuclear membranes reform around each collection of daughter chromosomes and the cell divides in two.

In mitosis, the result is the division of the original cell into two identical daughter cells. Meiosis involves two cell divisions, resulting, after the second telophase, in four haploid gametes, each containing a single copy of each chromosome.

Tentative Consensus Sequence A consensus sequence of amino acids or (more often) of nucleotides which is inexact, that is, where not every position can be completely characterized. Tentative consensus sequences may, nevertheless, be used to search databases, using programs such as Scansite. Tentative consensus sequences are often used to characterise gene promoter regions.

Termination Site, Termination Sequence, Terminator Sequence *See* Terminator

Terminator Genome Sequencing

Tertiary Structure The structure or fold of a single protein (or polypeptide) chain and, originally, the third level in the hierarchy of protein structure, between secondary and quaternary structure. Now it is widely known that protein chains may fold into one or many domains and that each domain may be assigned to a different fold category. The terms “supersecondary structure” and “domain” have now been added to the structural hierarchy between the secondary and tertiary levels.

Tetraploidy A cell is defined as tetraploid if it contains four copies of each chromosome – that is, twice the genetic content of a normal diploid cell, or four times that of a haploid gamete. Mosaic tetraploidy (where only some cells are tetraploid) is quite common in preimplantation diagnosis but very rare in implanted embryos and fetuses. It is not clear whether this is because the condition is embryonic lethal or whether it is, in fact, harmless due to selective growthn of normal cells. Complete tetraploidy is known to be embryonic lethal.

Tetrasomy The presence of four chromosomes or part-chromosomes of the same type instead of two in a diploid genome. It arises following errors of segregation during meiosis. In humans, a full tetrasomy of a whole chromosome would be incompatible with life, but mosaic tetrasomies of part chromosomes (i.e., where the aberration occurs in some cells only) occasionally occur. A mosaic tetrasomy of chromosome 12p has been associated with profound mental retardation.

TF, Trans-acting Factor, Nuclear Factor *See* Transcription Factor

Thermus aqutaticus DNA Polymerase, Taquenase *See* Taq Polymerase

Threading A method for predicting the structure of a protein from its sequence in cases where no obviously homologous proteins of known structure are available. The test sequence is “threaded” through a variety of protein fold templates and the sequence-structure match evaluated using, for example, an energy function. David Jones’ THREADER is a good example of a public domain threading program.

Threshold In any analysis where data is to be classified into two (or more) groups, but where the programs used produce numerical scores, the threshold is the score that marks the boundary between two groups. The threshold is normally set by the user and its value determines the number of false negatives and false positives that the experiment will produce (if the threshold is set too low there will be many false positives; if too high, there will be many false negatives).

Threshold Score *See* Threshold

Time of Flight The most widely used type of mass analyzer in mass spectrometry, at least as that is applied to the identification of separated proteins. The peptide ions are accelerated so ions of like charge have the same kinetic energy; therefore, from basic physical principles, there will be an inverse relationship between the time taken for an ion to travel to the detector and its mass/charge ratio. This enables that mass/charge ratio to be determined as a step toward peptide and protein identification.

TOF *See* Time of Flight

Topology The arrangement and linking of a group of elements; the properties of a figure that are unchanged by continuous distortion (strict mathematical definition). In protein structure, the topology of a protein describes its overall shape and the connectivity between the elements; it is the term that is used to describe the third (Fold Family) level of the CATH protein structure classification. The term may also be used to describe the orientation of a transmembrane helix bundle in the membrane.
Source: Hancks, P, The Collins English Dictionary (Collins, 1986) for strict mathematical definition.

Torsion Angle *See* Dihedral Angle

Toxicogenomics The interaction between genomics and toxicology, or the influence that genetic variation has on drug toxicity. Common genetic variations (SNPs) may lead to drugs causing toxic side effects in some people. For example, drugs may accumulate to toxic levels in people with less efficient variants of enzymes in the cytochrome p450 family of metabolic enzymes.

Transactivation It often happens that transcription factor–DNA complexes must be stabilized by the binding of other proteins before mRNA transcription can take place. This stimulation of transcription by a transcription factor and its associated adjacent proteins binding to a promoter region is termed transactivation.

Transcript Capture A technique of measuring and identifying the mRNA content of a cell by harvesting (or capturing) the transcripts present in that cell before they can be degraded. This method can be used to monitor gene expression within and between cell types and to identify splice variants, including those resulting from exon-skipping events.

Transcription Factor A protein that binds to the recognition sequence of a DNA molecule, upstream of a coding sequence, and facilitates transcription initiation. DNA-dependent RNA polymerases bind to the transcription factor–DNA complex that activates RNA poymerization. Transcription factors may also bind to upstream regulatory sequences or even to sequences within the coding regions.

Transcription Profiling The use of microarray or similar technologies to determine a profile of the mRNA molecules (transcripts) present in a cell type under particular conditions and at a particular time. Very briefly, short pieces of cDNA molecules are immobilized on a grid, mRNA from the cell type under study is tagged with fluorescent probes and hybridized to the stationary cDNAs. Fluorescence at a spot indicates the presence of an mRNA that is complementary to that cDNA molecule.

Transcription Unit The complete DNA sequence between the transcription initiation site and the transcription termination site, both sites as recognized by the DNA-dependent RNA polymerase. A transcripton may contain one gene or more than one; in the latter case the message produced is polycistronic, but only in prokaryotes is this ever translated into a single polyprotein.
Source: Kahl, G, The Dictionary of Gene Technology.

Transcriptome By analogy with "genome" , "proteome", and a large number of other "omes" , the set of mRNA transcripts that is present in a cell. Unlike the genome, but like the proteome, an individual organism's transcriptome is not constant but varies according to cell type, developmental stage and conditions (e.g., a disease state or the presence of a drug). However, the correlation between the transcriptome and the proteome is not particularly strong and the proteome is, self-evidently, more indicative of the metabolic processes that are taking place in the cell.

Transcripton *See* Transcription Unit

Transduction The transmission of a signal from the exterior surface of a cell or organelle into the interior of that system, leading to an internal response to the external signal. Signal transduction is initiated by a ligand binding to a surface receptor and carried out by a cascade of enzyme activity.

Transfection The uptake of viral nucleic acid by bacterial cells or speroplasts, resulting in the production of a complete virus. Alternatively, the integration of foreign DNA into the genome of cultured animal or plant cells via direct gene transfer.
Source: Kahl, G, The Dictionary of Gene Technology.

Transgene Any gene that has been transferred from one organism to another organism of a different species. The transformed organism is known as a transgenic organism. Transgenes may not be expressed, or may be expressed at very low levels, in the host organism. Transgenic modification must be strictly controlled by law.

Translation The process of protein synthesis at the ribosome is termed the translation of the RNA sequence into protein. The sequence of the resulting polypeptide is determined from that of the original RNA molecule via the genetic code. Although one code is used almost universally, alternate codes are used in mitochondria and some groups of organisms. The process of protein translation is a complex one in which the ribosome operates as a molecular machine.

Translocation The stepwise, codon-to-codon advance of a ribosome along a messenger RNA sequence with simultaneous transfer of the peptidyl-RNA from the A site to the P site of the ribosome. Each step exposes an mRNA codon for base pairing with its specific tRNA anticodon. Alternatively, any change in the position of a specific chromosome segment either within the chromosome ("shift") or to another nonhomologous chromosome (interchromosomal translocation).
Source: Kahl, G, The Dictionary of Gene Technology.

Transmission-disequilibrium Test A test of the role of genetic factors in disease states in which the genotypes of cases of a disease are compared to those of their parents to discover whether a genetic variant or marker is inherited by cases at frequencies higher than would be expected using classical Mendelian genetics. If the allele or marker is, in fact, transmitted in excess of what would be expected in cases of disease, it indicates that the allele is a risk factor for the disease.

Transposable Element, Mobile Element *See* Transposon

Transposon Generally, any sequence or segment of DNA that can change its location within a genome. However, in the strictest usage the term "transposon" is restricted to use in prokaryotes, with similar sequences in eukaryotes being termed "transposon-like elements". A transposon is flanked by short inverted repeat sequences; it encodes an enzyme that catalyzes its excision from its first site and insertion in a new site. Transposons can be used in the construction of certain types of cloning vectors.

Trans-splicing The ligation of exons from two different mRNA molecules to form one messenger RNA with a different combination of coding sequences that will therefore be translated into a different protein. Much of the complexity of vertebrate proteomes arises from the formation of multiple proteins from a simple gene set using mechanisms such as this one.
Source: Kahl, G, The Dictionary of Gene Technology.

Trinucleotide Repeat Expansion A sequence of three bases that is repeated a large and variable number of times at a specific position of a chromosome (and thus, a special case of microsatellite). The repeated sequence may occur in coding or noncoding DNA; where it occures in coding DNA, it gives rise to an amino acid repeat. Several rare single gene disorders arise from an expansion in a trinucleotide repeat. The best known of these is Huntington's disease, where the expansion is of the trinucleotide CAG in a coding region and therefore of the amino acid glutamine.

Trisomy The presence of three chromosomes of the same type instead of two in a diploid genome. It arises when one chromosome fails to segregate during meiosis. In humans, most trisomies are incompatible with life, but people with trisomy 21 (three copies of chromosome 21) can lead fairly satisfying lives, albeit with the mental and physical disabilities characteristic of Down's syndrome. Babies born with some other trisomies, including trisomy 13, may live a few months.

Tropism Tropism, in general, is the involuntary response of an organism to a stimulus. Viral tropism is the interaction between the virus and its host, and it can hamper gene therapy with viral vectors. Scientists are developing methods of modifying and decreasing tropism in viral vectors.

Tryptic Peptide *See* Proteolytic Peptide

Tumour Suppressor Gene Genes that code for signal transduction proteins that send signals that inhibit cell growth and division are known as tumor suppressor genes; they are, therefore, the opposite of oncogenes. When tumor suppressor genes are mutated they may lose their functionality, leading to a loss of control of cell proliferation and, potentially, the development of cancer.

Twilight Zone The accuracy of a model of the three-dimensional structure of a protein built from the structure of a similar sequence will depend on the percentage identity between the sequences. Between about 10% and 25–30% sequence identity, the degree of homology (evolutionary relationship) between the sequences cannot be inferred from the identity value alone, although a clear evolutionary relationship may be deduced from other means. In this case, the proteins are said to fall within the twilight zone, and successful homology modeling may not be possible.

Two-dimensional Gel Electrophoresis *See* 2D-PAGE

Two Hybrid *See* Yeast Two Hybrid

Underspliced Transcript If a pre-mRNA transcript is processed by fewer splicing events than would be required to produce the correct, mature mRNAs, that transcript is described as underspliced. This frequently occurs in viral processing; for example, the HIV virus originally produces a single transcript that is processed into some 30 mRNAs. If the transcript is underspliced, fewer mature mRNAs will be produced and the daughter virions will not be able to assemble.

Uniparental Disomy A genetic condition in which two copies of one or more chromosomes are inherited from one parent and none from the other parent, with the chromosome number remaining normal. The condition may be termed maternal or paternal uniparental disomy depending on the parent that provided both chromosomes. It may be silent (with the child appearing phenotypically normal); alternatively, it may result in developmental defects due to abnormal imprinting.

Untranslated Region, Untranslated Sequence *See* UTR

Upstream Toward the 5′ end of a DNA sequence. The term is most often used for sequence that is located 5′ of the coding sequence of a gene. The 5′-most region of that sequence that is transcribed into the pre-mRNA (but not translated into protein) is known as the 5′-untranslated region (5′-UTR). Control sequences that are bound by transcription factors and that are located upstream of the 5′-UTR are never transcribed into mRNA.

UTR Portions of sequence that are transcribed into RNA, but not translated into protein. Each transcribed gene includes a sequence upstream of the start codon (leader sequence, 5′ untranslated region or 5′-UTR) and one downstream of the stop codon (trailer sequence, 3′ untranslated region or 3′-UTR). Untranslated regions contain sequences that control expression. The poly-A tail is not part of the trailer sequence as it is not part of the original gene: it is added to the 3′ end of the trailer sequence after transcription.

Vector A plasmid or phage cloning vehicle specially constructed to achieve efficient transcription of a cloned DNA fragment and translation of its mRNA into protein. Cloning vectors often contain an expression cassette including a highly active promoter, to aid efficient gene expression. Genome sequences may be contaminated by vector sequence; this may be tested for by comparing the new genome sequence with a database of known vector sequences.
Source: Kahl, G, The Encyclopedia of Gene Technology.

Viral Capsid *See* Capsid

Viral Tropism *See* Tropism

Virulence Factor A gene or gene cluster in a microbial pathogen which increases the virulence of that pathogen to its human or animal host. Many virulence factors have been transferred between bacterial species via plasmids. Some very pathogenic bacteria, such as Vibrio cholerae (the causative agent of cholera) have been shown to contain “systems” of toxicity comprising 40 or more protein toxins and virulence factors.

Weight Matrix A statistical model used in sequence analysis in which each position in the sequence is modeled independently of the others. Typically, a score is allocated to each amino acid (or base) based on the likelihood of that amino acid (or base) being found at that position in the feature under consideration.

Western Blotting A technique, analogous to Southern blotting, for the detection of specific proteins that have been separated by 2D gel electrophoresis or a similar technique. The proteins are transferred to a membrane and visualized using specific radioactively labeled, fluorescence-conjugated or enzyme-conjugated antibodies.
Source: Kahl, G, The Dictionary of Gene Technology.

Whole Genome Shotgun, WGS *See* Shotgun Sequencing

Window *See* Specificity Window

Workflow In general terms, workflow is simply how a work procedure is organized. It is used in bioinformatics largely in applications that consist of a large number of relatively small or simple calculations, with later analyses chosen as a result of earlier ones. In these cases, workflow software may be used to automate at least some of the tasks and decisions. Examples might include the prediction of protein localization or immunogenicity from a sequence.

Yeast Two Hybrid A relatively new, powerful technique for detecting interactions between proteins. It is based on the dual-module composition of yeast transcriptional activators such as GAL4. The protein under test is linked (hybridized) to the DNA binding domain of GAL4, and a library of proteins to the GAL4 activation domain. DNA transcription only occurs if there is an interaction between the test protein and one of the library proteins.

z score *See* z-score

z-score A statistical parameter defined as the difference between a score achieved for one variable in a set and the mean score for the whole set, divided by the standard deviation in the scores. A high z-score indicates that the score for a particular variable is an outlier that may be of statistical significance (e.g., indicating a match). This concept is often used in protein fold recognition, where it is used to calculate whether there is a statistically significant match between a sequence and one particular fold when it is compared to the whole database of possible folds.